OUTCOME MEASURES IN CANINE CCL DISEASE
John F. Innes, BVSc, PhD, DSAS orth, MRCVS
University of Liverpool, Liverpool, UK

To improve the efficacy of treatments for cruciate disease, it is imperative that the efficacy of treatments can be measured. Objective measures of outcomes, such as force platforms, are desirable because of their lack of bias and potential for greater reliability. However, such objective measures may be limited in their availability because of the cost and time involved in collecting data. In addition, such measures may only capture certain dimensions of the disease. For example, the force platform is a robust outcomes measure but it is generally used to measure certain force parameters in the affected limb. Thus, there are aspects of stifle joint function (and dysfunction) that the force platform cannot measure, such as “inactivity stiffness”, “ability to climb stairs”. Therefore, in a chronic situation, such as the post-surgical stifle, it is likely to be useful to use clinical metrology instruments. If combined with web-based data collection, such instruments could provide the means to collect data from many thousands of cases.

Clinical metrology instruments

In the last 10-15 years there has been an interest in developing validated questionnaires, or “clinical metrology instruments” for use in canine orthopaedics outcomes assessment, including stifle disease. The advantages of using such instruments are that it should result in more robust and reliable data, one can understand the limitations of the instrument, and one can start to compare between published reports.

Designing a clinical metrology instrument can involve several steps. An important aspect of any evaluative index is a strategy for the measurement of all clinically important treatment effects. Item generation is followed by testing validity, reliability and responsiveness of the instrument. Initially, instrument development involves the generation of items (questions) that represent theoretic constructs (i.e. disease variables). Considering that the instrument will be completed by dog owners, it seems logical to involve dog owners in generation of the measures and this approach has been taken in one instrument, the Canine Brief Pain Inventory, designed to measure chronic pain in dogs. Alternatively, items may be generated by veterinarians with expert knowledge of the condition in question, and this approach has also been reported for the LOAD instrument. Item scaling refers to the various options for types of scale that are available for owners in answering each question. The simplest scale is a dichotomous one, e.g. is the dog lame or not? This type of scale is preferred by some of the regulatory authorities and has been used in clinical trials of NSAIDs for licensing purposes. However, such scale is inherently poorly responsive (see below). When an investigator requires grading of response or disease activity, there is a need for a discontinuous ordinal scales or visual analogue scales (VASs). Five, seven or nine-point rating scales are popular with five point scales perhaps representing the best compromise between responsiveness and reliability. Such scales are often presented as “none, mild, moderate, severe, extreme”, or numerical rating scales.

Once all items have been generated, one must consider the validity of the instrument. Validity refers to the extent to which the instrument is able to measure what is intended. There are various aspects to validity. Face validity is present if it "looks like" the instrument is going to measure what it is supposed to measure, and in this situation, this would be assessed by veterinarians. Content validity is also judgemental and is said to be satisfied when the measure, or combination of measures comprehensively captures the important areas of the domain which it is attempting to represent. Construct validity is concerned with the extent to which a particular measure relates to other measures in a manner which is consistent
with theoretically derived hypotheses concerning the constructs (or concepts) that are being measured. For instance, if other measures of the disease (e.g. veterinarian’s assessment) produce a similar signal (i.e. good correlation) to the new index, construct validity may be supported. Similar, but not synonymous, criterion validity refers to the extent to which an instrument produces the same results as a “gold standard”. This is a difficult issue in cruciate disease because there is no current consensus on a gold standard. One might consider the force platform a gold standard but, as already stated, it measures particular disease dimensions and the metrology instrument may be designed to capture more widely.

Reliability is the extent to which an instrument records the same numerical values on repeated occasions, assuming no underlying change in the index condition between assessment intervals. This is often evaluated in a ‘test-retest’ scenario when the disease has become stable. Reliability should be measured using a statistic that reflects agreement between the two assessments, such as the intraclass correlation coefficient (ICC). The ICC ranges from 0-1 with 1 representing perfect agreement. It should be noted that agreement is not the same as correlation in that there may be a good correlation with poor agreement due to systematic bias. In cruciate disease, one study demonstrated overall good reliability of an owner-administered questionnaire with concordance correlation coefficients between 0.549 and 0.916. That study also suggested that owners are more reliable for more generic items, and that very specific items showed poorer reliability.

Responsiveness is a measure of the extent to which an instrument can detect clinically important, statistically significant changes in health status; it is particularly relevant to clinical outcomes measures tools. Responsiveness should be tested using standard treatments, relevant types of patients and conventional sample sizes. There are a variety of approaches to assess responsiveness. Sensitivity is the ability to detect change statistically whereas relevant change is the change that is clinically meaningful. Therefore, responsiveness statistics are divided into sensitivity measures on the one hand, and methods designed to assess the ability of an instrument to separate clinically-relevant change from irrelevant change. Examples of sensitivity statistics are the standardised response mean (SRM) and the effect size. These indicate the magnitude of the change in comparison with the standard deviation of change, or the standard deviation at baseline, respectively. The larger the SRM or effect size, the greater the sensitivity to change. Although there are no absolute standards for effect size, it has been suggested that in comparative studies, examples of small, medium and large effect sizes might have values of 0.2, 0.5 and 0.8, respectively. Responsiveness of clinical outcomes measures has received a limited amount of attention in veterinary orthopaedics. One previous study looked at the responsiveness of an instrument (Bristol osteoarthritis in Dogs [BrOAD]) using surgical intervention for cruciate disease as the test treatment. In that study, the authors reported effect sizes of 0.68 to 2.75 for various items, but the design of that study was less than optimal in that clients had to recall the pre-treatment evaluation and there was no objective outcome measure. Thus the BrOAD instrument had sensitivity but without criterion and construct validity, one cannot be certain that this was relevant change. In another study, responsiveness of the LOAD instrument was considered low, as tested in a cohort of dogs with chronic elbow osteoarthritis using a licensed NSAID as the test treatment. However, concomitant force platform data confirmed that the response to the treatment in this population was small and did not reach statistical significance with any measure.

The instruments design process may generate multiple items and it may then be possible to apply statistical techniques to reduce the number of items without losing information. In this way, one can reduce respondent burden. There are various approaches to this including principal components analysis and internal consistency testing. However, for an evaluative index, it is particularly important that non-responsive items are deleted as they only contribute to random error.
Recent work in the author’s research group has compared the performance of three clinical metrology instruments: LOAD, CBPI and the Helsinki Chronic Pain Index. The results from these studies will be presented.