PHARMACOECONOMICS

From Theory to Practice

Edited by

RENÉE J. G. ARNOLD

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number: 978-1-4200-8422-1 (Hardback)

Library of Congress Cataloging-in-Publication Data

Pharmacoeconomics : from theory to practice / editor, Renee J.G. Arnold.
p. ; cm. -- (Drug discovery series ; 13)
Includes bibliographical references and index.
ISBN 978-1-4200-8422-1 (hardcover : alk. paper)
1. Drugs--Cost effectiveness. 2. Pharmacy--Economic aspects. 3. Decision making. I. Arnold, Renee J. G. II. Title. III. Series: Drug discovery series ; 13.
[DNLM: 1. Economics, Pharmaceutical. 2. Costs and Cost Analysis. 3. Decision Making. 4. Outcome Assessment (Health Care)--economics. QV 736 P5374 2010]

RS100.P433 2010
338.4'76151--dc22 2009032763

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

11 Patient-Reported Outcome Measures

Dianne Bryant, Gordon Guyatt, and Renée J.G. Arnold

CONTENTS

11.1 PATIENT-REPORTED OUTCOME MEASURES

A patient-reported outcome (PRO) is a direct subjective assessment by patients about aspects of their health, including symptoms, function, emotional well-being, quality of life, utility, and satisfaction with treatment. PROs ask patients to evaluate the impact and functional implications of the disease or treatment to reflect their interpretation of the experience, which is influenced by their internal standards, intrinsic values, and expectations. As such, PROs provide unique information that is unavailable from other sources.[1]

Direct measurement of health from the patient's perspective is an increasingly used outcome measure in clinical trial research. This phenomenon reflects a shift away from an exclusive emphasis on safety and efficacy, and from research that in the past focused narrowly on laboratory and clinical indicators of morbidity. Measuring patients' experience and the extent to which they can function in their daily activities is crucial when the primary objective of treatment is to improve how the patient is feeling. In fact, even when the goal of treatment is to reduce the incidence of seemingly straightforward outcomes like stroke or myocardial infarction, capturing the variability in patients' function and feelings will provide important complementary information if variability in the adverse morbid outcome varies in severity (e.g., a mild versus severe stroke).

11.2 HEALTH AND HEALTH MEASUREMENT

11.2.1 The World Health Organization

The World Health Organization (WHO) defines health as a state of complete physical, mental, and social well-being.[2] The WHO's International Classification of Functioning, Disability, and Health (ICF)[3] was developed to provide a standard language and framework to describe and measure health and health-related states. Within the ICF system, health outcomes are classified according to the effect upon body function, body structure, limitations in activities, and limitations in participation. Health outcomes that measure body function include measures of physiological functions of body systems (e.g., ejection fraction, glucose level, depression, pain, etc). Outcomes that measure body structures include measures of anatomical parts and their components (e.g., x-ray to measure fracture healing, computed tomography to measure tumor size, etc). Activity is defined as the performance of an action, whereas participation, more broadly, is defined as involvement in meaningful activities and fulfillment of roles that are socially or culturally expected of that person. Impairments are problems with body functions or structures. Having an impairment of a body structure (e.g., disc hernia) or function (e.g., reduced range of motion) may contribute to limitations in activities, including activities of daily living, walking, or driving a car, that might also contribute to restrictions in participation. Comprehensive assessment of an individual's health will include measures of body systems and function, as well as limitations in activities and participation.

11.2.2 Health-Related Quality of Life

Health-related quality of life (HRQoL) instruments measure the broad concept of health (physical, mental, and social well-being) by inquiring into the extent of difficulty with activities of daily living (including work, recreation, and household management) and how difficulties affect relationships with family, friends, and social groups, capturing not only the ability to function within these roles, but also the degree of satisfaction derived from doing them. HRQoL instruments often contain items that measure body function (e.g., pain, depression, anxiety) and limitations with activities and participation.

Within the construct of HRQoL, it is common to come across the terms *disease-specific* and *generic*. A disease-specific measure is tailored to inquire about specific aspects of health that are affected by the disease of interest (for example, specific to acne). In contrast, a generic instrument measures general health status, including physical symptoms, function, and emotional dimensions of health relevant to all health states, including healthy individuals.[4]

Disease-specific instruments are more responsive to small but important changes in health than are generic measures.[5] Because the items on a disease-specific HRQoL instrument are so focused on a particular disease, however, they cannot be used to compare the impact of one disease with another. In fact, in some cases, disease-specific measures are so specific that comparisons between different populations within the same disease are not possible (e.g., pediatric versus adult populations). On the other hand, generic HRQoL instruments are useful when measuring the impact of a specific illness or injury across different diseases, severities, and interventions.[4]

A number of previously widely used health profiles such as the Sickness Impact Profile (SIP)[6–11] and the Nottingham Health Profile (NHP)[12–16] are now of largely historical interest; health profiles developed from the Medical Outcomes Study, including the 36-Item Short-Form Health Survey (SF-36)[17–19] and 12-Item Short-Form Health Survey (SF-12)[20] have come to dominate the field of generic health status measurement.

