HTA 101
INTRODUCTION TO HEALTH TECHNOLOGY ASSESSMENT

Clifford S. Goodman, Ph.D.
The Lewin Group
Falls Church, Virginia, USA
clifford.goodman@lewin.com

January 2004

TABLE OF CONTENTS

TABLE OF CONTENTS (Cont)

TABLE OF CONTENTS (Cont)

ACKNOWLEDGMENTS

HTA 101: Introduction to Health Technology Assessment, is derived from an evolving set of seminars and other presentations that I have given on health technology assessment since the mid-1980s. This document is an update of *TA 101: Introduction to Health Care Technology Assessment,* which was made available on the web site of the National Information Center on Health Services Research and Health Care Technology (NICHSR) of the National Library of Medicine (NLM) in 1998.

Much of this material was first assembled as a single document for a conference, *Technology Assessment: A Tool for Technology Management and Improved Patient Outcomes,* held in January 1995 in Washington, DC. The conference was sponsored by the VA Health Services Research and Development Service and its Management Decision and Research Center, and the Association for Health Services Research, now known as AcademyHealth. The late Alice Hersh, who was Executive Vice President and Chief Executive Officer of AHSR, provided the impetus and considerable motivation for that conference.

HTA 101 draws from the work of the many people whose publications are cited in the references and reading lists. In particular, I would like to acknowledge the influence of David Banta, Robert Brook, the late Thomas Chalmers, David Eddy, the late John Eisenberg, Egon Jonsson, Bryan Luce, and Fred Mosteller on my understanding of the field and appreciation for the importance of involving others in it. Ione Auston of NICHSR has contributed to this work directly as well as indirectly through her continued efforts to strengthen and encourage sharing and coordination of HTA information resources in the US and around the world. Earl Steinberg, Kevin Marvelle, and the late Allan Meyers provided thoughtful and expert reviews of earlier versions of this work.

As was the 1998 version of this work, the updating, expansion, and preparation of *HTA 101* for distribution and viewing via the World Wide Web was funded by NICHSR, NLM. I wish to acknowledge the support and guidance of Catherine Selden, the NICHSR project officer for this effort.

Clifford S. Goodman
The Lewin Group
January 2004

ABOUT THE AUTHOR

Clifford S. Goodman is a Vice President and Senior Scientist at The Lewin Group, a health care policy and management consulting firm based in Falls Church, Virginia, near Washington, DC. He has more than 20 years of experience in health care evaluation, primarily devoted to assessment of health care technology and related policy analyses. Dr. Goodman's expertise involves technology assessment, evidence appraisal, outcomes research, health economics, decision analytic modeling, and studies pertaining to health care innovation, diffusion, and payment. He has worked with an international range of government agencies; pharmaceutical, biotechnology, and medical device companies; health care provider institutions; and professional and industry associations. On behalf of the US Agency for Healthcare Research and Quality (AHRQ), he is the Director of the Evidence-based Practice Centers Coordinating Center at The Lewin Group.

Dr. Goodman has worked in such diverse technological areas as heart assist devices, interventional cardiology, hypertension management, diagnostic testing, diagnostic imaging, organ and tissue donation, blood supply safety, national cancer policy, substance abuse treatment, and telemedicine. His work on databases in health technology assessment and health services research contributed to the development of the HealthSTAR (now incorporated into MEDLINE) and HSRProj databases of the National Library of Medicine. Dr. Goodman has testified before the US Congress on issues pertaining to Medicare coverage of health care technology. He is a member of the Medicare Coverage and Advisory Committee (MCAC) of the US Centers for Medicare and Medicaid Services (CMS).

Earlier in his career, as a National Research Council Fellow and later as director of the Council on Health Care Technology, he managed and staffed a series of technology assessment projects at the Institute of Medicine of the US National Academy of Sciences, including the landmark study, *Assessing Medical Technologies*. Subsequently, Dr. Goodman was a visiting researcher at the Swedish Council on Technology Assessment in Health Care in Stockholm.

Dr. Goodman is a board member of HTA International (HTAi) and is a Fellow of the American Institute for Medical and Biological Engineering. He did his undergraduate work at Cornell University, received a master's degree from the Georgia Institute of Technology, and earned his doctorate from the Wharton School of the University of Pennsylvania.

I. INTRODUCTION

Technological innovation has yielded truly remarkable advances in health care during the last three decades. In just the last several years, breakthroughs in such areas as antivirals, biotechnology, diagnostic imaging, molecular diagnostics, organ and tissue replacement, surgical techniques, wound care, and computer technology have helped to improve health care delivery and patient outcomes.

