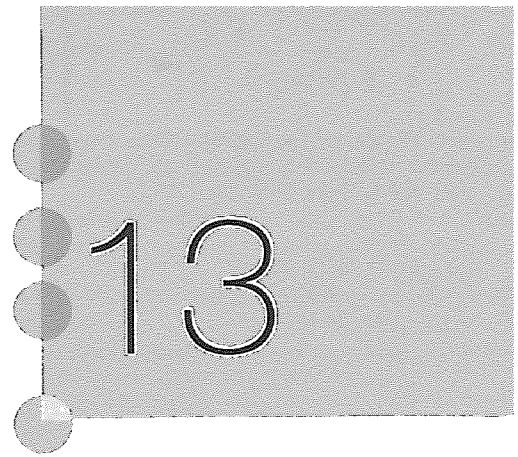


Clinical outcomes in sport and exercise physical therapies

James J. Irrgang and Robert G. Marx



CHAPTER CONTENTS

Introduction	206
Framework for identifying clinical outcomes	207
Selection of clinical outcome measures	211
Collecting, analyzing and interpreting clinical outcome measures	215
Summary	218

Introduction

Outcomes management is the process of data collection, analysis and interpretation of the efficiency and effectiveness of patient treatment, with the intent of improving quality of care and lowering health care costs (Dobrzykowski 1997), and is an integral component of the process of care provided by physical therapists (American Physical Therapy Association 2001). Outcomes data can be used to make patient management decisions, assess clinician and organizational performance and to provide evidence for the effectiveness of interventions provided by physical therapists and other rehabilitation specialists. The validity of the inferences made from outcomes data is dependent on the outcome measures themselves and the circumstances under which the data were collected.

A framework for assessing outcomes of sports physical therapy is presented in Fig. 13.1. Important outcomes of sports physical therapy include clinical outcomes, process outcomes, patient satisfaction and costs (Irrgang 1996). Clinical outcomes are usually the primary interest when attempting to demonstrate effectiveness of rehabilitation and reflect the clinical status of the patient. Disablement schemes, such as the Nagi Disablement Model (Nagi 1991) and the recent International Classification of Impairment, Disability and Health (ICIDH-2) proposed by the World Health Organization (2001) provide a useful framework for identifying relevant clinical outcome measures and will be discussed in greater detail below.

Process outcomes represent the utilization of resources and include measures such as the duration of care, number of visits and number and type of interventions provided to the patient. Evaluation of process outcomes can be used to answer the question 'Did the intervention provided to the patient match the patient's diagnosis (classification) based upon the findings of the examination?' Selection of the most appropriate intervention for a particular patient requires the ability of the clinician to examine, evaluate and diagnose the condition in order to select the most appropriate form of intervention. Achieving optimal process outcomes requires application of the principles of evidence-based practice (see Chapter 1). Evaluation of process outcomes can be used to assess clinician and organizational performance. Expert clinicians would be expected to choose the most optimal interventions given a patient's diagnosis. Sources of process outcomes data may include scheduling and billing databases and patient records.

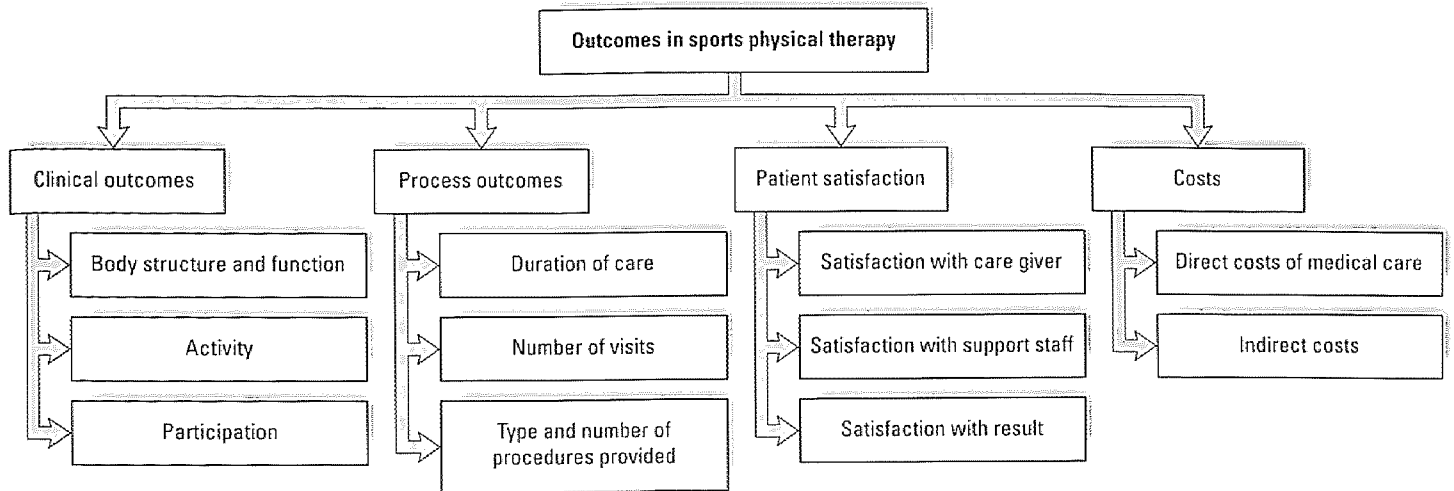


Figure 13.1 • Framework for assessing outcomes of sport physical therapy.

Patient satisfaction may also be an important outcome for sports physical therapy. Aspects of patient satisfaction include satisfaction with the caregiver, support staff and the clinical result. Patient satisfaction is usually measured anonymously using a written survey. Domains of patient satisfaction that are commonly measured in rehabilitation include satisfaction with access to care, physical environment, patient care, billing issues and overall satisfaction with the experience during the episode of care. Patient satisfaction instruments commonly use Likert-type rating scales that reflect the degree of satisfaction or dissatisfaction on a 5 to 7 point scale. Most patient satisfaction instruments have been developed by the user and have not undergone psychometric testing; however, Goldstein et al (2000) developed and tested a 26-item patient satisfaction survey to measure the outcome of physical therapy intervention. Patient satisfaction data can be aggregated to compare the levels of patient satisfaction achieved by individual clinicians or facilities, and may be useful for quality improvement initiatives to identify opportunities for improvement. Development and use of a standardized patient satisfaction instrument to measure the outcomes of sports physical therapy would be valuable to permit benchmarking between providers and organizations.

The costs of care may also be an important outcome of sports physical therapy. The total costs to an individual due to an injury include the direct costs for medical care as well as indirect costs. The direct costs for medical care for an injury go beyond the costs for rehabilitation and include the costs for diagnosis, medical/surgical management and any equipment that may be necessary to facilitate recovery. The indirect costs of an injury may be related to work time lost, decreased productivity or cost related to assistance required to perform activities of daily living or household activities. When discussing costs, one must be careful to distinguish costs from charges. From the perspective of the provider of sports rehabilitation, costs are the costs of providing a service including costs related to manpower, space, equipment and supplies. From the perspective of the patient or payer, costs are the charges for the

services that are rendered. Charges are closely related to the services (i.e. the process outcomes) that are provided during the episode of care. Costs can be used to calculate the value of sports physical therapy, which is defined as the ratio of the benefit of care divided by the costs of providing that care. The aim of sports physical therapy should be to provide high value, which is of large benefit at a relatively low cost. An important question that needs to be answered is 'Does sports physical therapy reduce the total costs associated with an injury?' The answer to this question requires detailed analyses and the expertise of a medical economist.

While process outcomes, patient satisfaction and costs are important outcomes of sports physical therapy to consider, the remainder of this chapter will focus on clinical outcomes. This will include a discussion of a framework for identifying important clinical outcomes, considerations for selecting outcome measures and collecting, analyzing and interpreting outcomes data.

Framework for identifying clinical outcomes

Disablement models provide a useful framework for identifying relevant clinical outcomes of sports physical therapy. Disablement models have been proposed by Nagi (1965, 1991), the World Health Organization (1980, 2001), and the National Center for Medical Rehabilitation Research (National Institutes of Health 1992). Disablement is the impact of injury or illness on the function of specific body systems, on basic human performance, and on an individual's role in society (Jette 1994, Verbrugge & Jette 1994). While the terminology used by the models differs somewhat, each of the models defines disablement at the tissue/cellular level, organ/ body system level, personal level and societal level (Table 13.1).

