FEATURES SECTION

How to do

Moving forward by looking back: 'Retrospective' clinical studies

L. E. Johnston, Jr.
University of Michigan, USA.

I have been asked to comment on 'how to do a retrospective study.' A truly authoritative response to this charge is perhaps best sought in one or more reputable epidemiology texts. The present communication, therefore, can be seen as a few comments on research design culled from the literature and a few decades of experience with 'retrospective' clinical research. These comments may be of use to the neophyte investigator; however, they are more likely to find an audience among those who wish to read the literature with an eye toward deciding 'what works and what doesn't.'

Current clinical practice recognizes a variety of mutually exclusive therapeutic alternatives, among which are two-stage growth modification treatment in children (functional appliances followed by fixed appliances), one-stage fixed-appliance therapy (extraction and non-extraction, in either adolescents or adults), and combined orthodontic/surgical treatment in adults. It is unlikely that they are as interchangeable as we might hope; however, experience shows that everything works well enough and often enough to support a clinical practice. Perhaps as a result, there is surprisingly little demand for evidence. Instead, it is common to opt for a single treatment (early, often, and without extraction?) for application to all patients. In effect, the problem of choosing an individualized strategy is avoided; however, this approach has a hidden cost that is passed through to the patient in the form of regret—the difference between what the patient gets and what he/she could have got from the most appropriate treatment. Unfortunately, the orthodontist who would provide evidence-based treatment to individual patients is faced with a literature largely bereft of interpretable content.

Many, if not most, clinical papers are either 'case reports' or purely anecdotal 'opinion pieces'. A few more are little more than extended case reports—carefully filtered 'explanatory' studies designed to extol the virtues of the author's favourite technique. The typical orthodontic clinical report, therefore, is popular, inoffensive, but largely incapable of supporting valid inference about clinical orthodontics. Indeed, when the sample size goes beyond a handful of highly filtered 'cases', the paper apparently ceases to be seen as clinical and, instead, becomes some sort of despised exercise in theory. Given this popular mindset, the recent call for 'evidence based dentistry' (EBD) is welcome, rational, and not a moment too soon. Not all agree, however.

Surprisingly, EBD has generated widespread controversy and contumely. Many feel threatened; others are insulted. Normally, one might expect only the more entrepreneurial to rail against a call for evidence; however, the EBD movement has raised the hackles of a surprisingly broad spectrum of the dental and orthodontic world. Part of the problem comes from a tendency by those who prosecute prospective randomized trials to publish divisive, self-serving 'hierarchies of research' that serve to denigrate all other forms of investigation, including the various types of 'retrospective' research, both good and bad. The goal of the present communication is to argue for a middle ground.

The randomized trial

In medicine, prospective randomized trials are seen as the gold standard for clinical investigation. This status is well deserved, given their unmatched capacity to avoid bias from sources both known and unknown. Indeed, O'Brien,[1] speaking in support of the EBD movement, has suggested that, in orthodontics, '... we must introduce a paradigm shift to accept that only poor evidence will be derived from retrospective investigations'. I would argue that, at least from the standpoint of orthodontics, such a 'paradigm shift' expects too much of clinical trials and too little of at least some forms of retrospective research. Clearly, the Devil is in need of an advocate.

Address for correspondence: Lysle E. Johnston, Jr, Department of Orthodontics and Pediatric Dentistry, School of Dentistry, University of Michigan, Ann Arbor, Michigan 48109–1078, USA. E-mail: lejjr@umich.edu

Whatever their strengths, prospective trials are also costly and, especially in orthodontics, they can take so long to conduct that they may well outlive the question they are designed to answer. Given that malocclusion is not a disease, perhaps the major problem for those who would conduct a non-trivial orthodontic randomized trial is the ethical recruitment of subjects.