11.2.3 Economic Evaluation of Health

When making decisions on behalf of patient groups, decision-makers must weigh the benefits and risks of treatment, but must also consider whether the benefits are sufficient to merit the health care resources that must be spent to provide them. Limited societal resources necessitate that in order to add a program, society must forgo some other benefit—if the envelope for health spending is fixed, than another health program must be reduced. An economic analysis can inform these decisions. The primary distinction between this paradigm and HRQoL is the inclusion of explicit valuation of both resource consumption and patient-important benefit and harm.

Economic analyses include methods to evaluate different effects (death, effects of stroke on HRQoL, effect of reduction in acne on HRQoL) in the same metric. One way to create the same units is through the concept of preferences. Utilities and values are different types of preferences. Whether you are dealing with utilities or values depends on how questions on measurement instruments are framed; are participants being asked to consider outcomes that are certain (values) or uncertain (utilities)?

The Standard Gamble is the classical method of measuring utility, based directly on the axioms first presented by von Neumann and Morgenstern (utility theory) that describes how a rational individual "ought" to make decisions when faced with uncertainty.[21] During administration of the Standard Gamble, the participant suffering from a health problem, such as severe hip osteoarthritis (in reality or hypothetically), imagines that there is an intervention that will result in a return to perfect health but that there is a risk of death associated with the intervention. Participants are asked to specify the largest probability of death they would be willing to accept before declining the intervention and choosing to remain in their current (suboptimal) health state. The larger the probability of death that the subject is willing to accept, the lower value they place on their current health state. The utility of the present health state—as in all utility measures—is placed on a continuum between death (typically give a value of 0) and full health (typically given a value of 1.0).

For instance, let us assume an individual suffering from severe hip osteoarthritis would be indifferent between his or her current health state and the gamble when the probability of dying is 50%. This would mean that the utility the individual places on a year in this health state is 0.5, in contrast to a year in perfect health, which would be worth 1.0—hence the concept of the QALY (quality-adjusted life year).

The Time Trade-Off[22] is a measure of values. It asks participants to imagine living their lives in their current health states and to contrast this with the alternative of perfect health in exchange for a shorter lifespan (preference-based measured). The administrator provides alternatives of years of life in the present health state versus years of life in perfect health. The more years a subject is willing to sacrifice in exchange for a return to perfect health, the worse the subjects perceive their current health state (see Figure 11.1 for an example with human immunodeficiency virus [HIV]). Utility is calculated by subtracting the number of years sacrificed from the number of years of life remaining divided by the number of years remaining. The number of years remaining is estimated using actuarial tables. So, for instance, if an individual with 30 years of life remaining with severe hip osteoarthritis was ready to trade off 15 of those years to achieve 15 years in full health, the QALYs allocated to 1 year with arthritis would be 0.5.

