OUTCOME ASSESSMENT FOR RESOURCE ALLOCATION IN PRIMARY CARE

ROBERT M. KAPLAN, THOMAS L. PATTERSON, AND ERIK J. GROESSL

MENTAL HEALTH AND PRIMARY CARE

Changes in Primary Care

It has been estimated that approximately one half of all medical care visits are made to primary care physicians (Stafford et al., 1999). Primary care practice has faced major changes, including changes in demographic trends (Golini, 2001), changes in health care reimbursement (Blumenthal, 2001), and an increased emphasis on primary care education in medical school (Burke, Baron, Lemon, Losh, & Novack, 1994). The scope of primary care practice continues to expand, often beyond services that have been the core of primary care training (St. Peter, Reed, Kemper, & Blumenthal, 1999). The role of psychology in primary care remains to be defined.

A significant amount of mental health care is delivered in primary care settings, yet psychologists have played a surprisingly small role in providing evidence on the value of psychological interventions in primary care (Coyne, Klinkman, & Nease, 2002). In less than 10 years, primary care physicians
doubled the number of prescriptions they wrote for antidepressant medications. The number now exceeds 15 million prescriptions per year (Hirschfeld, 1998). Although it is easy to argue that patients are better off with care provided by mental health providers, adequate evidence to support that argument is lacking. In this chapter we discuss measurement and analysis strategies that might be used to provide evidence for the benefits of behavioral interventions in primary care. We begin with a general overview of health care costs and the challenges of delivering mental health services. Next, we consider the need for systematic evaluation of mental health services in primary care settings. Specific approaches to quality-of-life assessment and cost-effectiveness analysis are then presented. We pay particular attention to the technical challenges of cost-effectiveness analysis for the evaluation of mental health services and to the application of standardized methodologies. In the final section of the chapter, we describe an example of cost-effectiveness analysis of screening for depression in primary care.

Changes in Mental Health Care

Like primary care, mental health care has undergone many changes in the last 20 years, some of which have resulted in cost reductions. During the 1990s, there was a substantial change in the way mental health care was delivered. For most patients, mental health care is now managed by for-profit "behavioral health" companies. These companies carved out a portion of the health insurance benefit by selling packaged mental health services at a low cost to health insurers or directly to employers. This relieved the health insurance companies of the responsibilities for mental health care (Iglehart, 1996) and resulted in a significant reduction in business for psychiatrists and psychologists. Instead of employing the services of psychiatrists, managed mental health companies rely heavily on the services of primary care physicians, social workers and, to some extent, psychologists (Meyer, 1993). These changes in policy have had profound effects on psychology. Psychiatrists have complained that the quality of patient care will suffer when nonpsychiatrists provide the care, yet there are simply no data available on whether patients fare better, worse, or the same with these new funding arrangements.

Psychiatrists were not the only professionals affected by the changes in the way mental health care was funded. Psychologists have felt crowded out by counselors with master's of social work and master's of family therapy degrees who might perform similar services for a lower wage. However, psychologists are beginning to play a bigger role in primary care medicine, and it is expected that this role will increase in the future (see chap. 14, this volume). However, to enhance this role, psychologists must be prepared to offer the data that support revisions in the health care system. As these changes take place, there is an increased need to assess outcomes.
WHAT ARE THE CHALLENGES?

Before discussing outcomes assessment, we must review a number of challenges for the analysis of health policy.

National Health Care Costs

Health care costs in the United States have grown exponentially since 1940. Although there was a temporary slowdown in the mid-1990s, the rate of increase began to accelerate again by the turn of the century. It has been estimated that health care in the United States consumed about 14% of the gross domestic product in 2002, while no other country in the world spends more than 10% (Heffler et al., 2001). Although the rate of growth has slowed, the Institute for the Future (2000) estimated that health care expenditures will increase at a rate of 6.4% annually and will account for 15.6% of the gross domestic product by 2010. Despite high expenditures, the U.S. system may not be producing exceptional health outcomes. Among 13 industrialized countries in one recent comparison, the United States ranked 12th when compared on 16 health indicators (Starfield, 2000). Therefore, improving health remains the top priority, but the challenge is to do so without continuing to increase expenses.

Opportunity Cost Problem

Although most provider groups understand that health care costs must be contained, few acknowledge that their own expenditures should be subject to evaluation. Successful lobbying to obtain reimbursement for a specific service may necessarily mean that another service is excluded. Suppose, for example, that the amount that can be spent on health care is fixed at $100 and that $3 of each $100 (3%) is devoted to behavioral health services. If psychologists are able to get $10 of each $100 spent on their services, then there will be less to spend on other, nonbehavioral health services. This is called the opportunity cost problem. Opportunity costs are the foregone opportunities that are surrendered as a result of using limited or fixed resources to support a particular decision. If more money is spent in one sector of health care, then by necessity, less is spent elsewhere. Therefore, an important question is “How do we decide which services should receive more and which should receive fewer resources?”

