CHAPTER 1

The State of the Evidence in Implant Prosthodontics

Gary R. Goldstein
New York University College of Dentistry, New York, New York, USA

Introduction

Okay, you have been placing and/or restoring implants for numerous years and are pleased with your clinical outcomes and your patient acceptance of this exciting treatment modality. You attend a lecture or read an article about a new product or technique that claims to have a higher insertion torque, less bone loss, etc.; so, how do you decide if you should switch? The rubrics are very simple, so, whether you read a paper in a peer-reviewed journal, a non-peer-reviewed journal, or hear it in a lecture, the rules are the same for all three.

There is a multitude of information available to the clinician, some evidence-based, some theory-based, some compelling and, unfortunately, some useless. Evidence-based dentistry (EBD) gives one the tools to evaluate the literature and scientific presentations. It constructs a hierarchy of evidence which allows the reader to put what they are reading, or hearing, into perspective. As we proceed on this short trail together, I want to state that there is no substitute for your own clinical experience and common sense, and hope that when you are done with this chapter you will understand why. I am not here to trash the literature, rather to propose that not all published works are equal.

Hierarchy of evidence

EBD is a relatively new phenomenon that was introduced in the 1990s. It evolved slowly due to misunderstandings and misrepresentations of what it is and what it means, and, despite a slow start, has picked up traction and is now an ADA Commission on Dental Accreditation (CODA) requirement, mandatory in dental education and the backbone of clinical research and practice. Journal editors and reviewers are well versed in the process and less likely to approve the methodologically flawed project for publication, putting more pressure on the researcher to pay heed to research design.

I could say, "Here is the hierarchy of evidence (Figure 1.1)," and save us, you the reader and me the author, a lot of time, but unfortunately things are not quite that simple. Routinely, if one is asked what the best evidence is, the response would be a meta-analysis or systematic review and, not having that, a randomized controlled trial (RCT). What is also obvious from the figure is the categorization of animal and laboratory studies. While these present critical contributions to our basic knowledge and the background information needed to design clinical studies, they cannot and should not be utilized to make clinical decisions.

According to the Cochrane Collaboration,[1] a Systematic Review (SR) "summarises the results of available carefully designed healthcare studies (controlled trials) and provides a high level of evidence on the effectiveness of healthcare interventions"; and a meta-analysis (MA) is a SR where the authors pool numerical data. I want to bring your attention to the fact that nowhere in the definition does it mention, or limit itself to, RCTs. SRs and MAs are different from the more typical narrative review where an investigator evaluates all, or much, of the available literature and tenders an "expert opinion" of the results. They usually have loose or no inclusion and exclusion criteria and no "ranking" of the articles being reviewed. For those interested in how one categorizes articles, the following websites would be helpful:

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1891443/

http://www.nature.com/ebd/journal/v10/n1/fig_tab/6400636f1.html#figure-title

http://www.ebnp.co.uk/The%20Hierarchy%20of%20Evidence.htm

We can break down studies into analytic or comparative, those that have a comparative group (randomized controlled trials, concurrent cohort studies, and case control studies) and descriptive, those that do not have a comparative group (cross-sectional surveys, case series, and case reports). Descriptive studies give us useful information about a material, treatment,

Evidence-based Implant Treatment Planning and Clinical Protocols, First Edition. Edited by Steven J. Sadowsky.

Companion website: www.wiley.com/go/sadowsky/implant

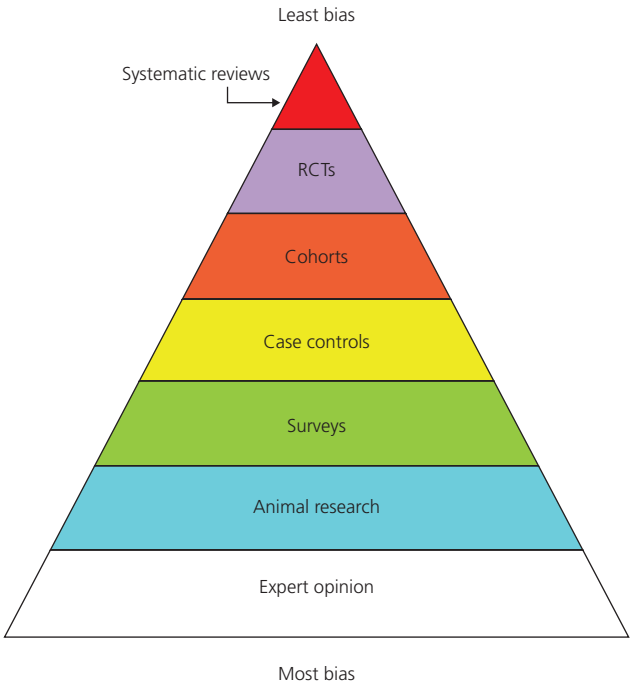

Figure 1.1 Hierarchy of evidence. *Source:* Adapted from http://consumers.cochrane.org.

etc., however, to determine if one material, treatment, etc. is better than another, requires a comparative study.

In addition, studies may be prospective or retrospective. In a prospective study the investigator selects one or more groups (cohorts) and follows them forward in time. In a retrospective study the investigator selects one or more cohorts and looks backwards in time. Prospective studies are considered superior since they can ensure that the cohorts were similar for possible confounding variables at the beginning of the study, that all participants were treated equally, and that dropouts are known and accounted for. Prospective studies allow for randomization or prognostic stratification of the cohorts. Retrospective studies can be very valuable and should not be minimalized, especially in uncovering adverse outcomes that have a low prevalence or take many years to become evident. The adverse effects from smoking have mostly been uncovered by retrospective investigation.

A randomized controlled trial (RCT) is a prospective, comparative study in which the assignment to the treatment or control group is done using a process analogous to flipping a coin. In reality, most projects are randomized utilizing a computer-generated random assignment protocol. The sole advantage of randomization is that it eliminates allocation bias. Feinstein[2] and Brunette[3] feel that the universal dependence on RCTs to achieve this is overestimated and prefer prognostic stratification of the matched cohorts for major confounding variables prior to allocation. What soon becomes obvious, however, is that prognostic stratification is not possible for every potential confounding variable, so only "major" ones are usually accounted for.