The proliferation of health care technology has accompanied burgeoning health care costs, and the former has been cited as "culprit" for the latter. However, the nature and strength of this relationship are complex and evolving (Cutler 2001; Medicare Payment Advisory Commission 2001; Newhouse 1992). Certainly, few patients or clinicians are willing to forego access to state-of-the-art health care technology. In the wealthier countries, and particularly in the US, adoption and use of technology has been stimulated by patient and physician incentives to seek any potential health benefit with limited regard to cost, and by third-party payment, malpractice concerns, provider competition and effective marketing of technologies. Some of the main factors that influence the development and demand for health technology are shown in **Box 1.**

Box 1. Factors That Reinforce the Market for Health Technology

- Advances in science and engineering
- Intellectual property, especially patent protection
- Aging population
- "Cascade" effects of unnecessary tests, unexpected results, or patient or physician anxiety
- Emerging pathogens and other disease threats
- Third-party payment
- Inability of third-party payers to limit coverage
- Financial incentives of technology companies, clinicians, and others
- Clinician specialty training at academic medical centers
- Malpractice avoidance
- Provider competition to offer state-of-the-art technology
- Public demand driven by consumer awareness, direct-to-consumer advertising, and mass media reports
- Strong economies, high employment

In this era of increasing cost pressures, restructuring of health care delivery and payment, and continued inadequate access to care for tens of millions in the US and many more millions around the world, technology remains the substance of health care. Culprit or not, technology can be managed in ways that improve patient access and health outcomes, while continuing to encourage innovation. The development, adoption, and diffusion of technology is increasingly mediated by a widening group of policymakers in the health care sector. Health product makers, regulators, clinicians, patients, hospital managers, payers, government leaders, and others increasingly demand well-founded information to support decisions about whether or how to develop technology, to allow it on the market, to acquire it,

to use it, to pay for its use and more. The growth and development of **health technology assessment** (HTA) in government and the private sector reflect this demand.

HTA methods are evolving and their applications are increasingly diverse. This document introduces fundamental aspects and issues of a dynamic field of inquiry. Broader participation of people with multiple disciplines and different roles in health care is enriching the field. The heightened demand for health technology assessment, in particular from the for-profit and not-for-profit private sectors as well as from government agencies, is pushing the field to evolve keener assessment processes and audience-specific reports. Like the information required to conduct most assessments, the body of knowledge about HTA cannot be found in one place and is not static. Practitioners and users of HTA should not only monitor changes in the field, but have considerable opportunities to contribute to its development.

Origins of Technology Assessment

Technology assessment (TA) arose in the mid-1960s from an appreciation of the critical role of technology in modern society and its potential for unintended, and sometimes harmful, consequences. Experience with the side effects of a multitude of chemical, industrial and agricultural processes, and such services as transportation, health, and resource management contributed to this understanding. Early assessments concerned such topics as offshore oil drilling, pesticides, automobile pollution, nuclear power plants, supersonic airplanes, and the artificial heart. TA was conceived as a way to identify the desirable first-order, intended effects of technologies as well as the higher-order, unintended social, economic and environmental effects (Brooks and Bowers 1970).

The term "technology assessment" was introduced in 1965 during deliberations of the Committee on Science and Astronautics of the US House of Representatives. Congressman Emilio Daddario emphasized that the purpose of TA was to serve policymaking:

> *[T]echnical information needed by policymakers is frequently not available, or not in the right form. A policymaker cannot judge the merits or consequences of a technological program within a strictly technical context. He has to consider social, economic, and legal implications of any course of action (US Congress, House of Representatives 1967).*

Congress commissioned independent studies by the National Academy of Sciences, the National Academy of Engineering (NAE), and the Legislative Reference Service of the Library of Congress that significantly influenced the development and application of TA. These studies and further congressional hearings led the National Science Foundation to establish a TA program and, in 1972, Congress to authorize the congressional Office of Technology Assessment (OTA), which was founded in 1973, became operational in 1974, and established its health program in 1975.

Many observers were concerned that TA would be a means by which government would impede the development and use of technology. However, this was not the intent of Congress or of the agencies that conducted the original TAs. In 1969, an NAE report to Congress emphasized that:

> *Technology assessment would aid the Congress to become more effective in assuring that broad public as well as private interests are fully considered while enabling technology to make the maximum contribution to our society's welfare (National Academy of Engineering 1969).*

With somewhat different aims, private industry used TA to aid in competing in the marketplace, for improving understanding of the future business environment, and for producing options for internal and external decisionmakers.

TA methodology drew upon a variety of analytical, evaluative and planning techniques. Among these were systems analysis, cost-benefit analysis, consensus methods (e.g., Delphi method), engineering feasibility studies, clinical trials, market research, technological forecasting, and others. TA practitioners and policymakers recognized that TA is evolving, flexible, and should be tailored to the task (US Congress, Office of Technology Assessment 1977).

