BIAS IN DENTAL RESEARCH CAN LEAD TO INAPPROPRIATE TREATMENT SELECTION

Rhonda F. Jacob, DDS, MS

In research, as in life, bias is the enemy of truth.

R. F. JACOB

Bias is a systematic error that distorts the true relationship between an event and its outcome. Bias will negatively affect the truth of the conclusions. In research, bias includes any systematic error in the design, conduct, or analysis of a study. Bias can occur at all stages of research, from the selection of the population, how treatment is provided, to how and when outcome measurements are made. One report reviewed more than 50 possible sources of bias in analytic research.[33] The various research designs differ in the features within the design that control bias. Specific maneuvers attempt to control bias by reducing opportunities for systematic errors and by encouraging impartial judgment by persons involved in the study. In health care research, bias can result in a mistaken estimate of a treatment's effect or an exposure's effect on the course of disease.[12] These mistaken estimates probably account for some of the conflicting conclusions observed in apparently similar studies. Mistaken estimates can lead to practitioners' offering ineffective or even harmful treatments. It is the clinician's obligation to continue professional education by reviewing current literature. To optimize continued learning and patient care, clinicians should understand and scrutinize the various biases that can exist in research reports.

From the Department of Head and Neck Surgery, MD Anderson Cancer Center, Houston, Texas

CLINICAL RESEARCH

How the human population as a whole behaves under natural conditions and how the entire population of humans responds to a particular treatment are the ultimate health care questions for researchers and clinicians. Because the entire human population cannot be entered into or managed in a study, researchers and clinicians rely on the laws of probability and inferential statistics, which allow smaller sample populations to be studied as representatives of the population as a whole. These studies of sample populations use a multitude of research methods to determine the relationship between event and outcome. If stringent research and design criteria are not maintained, the assurance is lost that the sample population and its event-to-outcome relationship accurately represent that relationship in the total population; the study lacks validity.

Health care research designs are broadly described as observational or experimental.[19, 38] In observational studies, a passive investigator usually observes subjects for exposures and outcomes. In experimental studies, an involved investigator usually prescribes an intervention to achieve a particular outcome. It is generally accepted that, because of the active participation of the investigator, experimental studies offer the best opportunity to control bias and that a correctly implemented experimental study offers the best available evidence to answer a specific research question.

Whether an observational or experimental design is chosen to answer a given health care question depends on the type of research question being asked. For many health care questions, an experimental research design may not be appropriate because of the constraints of population availability, population management, cost, time, and ethics. Various design strategies have evolved to overcome these constraints, but some of the strategies increase the possibility of bias.

A hierarchy of research design exists, based on study validity and the ability to control bias within certain study designs.[15, 32] Clinicians and researchers must understand that less confidence can be placed in the research conclusions derived from some study designs, and extreme caution must be exercised when using these study reports to influence decisions concerning patient care.

In addition to employing the appropriate study design, certain elementary research methods must be implemented in all studies to control bias. These include methods regarding patient selection, examiner training, intervention, data collection, and analysis. When bias is not controlled in these areas of clinical research, conclusions are highly suspect, no matter what the study design.

HIERARCHY OF RESEARCH DESIGN AND BIAS CONTROL

The hierarchy of research design is based on satisfying three main criteria: (1) randomized or nonbiased selection of target and control

subjects; (2) intervention or putative exposure under the control of the investigator, and (3) prospective gathering of outcomes after entry into the study.[7, 9, 11, 13, 14, 15, 34, 38, 39a] The control of bias in a given research project depends largely on whether these criteria are met.

One of the greatest biases of health care research arisis from the methods of selecting the sample populations targeted for the research. If research subjects are inappropriately selected, no amount of stringent research methodology can counter the bias of sample population selection. It has been suggested that the scope of population selection bias in health care literature poses "potential catastrophic damage to a study's inferential basis" and should be taken as a serious threat.[7] Some research designs have more inherent patient-sampling safeguards than others. When these safeguards are appropriately executed, the higher-quality design, with the higher-quality patient sampling safeguards, should be used to make health care decisions.

Randomized, Controlled Trials: "Best at Bias Control"

Randomized, control group trials (RCTs) offer the greatest opportunity for the investigator to identify subjects and then randomly assign them to the intervention group or the control group by a predetermined randomization protocol. Treatment is not rendered until the subjects are randomly assigned to the study groups. Patient data are collected in a prospective fashion to evaluate the intervention's effect on the outcome of interest. This study design is ideal for evaluating therapy. A sample group of subjects with the malady of interest is further narrowed in number by the use of specific inclusion and exclusion criteria appropriately based on the research question. Subjects are usually excluded from the study because their inherent characteristics are not relevant to the research question. For instance, adults would probably be excluded from an orthodontic treatment trial evaluating mandibular growth. Some characteristics or cointerventions may confound the research conclusion. Confounding characteristics may have or are suspected to have nearly as profound an effect on the outcome as the intervention, and including subjects with confounding characteristics makes it difficult to discern the true effect of the intervention. Subjects with these confounding characteristics are excluded from the study. For instance, when osseointegration of dental implants was first evaluated, diabetic patients were often excluded, because diabetes was thought to confound the ability to measure healing at the implant site. When evaluating a question related to in-office bleaching of teeth, researchers would probably exclude persons performing at-home bleaching (a cointervention), because this additional therapy would probably confound the true effect of the in-office study intervention.

