Evidence-Based Medicine and Outcomes Assessment

Dorothy Cimino Brown

For veterinary surgeons, making decisions about the optimal care of patients requires the integration of individual clinical experience with information gathered from the best available clinically relevant research. The proficiency and judgment that individual clinicians acquire through clinical experience and clinical practice are fundamental to the decision-making process; however, knowledge gained through unsystematically recorded personal experiences can be biased. Therefore, surgeons look externally to well-conducted clinical research to affirn their current decision-making process or suggest an alternative. Considering the merits of published research with which to answer clinical questions can be challenging. Major technical innovations tend to be reported as a series of cases, which form the core of the clinical surgical literature. New operations are adopted on the basis of these case series, before comparisons are ever made with previous techniques. The surgical procedures are in constant evolution, with each surgeon adding small technical refinements, making comparisons of case series over time difficult. Even if the surgical technique remains the same from one case series to the next, other variables often differ, such as presurgical screening, perioperative management, and postoperative follow-up, again confounding comparisons between techniques. One of the most significant problems associated with evaluating the efficacy of surgical interventions in the literature is the lack of a consistent, valid, reliable, and clinically relevant outcome assessment.

OUTCOME MEASURES

Outcome measures are the tools (aka instruments) used to measure the success of an intervention. For surgeons, the challenge is determining the definition of success for postoperative procedures to determine which outcome measures are most appropriate to apply to studies evaluating the efficacy of those procedures. Occasionally, the outcome of interest is clear and straightforward, for example, for ovariohysterectomy a successful outcome is sterility of the animal; for most procedures however, the outcome is less clearly defined. For example, for many orthopedic surgical procedures, a successful outcome is decreased pain and improved function of the animal. Ideally, therefore, outcome measures that capture pain and function would be used in efficacy studies. Many efficacy studies of orthopedic procedures use kinematic or force plate gait analysis. The theory behind the use of gait analysis is that it is an objective measure of lameness, and lameness, in turn, is an indirect indication of pain and function. Gait analysis is not a direct measure of function, which becomes obvious in the situation where a dog increases the amount of peak vertical force it generates on its limb following a surgical procedure, but is still unable to climb up the stairs or jump onto the bed unassisted. The procedure may have been a success in terms of gait analysis, but the improvement in function that the owner or surgeon desired was not obtained. So why use gait analysis as an outcome measure if it doesn’t directly assess the outcome of interest (i.e., the dog’s ability to function in its home environment)? Gait analysis is a valid, reliable, and objective measure of lameness. Two of those properties, validity and reliability, are required for all outcome measures, and objectivity in a measure is desirable whenever possible. If no valid and reliable measures of a dog’s function in its home environment are available, then another measure, such as gait analysis, must be chosen. The problem is that most direct measures of the success of a surgical intervention, such as the dog’s ability to function in its home environment, require assessment of a subjective attribute. The development of outcome assessment tools for subjective outcomes is not easy and requires considerable investment of both mental and fiscal resources.

OUTCOME ASSESSMENT IN VETERINARY MEDICINE

A sound method is available for the development and application of tools to assess subjective states. Unfortunately, much of this literature is virtually unknown to most veterinary researchers. Health science libraries do not routinely catalog journals or textbooks that address the concepts of measurement because they are predominantly directed at educational and psychologic audiences for the development of achievement, intelligence, or personality tests and scales. References on the topic that focus on those attributes of interest to researchers in health sciences, such as subjective assessment of health states and response to illness, are uncommon. It is not surprising therefore, to find that a vast majority of measures reported in the veterinary literature to collect data for a subjective outcome were devised specifically for each given study, with no clear indication that the process of questionnaire development included or assessed the reliability and validity of the measure. Given the number and variety of unpublished scales that appear in the veterinary literature as outcome measures, it is clear that most investigators believe that devising a series of reasonable looking questions and then averaging or summing the responses to get a score is all that is required to then use that tool (i.e., series of questions) as an outcome measure in their study. However, questions that appear very reasonable on the surface and have been used in published studies, on closer inspection are actually very problematic. Example questions are provided here:

Question 1: “How much difficulty does your dog have going up and down the stairs?”

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Difficulty</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>Extreme Difficulty</td>
</tr>
</tbody>
</table>

For this question, the focus will be on problems with the question itself:

Problem No.1: This is what is known as a double-barreled question, that is, it actually asks two different questions at once. It is asking an owner “How much difficulty does your
Problem No. 1: Some of the response options are asking for an assessment of change, while others are asking for a current assessment of health status. If a question is to be used to assess change, two options are available. The first option is to assess the current health status of the animal at two different points in time during the study, and then calculate the change. The second option is to ask about change at some point following an intervention or during the progression or resolution of disease. Mixing both types of responses in the same question causes confusion for the respondent.