Doing an RCT is ideal, but has the constraints of time (can we afford to wait the numerous years necessary to design, implement and publish?) and cost (where can you get the funding?). Furthermore, RCTs are only ideal for certain questions, for example one that involves therapy. If our question is one of harm, it would be unethical to randomize a patient to something with a known harmful effect. To my knowledge, there has never been a RCT that proved smoking was harmful. Could you get an Internal Review Board (IRB) or Ethics Committee approval to assign participants to a group that had to smoke two packs of cigarettes a day for 25 years? Yet, does anyone doubt, given the mass of clinical evidence, that it is better to not smoke? Ultimately, the design is determined by the question.

Sackett,[4] considered by many to be the father of evidence-based medicine, in response to the heated dialogue over which design was the best, and in an effort to refocus the time, intellect, energy, and effort being wasted, proposed that "the question being asked determines the appropriate research architecture, strategy, and tactics to be used – not tradition, authority, experts, paradigms, or schools of thought."

Causation is one of the most difficult things to prove. It is like approaching a single set of railroad tracks. One can feel the warmth of the track and know that a train passed, but in which direction? It is why many studies conclude a "correlation." In the EBM series authored by the McMaster faculty, the Causation section published in the Canadian Medical Journal had David Sackett using the pseudonym Prof. Kilgore Trout as the corresponding author.[5] One can only wonder what motivated him to use the pseudonym rather than his own name to write on this critical topic. Sackett's love of the works of Kurt Vonnegut is well known and one might wonder if, in fact, his Canadian home named the Trout Research & Education Centre, is based on Kilgore or the fish?

While the design is critical, one must also determine the validity of the methodology. According to Jacob and Carr,[6] internal validity is a reflection of how the study was planned and carried out and is threatened by bias and random variation; while external validity defines if the results of the study will be applicable in other clinical settings.

Bias

There are many types of biases and a full explanation of the multitude reported is beyond the scope of this chapter. Still, there are a few that are meaningful to us as clinicians. We can divide bias into the following groups: the reader, the author, and the journal.

The reader

Taleb[7] used the following quote to accent that past experience is not always the best method to judge what we are doing at the present time.

> But in all my experience, I have never been in an accident…of any sort worth speaking about. I have seen but one vessel in distress in all my years at sea. I never saw a wreck and have never been wrecked nor was I ever in any predicament that threatened to end in disaster of any sort.
>
> *E.J. Smith, 1907, Captain RMS Titanic.*

The reader is almost always subject to confirmation bias, which is to believe whatever confirms one's beliefs. It was best stated by Sir Francis Bacon[8]: "The human understanding, once it has adopted an opinion, collects any instances that confirm it, and though the contrary instances may be more numerous and more weighty, it either does not notice them or rejects them, in order that this opinion will remain unshaken." People seek out research in a manner that supports their beliefs. We have all invested time, energy, and money getting a dental education. We have successfully treated patients and are loath to admit that something we have been doing is not as useful, successful, good, etc., as another product, technique, or procedure. This is a form of cognitive dissonance and a common human reaction. It is difficult for a clinician, and especially an educator, to admit that what they have been doing and/or teaching is not currently the best for our patients. Remember, we performed the procedure with "older" information and materials and are evaluating our outcomes or planning new treatment with "newer" evidence. The best recourse is self-reflection. Keeping up to date with clinically proven advances is our obligation as health providers.

The author

Allocation bias, a type of selection bias, is present when the two or more groups being compared are not similar, especially for confounding variables that could affect the outcome of the study. Familiar examples could be smoking, diabetes, osteoporosis, etc. Theoretically, randomization will account for this and is its major advantage, but only in the presence of a compelling number of participants (N).

The problem of allocation bias was demonstrated in a recent study.[9] The investigators were attempting to compare a one-stage protocol with a two-stage protocol with respect to marginal bone loss after 5 years; unfortunately the patients in the two-stage cohort were those who did not have a predetermined insertion torque at placement. As such, the two cohorts were not similar (one-stage = high insertion torque, two-stage = low insertion torque) for a major confounding variable and the internal validity of the study is in question.

Chronology bias refers to how long a clinical study ran and whether you, the reader, feel the time span was sufficient to justify the results and/or reveal expected or unexpected untoward responses. For example, company A has introduced a new implant surface that supposedly allows for faster osseointegration. How long would you expect the trial to run in order to accept the results as meaningful? What was their outcome assessment for success? Let's assume that there was a matched control with an adequate N. Since this is a human study, sacrificing the subjects to get histology would not sit well with your local IRB, but you have confidence that the selected outcome assessment is reliable. They have compelling evidence that a range of 3–6 months proved verifiable in their control group. They run a 6-month study with all subjects completing the full 6-month protocol. Do you feel the time is sufficient? Some will say yes and some would feel more comfortable allowing the study to run for 1 year to be certain of the external validity. Some might question whether the new surface will function under occlusal load and the biologic burden of the oral cavity and feel a multiyear protocol is needed.

In a study examining bone loss around implants, what time sequence would you require, 1 year, 2–3 years, 4–5 years, 5+ years? If a study examined the periodontal response to varying emergence profiles on implant-retained restorations, would you accept a shorter clinical trial than with the previous example? If the study was looking at monolithic zirconia that had surface custom staining and you were concerned about the outer glaze/stain wearing off, how many years would you expect the study to run? If you polled a group of experienced clinicians, you would get different answers to each, so who is correct? Unfortunately, EBD does not give you a definitive answer to this problem. It all comes back to your comfort with the premise and methodology, clinical experience, and need to alter your clinical regimen.

Referral filter bias is a type of selection bias and refers to where the patient is to be treated. For example, tertiary care cancer hospitals like M.D. Anderson or Memorial Slone Kettering have a different patient pool than you would expect to see in your private office. People who get on a plane and travel to the Mayo Clinic are not similar to the ones who are in your office because you practice close to where they work or live. Will the dental school patient be similar to yours? Will the patients in the office of a clinician who does external marketing be similar to yours, or vice versa?

