EVIDENCE-BASED DENTISTRY SERIES

Measurement in dentistry

Alan B. Carr, DMD, MS,[a] and Glen P. McGivney, DDS[b]

CLINICAL SCENARIO

You have been working with a new hygienist for the past 6 months and have expanded her responsibilities to include taking a more active role in screening patients for needed treatment. Although you like the opportunity this provides to the growth of the practice, you frequently differ with her opinion on needed treatment. You feel she is routinely suggesting treatment options that are not appropriate and suggesting more treatment than needs to be considered. You share this concern with your practice partner who tells you that his experience with the new hygienist has been just the opposite, that she does not suggest enough treatment.

These differences of opinion cause some concern so you decide to monitor the situation more closely for the next few months. During this period you discover that the teeth most often responsible for the differing opinions are teeth that have existing restorations and you wonder whether this is a common occurrence. During the monitoring period, an article appears in a national weekly publication describing the variability in treatment decisions by different dentists. This has created a significant concern in your community and you are asked by the local newspaper to comment on the article. You now face a problem similar to the one you are experiencing in your office, except now you must respond to the local community on behalf of your profession. You decide to research "variations" associated with "diagnosis and treatment planning" in dental health care. Specifically, you would like to be able to respond as to how common clinical disagreement is, the reasons for and characteristics of clinical disagreement, and how the level of disagreement might be reduced. It seems obvious to you that answering these questions will also help solve the problem in your office.

You spend some time at the computer conducting a search of the literature using the MESH headings: dentistry, dental history and physical examination, dental examination, diagnosis, treatment planning, patterns of treatment, examiner variability, and examiner reliability. Your search found 12 articles that fit these MESH headings. A review of the abstracts, however, revealed only 1 article ("Agreement Among Dentists' Recommendations for Restorative Dentistry," by Bader and Shugars[1]) that could be helpful in answering your questions and, at the same time, prepare you to respond to the press as well as solve the problem with your hygienist regarding the treatment needs of the patients she is screening.

INTRODUCTION

Health care providers are trained to recognize the health status of patients and to prescribe sequenced treatment plans for their health restoration or maintenance. Observations from clinical and laboratory examinations are used for the purposes of making diagnoses, prescribing treatment, and monitoring patient responses. These observations provide the recorded characteristics needed for comparison with existing health standards or previous observations. The rules and guidelines that provide the qualitative and quantitative order to these recorded characteristics are presented and discussed as principles of measurement. In practice and in the medical/dental literature, measurements of clinical signs and symptoms provide the basis for analysis of any clinical condition or its reporting in the literature. Methods used for obtaining measurements can be simple or sophisticated, but this distinction does not impart any protection from errors associated with the measures obtained.[2,3]

This unit in the Evidence-Based Dentistry series discussed the basic principles associated with making clinical decisions based on measurements. The reader is encouraged to review or have available for reference, the glossary published in the introductory unit of this series (January issue).

CLINICAL OBSERVATION: MEASUREMENT AND DISAGREEMENT

Disagreements over clinical findings, diagnoses, and management decisions are common in medicine and dentistry, and can result in failure to provide the best care for patients.[4] For example, the incorrect diagnosis of a temporomandibular joint or pulpal condition can result in providing a surgical or restorative treatment that has a negative effect on the well-being of the patient.

Clinical disagreements occur in 2 ways. A clinical decision based on a sign or symptom can be shown to be wrong when compared with more valid evidence such as a biopsy, radiograph, or arthroscopic examination. Also, clinical decisions can be shown to be inconsistent if an examination of the same patient by other clinicians (or a second time by the same clinician) reveals disagreement. The first example of disagree-

[a]Director and Professor, Advanced Prosthodontics and Maxillofacial Prosthetics, College of Dentistry, The Ohio State University, Columbus, Ohio.

[b]Former Professor and Director of Postgraduate Prosthodontics, Department of Restorative Dentistry, School of Dental Medicine, SUNY at Buffalo, Buffalo, N.Y.

ment is an issue of validity, whereas the second example is a problem of reliability. In both situations, disagreement has occurred. These same phenomena occur in clinical practice and research arenas and, if not understood and controlled, such disagreements can render the best intentions to be incorrect. Patient management decisions are improved when error and variations that can be associated with observational measures are understood, and methods to improve measurement reliability are part of a routine.