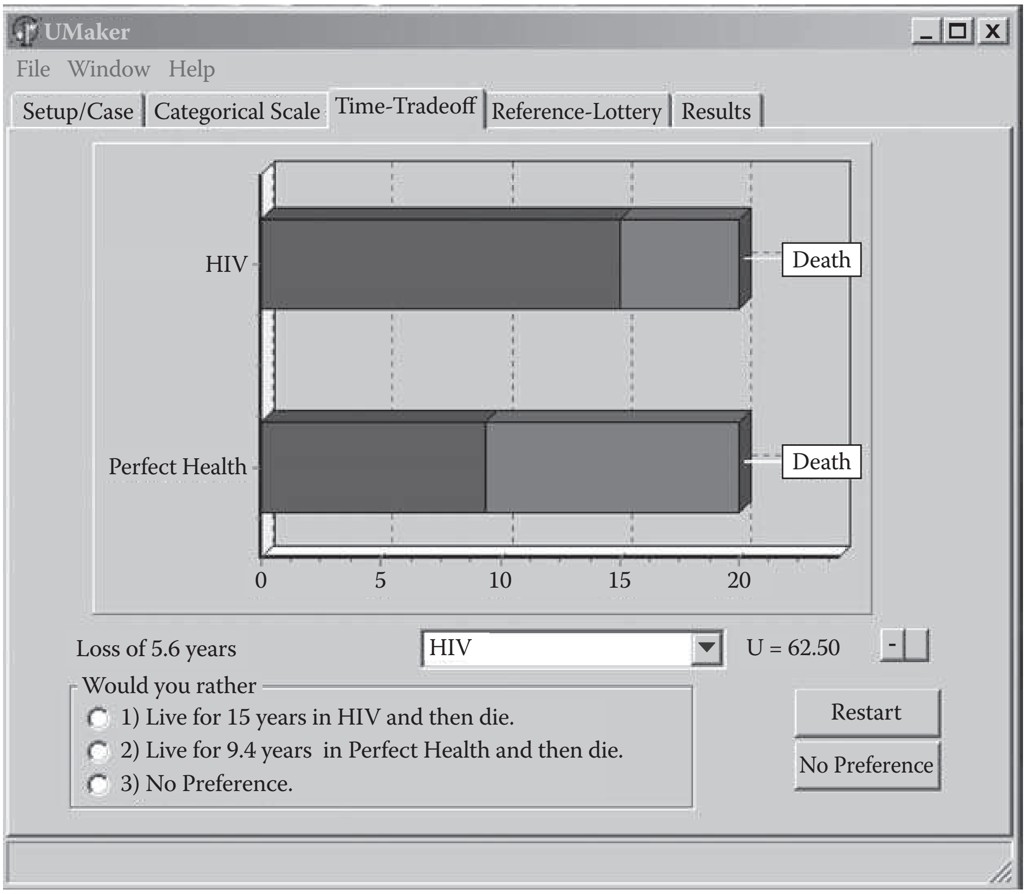

FIGURE 11.1 Time trade-off with HIV health states. Participants are asked to express their preference for living with HIV for 15 years and then dying or living in perfect health for an increasing number of years (less than 15 years) and then dying, until the point of indifference (no preference). Reproduced with permission from U–Maker (Sonnenberg).

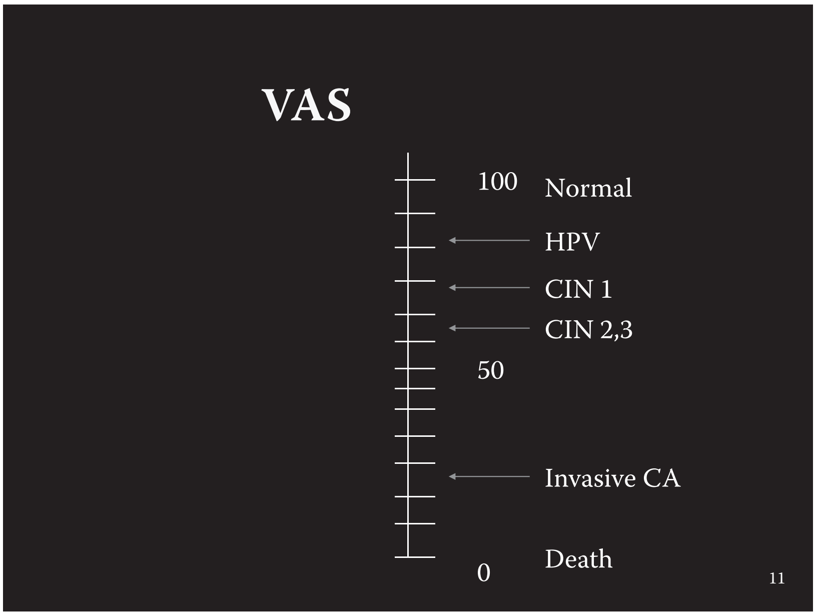

FIGURE 11.2 Visual analog scale.

Another common value-based measure is the Feeling Thermometer (FT). When completing the FT, participants rate their health status using a visual analog scale presented in the form of a thermometer from 0 (worst) to 100 (best)[23–25] (see Figure 11.2 for an example of a visual analog scale for human papillomavirus [HPV] health states).

Measuring preferences for health states using the Standard Gamble or Time Trade-Off is time consuming and can be complex. An alternative method is to use a pre-scored multi-attribute health status classification system. Some common systems include the Quality of Well-Being Scale,[26] Health Utilities Index (HUI),[27–30] European Quality of Life Scale (EQ-5D),[31] and Short Form 6D (SF-6D).[32–35] In general, patients are asked to rate their ability to function in physical, emotional, and social aspects of life, reporting on their health state rather than on their preference for different health states. The patient's preference is assigned on the basis of a mathematical model using preference ratings of health states that have been derived from a random sample of the general population.

11.3 MEASURING PATIENT SATISFACTION

Measurement of patient satisfaction is commonly used to evaluate treatment outcomes. Studies document that satisfied patients are more likely to comply with treatment protocols,[36,37] to use medical care services,[38,39] and to maintain a relationship with a specific provider.[40] Lack of clarity concerning the meaning of satisfaction has, however, been identified as a major weakness.[41–47] Patient ratings of satisfaction are generally directed at either the process of care or treatment outcome,[48] the latter of which is of most interest to clinicians.

Satisfaction may be best thought of as a construct, like health, that cannot be measured directly. Those who have investigated items that are important to patients in determining satisfaction have recommended going beyond inquiry about physical symptoms and function of the diseased body part to include items that probe satisfaction with resolution of social effects of the disease.[49,50] Some have suggested that patient expectations and experiences play a role in defining satisfaction, though the evidence is inconsistent.[51,52]

Experts in the field of measurement of patient satisfaction with treatment outcomes suggest that researchers should develop satisfaction instruments in much the same way they would approach the development of a new measure of quality of life, including the use of qualitative methods for item generation.[48,53] In consulting with the patients, the main objective should be to identify particular contexts in which the affected body part has different meanings, and tailor questions about satisfaction accordingly.