Ideally, clinical projects should be triple blinded in that the person administering the therapy, drug, etc., the patient, and the person doing the outcome assessment are not familiar with what is being tested. The rationale is obvious. I am looking at my work and it all looks great, but you might not be so kind. It is one of the reasons that a case series performed and evaluated by the same group has a lower external and internal validity. In implant therapy studies, blinding often becomes quite difficult. If you are testing a zirconia abutment vs. a titanium abutment, there is a distinct visual difference that is hard to hide. Comparing a locator attachment with a ball attachment is another example that would be impossible to blind, as is one-stage vs. two-stage surgery. But, because the study is not blinded does not mean it is not well done and useful. Here we rely on the integrity of the researcher.

Conflict of interest (COI) is easily understood and it is now mandatory to disclose this in most journals. In a 2013 article in JADA,[10] the authors examined RCTs in 10 journals, three of which did not have mandatory reporting of COI, and found that "RCTs in which authors have some type of COI are more likely to have results that support the intervention being assessed." Here we have a major issue that needs to be addressed. Much of the implant clinical research we see is funded by commercial companies. In the US, it is apparent that the National Institute of Dental and Craniofacial Facial Research's (NIDCR) policy is to fund basic science research and allow companies to fund clinical trials. While this seems counter-intuitive, it is a fact of life. So we as clinicians, and the patients that we treat, are starting with a decided bias in the research being presented. Often SRs have eliminated RCTs for high risk of bias in other parameters but accept those with industry support. Burying our heads in the sand is not appropriate either. We must be realistic in how we evaluate all forms of bias. As with a lack of blinding, the internal validity suffers, but with both, one must assume the integrity of the researcher is intact. While there is a bad egg in every field, for the most part our colleagues are honest and sincere in their desire to do a study that will answer a needed question, undergo peer review, and stand the test of time. In today's digital world, once it is written it is there for all to see for all time.

I wouldn't have seen it if I didn't believe it. Sherlock Holmes, in the novel *A Scandal in Bohemia*,[11] said, "It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts." This is an unfortunate but sometimes unavoidable consequence for the researcher who has invested time, effort, and money in a project and perhaps is unable to clearly see what was happening. Sometimes it is innocuous, like framing the data in a positive manner. An example of this is the researcher who states they had a 70% success rate instead of a 30% failure rate, or justifying the lower success rate by saying that we save the patient time, money etc. But other times it takes on a more disconcerting approach, which could be a type of apophenia, which, according to the Merriam-Webster Online Dictionary is "the tendency to perceive a connection or meaningful pattern between unrelated or random things (such as objects or ideas)." One of the best examples was given by Cotton[12] in a 1988 editorial in the Journal of Dental Research, in which he described an experiment where a frog was trained to jump when told "jump". After one leg was amputated the frog was still able to jump when told. After two legs were amputated the frog was still able to jump when told. The same occurred after three legs were amputated. After the fourth amputation the frog could not jump so the researchers concluded that quadruple amputation in frogs created deafness. His example explains it all.

The journal

Publication bias is often defined as a preference to publish studies that have a positive finding, and it is true that most studies have historically been positive or neutral. In fact, a recent Cochrane Review[13] found that trials with positive findings were "published more often, and more quickly, than trials with negative findings." There are numerous potential reasons for this finding. Many researchers are reluctant to admit that their premise was incorrect, but these studies have just as much clinical value as positive ones. The negative result always creates a conundrum when the study was funded by a company who now wants to squash the publication and the researcher does not want to risk the loss of future grants. We, as clinicians, should be accepting and thankful to our research colleagues who publish despite these concerns.

There is another form of publication bias which may be imposed by the editor and/or journal board. There have been numerous articles that were rejected by one journal and, when published by another, have gone on to be highly quoted and generated a respectable citation index. Our history is riddled with unfortunate examples of work that had difficulty getting published: gastric ulcers are caused by bacteria; lactic acid buildup in muscles being exercised is good; lobotomies on patients suffering from chronic pain are unacceptable; carotid ligation in people who have suffered a stroke is not indicated; and all teeth that have had root canal therapy, regardless of the remaining tooth structure, should not be decoronated and a post placed 3 mm from the apex to support a core that replaces the removed tooth structure.

Some editors are extremely rigid, in that they require adherence to editorial demands, and others are more accepting, feeling that they have performed their duty in helping to upgrade the quality of the publication and if the authors refuse they would allow reader opinion and history to resolve the conflict. Either way, the journal is not giving a "stamp of approval" to each article published and, again, it is up to the reader to decide if the results are applicable to their patients.

Statistics

Pundits have been famous for analyzing data and coming up with fallacious conclusions. The most famous example is the Chicago Tribune's headline in 1948 predicting that Dewey won the presidential election over Truman. Issues like the mortgage crisis of 2008–2009 and the current debate over the US student loan crisis, show the constant problems with data analysis.

Statisticians are a testy and argumentative group and are almost as bad as prosthodontists. Much of the early work on statistics was derived by intellectuals trying to improve their gambling odds.[14] Many of their theories are based on non-medical protocols which do not translate into the clinical

milieu. Statisticians are like bad relatives; we may not want them but we have no choice. They are a critical part of our lives. They help us to determine the N and power of a project and to clarify data collection methodology. They analyze the data to help us interpret and understand the results, but statisticians do not make clinical decisions. Clinicians make the decisions by utilizing the best available evidence to treat the patient. The statistics section should be the shortest and least obtrusive part of the publication, unless some new statistical method is being presented.

Not all statisticians agree on which tests to run on specific issues and many of us have had a journal ask us to rerun data because "their" statistician did not agree with "our" statistician. Some tests are considered more robust than others and sometimes a more rigorous test might show something different. Sometimes the wrong statistics were performed. But sometimes the request is unwarranted or even unreasonable.