It is axiomatic that randomized trials may be prosecuted only if there exists a state of clinical equipoise, '... an honest, professional disagreement among expert clinicians about the preferred treatment'.[2] Studies designed to determine the 'preferred treatment' often feature options that differ greatly in terms of morbidity and inconvenience to the patient: orthodontics versus surgery; long treatments versus short; extraction versus non-extraction. Furthermore, Kodish *et al*.[3] have noted that 'The autonomy principle dictates that patients' personal values and motivations be given the highest priority in reaching a treatment decision.' Would potential subjects permit themselves to be randomized to surgery (rather than orthodontics) or to extraction (rather than non-extraction) if:

(1) they were *fully informed* that all of the treatments under investigation are assumed at least provisionally to be equally effective;
(2) pains were taken to elicit a preference?

Probably not. Unfortunately, even if one were able to recruit informed samples, there is no guarantee that the results would be worth the cost and effort.

The scope of the samples (and, hence, the general significance to the choice of treatments) would *decrease* in direct proportion to the diligence and skill of the treatment planning: the more care, the narrower the range of patients for whom there is true uncertainty about the relative merits of the competing treatments. As noted by Freedman,[2] 'Overly "fastidious" trials, designed to resolve some theoretical question, fail to satisfy the second ethical requirement of clinical research, since the special conditions of the trial will render it useless for influencing clinical decisions, even if it is successfully completed.'

Orthodontic treatments, moreover, are not a series of pills that can be administered and evaluated blindly. Both the patient and the clinician would presumably know of their participation in a clinical trial and thus might be expected to act with unusual diligence during the course of treatment. Moreover, should an assigned treatment prove to be ineffective, equipoise would be disturbed, and the clinician would be obligated to switch treatments, provided that a change is still possible (i.e. from orthodontics to surgery or from non-extraction to extraction). Alternatively, having been assigned an irreversible, but failing treatment, would a clinician proceed with the skill and confidence that would have been employed had the choice of therapy been his/her own? On balance, it is not obvious that the general problem of evaluating orthodontic strategies warrants the stern and perhaps ethically ambiguous measures that might well be required to achieve interpretable data from a prospective study.

The few orthodontic trials that have managed—occasionally at the cost of millions—to generate data have compared simple technical details (e.g. 'headgear versus functionals'), not basic strategies (e.g. orthodontics versus surgery). Although there are a number of simple questions that might best be explored prospectively (comparisons of wires, brackets, methods of retention, and the like), the most vexing of our clinical questions go well beyond the scope of an ethical randomized trial. Like it or not, we will be forced to rely on less exalted 'retrospective' alternatives. This is not to say, however, that the *goals* of randomized trials are entirely out of reach. As noted by Feinstein:[4]

> To obtain cogent scientific answers to cogent clinical questions, and to improve both the scientific and humanistic qualities of clinical practice, clinicians will usually have to rely on evidence obtained without experimental assignment of the compared agents.
>
> A ... misconception is to give randomization credit for certain scientific standards and precautions for which it is really not responsible... . The misconception just cited, which confuses the tactic of randomized assignment and the strategy of a scientific plan, is particularly important, because many of the desirable scientific features associated with randomized clinical trials ... are really attributable to advance scientific planning, not to randomization. These desirable features can therefore be obtained with suitable planning even when randomization is not used.

Unfortunately, 'suitable planning' is relatively rare in the orthodontic literature. Instead, our retrospective observational studies commonly feature biases that tend to overstate the effectiveness of a treatment. These obviously flawed studies in turn are invoked to discredit the entire genre. Well designed medical studies, however, have been shown to mirror the results of prospective randomized trials. Benson and Hartz[5] compared the results of observational studies with those of randomized trials. From a survey of 136 reports on 19 treatments, they '... found little evidence that estimates of treatment effects in observational studies reported after 1984 are either consistently larger than or qualitatively different from those obtained in randomized, controlled trials.' Indeed, meta-analyses have demonstrated that, when prospective and retrospective studied disagree, it is the prospective trial that is more likely to be discrepant. As noted by Concato and associates:[6]

> ... Previous studies have shown that observational cohort studies can produce results similar to those of randomized controlled trials when similar criteria are used to select study subjects. In addition, data from non-medical research do not support a hierarchy of research designs. Finally, the finding that there is substantial variation in the results of randomized, controlled trials is consistent with prior evidence of contradictory results among randomized, controlled trials.