After the subjects are selected, they are queried about their willingness to undergo the study. Ideally, a study would report data on the subjects who were eligible for the study, but refused to enter it.[7] The

investigators should then evaluate the characteristics of those persons who entered and those who refused. This evaluation can establish that the persons entering or not entering the study are alike in measurable respects, and therefore the subjects entering the study are representative of the total population of those subjects. For instance, in a dental trial in which subjects are required to pay for therapy, persons of a lower socioeconomic status may be eligible for the study but consistently refuse to enter the trial because of financial concerns. This financial issue becomes a selection bias, before the trial subjects are enrolled. The subjects in the study should be recognized as representing the population within a socioeconomic stratum, rather than the population as a whole.

The manner in which subjects are recruited before screening can produce a selection bias. If an implant study is advertised, only persons interested in implants report to the recruiting site. If all dental school denture patients are queried about their desire to enter an implant study, a number will probably refuse. Something is inherently different in subjects who volunteer for studies versus those who do not. Patients who actively seek implants and those who are offered and accept implants as an option to new dentures are likely to be from different subsets of the population. This selection bias of recruitment at the outset of a study could greatly affect how patients report their satisfaction outcomes and could account for implant studies' reporting contradictory results.[2, 20]

After subjects are screened and found to be eligible for the study, they will be randomly assigned to the treatment groups. Randomization allows the patients an equal opportunity to be assigned to either intervention group, thereby reducing selection bias and allowing the study to be representative or generalized to the total population of other patients with similar maladies and characteristics. Randomization should be generated by computer programs,[9] and the entry schedule should be kept blind to investigators and study accrual personnel. Assigning subjects to study groups by birthdate, entry date, hospital number, or an alternation schedule is haphazard, but is not randomization. These "haphazard or quasi-randomized" methods allow study personnel or referring clinicians to have prior knowledge of which group the patient will enter. Well-meaning assistants have been known not to enter a subject in a trial when they believe the subject would receive little benefit from the assigned therapy. A system of alternating assignment allows one to guide the order of accrual of subjects and to place a subject in a specific study group, based on the desires of the subject. Assistants responsible for accruing subjects might guide some subjects to a particular group because the morbidity rate is lower and the subject might be more likely to finish the study. If accrual personnel know the new therapy is next to be assigned, they might give positively slanted information to a prospective subject, thereby ensuring the subjects' entry into the study. (Clinicians, too, can be influenced by their perception of what offers the best treatment opportunity for their patients.) These systematic biases can distort treatment outcomes. Blind randomization allows equal

distribution of variables known to affect treatment outcomes. Perhaps more importantly, randomization allows equal distribution of the unknown characteristics that might affect treatment outcomes. Having interventions under the control of the examiner at the outset of the study allows treatment methods to be standardized by trained practitioners. This control also allows standardization of follow-up regimens, record keeping, and measurements of outcomes. This standardization of methods, along with adequate training for practitioners and examiners, helps minimize bias, thereby making the RCT the definitive clinical trial.

Observational Comparison Studies: "Ranges of Bias Control"

Observational studies, such as cohort and case-control studies, have been used in epidemiologic surveys to determine the natural history of a disease and exposures associated with a disease. Observational studies, in which the investigators do not actually manipulate the subjects' exposure to a treatment or event, but only observe outcomes and often retrospectively determine exposures, are often used to discern the prevalence of a disease. Such studies are also used to determine population characteristics that might be risk factors for disease. Observational studies rank lower in the hierarchy of evidence because they can meet few if any of the three criteria for bias control. Observational comparison studies, however are usually the studies of choice when risk factors or harmful exposures are being evaluated. The major weakness of these studies is that the patients are not randomly selected; but rather are selected because they were exposed to a particular event, had a specific lifestyle choice, or were noted to have a particular outcome.

A number of study designs fall under the description of observational. The strongest of the observational design strategies is the inception cohort study, in which the investigator is present immediately after the exposure or event occurs (at the inception) and follows the subjects for outcomes using prospective and standardized methods. A control group, whose subjects were not exposed to the event, must be followed with the same prospective methods to determine comparatively how many subjects develop the outcome of interest. Great care must be taken in selecting subjects for the concurrent control group. The control group must be as nearly equivalent as possible to the exposed population in every measurable characteristic that might affect the outcome. The characteristics commonly considered are age, sex, socioeconomic status, and educational background. Depending on the outcome of interest, other characteristics, such as geographic area, concurrent medical conditions, cointerventions, occupational exposures, among others, must also be examined in selecting the control population. Unfortunately, in all circumstances there are unknown characteristics that may influence the outcome of interest. Unlike the random assignment of subjects to the test or control group, there can be no safeguard to assure that these

unknown characteristics are equally distributed in the control population. Therefore, it is understood that considerable bias can occur in selecting the control population. An inception cohort trial might be used to determine whether persons with and without amalgam restorations have an equal risk of developing multiple sclerosis. Another current health question that might be considered using the inception cohort design is whether the risk of autoimmune diseases is equal in women who undergo silicone breast augmentations versus those who do not. Both these issues have been hotly debated in health care and media arenas. There is probably an unknown multifactorial cause for multiple sclerosis and autoimmune diseases; therefore, selecting a control population that is similar for these unknown characteristics is nearly impossible and fraught with bias.