There is a difference between statistical significance and clinical significance. Feinstein[15] opined that "statistical significance has become a malignant mental pathogen" as it does not take into consideration the methodology, the clinical implications, or the cause of the difference. Say you now possessed an extremely accurate instrument that could measure bone loss in the nanometer range and were able to show that the mean bone loss after 5 years was 10 nanometers more with implant B when compared to A and the difference was statistically significant, is that clinically significant? In this case no, but in many projects clinicians will have different opinions as to the clinical relevance of the data presented. If we now had a 0.5 mm difference, how would you respond? Or a 1 mm difference?

What does the mean mean? To paraphrase an example given by Wheelan,[16] after reading the latest contract for the sanitation workers in New York City, nine senior dental faculty members went to happy hour at a local bar to share their angst over the current data. When adding up their clinical supervision time, lecture, and/or seminar time, preparation time, research, and personal development time, and comparing it with the hours worked by the sanitation people, it was obvious that the mean salaries of those at the bar were less than the sanitation workers and required another round of drinks. Unbeknownst to them, Warren Buffett walked in and ordered a drink; suddenly the mean salary of the 10 people sitting at the bar skyrocketed to an obscene number. Ok, so you say eliminate Warren as he is the outlier, definitely an accepted technique for statisticians and it indeed makes sense. But what if we are doing a medical or dental clinical trial?

Biostatistics should have different rules than other forms of statistics. Should the data be manipulated? Homogenizing the data by log transforming it or eliminating outliers is acceptable in large survey studies to conform to the Gaussian curve, but is not appropriate in medical and dental clinical studies, especially those with small Ns.

Let's say I'm doing a measurement of pocket depth around implants after 1 year of clinical loading and I decide to take three readings with a periodontal probe and average the three for each patient that I am analyzing. I have 10 patients on whom I am going to take the measurements. If my data set is running between 5 and 6 mm for all of the patients, then using the average of three readings (5.5 mm) would probably suffice. But picture the scenario where I'm getting readings of 0 mm, 5 mm, and 10 mm. It is obvious that by using the mean (5 mm) I've eliminated a major variation (was it the probe, me, or variations in the depth manifested by not being able to be in the same exact location?) which could potentially affect the statistical validity of my sample. What should be done is to enter all three measurements as a subset of the patient. The mean will still be the same but the standard deviation will be much larger in the 0, 5, 10 group which could affect the statistical significance. The statistical programs easily handle this, but many research projects are not designed in this manner.

The over-reliance on statistics that cannot truly assume the effect of chance is also questionable. Derek Richards[17] stated that "when testing a new treatment in a clinical trial, there are three possible explanations for why it did or did not work as expected – chance, bias or the truth."

Chance and the N have a way of intertwining. If I asked you to flip a coin, you would say you have a 50-50 chance of getting a head or tail. If I told you that I flipped the coin and got five heads in a row you might not be surprised. Would you still bet 50-50 on the next flip? If I told you that I had flipped nine heads in a row, some of you might say *okay it's possible* and some of you might say *no way*. But, if I told you that we were going to flip the coin 1000 times everybody would pretty much agree that we would get 500 heads and 500 tails. Would you put a wager on that? John Kerrich, a mathematician who had the bad luck to be spending time in a German prison during World War II, had the time to do 10 000 coin tosses. He had 44% heads after 100 throws and 50.67% at 10 000.[18] So how does chance come into play in a clinical trial with an N of 10 or 20?

The N is determined *a priori* by the clinician who helps determine the expected clinical difference with input from the statistician who helps determine the estimated number. It can also be verified *post hoc* by a power analysis. But does the N make sense to the clinician? One of the biggest problems we see, especially in implant literature, is the inadequate N despite the fact that large numbers of patients are in the clinical trial. Let's look at a clinical trial to determine if implant A will integrate faster using bone grafting material X or Y in the maxillary premolar area, utilizing a split-mouth design, where the patient is their own control, with an N of 40. In this situation the age, sex, medical and dental history, medications etc. are the same for the test group and the control group. Since the confounding variables are equally

distributed, the split mouth has the advantage of requiring a smaller N and you may agree that 40 patients are adequate.

In another clinical example, with an N of 100 patients, implants were placed in the maxilla, in the mandible, some were premolars, some were molars, and some were anteriors. In addition, some implants were 8 mm, some were 10, and some were 12. Some implants were wide body, some regular body, and some narrow body. And then, to make matters worse, different brands of implants were also used. So if I am looking for data to determine whether or not I can use a regular body, 10 mm, brand X implant in a maxillary premolar on a 40-year-old female nonsmoker, there may only be an N of one or two whose results are applicable to my patient. This shotgun technique of patient allotment violates basic EBD principles in that a well-defined question was not established or that too many questions (Is brand A better than brand B? Does implant width make a difference? Does implant length make a difference? Does the arch or tooth position matter and what about smoking, age or sex?) were trying to be answered. Each time you add another variable you have to double the N needed, so here 100 patients is not an adequate number.

How do you handle dropouts? One of the key questions in evaluating the validity of a research project is "Were all patients who entered the trial properly accounted for and attributed at its conclusion?"[19] Patients pass away (hopefully not from our dental treatment), move to other areas, or become too sick to return for follow-up. But patients can also drop out because they are unhappy with the treatment, the clinician, or the environment in which the treatment occurred. Patients can also be noncompliant with the protocols such as taking medication or using a prescribed home-care regimen to which they were randomized.

The classical manner of handling dropouts is the "intention to treat" method in which all subjects are followed regardless of adherence. Sackett[20] espoused it and Montori and Guyatt,[21] in a more recent commentary, lambasted alternative strategies. On the other hand, Gerard Dallal[22] in his *Handbook of Statistical Practice* called the "intention to treat" a fraud, and gives many examples where it is severely flawed. But he also questioned the per protocol, in which only data from adherent subjects are analyzed, as well as some other varied attempts to deal with the problem.