The Nagi scheme is the model of disablement accepted in the Guide to Physical Therapist Practice (American Physical

Table 13.1 Comparison of disablement schemes

System	Tissue/cellular level	Organ/system level	Personal level	Societal level
Nagi	Active pathology	Impairment	Functional limitation	Disability
ICIDH	Disease	Impairment	Disability	Handicap
NCRMM	Pathophysiology	Impairment	Disability	Societal limitation
ICIDH-2	Impairment of body structure and function	Impairment of body structure and function	Active restriction	Participation restriction

Therapy Association 2001). The Nagi disablement model includes active pathology, impairment, functional limitations and disability. Active pathology may result from infection, trauma, metabolic imbalance and/or degenerative disease conditions, and may interrupt or interfere with normal cellular processes. Active pathology includes the simultaneous efforts by the organism to regain homeostasis. Impairment is a loss or abnormality of an anatomical, mental or emotional nature that results in loss or abnormal function at the organ or body system level. Functional limitations refer to the manifestations of pathology and impairment on the function of the individual as a whole and may include limitation in physical or psychological function. Disability refers to the function of the individual within society and is defined as 'the inability or limitation experienced by the individual in performing socially defined roles and tasks within the context of a socio-cultural and physical environment' (Nagi 1991). Disability may affect family and other interpersonal interactions, work and other economic pursuits, education, recreation and/or self-care.

Recently, the World Health Organization (WHO) introduced a revision of the International Classification of Impairments, Disabilities and Handicaps (ICIDH), called the International Classification of Functioning and Disability (ICF) (World Health Organization 2001). The ICF provides a unified and standard language and framework for the description of health and health-related states that can be used as a framework to measure health outcomes.

In the ICF, health domains are described from the body, individual and societal perspectives in terms of (1) impairment (i.e., to body structure and function), (2) activity and (3) participation. In the ICF, functioning is an umbrella term that refers to all body functions, activities and participation, while disability is the umbrella term for impairments, activity limitations and participation restrictions. Body structures are the anatomical parts of the body, such as organs, limbs and their components. Body function refers to the physiological functions of the body systems including psychological function. Impairments are problems in body structure or function. Activity is the execution of a task or action by an individual, while participation is involvement in life situations. Activity

limitations are difficulties an individual may have in executing activities, and participation restrictions are problems an individual may experience in involvement in life situations. The ICF model of functioning and disability provides a detailed description of body structure and function, activity and participation. For example, the activity and participation domain includes learning and applying knowledge, general tasks and demands, communication, mobility, self-care, domestic life, interpersonal interactions and relationships, major life areas and community, social and civic life.

The descriptions of body structure and function, activity and participation provided by the ICF can be used to identify important clinical outcomes of sports physical therapy. To illustrate this, consider an athlete with an acute knee sprain. Impairment of body structure may include disruption of the anterior cruciate ligament (ACL) or injury to the meniscus, articular cartilage or subchondral bone. Clinical outcome measures to evaluate body structure may include radiographs and magnetic resonance imaging. Impairment of body function may include limited range of motion, weakness or laxity of the knee. Measures of clinical outcome at the level of impairment of body function for this individual may include goniometry to measure the range of knee motion, isometric or isokinetic testing to measure quadriceps performance, or use of the KT-1000 (MedMetric, San Diego, CA) to measure anterior tibial laxity. Activity limitations experienced by this individual may include difficulty walking, climbing stairs, running, jumping and landing or cutting and pivoting. The resulting participation restrictions may include the inability to participate in sports such as football, soccer or basketball. Clinical outcome in terms of activity and participation can be measured by observing and rating the performance of the individual while executing a variety of activities, or by the use of standardized self-reports of activity limitations and participation restrictions. In summary, clinical outcome measures of body structure and function may include the results of diagnostic studies such as laboratory tests and imaging studies, as well as the findings from clinical examination of the involved structure or region. Clinical outcome measures of activity and participation may include observation of the individual or use of standardized self-reports of activity limitations and participation restrictions.

Health-related quality of life

Health-related quality of life is an individual's perception of his or her health. Broadly, health-related quality of life encompasses an individual's perception of his or her physical, emotional and social function. Health-related quality of life deals with what people perceive their health condition to be and the consequences of it; hence, it is the individual's subjective sense of wellbeing (World Health Organization 2001). Because health-related quality of life encompasses an individual's physical, emotional and social function, it overlaps with the activity and participation domains of the ICF model. As such, health-related quality of life measures can be used to measure the individual's perception of his or her activity and participation.

Many health-related quality of life measures have been developed. These can be classified as general or specific measures of health-related quality of life. General measures of health-related quality of life are designed to be applicable across a number of disease processes and interventions, and across demographic and cultural subgroups (McSweeney & Creer 1995). General health-related quality of life instruments are designed to give a comprehensive and general overview of health-related quality of life. General health-related quality of life measures are usually multidimensional and scores can be obtained for each dimension, or they can be combined to provide an overall measure. The most widely known and accepted general measure of health-related quality of life is the Medical Outcomes Study Short Form - 36 (McHorney et al 1993, 1994, Ware & Sherbourne 1992).

General measures of health-related quality of life permit comparisons across populations with different health conditions (Guyatt et al 1993, McSweeney & Creer 1995) and are more likely to detect unexpected effects of intervention (Kessler & Mroczek 1995, McSweeney & Creer 1995). An important limitation of general health-related quality of life measures is that they tend to be less responsive than specific measures of health-related quality of life, to changes in health status (Guyatt et al 1993). Therefore, use of general health-related quality of life measures might make it more difficult to detect the effects of an intervention for a specific condition. General measures of health-related quality of life are susceptible to ceiling effects. The presence of ceiling effects limits the ability to detect the effects of intervention, especially when used by young, healthy, high-level functioning individuals such as athletes. Because general measures of health-related quality of life measure a broad range of health including emotional function, the content may appear less relevant to patients and clinicians. Finally, general measures of health-related quality of life tend to be longer and more difficult to score.

Specific health-related quality of life measures are designed to focus on aspects of health that are specific to the primary condition or population of interest, with the intent of creating a more responsive measure (Guyatt et al 1993). Specific measures of health-related quality of life have been developed for specific diseases (e.g. osteoarthritis of the knee), specific populations of patients (e.g. the frail elderly), specific functions (e.g. physical function) or for symptoms (e.g. pain) (Guyatt et al 1993).

Specific measures of health-related quality of life are responsive to small changes in the patient's condition and are easy to administer and interpret (McSweeney & Creer 1995). The increased responsiveness of specific measures of health-related quality of life stems from the fact that they include only those important aspects of health-related quality of life that are relevant to the condition or population being studied (Guyatt et al 1993). Specific health-related quality of life measures usually relate closely to areas commonly assessed by clinicians, therefore they are more likely to be accepted by clinicians for routine use. Since specific health-related quality of life measures relate more closely to a particular condition, they are also more likely to be accepted by patients. Disadvantages of specific measures of health-related quality of life are that they do not measure all

aspects of health status and they do not allow for comparisons between different disease states and/or populations.

Specific health-related quality of life measures include disease-specific, region-specific and patient-specific measures. Disease-specific measures of health-related quality of life are developed for a particular injury or illness. The content of disease-specific, health-related quality of life measures includes the symptoms, activity limitations and participation restrictions commonly experienced by individuals with the injury or illness for which the instrument was developed. Examples of disease-specific health-related quality of life measures include the Lysholm Knee Score (Tegner & Lysholm 1985), the Cincinnati Knee Rating System (Noyes et al 1984, Barber-Westin et al 1999) and the Quality of Life Assessment in Anterior Cruciate Ligament Deficiency (Mohtadi 1998) for knee ligament injuries. Also, there is the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) (Bellamy et al 1988) for osteoarthritis of the knee and hip, and the Western Ontario Shoulder Instability Index (Kirkley et al 1998) for shoulder instability.