The inception cohort study is the premier observational study because of its prospective nature. Unfortunately, waiting for outcomes to occur may take many years, leading to loss of subjects, loss of trained study personnel, and prohibitive costs. The difficulty of maintaining the validity of a protracted study adds additional biases. Other observational designs using retrospective, one-point-in-time evaluation of comparative populations with and without the outcome of interest offer a more immediate answer to the research question. The price for immediacy, however, is increased bias and risk of distorting the true relationship between event and outcome.

Three types of retrospective studies are cross-sectional, ex post facto, and case-control studies.[15, 39a] In these research designs, the outcome has already occurred in the test population. The selection of the control population is critical to reduce bias. Control subjects should be as equivalent as possible in all characteristics to the test population, with the exception of the exposure of interest. The comparison is the incidence of the outcome in the test population and in the control population. The control subjects may come from the same population pool as the test subjects or from a different population pool. For instance, when investigating whether a particular dental assistant chair may increase the risk for lower back pain, the same-pool subjects might be drawn from all dental assistants at one dental school, but they would be allocated into the control or test group based on whether they used a specific design of chair. Control subjects drawn from a different-pool population could be assistants working at a different dental school where a different chair design is used. Same-pool populations are more likely to have similar demographic and workplace characteristics, both known and unknown. Regardless of whether same- or different-pool subjects are selected for the control group, the processes for identifying possible subjects and the final selection of each subject must be consistent. In either design, the subjects would be queried about their present or past history of back pain.

Selection bias is quite difficult to control in observational studies. Because investigators are often gathering data on exposures that have already occurred, the existence of an exposure or outcome must often

be confirmed by patient report or past medical records. Medical records are often incomplete, because practitioners may not document the specific findings required for the study. Alternately, an investigator may infer exposure or outcome from other clinical findings (not the outcome of interest) or tangential records such as insurance claims. These methods may lead to a biased selection of subjects that do not represent the totality of exposed patients. Investigators often evaluate characteristics in the two populations to show that they are similar in all respects except the exposure of interest. Even though the two groups being evaluated may seem comparable, there is always the possibility that one or more unidentified characteristics are responsible or at least influence the outcome of interest; these other characteristics are unlikely to be distributed equally between the two groups.

Subjects may also be selected based on their recall of an exposure or event, thereby creating a recall bias. Subjects who have the outcome of interest, or fear they will develop the outcome of interest, are more likely to recall that the exposure occurred. During subject interviews, investigators should blind the subjects to the outcome of interest and the exposure of interest. This blinding can be accomplished by asking the subjects many questions regarding various outcomes and exposures to decrease their awareness of the possible interactions.

Besides the difficulties inherent in population selection, the one-point-in-time studies have other biases. The assessment of outcomes represents a snapshot of the subjects' daily lives. Outcomes that are identified by waxing and waning signs and symptoms may not be present during the study evaluation. At evaluation, the outcome may be at an early, barely detectable level. Subjects aware of the possible outcome and exposure relationship may have a biased response when asked to recall their symptoms and exposure data and their cointerventions. Cointerventions or confounding signs and symptoms are also likely to wax and wane, thereby affecting the outcomes during the evaluation and affecting recall by the subjects.

Observational studies have been used extensively to evaluate harmful exposures. Smoking risks for cardiovascular disease and lung cancer have been universally accepted only in the past decade. Many investigators from many countries have reported increased health risks in persons who smoke. Because of the inherent weaknesses in observational studies and the political and monetary implications of these findings, many years and hundreds of confirmatory studies were required before the risks of smoking were accepted. The few studies that have been conducted on the health issues of amalgam restorations or breast implants reveal a wide range of risks, including no increased risk, for persons undergoing these treatments. Currently, the literature regarding these controversies include more letters to the editor than clinical trials. A MEDLINE search of reports associating amalgam restorations with multiple sclerosis reveals only three small, case-control trials in the past 20 years; with inconclusive suggestions of an increased risk of multiple sclerosis or alterations in immune parameters. One study reported that

their multiple sclerosis population did not have an increased number of amalgam restorations, but did have an increased number of caries compared with the control population.[25] The other study reported that the multiple sclerosis group had an increased number of amalgam restorations.[1] This report shows another problem with one-point-in-time evaluations: the inability of such studies to establish a cause-and-effect relationship between exposure and outcome. Patients with multiple sclerosis may have poorer oral hygiene because of possible physical constraints of their disease. Therefore, they may have more caries, and if their caries are treated, they are likely to have more amalgam restorations. This chicken-or-the-egg problem is common in case-control trials that identify possible associations between two processes. Associations can be revealed, but not causation; a cause must always precede the outcome.

Case Reports and Case Series: "Bias Out of Control"

Having a comparison control group is an absolute criterion for research. Studies that do not have a comparison group are relegated to an inferior position in the research hierarchy. These reports are most commonly referred to as case reports or case series. It has been stated that in these studies the only basis for comparison is "implicit, intuitive, and impressionistic."[14] Sackett states that inductive reasoning gives way to seductive reasoning.[32] Rather than controlling bias, case reports and case series are more likely to be *bias out-of-control*. Reports of a single patient outcome or a series of patient outcomes are subject to extreme bias in patient selection and treatment delivery decisions and methods. Subjects in case series do not represent a random sample of the total population, patients within the treatment group often have many pretreatment characteristics besides the malady of interest, and subjects are rarely treated with a standardized protocol of therapy. Often, data are gathered in a retrospective review of charts with nonstandardized measurement and outcomes assessment criteria. Despite their best intentions, reporting clinicians are biased by the very fact that they rendered the care and analyzed the outcome. Clinicians should never predict treatment outcomes based on reports that do not have a comparison group.