As with HRQoL, the challenge in developing an instrument to measure satisfaction is capturing the necessary content to appropriately measure the construct. In fact, several authors who have compared satisfaction ratings between measures on the same patients have found substantial differences.[54,55] To date, most existing instruments were developed from the perspective of the provider or institution and not the patient.

Like HRQoL, several types of satisfaction measures exist. For example, there are global ratings that contain one or two general questions about overall satisfaction, or multidimensional indexes that probe different aspects of satisfaction, including such things as emotions, desires, perceptions, and expectations.

One disadvantage of global ratings is that they do not capture what patients are considering when reporting their satisfaction. Because of this, global instruments are generally found to be unreliable and tend to be highly skewed.[43,55–57] As with HRQoL, there are also generic and disease-specific instruments to measure satisfaction. Generic instruments can be used to assess satisfaction in any population, whereas disease-specific scales are designed for use in specific patient populations. The pros and cons of generic versus disease-specific instruments are similar to those outlined in Section 11.2.2.

11.4 WHAT ARE THE PROPERTIES OF A GOOD MEASUREMENT INSTRUMENT?

The choice of instrument should align itself with the objectives of the clinician, researcher, or policy-maker. The intent may be to (1) discriminate between patients with different disease severity at a point in time (e.g., whose asthma is impairing function to a greater degree and who to a lesser degree), (2) to predict patient outcome (e.g., functional status may predict mortality in heart failure patients), or (3) to evaluate change following an intervention (e.g., which stroke patients have improved and which have not). To be useful for application in a research and clinical setting for the first two intentions, instruments must be valid (measure what they are supposed to measure—discriminative validity) and reliable (provide consistent ratings between repeated measures in a stable population). If the intention is to evaluate change following treatment, the instrument must be valid (longitudinal validity) and responsive (able to detect important change, even if the magnitude of the change is small).

11.4.1 Validity

An assessment of the validity of a new instrument is an evaluation of whether the instrument measures what was intended. Instruments with the greatest potential for validity will have, in choosing items, consulted with patients, and perhaps clinician experts or patients' family members who have experience with the disease to ask how the disease affects their lives.

One of the first steps in selecting an instrument is to review the items that make up the questionnaire. In some cases, the authors of an instrument will describe its content or include the instrument in an appendix (more common in online publications than in hard copy) so that clinicians can use their own experiences to decide whether what is being measured reflects what is important to patients (*face validity*) in a comprehensive way (*content validity).*

Readers or researchers can use several strategies to provide empirical evidence of the validity of the outcome measure. For example, they can investigate the *criterion validity* of the instrument, which is an assessment of whether the instrument behaves the way it should when compared with a gold standard measurement of the construct (e.g., the gold standard for virtual colonoscopy using imaging approaches is standard colonoscopy). Although measures of body function and structure are likely to have a gold standard reference, there is no gold standard for quality of life.

Construct validity assesses the extent to which the instrument relates to other measures of theoretical concepts (constructs) in the way that it should. Types of construct validity include convergent and discriminant validity. *Convergent validity* examines the degree to which interpretations of scores on the instrument being tested are similar to the interpretation of scores on other instruments that theoretically measure similar constructs. For example, we would expect that patients with poorer performance on a 6-minute walk test will have more dyspnea in daily life than those with better walk test scores, and we would expect to see substantial correlations between a new measure of emotional function and existing emotional function questionnaires.

Discriminant validity predicts weaker correlations with less closely related measures. For instance, one might expect a lower correlation between spirometry and daily dyspnea than between the walk test and daily function. To improve the strength of the inference, investigators pre-specify the magnitude of the correlation that is expected (e.g., no correlation $r<0.20$; weak $r>0.20$—0.35; moderate $r>0.35$—0.50; strong $r>0.50$). They would then administer multiple instruments (spirometry, walk test, other dyspnea questionnaires, global ratings of function) to a group of patients suffering from chronic obstructive pulmonary disease (COPD) to determine the agreement between predicted and observed correlations. The better the agreement between predicted and observed correlations, the stronger is the evidence for construct validity.

The appropriate way to design a study to investigate these types of validity for a discriminative instrument is by looking at the correlations between measures at a single point in time. Such correlations reflect an instrument's *cross-sectional construct validity.*

Conversely, the appropriate way to measure validity for evaluative instruments is by looking at the correlations in change over time between measures. For example, COPD patients who deteriorate in their six-minute walk test score should, in general, show increases in dyspnea, whereas those whose exercise capacity improves should experience less dyspnea; a new emotional function measure should show improvement in patients who improve on existing measures of emotional function. Such correlations reflect an instrument's *longitudinal construct validity.*

11.4.2 Reliability

Reliability is defined as the extent to which an instrument is free from measurement (random) error. In practice, reliability refers to the extent to which an instrument discriminates between individuals in a population in a consistent manner when respondents are in stable health.

The mathematical relationship that defines reliability can be explained by the ratio of the variability in scores between patients to the total variability (i.e., between and within patient variability). Scores obtained on a reliable instrument will demonstrate relatively small differences between scores upon repeated administrations in patients who are stable in their condition (i.e., small within-person variability). Reliability will always appear to be greater when measured in a heterogeneous population with greater variability in scores between patients (e.g., includes patients with no limitations to those with severe limitations) than in a homogeneous population.