For our purposes, say you are doing a study to determine which postoperative antibiotic "regimen", in patients getting immediate placement of an implant, is more efficacious. You have three groups: test group regimen A; test group regimen A and regimen B; and placebo group C. Mr. Smith, who was randomized to group A, never took his medication. "Intention to treat" demands he be included in the data for group A. Are you comfortable with that? This argument will endure for many years before there is any hope of a settlement. It is the researcher's obligation to decide how dropouts will be handled in the protocol stage and however they choose to do this, the number and reason for the dropouts needs to be clearly reported. At the end of the day, dropouts are a problem that affects the internal and external validity of a study. Sackett[23] has said that "it would be unusual for a trial to withstand a worst-case scenario if it lost more than 20% of its patients." It is you the clinician who must decide if you are comfortable with the number of dropouts and how they were handled.

How long should you run a clinical trial to avoid "follow-up not complete"? Some patients will start in the first year of the study, some in the second, etc. The reality is that not every patient starts on day 1, as this is a clinical trial and not a horse or a car race where everyone starts at the same time. If you do a 5-year study and only a small percentage of the patients were treated for 5 years, should you still call that a 5-year study? Statisticians will say okay and that there are formulae that they could use to predict what will actually happen, but how do you feel? In the typical dental studies with small Ns, where chance and outliers wreak havoc, why not just wait until everyone completes the study? It creates numerous problems, not the least of which is time and money, for the research staff, but if everyone does not complete the study, once again we risk putting the decision-making process in the hands of the statistician rather than the clinician.

A major concern in evaluating the outcome of a study, especially on implants, is what was the outcome assessment? Was it the implant, or was it the patient? It all depends on the question. If your question is *Will this new implant integrate?*, then the implant should be the outcome measure. As an example, you are reading a study on implant survival in 30 patients who have six fixture prostheses in the maxilla. At the end of 5 years, seven patients lost two implants, seven patients lost one implant, and one patient lost five implants. Only one patient had the prosthesis compromised sufficiently to require a redo, the patient with five lost implants. The other patients had the failed implants removed from the mouth and the existing prosthesis was deemed usable. If you use the patient as the outcome assessment then only one failure occurred. But, if you use the implant as the outcome assessment, then 26 out of 180 implants failed. The result is two totally different data sets; two totally different conclusions.

In a recent article in the Journal of Dental Research,[24] a retrospective cohort study was conducted to determine the effect of selective serotonin reuptake inhibitors (SSRIs) on implant survival. In this study the question really is *Will SSRIs inhibit osseointegration?*, so the patient should be the outcome measure since the medication affects the patient. The study showed a failure rate in the SSRI group of 10.6% (10/94 failed) and 4.6% (38/822) in the nonuser group, utilizing implants as the outcome assessment. The researchers, to their credit, understanding that SSRIs would affect the patient, ran a separate statistical analysis "to account for cluster effects of multiple implants when placed and evaluated in a single patient." Here you get to see the data both ways and, regardless of your opinion on the outcome assessment, the ability to apply the conclusions.

We seem to have an immense fascination with the Gaussian bell-shaped curve which has lately come under a great deal of criticism. Carl Friedrich Gauss, the German mathematician for whom the bell-shaped curve is named, placed his "proof" inconspicuously in a section at the end of his book *The Theory of the Motion of Heavenly Bodies Moving about the Sun in Conic Sections*. Interestingly, he later considered that proof invalid.[25] Both Feinstein[26] (*On Exorcizing the Ghost of Gauss and the Curse of Kelvin*) and Taleb[27] (The *Bell Curve, That Great Intellectual Fraud*), devoted entire chapters to this.

While data dredging has always been a concern, the user friendly statistical programs currently available have enabled the dredger to easily run a multitude of tests until they finally find one that proves a premise. This has become a more prevalent issue and unfortunately, it may also involve culling or manipulating data and is commonly referred to as p-hacking. While most of the time it is used to try and prove a difference, it can also be used to do the opposite.

Understanding that the P value is a probability of the likelihood that what is seen occurred by chance, it does seem counter-intuitive to say that a P = .05 is statistically significance but a P = .51 is not. Today it is recommended to look at confidence intervals and see if, and how much, they overlap. While there are formulae to help derive this, it should not be the reader's obligation to do math when analyzing an article. If not provided by the author, the confidence interval (CI) can be simply viewed as two standard errors. For example, let's take a data set where we have a mean of 10 in group A and a mean of 20 in group B. If the standard error (SE) around group A and B are plus or minus 2, then the confidence interval around group A would go from 6 to 14 (2× the SE) and the confidence interval around B would go from 16 to 24. Since the confidence intervals are not overlapping, one can be assured that the groups are different. But let's create a sample where the standard error is plus or minus 5 for both groups. Now the confidence intervals stretch from 0 to 20 in group A and 10 to 30 in group B. Since the confidence intervals are overlapping, in this case severely to demonstrate the point, even if the data were statistically significant (which I doubt, given the exaggerated standard errors in this scenario), one would be tempted to be concerned about the data sets. If there is a large overlap there is a large concern, if there is a small overlap there is much less concern, and, again, if there is no overlap there is no concern.

Classical statistics usually follows the Neyman–Pearson approach, but there is much controversy in that and many people are looking to the Bayesian approach which was first proposed by the Reverand Thomas Bayes in 1763.[28] It states that the probability that A will occur if B occurs is often different than the probability that B will occur if A occurs. The Bayesian theory contradicts standard statistical analysis by bringing prior probability into the equation. Simplistically, it means that if you just look at the data set in and of itself without having background information upon which to apply or how to apply that data set, you will come to a potentially incorrect conclusion. An example is given by Mlodinow.[29] He applied for life insurance and took a routine blood test which came back HIV positive. His doctor told him that he had 1 in 1000 chances of being healthy, since HIV tests will give a false positive in only 1 out of 1000 samples. But, the confusion is that his doctor assumed he would test positive if he was not HIV positive with the chances that he would not be HIV positive if he tested positive. Hence, he was looking at the chance he was not infected out of all negative and positive tests, rather than the chance that he was not infected just out of all positive tests. In order to understand the example, it is important to note that he is a white American, heterosexual male, non-IV drug user and according to the Center for Disease Control (CDC) data, only 1 in 10 000 people in that data set was infected with HIV. Therefore, given that the false-negative rate is almost zero, we can deduce that in the 10 000 men in the proposed sample, 9989 will be testing negative. If we look at the people who tested positive, 10 will be false-positives (1 in 1000 false-positive rate) and one will be a true positive (1 in 10 000 prevalence); so rather than a 1 in 1000 chance that he is HIV positive, his chances are 10 out of 11 that he is not.