Despite their unreliability as predictions of treatment outcomes, unusual case reports and case series have value. These case reports call attention to little-known maladies, reveal complications of proposed therapies, and document outcomes that may have occurred because of exposures and proposed therapies. Precisely documented characteristics and descriptive data from case series and case reports are often used to plan subsequent research with control groups.

Historical Control Groups

Control groups are required to assess the value of a therapy. Dental and medical reports have commonly used data from patients who were

treated earlier at the same institution with a different modality of treatment. As a new therapy is introduced to the profession, some practitioners will begin using it. To assess the value of the new therapy, the practitioners will compare the group receiving new therapy with the group receiving the older therapy. When comparing the outcomes of the old and new therapy, the patients who received older therapy would become the historical control group. Rarely is this historical control group an arm of a RCT with specific population criteria and prospective data protocols. Instead, the historical control group usually consists of patients who were given the older therapy based on a number of decisions made by the patient and the practitioner, and specific treatment and outcome analysis methods were not standardized. Often, these data are gathered from chart reviews. Even if the two groups are treated during the same time-frame, a multitude of biases exist in this type of patient assignment and in the non-standardized methods. When patients in a historical control group were treated many months or even years previously, unknown variables and unknown cointerventions can create additional bias that is likely to affect outcome.

An analysis of the literature was performed to compare findings in therapies when data reports are based on RCTs versus historical control trials. A total of six therapies had reports of both study designs evaluating similar outcome endpoints, for a total of 50 RCTs and 56 historical control trials. The historical control trials found that the new therapy was better in 78% of the trials, where the RCTs found the new therapy was better in only 20% of the trials. When comparing the control group in the RCT with the control group in the historical controlled trial, the control group in the historical control trial not only fared worse than the experimental group in that trial, but often fared worse than the control group in the RCT. This finding supports the lack of equivalence in the two populations.[34] The two groups are rarely equivalent, except for the primary diagnosis. When a new therapy is developed, there are often conscious or unconscious efforts to narrow the criteria in the treatment group to include only those who are considered most likely to benefit or most likely to comply with the new methods. The others receive the traditional or historical treatment. Also, when historical controls are used, not all participants are included in the evaluation. The finding that control groups in the historical control trials faired worse than control groups in the RCTs suggests that bias in patient selection may ''irretrievably weight the outcome of HCT in favor of new therapies.''[34]

In a retrospective chart review of patients receiving palatal obturator prostheses to restore palatal defects following maxillectomy, it was hypothesized that patients had shorter hospital stays when they were given this prosthesis at time of surgery rather than several days after surgery. A review of 120 patients from 1960 to 1971 revealed that nearly 58% of patients did not receive surgical prostheses, and an evaluation of 151 patients from 1980 to 1984 revealed that 45% did not receive a surgical

prosthesis. In the earlier trial, there was a significant difference in the duration of hospitalization of the two groups studied (22.7 and 14.2 days, respectively), but no significant difference was observed in the later trial (10.6 and 8.0 days, respectively). The practice of dentistry and medicine has changed remarkably from 1960 to 1984, but the cause of the difference in hospital stay in the two groups in the earlier trial and the cause of the magnitude of difference of hospital stay between the two trials remains undetermined. Thus, using historical controls, even within the same institution, presents difficulty in distinguishing treatment effects from changes in ancillary care, manpower, referral patterns, patient support methods, health care reimbursement, and so forth. Historical controls derived from published reports present the same difficulties.

BIAS IN RESEARCH METHODS

Bias control continues beyond design selection and population selection. Specific methods of bias control should be implemented in the conduct and analysis of the investigation. These methods are applicable to all research designs.

Blind Participants

Blinding the investigators, examiners, and subjects to the intervention and the outcome is a significant controller of bias. Double-blind methods are the ideal situations. Subjects and study personnel are blind to the treatment assignments and to any study events or information that might influence outcome assessments. Single-blind methods blind either the examiner or the patient. When procedures are performed, the persons who examine subjects for outcomes or collect data from subjects should not be the same individuals who perform the procedures. Dentists have been trained to perform various treatment alternatives. For example, fixed partial dentures, removable partial dentures, and implants have all been used to replace the same missing tooth. Most dentists prefer one restorative method over another, and no dentist can state that the preference is solely based on scientific evidence. If the preference is not solely based on scientific evidence, there is an element of bias, and this bias can affect the outcome assessment if outcome data are collected by the practitioner.

Those who collect data should be blind to the hypotheses of the study. This blinding is likely to be easier than blinding the clinician who performed the dentistry. Data on oral conditions, restorative conditions, and function could be collected; however, only some of the data would be relevant to a given study. Some institutions have established data

collection facilities, where routine data is collected under strict protocol for all subjects sent to the data collection facility, irrespective of the study in which the subjects are involved. Subjects can be queried about a number of oral conditions without knowing specifically what condition or exposure is relevant to the hypothesis. Blinding subjects to their treatment, especially in dentistry, requires ingenuity. Sham treatments are often unconvincing, and the informed consents required today are so explicit that study subjects may be biased by the description of the procedures and the list of possible complications. Preconceived notions that subjects form during the informed consent process may influence their outcome responses. This influence may be a problem when a model consent form, with its blanks to be completed, has been approved by an institutional review board and is expected to serve as the consent form for all studies. Investigators should campaign for wording in their specific consent form that avoids biasing study participants. When informing subjects of the comparative treatments in the study, clinicians and research assistants should strive to control their own biases. When screening persons for study entry, applicants should be reminded that the study is being conducted because the dental community is not convinced which treatment functions better, is faster, is more esthetic, has greater longevity, and so forth.