When confronted with the choice between two good programs, it is always tempting to support both. The difficulty is that it costs more to offer multiple programs. The cost of programs is represented in the fees for health insurance or the cost of health care to taxpayers. A society can choose to offer as many health programs as it wants; however, more programs require
more funding, and health care consumers do not want higher fees for their health insurance or higher taxes.

**Measures of Outcome**

From a societal perspective, there is a need for measures that allow clinicians and policymakers to determine the efficacy of various treatments while taking into consideration costs, potential side effects that may result from treatments, and the dimensions of outcome that may or may not be improved by the treatments. Outcomes researchers have provided a variety of methods to quantify the benefits of health care interventions. In this section we review some of the concerns of outcome measurement, and then we examine methods that can be used for these evaluations.

**What Populations Are We Considering?**

Primary care physicians see a heterogeneous group of patients. In 1994, the top 10 diagnostic clusters represented less than half of all primary care visits (43.4%; Crews, Batal, Elasy, Casper, & Mehler, 1998). These diagnoses range from psychiatric diagnoses to sprains and strains; each disorder represents a unique set of symptoms, which last for varying lengths of time, and the treatment for each of these diagnoses has unique side effect profiles and widely divergent costs. These complexities suggest that outcome measures either must be specific to the disorder in question, which means that direct comparisons between disorders would be difficult, if not impossible, or be able to capture dimensions that are common to all of these disorders.

In addition to treating specific diagnoses, primary care physicians are cast in the role of providing preventive services such as dietary counseling and blood pressure management. It is well understood that the morbidity and mortality associated with chronic diseases such as cancer, heart disease, hypertension, HIV, and diabetes mellitus could be significantly reduced through lifestyle modifications. The Healthy People 2010 project suggests that reduction of morbidity and mortality is a shared responsibility of the individual, family, community, media, government, and health professionals (U.S. Department of Health and Human Services & Healthy People 2010 [Group], 2000). In fact, the majority of the average 18.1-min visit to primary care physicians can be taken up with attention to these matters (Stafford et al., 1999). It may appear that there are no immediate benefits from such activities, but quality preventive care might improve long-term outcomes. This suggests that primary care outcomes should consider measures that demonstrate long-term benefits such as quality-adjusted life years, a concept we describe later in this chapter.

**Comorbidities and Side Effects**

Measurement of outcome is complicated by patient comorbidities. For example, there is a high incidence of both acute and chronic medical prob-
lems among patients with chronic mental illness. Studies of ambulatory psychiatric patients, who are most likely to be seen in primary care settings, reveal that up to 93% have medical problems, many of which were undiagnosed or inadequately treated (Bartsch, Shern, Feinberg, & Fuller, 1990; Felker, Yazel, & Short, 1996). Despite this evidence, the use of general care services is limited (Worley, Drago, & Hadley, 1990). The underdiagnosis of medical problems may be influenced by a lack of ability and resources of people with major psychiatric illness to access care and follow through with treatment (Crews et al., 1998). In addition, the behavior of psychiatric patients is sometimes seen as disruptive to medical personnel (Hall, Beresford, Gardner, & Popkin, 1982). Psychiatric patients also are less likely to engage in behaviors that enhance health and well-being. For example, Holmberg and Kane (1999) used the Health Promoting Lifestyle Profile (S. N. Walker, Sechrist, & Pender, 1987) and found that psychiatric patients engaged in more behaviors linked to premature death, such as overeating and smoking, than did comparison participants.

In addition, most medications and some behavioral treatments affect other areas of health and well-being. For example, an exercise-based intervention may lower depression and may also improve sleep, cause weight loss, or increase joint pain. An antidepressant may not only reduce depression but also increase a patient's attendance and productivity at work. Once again, these examples suggest that interventions have multiple, sometimes unknown, impacts that are not captured with measures specific to the main presenting complaint. Therefore, finding comprehensive, generic measures to assess all areas of health in each patient appears optimal.

Health-Related Quality of Life

It is clear that an individual's health affects his or her quality of life, and it is the goal of health care to maintain optimal functioning and decrease disabilities associated with chronic illnesses (Field & Gold, 1998). This emphasis has led to a focus on health-related quality of life (HRQOL). Broadly defined, HRQOL is a state of complete physical, mental, and social well-being and not merely the absence of disease or infirmity (World Health Organization, 1948). A number of domains of HRQOL have been identified, including physical health, emotional health, cognitive functioning, sexual functioning, social role performance, and work productivity (Ware & Sherbourne, 1992). As decisions are made regarding which treatments produce improvements in HRQOL, it is imperative that careful consideration be given to the measure(s) chosen to quantify changes in HRQOL.

Identifying the target behavior is an important first step in choosing a measure. Although program planners may not have the resources to collect the depth and breadth of information they would like, doing assessments regarding the impact of specific conditions, and having the ability to com-
pare those conditions and translate findings in common units, are important considerations.