I would argue, therefore, that the key to the generation of timely, cost-effective answers to many of today's most pressing clinical questions is the recognition and, to the extent possible, the elimination of the various biases that beset the average retrospective clinical investigation. To this end, I would suggest a few homely rules-of-thumb.

Retrospective alternatives

Control of susceptibility bias

Perhaps the major problem to be overcome in a retrospective 'case-control' orthodontic study is that of achieving an approximation of the bias-free sample selection that is a prominent feature of a prospective trial. Given that we are limited to patients who already have been treated, the various groups to be studied/compared will have been defined by past therapeutic decisions. The nature of the treatment planning, execution, and documentation, therefore, are filters that must be understood if sample selection is to be adequate to the task of generating interpretable data.

A malocclusion is not a disease; rather, it is a relatively non-specific *sign* that can result from a variety of formal, material, and efficient causes, none of which may be remarkable in and of itself: early loss of deciduous teeth, thumb-sucking, mid-facial protrusion, mandibular under-development, to name but a few. Given this great within-Class variability, the random, ethical assignment of treatments becomes a major problem. For example, it is unlikely that many orthodontists would opt for surgery or agree to extract bicuspids in a patient with a straight profile and generalized spacing. Conversely, only the most adventuresome would be willing to treat a severe tooth-arch discrepancy purely by expansion. As a result, a prospective study would have to identify (say, by the vote of a panel of experts) a prognostically homogeneous stratum of patients who would be equally eligible for each alternative under investigation (i.e. meet the ethical requirement that the trial begin with an honest null hypothesis). If there are ethical or practical problems in achieving a random assignment among these subjects, an obvious alternative would be to employ groups of patients who have already been treated according to the best judgment of a coterie of experienced clinicians. Given that all groups in a retrospective case-control study will have been formed by past clinical decisions, it is necessary to understand the basis of these decisions so that matched, equally susceptible samples can be identified.

A good initial rule of thumb is only to consider subjects who could have been enrolled in a prospective trial at the time treatment was begun. This simple criterion immediately eliminates the crudest and most common of biases, namely the temptation to choose subjects on the basis of outcome (or aliases such as 'co-operation'). As with a prospective trial, samples can be selected on the basis of a wide variety of pretreatment characteristics; however, because comparison of treatments can have meaning only in and for patients who are eligible for each alternative these characteristics must be stated precisely and must be common to all groups. Sample selection, therefore, is the key to a retrospective design that aspires to 'quasi-experimental' status.

Why has a given patient been treated without extraction or with functional appliances or with surgery? Probably because the malocclusion or the skeletal deformity was thought to be particularly susceptible to a given form of treatment. As a result, groups defined by treatment decisions tend to feature susceptibility bias: at the outset, patients treated in one way tend to be different from those treated another. Extraction patients, for example usually exhibit crowding, whereas non-extraction patients do not. When samples are different at the outset, differences at the end of treatment are largely uninterpretable. Thus, if meaningful between/among treatment comparisons are to be conducted, all patients must be eligible for all treatments. The problem then reduces to one of identifying groups of treated subjects who were, at the outset, morphologically similar, at least from the standpoint of the characteristics upon which the treatment decisions were based.

