Methods used to evaluate growth modification in Class II malocclusion

Joan Frances Camilla Tulloch, BDS, FDS, D Orth, RCS (Eng),* William Medland, BDS, MSc, and Orhan Celil Tuncay, DMD*****
Chapel Hill, N.C.

The methods used to study growth modification in orthodontic patients can have considerable impact on the conclusions that may be drawn. Because of the large "between-patient" variation and small mean changes usually observed, apparent differences in response may sometimes be more attributable to study design than to treatment effectiveness. A systematic review of four major orthodontic journals (1980 to 1987) identified 50 studies reporting treatment of young patients with Class II malocclusion. Variables defined to classify the studies included appliance systems, patient selection, comparison groups, research design, data collection, analysis, and reporting. The appliance systems most frequently investigated were the function regulator and the activator, used with and without headgear. Only 11 (22%) studies were prospective, and random assignment to alternative treatments was never used in this sample. Comparison groups used in 76% of the studies were untreated Class II patients (n = 18) and/or patients with alternate appliance systems (n = 17). In only 24% of the reports were groups tested for pretreatment equivalence. Few studies reported fully how patients had been selected, how decisions had been made to discontinue or change treatment, or whether patients had been lost to study. While most studies reported "p values," in only four were alpha levels adjusted for the number of tests (type I error), and no study included a post beta estimate (type II error). Age, sex, maturation, and duration of treatment were usually reported but seldom adjusted for in the analyses. Given the multiple indices of treatment effect, the generally small sample sizes, weak research designs, and incomplete reporting of important data, we cannot yet conclude whether orthodontic treatment influences the growth of Class II patients. (AM J ORTHOD DENTOFAC ORTHOP 1990;98:340-7.)

Class II malocclusion is one of the most common problems seen by orthodontists. Even though successful treatment for this condition has been demonstrated repeatedly, clinicians and patients continue to seek either simpler or better methods of correcting the occlusion while maintaining or improving facial appearance. A large proportion of Class II patients have a significant skeletal imbalance, and much of orthodontic treatment is aimed at correcting or masking this discrepancy. Treatment that has the ability to alter a patient's facial growth is of great interest. Numerous investigations have been carried out over the years to evaluate the possibilities of growth modification with orthodontic appliances; however, the results have generally been equivocal. While some studies have reported significant effects, others have failed to demonstrate any consistent change. Despite the volume of literature, some basic questions remain to be answered. Do functional appliances produce any measurable change in the amount, direction, and duration of growth, or do they correct a Class II occlusion by some other means? Are the changes predictable and significantly different from those that would occur either with no treatment or with a conventional appliance system? Clear answers to these questions would represent a significant advance in our knowledge.

The usual way to study these questions is to compare growth in patients treated with "growth-modifying appliances" with growth in patients who have either received no treatment or have been treated with other appliances. Although growth modification has been attempted since the turn of the century, the precise effects of such treatments remain unclear. Little is known about how such appliances work, which tissue systems are influenced, or the magnitude and likelihood of their effects. The difficulty of establishing the relationship between treatment and growth changes is considerable and stems from the variations in timing, amount, and

Supported by the Robert Wood Johnson Foundation Dental Services Research Scholars Program, and by National Institutes of Health Grants DE5537 and DE08708.

*Department of Orthodontics, School of Dentistry, University of North Carolina.

**Brisbane, Australia.

***Department of Orthodontics, University of Mississippi Medical Center.

8/1/16230

direction of growth seen both with and without treatment, the changes in the different tissue systems, and the different appliance regimens used. Independent of treatment, the expected growth increments depend on a patient's maturational age and sex and the length of time considered.

Although not unanimous, a consensus seems to be appearing in the orthodontic literature that, in spite of large variations within groups and small mean differences, functional appliances do produce modest (and sometimes statistically significant) amounts of increased mandibular growth, decreased maxillary growth, and dental movement during the correction of Class II malocclusion.[1] However, no matter how numerous the reports or how uniform the conclusions drawn, poorly performed, ineptly analyzed, and inadequately reported studies do not establish a sound basis from which to make treatment decisions. A review of the conduct of recent published studies was undertaken to determine whether methodologic considerations—including sampling procedures, selection of comparison groups, choice of research design, consideration of confounding variables, and statistical methods—might suggest that conclusions either supporting or refuting a growth modification effect should be regarded with caution.