A number of approaches have been used in the measurement of HRQOL, both in psychiatric and medical conditions. Approaches to measurement of outcome include generic profile measures that yield dimension-specific scores (e.g., the Sickness Impact Profile; Bergner, Bobbitt, Kressel, et al., 1976; Bergner, Bobbitt, Pollard, Martin, & Gilson, 1976); the Medical Outcomes Study health survey, which is sometimes referred to as the SF-36; Ware & Gandek, 1998) or the use of single indexes (e.g., Karnofsky Performance Status and the Functional Living Index; Ganz, Haskell, Figlin, La Soto, & Siau, 1988). An alternative approach to the measurement of outcomes focuses on specific populations or diseases. One such instrument, developed specifically for use with psychiatric patients, is a broad-based assessment of recent and current life experiences in a variety of life areas that was developed by Lehman, Slaughter, and Myers (1991). Alternatively, investigators may use customized batteries of individual measures that attempt to capture specific dimensions of quality of life thought to be important in particular disorders (e.g., the Social Adjustment Scale, Patterson et al., 1997; Paykel, Weissman, & Prusoff, 1978; the Scales for Assessment of Positive and Negative Symptoms; Andreasen & Olsen, 1982).

Each of these approaches has limitations, including difficulties comparing across dimensions (i.e., weighting or nonweighting of specific dimensions) and across populations, diseases, or both. For the last 10 years, Kaplan and his colleagues have argued that mental and physical health should be assessed using a common measurement unit (Kaplan, 1990, 1994, 1996; Kaplan et al., 1995; Kaplan & California Policy Seminar, 1993; Kaplan, Feeny, & Revicki, 1999; Kaplan, Ganiats, Sieber, & Anderson, 1998). In fact, comparisons between any competing interventions or treatments in health care require that outcomes be expressed using some common denominator. Not all health care interventions are equally efficient in returning benefits for the expended dollar. In the next section, we review several generic approaches to outcomes evaluation and economic analysis in health care.

REVIEW OF SPECIFIC MEASURES

There are numerous methods for the assessment of HRQOL. There is now an entire journal devoted to HRQOL measurement and several professional societies that focus on the topic. Methods for assessing HRQOL represent at least two different conceptual traditions. One grows out of the tradition of psychometric theory, and the other has its roots in decision theory. Several efforts to develop measures of health status were launched in the late 1960s and early 1970s. All the projects were guided by the World Health Organization's definition of health status as a "complete state of physical,
mental, and social well-being and not merely absence of disease" (World Health Organization, 1948). The projects resulted in a variety of assessment tools, including the Sickness Impact Profile (SIP; Bergner, Bobbitt, Kressel, et al., 1976), the Quality of Well-Being Scale (Kaplan, 1996; Kaplan et al., 1998), the SF-36 (Ware & Gandek, 1998), and the Nottingham Health Profile (Lowe, O'Grady, McEwen, & Williams, 1990). Many of the measures examine the effect of disease or disability on performance of social roles, ability to interact in the community, and physical functioning. Some of the systems have separate components for the measurement of physical, social, and mental health. The measures also differ in the extent to which they consider subjective aspects of life quality (Brown, Gordon, & Haddad, 2000).

The psychometric approach attempts to provide separate measures for the many different dimensions of quality of life. Perhaps the best known example of the psychometric tradition is the SIP, a 136-item measure that yields 12 different scores displayed in a format similar to a Minnesota Multiphasic Personality Inventory profile (Bergner, Bobbitt, Pollard, et al., 1976).

The decision theory approach attempts to weight the different dimensions of health in order to provide a single expression of health status. Supporters of this approach argue that psychometric methods fail to consider that different health problems are not of equal concern. A runny nose is not the same as severe chest pain. Experimental trials using the psychometric approach often find that some aspects of quality of life improve while others get worse. For example, a medication might reduce high blood pressure but also produce headaches and impotence. Many argue that the quality-of-life notion is the subjective evaluation of observable or objective health states. The decision theory approach attempts to provide an overall measure of quality of life that integrates subjective function states, preferences for these states, morbidity, and mortality.

Common Methods for the Measurement of Quality of Life

A variety of methods have been proposed to measure quality of life, but we cannot review and critique them all here. Instead, we present some of the most widely used psychometric and decision theory based methods. Readers interested in more detailed reviews should consult McDowell and Newell (1996) and S. R. Walker and Rosser (1993).

Psychometric Methods

SF-36. Perhaps the most commonly used outcome measure in the world today is the Medical Outcome Study Short Form–36 (SF-36). The SF-36 grew out of work by the RAND Corporation and the Medical Outcomes Study (Ware & Gandek, 1998). It was originally based on the measurement strategy from the RAND Health Insurance Study (Manning et al., 1987;
Newhouse et al., 1987). The SF-36 includes eight health concepts: (a) physical functioning, (b) role—physical, (c) bodily pain, (d) general health perceptions, (e) vitality, (f) social functioning, (g) role—emotional, and (h) mental health (Kosinski, Keller, Hatoum, Kong, & Ware, 1999). The SF-36 can be either administered by a trained interviewer or self-administered. It has many advantages. For example, it is brief, and there is substantial evidence for its reliability and validity. The SF-36 can be machine scored and has been evaluated in large population studies. The reliability and validity of the SF-36 are well documented (Keller, Ware, Hatoum, & Kong, 1999; Scott-Lennon, Wu, Boyer, & Ware, 1999; Stewart & Ware, 1992).