Early Health Technology Assessment

Health technologies had been studied for safety, effectiveness, cost, and other concerns long before the advent of HTA. Development of TA as a systematic inquiry in the 1960s and 1970s coincided with the introduction of health technologies that prompted widespread public interest in matters that transcended their immediate health effects. Health care technologies were among the topics of early TAs. Multiphasic health screening was one of three topics of "experimental" TAs conducted by the NAE at the request of Congress (National Academy of Engineering 1969). In response to a request by the National Science Foundation to further develop the TA concept in the area of biomedical technologies, the National Research Council conducted TAs on in vitro fertilization, predetermination of the sex of children, retardation of aging, and modifying human behavior by neurosurgical, electrical or pharmaceutical means (National Research Council 1975). The OTA issued a report on drug bioequivalence in 1974, and the OTA Health Program issued its first formal report in 1976.

Since its early years, HTA has been fueled in part by emergence and diffusion of technologies that have evoked social, ethical, legal, and political concerns. Among these technologies are contraceptives, organ transplantation, artificial organs, life-sustaining technologies for critically or terminally ill patients, and, more recently, genetic testing, genetic therapy, and stem cell research. These technologies have challenged certain societal institutions, codes, and other norms regarding fundamental aspects of human life such as parenthood, heredity, birth, bodily sovereignty, freedom and control of human behavior, and death (National Research Council 1975).

Despite the comprehensive approach originally intended for TA, its practitioners recognized early on that “partial TAs” may be preferable in circumstances where selected impacts are of particular interest or where necessitated by resource constraints (US Congress, Office of Technology Assessment 1977). In practice, relatively few TAs have encompassed the full range of possible technological impacts; most focus on certain sets of impacts or concerns, depending upon who conducts the assessment and what data and other resources are available for the assessment. Although there have been important instances of comprehensive HTAs, partial ones have been typical (Goodman 1992).

Box 2 shows various definitions of TA and HTA.

Box 2. Some Definitions of TA and HTA

We shall use the term assessment of a medical technology to denote any process of examining and reporting properties of a medical technology used in health care, such as safety, efficacy, feasibility, and indications for use, cost, and cost-effectiveness, as well as social, economic, and ethical consequences, whether intended or unintended (Institute of Medicine 1985).

Technology assessment (TA) is a category of policy studies, intended to provide decision makers with information about the possible impacts and consequences of a new technology or a significant change in an old technology. It is concerned with both direct and indirect or secondary consequences, both benefits and disbenefits, and with mapping the uncertainties involved in any government or private use or transfer of a technology. TA provides decision makers with an ordered set of analyzed policy options, and an understanding of their implications for the economy, the environment, and the social, political, and legal processes and institutions of society (Coates 1992).

Technology assessment is a form of policy research that examines short- and long-term social consequences (for example, societal, economic, ethical, legal) of the application of technology. The goal of technology assessment is to provide policy-makers with information on policy alternatives (Banta 1993).

Health technology assessment considers the effectiveness, appropriateness and cost of technologies. It does this by asking four fundamental questions: Does the technology work, for whom, at what cost, and how does it compare with alternatives? (UK National Health Service R&D Health Technology Assessment Programme 2003)

Health technology assessment ... is a structured analysis of a health technology, a set of related technologies, or a technology-related issue that is performed for the purpose of providing input to a policy decision (U.S. Congress, Office of Technology Assessment 1994).

An [HTA] describes: about the technology and its use, which technology is clinically effective, for whom, how it compares with current treatments, [and] at what cost (Canadian Coordinating Office for Health Technology Assessment, 2002)

[HTA] is a multidisciplinary field of policy analysis. It studies the medical, social, ethical, and economic implications of development, diffusion, and use of health technology (International Network of Agencies for Health Technology Assessment 2002).

II. FUNDAMENTAL CONCEPTS

Health Technology

Technology is the practical application of knowledge. Three ways to describe health care technology include its material nature, its purpose, and its stage of diffusion.

Material Nature

For many people, the term "technology" connotes "hardware" or other mechanical devices or instrumentation; to others, it is a short form of "information technology" such as computer software. However, the practical application of knowledge in health care is quite broad. Broad categories of health technology include the following.

- *Drugs:* e.g., aspirin, beta-blockers, antibiotics, HMG-CoA reductase inhibitors ("statins")
- *Biologics:* vaccines, blood products, cellular and gene therapies
- *Devices, equipment and supplies:* e.g., cardiac pacemakers, CT scanners, surgical gloves, diagnostic test kits
- *Medical and surgical procedures:* e.g., psychotherapy, nutrition counseling, coronary angiography, gall bladder removal
- *Support systems:* e.g., electronic patient record systems, telemedicine systems, drug formularies, blood banks, clinical laboratories
- *Organizational and managerial systems:* e.g., prospective payment using diagnosis-related groups, alternative health care delivery configurations, clinical pathways, total quality management programs