When reviewing the literature, clinicians should evaluate whether blind methods were used in data collection, and the methods for assuring blinding should be explained. With this information, the clinician can determine if blinding truly occurred. If blind data collection was not employed, clinicians should search for other studies that address the research question.

Treat All Subjects the Same

Specific methods for delivery of interventions, data collection, and analyses should be determined before initiating an investigation. These protocols should assure that study participants in both treatment and control groups are treated and assessed equally. Doing so requires that the same follow-up regimen, follow-up data, and tests be performed on all subjects. Questionnaires and quality-of-life analyses should be administered in the same fashion to all participants. Follow-up examinations should be scheduled as often as needed to gather the data necessary to answer the research question and as often as needed to anticipate complications, complaints, and compliance issues. Bias can result if patients with complications must alter follow-up regimens because the follow-up examinations were not scheduled frequently enough. Subjects with less tolerance or with more complaints have potential for more frequent follow-up and have the potential of being evaluated differently. It is likely that more data will be gathered on these subjects. Pertinent data may be missed on subjects who return sporadically; their complications and improvements may need to be assessed by history taking

rather than by examiner observation. The inequities in such data gathering should be recognized as potential biases.

Although prospective interventions are not employed in observational studies, specific protocols for data review of records, patient interviews, and tests to evaluate outcomes should be designed in a prospective fashion. Before the investigation is begun, and even before the populations are selected, methods must be established so all subjects are tested and queried similarly. In some observational designs, the outcome of interest is often present before the study is initiated. The investigator queries subjects about exposure history. There is potential for investigators to interview subjects more vigorously to uncover the exposure when the subjects exhibit the outcome. This difference in the level of interrogation potentially biases towards a positive correlation between the exposure and outcome. This problem underscores the need for established methods for data gathering, as well as the need to blind the examiners to the outcomes.

Often, subjects are not treated similarly because of missing data. In dentistry, outcomes or baselines may be retrospectively assessed using existing radiographs, photographs, or study casts. Records that were not made for the purpose for which they are currently being used often fall short of meeting various criteria. Frequently, subjects who are otherwise eligible for the study cannot be enrolled because these previously collected records are not available or are nondiagnostic. Records made during a routine clinical examination may serve the purpose for a patient's treatment or evaluation on that occasion but are often not detailed enough for a later research project. For instance, casts made for custom trays may not be of adequate quality to serve as baseline for studies that require anatomic detail of all tooth surfaces. Less than ideal radiographs may not be remade if patients complain of discomfort, and appropriate angulations of film and beam may be sacrificed. Photographs may be missing; in a busy practice, clinicians may not retain serial photographs of specific patient outcomes that were unsuccessful or unesthetic. Investigators must decide either to extrapolate data from these less-than-ideal sources of documentation or to exclude these potential research subjects. Although it might seem that the better solution is to exclude subjects with missing documentation, doing so may create a serious selection bias. One study sought to evaluate the esthetic outcomes of a specific surgical method of closing cleft lip and palate. Subjects came from one surgeon's practice, were treated by one of two surgical methods, and were included only if they had had a clinical photograph made after age 15 years. The esthetics of the lip closure were evaluated by a panel of lay judges blind to the surgical method. Subjects with missing photographs or poor-quality photographs were excluded from the investigation. Twenty subjects were included in each group for analysis. No data were supplied as to the number of subjects who never returned before age 15 years, how many subjects failed to have quality photographs, or the percentage of the entire population these 40 patients represented. In this investigation, a population selection

bias occurred based on whether photographic documentation was available on the subjects.[31]

Calibration and Training of Examiners

Innumerable studies are available in the health care literature that specifically test the level of agreement among multiple examiners who are evaluating a clinical test, making a diagnosis, reading radiographs, or measuring treatment outcomes. More than 300 clinical reports evaluating observer variability in health care published from 1985 to 1989 were complied in a pre-MEDLINE bibliography.[8] A MEDLINE search found 57 clinical trials that evaluated observer variation between 1990 and 2001. Various indices of agreement have been formulated based on percentage, probabilities, correlation coefficients, the kappa statistic (κ), and others.[6] The κ statistic is preferred because it provides for an adjustment of agreement beyond chance and is appropriate for category scales and continuous data. (Kappa is affected by prevalence, and it cannot be calculated when one of the investigators constantly uses the same score. Variations on the original formulations by Cohen are frequently employed. Kappa is widely used and widely debated. Continued variations and other models for measuring agreement are being evaluated in statistical arenas.) It has been estimated that for many medical decisions, clinical agreement is at a suboptimal level, with κ below 0.35.[23] It has been proposed that κ less than 0.4 is poor agreement, κ of 0.40 to 0.75 is fair to good agreement, and κ above 0.75 to 1.00 is excellent agreement.[10]