Region-specific, health-related quality of life measures have been developed to determine the effects of a variety of pathologies and impairments affecting a particular region. The content of region-specific measures of health-related quality of life reflects the symptoms, activity limitations and participation restrictions commonly experienced by individuals with impairment of the particular region for which the instrument was developed. Examples of region-specific measures of health-related quality of life include the following:

- Disabilities of the Arm Shoulder and Hand Index (DASH) (Beaton et al 2001a) for the upper extremity
- American Shoulder and Elbow Surgeons Patient Self-Evaluation Form (Richards et al 1994) for the shoulder
- Simple Shoulder Test (Lippitt et al 1993) for the shoulder
- Oswestry Low Back Pain Disability Questionnaire (Fairbank et al 1980) for the lumbar spine
- Quebec Back Pain Disability Scale (Kocec et al 1995) for the lumbar spine
- Lower Extremity Function Scale (Binkley et al 1999) for the lower extremity
- Knee Outcome Survey for the knee (Irrgang et al 1998)
- International Knee Documentation Committee Subjective Knee Form (Irrgang et al 2001) for the knee
- Foot and Ankle Ability Measure (Martin et al 2005) for the foot and ankle.

Patient-specific measures of health-related quality of life are defined by the patient. Patients are requested to provide a list of three to five relevant activities that they are either unable to do or have difficulty doing as a result of their problem, and then provide a rating of the difficulty they have doing each activity on an 11-point scale that ranges from 'unable to do' to 'able to do at preinjury level' (Stratford et al 1994, Westaway et al 1998). Patient-specific measures have been tested for the low

back (Stratford et al 1994), knee (Chatman et al 1997) and cervical spine (Westaway et al 1998). Patient-specific measures of health-related quality of life are applicable to a large number of clinical conditions, efficient and easy to administer and record and have been found to have adequate psychometric properties. While patient-specific measures of health-related quality of life are responsive to within-subject change, between-subject comparisons are not possible because the content of each is determined by each patient.

Clinical outcomes that should be measured

The most important clinical outcome of rehabilitation for athletes is whether they can return to their prior level of activity and participation with the same intensity, frequency, duration and skill without symptoms and risk of reinjury. Furthermore, this outcome should be achieved in the shortest period of time possible. While on the surface this outcome appears easy to determine, it is difficult to quantify due to the varying demands of sports and levels of participation. Thus, function (i.e. the activity and participation) of the athlete is an important clinical outcome of sports physical therapy.

In addition to being an important outcome of treatment, sports activity is also an important prognostic factor in the sports medicine population because people who are very active have different expectations and demands than those who are active. In terms of outcome, it is important to know if a patient has returned to his or her preinjury level of sports activity in terms of the frequency, intensity and duration of participation as well as the length of time needed to return to this level. Defining outcome only in terms of absence of symptoms may be misleading if the level of sports activity is not considered. For example, the outcome would be considered suboptimal if prior to injury, an athlete could participate in very strenuous sports that require sprinting, cutting, pivoting, jumping and landing with no symptoms, but after surgery and rehabilitation can only return to moderate sports that involve running and turning without symptoms.

Because the frequency, intensity and duration of sports participation vary widely among patients, it is important for studies evaluating such individuals to clearly describe the patient's activity profile. For example, a study describing a new surgical technique for a knee disorder should document the patients' activity level to ensure that the results can be applied to the appropriate patient population. For studies comparing two groups of patients, it is important for the activity levels of the two groups to be similar in order to avoid a biased comparison.

In a systematic review (Marx et al 2001), five activity level rating scales that are potentially applicable to outcomes studies for knee injuries in sports medicine were identified (Daniel et al 1994, Noyes et al 1989, Seto et al 1988, Straub & Hunter 1988, Tegner & Lysholm 1985). There were inherent problems with each of the available instruments, which led to the construction of a new rating scale for this purpose (Marx et al 2001). This activity rating scale consists of four questions relating

to the frequency with which the patient runs, cuts, pivots and decelerates. It has been demonstrated to be reliable and valid (Marx et al 2001). This scale is recommended in addition to a knee-specific health-related quality of life instrument for the evaluation of athletic patients with disorders of the knee. An activity rating scale for the upper extremity has recently been developed and has demonstrated acceptable levels of test-retest reliability and validity (Brophy et al 2005).

Using the ICF model of functioning and disability, the range of clinical outcome measures should include measures of body structure and function, activity, and participation. Whyte (1994) suggested that the level of outcome measurement should be at or higher than the level of intervention. For example, the aim of ACL reconstruction is to restore stability of the ACL deficient knee. Thus an appropriate clinical outcome measure of ACL reconstruction is anterior and rotational laxity of the knee. One would expect that if surgery were successful, it would reduce anterior tibial translation as measured with the KT-1000 and eliminate the pivot shift. Also, restoring stability of the knee should allow the athlete to return to running, jumping and landing, cutting and pivoting, and ultimately to sports such as football, soccer or basketball. Thus, potential clinical outcome measures following ACL reconstruction include measurement at the levels of body structure and function, activity, and participation.

As another example, consider an athlete with a grade II posterior cruciate ligament injury with 6 to 10mm of increased posterior tibial translation compared to the non-involved knee. Non-operative management for this individual may include quadriceps strengthening and a functional exercise progression. In this case, measurement of laxity would not be an appropriate outcome measure for this intervention because the intervention would not be expected to reduce posterior tibial translation. Appropriate clinical outcome measures for this case include strength testing of the quadriceps as well as the ability of the athlete to return to sports activities and participation. Therefore, clinical outcome measures should be thoughtfully selected and be appropriate for the intervention that was provided.

In the past it was believed that there was a direct link between impairment of body structure and function, activity limitations and participation restrictions; however, there is a growing body of literature to the contrary. For example, Snyder-Mackler et al (1997) found no relationship between laxity measured with the KT-1000 and activity and participation measured with the Knee Outcome Survey in ACL deficient copers and non-copers. Pantano et al (2001) obtained similar results 3–5 years after ACL reconstruction. Laxity, range of motion and isokinetic quadriceps and hamstring strength were not significantly related to function and disability as measured with the Knee Outcome Survey. Thus, at least for the knee, there does not appear to be a direct relationship between impairment of body structure and function and the resulting activity limitations and participation restrictions.

This lack of a direct relationship between impairment of body structure and function and activity and participation limitations is inherent in the ICF model (Fig. 13.2). In this model, disability is the outcome of a complex interaction between the

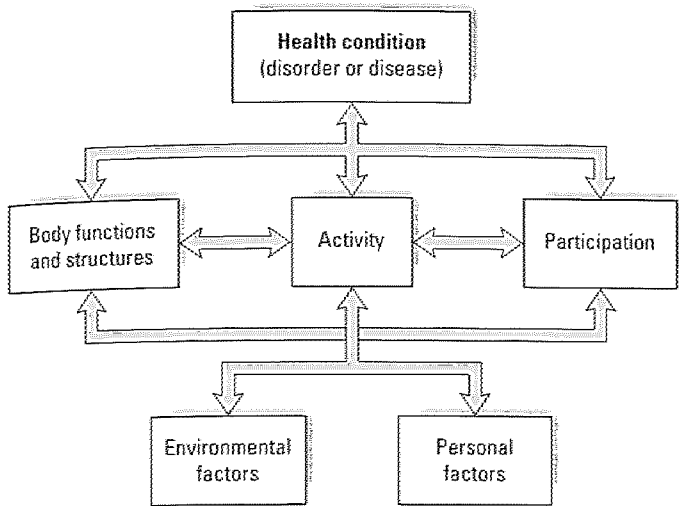


Figure 13.2 • Model of functioning and disability proposed by the World Health Organization in the International Classification of Functioning Disability and Health. (Reproduced with permission from ICF – International Classification of Functioning and Disability, World Health Organization, 2001.)

individual's health condition and contextual factors (World Health Organization 2001). Contextual factors include environmental and personal factors. Environmental factors are the make up of the physical, social and attitudinal environment in which individuals live and conduct their lives. Personal factors include gender, race, age, other health conditions, fitness, lifestyle, habits, upbringing, coping styles, social background, education, profession, past and current experiences, overall behavior pattern and character style, individual psychological assets and other characteristics, all or any of which may play a role in disability (World Health Organization 2001). Research is needed to explore these complex interactions between health conditions and contextual factors and the resulting disability.

The lack of a direct relationship between impairment of body structure and function and the resulting activity limitations and participation restrictions implies that measures of impairment should not be combined with measures of activity limitations and participation restrictions into a single composite score. Rather, reports of clinical outcome should include separate summaries of relevant measures of impairment of body structure and function that are appropriate for the interventions that were provided and valid measures of activity limitations and participation restrictions. Furthermore, because the relationship between impairment and the resulting activity limitations and participation restrictions is not direct, and because activity limitations and participation restrictions are of the utmost concern to the athlete, the primary clinical outcome should be measures of activity limitations and participation restrictions. Activity limitations and participation restrictions may be measured either through direct observation of performance or by general or specific measures of health-related quality of life.