METHOD

A systematic review of four of the major English-language orthodontic journals—THE AMERICAN JOURNAL OF ORTHODONTICS AND DENTOFACIAL ORTHOPEDICS, *The Angle Orthodontist, The British Journal of Orthodontics,* and *The European Journal of Orthodontics*—was carried out. Journals published between the years 1980 and 1987 were reviewed to identify articles that reported treatment of Class II malocclusion with some appliance system designed to effect changes in skeletal and dental relationships, possibly by growth modification. Only articles that reported original data obtained from radiographs of growing patients were selected for further review. Articles that were purely technique-oriented, case reports in which no attempt had been made to quantify and aggregate treatment changes, and articles reporting data on fewer than 10 subjects were excluded. The selected articles were then differentially photocopied to remove all identifying characteristics and were scored independently by each author. The scores were compared, and where disagreement occurred the articles were reread and discussed until conflicts were resolved. Four groups of variables were defined to describe the important characteristics of each study.

The first group related to the type of appliance system used, the number of patients investigated, and the selection of control or comparison samples. Since there is considerable overlap between the design of many orthodontic appliances and continued debate about how each works, the appliance types were subdivided to keep the groups as discrete as the reports would allow. Many different control or comparison groups can be used to evaluate the effects of treatment on growth, and the selection of comparison group can have a great impact on the value of the information obtained and the degree to which such information lends itself to generalization. Studies were divided into those that made comparisons with historic population-based data sets, such as the Michigan or Burlington growth studies; those that considered untreated patients, either concurrently enrolled or previously observed; those that compared alternate appliance systems; and those that used no control groups. Within each of these groups, comparisons could be made against "normal" (Class I) occlusion or Class II occlusion, or against a mixed sample of patients.

The second area of review related to research designs and methods of data collection. All studies used a longitudinal research design, but there were differences in the methods of patient selection and data collection. The design of a specific study was categorized as "prospective" if the patients had been enrolled at the start of the study and were followed to determine the outcome of treatment, with the treatment and data collection carried out prolectively,[2] generally in accord with some predetermined protocol. Where possible, the assignment method was determined and recorded as being based on clinicians' beliefs and preference, by random allocation, or by non-random methods such as choosing consecutively enrolled patients. If the data were obtained from the records of previously treated patients, reflecting only the information available in the general treatment record, then the data collection was classified as retrolective.[2] With this type of data collection, the methods by which the individual cases are identified and selected for analysis can generate powerful biases that have a considerable impact on the conclusions and inferences that can be drawn and the comparability of different studies. Where possible, the selection criteria were categorized as being based either on the patient's initial condition or on the patient's having completed a certain phase of treatment. A number of additional exclusion criteria were also considered: the degree of patient cooperation, the availability of adequate records, and evidence of successful correction of molar relationship.

The third broad area considered a number of additional factors that can have a subtle bearing on the

Table I. Number of studies reporting data on different appliance systems for the correction of Class II malocclusion, by sample size

Appliance type	Sample size		
	10-15	16-30	>31
Fränkel	2	8	4
Activator	2	8	9
Bionator	—	2	—
Herbst (± headgear)	1	3	3
Activator + headgear	—	—	2
Headgear only	—	1	8
Headgear + fixed appliance	3	3	2
Fixed appliance only	1	2	1
Other	1	4	1