An instrument free of random error will have a reliability of 1.0 as long as there is some between-patient variability. As the amount of random error increases in relation to the between-patient variability, the measure of reliability will approach 0. Common expressions of the magnitude of reliability are *Kappa*, when the scale is categorical and *intraclass correlation coefficient* (ICC) when the scale is continuous. Several potential influences may affect the reliability of an instrument, including learning effects, regression to the mean, alterations in mood, circumstance and conditions of administration, and the length of time between assessments. It is also possible that real changes have occurred between consecutive assessments. The most important frequently neglected determinant of reliability is the variability in patient's status on the underlying attribute.

Different techniques to measure the reliability of an instrument include test-retest and inter-rater. *Test-retest reliability* is a measure of the magnitude of the agreement between ratings in repeated administrations of the instrument in a population with a stable health condition. There is no gold standard timeframe between subsequent administrations of the instrument; repeated administrations too close together face criticisms that high levels of agreement reflect patients' ability to remember previous responses, whereas administrations at large intervals run the risk of real changes having occurred within the sample of patients. In general, convention would suggest that any time from 1 to 4 weeks is appropriate, but this will be largely determined by the length of time that patients are expected to remain stable in their condition.

Inter-rater reliability is a measure of the magnitude of the agreement between ratings given by different raters administering the same instrument in a population with a stable health condition. The literature contains some discussion around study design for inter- and intra-rater reliability that suggests that the timing of ratings (e.g., time of day), by different raters, location, and patient position may influence agreement between raters.[58] Depending on the instrument, raters may be able to assess the same patient at fairly tight intervals whereas other outcomes may need to be measured on different days (e.g., measuring maximum strength that requires recovery time).

Internal consistency reliability is quite different from test-retest and inter-rater reliability, and measures the extent to which items in an instrument yield similar scores in the same patients on a single administration. The internal consistency reliability coefficient (R) is used to calculate the standard error of measurement (SEM), which provides an easily defined estimate of the reproducibility of individual measurements ($SEM = \sigma(1 - R)^{1/2}$) and can be used to determine whether true change has occurred within an individual ($\sqrt{2}$ x SEM).[59] Internal consistency is very limited as a measure of reliability because it relates only to the correlation between items on a single administration, and makes no attempt to assess the degree of variability on repeated administration of a measure.

11.4.3 Sensitivity to Change, Responsiveness, and Minimally Important Difference

Many people use the terms "sensitivity to change" and "responsiveness" interchangeably, but by some definitions there are important differences. Sensitivity to change has been defined as the ability of an instrument to measure true change in the state being measured regardless of whether it is relevant or meaningful to the patient or clinician.[60] In contrast, responsiveness has been defined as the ability of the instrument to detect change that is important to the patient in the state being measured even if that difference is small.[60,61] It follows that the minimally important difference (MID) is defined as the smallest difference in score in the outcome of interest that informed patients or informed proxies perceive as important, either beneficial or harmful, and that would lead the patient or clinician to consider a change in management.[62,63]

The magnitude of change that constitutes an MID for many objective outcomes may be intuitive to the clinician (changes in platelet count or serum creatinine). For most PRO measures however, the magnitude of change that constitutes an MID is not self-evident, creating difficulties with interpreting the results of studies that report changes in PRO. In studies that show no difference in HRQoL when patients receive a treatment versus a control intervention, clinicians should look for evidence that the instrument has been shown to be responsive to small or moderate-sized effects in a similar population in previous investigations. In the absence of this evidence, it is unknown whether the intervention was ineffective or whether the instrument was not responsive.

11.5 INTERPRETING THE RESULTS OF A STUDY THAT REPORTS PATIENT-REPORTED OUTCOMES

Physicians often have limited familiarity with methods of measuring how patients feel or their ability to do the things they need or want to do. At the same time, published articles recommend administering or withholding treatment on the basis of its impact on patients' well-being. Thus, if a measure is to be clinically useful, its scores must be interpretable. Interpretability is greatly enhanced if we know the magnitude of the change in score that is important—the MID.

Strategies to define important change have included distribution-based approaches and anchor-based approaches. In general, distribution-based approaches relate the magnitude of the effect to some measure of variability. For example, in a simple before–after comparison, one could calculate the difference between scores before and after treatment divided by the standard deviation of scores at baseline; the resultant statistic is coined the "effect size." In a parallel groups design, the effect size is generated by calculating the difference in scores between the treatment and control group divided by the standard deviation of the change that patients experienced during the study.

A rough rule of thumb for interpreting effects sizes is that changes of a magnitude of 0.2 represent small changes, 0.5 moderate changes, and 0.8 large changes.[64] Interpretation using effect sizes remains problematic because it is sensitive to the homogeneity of the distribution of the sample of patients who participated in the study (i.e., estimates of variability will vary from study to study). In other words, the same difference between treatment and control will appear as a large effect size if the sample is homogenous (patients are similar and thus there is a small between-patient variability, which defines the standard deviation) and as a small effect size if the sample is heterogeneous (patients are dissimilar and thus there is large between-patient variability).