Despite its many advantages, the SF-36 also presents some disadvantages. For example, it does not have age-specific questions, and one cannot clearly determine whether it is equally appropriate at each level of the age continuum. The items for older retired individuals are the same as those for children (Stewart & Ware, 1992). Nevertheless, the SF-36 has become the most commonly used behavioral measure in contemporary medicine.

**Dartmouth Primary Care Cooperative Information Project charts.** Several methods have been developed for the assessment of health outcomes in primary care settings. The best developed among these is the Dartmouth Primary Care Cooperative Information Project (COOP; Wasson, Kairys, Nelson, Kalishman, & Baribeau, 1994). The purpose of this effort was to develop a practical system for measuring health status in primary care medicine (Nelson, Landgraf, Hays, Wasson, & Kirk, 1990). The system uses simple, self-rating charts in which physical, mental, and role functioning are self-rated. The COOP charts include the descriptive title and a question asking about functioning during the last 4 weeks. The 5-point response scale is presented in the form of a simple picture. The group developing the measure has performed extensive evaluative research and has documented the reliability and validity of the COOP measures (Larson, Hays, & Nelson, 1992).

**Decision Theory Approaches**

Quality-adjusted life years (QALYs) are generic measures of life expectancy with adjustments for quality of life (Gold, 1996). QALYs consider both benefits and side effects of programs in terms of the common outcome units. Although QALYs are typically assessed for patients, they can also be measured for others, including caregivers who are placed at risk because they experience stressful life events. The Institute of Medicine recommended that population health metrics be used to evaluate public programs and to assist the decision-making process (Field & Gold, 1998).

The need to integrate mortality and quality-of-life information is clearly apparent in studies of heart disease. Consider hypertension. People with high blood pressure may live shorter lives if untreated, longer if treated. Thus, one benefit of treatment is to add years to life. However, for most patients, high
blood pressure does not produce symptoms for many years. Conversely, the treatment for high blood pressure may cause negative side effects. If one evaluates a treatment only in terms of changes in life expectancy, then the benefits of the program will be overestimated, because one has not taken side effects into consideration. On the other hand, considering only current quality of life will underestimate the treatment benefits, because information on mortality (death) is excluded. In fact, considering only current function might make the treatment look harmful, because the side effects of the treatment might be worse than the symptoms of hypertension. A comprehensive measurement system takes into consideration side effects and benefits and provides an overall estimate of the benefit of treatment (Russell, 1986).

How does one integrate side effects and benefits to estimate the overall impact of a disease and its treatment? If a man dies of heart disease at age 50, and one expected him to live to age 75, then one might conclude that the disease precipitated 25 lost life years. If 100 men died at age 50 (and also had a life expectancy of 75 years), one might conclude that 2500 (100 men x 25 years) life years had been lost. Yet death is not the only relevant outcome to consider. Many adults suffer myocardial infarctions that leave them somewhat disabled for long periods of time. Although they are still alive, they suffer diminished quality of life. QALYs take into consideration such consequences. For example, a disease that reduces quality of life by one half will take away 0.5 QALY over the course of each year. If the disease affects two people, it will take away 1 year (2 x 0.5) over each year. A medical treatment that improves quality of life by 0.2 for each of five individuals will result in the equivalent of 1 QALY if the benefit persists for 1 year. This system has the advantage of considering both benefits and side effects of programs in terms of the common QALY units.

Of the several different approaches for obtaining QALYs, most are similar. The three most commonly used methods are the EQ-5D (Feeny, Furlong, Mulhurn, Barr, & Hudson, 1999), the Health Utilities Index (HUI; Feeny et al., 1999), and the Quality of Well-Being Scale (QWB; Feeny et al., 1999).

EQ-5D. The approach most commonly used in the European community is the EQ-5D. This method, developed by Paul Kind and his associates (1997), has been developed by a collaborative group from Western Europe known as the EuroQol group. The intention of this effort was to develop a generic currency for health that could be used commonly across Europe. The concept of a common EuroQol was stimulated by the desire for a common European currency: the Euro dollar. The original version of the EuroQol included 14 health states in six different domains. In addition, respondents placed their health on a continuum ranging from death (0.0) to perfect health (1.0). The method was validated in postal surveys in England, Sweden, and the Netherlands. More recent versions of the EuroQol, known as the EQ-5D, are now in use in a substantial number of clinical and population studies (Gudex, Dolan, Kind, & Williams, 1996; Hurst, Kind, Ruta, Hunter, &
Although the EQ-5D is easy to use and comprehensive, there have been some problems with ceiling effects. Substantial numbers of people obtain the highest possible score.