Purpose or Application

Technologies can also be grouped according to their health care purpose, i.e.:

- *Prevention*: protect against disease by preventing it from occurring, reducing the risk of its occurrence, or limiting its extent or sequelae (e.g., immunization, hospital infection control program, fluoridated water supply)
- *Screening*: detect a disease, abnormality, or associated risk factors in asymptomatic people (e.g., Pap smear, tuberculin test, mammography, serum cholesterol testing)
- *Diagnosis*: identify the cause and nature or extent of disease in a person with clinical signs or symptoms (e.g., electrocardiogram, serological test for typhoid, x-ray for possible broken bone)
- *Treatment*: designed to improve or maintain health status, avoid further deterioration, or provide palliation (e.g., antiviral therapy, coronary artery bypass graft surgery, psychotherapy, drugs for cancer pain)
- *Rehabilitation*: restore, maintain or improve a physically or mentally disabled person's function and well-being (e.g., exercise program for post-stroke patients, assistive device for severe speech impairment, incontinence aid)

Not all technologies fall neatly into single categories. Many tests and other technologies used for diagnosis also are used for screening. (The probability that a patient has a disease or other health condition is greatly affected by whether these technologies are used for screening asymptomatic patients or diagnosing symptomatic patients.) Some technologies are used for diagnosis as well as treatment, e.g., coronary angiography to diagnose heart disease and to monitor coronary angioplasty. Implantable cardioverter defibrillators detect potentially life-threatening heart arrhythmias and deliver electrical pulses to restore normal heart rhythm. Electronic patient record systems can support all of these technological purposes or applications.

Certain "boundary-crossing" or "hybrid" technologies combine characteristics of drugs, devices or other major categories of technology (Goodman 1993; Lewin Group 2001). Among the many examples of these are: photodynamic therapy, in which drugs are laser-activated (e.g., for targeted destruction of cancer cells); local drug delivery technologies (e.g., implantable drug pumps and drug inhalers); spermicidal condoms; and bioartificial organs that combine natural tissues and artificial components. Examples of hybrid technologies that have complicated regulatory approval and coverage decisions in recent years are gallstone lithotripters (used with stone-dissolving drugs) (Zeman 1991), positron-emission tomography (PET, used with radiopharmaceuticals) (Coleman 1992), and metered-dose inhalers (Massa 2002).

Stage of Diffusion

Technologies may be assessed at different stages of diffusion and maturity. In general, health care technologies may be described as being:

- *Future:* in a conceptual stage, anticipated, or in the earliest stages of development
- *Experimental:* undergoing bench or laboratory testing using animals or other models
- *Investigational:* undergoing initial clinical (i.e., in humans) evaluation for a particular condition or indication
- *Established:* considered by providers to be a standard approach to a particular condition or indication and diffused into general use
- *Obsolete/outmoded/abandoned:* superseded by other technologies or demonstrated to be ineffective or harmful

Often, these stages are not clearly delineated, and technologies do not necessarily mature through them in a linear fashion. A technology may be investigational for certain indications, established for others, and outmoded or abandoned for still others, such as autologous bone marrow transplantation with high-dose chemotherapy for certain types of advanced cancers. Many technologies undergo multiple incremental innovations after their initial acceptance into general practice (Gelijns and Rosenberg 1994; Reiser 1994). Further, a technology that was once considered obsolete may return to established use for a better defined or entirely different clinical purpose. A prominent example is thalidomide, whose use as a sedative during pregnancy was halted more than 40 years ago when it was found to induce severe fetal malformation, but which is now used to treat such conditions as leprosy, advanced multiple myeloma, chronic graft vs. host disease, and certain complications of HIV infection (Baidas 2002).

Health Technology Assessment

Health technology assessment is the systematic evaluation of properties, effects or other impacts of health technology. The main purpose of HTA is to inform policymaking for technology in health care, where policymaking is used in the broad sense to include decisions made at, e.g., the individual or patient level, the level of the health care provider or institution, or at the regional, national and international levels. HTA may address the direct and intended consequences of technologies as well as their indirect and unintended consequences. HTA is conducted by interdisciplinary groups using explicit analytical frameworks, drawing from a variety of methods.