Selection of clinical outcome measures

When selecting clinical outcome measures, one must consider the purpose for which the information will be used as well as practical and psychometric considerations of the instrument. Kirshner & Guyatt (1985) classified health indices according to their purpose as discriminative, predictive or evaluative measures of health. Discriminative health measures are those that distinguish individuals or groups of individuals on an underlying dimension when no external criterion is available for validating the measure. Predictive health indices are those which attempt to classify individuals into a set of predefined categories when an external criterion measure exists concurrently or in the future to determine if the individual has been correctly classified. Evaluative health indices are those that have been developed to measure change within an individual or of a group over time on the dimension of interest. To demonstrate the effects of sports physical therapy, one is interested in measuring change in body structure and function, activity and participation over time. Thus, demonstration of clinical outcomes requires the use of evaluative health indices.

Practical considerations for selection of clinical outcome measures include ease of use, acceptance and costs (McSweeney & Creer 1995). Self-administered health-related quality of life measures are attractive clinical outcome measures because they require minimal resources for administration; however, the reading level must be appropriate for the intended audience. Ease of use of self-administered health-related quality of life measures is determined by the time required for administration and scoring, as well as the effort required to interpret and use the data. Costs for administering and scoring a measure of health-related quality of life will greatly influence its acceptance in clinical practice and research. Clinical outcome measures that require special expertise or equipment for administration, scoring and interpretation are likely to be more costly and less readily accepted than self-administered health-related quality of life measures that can be scored manually (McSweeney & Creer 1995).

Psychometric considerations for selection of a clinical outcome measure

A number of authors have discussed psychometric considerations for selecting a clinical outcome measure (Guyatt et al 1993, Hoffman et al 1995, Kessler & Mroczek 1995, Kirshner & Guyatt 1985, Lohr et al 1996, McSweeney & Creer 1995, Testa & Nackley 1994, Testa & Simonson 1996). Important psychometric considerations when selecting an evaluative clinical outcome measure include reliability, validity and responsiveness. According to contemporary validity theory, in which validity is defined as 'the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of the inferences and actions based on the tests scores' (Messick 1989), these psychometric considerations all fall within the

realm of validity. Knowledge of the psychometric characteristics of a clinical outcome measure allows one to interpret the appropriateness and usefulness of the inferences and actions that are based on the scores from the measure. Furthermore, the nature of the evidence that is required to interpret the appropriateness and usefulness of scores from a particular clinical outcome measure is dependent on the intended purpose and/or use of the measure (Kirshner & Guyatt 1985). When selecting a clinical outcome measure, one must consider the psychometric evidence to support the validity of the inferences that will be made and the actions that will be taken based upon the clinical outcome measure.

Validity evidence to support use and interpretation of a clinical outcome instrument to measure change should include evidence that: the items are all related to the construct being measured (i.e. the items are unidimensional), the score is related to other measures of the construct and not unduly related to other constructs (i.e. the score demonstrates convergent and divergent evidence of validity), the score remains stable when the underlying condition measured by the outcome instrument remains stable (i.e. the score demonstrates test re-test reliability) and the score changes with improvement or deterioration of the condition measured by the instrument (i.e. the score is responsive). Among these, evidence for reliability and responsiveness is most important and will be discussed in greater detail below.

Reliability

Reliability of a clinical outcome measure implies consistency of measurement. Measures of reliability include internal consistency and test re-test reliability. Internal consistency is applicable to outcome measures such as measures of health-related quality of life that consist of multiple items. Internal consistency is the degree to which all items on the scale consistently measure the underlying condition and is concerned with measurement errors related to sampling of items that are included on the instrument (Crocker & Algina 1986). Internal consistency is most commonly estimated with coefficient alpha.

Test re-test reliability is the degree to which scores remain stable when there is no change in the underlying construct that is being measured. Test re-test reliability for a clinical outcome measure is estimated by measuring individuals two or more times over a period of time when the individual's condition is expected to remain stable. The amount of time between repeat measurements is an important issue when determining test re-test reliability (Marx et al 2003). The length of time should not be so short that memory or recall artificially inflates the test re-test reliability estimate. Conversely, the length of time between repeat administrations of the clinical outcome measure should not be too long to avoid change in the condition that is being measured. In general, the length of time between repeat administrations of the clinical outcome measure should be relatively short (e.g. 1–3 days) when the condition being measured is expected to change rapidly (e.g. patients within the first 4 weeks after ACL reconstruction). The time between

repeat administrations of the clinical outcome measure should be longer (e.g. 4 weeks or more) when the condition is not expected to change (e.g. individuals who are 3 to 5 years status post-ACL reconstruction). Given this, it is evident that test re-test reliability should not be considered a property of the instrument itself, but rather the degree of consistency of measurement when applied to certain populations under particular measurement conditions (Streiner & Norman 1995).

The type of reliability coefficient used to estimate test re-test reliability is dependent on the nature of the data. Percent agreement and Cohen's kappa statistic, which is the agreement above chance agreement, are recommended for nominal or ordinal level data, while the intraclass correlation coefficient (Shrout & Fleiss 1979) is recommended for interval or ratio level data. Because test re-test reliability is concerned not only with the relative standing of individuals on repeated measurement, but also the degree to which the repeated measurement yields the same score, the intraclass correlation coefficient is recommended over the Pearson correlation coefficient when estimating test re-test reliability for continuous data.

The standard error of measurement can be used to interpret estimates of internal consistency and test re-test reliability. The standard error of measurement (SEM) is defined as:

$$SEM = \sigma\sqrt{1 - r}$$

where σ is the standard deviation of the scores and r is the reliability coefficient. When coefficient alpha is used to determine the standard error of measurement, the interpretation should be limited to the error associated with a measure at a single point in time. If one is concerned with error over repeated measurements, the intraclass correlation coefficient should be used to determine the standard error of measurement. The standard error of measurement can be used to determine a confidence interval that can facilitate interpretation of a score. For example, coefficient alpha for the Activities of Daily Living Scale of the Knee Outcome Survey was found to be 0.92 in a sample of 397 patients presenting for outpatient physical therapy with a variety of knee impairments and the corresponding standard error of measurement was approximately 6 points on the Activities of Daily Living Scale (Irrgang et al 1998). Thus, the 95% confidence interval (i.e. ± 1.96 SEM) for an observed score of 70 ranges from 58 to 82. This means that if an individual has a true score of 70, 95 out of 100 times that the individual is measured the individual would have an observed score between 58 and 82.

When one is concerned about the magnitude of the expected difference between scores on repeat testing, the intraclass correlation coefficient should be used to calculate the SEM. Furthermore, the standard error of measurement should be multiplied by $\sqrt{2}$ to take into account that there is error associated with both the first and second scores (Streiner & Norman 1995). To illustrate this, consider test re-test reliability of the International Knee Documentation Committee Subjective Knee Score, which was found to be 0.94 when estimated over an average interval of 50 days in 33 individuals who were participating

in long-term outcome studies following ACL reconstruction, meniscal replacement or proximal and distal realignment of the knee extensor mechanism (Irrgang et al 2001). The standard deviation of the scores for this data was 18. Given this, the 95% confidence interval of the expected difference between the scores was ± 12.2 points. Thus, a change in the score over a 7-week interval greater than 12.2 points represents a true change beyond measurement error, while a change less than 12.2 points may occur due to chance alone because of the measurement error at each point in time. The expected difference between scores due to error associated with repeated testing is called the minimal detectable change and is interpreted as the amount of change that needs to be observed before it can be considered beyond the bounds of measurement error for an instrument in a particular application (Beaton 2000).

Responsiveness

Responsiveness is the degree to which a clinical outcome score changes as the underlying condition that is measured by the scale changes. A clinical outcome measure that is responsive will reflect improvement as an individual's condition improves and deterioration as the individual's condition worsens. Demonstration of responsiveness for a clinical outcome measure requires evidence that the measure accurately detects change when change has occurred. Studies to demonstrate responsiveness of a clinical outcome measure should link the amount of change in the outcome score to a construct of change. The construct of change is the way that was used to demonstrate that change has in fact occurred (Beaton 2000). Examples of a construct for change include change from before to after treatment of a known efficacy or change in those deemed to be better or worse, based on an external marker of change (Beaton 2000, Stratford et al 1996).