Even calibrated examiners in dental investigations have not consistently reached good agreement in clinical measurement. Observer agreement was reported among seven calibrated observers of various dental specialties, who evaluated quality of bone trabeculation from 100 panoramic radiographs using a five-point scale. This scale was similar to that used in various implant studies and ranged from lack of trabeculation to bone as dense as cortical bone. The mean intraobserver agreement was $\kappa = 0.61$. The observers were paired in 21 pairs, with interobserver agreement ranging from $\kappa = 0.23$ to 0.56. Comparison of all seven examiners measuring all 100 sites and grades revealed $\kappa = 0.38$. Grade 1, representing no trabeculation, had the most agreement of $\kappa = 0.76$. Grades 2, 4, and 5 were $\kappa = 0.38$ to 0.39. The worst agreement was for normal trabeculation with $\kappa = 0.23$. A grade of 5, representing dense trabeculation, was given 230 times, but 25 subjects were regraded to level 2 on a repeat examination by the same examiners.[39] These measurement methods have been used to qualify boney trabeculation and subsequently enroll or exclude patients from implant studies. These same bone qualification methods have been used retrospectively to explain implant failures.

Another investigation considered 11 parameters of fixed restorations evaluated on a five-point scale by two calibrated examiners from each of six participating centers. The two examiners from each institution

were evaluated for agreement on each of the 11 parameters. The agreement of each pair ranged from $\kappa = 0.16$ to 0.95. The mean of the κ values from all six institutions for each parameter ranged from $\kappa = 0.56$ to 0.91. Marginal integrity had the lowest level of agreement.[27] An evaluation of four calibrated examiners investigating the efficacy of dental radiography found intraexaminer agreement was $\kappa = 0.75$ or higher at baseline and remained at approximately the same level (0.80) throughout the 24-month period of the study. The interexaminer agreement among the six pairings of the four examiners ranged from $\kappa = 0.68$ to 0.80 for caries and 0.72 to 0.83 for periodontal disease.[40]

As in other health care clinical measurements, various dental measurements result in ranges in practitioners' level of agreement. This lack of agreement indicates how critical it is to decrease bias created by systematic errors in measurement by training multiple examiners in the appropriate use of measurement instrumentation and in the implementation of clinical criteria. The more explicitly each measurement technique and category is defined, the less ambiguous are the demarcations between categories, and the higher is the observer agreement. The level of agreement of multiple examiners should be tested before an investigation to assure that the examiners have reached an understanding of measurement criteria and an acceptable level of agreement. During the investigation, continued calibration is often necessary, and the final level of agreement achieved during the investigation should be reported.

Accounting for all Subjects

It is disconcerting to an investigator to have subjects not complete a study. Statistical tests (power analysis) are often performed before the investigation to determine how many subjects are necessary to detect a difference in outcome between the groups. When subjects do not finish the trial, a result may be inconclusive because the lower number of study subjects causes a lack of statistical power. In prospective trials that require a long follow-up to determine the outcome of interest, there is an increased chance of losing subjects for a myriad of reasons: noncompliance, moving away from area, loss of contact, inability to travel to test site, and unrelated death, among others. It is important to determine the characteristics of the subjects who left the study and to perform another analysis of the remaining subjects to determine if the two groups are still equivalent in the variables that might influence the treatment effect. In addition, one should determine if the dropouts are more common in one group than the other. Uneven loss of subjects was found in a study evaluating the effectiveness of vitamin C in decreasing cold signs and symptoms.[20a] The caplets often broke, allowing subjects to taste the medication, and subjects discussed this occurrence among themselves while waiting for study evaluations. Persons in the placebo group realized they were not tasting ascorbic acid and began to drop out of the study, anticipating no benefit, where as the subjects "tasting

the benefit of treatment'' continued the study. More drop-outs in one group than another can signal a loss of blindness to therapy or may indicate untoward side effects. Uneven loss makes the study groups unequal in numbers and in known and unknown study variables.

Too often, reports simply change the number of subjects (N) at the end of the study, with minimal or no reference to the subjects lost to follow-up. It is assumed that these lost subjects have experienced the outcomes at the same rate as those subjects remaining in the study. For example, a systematic literature review of the English-language reports published since 1960 evaluated the survival rate of fixed partial dentures (FPD). Difficulty arose in performing a meta-analysis of the reports because many of the reports did not have any follow-up data on a large portion of the subjects after insertion of the prostheses.[36] As follow-up continued, even more subjects were lost to follow-up. One report quantified 255 FPD inserted over 10 years but only had 121 available for evaluation at year 11.[30] Another considered a one-point-in-time evaluation of 77% of an original 184 FPD. No data were reported on the 33% of lost subjects.[5] Eighteen years after insertion of 122 FPD, 66 persons were available for a follow-up analysis. No data were reported on the 54% of lost subjects.[28] A large database of 642 FPD inserted in 1974 was randomly selected from a national dental insurance registry. A 10-year evaluation was made, but only 164 persons presented for examination.[21] The subjects were evaluated again at year 14, with only 97 of the original 642 subjects reporting.[22] It is inappropriate to assume that 30% to 50% of subjects lost to follow-up would have the same outcomes as those subjects remaining in the study.

Investigations are often undertaken to determine differences in treatment outcomes that are usually quite small. Often, the difference in outcomes between the therapies is only 10%. Loss of subjects will reduce the statistical ability to detect these small differences in outcome. Losing only 10% to 15% of subjects can render a study inconclusive. Altering the final N of the study risks drawing the wrong conclusion about the value of the therapy.