interpretation of these studies, such as where and by whom the treatment was carried out—whether in a private office by one clinician, in a dental school, or in multiple locations. This variable can have an important influence on the timing and acquisition of serial records, the goals and intensity of treatment, and the comparability of some radiographic measures. Second, because the length of time that a patient is observed is an important variable in determining the expected amount of growth, the "stopping rule"—that is, the criteria by which it is decided to discontinue or change treatment—becomes an important issue and should be presented clearly. For example, was the treatment discontinued when certain morphologic changes had occurred, after a certain time had elapsed, when the patient or clinician tired of the treatment, or when all hope of change had been abandoned? The response to treatment is, at minimum, a function of the patient's maturational level, sex, age at start of treatment, and the duration of treatment. In addition, the time between the records and the different phases of treatment should also be considered. For example, can the observed changes be attributed to the appliance, or is a substantial proportion of the change attributable to a period of growth when no "treatment" was actually under way? These variables, at a minimum, should be reported in every study and, if possible, the data should be handled in a way that can control for, or allow analysis of, their impact. Each study was scored as to whether these variables had been mentioned and whether any attempt had been made to allow for or analyze their influence.

The final grouping relates to statistical analysis, which posed some special problems in these studies. The following were important considerations, we believed: (a) whether there had been testing for pretreatment equivalence between groups, (b) whether "alpha" levels (the probability of rejecting a null hypothesis when there is, in fact, no difference between groups, or type I error) had been adjusted to account for the number of variables tested,[3] (c) whether "post-beta" estimates had been made to determine the likelihood of type II error (the probability of accepting the null hypothesis when a real and important difference does exist),[4] (d) whether statistical methods that can adjust for multiple and interdependent measures had been used,[3] (e) whether the data had been analyzed as longitudinal or cross-sectional data, and (f) whether some assessment of the measurement error had been included.

RESULTS

The results reported here are based on the 50 studies identified. Each study is cited as the unit of analysis, though it was clear, even in the "blind" review, that data from the same group of patients were, in several instances, being presented in more than one article.

The appliance systems most frequently reported on during the 1980s were the function regulator and the activator, used either with or without headgear (Table I). The effect of headgear alone continues to be investigated in several studies that report data on large samples. Surprisingly few studies considered the growth-modifying potential of conventional fixed appliances. The most common comparison made in this period was between treated patients and persons with Class II malocclusion who were concurrently enrolled but untreated (Table II). Comparisons with alternative appliance systems are also frequently reported, and in several studies both comparison groups are used.

The research design most frequently adopted (in 74% of the studies) makes use of data from previously treated patients. Only 11 (22%) of the studies employed prospective data collection, and in two studies it was unclear exactly how data had been obtained. The method by which patients were identified and then allocated to treatment was unclear in seven of the prospective studies. Consecutive patient assignment was claimed in four studies, although three of these also stated that only patients with a "good growth potential" had been included. In the retrolective studies, the criteria for sample selection were most generally reported as being based on the patients' initial condition, though a number indicated that additional inclusion criteria such as adequate records or a history of compliance had also been required (Table III).

Inclusion of information on the location of treatment and management of patients was not particularly well reported in this sample (Table IV). In 21 (42%) of the reports it was unclear where or by whom patients had been treated. In 30 (60%) of the studies the criteria by which decisions had been made to discontinue or change treatment were unclear. In 34 (68%) of the stud-

Table II. The selection of comparison groups in the studies reviewed (some studies used more than one comparison group)

	Type of occlusion in comparison group		
Selection of group	*Class I*	*Class II*	*Mixed sample*
Historic data set	2	3	4
Previously observed untreated patients	—	2	1
Concurrently enrolled untreated patients	1	18	1
Alternate appliance system	—	17	—
No control used	12	—	—

Table III. Selection and inclusion criteria for individual cases in studies using retrolective data collection (n = 37)

	Sample selection		
Inclusion criteria	*Initial condition*	*Completion of phase of treatment*	*Unclear*
All	9	—	—
Cooperation	5	1	—
Adequate records	4	1	—
Successful molar correction	8	1	—
Other*	3	1	—
Unclear	7	—	3

*"Nonextraction and no other appliances," "reasonable amount of time," and "random."