Factors that will affect the magnitude of change for a given instrument include the patient group under study, the type of treatment being studied, timing of the data collection and the construct for change (Beaton 2000). For example, one would expect greater change over a similar time frame for those with an acute condition compared to those that have a chronic condition. The patient group, type of treatment and timing of data collection must be comparable before the results of a responsiveness study can be applied to a particular clinical setting or used to judge the meaningfulness of a change score (Beaton 2000). The construct for change must also be considered when interpreting the results of a responsiveness study. The construct of change is defined by the answers to: (1) Who is the focus of the analysis? (2) Which scores are being compared? and (3) What kind of change or difference is being examined? (Beaton 2000, Beaton et al 2001b).

The focus of the analysis can be at the group or individual level. When the focus of the analysis is at the group level, summary statistics, such as the mean, standard deviation, median and range or combinations of these statistics in the form of an effect size, standardized response mean or Guyatt's responsiveness index are reported for the entire group or the subgroup of

individuals who improved (Beaton 2000). While these group statistics provide useful information about what should be expected for a group of individuals, they do not provide meaningful information that can be used to interpret change of a particular individual.

When the focus of analysis is at the individual level, receiver operating characteristic curves are used to determine the cut-off value for the change score that has the highest sensitivity and specificity for change (Beaton 2000). Sensitivity of change is the proportion of subjects that have improved according to a criterion of change that have a change score above the cutoff point. Specificity of change is the proportion of subjects that have not improved according to a criterion measure of change that have a change score below the cutoff point (Deyo & Centor 1986). When the focus of the analysis is at the individual level, the challenge is to select the cutoff point in the change score that best discriminates between those that have improved and those that have not improved. The results of a responsiveness study, where the focus of analysis has been at the individual level, can be used to determine if a particular individual has improved or not, given the similarities of other aspects of the responsiveness study (e.g. patient group, type of treatment and timing of data collection) (Beaton 2000).

The construct of change for a responsiveness study is also defined by the scores that are being compared (Beaton 2000). Possibilities include comparison of scores within subjects, between subjects or between group differences of within-subject change. A within-subjects comparison of scores is made over time by comparing before and after scores within an individual or group of individuals. A between-subjects comparison entails a comparison of health states between persons at a single point in time. Differences in the scores between pairs of subjects, one of whom states he or she is healthier than the other, can be used to determine minimally clinically important differences; however, it is doubtful that this construct for change is comparable to constructs that link responsiveness to longitudinal change over time (Beaton et al 2001b). A between-groups difference of within-subject change compares the within-group change over time between two or more groups. This comparison is concerned with the amount of change in one group over time compared to the amount of change over the same period of time in another group. An example of this is a clinical trial, in which the within-subject change in the experimental group is compared to the within-subject change in the control group. Demonstration of responsiveness would entail demonstrating that the within-subject change is greater in the experimental group than in the control group. In essence, this comparison is a combination of a within-subjects and between-subjects comparison.

The third determinant of the construct for change is what kind of change is being quantified in the study to provide evidence for responsiveness. The change that is being quantified can include: (1) change that is considered greater than measurement error; (2) change that is observed over time before and after a treatment; (3) change in those that have improved according to a criterion measure of change; or (4) change in

those that have had a major improvement according to a criterion measure of change (Beaton 2000, Beaton et al 2001b).

The minimal detectable change is the amount of change that is considered to be greater than measurement error (Christensen & Mendoza 1986). It is calculated as the confidence interval for the standard error of measurement for the expected difference between before and after scores (see Reliability section above). Calculation of the standard error of measurement to determine the minimum detectable change requires the use of the test re-test reliability coefficient and multiplication by $\sqrt{2}$ to correct for error associated with measurements made at two points in time. The minimum detectable change can be considered to be the lowest change score that can be confidently considered to be beyond measurement error. It provides an anchor for interpretation of a change score because only when a change score exceeds the minimum detectable change can the clinician be confident that the change score represents true change of the individual and not measurement error (Beaton 2000).

Observed change is the change that occurs in an individual before and after a treatment or over a period of time that is expected to result in improvement for most individuals (Beaton 2000, Beaton et al 2001b). No criterion measure of change is used to determine if individuals have truly changed. This type of change assumes that with treatment or the passage of time, most subjects will improve and the analysis is performed on all subjects being studied. The observed change provides information on how much change should be expected for individuals receiving a particular treatment or being observed for a similar period of time.

Estimated change utilizes a criterion measure of change to determine whether a change has occurred (Beaton 2000, Beaton et al 2001b). The criterion measure of change is used to separate a group of individuals into those that have improved and those that have not improved. Many different criterion measures of change have been proposed from different perspectives including the patient, clinician, payer or society (Beaton 2000). The analysis involves contrasting the change score between those that have improved and those that have not improved. The estimated change provides an estimate of the magnitude of the change score in individuals who improve, and because a criterion measure of change is used, it provides an opportunity to determine the best change score to use as a threshold for improvement in similar individuals.

Important change is similar to estimated change; however, it implies that not only has change occurred, but the change that has occurred is important to the patient, clinician, payer or society (Beaton 2000, Beaton et al 2001b). Important change is often referred to as the minimum clinically important change. It is determined by using a criterion measure that determines that important change has occurred.

The above taxonomy developed by Beaton (Beaton 2000, Beaton et al 2001b) can be used when judging the usefulness of a clinical outcome measure to evaluate change over time. When selecting an outcome instrument to measure change, one should review the evidence to support its reliability and validity for a particular application. Ideally, the evidence to support

use of a clinical outcome measure should be determined under conditions that are similar to those under which the measure will be utilized. This requires careful analysis to determine if the patient group, type of treatment, timing of data collection and construct of change that were used in the study to provide evidence for the usefulness of the clinical outcome measure matches the patient group, type of treatment, timing of data collection and construct for change of the application for which the instrument will be used. To facilitate this, Beaton (2000) has provided guidelines for evaluating evidence to support the use and interpretation of a clinical outcome instrument to measure change (see Box 13.1).

Box 13.1 Guidelines for evaluating evidence to support use and interpretation of a clinical outcome instrument to measure change (reproduced with permission from Beaton D E 2000 Understanding the relevance of measured change through studies of responsiveness. *Spine* 25:3192–3199)

First, be clear about the type of information you need to know about. Define your patient group and type of change you need to understand. Second, appraise whether this study offers you the right kind of information by asking these questions:

1. Are the patients similar enough to my own?
 - Yes
 - No
2. Are they looking at a similar type of treatment?
 - Yes
 - No
 - Time between treatment _____
3. What category of responsiveness is being studied?
 - A. Who is the focus of the analysis and the results presented?
 - Individuals or groups of patients?
 - group level
 - individual level
 - B. Which data are being contrasted?
 - over time (same patients, over time)
 - one point in time (between persons)
 - hybrid (between group differences of within person change)
 - C. What type of change is being quantified?
 - minimum change detectable given measurement error
 - observed change in a given population
 - observed change in those deemed to have changed (estimated change) – according to:
 - patient
 - clinician/researcher
 - payer
 - society
 - observed change in those deemed to have had an important change – according to:
 - patient
 - clinician/researcher
 - payer
 - society

Collecting, analyzing and interpreting clinical outcome measures

Outcomes management requires collection, analysis and interpretation of data to improve the quality of care that is provided and to reduce health care costs. The ultimate aim of an outcomes management system should be to increase value, which is defined as the ratio of quality to costs. High value is achieved by maximizing improvements in quality while minimizing costs. Outcomes data can be used to make patient management decisions, and to provide evidence for the effectiveness of interventions provided by physical therapists and other rehabilitation specialists. When utilized as part of a quality improvement initiative, outcomes data can be used to assess performance of individual clinicians as well as of organizations as a whole. In addition to the quality of the outcome measures, validity of the inferences made from outcomes data is dependent on the circumstances under which the data were collected and interpreted.