**HUI.** Torrance, Feeny, Furlong, and their Canadian associates developed the HUI, which is derived from micro-economic theory (Feeny et al., 1999). There have been several versions of the measure, typically identified by “Mark.” The HUI Mark I was developed for studies in the neonatal intensive care unit and had 960 unique health states. In 1992, the HUI Mark II was developed and included 24,000 unique health states. The HUI Mark III, released in 1995, had 972,000 health states. The eight components of the HUI Mark III include (a) vision (six levels), (b) hearing (six levels), (c) speech (five levels), (d) ambulation (six levels), (e) dexterity (six levels), (f) emotion (five levels), (g) cognition (six levels), and (h) pain (five levels). Multiplying the number of levels across the eight dimensions yields the 972,000 states. Using multi-attribute utility scaling methods, judges evaluate levels of wellness associated with each level of each domain. A multi-attribute model is used to map preference for the 972,000 possible states onto the 0.0–1.0 continuum. The HUI has been used in many population and clinical studies. Figure 16.1 shows estimates of the HUI for men and women in the American population. For overall health status, men obtain higher scores early in the life cycle; however, after about age 45, women obtain higher scores, and this difference grows systematically through the remainder of the life span (Kaplan & Erickson, 2000).

**QWB.** A third method, the QWB, integrates several components into a single score. First, patients are classified according to objective levels of functioning that are represented by scales of mobility, physical activity, and social activity. Once observable behavioral levels of functioning have been classified, each individual is placed on the 0–1.0 scale of wellness, which describes where a person lies on the continuum between optimum functioning and death (Kaplan, 1990, 1994, 1996; Kaplan et al., 1995, 1998, 1999; Kaplan & Bush, 1982; Kaplan & California Policy Seminar, 1993).

Most traditional measures used in medicine and public health consider only whether a person is dead or alive. In other words, all living people are assigned the same score. Yet there are different levels of wellness, and there is a need to quantify these levels. To accomplish this, the observable health states are weighted by quality ratings for the desirability of these conditions. Human value studies have been conducted to place the observable states onto a preference continuum, with an anchor of 0 for death and 1.0 for completely well. Studies have shown that the weights are highly stable over a 1-year period and that they are consistent across diverse groups of raters. Finally, one must consider the duration of stay in various health states. Having a cough or a headache for 1 day is not the same as having the problem for 1 year. A health measure must take these durations into consideration. Using this information, one can describe HRQOL in terms similar to years of
life. For example, 1 year in a state assigned the weight of .5 is equivalent to 0.5 of a QALY.

Rating Methods

Cost–utility analysis requires an assessment of utilities for health states. A variety of different techniques have been used to assess these utilities. Some analysts do not measure utilities directly; instead, they evaluate health outcomes by simply assigning a reasonable utility. However, most current approaches have respondents assign weights to different health states on a scale ranging from 0 (for dead) to 1.0 (for wellness). The most common techniques include rating scales, the standard gamble, and the time trade-off. Rating scales provide simple techniques for assigning a numerical value to an object. There are several methods for obtaining rating scale information. One is the category scale. This is a simple partition method in which respondents are requested to assign each case a number selected from a set of numbered categories representing equal intervals. This method, exemplified by the familiar 10-point rating scale, is efficient, easy to use, and applicable in a large number of settings. In a typical administration, the respondent reads
the description of a case and rates it on a 10-point scale ranging from 0 for dead to 10 for asymptomatic, optimum function. The endpoints of the scale are typically well defined.

Another common rating scale method is the visual analog scale. In the visual analog method, the respondent is presented with a line, typically 100 cm long, with well-defined endpoints. The respondent’s task is to mark the line to indicate where his or her preference rests in relation to the two poles. The standard gamble offers a choice between two alternatives: Choice A—living in healthy state with certainty, or Choice B—taking a gamble on a new treatment for which the outcome is uncertain. The respondent is told that a hypothetical treatment will lead to perfect health with a probability of \( p \) or immediate death with a probability of \( 1 - p \). He or she can choose between remaining in a state that is intermediate between wellness and death or taking the gamble and trying the new treatment. The probability \( (p) \) is varied until the respondent is indifferent between Choices A and B. The concept of probability is difficult for most respondents and requires the use of visual aids or props to assist in the interview. Thus, an alternative to the standard gamble uses a trade-off in time, in which the respondent is offered a choice of living for a defined amount of time in perfect health or a variable amount of time in an alternative state that is less desirable. It is presumed that all respondents would choose a year of wellness versus a year with some health problem. However, by reducing the time of wellness and leaving fixed (e.g., 1 year) the time in the suboptimal health state, an indifference point can be determined. For example, a respondent may rate being severely depressed for 2 years as equivalent to perfect wellness for 1 year.

Internet Data Collection

A variety of new techniques are available to assess patient health status over the Internet. One good example is provided by Impact 3, an Online Survey Generation Instrument. Impact 3 was developed by Leslie Lenert at the University of California, San Diego. The site allows the user to build custom-based questionnaires that can collect a wide variety of data. It has standard data collection forms for background information, medical and health history, and demographic characteristics. The site also allows the user to select from a variety of standardized questionnaires. Furthermore, it allows data collection via common instruments such as the visual analog scale, the time tradeoff, willingness to pay, and the standard gamble (Lenert & Kaplan, 2000). The site is available to the public and has the attractive feature of allowing a researcher or clinician to quickly build his or her own Web site (see www.preferences.ucsd.edu).