On the other hand, anchor-based approaches involve comparing the magnitude of the change observed on a PRO to an anchor or independent standard that is itself interpretable. The anchor may be defined by achieving change on some external criteria, for example, changing category increasing on a well-known classification system for disease or functional severity (e.g., moving from New York Heart Association Functional Classification III to II) or moving in or out of a diagnostic category (e.g., from depressed to non-depressed, or the reverse).

Another common anchor-based approach, the global rating of change, follows patients longitudinally and asks them to report whether they got better, stayed the same, or got worse. If better or worse, patients rate how much change has occurred—for example, they may rate the degree of change from 1 (minimal change) to 7 (a very large change), where 1 to 3 indicates a small but important change. In the most common way of using this approach, the investigators estimate the MID as the average of the change scores on the PRO that corresponds to a small but important change (that is, the average change in patients who have rated themselves as 1 to 3 on the degree of change rating).

11.6 EXAMPLE OF USE OF HRQOL IN HPV DECISION-ANALYTIC MODELING

Goldie and colleagues[65] used age-specific quality weights for non-cancer states (range from 0.92 in women aged 25–34 years to 0.74 in women older than 85 years) based on data from the Health Utilities Index (Mark II Scoring System) and quality weights for the time spent in cancer health states (range 0.65 for Stage I to 0.48 for Stage IV invasive cervical cancer) from utility estimates by the Institute of Medicine's Committee to Study Priorities for Vaccine Development. These weights were then multiplied by the time spent in the health state and then summed to calculate the number of QALYs in the cost-effectiveness model (see Chapter 9 on use of utilities in HPV modeling).

11.7 SUMMARY

Patient-reported outcome measures provide information gathered directly from the patients about their experiences with the disease and its treatment. Because of the unique perspective offered by patient-reported instruments, direct measurement of health from the patient's perspective is popular and has replaced more objective measures as the primary outcome of interest for a broad spectrum of clinical conditions. For the purpose of evaluating studies that include patient-reported outcomes, it is important to understand the fundamentals of reliability, validity, and responsiveness of the outcome measure being used in addition to appraising the validity of the study. To make wise management decisions, patients and clinicians need to know the magnitude of the effect of treatments on a variety of outcomes, including patient-reported outcomes. Investigators must choose an informative method to present their findings to enhance the interpretability and applicability of their results in a clinical setting.

REFERENCES

1. Rothman ML, Beltran P, Cappelleri JC, Lipscomb J, Teschendorf B. 2007. Mayo/FDA Patient-Reported Outcomes Consensus Meeting Group. Patient-reported outcomes: Conceptual issues. *Value in Health* 10 (Nov–Dec)(Suppl 2):S66–75.
2. Preamble to the Constitution of the World Health Organization as adopted by the International Health Conference. Official Records of the World Health Organization, no. 2, p. 100 and entered into force on 7 April 1948 19–22 June, 1946. signed on 22 July 1946 by the representatives of 61 States. New York.
3. World Health Organization. Towards a common language for functioning, disability, and health: ICF The International Classification of Impairment, Disability, and Health. Geneva: World Health Organization. 2002. Report No: WHO/EIP/GPE/CAS/01.3.
4. Jackowski D, Guyatt G. 2003. A guide to health measurement. *Clin Ortho Related Res* 413:80–9.
5. Wiebe S, Guyatt G, Weaver B, Matijevic S, Sidwell C. 2003. Comparative responsiveness of generic and specific quality-of-life instruments. *J Clin Epidemiol* 56 (1):52–60.
6. Bergner M, Bobbitt RA, Carter WB, Gilson BS. 1981. The sickness impact profile: Development and final revision of a health status measure. *Med Care* 19(8):787–805.