Purposes of HTA

HTA can be used in many ways to advise or inform technology-related policymaking. Among these are to advise or inform:

- Regulatory agencies such as the US Food and Drug Administration (FDA) about whether to permit the commercial use (e.g., marketing) of a drug, device or other technology
- Health care payers, providers, and employers about whether technologies should be included in health benefits plans or disease management programs, addressing coverage (whether or not to pay) and reimbursement (how much to pay)
- Clinicians and patients about the appropriate use of health care interventions for a particular patient's clinical needs and circumstances
- Health professional associations about the role of a technology in clinical protocols or practice guidelines
- Hospitals, health care networks, group purchasing organizations, and other health care organizations about decisions regarding technology acquisition and management
- Standards-setting organizations for health technology and health care delivery regarding the manufacture, use, quality of care, and other aspects of health care technologies
- Government health department officials about undertaking public health programs (e.g., vaccination, screening, and environmental protection programs)
- Lawmakers and other political leaders about policies concerning technological innovation, research and development, regulation, payment and delivery of health care
- Health care product companies about product development and marketing decisions
- Investors and companies concerning venture capital funding, acquisitions and divestitures, and other transactions concerning health care product and service companies

HTA contributes in many ways to the knowledge base for improving the quality of health care, especially to support development and updating of a wide spectrum of standards, guidelines, and other health care policies. For example, the Joint Commission on Accreditation of Healthcare Organizations (JCAHO) and the National Committee for Quality Assurance (NCQA) set standards for measuring quality of care and services of hospitals, managed care organizations, long-term care facilities, hospices, ambulatory care centers, and other health care institutions. Health professional associations (e.g., American College of Cardiology, American College of Physicians, American College of Radiology,

American Medical Association) and special panels (e.g., the US Preventive Services Task Force of the Agency for Healthcare Research and Quality) develop **clinical practice guidelines,** standards, and other statements regarding the appropriate use of technologies. Standards-setting organizations such as the American National Standards Institute and the American Society for Testing and Materials coordinate development of voluntary national consensus standards for the manufacture, use and reuse of health devices and their materials and components.

As noted above, HTA can be used to support decisionmaking by clinicians and patients. The term **evidence-based medicine** refers to the use of current best evidence from scientific and medical research, and the application of clinical experience and observation, in making decisions about the care of individual patients. This has prompted the appearance of many useful resources, including:

- *Evidence-Based Medicine* (Sackett 1997), a guide to the field
- *Evidence-Based Medicine* (a joint product of the American College of Physicians and the BMJ Publishing Group), a journal digest of articles selected from international medical journals
- "Users' guides to the medical literature," a series of more than 30 articles by the Evidence-Based Medicine Working Group published in the *Journal of the American Medical Association,* ranging from (Oxman 1993) to (Guyatt 2000)
- Centre for Evidence-Based Medicine [www.cebm.net/]

Basic HTA Orientations

The impetus for an HTA is not necessarily a technology. Three basic orientations to HTA are as follows.

- *Technology-oriented* assessments are intended to determine the characteristics or impacts of particular technologies. For example, a government agency may want to determine the clinical, economic, social, professional, or industrial impacts of population-based cancer screening, cochlear implants, or other particular interventions.
- *Problem-oriented* assessments focus on solutions or strategies for managing a particular problem for which alternative or complementary technologies might be used. For example, clinicians and providers concerned with the problem of diagnosis of dementia may call for the development of clinical practice guidelines involving some combination or sequence of clinical history, neurological examination, and diagnostic imaging using various modalities..
- *Project-oriented* assessments focus on a local placement or use of a technology in a particular institution, program, or other designated project. For example, this may arise when a hospital must decide whether or not to purchase a magnetic resonance imaging (MRI) unit, considering the facilities, personnel, and other resources needed to install and operate an MRI unit; the hospital's financial status; local market potential for MRI services; competitive factors; etc.

These basic assessment orientations can overlap and complement one another. Certainly, all three types could draw upon a common body of scientific evidence and other information. A technology-oriented assessment may address the range of problems for which the technology might be used and how appropriate the technology might be for different types of local settings (e.g., inpatient versus outpatient). A problem-oriented assessment, examining the effects or other impacts of alternative technologies on a given problem, may incorporate multiple, focused (i.e., on the problem at hand) technology-oriented assessments. A project-oriented assessment would consider the range of impacts of a technology or its alternatives in a given setting, as well as the role or usefulness of that technology for

various problems. Although the information used in a project-oriented assessment by a particular hospital may include findings of pertinent technology- and problem-oriented assessments, local data collection and analysis may be required to determine what is sensible for that hospital. Thus, many HTAs will blend aspects of all three basic orientations.

Timing of Assessment

There is no single correct time to conduct an HTA. It is conducted to meet the needs of a variety of policymakers seeking assessment information throughout the lifecycles of technologies. Investors, regulators, payers, hospital managers and others tend to make decisions about technologies at particular junctures, and each may subsequently reassess technologies. Indeed, the determination of a technology's stage of diffusion may be the primary purpose of an assessment. For insurers and other payers, technologies that are deemed experimental or investigational are usually excluded from coverage, whereas those that are established or generally accepted are usually eligible for coverage (Newcomer 1990; Reiser 1994; Singer 2001).