Cost-Effectiveness Evaluations

In addition to health benefits, programs also have costs. Resources are limited, and good policy decisions require allocations that maximize life ex-
**Cost/QALY for Selected Interventions**

<table>
<thead>
<tr>
<th>Program</th>
<th>Cost/QALY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seat Belt Laws</td>
<td></td>
</tr>
<tr>
<td>Pneum. Vaccine</td>
<td></td>
</tr>
<tr>
<td>CABG, (L Main)</td>
<td></td>
</tr>
<tr>
<td>Oral Gold (Arthritis)</td>
<td></td>
</tr>
<tr>
<td>Behavior Mod, COPD</td>
<td></td>
</tr>
<tr>
<td>Hemodialysis</td>
<td></td>
</tr>
<tr>
<td>Mammography</td>
<td></td>
</tr>
<tr>
<td>CABG, 1 Vessel</td>
<td></td>
</tr>
</tbody>
</table>

**Cost/QALY**

Figure 16.2. Comparison of several programs using cost-quality-adjusted life years (QALY) ratios. (From Kaplan 2001). Pneum. = pneumonia; CABG = coronary artery bypass graft; Mod = modification; COPD = chronic obstructive pulmonary disease.

Expectancy and HRQOL. Methodologies for estimating costs have now become standardized (Gold, 1996). From an administrative perspective, cost estimates include all costs of treatment as well as costs associated with caring for any side effects of treatment. From a social perspective, costs are broader and may include costs of family members who are not working in order to provide care. When one is comparing programs for a given population with a given medical condition, cost-effectiveness is measured as the change in costs of care for the program compared with the existing therapy or program, relative to the change in health measured in a standardized unit such as the QALY. The difference in costs over the difference in effectiveness is the incremental cost-effectiveness and is usually expressed as the cost-QALY ratio. Because the objective of all programs is to produce QALYs, the cost-QALY ratio can be used to show the relative efficiency of different programs (Gold, 1996).

Figure 16.2 shows a comparison of different programs that have been analyzed using the cost-QALY ratio. Some traditional interventions, such as
bypass surgery for one-vessel coronary heart disease, may cost as much as $700,000 to produce a QALY. Screening programs, such as mammography, may also require many resources to produce a QALY. On the other hand, public health programs, such as pneumonia vaccines for the elderly or laws requiring children to be in infant seats and adults to use seatbelts, may produce a QALY at a very low cost. The figure shows a hypothetical “pay line.” It might be argued that programs to the left of the pay line should be funded, but those with cost–QALY ratios to the right of the line be examined more carefully.

Standards for Cost–Effectiveness Analysis

Contrary to the portrayal of cost–effectiveness analysis in the popular media, the purpose of the analysis is not to cut costs; rather, cost–effectiveness analysis attempts to identify which interventions produce the greatest amount of health using the resources that are available. Because of the confusion about cost–effectiveness analysis, the Office of Disease Prevention and Health Promotion in the Public Health Service appointed a panel to develop standards for cost–effectiveness analysis (Gold, 1996; Russell, Gold, Siegel, Daniels, & Weinstein, 1996; Siegel, Weinstein, Russell, & Gold, 1996; Weinstein, Siegel, Gold, Kamlet, & Russell, 1996). In the following sections, we review some of the major elements of cost–effectiveness analysis as defined by this panel.

Perspective. The results of cost–effectiveness analysis may depend on perspective. From the societal perspective, all health care benefits and costs are considered, regardless of who experiences them or pays for them. The administrative perspective evaluates the problem through the eyes of a specific agency. Individual perspectives consider costs and benefits from the viewpoint of an individual citizen or patient. There may be occasions on which results differ dramatically as a function of perspective. A health maintenance organization, for example, may save money by denying a particular mental health service. So, from an administrative perspective, costs may be reduced; however, from a societal perspective, costs may increase, because other agencies may be required to pay for this service or for the consequences of conditions being left untreated. The panel concluded that most analyses should incorporate the societal perspective.

Comparators. Virtually all decisions involve evaluation in comparison to some alternative. Evaluations of innovative new therapies should compare the new approach with care that was standard before the new intervention was available.

Accounting for costs. From the societal perspective, the cost component considers all resources required for the intervention and for the comparator. These include all costs for all people exposed to the program regardless of whether they eventually developed a health problem. In cost–benefit analysis, the cost savings in reduced health care are subtracted
from the cost of an intervention. For example, a psychological intervention may reduce the number of visits to health care providers. If the resources saved by reduced visits exceed the costs of the programs, a cost offset has been achieved. Behavioral programs may offer cost offsets, but careful analysis rarely shows that intervention programs actually save money (Russell, 1986). Some cost–effectiveness analyses examine how changes in utilization as a result of a treatment or intervention affect health care costs.

**Sensitivity analysis.** Many analyses estimate values for variables, and there is uncertainty about whether these estimates are correct. A sensitivity analysis examines how the results of the cost-effectiveness analysis would change if these estimated values were allowed to vary between a realistic upper and lower bound. In other words, researchers examine and report how sensitive their results are to the estimates contained in their analyses.