7. Bergner M, Bobbitt RA, Kressel S, Pollard WE, Gilson BS, Morris JR. 1976. The sickness impact profile: Conceptual formulation and methodology for the development of a health status measure. *Int J Health Serv* 6(3):393–415.
8. Bergner M, Bobbitt RA, Pollard WE, Martin DP, Gilson BS. 1976. The sickness impact profile: Validation of a health status measure. *Med Care* 14(1):57–67.
9. de Bruin AF, Buys M, de Witte LP, Diederiks JP. 1994. The sickness impact profile: SIP68, a short generic version. First evaluation of the reliability and reproducibility. *J Clin Epidemiol* 47(8):863–71.
10. de Bruin AF, Diederiks JP, de Witte LP, Stevens FC, Philipsen H. 1997. Assessing the responsiveness of a functional status measure: The sickness impact profile versus the SIP68. [Review] [38 refs]. *J Clin Epidemiol* 50(5):529–40.
11. de Bruin AF, Diederiks JP, de Witte LP, Stevens FC, Philipsen H. 1994. The development of a short generic version of the sickness impact profile. *J Clin Epidemiol* 47(4):407–18.
12. Hunt SM, McEwen J. 1980. The development of a subjective health indicator. [Review] [62 refs]. *Sociol Health Illn* 2(3):231–46.
13. Hunt SM, McKenna SP, McEwen J, Backett EM, Williams J, Papp E. 1980. A quantitative approach to perceived health status: A validation study. *J Epidemiol Commun Health* 34(4):281–6.
14. Hunt SM, McEwen J, McKenna SP. 1985. Measuring health status: A new tool for clinicians and epidemiologists. *J R Coll Gen Pract* 35(273):185–8.
15. Hunt SM, McKenna SP, McEwen J, Williams J, Papp E. 1981. The Nottingham health profile: Subjective health status and medical consultations. *Soc Sci Med [A]* 15(3 Pt 1):221–9.
16. Hunt SM, McKenna SP, Williams J. 1981. Reliability of a population survey tool for measuring perceived health problems: A study of patients with osteoarthrosis. *J Epidemiol Commun Health* 35(4):297–300.
17. McHorney CA, Ware JE, Jr, Raczek AE. 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. *Med Care* 31(3):247–63.
18. McHorney CA, Ware JE, Jr, Lu JF, Sherbourne CD. 1994. The MOS 36-item short-form health survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. *Med Care* 32(1):40–66.
19. Ware JE, Jr, Sherbourne CD. 1992. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med Care* 30(6):473–83.
20. Ware J, Jr, Kosinski M, Keller SD. 1996. A 12-item short-form health survey: Construction of scales and preliminary tests of reliability and validity. *Med Care* 34(3):220–33.
21. Von Neumann J, Morgenstern O. *Theory of Games and Economic Behaviour.* Princeton, NJ: Princeton University Press. 1944.
22. Torrance GW, Thomas WH, Sackett DL. 1972. A utility maximization model for evaluation of health care programs. *Health Serv Res* 7(2):118–33.
23. Schunemann HJ, Griffith L, Jaeschke R, Goldstein R, Stubbing D, Guyatt GH. 2003. Evaluation of the minimal important difference for the feeling thermometer and the St. George's respiratory questionnaire in patients with chronic airflow obstruction. *J Clin Epidemiol* 56(12):1170–6.
24. Schunemann HJ, Griffith L, Stubbing D, Goldstein R, Guyatt GH. 2003. A clinical trial to evaluate the measurement properties of 2 direct preference instruments administered with and without hypothetical marker states. *Med Decision Making* 23(2):140–9.
25. Puhan MA, Guyatt GH, Montori VM, Bhandari M, Devereaux PJ, Griffith L, et al. 2005. The standard gamble demonstrated lower reliability than the feeling thermometer. *J Clin Epidemiol* 58(5):458–65.

26. Kaplan RM, Anderson JP. The quality of well–being scale! Rationale for a single quality of life index. In: Walkee SR, Rosser R, Eds. *Quality of Life: Assessment and Application.* London: MTP Press. 1988. p. 51–77.
27. Torrance GW, Feeny DH, Furlong WJ, Barr RD, Zhang Y, Wang Q. 1996. Multiattribute utility function for a comprehensive health status classification system. Health utilities index mark 2. *Med Care* 34(7):702–22.
28. Boyle MH, Furlong W, Feeny D, Torrance GW, Hatcher J. 1995. Reliability of the health utilities index—mark III used in the 1991 cycle 6 Canadian general social survey health questionnaire. *Qual Life Res* 4(3):249–57.
29. Feeny D, Furlong W, Boyle M, Torrance GW. 1995. Multi-attribute health status classification systems. Health Utilities Index. [Review] [58 refs]. *Pharmacoeconomics* 7(6):490–502.
30. Torrance GW, Furlong W, Feeny D, Boyle M. 1995. Multi-attribute preference functions. Health utilities index. [Review] [83 refs]. *Pharmacoeconomics* 7(6):503–20.
31. EuroQol—a new facility for the measurement of health-related quality of life. The EuroQol Group. *Health Policy* 16(3) 1990: 199–208.
32. Brazier J, Roberts J, Deverill M. 2002. The estimation of a preference-based measure of health from the SF–36. *J Health Econ* 21(2):271–92.
33. Brazier J, Roberts J, Tsuchiya A, Busschbach J. 2004. A comparison of the EQ-5D and SF-6D across seven patient groups. *Health Econ* 13(9):873–84.
34. Tsuchiya A, Brazier J, Roberts J. 2006. Comparison of valuation methods used to generate the EQ-5D and the SF-6D value sets. *J Health Econ* 25(2):334–46.
35. Kharroubi SA, Brazier JE, Roberts J, O'Hagan A. 2007. Modeling SF-6D health state preference data using a nonparametric Bayesian method. *J Health Econ* 26(3):597–612.
36. Raper J, Davis BA, Scott L. 1999. Patient satisfaction with emergency department triage nursing care: A multicenter study. *J Nurs Care Qual* 13(6):11–24.
37. Thomas JW, Penchansky R. 1984. Relating satisfaction with access to utilization of services. *Med Care* 22(6):553–68.
38. Lee Y, Kasper JD. 1998. Assessment of medical care by elderly people: General satisfaction and physician quality. *Health Serv Res* 32(6):741–58.
39. Marquis MS, Davies AR, Ware JE, Jr. 1983. Patient satisfaction and change in medical care provider: A longitudinal study. *Med Care* 21(8):821–9.
40. Wartman SA, Morlock LL, Malitz FE, Palm EA. 1983. Patient understanding and satisfaction as predictors of compliance. *Med Care* 21(9):886–91.
41. Abramowitz S, Cote AA, Berry E. 1987. Analyzing patient satisfaction: A multianalytic approach. *Qrb Qual Rev Bull* 13(4):122–30.
42. Fitzpatrick R, Hopkins A. 1983. Problems in the conceptual framework of patient satisfaction research: An empirical exploration. *Sociol Health Illn* 5(3):297–311.
43. Locker D, Dunt D. 1978. Theoretical and methodological issues in sociological studies of consumer satisfaction with medical care. *Soc Sci Med* 12(4A):283–92.
44. Sitzia J, Wood N. 1997. Patient satisfaction: A review of issues and concepts. *Soc Sci Med* 45(12):1829–43.
45. Williams B. 1994.Patient satisfaction: A valid concept?. *Soc Sci Med* 38(4):509–16.
46. Williams B, Coyle J, Healy D. 1998. The meaning of patient satisfaction: An explanation of high reported levels. *Soc Sci Med* 47(9):1351–9.
47. Hudak PL, McKeever PD, Wright JG. 2004. Understanding the meaning of satisfaction with treatment outcome. *Med Care* 42(8):718–25.
48. Hudak PL, Wright JG. 2000. The characteristics of patient satisfaction measures. *Spine* 25(24):3167–77.
49. Hudak PL, McKeever P, Wright JG. 2007. Unstable embodiments: A phenomenological interpretation of patient satisfaction with treatment outcome. *J Med Humanit* 28(1):31–44.