Clinical measurement and clinical disagreement are presented together because they share considerations that apply to both patient care and research. Clinical measurement involves clinical observations, with or without special instrumentation, which are subsequently used for decision making. Variations in these observations, as they relate to patient care, are regarded as clinical disagreements. These include diagnostic, prognostic, and therapeutic disagreements.

Webster defines *measure* as, "the extent, dimensions, and capacity, of anything, especially as determined by a standard." Attributes that are measurable can be classified into certain types of measurement scales. Usually 4 types of scales are identified[4] and can be remembered by the acronym "NOIR": nominal, ordinal, interval, and ratio. *Nominal scales* consist of numbers or names used to represent a set of mutually exclusive and exhaustive classes to which assignment is made (male / female, race, blood group, prosthesis). *Ordinal scales* are like the nominal scales with the exception that the classes can be ordered or ranked. Ranking implies directionality but neglects distances between categories along the scale, as in as psychosocial scales (strongly agree, agree, disagree, strongly disagree). *Interval scales* are similar to ordinal scales with the additional feature that distances between all adjacent classes are equal and, conceptually, these scales are infinite, without beginning or ending (such as temperature, calendar date, or time). *Ratio scales* are like interval scales with the additional features that a meaningful zero point exists without any minus values (eg, weight, blood pressure, carious surfaces, DDS visits), allowing for judgments of magnitude differences to be made (eg, 2 or 3 times as large). The differences between scales are important to the manner in which measures are summarized and analyzed.

On the basis of the above, measurement consists of rules for assigning classifications or numbers to observable attributes.[4] These attributes include distinctive qualities and elements, such as properties, characteristics, dimensions, or behaviors. *Quantitative variables*, those most often and easily considered when discussing measurement, are variables that are measured by some physical device or objective means such as counting. The subclasses for quantitative variables are labeled by meaningful numbers and are classified by interval or ratio scales. Quantitative variables can be *discrete*, those that take on a finite or countable set of values such as sulcus depths or Periotest values, or they can be *continuous*, those that take on an infinite or uncountable set of values such as patient visits for a large population. *Qualitative variables*, frequently used in clinical evaluations, are measured by assigning the findings or subjects to mutually exclusive and exhaustive subclasses. These subclasses are labeled by names or numbers that can be numerically meaningless or may sometimes indicate an ordering of subclasses. Both nominal and ordinal scales are frequently referred to as qualitative.

Clinical measurement and research considerations

The focus of this evidence-based series is to inform clinicians how to recognize and use the best available evidence for clinical decisions. As clinicians we use a variety of measurement scales to make decisions. These same measurement scales are frequently used in the literature to determine important diagnostic, therapeutic, and prognostic differences. How can we decide what measures in dentistry are the most important to focus on? As previously mentioned,[5] the focus of your clinical question provides clues as to what measures are important to answer your question and also what measures are important to researchers investigating clinical questions.

Adequate formulation of a research question is fundamental to measurement because the researcher must consider the best means (ie, measure) to answer the question. When reporting on the research in manuscript form, the introduction section presents the problem that concerns the clinician and provides background information from previous, related research to help focus the research question. Clear justification for and definition of the attributes of interest is essential for the reader to know what it is that is being measured and why such a measure is useful for the question at hand. Describing the attribute that is to be measured is a prerequisite to knowing how it is to be measured. The how of measurement is described in the measurement procedure and consists of various techniques for data collection ("assigning numbers") in the study. The "rules" in the definition of measurement specify the requirements for explicit delineation of the procedure. If a measure is to be used to compare persons, it is mandatory that the rules are unambiguous and practical for different observers to apply.