Another example given by Siegfried in Science News[30] had to do with steroid testing of baseball players. Using an assumption that the test is 95% accurate and one of the players on your team tested positive, the probability of guilt should be 95%. But using the Bayesian approach you need to know some additional information. Previous data on this type of testing showed that 5% of professional baseball players use steroids. He proposes on a test of 400 players, 20 would be users (the 5%) and 380 would not be users. So, giving a test to all 400 that is 95% accurate, of the 20 users 19 would be identified and, of the 389 nonusers, 19, or 5%, would be incorrectly identified. So testing 400 players would give you 38 positives, 19 of whom were users and 19 of whom were not users. Your player has a 50% chance of being guilty.

Since the classical versus the Bayesian disagreement transcends my pay grade, those of you with interest in this topic can pursue it with the references stated, and others, which can easily be found on the Internet, and I will allow those with more knowledge than me to continue this debate.

Researchers devise questions and in simplistic terms, "will A be better than B?" or, "will A last longer than B?". The statistical consultant desires a null hypothesis to do their data analysis, but why must the reader have to deal with inverted statistical logic? Which is more intuitive, "A is better than B" or the "null hypothesis was rejected"? Perhaps the null hypothesis, which is confusing jargon forced upon us by statisticians, needs to be "rejected" and be null and void in the manuscript conclusion.

Yogi Berra, the great philosopher and Hall of Fame baseball player and coach said, "It is tough to make predictions, especially about the future."[31] Few of us have the background to truly analyze the statistics being utilized in today's clinical studies. What are needed are meaningful answers to our clinical questions,

not fancy data manipulation that could possibly obscure the facts we seek. A memorable quote from an esteemed mentor and friend, Dr. Louis Blatterfein, was "if you have nothing to say dazzle them with your footwork." Statistical analyses have to make sense. If you are concerned that the methodology is questionable then don't worry about the statistics; you are not accepting the premise and/or the results. A flawed project cannot be salvaged by exotic statistical manipulation. Remember, "garbage in – garbage out". If the methodology is sound, you can assume the statistics are also.

Evaluation

Sackett[32] states, "Evidence-based medicine is not restricted to randomized trials and meta-analyses. It involves tracking down the best external evidence (from systematic reviews when they exist; otherwise from primary studies) with which to answer our clinical questions." In addition, not all SRs are well done and articles have been written on how to evaluate them.[17,33]

If what you are doing has a 95% success rate, unless you had compelling evidence from a well-designed RCT, why would you change? If what you are doing has a 30% failure rate, why are you still doing it? But, suppose you prefer to frame it in a more positive manner and say a 70% success rate, and you have no other treatment options, can you or your patient wait for the RCT? Each one of us might put a different number for the percentage success rate you would accept or not accept. Here we are directed to the best available evidence, which unfortunately in dentistry may be a case series.

The key aspect to having an evidence-based practice is to be able to critically appraise the article you have been reading. When evaluating an article you look at the methodology first. In the study, was the patient population similar to the patient you are treating? Is the operator expertise similar to your own? Is your environment similar to the one in which the study was performed? Are your inclusion and exclusion criteria for treatment the same? You need to look at the design, the biases, the statistical methods, and the conclusions to make a judgment using your EBD tools to evaluate the internal and external validity and determine *are the results applicable to my patient*? Two people can read an article and after critically appraising it come up with different opinions of its clinical usefulness. And there is nothing wrong with this.

Clinicians function in a different environment than researchers. The researcher is looking for statistical significance, while the practitioner demands clinical significance. In an effort to standardize their cohorts, the researcher has stringent inclusion and exclusion criteria and works with mean populations. Clinicians treat the standard deviation, rarely the mean, and have people in their practice who fall outside of the inclusion/exclusion criteria of the study. We treat the diabetic, the smoker, the pregnant women, the neurologically impaired, and the patient on a wide assortment of medications, etc. What if the patient was referred by your best referring doctor or is the relative of an existing patient who has sent many friends, relatives, and business associates to you? Clinicians also have the concerns of litigation, since they are not under a university or hospital umbrella, and the possibility of the everlasting and insidious negative internet review.

When evaluating a new procedure or product we should never be using only one outcome assessment. Let's look at implant placement procedures A and B. Certainly, implant failure or success is a primary outcome, but what other clinical parameters does the clinician need to take into consideration. For our discussion, A equaled B in terms of failure/success. But, how many surgical procedures were involved? What was the morbidity? What was the cost to the patient? Did one procedure require a shorter treatment time than another? While the implants were still in place, was there more bone loss or soft-tissue issues for one procedure? What other patient management issues are important to you?

Many factors come into play when trying to determine treatment for our patients. Fretwurst *et al*.[34] in an October 2014 article, found residual DNA in the allografts tested. So, what is the clinician to do? Is there a critical mass for residual DNA? What do the numbers mean? Is there clinical evidence of any harm? Is that because we never looked for it? How do public perception and legal consequences enter into our clinical decision-making process?

Is a poorly-done RCT of more value than a well-done non-randomized controlled trial? The answer is unequivocally no. But, is it better or worse than a well-done case series? Here we will get into some disagreements. If you have a few case series that have shown a 95% success rate over 10 years, that is compelling evidence that needs to be accounted for. If they present with a 50% failure rate that is also critical evidence that should not be discarded. Sackett tells us that we should use the best available clinical evidence.[4]

A very significant problem in the literature and one that can have unfortunate consequences in the medical–legal and insurance arenas is misstating the conclusions of a MA or SR. In an article published in Evidence Based-Dentistry in 2010, *Indirect or direct restorations for heavily restored posterior adult teeth*, one RCT that compared composites with crowns on root canal treated premolars was rejected because the clinical scenario was a vital tooth. Two prospective studies comparing large amalgams with crowns (a 5-year and a 17-year follow-up) were rejected because they were not randomized. The author concluded, "The clinician can only say that there is no high quality clinical evidence to suggest that placing a crown on a posterior tooth would lead to its longer retention than a composite or amalgam."[35] This is extremely dangerous as insurance companies, and perhaps government agencies and the press, latch onto this information and misuse it, preventing practitioners from providing the care they feel appropriate for their patients. *Sorry Doc, we won't pay for crowns since there is no evidence that they are better than a composite or amalgam*.