Problem No. 2: Response options do not cover all of the possible choices for which the owner may be looking. Two "change" options are provided: "no change" and "increased," but no option is given for "decreased." Three health status options are listed: "indifferent," "little attention," and "needy"; no options are offered for a normal or average amount of attention, or for a lot of attention without the negative connotation of being needy.

Problem No. 3: The response options are not mutually exclusive. If a dog pays little attention to the family, but that is not any different than the dog’s normal behavior, will the owner choose option number 2 or option number 4? If the owner circles both on the questionnaire, how will the investigator manage the data?

Designing clear questions with appropriate response options is more challenging than it would appear at first glance. In addition, it is only the first step in the development of an appropriate outcome assessment tool. Devising the questions must be followed by confirmation that those questions make up a valid and reliable instrument.

**STEPWISE DEVELOPMENT OF A HEALTH MEASUREMENT INSTRUMENT**

**Step One: Devising the Items (Questions)**

The first step in designing a scale or questionnaire is devising the questions themselves. This is far from a trivial task in that no amount of statistical manipulation after the fact can compensate for poorly chosen questions, that is, those that are badly worded, ambiguous, irrelevant, or even not present. Two of the more relevant techniques for developing questions in a rigorous and systematic manner include the use of focus groups and key informant interviews. Focus groups are discussions in which a small group of people (typically six to twelve) with traits of interest are guided by a facilitator to talk about themes that are important to the investigation. For example, if one wanted to develop an outcome assessment scale to measure the ability of dogs with osteoarthritis to function in their home environment, owners of dogs with osteoarthritis would be assembled to discuss the behaviors they associate with their dog’s ability to function and how their dog’s osteoarthritis affects those behaviors. Key informant interviews are interviews with a small number of people who possess unique knowledge. These could be owners of animals with the condition of interest, but usually the group consists of three to ten clinicians who have extensive experience evaluating and managing those patients.

Once a series of questions has been devised, a method by which the responses will be obtained must be chosen. This is dictated in part by the nature of the question, but a very large body of research in this area describes the advantages and disadvantages of the wide variety of scaling methods that can be used. One form of question frequently used requires only a categorical judgment by the respondent, indicated by a "yes/no" response or a simple check. This results in a nominal scale (Figure 11-1). However, many of the variables of interest to health care researchers are continuous rather than categorical, and many options for scaling these responses may be chosen. The visual analogue scale is a line of fixed length with anchors at the extreme ends and with no words describing the intermediate positions (Figure 11-2). Respondents place a mark on the line corresponding to their perceived state. This approach is simple for investigators but often is not well understood or completed appropriately by respondents, and other methods may yield more precise measurement. To minimize the problems that arise when some respondents inappropriately complete visual analogue scales, those response options can be converted relatively easily to numeric rating scales by converting the visual analogue line to a 0 to 10 choice, as in the

*References 1, 2, 4, 5, 8–12, and 14.*
example (Figure 11-3). Adjectival scales use descriptors along a continuum rather than simply labeling the endpoints (Figure 11-4). Likert scales are bipolar scales measuring a continuum of positive to negative responses to a statement (Figure 11-5). When Likert scales are constructed, consideration should be given to several issues, such as the number of scale divisions to provide, and whether or not a neutral category should be included. For example, sometimes it is desirable to use a four- or six-point scale and exclude the neutral response option, to force the respondent to make a choice in the negative or positive direction.

Regardless of which specific scaling approach is chosen, many questions must be addressed when continuous scales are designed. For example,

1. How many divisions or boxes should be provided?
   The choice of the number of steps or boxes on a scale should be based on the rater’s ability to discriminate. Ratings that use only two categories to express a continuum will result in loss of information. For example, the question in Figure 11-6 might be responded to in two ways: (a) or (b). The first method (a) reduces all positive opinions from mild to strong to a single option with potential loss of information. An owner who believes his dog has had only a very mild decrease in quality of life would have an obvious choice in response option (b), but might have a difficult time choosing between the only two categories available in response option (a). He would have to decide whether the mild decrease that he believes exists is “enough” to warrant an “agree” response in option (a). This loss of more precise information leads to a reduction in the reliability of the question. Although it is intuitive to believe that the rater’s ability to discriminate among multiple categories would be highly contingent on the particular situation, multiple studies suggest that this is not the case. In fact, as a rule, reliability coefficients go down as fewer categories are used. Studies suggest that the minimum number of categories should be in the range of five to seven. As far as a maximum number of categories, it appears that people are unable to discriminate much beyond seven levels but often avoid using the two extreme ends of a scale, suggesting that a five- to ten-category rating scale is preferred in the vast majority of cases.