OUTCOMES IN DENTISTRY

It is often observed that different research reports, which address similar clinical questions, use different outcome measures. Valid clinical parameters chosen as care outcome measures should possess acceptable levels of *reliability and responsiveness, and be clinically meaningful.*[6]

In an effort to answer the question, "What measure is important to the specific decision-making process?", it is

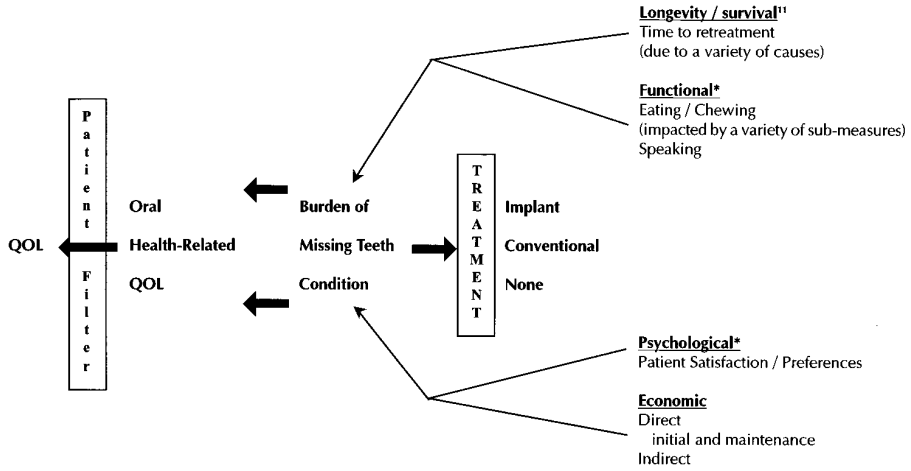


Fig. 1. Dental outcome hierarchy-based on patient-perceived quality of life. *In example of missing teeth, it is often functional, esthetic, and/or psychological (discomfort/pain) concerns that most define patient burden and compels them to seek care. This is because of the impact the burden has on their QOL through its effect on Oral Health-Related QOL. The patient judges care to be successful if sufficient positive outcomes, among 4 major domains listed, outweigh negative outcomes and remove original burden to Oral Health-Related QOL influence on overall QOL.[11]

important to recognize that the measure must have some established relevance to a primary patient concern. It is possible to have a highly accurate measure that has no relevance to the actual concern of the patient and is therefore not beneficial for answering clinical questions. In considering why a measure should have relevancy to the patient's concern, it is helpful to consider the difference in measures of process and outcomes.[7]

Outcomes, which are an end result of a process, are what concern the patient, and what is used by them to judge whether adequate "care" has been provided. Measures of process (such as pocket depth or bone loss) are useful to the extent that they are good proxies or surrogates for an outcome that is important to the patient.[8] This can present a dilemma, as correlations between measures of medical / dental processes and patient outcomes are often weak.[9,10]

Spilker describes a relationship of outcomes in a hierarchy in which an overall assessment of well-being, referred to as the quality of life (QOL), is the ultimate basis for judgment. This QOL judgment is made by considering broad areas of life experiences that include physical, psychological, economic, and social domains. Each of these domains are composed of various component parts that can include submeasures for the domains (ie, mastication, swallowing, and speaking for the physical domain).[11]

Disease or conditions of ill health place a burden on the patient's well-being. It is because of a desire to remove the burden and reestablish their well-being that patients are compelled to seek some form of health care (Fig. 1). A strong argument favoring this QOL-based hierarchy is the fact that patients are the best judges of the impact of health care on their QOL. This is because clinicians judge clinical response rather than how the clinical response is filtered through a patient's values and beliefs to form their specific QOL assessment. It is this final judgment of the "value of care-belief filter" that ultimately determines which factors of a treatment (ie, the benefits vs the risks) sum together, and whether the overall change represents a positive or negative effect on QOL.[11] This is the basis for the relational "strength" associated with validity characteristics of outcomes within such a hierarchy framework.

This outcome hierarchy concept, with the ultimate patient QOL measure, is especially important to understand for diseases / conditions that are not life threatening, for treatments that are often elective, and where treatment options vary in the required methods (ie, surgical vs nonsurgical). The outcome hierarchy framework provides a means to relate a variety of clinical measures (both process and outcome) referenced to their impact on patient QOL. For example, the hierarchy concept lends itself to an understanding of how a condition of missing teeth impacts the patient as a whole and what measures can be used to judge successful care (Fig. 1).[12] The burden associated with tooth loss is what compels the patient to seek care. The clinical endpoint of treatment (ie, the treatment target) is a reduction in the burden of the missing teeth for the patient in question.