Here we must decide what the question was. "Does it work?" only requires a case series. "Is it better?" requires a comparative study. Since, crowns and amalgams have been in use for many years the onus would be to prove that the composite was as good as a crown or an amalgam in the tooth involved. A major variable in determining which restorative material is indicated is the amount of remaining tooth structure and the opposing occlusion. Also, is "longer retention" the needed outcome assessment? What if the restoration or residual tooth structure fractured? What if the tooth devitalized? What if the tooth wore down occlusally and the opposing tooth extruded? What if there was recurrent decay or periodontal issues? Would "there is no high quality evidence that supports or rejects the practice of placing a crown or onlay on a vital posterior tooth rather than a composite or amalgam restoration to ensure longer tooth survival" have been a better conclusion? Or perhaps, *more clinical research, especially RCTs, is needed*. Absence of evidence is not evidence of absence,[36] especially if a good portion of the evidence has been excluded.

A positive example was a review on whether or not occlusal splints should be routinely prescribed for bruxers undergoing implant therapy.[37] The authors concluded, "The absence of evidence-based studies to recommend occlusal splints in bruxers who have received implant-supported rehabilitation emphasizes the need for well-designed randomized controlled clinical trials." So why am I pointing this out? The conclusion seems valid and well founded and I agree. Unfortunately, many authors would say something like "there is no evidence that an occlusal splint should be prescribed in bruxers undergoing implant therapy," which, if someone was just reading the conclusion would lead them to believe that occlusal splints are contraindicated. This occurs quite often where authors conclude that, since there is inadequate evidence to support a premise, that the premise is fallacious.

Another problem is that someone can perform a systematic review with only RCTs as inclusion criteria and, despite a multitude of clinical trials that were not RCTs, draw a conclusion that there is no evidence to support the question asked. This has limited value, especially in the US, as NIDCR has been reluctant to fund RCTs. Yes, we need them, but who will fund them? Should we depend on industry to fund our RCTs? Will they fund the project whose premise is that the product may not be good? Well-done cohort studies or even case series may be the best available evidence and have significant value. If there are no RCTs, or no well-done RCTs, then the author is obligated to follow the trail to the best available evidence. There is no requirement that a SR needs to only look at RCTs! If you have numerous case series that have shown a 95% success rate over 10 years that is compelling evidence that needs to be accounted for. If they present with a 50% failure rate that is also critical evidence that should not be discarded.

When reading a SR or a MA, what were the inclusion and exclusion criteria? Do you agree with them? Are you comfortable with the ones excluded? Should they have been?[33] Some SRs will include discussions with recognized experts, especially ones who may be performing current research on the topic being reviewed, as well as Grey Literature, which are articles not published in peer-reviewed journals. What is your position on that?

We also have the possibility of committee bias. Are evaluators on a committee evaluating their own work? Are they receiving grants, stipends, or other forms of corporate support? They may be esteemed experts, but should they be on the committee? Even if they recuse themselves from the discussion on their particular paper, what is the risk of bias for the committee evaluations?

Evidence-based practice does not mean that you have to wait for a MA to make decisions. In order to do a MA you need data that can be pooled. Because there are no or few randomized controlled trials available, doesn't mean there is no evidence. If evidence cannot be pooled then the SR is more than adequate. If one finds that there are a few or no RCTs available then one can broaden the inclusion criteria. This is where a critically appraised topic (CAT), which is a defined critical summary of research evidence that answers a clinical question, may be more helpful. Evaluating the available evidence is more fruitful than saying we need more RCTs.

Some journals are moving away from the case series and/or case presentations in favor of RCTs and SRs, and, while it is improving the status of the journal, is it really improving dental care? Where does the innovator publish? In order to get funding for RCTs, there needs to be justifiable evidence to support the researcher's premise. A problem facing dentistry is the inability, given demands of EBD, for imaginative thoughts to have a place to be published. Medicine has recognized this problem and created a Journal of Medical Hypothesis which caters to original ideas that can be the basis for future research rather than RCTs, SRs, and MAs. We in dentistry need to follow suit. Historically most of our articles were expert opinion and much of it did not stand the test of time but, if we cut off the essay or the case presentation because there is no place for it to be published, are we losing the innovation necessary for us to grow? Prospective clinical studies are needed. So please do not misunderstand the thought process here. We still are obligated to make clinical decisions on the highest level of research available, but can we risk cutting off the innovative sparks that may lead us in the future?

Conclusion

EBD gives you, the clinician, the tools to run an evidence-based practice. Once you have ascertained that the results are applicable to your patient, you now have to determine if the results are valid and compelling enough to allow you to feel comfortable applying them. You have earned a BS or BA and a DDS or DMD. Many of you have Master's degrees, and perhaps PhDs. You have taken general practice residency programs, specialty programs, and perhaps specialty certification exams. You have a lifelong commitment to continuing education, attend lectures

and seminars, and read professional journals. You have spent years honing your clinical and patient-management skills. You are indeed the real clinical scholars and you make the clinical decisions. EBD is a tool; it can never replace your skill, experience or judgment.[38]