Collection of outcomes data

To minimize selection bias, outcomes data should be collected from all individuals with the condition of interest. If this is not possible, then outcomes data should be collected from a representative random sample of individuals. Selection bias in outcomes data can arise when the sample is not representative of the population of individuals that are of interest to the clinician. For example, excluding individuals from the analysis with incomplete follow-up data due to individuals terminating treatment before the course of care is complete may result in an overly favorable outcome. While it is acknowledged that collection of outcomes data from all individuals with the condition of interest or collection of data from a truly random sample may not be possible, attempts should be made to detect bias by determining if the sample of individuals for whom outcomes data is complete is different at the start of care from the sample of individuals with incomplete outcomes data. This requires consistent collection of outcomes data from all individuals at the start of care.

An important concept in the ICF classification system is that contextual factors interact with an individual's health condition to determine the individual's level of functioning (i.e. the individual's activity and participation) (World Health Organization 2001). Contextual factors include both personal and environmental factors. Personal factors include features of the individual that are not part of a health condition such as gender, race, other health conditions, fitness, lifestyle, habits, upbringing, coping styles, social background, education, profession, past and current life experiences, character and psychological status (World Health Organization 2001). Environmental factors include the individual's immediate environment, such as the home, workplace or school, as well as services and systems in the community that can have an impact on the functioning of the individual. Because an individual's level of function and disability are the result of an interaction between the individual's

health condition with personal and environmental contextual factors, attempts to describe the clinical outcome of care provided by the physical therapist should include collection of information that may mediate or modify clinical outcome. Thus, an outcomes data collection system should attempt to capture personal and environmental factors that may impact on clinical outcome.

To make valid inferences, the outcomes data collection system should make use of valid (as broadly defined above) clinical outcome measures. When selecting an outcome measure, use of a 'common currency' will facilitate the ability to compare results between organizations as well as to data in the literature for benchmarking purposes. Linking the outcomes data collection system to scheduling and billing systems will facilitate incorporation of process and cost outcomes. Procedures should be established to facilitate systematic collection of outcomes data to allow for assessment of the individual over time. These procedures should minimize burden on staff and patients required to collect the outcomes data. The use of computer technology can greatly facilitate this process.

Components of a comprehensive outcomes data collection system should include relevant clinical outcome measures. This should include both general and specific health-related quality of life measures to quantify disability in terms of activity limitations and participation restrictions. A general measure of health-related quality of life, such as the SF-36, should be included to provide a comprehensive assessment of health, including physical, emotional and social functioning. Additionally, a general measure of health-related quality of life permits comparisons across populations with different health conditions (Guyatt et al 1993, McSweeney & Creer 1995) and is more likely to detect unexpected effects of intervention (Kessler & Mroczek 1995, McSweeney & Creer 1995). A specific measure of health status, such as the DASH for conditions affecting the upper extremity or the Oswestry Low Back Disability Questionnaire for conditions affecting the low back, should be included to enhance detecting the effects of intervention on an individual's level of function and disability.

As interventions provided by physical therapists are often directed at impaired body structure and function, a comprehensive outcomes data collection system may also include relevant measures of impairment. Inclusion of impairment measures will enable clinicians to determine if interventions directed at impairments were effective at alleviating the impairment and will allow for exploration of relationships between impairment, activity limitations and participation restrictions. However, as noted above, because the relationship between impairment and disability is not direct, and because function and disability are of the utmost concern to the athlete, measures of impairment should not be the only clinical outcome measure included in the outcomes database.

To account for factors that may mediate or modify clinical outcome, contextual factors including characteristics of the individual as well as characteristics of the environment in which the individual must function should be included in a comprehensive outcomes data collection system. Examples of personal

factors that may be included are gender, age, level of education and comorbidities. Environmental factors that may be included in an outcomes data collection system are the nature and demands imposed by work or the type of sports activities in which the individual is able to participate. To describe process outcomes, the outcomes data collection system should include information regarding the duration of care, number of visits and type of procedures provided to the individual. This information may be obtained from the scheduling and/or billing systems.

Systematic collection of outcomes data requires the development of forms that are either completed by the patient or clinician. These forms should be user-friendly and utilize a type font size that is easy to read. The forms should include clear instructions to facilitate accurate completion and to minimize the burden of administration. To minimize the burden of completing the form, the form should maximize the use of check boxes. The form should be organized to facilitate scoring and entry of data into a computerized database. Development of forms that can be scanned or faxed into a computerized database can facilitate data entry. Once in a database, routines can be written to automatically score the data and to store it for analysis at a later time.

An alternative to the use of paper forms to collect outcomes data is the use of computers with user-friendly interfaces. Advances in computer technology including the use of touch screens, wireless networks and the internet have opened new avenues for collection of outcomes data.

In the future, it is envisioned that computerized data collection systems that are internet compatible will be available to collect patient-reported outcomes data during an office visit or from the patient's home. In the clinic, the system will make use of touch screen tablet computers that are wirelessly connected to a local server to efficiently and effectively collect clinical outcomes data. Using a computerized system to collect patient-reported outcomes data while the patient is in the clinic takes advantage of unutilized patient time, helps to increase operational efficiency, and provides the clinician and patient with an instantaneous summary of the patient's status. Providing a summary of the patient-oriented data to the clinician at the time of the encounter should facilitate the encounter and should improve the clinician's efficiency. At home, patients could log-on to a secure website to provide patient-reported outcomes data. Being able to collect patient-reported outcomes data from home simplifies data collection and enhances patient follow-up. Clinical data could be input directly into the system using computers located throughout the clinic. This allows for efficient recording of findings from the physical examination as well as the treatment that was provided.

To facilitate efficient longitudinal data collection, the system should make use of condition-, time- and/or protocol-specific algorithms to determine the data that needs to be collected and/or updated at any given encounter. The system should also be able to accommodate algorithms to administer unique questions based upon responses to prior questions. This will allow the data collection process to be tailored to the unique responses

from the patient or clinician, which improves efficiency of data capture. These algorithms should be easily adjustable to tailor the data collection process to unique clinical populations. Finally, the system should be designed to meet all HIPAA regulations. As such, each patient should be assigned a unique identification number that allows linking of data from multiple encounters and all data should be encrypted when transmitted over the internet.

Outcomes data should be collected at the initiation of care, at regular intervals during the course of care and at the conclusion of care. Data that should be collected at the initiation of care include general and specific measures of health-related quality of life, as well as personal and environmental factors that may mediate or moderate the outcome of care. To measure the response to intervention during the course of care, specific measures of health-related quality of life should be collected on a regular basis (e.g. every 1–2 weeks) during the course of care. Collection of specific measures of health-related quality of life at regular intervals during the course of care will also ensure that some follow-up outcomes data is available for analysis should an individual terminate care before the episode of care is completed. At the conclusion of care, general and specific measures of health-related quality of life, as well as a measure of patient satisfaction, should be collected. To assess the lasting effects of intervention after the course of care is completed, efforts should be made to collect follow-up data, including general and specific measures of health-related quality of life. Alternatives to having an individual return to the clinic to collect this data include the use of telephone or mailed surveys that may include the use of electronic mail and the world wide web.

Analyzing and interpreting clinical outcomes data

Clinical outcomes data can be analyzed and interpreted at either the individual or group level. At the individual level, clinical outcomes data can be used to determine if a particular individual is better, worse or unchanged as a result of intervention. This entails calculation of a change score, which is the difference in the clinical outcome measure from initiation of care to follow-up. To determine if the individual is better, worse or unchanged, the change score can be compared to either the minimum detectable change or minimal clinically important difference. If the change score exceeds the minimal detectable change, then one can be confident that the amount of change as measured by the clinical outcome instrument exceeds the bounds of measurement error for the instrument in a particular application. If the change score exceeds the minimal clinically important difference, then we can be confident that the change, as measured by the clinical outcome instrument, exceeds the amount of change described by other similar individuals as being an important amount of change. The validity of the decisions made at the individual level is dependent on how well the evidence for interpretation of the change score matches the condition under which the decision is applied. As discussed in the above section on responsiveness, application of

evidence supporting responsiveness of a clinical outcome measure should consider the degree to which the patient group, type of treatment, timing of data collection and construct for change used to provide evidence of responsiveness for the clinical outcome measure matches the patient group, type of treatment, timing of data collection and construct for change under which the evidence will be applied.

Measures of central tendency and dispersion are used to summarize data at the group level. Measures of central tendency include the mean, median and mode. The mean is the average score for the group. The median is the middle score or the score at the 50th percentile. The most frequent score is the mode. When data are skewed, such as when there are a few extreme scores, the median will provide a better representation of the entire group than will the mean. This is because the few extreme scores affect the mean, while the median will not be affected as much.