Tables IV A-C. Reporting on locations and management of treatment in the studies reviewed (n = 50)

A. *Location of treatment*

Private office	*Dental School*	*Multiple locations*	*Unclear*
8	10	11	21

B. *Reason for ending treatment*

Elapsed time	*Morphologic change*	*Unclear*
8	12	30

C. *Patients lost to study*

Number mentioned	*Reason analyzed*	*Not applicable*	*Unclear*
13	0	3	34

ies no mention was made of whether and why patients who had started in treatment had been lost to further study. In a similar fashion, while most studies mentioned the important variables of age, sex, and duration of treatment, analyses did not, in general, make full use of this information (Table V).

The statistical handling of data in these reports was not uniform (Table VI). While the majority of studies

Table V. Considerations in statistical analysis in the reviewed studies

	Yes	*No*	*N/A*	*Unclear*
Groups tested for pretreatment equivalence	12	29	9	0
"*p* values" reported	44	5	1	0
Alpha levels adjusted for number of tests	4	41	5	0
Post-beta estimate	0	50	—	0
Longitudinal data analysis	40	7	—	3
ANOVA/correlation/regression	15	33	2	0
Withdrawals considered in analysis	1*	41	8	0
Measurement error	24	26	—	0

*One study compared pretreatment variables of those who complied with those who did not.

Table VI. Studies reporting and controlling for variables that could modify or confound the effects of treatment (n = 50)

Variables	*Reported*	*Adjusted in analysis**
Age	48	20
Sex	46	15
Maturation level	17	14
Treatment duration	48	31
Time between records and treatment	4	2

*"Adjusted in analysis" included studies in which it was obvious that efforts had been made to select the comparison groups to "match" the treatment group on important variables.

analyzed the data longitudinally by methods that considered individual patient variation rather than the "group-mean" changes, surprisingly few tested the groups for pretreatment equivalence. Although "*p* values" were almost always reported, very few studies considered adjusting the "alpha" levels to reflect the number of variables tested, thereby ignoring consideration of type I errors. No studies performed a "post-beta" estimate to determine the probability of a type II error. Fewer than half of the studies used the more powerful statistical analyses such as analysis of variance, correlation, and regression that can help with problems of multiple, interdependent, and repeated measures.

DISCUSSION

This review was undertaken in an attempt to characterize and evaluate the quality of studies investigating growth modification in Class II patients. We hope that our findings may suggest how differences in results may easily be attributable to study design, analysis, and reporting rather than to treatment effectiveness, and we wish to emphasize how more attention to methodologic issues and uniformity in reporting data could improve the practitioner's access to information and his or her ability to interpret apparently contradictory findings.

It appears from this review that a wide variety of appliances continues to be used in an attempt to modify growth. In the absence of any specified selection criteria, it is difficult to conclude whether this diversity reflects very specific indications for each appliance, that all appliances work equally well, or that no one appliance has yet been demonstrated to have a consistent effect. It is interesting to note that each of the appliances examined has been used, with varying degrees of enthusiasm, for many years—a finding which suggests that, at the very least, previous attempts to compare or evaluate effectiveness have been less than conclusive. If all appliances have equivalent effects on Class II correction, then perhaps additional measures such as ease of use, reduction in treatment burden or clinician time, or improvement in the quality of the result should be included with the more traditional measures of cephalometric change.

Sample size

Consideration of sample size in these studies seldom goes beyond a simple statement of the number of patients in each group. Determining the number of persons needed to establish treatment effectiveness depends on the magnitude of the treatment effect that is considered worth demonstrating, the variability in response, and the alpha and beta levels selected as acceptable probabilities of a type I error (claiming that a difference exists when it does not) or a type II error (missing a real and important effect). While it is generally true that large effects can be detected with few subjects, particularly when there is little variation in response, the changes that orthodontists seek to demonstrate are small in comparison with the variability seen among patients, and they therefore mandate the study of large samples. Published reports that give few details of prior planning make it difficult to determine to what extent the hazards of insufficient sample size were considered before the researchers concluded that there was no difference in treatment effect. None of the studies we reviewed provided reasonable assurance that a meaningful difference in treatment effect would not have been missed simply on the basis of insufficient sample size. For studies that are based on more than one hypothesis, the determination of sample size becomes complex. It may be necessary to calculate the sample size for each hypothesis and then select the maximum size computed. While a sample may have sufficient power to determine an effect for one particular parameter, another variable with a smaller mean treatment effect or larger variance might require a sample several times larger to reveal a statistically significant difference. Given the difficulty of following patients over prolonged periods, it would seem that more attention should be paid to identifying those measures for which an effect can be tested within the sample sizes available. An alternative would be to report the data in a complete and meticulous fashion so that small but similar studies may be combined in a meta-analysis[5-7] to improve the statistical power.