Any research that is not directly associated with clinical care (ie, bench-top studies common in prosthodontics) is deemed important only to the extent that it

can be shown to impact some feature of clinical care that is important to the patient. Discussion of clinical outcomes for treatments directed toward reducing the burden of tooth loss has stressed that, because many factors impact on the patient-perceived outcome of "time-to-retreatment" for all prostheses, such an outcome has use because it can allow cross comparison between treatments and it allows assessment of which submeasures actually have impact at the patient level.[11] For common dental "technical" procedures, the association of the technical quality with time-to-retreatment is important to establish and understand. But a failure to recognize the primary importance of the patient-perceived outcome allows research activity to miss the potential benefit of its impact (or lack of impact) on oral health care, and it confuses the reader who seeks answers to problems that are anchored to patient benefit. This dilemma is one reason many journals require authors to provide statements directed at describing the clinical significance of the research findings.

Measurement error: Why is it important?

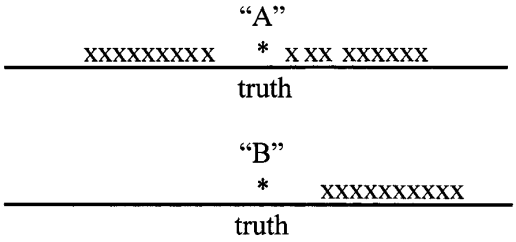
A measure is accurate if it reflects the "true" value of the attribute being measured. This "truth in measurement" is also considered "validity" and not only represents the true state of the attribute being measured, free of measurement error, but also encompasses concerns of what is being measured and the method of measurement. It is possible that a measure can be made that is free from measurement error but is inappropriate for the intended use because what is actually measured is not what was intended to be measured. An example of such an inappropriate measure would be the use of plaque-staining dyes to provide an estimate of the plaque amount when it is known that such dyes also stain pellicle.[13] It has also been suggested that many cephalometric landmarks have been defined more for the convenience of identification and reproducibility than to represent actual anatomic validity.[14]

A measure is inaccurate to the extent that it deviates from the "truth." The variation in measurement of an attribute may imply either that true differences of the attributes exist within or between subjects or there is random error associated with the repeated measurements. Accuracy can be represented by the following equation, which illustrates that the usefulness of the measurement depends on the extent to which the 2 types of error obscure the true value:

Measured Value = True Value + Systematic Error + (Bias) Random Error (Noise)

Random error, also referred to as noise or random variation, is the variation in a sample that can be expected to occur by chance. This type of error occurs when repeated measurements of the same attribute under similar conditions do not agree, but there is no systematic deviation from the true state of the attribute. Such an error is always present and large sample sizes minimize its effect. The *systematic error(s)* is an error that systematically alters every observation made by that observer or every observation made by that device, and is referred to as a type of *bias*. It can result from a machine calibrated to the wrong standard, or results when observers that gather data are not blind to treatment assignments. Because this error causes a difference between what is intended to be measured and what is actually measured, this form of measurement error not only threatens accuracy, but also threatens the validity of the study. Measurement error can therefore be considered to be the amount by which a measurement is incorrect due to the influence of many potential problems inherent in the measuring process.

By way of illustration, the following are 2 sets of repeated measures that are sample estimates from a larger parent population. If the entire population could be measured, the mean would be represented by "A" and "B":



As illustrated, repeated measures "A" exhibit a relatively high random error, which can be stated as poor precision, compared with repeated measures "B," but "A" is (randomly) centered around the "truth." Repeated measures "B" has a relatively low random error (more precise than "A") but repeated measures "B" are biased in that their mean value does not equal the true value. The mean estimate provided by the repeated measures "A" is more useful, given a large enough sample size, because decisions based on these results are likely to provide more benefit than harm since they represent a true sample of an attribute. To the contrary, a mean estimate, provided by the repeated measures "B," could potentially be harmful because the estimate is inaccurate and does not represent a true estimate of an attribute. Decisions based on analysis using this inaccurate measure may actually do more harm than good.