There are tradeoffs inherent in decisions regarding the timing for HTA. On one hand, the earlier a technology is assessed, the more likely its diffusion can be curtailed if it is unsafe or ineffective (McKinlay 1981). From centuries' old purging and bloodletting to the more contemporary autologous bone marrow transplantation with high-dose chemotherapy for advanced breast cancer, the list of poorly evaluated technologies that diffused into general practice before being found to be ineffective and/or harmful continues to grow. **Box 3** shows examples of health care technologies found to be ineffective or harmful after being widely diffused.

On the other hand, to regard the findings of an early assessment as definitive or final may be misleading. An investigational technology may not yet be perfected; its users may not yet be proficient; its costs may not yet have stabilized; it may not have been applied in enough circumstances to recognize its potential benefits; and its long-term outcomes may not yet be known (Mowatt 1997). As one technology assessor concluded about the problems of when-to-assess: "It's always too early until, unfortunately, it's suddenly too late!" (Buxton 1987). Further, the "**moving target problem**" can complicate HTA (Goodman 1996). By the time a HTA is conducted, reviewed and disseminated, its findings may be outdated by changes in a technology, in how it is used, or in its technological alternatives for a given problem.

Some payers provide conditional coverage for selected investigational technologies in order to compile evidence on safety, effectiveness, cost, etc., for making more informed coverage policies. In these instances, payers cover the use of a technology only under certain conditions, such as where patients are enrolled in an RCT at certain participating medical centers. This arrangement offers a way to balance the need for evidence with the demand for access and financially compensated care. Depending on the type of technology involved, it further enables refinement of technique or delivery, and building of experience and expertise among physicians and other providers (Beebe 1997; Brenner 2002; McGivney 1992; Sheingold 1998; Medical Technology Leadership Forum 1999; Wood 2001).

Box 3. Examples of Health Care Technologies Found to be Ineffective or Harmful After Being Widely Diffused

- Autologous bone marrow transplant with high-dose chemotherapy for advanced breast cancer
- Colectomy to treat epilepsy
- Diethylstilbestrol (DES) to improve pregnancy outcomes
- Electronic fetal monitoring during labor without access to fetal scalp sampling
- Episiotomy (routine or liberal) for birth
- Extracranial-intracranial bypass to reduce risk of ischemic stroke
- Gastric bubble for morbid obesity
- Gastric freezing for peptic ulcer disease
- Hormone replacement therapy for healthy menopausal women
- Hydralazine for chronic heart failure
  Intermittent positive pressure breathing
- Mammary artery ligation for coronary artery disease
- Optic nerve decompression surgery for nonarteritic anterior ischemic optic neuropathy
- Quinidine for suppressing recurrences of atrial fibrillation
  Radiation therapy for acne
- Sleeping face down for healthy babies
- Supplemental oxygen for healthy premature babies
- Thalidomide for sedation in pregnant women
- Thymic irradiation in healthy children
- Triparanol (MER-29) for cholesterol reduction

Sources: Coplen 1990; Enkin 1995; Feeny 1986; Fletcher 2002; Grimes 1993; Mello 2001; The Ischemic Optic Neuropathy Decompression Trial Research Group 1995; Passamani 1991; Rossouw 2002; US DHHS 1990, 1993; others.

Despite the value of conditional coverage in principle, some observers have raised practical and ethical concerns about their implementation. Among these are that: (1) if conditional coverage is initiated after a technology has diffused, some patients who had expected to get a procedure may be denied it if they are not enrolled in a trial; (2) some patients who would be interested in enrolling in a covered trial are not located near a participating center and are therefore denied access; (3) patients and physicians who believe in the effectiveness of the technology may be unwilling to be involved in an RCT, including some who decide to finance the technology outside of the trial and therefore diminish enrollment, (4) the indications for using the technology in the conditional coverage trial may be too broad or too narrow to properly reflect the potential safety and effectiveness of the technology; and (5) the technology continues to evolve during the conditional coverage process, to the point where the trial findings are of diminished relevance (Berger 2001; Cooper 2001).

Properties and Impacts Assessed

What is assessed in HTA? HTA may involve the investigation of one or more properties, impacts, or other attributes of health technologies or applications. In general, these include the following.

- Technical properties
- Safety
- Efficacy and/or effectiveness
- Economic attributes or impacts
- Social, legal, ethical and/or political impacts

Technical properties include performance characteristics and conformity with specifications for design, composition, manufacturing, tolerances, reliability, ease of use, maintenance, etc. **Safety** is a judgment of the acceptability of risk (a measure of the probability of an adverse outcome and its severity) associated with using a technology in a given situation, e.g., for a patient with a particular health problem, by a clinician with certain training, and/or in a specified treatment setting.