Sample selection

The selection and composition of comparison groups has a great impact on the conclusions that may be drawn, the appropriateness of generalization, and the value of a study. Several types of control have been used in these investigations. Comparison of groups of patients treated with different appliance systems provides a contrast of treatments but not a determination of the effectiveness of treatment versus no treatment. To establish the effect of an appliance on growth, the best practical comparison would be with a concurrently enrolled group of untreated persons with Class II malocclusions. However, there are obvious ethical difficulties in conducting prolonged longitudinal studies that follow untreated persons over a period when treatment is normally provided. It has been argued that if the course of treatment with standard therapy or the pattern of growth with "no treatment" is already known, then controls are unnecessary; the effectiveness of any new treatment can simply be evaluated against the already existing knowledge. However, this point of view can be justified only if the expected changes are known rather precisely and the variability is small. Our current knowledge concerning growth in general, or specific to persons with Class II malocclusion does not suggest that these criteria are met.[8,9] Comparisons against previously observed patients or historic data sets such as the Michigan, Burlington, or Bolton studies should be regarded with caution for another reason. Growth data collected some years ago may no longer be valid for today's population, since secular trends may occur in the craniofacial region as well as in such dimensions as height, weight, and onset of puberty. Contemporary controls need to be compared to these historic data sets to determine whether they are, indeed, cohorts that can be considered as valid controls for more recent studies.

Research design

In general, the most common research design used during the 1980s was an analysis of previously treated patients whose records had been selected for further study after treatment had been completed. Such designs are inherently weaker than prospective studies, since they inevitably identify samples that reflect the clinicians' beliefs about the type of patient most likely to respond favorably to treatment. In addition, and perhaps more important, patients who do not respond or cooperate are usually transferred to another treatment method, or treatment is discontinued. This selection process leaves only those patients who showed the expected desirable responses. Since growth is known to affect treatment, this process also obscures the question of whether treatment or optimal growth caused the effect. Selecting only those patients who completed treatment, successfully or not, or only those patients who were treated with a single appliance and for whom a complete set of records is available, certainly biases a study in favor of those who experienced favorable growth and begs the question of whether or not the treatment caused the growth. Even when a deliberate attempt is made not to select patients on the basis of the success of their treatment results, followup of patients who did not respond to treatment are seldom available for further consideration. This loss to followup is probably more common with respect to patients treated in private practice, where records are perhaps less likely to be obtained according to a research protocol and are produced, instead, in response to clinical needs. Information on how many patients were lost to followup, and why, is infrequently given, and the probable impact of this loss on the results was never analyzed during the period covered by our review.

Perhaps the greatest potential source of bias in these studies arises from the possibility that patients in alternate groups may differ in some important way. Testing for general pretreatment equivalence, even though extremely important, does not guard against differences in some previously unsuspected and therefore unmeasured variable. Even when comparison groups are carefully selected to match on initial conditions, differences in age, sex, maturation level, and duration of treatment can confound the results. This is particularly true when controls are chosen from patients treated at another time, by another doctor, in another setting. Prospective studies, for all their difficulty, are considered more powerful than retrospective studies when comparisons between treatments must be made. However, even when experimental and control groups are managed by the same clinician, in the same place, during the same time period, by a previously determined treatment protocol, any assignment to therapy that is not random can introduce dissimilarities in the groups. Both patients' preferences and clinicians' judgments make it likely that while some patients will be more actively recruited than others, some patients will not be considered for all alternatives. The randomized prospective clinical trial remains the only method of treatment assignment that provides strong assurance, but still no guarantees, about the comparability of groups. A randomized trial has the advantage of allowing comparisons of treatments among groups for which pretreatment equivalence has been statistically achieved by a process that avoids selection bias and defines the groups so that differences in their observed experience can be attributed to treatment effect. There have been many examples in the health sciences where opinions based on the observation of a few patients or studies that used relatively weak design have been responsible for suggesting spurious associations and promoting useless treatments, which have later been discarded after further research with stronger designs and better comparison groups.[10,11]