Measures of dispersion include the range, standard deviation and variance. The range is the distance between the highest and lowest score for the entire group. The standard deviation is the square root of the variance. The variance is equal to the average squared distance between each individual score and the group average as given by the following formula:

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

In this formula σ^2 is the variance, x_i is the score of the i -th individual from 1 to n , \bar{x} is the group average and n is the number of individuals in the sample. Scores that are spread over a larger range will have a larger variance and standard deviation, while scores that are very closely distributed to the group mean will have a smaller variance and standard deviation.

A number of indices have been developed to summarize change in the clinical outcome measure over time. These include a change score, effect size, standardized response mean and Guyatt's responsiveness index. A complete description and use of these statistics can be found in Stratford et al (1996); however, the effect size and standardized response mean deserve additional comment. The effect size (ES) is calculated as:

$$ES = \frac{\text{Average change}}{\text{Standard deviation of initial scores}}$$

and the standardized response mean (SRM) is calculated as:

$$SRM = \frac{\text{Average change}}{\text{Standard deviation of change scores}}$$

Both the ES and SRM relate the average change in the clinical outcome score from the initial to follow-up measure to the variability of the scores. The ES relates the average change to the variability of the initial measures, while the SRM relates the average change to the variability of the change scores. The resulting statistics interpret the magnitude of change in terms of the variability of the scores. For example, an ES of 0.5 is

interpreted to mean that the average change from the initial to follow-up measure is equal to one-half of a standard deviation of the initial scores. Cohen (1969) provides guidelines for interpreting the ES: an ES greater than or equal to 0.8 is considered large, 0.5 is considered medium and 0.2 is considered small.

Use of an ES to describe clinical outcomes data is described in Fig. 13.3. This figure is a radar graph that displays the ES over the episode of care for patients with a variety of knee impairments. Each axis of the graph represents the ES for one of the 8 scales of the SF-36. Graphs such as this can be used to benchmark (i.e. compare) the performance of an individual or organization against data from an external organization. For example, in Fig. 13.3, the effect of physical therapy intervention provided by a large outpatient rehabilitation organization in the Pittsburgh, PA, region is compared to data published by Jette & Jette (1996). On average, the improvement in physical function achieved by the outpatient rehabilitation organization in Pittsburgh is approximately 0.1 of a standard deviation larger than that reported by Jette & Jette (1996), and the improvement in social function is approximately 0.3 of a standard deviation larger. While this implies that individuals with knee impairments treated in Pittsburgh may experience slightly better

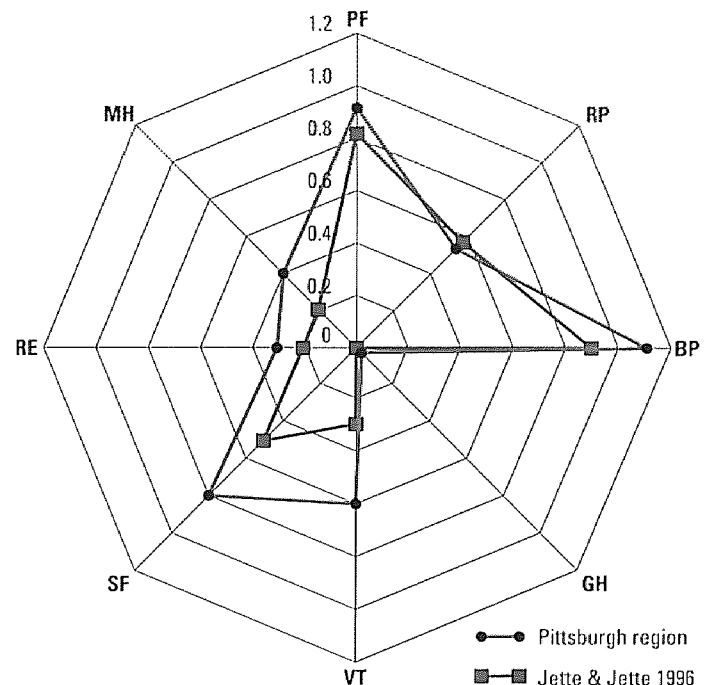


Figure 13.3 • Radar graph of the effect size of each of the 8 scales of the SF-36 following a course of physical therapy in patients with a variety of knee impairments. Data from patients receiving physical therapy from a large multicenter nonprofit outpatient physical therapy organization in the Pittsburgh, PA, region are benchmarked against data reported by Jette & Jette (1996). Each of the axes represents a scale of the SF-36. PF = Physical Function Scale. RP = Role Limitations Due to Physical Function. BP = Bodily Pain. GH = General Health. VT = Vitality. SF = Social Function. RE = Role Limitations Due to Emotional Health. MH = Mental Health.

improvement from a course of physical therapy in health-related quality of life as measured by the SF-36, the results must be interpreted carefully because the validity of this inference is dependent on a number of factors including differences in patient and therapist characteristics, medical diagnosis and length of care between the two samples.

Several statistics can be calculated to determine if the change over time, either within or between groups of individuals, is statistically different from zero. The paired t-test can be used to answer the question 'Is the magnitude of change within a group over time significantly different from 0?', while the independent t-test of change scores can be used to answer the question 'Is the magnitude of change over time in one group larger than the magnitude of change for another group?' The statistical significance of these comparisons is largely dependent on the sample size. A large change may not be significant if the sample size is too small, and a small change may be statistically significant but not clinically meaningful if the sample size is large. Therefore, the significance of any statistical results must be interpreted clinically.

Summary

Clinical outcomes data can be used to facilitate patient management decisions, to assess clinician and organizational performance and to provide evidence for the effectiveness of interventions provided by physical therapists and other rehabilitation specialists. The validity of the inferences made from outcomes data is dependent on the validity of the outcome measures themselves and the circumstances under which the data were collected, analyzed and interpreted. Clinical outcomes may include measures of impairment of body structure and function, activity limitation and participation restriction. However, because the relationship between impairment and the resulting activity limitation and participation restriction is not direct, and because activity limitations and participation restrictions are of the utmost concern to the athlete, the primary clinical outcome should be measures of activity limitation and participation restriction. Activity limitation and participation restriction may be measured either through direct observation of performance or by general or specific measures of health-related quality of life. Clinical outcomes data must be collected systematically to ensure valid inferences from the data.