50. Hudak PL, Hogg-Johnson S, Bombardier C, McKeever PD, Wright JG. 2004. Testing a new theory of patient satisfaction with treatment outcome. *Med Care* 42(8):726–39.
51. Linder-Pelz S. 1982. Social psychological determinants of patient satisfaction: A test of five hypothesis. *Soc Sci Med* 16(5):583–9.
52. Kane RL, Maciejewski M, Finch M. 1997. The relationship of patient satisfaction with care and clinical outcomes. *Med Care* 35(7):714–30.
53. Lynn MR, McMillen BJ. 2004. The scale product technique as a means of enhancing the measurement of patient satisfaction. *Can J Nursing Res* 36(3):66–81.
54. Ross CK, Steward CA, Sinacore JM. 1995. A comparative study of seven measures of patient satisfaction. *Med Care* 33(4):392–406.
55. Ware JE, Jr. 1978. Effects of acquiescent response set on patient satisfaction ratings. *Med Care* 16(4):327–36.
56. Blais R. 1990. Assessing patient satisfaction with health care: Did you drop something? *Can J Prog Eval* 5:1–13.
57. A guide to direct measures of patient satisfaction in clinical practice. Health services research group. *CMAJ* 146(10) 1992:1727–31.
58. Hays RD, Anderson RT, Revicki D. Assessing the reliability and validity of measurement in clinical trials. In: Staquet MJ, Hays RD, Fayers PM, Eds. *Quality of life assessment in clinical trials: Methods and practice*. Oxford: Oxford University Press. 1998. p. 169.
59. Stratford PW, Goldsmith CH. 1997. Use of the standard error as a reliability index of interest: An applied example using elbow flexor strength data. *Phys Ther* 77:745–50.
60. Liang MH. 2000. Longitudinal construct validity: Establishment of clinical meaning in patient evaluative instruments. [see comment]. [Review] [37 refs]. *Med Care* 38(9 Suppl):84–90.
61. Kirshner B, Guyatt G. 1985. A methodological framework for assessing health indices. *J Chronic Dis* 38(1):27–36.
62. Schunemann HJ, Puhan M, Goldstein R, Jaeschke R, Guyatt GH. 2005. Measurement properties and interpretability of the chronic respiratory disease questionnaire (CRQ). *COPD: J Chronic Obstruct Pulmon Dis* 2(1):81–9.
63. Schunemann HJ, Guyatt GH. 2005. Commentary—goodbye M(C)ID! Hello MID, where do you come from? comment. *Health Serv Res* 40(2):593–7.
64. Cohen J. *Statistical power analysis for the behavioral sciences.* 2nd ed. Hillsdale, NJ.: Lawrence Erlbaum Associates. 1988.
65. Goldie SJ, Kohli M, Grima D, et al. 2004. Projected clinical benefits and cost-effectiveness of a human papillomavirus 16/18 vaccine. *J Natl Cancer Inst* 96(8):604–15.