No matter what research design is used, variations in the type and timing of the records obtained, the diagnostic criteria used, the management of patients, and guidelines used to determine when to discontinue or change treatments are all important issues that must be clearly presented if studies are to be compared. These data are seldom reported and even less frequently analyzed. However, such variables have a real impact on the accuracy of estimates of incremental growth that may result from treatment. Without clear delineation of the protocol for treatment, data gathering, and analysis, it becomes almost impossible to make meaningful assessments of the effects of different therapies. Careful consideration of treatment conditions and data collection could simplify comparisons and perhaps enhance the value of similar small studies by allowing combination of data.

CONCLUSION

The primary reason for this review was to determine whether endorsement of the gradually emerging consensus that orthodontic appliances might produce modest influences on growth in patients with Class II malocclusion is warranted, or whether further investigations need to be undertaken. The volume, diversity, and contradictory nature of the current literature suggest that, before embarking on additional research, researchers need to clarify some of the apparent conflicts and identify important methodologic issues through a review of recent studies. Given the problems of the multiple and limited indices of growth used by orthodontists—coupled with the weak research designs most frequently employed, the small sample sizes studied, and the ambiguous and incomplete reporting of important information—it is difficult to endorse conclusions supporting or refuting the growth-promoting influence of orthodontic appliances. Even the best of these studies suffer from methodologic limitations that make the interpretations of the results difficult. This review suggests that increased consideration of design, analysis, and reporting would strengthen the value of clinical studies and permit more ready access to important information.

REFERENCES

1. Bishara SE, Ziaja RR. Functional appliances: a review. AM J ORTHOD DENTOFAC ORTHOP 1989;95:250-8.
2. Feinstein AR. Clinical Biostatistics XLIV. A survey of the research architecture used for publications in general medical journals. Clin Pharmacol Ther 1978;24:117-25.
3. Godfrey K. Comparing the means of several groups. In: Bailar JC III, Mostellar F, eds. Medical uses of statistics. Westford, Mass: Murray Printing, 1986;205-44.
4. Freimann JA, Chalmers TC, Smith H Jr, Kuebler RR. The importance of beta, the type II error and sample size in the design

and interpretation of the randomized control trial. N Engl J Med 1978;299:690-4.
5. Glass GV, McGaw B, Smith ML. Meta-analysis of social research. Beverly Hills, California: Sage, 1981.
6. Light RJ, Pillemer DB. Summing up: the science of review research. Cambridge: Harvard University Press, 1984.
7. Wolf FM. Meta-analysis: quantitative methods for research synthesis. Beverly Hills: Sage, 1986.
8. Bjork A. Prediction of mandibular growth rotation. AM J ORTHOD 1969;55:578-99.
9. Bookstein FL. Notes on the logic of statistics in orthodontic research. In: Vig PS, Ribbens KA, eds. Science and clinical judgment in orthodontics. Monograph 19, Craniofacial Growth Series. Ann Arbor: Center for Human Growth and Development, University of Michigan, 1986.
10. Chalmers TC, Matta RJ, Smith H Jr, Kunzler A-M. Evidence favoring the use of anticoagulants in the hospital phase of acute myocardial infarction. N Engl J Med 1977;297:1091-6.
11. Chalmers TC, Celano P, Sacks HS, Smith H Jr. Bias in treatment assignment in controlled clinical trials. N Engl J Med 1983; 309:1358-61.

Reprint requests to:
Dr. Joan Frances Camilla Tulloch
Department of Orthodontics
School of Dentistry
University of North Carolina
Chapel Hill, NC 27514