Efficacy and effectiveness both refer to how well a technology works to improve patient health, usually based on changes in one or more pertinent health outcomes or "endpoints" as described below. A technology that works under carefully controlled conditions or with carefully selected patients under the supervision of its developers does not always work as well in other settings or as implemented by other practitioners. In HTA, **efficacy** refers to the benefit of using a technology for a particular problem under ideal conditions, e.g., within the protocol of a carefully managed **randomized controlled trial**, involving patients meeting narrowly defined criteria, or conducted at a "center of excellence." **Effectiveness** refers to the benefit of using a technology for a particular problem under general or routine conditions, e.g., by a physician in a community hospital for a variety of types of patients.

Clinicians, patients, managers and policymakers are increasingly aware of the practical implications of differences in efficacy and effectiveness. Researchers delve into **registers**, **databases** (e.g., of third-party payment claims and administrative data) and other epidemiological and observational data to discern possible associations between the use of technologies and patient outcomes in general or routine practice settings. The **validity** of any findings regarding causal connections between interventions and patient outcomes may be weakened to the extent that these data are not derived from prospective, randomized, controlled studies (US Congress, OTA 1994). As discussed below, some newer prospective trials are designed to incorporate varied groups of patients and settings.

Box 4 shows certain distinctions in efficacy and effectiveness for diagnostic tests. Whereas the relationship between a preventive, therapeutic, or rehabilitative technology and patient outcomes is typically direct (though not always easy to measure), the relationship between a technology used for diagnosis or screening and its patient outcomes is typically indirect. Also, diagnostic and screening procedures can have their own short-term and long-term adverse health effects, e.g., biopsies and certain radiological procedures.

Health technologies can have a wide range of microeconomic and macroeconomic attributes or impacts. Microeconomic concerns include costs, prices, charges, and payment levels associated with individual

technologies. Other concerns include comparisons of resource requirements and outcomes (or benefits) of technologies for particular applications, such as cost effectiveness, cost utility, and cost benefit. (Methods for determining these are described below.)

Box 4. Efficacy vs. Effectiveness for Diagnostic Tests

	Efficacy	**Effectiveness**
Patient Population	Homogeneous; patients with coexisting illness often excluded	Heterogeneous; includes all patients who usually have test
Procedures	Standardized	Often variable
Testing Conditions	Ideal	Conditions of everyday practice
Practitioner	Experts	All users

Adapted from: Institute of Medicine 1989.

Examples of macroeconomic impacts of health technologies are the impact of new technologies on: national health care costs, resource allocation among different health programs or among health and other sectors, and shifts in the site of care, such as from inpatient to outpatient settings. Other macroeconomic issues that pertain to health technologies include the effects of intellectual property policies (e.g., for patent protection), regulation, third-party payment, and other policy changes on technological innovation, investment, competitiveness, technology transfer, and employment.

A variety of technologies raise social and ethical concerns. Such technologies as genetic testing, use of stem cells to grow new tissues, allocation of scarce organs for transplantation, and life-support systems for the critically ill challenge certain legal standards and societal norms. For example, the small and slowly increasing supply of donated kidneys, livers, hearts, and other organs for transplantation continues to fall behind the rapidly expanding need for them, raising ethical, social, and political concerns about allocation of scarce, life-saving resources (Miranda 1998; Yoshida 1998). In dialysis and transplantation for patients with end-stage renal disease, ethical concerns arise from patient selection criteria, termination of treatment, and managing non-compliant and other problem patients (Rettig 1991).

Ethical questions continue to prompt improvement in informed consent procedures for patients involved in clinical trials. Allocation of scarce resources to technologies that are expensive, inequitably used, or non-curative raises broad social concerns (Gibson 2002). Ethical considerations arise in HTA in the form of normative concepts (e.g., valuation of human life); applications of technology (prevention, screening, diagnosis, therapy, etc.); research and the advancement of knowledge; allocation of resources; and the integrity of HTA processes themselves (Heitman 1998). Methods for assessing ethical and social implications of health technology remain relatively underdeveloped, and the means of translating these implications into policy are often unclear (Van der Wilt 2000). Even so, greater efforts are being made to involve different perspectives in the HTA process in order to better account for

identification of the types of effects or impacts that should be assessed, and for values assigned by these different perspectives to life, quality of life, privacy, choice of care, and other matters (Reuzel 2001).

The terms "appropriate" and "necessary" often are used to describe whether or not a technology should be used in particular circumstances. For example, the appropriateness of a diagnostic test may depend on its safety and effectiveness compared to alternative available interventions for particular patient indications, clinical settings, and resource constraints. A technology may be considered necessary if withholding it would be deleterious to the patient's health (Hilborne 1991; Kahan 1994; Singer 2001).

The properties, impacts, and other attributes assessed in HTA pertain across the wide range of types of technology. Thus, for example, just as drugs, devices, and surgical procedures can be assessed for safety, effectiveness, and cost effectiveness, so can hospital infection control programs, computer-based drug-utilization review systems, and rural telemedicine networks.