Data Used Appropriately: Chart Reviews and Errors of Omission

In health care research, review of patient treatment records is a common method of describing disease prognosis and determining therapeutic outcomes. Often, historical control data are collected from treatment records to compare previous therapies with current therapies. Some studies have used insurance records or national health care registries to gather data on the prevalence of a disease. When patients are treated as subjects in a research protocol, the data recorded are driven by the research question. In a well-designed trial, measurements or tests required for the protocol are documented and read with strict attention to minimizing bias, using many of the methods previously described.

Records kept for routine treatment in a clinical setting, however, are often incomplete. Tests may be read, but not recorded. Not all subjects will receive the same tests, and techniques may be modified based on factors unrelated to the disease process. Patient compliance is often not considered. Cointerventions are rarely recorded. Follow-up examinations are often scheduled at patients' requests; therefore, unless patients have a specific complaint, their follow-up schedule will be abbreviated compared with other patients receiving the same treatment. Often, the notes are influenced by a patient's complaint; unless the patient complains, the follow-up note is an array of summary statements of "patient satisfied, within normal limits, normal diet, good esthetics, good occlusion, watch tooth # 3," and so forth. Treatment records are maintained by the treating clinician, and often patients are reluctant to complain to their practitioners, lest that complaint negatively affect the practitioner–patient relationship. For the same reason, patients may tend to overemphasize the positive outcomes of their treatment. Clinicians are also likely to overestimate the positive outcomes of therapies they deliver, waiting for patients to bring forward complaints, rather than asking whether patients experience particular difficulties.

Without standard treatment protocols and documentation, omission of data or ambiguous interpretation of data to fit a research question is problematic. A concurrent investigation was undertaken to evaluate temporomandibular disorder on a group of patients receiving orthognathic surgery. An RCT evaluating the cost, risks, and efficacy of two jaw fixation techniques was performed, and pertinent data were documented by the treating clinicians in the patients' records. The second study involved specific evaluations of patients with temporomandibular disorder performed by blind examiners with specific examination protocols performed on the same patients. The authors then examined the disagreement between data taken from the treatment records and data taken from the temporomandibular disorder examination. Four parameters were evaluated: (1) a vertical opening of more than or less than 40 mm, (2) the presence or absence of clicking, popping, or locking of a joint, (3) the presence or absence of pain, and (4) the presence or absence of crepitus. Although both studies were prospective, it became apparent that the surgeons focused more on efficacy of treatment than on secondary outcomes. Often, no data in the treatment records addressed the criteria for temporomandibular disorder. In other instances, it was necessary to create operational definitions of the four criteria that would allow interpretation of the surgeons' notes to categorize the outcomes. At 2- and 24-month surgical follow-ups, surgeons stated that 23% and 0% of subjects, respectively, had a vertical opening below 40 mm, whereas the temporomandibular disorder examiners reported 90% and 21%, respectively. The surgeons reported pain in 8.6% and 1.7% of the subjects, respectively, whereas the temporomandibular disorder examiners reported 47% and 29%, respectively. These differences show the level of disagreement that can occur when data from routine treatment records are used for research purposes as compared with data gathered by blind, calibrated examiners.[85]

SUMMARY

The first RCT was instituted in the early 1950s, evaluating streptomycin and bed rest compared with bed rest alone for tuberculosis.[26] This research design has become the reference standard for comparative evaluations of therapies because of its prospective nature and the ability to control bias. Because it is easier to conduct observational studies, they have often been inappropriately substituted for the better experimental study designs. Since the 1950s, however, readers of the medical literature have slowly come to demand quality clinical research to assist them in caring for their patients. Dentists are somewhat behind their medical colleagues in using the strongest research designs to answer clinical questions. In dentistry, observational studies with convenience samples of patients have been commonly used. It is often argued that few dental ailments affect a person's life as negatively as most medical maladies; therefore, experimental rigors are not required of dental research. Although most dental care does not involve life-and-death issues, dentists are as eager as physicians to offer their patients optimal care. Optimal care is best defined through nonbiased research strategies.

References

1. Bangsi D, Ghadirian P, Ducic S, et al: Dental amalgam and multiple sclerosis: A case-control study in Montreal, Canada. Int J Epidemiol 27:667–671, 1998
2. Boerrigter EM, Geertman ME, van Oort RP, et al: Patient satisfaction with implant-retained mandibular overdentues: A comparison with new complete dentures not retained by implants—a multicentre randomized clinical trail. Br J Oral Maxillofac Surg 33:282–288, 1995
3. Carr AB, McGivney GP: Users' guides to the dental literature: How to get started. J Prosthet Dent 83:13–15, 2000
4. Chalmers TC, Celano P, Sacks HS, et al: Bias in treatment assignment in controlled clinical trials. N Engl J Med 309:1358–1361, 1983
5. Cheung GS, Dimmer A, Mellor R, et al: Gale M. A clinical evaluation of conventional bridgework. J Oral Rehabil 17:131–136, 1990
6. Cohen J: A coefficient of agreement for nominal scales. Educational Psychology and Measurement 20:37–46, 1960
7. Ellenberg JH: Selection bias in observational and experimental studies. Stat Med 13:557–567, 1994
8. Elmore JG, Feinstein AR: A bibliography of publications on observer variability (final installment). J Clin Epidemiol 45;567–580, 1992
9. Feinstein AR: Clinical Epidemiology: The Architecture of Clinical Research. Philadelphia, WB Saunders, 1985
10. Fleiss JL: The measurement of interrater agreement. *In* Fleiss JL: Statistical Methods for Rates and Proportions, ed 2. New York, John Wiley & Sons, 1981
11. Friedman GD: Primer of Epidemiology, ed 4. New York, McGraw-Hill, 1994
12. Gordis L: Epidemiology. Philadelphia, WB Saunders, 1996
13. Hulley SB, Cummings SR: Designing Clinical Research in Epidemiologic Research, ed 2. Baltimore, Lippincott Williams & Wilkins, 2001
14. Isaac S, Michael WB: Handbook in Research and Evaluation: A Collection of Principles, Methods, and Strategies Useful in the Planning, Design and Evaluation of Studies in Education and the Behavioral Sciences, ed 2. San Diego, CA, Edits Publishers, 1981
15. Jacob RF, Carr AB: Hierarchy of research design used to categorize the "strength of evidence" in answering clinical dental questions. J Prosthet Dent 83:137–152, 2000
16. Jacob RF: [abstracts/commentary]. Journal of Prosthodontics 6:325–327, 1997