References

1. Cochrane Consumer Network. What is a systematic review [internet]. [updated 2012 Jun 25; cited 2015 June 22]. Available from: http://consumers.cochrane.org/what-systematic-review
2. Feinstein AR. An additional basic science for clinical medicine: II. The limitations of randomized trials. *Ann Intern Med* 1983; **99**(4): 544–550.
3. Brunette DM. *Critical Thinking: Understanding and Evaluating Dental Research*. Chicago: Quintessence Publishing Co 1996; 152.
4. Sackett DL, Wennberg JE. Choosing the best research design for each question: it's time to stop squabbling over the "best" methods. *BMJ* 1997; **315**(7123): 1636.
5. Department of Clinical Epidemiology and Biostatistics, McMaster University Health Science Center. How to read clinical journals, IV: to determine etiology or causation. *Can Med Assoc J* 1981; **124**(8): 985–990.
6. Jacob RF, Carr AB. Hierarchy of research design used to categorize the "strength of evidence" in answering clinical dental questions. *J Prosthet Dent* 2000; **83**(2): 137–152.
7. Taleb NN. *The Black Swan*. New York: Random House Trade Paperbacks 2010; 42.
8. Mlodinow L. *The Drunkard's Walk: How Randomness Rules Our Lives*. New York: Pantheon Books 2008; 189.
9. Berberi AN, Tehini GE, Noujeim ZF, *et al*. Influence of surgical and prosthetic techniques on marginal bone loss around titanium implants. Part I: Immediate loading in fresh extraction sockets. *J Prosthodont* 2014; **23**(7): 521–527.
10. Brignardello-Peterson R, Carrasco-Labra A, Yanine N, *et al*. Positive association between conflicts of interest and reporting of positive results in randomized clinical trials in dentistry. *JADA* 2013; **144**(10): 1165–1170.
11. Kameron Kent Searle. Sherlock Holmes Quotes. 2012–2015 [cited 2015 June 22]. Available from: http://sherlockholmesquotes.com/
12. Cotton WR. How far can a frog jump? A current assessment of pulp biology research. *J Dent Res* 1988; **67**(9): 1251.
13. Hopewell S, Loudon K, Clarke MJ, *et al*. Publication bias in clinical trials due to statistical significance or direction of trial results. *Cochrane Database Syst Rev* 2009; **1**: MR000006. doi: 10.1002/14651858.
14. Mlodinow L. *The Drunkard's Walk: How Randomness Rules Our Lives*. New York: Pantheon Books 2008; 48–50.
15. Feinstein A. *Clinical Biostatistics*. St. Louis: The C.V. Mosby Company 1977; 11.
16.. Wheelan C. *Naked Statistics: Stripping the Dread from the Data*. New York: W.W. Norton & Company 2013; 19.
17. Richards D. Critically appraising randomised trials. *Evid Based Dent* 2009; **10**(3): 88–90.
18. Mlodinow L. *The Drunkard's Walk: How Randomness Rules Our Lives*. New York: Pantheon Books 2008; 95.
19. Goldstein GR, Preston JD. How to evaluate an article about therapy. *J Prosthet Dent* 2000; **83**(6): 599–603.
20. Sackett DL, Richardson WS, Rosenberg W, Haynes RB. *Evidence-Based Medicine: How to Practice & Teach EBM*. New York: Churchill Livingstone 1997; 96.
21. Montori VM, Guyatt GH. Intention-to-treat principle. *Can Med Assoc J* 2001; **165**(10): 1339–1341.
22. Dallal G. *The Little Handbook of Statistical Practice*. [Internet]. 2015 [cited 2015 June 22]. Available from: http://www.jerrydallal.com/LHSP/LHSP.HTM.
23. Sackett DL, Richardson WS, Rosenberg W, Haynes RB. *Evidence-Based Medicine: How to Practice & Teach EBM*. New York: Churchill Livingstone 1997; 95.
24. Wu X, Al-Abedalla K, Rastikerdar E, *et al*. Selective serotonin reuptake inhibitors and the risk of osseointegrated implant failure: a cohort study. *J Dent Res* 2014; **93**(11): 1054–1061.
25. Mlodinow L. *The Drunkard's Walk: How Randomness Rules Our Lives*. New York: Pantheon Books 2008; 143.
26. Feinstein A. *Clinical Biostatistics*. St. Louis: The C.V. Mosby Company 1977; 229–242.
27. Taleb NN. *The Black Swan*. New York: Random House Trade Paperbacks 2010; 229–252.
28. Brunette DM. *Critical Thinking: Understanding and Evaluating Dental Research*. Chicago: Quintessence Publishing Co 1996; 172–174.
29. Mlodinow L. *The Drunkard's Walk: How Randomness Rules Our Lives*. New York: Pantheon Books 2008; 114–117.
30. Siegfried T. Odds Are, It's Wrong: Science fails to face the shortcomings of statistics. Science News [internet]. 2010 March 27 [cited 2015 June 22];177(7):26. Available from: https://www.sciencenews.org/article/odds-are-its-wrong.
31. Taleb NN. *The Black Swan*. New York: Random House Trade Paperbacks 2010; 136.
32. Sackett DL, Richardson WS, Rosenberg W, Haynes RB. *Evidence-Based Medicine: How to Practice & Teach EBM*. New York: Churchill Livingstone 1997; 4.
33. Felton DA, Lang BR. The overview: an article that interrogates the literature. *J Prosthet Dent* 2000; **84**(1): 17–21.
34. Fretwurst T, Spanou A, Nelson K, *et al*. Comparison of four different allogeneic bone grafts for alveolar ridge reconstruction: a preliminary histologic and biochemical analysis. *Oral Surg Oral Med Oral Pathol Oral Radiol* 2014; **118**(4): 424–431.
35. Hurst D. Indirect or direct restorations for heavily restored posterior adult teeth? *Evid Based Dent* 2010; **11**(4): 116–117.
36. Sedgwick P. Understanding why "absence of evidence is not evidence of absence". *BMJ* 2014; **349**: g4751.
37. Mesko ME, Almeida RCCR, Porto JAS, *et al*. Should occlusal splints be a routine prescription for diagnosed bruxers undergoing implant therapy? *Int J Prosthodont* 2014; **27**(3): 201–203.
38. Sackett DL, Richardson WS, Rosenberg W, Haynes RB. *Evidence-Based Medicine: How to Practice & Teach EBM*. New York: Churchill Livingstone 1997; 5.