**Computer simulation and decision modeling.** Although not a guideline in itself, the use of computers in simulating future outcomes of intervention in cost–effectiveness analysis has become popular. It involves developing a decision tree of all known outcomes of a treatment and using estimates of their likelihood to predict what the result would be if all patients were followed until they reach a certain age or their life expectancy is reached. The likelihood of each outcome is based on results from previous studies or the best estimates available. Sensitivity analyses become very important for evaluating the stability of findings based on these estimates, because the outcomes are estimated and are not observed. Published cost–effectiveness simulations using decision models are continuing to increase and are predicted to play an increasing role in health decision making.

The panel standards for cost–effectiveness are now widely recognized; however, they rarely have been applied in behavioral outcome studies. Kaplan and Groessl (2002) reviewed current cost–effectiveness studies in behavioral medicine and found only two studies consistent with the panel’s criteria for high-quality analyses. However, several studies in mental health care in primary care settings have used current methodologies. We describe one example in the following section.

**SCREENING FOR DEPRESSION IN PRIMARY CARE**

A recent example of a cost–utility analysis is provided by an evaluation of screening for depression in primary care. Depression provides a good example because of its high prevalence. Between 5% and 12% of men and between 10% and 25% of women experience at least one major depressive episode during their lifetimes (Kessler et al., 1994). Most of these individuals, if treated at all, are cared for in the primary care system rather than the mental health system. It has been suggested that patients would benefit if primary care physicians screened for depression and initiated early treatment.
However, resources available to primary care physicians are limited, and it is not known whether screening for depression is a good use of the resources in relation to other alternatives. We can no longer assume something should be implemented just because it intuitively makes sense.

Valenstein, Vijan, Zeber, Boehm, and Buttar (2001) performed a cost-utility analysis of the benefits of screening for depression in primary care. The purpose of this analysis was to estimate the cost to produce a QALY (utility). This outcome typically is estimated using one of the decision-based outcome measures such as the QWB, the HUI, or the EuroQol. The comparator was no depression screening. The investigators created a hypothetical cohort of 40-year-old primary care patients and assumed that these patients would either be screened for depression every year or that no screening would take place. The computer simulation followed the patients until they reached either age 90 or were assumed to have died of other causes. Screening consisted of the administration of a self-administered depression questionnaire with follow-up of positive cases by a nurse and primary care provider.

For each time period in the model, it was assumed that the patient was in one of eight states: (a) never depressed; (b) history of depression, in remission; (c) history of depression, still in treatment; (d) significant depressive symptoms; (e) significant depressive symptoms, in treatment; (f) major depression; (g) major depression, in treatment; (h) deceased. To meet the definition of an undetected depressive episode, a patient would need (a) two or more depressive symptoms for at least 2 weeks, (b) to be experiencing functional impairment, and (c) to not meet the Diagnostic and Statistical Manual for Mental Disorders (American Psychiatric Association, 1994) criteria for major depression. All patients who had a diagnosis of minor depression, dysthymia, or major depression were also considered to be depressed and were included. Estimates of the effectiveness of depression treatment and screening were obtained from a review of 350 articles published in the peer-reviewed literature. The prevalence of depression was estimated from published studies that examined patients in primary care settings.

The analysis suggested that, on average, depression screening measures had a sensitivity of 84% and an average specificity of 72% for major depression. For minor depression, sensitivities were significantly lower. Valenstein et al. (2001) used a sensitivity of 35% for depressive symptoms. It was assumed that primary care physicians would initiate treatment in 45% of the patients with major depression and 20% of the patients with depressive symptoms. Furthermore, it was assumed that 3%-5% of the patients would be referred to a mental health specialist and that 26% of those with major depression and 13% of those with symptoms would self-refer during a 12-month interval. Using the QWB, Valenstein et al. used utilities for depression of between .55 and .68. It was assumed that utilities for the general population who do not have depression were between .81 and .90.
The results suggest that, compared with no screening, the annual costs of screening in primary care are $192,444/QALY. If screening is reduced from once a year to once every 5 years, costs are reduced to $50,988/QALY. A one-time screening program (not repeated on a regular time interval) reduced that cost–QALY ratio to $32,053. Valenstein et al. (2001) concluded that regular screening may not be competitive with other health care programs. However, one-time screening competes favorably with other options in health care (Valenstein et al., 2001). A variety of variables affect the analysis. One is the assumption that screening measures are not very good, particularly for minor depression. For example, given that there is a 5% prevalence of major depression in primary care, a sensitivity of .84 and a specificity of .72 lead to many false diagnoses. For every 100 patients screened, there would be 4 true positive cases and 27 false positive results (Kroenke, 2001). Each of these false positives adds significantly to the costs of the program, because each one must be followed up. The analysis also assumes that treatments are not very effective and that adherence to treatment is only about 50%. The study did not evaluate the benefits of cognitive–behavioral therapy in relation to pharmacological management of the condition.