The most common method of finessing the problem of susceptibility bias is to 'match' the various groups according to a limited set of pretreatment characteristics (A–N–B, overjet, and the like). The matching process often involves untreated controls or data from 'normative' standards, in which case the criteria for the match tend to be age and sex. Although this approach is a step in the right direction, I suspect that the results are often compromised by the fact that the criteria employed in the matching may seem important, but may not actually be those that determined the original choice of treatment. Parenthetically, it may be noted that this same problem would also affect the identification of subjects eligible to participate in prospective trials. In response to this uncertainty in retrospective research, discriminant analysis can be applied to pretreatment data to identify the characteristics that were different between/among the various treatment groups. It is assumed that these differences influenced the assignment of treatments and thus can be used to define groups that would have been eligible for all treatments.[7]

By way of illustration, a discriminant analysis of extraction and non-extraction edgewise patients showed that crowding and protrusion were the main pretreatment differences.[8] Presumably, these also were the characteristics upon which the decision to extract was based. The resulting standardized discriminant scores then were used to identify 'borderline' subjects (those with discriminant scores near zero) who presumably could have been treated either way. The results of this (and other) studies show that, when faces are matched for a few characteristics, the overall match tends to be good as well.[8–10] Given samples that start out essentially the same, long term differences can be ascribed to differences between/among-treatments. Moreover, natural clinical biases (e.g. some orthodontists choose to extract in almost every case, whereas other seldom do) can be counted on to ensure that the various techniques will have been applied ethically, but perhaps not always wisely, to a spectrum of patients that might be well beyond the range of subjects to whom random treatments could be assigned prospectively. As outlined here, retrospectively defined equipoise is not 'balanced on a knife's edge',[2] but rather is broad and robust. It would permit the generation of clinically meaningful treatment comparisons throughout much of the range of actual clinical application, not merely the smaller segment for which ethical randomization can be rationalized.

Comparisons of extraction and non-extraction treatments in which no attempt is made to eliminate susceptibility bias commonly show outcomes that are, on average, quite similar. As judged by their impact on *matched* samples, however, the two treatments actually can be seen to differ in their impact on the profile: extraction produces a reduction in protrusion, whereas non-extraction treatments generally do not.[8] Thus, on application to the usual protrusive, crowded patients, extraction brings them into line with the relatively unchanged non-extraction patients, thereby leading to the faulty conclusion that the two treatments are interchangeable. Interestingly enough, when susceptibility is controlled, the two strategies seem to have a significantly different impact on PAR scores and thus are anything but interchangeable.[11]

Careful definition of the exclusion/inclusion criteria are usually said to be necessary so that the study can be replicated. I would argue that an equally good reason for a careful characterization of inclusion and exclusion criteria is to permit a clinician to know whether or not a study's findings apply to the next patient awaiting treatment. For example, it is common to see papers that feature exceptional results achieved by a variety of non-extraction treatments. Can these results be obtained in all patients or are these techniques applicable only to patients who have minimal protrusion and crowding? If you pick and choose from enough treatments you can demonstrate anything and thus prove nothing. Whether you are a performer of research or a consumer of its findings, it is necessary to know exactly what kind of patients were studied.

A few more obvious, avoidable biases

There are some groups of patients for whom growth and treatment comparisons might be desired, but whose prognostic indicators rarely overlap. For example, functional appliances are generally used in children, whereas fixed appliances, the obvious alternative, are used mostly in adolescents. Skeletal change, however, is dependent on the intensity of growth, which in turn is a variable function of age, sex, treatment time, and perhaps the choice of treatments. Thus, differences between the jaw growth accompanying functional therapy in younger patients and fixed-appliance treatment in older patients would be confounded with systematic differences in growth intensity that would be anticipated even in the absence of treatment. To address this problem, an index of expected growth intensity can be developed from the integration of sex-specific normative growth curves. The resulting dimensionless 'expected growth units' bear a significantly better relationship to increments of skeletal growth than does treatment time and thus can fuel analyses of covariance designed to compare growth effects in disparate samples.[12]

Once a treatment is instituted, why is a given patient documented well enough even to be included in a sample (or, for that matter, in a slide at a meeting)? This simple question is one of the key steps in the practice of evidence-based dentistry. Orthodontists sometimes neglect to take routine follow-up records, except perhaps for those patients whose treatment has turned out especially well (detection bias). Alternatively, the investigator may well have gathered complete records, but may still select the sample on the basis of outcome. In this context, it is common to argue that if you want to explain *how* an appliance works, you have to study it *when* it works. However it is justified, the end result is exclusion bias.