2. Should an even or odd number of categories be given?
   As described previously, an even number of response options (i.e., not providing a neutral option, Figure 11-6, b) forces respondents to commit themselves to the positive or the negative. No absolute rule applies as to whether this is recommended. It depends on the needs of the particular research as to whether it is...
The financial cost of veterinary care for my dog is too high. (Fill in the oval above the words that most accurately describe your response.)

- Strongly agree
- Agree
- Cannot say
- Disagree
- Strongly disagree

Figure 11-7 Example of a question with a response option that inaccurately aims to document a neutral response to the question.

The choice may not be clear as to how to word a question or scale the response options. The key principles are to avoid ambiguous, complex, or double-barreled questions that contain technical jargon or uncommon or vague words; and to provide response options that are appropriate to the question and are designed to collect complete and accurate data. Once all of the best attempts are made to design appropriate questions and response options, the “best” questions are then selected (Step 2) to make up a preliminary questionnaire that is then tested (Step 3) to ensure that the questions do indeed collect valid and reliable data.

**Step Two: Selecting the Items (Questions)**

Typically, not all of the items that are developed are ultimately included in the new instrument. Some may be confusing, may be interpreted differently by different respondents, or may not deliver the desired information. Various criteria can be used to determine which of the developed questions should be retained for the preliminary instrument. Items that are difficult to interpret are removed or rephrased. Generally, it is recommended that instruments not require reading skills beyond a sixth grade level. The readability of an instrument can be measured by several different methods, but readability alone does not ensure the absence of ambiguity. One way to evaluate the ambiguity of questions is to ask a group of respondents to rephrase the questions in their own words, and see how close to the intended meaning they are. Questions must be screened for the use of jargon (e.g., “NSAID” or “range-of-motion”) that may not be understood by all respondents, or value-laden words (e.g., “trivial” or “too much”) that could prejudice the respondents and distract from clear interpretation of a question.

Once the items have been screened for readability and ambiguity, they can be presented to a panel of “experts” for determination of whether the items look reasonable. This is typically a handful of people with experience managing the condition to be captured by the questionnaire. They can make an assessment of whether the items appear, on the surface, to be measuring what is intended (face validity), and whether the questions cover all relevant aspects of the condition (content validity). Once a preliminary set of items is agreed upon, the instrument can be pretested.

Pretesting involves administration of the preliminary instrument to a small group of respondents to determine whether some items may not perform well. In most situations, when a trait is measured, the instrument should be homogeneous, that is, all of the items should be tapping into different aspects of the same trait, not different aspects of different traits. Therefore, the items should be moderately correlated with each other. An inter-item correlation matrix can be analyzed from the data collected from this group of respondents to identify items that have consistently low correlations with other items in the instrument. These items can then be revised and the preliminary instrument pretested again in a different group of respondents, or the items can be removed from the instrument before moving onto large-scale testing for reliability and validity.

Desirable for raters to be allowed a neutral response. If raters are “forced” to choose between a positive and a negative response, they may opt not to respond to the question at all, leaving missing data points. However, if the neutral response does not contribute to an understanding of the question or outcome being assessed, it is not worthwhile to offer it as an option. For example, if an investigator is interested in quantifying owners’ perceptions of the cost of veterinary care, it may not be very informative to have the “easy” option of “no opinion” as a neutral category.

3. **Should all the points on the scale be labeled, or just the ends?**

Most studies suggest that relatively little difference exists between scales with adjectives under each box and scales with adjectives only at the ends. Respondents prefer scales where many or all of the points on the scale are labeled, but this usually is offset by the inability of the scale constructor to define enough appropriate adjectives.