References

- American Physical Therapy Association 2001 Guide to physical therapist practice: 2nd edition. *Physical Therapy* 81:9-744
- Barber-Westin S D, Noyes F R, McCloskey J W 1999 Rigorous statistical reliability, validity, and responsiveness testing of the Cincinnati Knee Rating System in 350 subjects with uninjured, injured, or anterior cruciate ligament-reconstructed knees. *American Journal of Sports Medicine* 27:402-416
- Beaton D E 2000 Understanding the relevance of measured change through studies of responsiveness. *Spine* 25:3192-3199
- Beaton D E, Katz J N, Fossel A H et al 2001a Measuring the whole or the parts? Validity, reliability, and responsiveness of the Disabilities of the Arm, Shoulder and Hand outcome measure in different regions of the upper extremity. *Journal of Hand Therapy* 14:128-146
- Beaton D E, Bombardier C, Katz J N et al 2001b Looking for important change/differences in studies of responsiveness. *Journal of Rheumatology* 28:400-405
- Bellamy N, Buchanan W W, Goldsmith C H et al 1988 Validation study of WOMAC: a health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. *Journal of Rheumatology* 15:1833-1840
- Binkley J M, Stratford P W, Lott S A et al 1999 The lower extremity functional scale (LEFS): scale development, measurement properties, and clinical application. *Physical Therapy* 79:371-383
- Brophy R H, Beauvais R L, Jones E C et al 2005 Measurement of shoulder activity level. *Clinical Orthopaedics* 439:101-108
- Chatman A B, Hyams S P, Neel J M et al 1997 The patient-specific functional scale: measurement properties in patients with knee dysfunction. *Physical Therapy* 77:820-829
- Christensen L, Mendoza J L 1986 A method of assessing change in a single subject: an alteration of the RC Index. *Behavior Therapy* 17:305-308
- Cohen J 1969 Statistical power analysis for the behavioral sciences. Academic Press, New York
- Crocker L, Algina J 1986 Introduction to classical and modern test theory. Harcourt Brace Jovanovich College Publishers, Fort Worth, TX
- Daniel D M, Stone M L, Dobson B E et al 1994 Fate of the ACL-injured patient: a prospective outcome study. *American Journal of Sports Medicine* 22:632-644
- Deyo R A, Centor R M 1986 Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. *Journal of Chronic Diseases* 39:897-906
- Dobrzykowski E A 1997 The methodology of outcomes measurement. *Journal of Rehabilitation Outcomes Measurement* 1:8-17
- Fairbank J C T, Couper J, Davies J B et al 1980 The Oswestry Low Back Pain Disability Questionnaire. *Physiotherapy* 66:271-273
- Goldstein M S, Elliott S D, Guccione A A 2000 The development of an instrument to measure satisfaction with physical therapy. *Physical Therapy* 80:853-863
- Guyatt G H, Feeny D H, Patrick D L 1993 Measuring health-related quality of life. *Annals of Internal Medicine* 118:622-629
- Hoffman L G, Rouse M W, Brin B N 1995 Quality of life: a review. *Journal of the American Optometric Association* 66:281-289
- Irrgang J J 1996 Outcomes in sports medicine: classification schemes for physical impairments, functional limitations, and disability. Project Focus 96. Conference on Sports Related Injury. Presented by the Foundation for Physical Therapy. American Physical Therapy Association, Alexandria, VA
- Irrgang J J, Synder-Mackler L, Wainner R S et al 1998 Development of a patient-reported measure of function of the knee. *Journal of Bone and Joint Surgery (Am)* 80A:1132-1145
- Irrgang J J, Anderson A F, Bohland A L 2001 Development and validation of the International Knee Documentation Committee Subjective Knee Form. *American Journal of Sports Medicine* 29:600-613
- Jette A M 1994 Physical disablement concepts for physical therapy research and practice. *Physical Therapy* 74:380-386
- Jette D U, Jette A M 1996 Physical therapy and health outcomes in patients with knee impairments. *Physical Therapy* 76:1178-1187
- Kessler R C, Mroczek D K 1995 Measuring the effects of medical interventions. *Medical Care* 33:AS109-AS119
- Kirkley A, Griffin S, McLintock H et al 1998 The development and evaluation of a disease-specific quality of life measurement tool for shoulder instability. The Western Ontario Shoulder Instability Index (WOSI). *American Journal of Sports Medicine* 26:764-772
- Kirshner B, Guyatt G 1985 A methodological framework for assessing health indices. *Journal of Chronic Diseases* 38:27-36
- Kopoc J A, Esdaile J M, Abrahamowicz M et al 1995 The Quebec Back Pain Disability Scale: measurement properties. *Spine* 20:341-352

- Lippitt S B, Harryman D T II, Matsen F A III 1993 A practical tool for evaluating function: the Simple Shoulder Test. In: Matsen III FA, Fu FH, Hawkins RJ (eds) *The shoulder: a balance of mobility and stability*. American Academy of Orthopaedic Surgeons, Rosemont, IL, p 501-518
- Lohr K N, Aaronson N K, Alonso J et al 1996 Evaluating quality-of-life and health status instruments: development of scientific review criteria. *Clinical Therapeutics* 18:979-992
- McHorney C A, Ware J E Jr, Raczek A E 1993 The MOS 36-Item Short-Form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. *Medical Care* 31:247-263
- McHorney C A, Ware J E Jr, Lu J F R et al 1994 The MOS 36-item short-form health survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. *Medical Care* 32:40-66
- McSweeney A J, Creer T L 1995 Health related quality-of-life assessment in medical care. *Disease of the Month* 41:6-71
- Martin R L, Irrgang J J, Burdett R G et al 2005 Evidence of validity for the Foot and Ankle Ability Measure (FAAM). *Foot and Ankle International* 26:968-983
- Marx R G, Stump T J, Jones E C et al 2001 Development and evaluation of an activity rating scale for disorders of the knee. *American Journal of Sports Medicine* 29:213-218
- Marx R G, Menezes A, Horovitz L et al 2003 A comparison of two time intervals for test-retest reliability of health status instruments. *Journal of Clinical Epidemiology* 56:730-735
- Messick S 1989 Validity. In: Linn R L (ed) *Educational measurement*, 3rd edn. American Council on Education/Macmillan Series on Higher Education, New York, p 13-103
- Mohtadi N 1998 Development and validation of the quality of life outcome measure (questionnaire) for chronic anterior cruciate ligament deficiency. *American Journal of Sports Medicine* 26:350-359
- Nagi S Z 1965 Some conceptual issues in disability and rehabilitation. In: Sussman M B (ed) *Sociology and rehabilitation*. American Sociological Association, Washington, DC, p 100-113
- Nagi S Z 1991 Disability concepts revisited: implication for prevention. In: Pope A M, Tarlov A R (eds) *Disability in America: toward a national agenda for prevention*. National Academy Press, Washington, DC, p 309-327
- National Institutes of Health 1992 National Advisory Board on Medical Rehabilitation Research. Draft V: report and plan for medical rehabilitation research. National Institutes of Health, Bethesda, MD
- Noyes F R, McGinniss G H, Mooar L A 1984 Functional disability in the anterior cruciate insufficient knee syndrome: review of knee rating systems and projected risk factors in determine treatment. *Sports Medicine* 1:278-302
- Noyes F R, Barber S D, Mooar L A 1989 A rationale for assessing sports activity levels and limitations in knee disorders. *Clinical Orthopaedics* 246:238-249
- Pantano K J, Irrgang J J, Burdett R et al 2001 A pilot study on the relationship between physical impairment and activity restriction in persons with anterior cruciate ligament reconstruction at long-term follow-up. *Knee Surgery, Sports Traumatology, Arthroscopy* 9:369-378
- Richards R R, An K-N, Bigliani L U et al 1994 A standardized method for the assessment of shoulder function. *Journal of Shoulder and Elbow Surgery* 3:347-352
- Seto J L, Orofino A S, Morrissey M C et al 1988 Assessment of quadriceps/hamstrings strength, knee ligament stability, functional and sports activity levels five years after anterior cruciate ligament reconstruction. *American Journal of Sports Medicine* 16:170-180
- Shrout P E, Fleiss L 1979 Intraclass correlation: uses in assessing rater reliability. *Psychological Bulletin* 86:420-428
- Snyder-Mackler L, Fitzgerald G, Bartolozzi III A R et al 1997 The relationship between passive joint laxity and functional outcome after anterior cruciate ligament injury. *American Journal of Sports Medicine* 25:191-195
- Stratford P W, Binkley J, Solomon P et al 1994 Assessing change over time in patients with low back pain. *Physical Therapy* 74:528-533
- Stratford P W, Binkley J M, Riddle D L 1996 Health status measures: strategies and analytic methods for assessing change scores. *Physical Therapy* 76:1109-1123
- Straub T, Hunter R E 1988 Acute anterior cruciate ligament repair. *Clinical Orthopaedics* 227:238-250
- Streiner D L, Norman G R 1995 *Health measurement scales: a practical guide to their development and use*. Oxford University Press, New York, p 1-231
- Tegner Y, Lysholm J 1985 Rating systems in the evaluation of knee ligament injuries. *Clinical Orthopaedics* 198:43-49
- Testa M A, Nackley J F 1994 Methods for quality of life studies. *Annual Review of Public Health* 15:535-559
- Testa M A, Simonson D C 1996 Assessment of quality-of-life outcomes. *New England Journal of Medicine* 334:835-840
- Verbrugge L M, Jette A M 1994 The disablement process. *Social Science and Medicine* 38:1-14
- Ware J E Jr, Sherbourne C D 1992 The MOS 36-item short-form health survey (SF-36): 1. Conceptual framework and item selection. *Medical Care* 30:473-483
- Westaway M D, Stratford P W, Binkley J M 1998 The patient-specific functional scale: validation of its use in persons with neck dysfunction. *Journal of Orthopaedic and Sports Physical Therapy* 27:331-338
- Whyte J 1994 Toward a methodology for rehabilitation research. *American Journal of Physical Medicine and Rehabilitation* 73:428-435
- World Health Organization 1980 *International classification of impairments, disabilities and handicaps: a manual of classification relating to the consequences of disease*. World Health Organization, Geneva
- World Health Organization 2001 *International classification of functioning disability and health*. World Health Organization, Geneva