If only a limited number of clinicians generate the records, the results may depend more on the skill of the clinician than on the general worth of the treatment (proficiency bias). Proficiency bias can work the other way: clinical studies performed in universities commonly are indicted (verbally, but only rarely in print) for having studied treatments (functional appliances, articulator mountings, etc.) that have been executed less skillfully than would be the norm 'on the outside'. Whatever the merit of this criticism, it can be argued that university studies may feature better and more systematic documentation than might be seen in some private practices. In any event, when the appliance being examined is one that is generally employed on a minority of patients, or the source of records is a small group of 'co-operating orthodontists', or the samples have been formed on the basis of the quality of the records, or the samples to be compared are small and each has been obtained from a separate performance site, then the possibility/probability of bias must be considered.

Finally, once a sample has been defined, all of its members must be accounted for at the end of the study. Although the popular 'consecutively treated' criterion sounds good, the phrase actually conceals a host of potential evils. Instead, it may mean 'consecutively finished'—all patients who, for whatever reason, responded well enough to complete the course of treatments. The patients who fail to respond commonly are shifted to another treatment, often the very one (e.g. extraction or surgery) that the modality under investigation was designed to supplant.

Conclusions

There are a myriad of important clinical questions that vex contemporary clinical practice. The search for answers can unite the specialty; the answers, in turn, will have a profound impact on patient care. In the process, we should be satisfied with nothing less than the best evidence that circumstances and the nature of the questions will allow. This best evidence sometimes will come from randomized trials; however, as noted recently by Concato and associates,[6] 'The popular belief that only randomized, controlled trials produce trustworthy results and that all observational studies are misleading does a disservice to patient care, clinical investigation, and the education of health care professionals.' From the standpoint of a specialty that can attract only limited research funding, it is a conceit we can ill afford.

References

1. O'Brien K. Editorial: is evidence-based orthodontics a pipe-dream? *J Orthod* 2001; **28:** 313.
2. Freedman B. Equipoise and the ethics of clinical research. *N Engl J Med* 1987; **317**: 141–145.
3. Kodish E, Lantos JD, Siegler M. The ethics of randomization. *CA Cancer J Clin* 1991; **41**: 180–186.
4. Feinstein AR. An additional basic science for clinical medicine: III. The challenges of comparison and measurement. *Ann Intern Med* 1983; **99**: 705–712.
5. Benson K, Hartz AJ A comparison of observational studies and randomized, controlled trials. *N Engl J Med* 2000; **342**: 1878–1886.
6. Concato J, Shah N, Horwitz RI. Randomized controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med* 2000; **342**: 1887–1892.
7. Cassidy DW Jr, Herbosa EG, Rotskoff KS, Johnston LE Jr. A comparison of surgery and orthodontics in 'borderline' Class II, division 1 adults. *Am J Orthod Dentofac Orthop* 1992; **104**: 455–470.
8. Hannapel ED, Johnston LE Jr. Extraction vs. non-extraction: PAR-score reduction as a function of initial susceptibility. *Prog Orthod* 2002; **3:** 1–5.
9. Johnston LE Jr. Growth and the Class II patient: rendering unto Caesar. *Semin Orthod* 1998; **4:** 59–62.
10. Livieratos FA, Johnston LE Jr. A comparison of one- and two-stage non-extraction alternatives in matched

Class II samples. *Am J Orthod Dentofacial Orthop* 1995; **108**: 118–131.

11. Miettinen OI. Stratification by a multivariate confounder score. *Am J Epidemiol* 1976; **104:** 609–620.

12. Paquette DE, Beattie JR, Johnston LE Jr. A long-term comparison of non-extraction and bicuspid-extraction edgewise therapy in ‘borderline’ Class II patients. *Am J Orthod Dentofacial Orthop* 1992; **102:** 1–14.