Measurement error can be categorized and evaluated by considering the source of error in the examiner, the examined, or the examination.[15]

When error is associated with repeated measures by the same examiner, it is called *intraexaminer or*

intraobserver error. Within the observer or examiner, there are biologic variations in the senses as well as variations in perception, cognition, and mood that can influence physical examination. Error that is associated with measures made by different examiners of the same attribute is called *interexaminer or interobserver error*. Examiner disagreement can be influenced by prior expectation, even when sophisticated diagnostic tools are used, and by diagnostic classification schemes in which overlapping categories or "breakpoints" are interpreted differently by clinicians.

Sometimes due to the subject examined, the results may be influenced in a consistent or systematic direction (such as, by the time of day or room temperature), while at other times, the influence is unsystematic (for example, a response to normal variations in a material). These are examples of within-subject error. Error associated with the individual being examined can be due to biologic variation in the attribute (such as blood pressure, growth, response to pain), the effects of illness or medication, and patient memory regarding important historical predictors related to physical findings (past health care, trauma, response to previous interventions).

Error can also be introduced by the tool, which is being used to produce the measures. Whether these tools are equipment (microscopes, thermometers, radiographs, pulp testers) or "paper" tools used to evaluate behavioral or psychological characteristics (surveys, patient satisfaction\quality of life), an index of the reliability of the measure is important to assess the reproducibility and hence the usefulness of the results. Also, the actual examination procedure can produce error if the environment hinders the examiner ability to adequately use his/her senses, if the interactions between patient and clinician are disruptive, or if the diagnostic tools are either inaccurate or incorrectly used.

Steps used to reduce clinical disagreement or variability

The strategies for preventing or reducing error and clinical disagreement demand time and effort. For clinicians, these strategies should include but not be limited to: the evaluation of the patient history, the physical examination, and the diagnostic evaluation that determines a specific diagnosis, prognosis or management decision related to the cost of patient care for usual and customary procedures as well as potentially harmful procedures.

Steps used to reduce clinical disagreement or variability include conducting the examination in a proper environment and being aware of individual strengths and weaknesses. Blind assessment, especially relative to the more subjective characteristics, and repeating the more questionable aspects of an examination can reduce the impact of prior expectation and improve clinical diagnostic skills by the strategy of confirmation. When using classification schemes, the examiner should have well-defined operational guidelines for each classification to minimize ambiguity. All diagnostic equipment should be routinely monitored, instruments should be calibrated, and proficiency in the use of the equipment is necessary to assure consistently reliable outcomes.[15]

For researchers, to improve the validity of the measures, the main concern should be to assure that the selected measures are tangible patient outcomes or surrogates that have been established as valid indicators of the outcomes. Establishing validity requires measurement of known quantities, or a gold standard, and evaluating whether any new method accurately measures what it is intended to measure. This form of calibration is different from using a new technique, along side an existing technique to determine agreement between methods.[16] The former checks a method against a gold standard to "calibrate" it to the "truth," while the latter checks the reliability of the new technique against an existing technique in the absence of any consideration of the correctness of the measures. In research, Bland and Altman[16] suggest that *calibration* of a measurement technique or of multiple examiners is an attempt to bring measures toward a gold standard, whereas *training for reliability* simply attempts to bring measures to better precision, as shown in the illustration of repeated measures above.

Controls made for both systematic and random measurement errors should be explicitly described as a part of every research effort. This provides some assurance to the reader that the measures were carefully produced and have the best chance to provide a true estimate for answering the question. Establishing that no systematic overestimation or underestimation of the measure occurs is crucial whether the measure is orthodontic radiographs,[14] periodontal pocket measurements,[17] or distortions common to indirect prosthodontic procedures.[18]

An important way to help control for systematic errors is to randomize the order in which measures are made.[14] Random errors can be very important, as they add to the natural variability of any measurement and can obscure any real differences between groups being measured. Random error can be reduced by replicating and averaging measures, paying attention to replicating the entire process involved in the measurement. It is important for the reader to have some estimate of measurement error, which is often provided by a standard deviation obtained from multiple measures of the same specimen or object. In a discussion on the effects of error in interpretation of research results, Houston[14] describes that if a study shows there is a difference between group means that would be of clinical importance but is not statistically significant, then either too few cases were included to demonstrate whether there

was a real difference, or that the results were so variable that they might have had little practical application to the individual sample.