17. Jacob RF: [abstracts/commentary]. Journal of Prosthodontics 7:210–213, 1998
18. Jacob RF: [abstracts/commentary]. Journal of Prosthodontics 7:68–69, 1998
19. Jaeschke R, Sackett DL: Research methods for obtaining primary evidence. Int J Technol Assess Health Care 5:503–519, 1989
20. Kapur KK, Garrett NR, Hamada MO, et al: Randomized clinical trial comparing the efficacy of mandibular implant-supported overdentures and conventional dentures in diabetic patients. Part III: Comparisons of patient satisfaction. J Prosthet Dent 82:416–427, 1999
20a. Karlowski TR, Chalmers TC, Frenkel LD, et al: Ascorbic acid for the common cold. A prophylactic and therapeutic trial. J Am Med Assoc 231:1038–1042, 1975
21. Karlsson S: A clinical evaluation of fixed bridges, 10 years following insertion. J Oral Rehabil 13:423–432, 1986
22. Karlsson S: Failures and length of service in fixed prosthodontics after long-term function. A longitudinal clinical study. Swed Dent J 13:185–192, 1989
23. Koran LM: The reliability of clinical methods, data and judgments. N Engl J Med 293:695, 1975
24. Kramer MS: Clinical Epidemiology and Biostatistics: A Primer for Clinical Investigators and Decision-makers. Berlin, Springer-Verlag; 1988
25. McGrother CW, Dugmore C, Phillips MJ, et al: Multiple sclerosis, dental caries and fillings: A case-control study. Br Dent J 187:261–264, 1999
26. Medical Research Council: Streptomycin treatment of pulmonary tuberculosis. BMJ 2:769–782, 1948
27. Morris HF: Department of Verterans Affairs cooperative studies project number 147: Level of examiner reliability over seven years. Implant Dentistry 2:245–249, 1993
28. Palmqvist S, Swartz B: Artificial crowns and fixed partial dentures 18 to 23 years after placement. International Journal of Prosthodontics 6:279–285, 1993
29. Phillips C, Tulloch JF: The randomized clinical trial as a powerful means for understanding treatment efficacy. Seminars in Orthodontics 1:128–138, 1995
30. Reuter JE, Brose MO: Failures in full crown retained dental bridges. Br Dent J 157:61–63, 1984
31. Ross RB. MacNamera MC: Effect of presurgical infant orthopedics on facial esthetics in complete bilateral cleft lip and palate. Cleft Palate Craniofac J 31:68–73, 1994
32. Sackett DL, Haynes RB, Guyatt GH, et al: Clinical Epidemiology. A Basic Science for Clinical Medicine, ed 2. Boston, Little, Brown and Co, 1991
33. Sackett DL: Bias in analytic research. Journal of Chronic Diseases 32:51–63, 1979
34. Sacks H, Chalmers TC, Smith H Jr: Randomized versus historical controls for clinical trials. Am J Med 72:233–240, 1982
35. Scott BA, Clark GM, Hatch JP, et al: Comparing prospective and retrospective evaluations of temporomandibular disorders after orthognathic surgery. J Am Dent Assoc 128:999–1033, 1997
36. Scurria MS, Bader JD, Shugars DA: Meta-analysis of fixed partial denture survival: Prostheses and abutments. J Prosthet Dent 29:459–464, 1998
37. Shugars DA, Bader JD, White BA, et al: Survival rates of teeth adjacent to treated and untreated posterior bounded edentulous spaces. J Am Dent Assoc 129:1089–1102, 1998
38. Stamm JW: Types of clinical caries studies: Epidemiological surveys, randomized clinical trials, and demonstration programs. J Dent Res 63:701–707, 1984
39. Taguchi A, Tanimoto K, Suei Y, et al: Observer agreement in the assessment of mandibular trabecular bone pattern from panoramic radiographs. Dentomaxillofacial Radiology 26:90–94, 1997
40. Valachovic RW, Douglass CW, Berkey CS, et al: Examiner reliability in dental radiography. J Dent Res 65:432–436, 1986

Address reprint requests to
Rhonda F. Jacob, DDS, MS
MD Anderson Cancer Center
1515 Holcombe Boulevard,
Box 0441
Houston, TX 77030

e-mail: rjacob@mail.mdanderson.org