4. **Which adjectives should be used?**

Choosing the appropriate adjectives to describe the scale can be challenging. Most of the terms used have an inherent vagueness. For example, phrases like “Often” or “Rarely” are used to elicit judgments about the frequency of an event. The question that arises from the use of such terms is this: To what extent do most people agree on the meanings of such adjectives? Part of the problem is the vagueness of the terms themselves, and another is the fact that those meanings can change, depending on context. For example, “Often” connotes a higher frequency for common events compared with rare ones. Because these descriptors are prone to individual interpretation, the reliability of these questions (discussed later in the chapter) is pinned to the same from one question to the next to make it easier for the respondents, but it is far more important that the descriptors be appropriate for the question that is being asked. Forcing the response options to fit the question can lead to ambiguity in interpretation of the question. In cases where the same response option is appropriate for an extended series of questions, some scales reverse the order of responses at random, so that successive questions may have response categories that go from low to high or high to low. This can be done to combat response fatigue. Toward the end of a lengthy set of questions with repetitive response options, respondents tend to repeat the same answer to the remaining questions, resulting in uniform and inaccurate answers. The downside to reversing the order of scale responses is that a careless respondent may not notice the change, resulting in uninterpretable answers.

5. **Should the scaling change for successive questions or remain as consistent as possible?**

Investigators often want to keep the labels under boxes the same from one question to the next to make it easier for the respondents, but it is far more important that the descriptors be appropriate for the question that is being asked. Forcing the response options to fit the question can lead to ambiguity in interpretation of the question.
Step Three: Assessing Reliability and Validity

Once a data-gathering instrument is developed, it must be established that it will target what it is supposed to measure. This is defined as the validity of the instrument. In addition, the instrument must measure what it is supposed to measure in a consistent manner. The tendency toward consistency is referred to as reliability. Before evidence that an instrument is measuring what it is intended to measure can be obtained, it is first necessary to gather evidence that it is measuring something reliably. An instrument can be tested for reliability in several ways. An assessment of internal consistency can be based on the data collected from a single administration of the instrument to a large group of respondents. Although various methods of consistency calculation may be used (item-total correlation, split-half reliability, Kuder-Richardson 20, coefficient α), all represent the average of the correlations among all items. Assessment of internal consistency alone, however, is not sufficient to declare an instrument reliable, because it relies on only a single administration.

The reproducibility of a measure administered on different occasions can be analyzed in various ways. One might demonstrate the degree of agreement between different observers (interobserver, aka intrarater, reliability); the degree of agreement between observations made by the different occasions (intraobserver, or intrarater, reliability); or observations on the patient on two occasions separated by some interval of time (test-retest reliability). For example, if an instrument designed to measure chronic pain in dogs with osteoarthritis is reliable, it should deliver very similar results when administered on two different occasions, 1 week apart, to owners of dogs that have stable disease and no change in treatment. If instead, the results between the two administrations are very different, the instrument would not be useful for testing the efficacy of an intervention. How would one interpret a change in instrument score following an intervention if a change can be documented when there is no intervention at all? It is only when it has been demonstrated that the instrument is measuring something reliably that the process of determining what that something is can begin.

To confirm that the instrument is measuring what is intended requires more than peer judgments (face validity). Validating an instrument is a process by which it is determined what degree of confidence can be placed on conclusions drawn about an animal based on the score from that instrument. If other validated instruments designed to measure the same attribute but, then an obvious approach is to administer the two experimental instruments along with the existing one, and see whether a strong correlation between the two is evident. More likely, however, no other measure exists, and developers must test "construct" validity.

Construct validity is evaluated when the attribute being measured cannot be directly observed. For example, chronic pain cannot be "seen," but behaviors can be observed that, according to theories about chronic pain in companion dogs, result from chronic pain. No one single experiment or statistic can unequivocally "prove" a construct. It is through multiple analyses and assessments that a construct appears to be valid. For example, given a construct of chronic pain, dogs with osteoarthritis that score high on a newly developed instrument differ from dogs that score low on it, in terms of attributes such as the results of force plate gait analysis, assessment on two of health-related quality of life, response to NSAID therapy, and daily activity monitoring. No one single experiment can "prove" the validity of the instrument, but multiple, well-designed, hypothesis-driven studies can build the body of evidence that the instrument is measuring what is intended.

It is necessary to conduct validation studies for each new instrument that is developed, and the task is an ongoing one.

If the instrument is to be used in a group of animals for which it was not initially validated (e.g., dogs with bone cancer), we must demonstrate that the inferences made for them are as valid as for the original population (e.g., dogs with osteoarthritis). In addition, modifications of the instrument, such as changes in wording of questions or responses, changes in the order of questions, or removal or addition of questions, often require new validity studies.