When the concern is interexaminer agreement, the objective is to have multiple examiners function as one. For example, for agreement associated with observations of periodontal disease presence / absence, gingival recession, tooth mobility, and other related periodontal measures, it is important to measure agreement beyond what would be expected due to agreement because of chance alone. The statistic kappa[20] is used for such a measure of agreement. It has been demonstrated that for nominal scale periodontal measures, improvements in kappa were found to be related to an improvement in examiner concordance on criteria for assignment to individual categories of the variables.[21]

CONCLUSION

The relationship between clinical observations and measurements found in dental literature were described within the context of measurement principles as a means to focus our critical appraisal attention to the important measures used for clinical decisions. Outcomes were discussed from the recognition that dental care, which is largely elective, is an endeavor that seeks to satisfy patient-perceived concerns and therefore needs to consider measures that impact patient-perceived dimensions of care.

Returning to our scenario

From the article you reviewed you have gained a better understanding of the complexities and reasons for the variability associated with health care diagnoses and treatment decisions. This gives you some confidence for your discussion with the local newspaper concerning the national magazine findings. Also, you have decided that you will work with your hygienist to protect your practice from suggestions of overtreatment that could come from either of you. To help accomplish this, you set up a schedule to review information describing the life expectancy of typical restorations you see in your patients, you and your hygienist will review common clinical presentations of deteriorating restorations and recurrent caries to be more in agreement with each other and the current evidence describing these areas, and you will review typical failure patterns for the restorations you commonly replace. Finally, you plan to conduct this review and follow your practice performance for the next 6 months to determine whether additional collaborative training is required to improve your agreement.

REFERENCES

1. Bader JD, Shugars DA. Variation, treatment outcomes, and practice guidelines in dental practice. J Dent Educ 1995;59:61-95.
2. Reddy MA. The use of periodontal probes and radiographs in clinical trials of diagnostic tests. Ann Periodontol 1997;2:113-22.
3. Pihlstrom BL. Measurement of attachment levels in clinical trials: probing methods. J Periodontal 1992;63:1072-7.
4. Evidence Based Medicine Workgroup, Department of Clinical Epidemiology, McMaster University, Hamilton, Ontario, Canada 1992. Workshop publication.
5. Carr AB, McGivney G. Users' guides to the dental literature. How to get started. J Prosthet Dent 2000:83:13-20.
6. Polson AM. The research team, calibration, and quality assurance in clinical trials in periodontics. Ann Periodontol 1997;2:75-82.
7. Fries JF. Toward an understanding of patient outcomes measurement. Arthritis Rheum 1983;26:697-704.
8. Hujoel PP, Leroux BG, DeRouen TA, Kiyak HA. Evaluating the validity of probing attachment loss as a surrogate for tooth mortality in a clinical trial on the elderly. J Dent Res 1997;76;858-66.
9. Carlsson GE, Otterland A, Wennstrom A. Patient factors in appreciation of complete dentures. J Prosthet Dent 1967;17:322-8.
10. Bergman B, Carlsson GE. Clinical long-term study of complete denture wearers. Acta Odontol Scand 1985;53:56-61.
11. Spilker B. Introduction. In: Spliker B, editor. Quality of life assessment in clinical trials. New York: Raven Press, Ltd; 1993. p. 3-9.
12. Carr AB. Successful long-term treatment outcomes in the field of osseointegrated implants: prosthodontic determinants. Int J Prosthod 1998;11: 502-12.
13. Brunette DM. Critical thinking. Understanding and evaluating dental research. Chicago: Quintessence; 1996. p. 65.
14. Houston WJB. The analysis of errors in orthodontic measurements. Am J Orthod 1983;83:382-9.
15. Department of Clinical Epidemiology and Biostatistics, McMaster University. Clinical disagreement: I. How often it occurs and why. Can Med Assoc J 1980;123:499-504.
16. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;2:307-10.
17. Hefti AF. Periodontal probing. Crit Rev Oral Biol Med 1997;8:336-56.
18. Nicholls JI. The measurement of distortion: theoretical considerations. J Prosthet Dent 1977;37:578-86.
19. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 1968;70:213-20.
20. Fleiss JL, Chilton NW. The measurement of inter-examiner agreement on periodontal disease. J Periodont Res 1983;18:601-6.

doi:10.1067/mpr.2000.105799