The sensitivity analysis showed that cost of treatment was not an important factor. Varying the cost of professional visits and antidepressant medication by ±50% had only a small impact on the cost–effectiveness ratio. Similarly, cost offset did not have a large effect. For example, assuming that treatment of depression would reduce general medical expenses by 20% had only a minor effect on the cost–QALY estimates.

Although the Valenstein et al. (2001) study is important, it also suggests important directions for future research. In particular, the study makes clear that much work remains to be done. For example, studies have not prospectively evaluated the outcomes for patients who have been assigned to screening or to usual care. Valenstein et al. did the best they could with computer simulation, given the data published in the literature. However, considerably more needs to be known about health outcomes for patients treated for depression. At present, this information is usually estimated rather than measured.

**SUMMARY AND CONCLUSIONS**

The primary care setting has become an important venue for mental health care. This change in the delivery of health care has created many new problems. Health care has become expensive, and there is clearly a need to use resources in a way that provides the most benefit for the most people. Policy shifts have redistributed the way care is delivered. Psychiatrists and, to a lesser extent, psychologists, have lost market share to lower cost providers. Many providers have argued that the change in the way care is delivered will harm patients.

**OUTCOME ASSESSMENT FOR RESOURCE ALLOCATION**

309
Unfortunately, we have little evidence that the changes in the delivery of care are better or worse for patients. To establish this, we need significantly more information on patient outcomes. A variety of measures are available to assess patient outcomes, each from a different perspective, with advantages and disadvantages. These measures include mental health specific measures and generic measures such as the SF-36, the COOP charts, the HUI, and the QWB.

Methodologies for evaluating costs should also be considered in health care decisions. The panel on cost-effectiveness in medicine and health care has offered standards for cost-effectiveness evaluations. Although a growing number of studies apply these methodologies, there are still relatively few systematic cost-effectiveness evaluations of mental health care. Given the lack of data, analyses must depend on the available information. Studies published in the literature offer some surprising findings. For example, regular screening for depression is relatively expensive to produce a QALY. On the other hand, one-time screening competes well with other health care services.

How the information gathered through the various assessment methods discussed in this chapter will shape the future of primary health care remains to be seen. Although it is clear that a large portion of mental health care is currently, and will undoubtedly continue to be, delivered in primary care settings, there is no agreed-on model for how this will be accomplished. A number of ways to manage patients have been discussed, including the referral model and a collaborative care model. Patients with serious mental illness are certainly best cared for in a referral model. The severity of the illness (e.g., schizophrenia vs. mild depression) will in part undoubtedly determine which model to follow. Lower spectrum anxiety and depression might be best cared for in a collaborative care model in which primary care physicians and mental health providers share management (Maser & Patterson, 2002). However, the costs and benefits of these various models of care depend, in part, on the perspective one takes. For example, it may be more cost-effective from the health care system's perspective for the primary care physician to provide the care. In contrast, from the patient's perspective, where improved quality of life is paramount, it is unclear which model is most useful. The data to conduct such cost-effectiveness analyses from each perspective do not currently exist. We encourage investigators to gather the data necessary to address these critical issues.

Many aspects of cost-effectiveness analysis are not covered in this chapter. For more details, readers are referred to Gold (1996), a much more comprehensive source. It is worth mentioning that our presentation of methods for making important health policy decisions may seem detached or insensitive to some readers. Placing dollar and QALY values on human lives raises important ethical issues (Dranove, 2003). Although we argue that these methods should be considered, we do not believe that cost-effectiveness
should be the only information that is used to make these major health decisions. Cost-effectiveness analysis is an objective tool for the evaluation of health interventions that may maximize the overall health of a population. It should be one component of health policy decisions.

Mental health outcomes research offers unusual opportunities for basic methodological and applied research. We believe the most important opportunities are in demonstrations that mental health services in primary care result in benefits to patients. This can be demonstrated through studies that measure patient reported quality-of-life outcomes. The best examples of benefit will result from systematic randomized clinical trials. In addition to demonstration of patient benefit, more studies are needed that demonstrate that mental health services offer good value. More studies are necessary, too, to show that investments in mental health care produce an increment in benefit over more traditional care. It will be necessary to demonstrate that adding another service, which adds cost, can still be justified, because patients experience better outcomes and overall health care costs are reduced. To date, there have been very few systematic studies in this field. Although the work will be difficult, there are plenty of important opportunities in mental health services research.

REFERENCES


OUTCOME ASSESSMENT FOR RESOURCE ALLOCATION


Keller, S. D., Ware, J. E., Jr., Hatoum, H. T., & Kong, S. X. (1999). The SF-36 Arthritis-Specific Health Index (ASHI): II. Tests of validity in four clinical trials. Medical Care, 37(5, Suppl.), MS51–MS60.


Kosinski, M., Keller, S. D., Hatoum, H. T., Kong, S. X., & Ware, J. E., Jr. (1999). The SF-36 Health Survey as a generic outcome measure in clinical trials of patients with osteoarthritis and rheumatoid arthritis: Tests of data quality, scaling assumptions and score reliability. Medical Care, 37(5, Suppl.), MS10–MS22.