STUDY DESIGN AND SUBJECTIVE VERSUS OBJECTIVE MEASURES

Because the method of development of valid and reliable outcome measures is not fully understood by most in veterinary medicine, it is incorrectly assumed by many that subjective outcome measures are inferior to objective ones. If the validity and reliability of a subjective measure have been confirmed through appropriate processes, it is just as reasonable a choice for an outcome measure as an objective one, and may actually be a better choice if it truly measures the outcome of interest. The implications for the use of a validated measure for a subjective outcome, then does not concern the validity of the data that will be collected from it, but rather focuses on the design of the study from which the measure will be used. For example, if an owner assessment of chronic pain is to be used as the outcome, it is preferable that the owner be blinded to the treatment group, so he can give an unbiased assessment. Blinding may not be important for some objective outcome measures, but this certainly is not true for all of them. For example, mean arterial blood pressure might be a reasonable choice of outcome for an antihypertensive intervention and would be considered an objective measure; however, such variables as the tightness of the cuff, the stress level of the animal, and the acclimation environment of the animal can affect the reading from the machine and are controlled to some extent by the operator. The operator therefore should be blinded to the treatment group, so an unbiased assessment can be obtained from this objective measure.

The use of a control group is also important for obtaining an unbiased estimate of the effect of an intervention, regardless of whether the outcome measure is subjective or objective. Documented improvements in placebo-treated animals are common for both types of measures and are likely due to regression to the mean, which is a ubiquitous phenomenon that can occur when repeated measurements are taken on the same subject.15 Relatively high (or low) observations are likely to be followed by less extreme ones nearer the subject’s mean because of natural variation in disease symptoms. For example, dogs with osteoarthritis will oscillate between "good" days and "bad" days, depending on such things as weather and activity level. These oscillations may influence when an owner seeks to include the dog in a study. For example, the owner may be more likely to seek out a study when the animal is experiencing a period of “bad” days. These dogs would score more poorly at the baseline outcome assessment for pain and function than they would during a period of “average” days. As these dogs progress through the natural variability of the symptoms, they would regress back to their average level of discomfort and functionality, and their scores would improve in spite of the fact that no intervention for the condition existed. An improvement in control groups has been documented using objective measures such as force plate gait analysis. The fact that dogs in a control group can show improvement with no intervention at all may indicate that not all of the improvement seen in an active treatment group is necessarily attributable to the intervention. Therefore, the use of a control group is very important to the assessment of intervention efficacy.
The use of blinding and control groups in studies of surgical interventions is not always possible. In those studies where their inclusion is not feasible, data collected must be evaluated with the knowledge that any documented treatment effect is likely exaggerated, regardless of whether subjective or objective outcomes are used. At a minimum, however, if the same validated outcome measures are used by most investigators, comparisons can be made between studies, with awareness that the same biases could be present in each.

**PROCESS MEASURES**

Process measures are distinct from outcome measures. Process measures evaluate the performance of the process by which the animals' outcome is achieved. Examples of process measures in surgery include surgery time, days of hospitalization, and cost of care. If two interventions achieve the same outcome, process measures can be used to help determine which of those interventions might be preferable. For example, if two different surgical interventions for thoracolumbar disk herniation lead to the same improvement in pain and function, the choice of procedure may be driven by a differential in surgery time.

**CONCLUSIONS**

A major problem facing veterinary surgeons is the lack of high-quality evidence available for optimal decision making. Does wide variation in operative procedures for a given problem make a difference? Can a large number of reported operative techniques for a given problem testify to the reality that, in many situations, the technique may not matter despite strenuous and occasionally acrimonious discussions in support of one operation or another? Ideally, fair and direct head-to-head comparisons of diagnostic and therapeutic interventions would allow surgeons to make well-informed decisions for their patients. Ideally, comparisons would be made on the basis of clinically relevant, widely available, easy to use, and well-validated outcome measures. Questions regarding whether the veterinarian and/or the owner believes that the animal has benefited from an intervention are clinically relevant and can be addressed in carefully designed outcome measures. The fact that an owner or veterinarian assessment is inherently subjective does not preclude its use as an outcome measure. If proper methods are followed, hard measures of subjective outcomes can be developed, and subjective states can be reliably quantified. Outcome assessment is critically important in evaluating the efficacy of surgical procedures. Such an assessment assists surgeons in distinguishing between various treatment methods and helps to identify effective treatment options, which, in turn, improves patient care.

**REFERENCES**

For a complete list of references, log onto www.veterinarysurgerysmallanimal.com.
REFERENCES

AUTHOR QUERY FORM

Dear Author

During the preparation of your manuscript for publication, the questions listed below have arisen. Please attend to these matters and return this form with your proof.

Many thanks for your assistance.

<table>
<thead>
<tr>
<th>Query References</th>
<th>Query</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AU: Reference 19,20 has not been included in the Reference List, please supply full publication details.</td>
<td></td>
</tr>
</tbody>
</table>