Measuring Health Outcomes

Health outcome variables are used to measure the safety, efficacy and effectiveness of health care technologies. Health outcomes have been measured primarily in terms of changes in mortality (death rate) or morbidity (disease rate). For a cancer treatment, the main outcome of interest may be five-year survival; for treatments of coronary artery disease, the main endpoints may be **incidence** of fatal and nonfatal acute myocardial infarction and recurrence of angina. Increasingly, health outcomes are being measured in the form of health-related quality of life and functional status.

In a clinical trial comparing alternative treatments, the effect on health outcomes of one treatment relative to another (e.g., a control treatment) can be expressed using various measures of **treatment effect**. These measures compare the probability of a given health outcome in the treatment group with the probability of the same outcome in a **control group.** Examples are **absolute risk reduction, odds ratio, number needed to treat,** and **effect size. Box 5** shows how choice of treatment effect measures can give different impressions of study results.

Health-Related Quality of Life Measures

Although mortality and morbidity are usually the outcomes of greatest concern, they are not the only outcomes of importance to patients nor to others. Many technologies affect patients, family members, providers, employers, and other interested parties in ways that are not reflected in mortality or morbidity rates; this is particularly true for many chronic diseases.

Health-related quality of life (HRQL) measures (or indexes) are increasingly used along with more traditional outcome measures to assess health care technologies, providing a more complete picture of the ways in which health care affects patients. HRQL measures capture such dimensions as: physical function, social function, cognitive function, anxiety/distress, bodily pain, sleep/rest, energy/fatigue and general health perception. HRQL measures may be disease-specific (e.g., heart disease or arthritis) or general (covering overall health). They may be one-dimensional (concerning one aspect such as distress) or multidimensional (Patrick and Deyo 1989). They may provide a single aggregate score or yield a set of scores, each for a particular dimension. HRQL measures are increasingly used by health product companies to differentiate their products from those of competitors, which may have virtually

indistinguishable effects on morbidity for particular diseases (e.g., hypertension and depression) but may have different profiles of side effects that affect patients' quality of life (Gregorian 2003).

Box 5. Choice of Treatment Effect Measures Can Give Different Impressions

A study of the effect of breast cancer screening can be used to contrast several treatment effect measures and to show how they can give different impressions about the effectiveness of an intervention (Forrow 1992). In 1988, Andersson (1988) reported the results of a large RCT that was conducted to determine the effect of mammographic screening on mortality from breast cancer. The trial involved more than 42,000 women who were over 45 years old. Half of the women were invited to have mammographic screening and were treated as needed. The other women (control group) were not invited for screening.

The report of this trial states that "Overall, women in the study group aged ≥55 had a 20% reduction in mortality from breast cancer." Although this statement is true, calculation of other types of treatment effect measures provides important additional information. The table below shows the number of women aged ≥55 and breast cancer deaths in the screened group and control group, respectively. Based on these figures, four treatment effect measures are calculated.

For example, *absolute risk reduction* is the difference in the rate of adverse events between the screened group and the control group. In this trial, the absolute risk reduction of 0.0007 means that the absolute effect of screening was to reduce the incidence of breast cancer mortality by 7 deaths per 10,000 women screened, or 0.07%.

Group	Number of Patients	Deaths from breast cancer	Probability of death from breast cancer	Absolute risk reduction[1]	Relative risk reduction[2]	Odds ratio[3]	Number needed to screen[4]
Screened	13,107	35	$P_C = 0.0027$	0.0007	21%	0.79	1,429
Control	13,113	44	$P_C = 0.0034$				

Women in the intervention group were invited to attend mammographic screening at intervals of 18-24 months. Five rounds of screening were completed. Breast cancer was treated according to stage at diagnosis. Mean follow-up was 8.8 years.

1. Absolute risk reduction: $P_C - P_S$
2. Relative risk reduction: $(P_C - P_S) \div P_C$
3. Odds ratio: $[P_S \div (1 - P_S)] \div [P_C \div (1 - P_C)]$
4. Number needed to screen: $1 \div (P_C - P_S)$

Source of number of patients and deaths from breast cancer: Andersson 1988.

HRQL measures can be used to determine the effects of a technology on patients, to compare alternative technologies for their effects on patients with a particular problem or disability, or to compare different

technologies' respective abilities to improve the quality of life of patients with different problems. Reflecting, in part, the need to demonstrate the effectiveness of many new technologies for chronic conditions such as rheumatoid arthritis, migraine, and depression, considerable advances have been made in the development and validation of these measures in the last 25 years. **Box 6** shows dimensions of general HRQL measures that have been used extensively and that are well validated for certain applications. **Box 7** shows aspects of selected disease-specific HRQL measures.

Quality-Adjusted Life Years

A unit of health care outcome that combines gains (or losses) in length of life with quality of life is the **quality-adjusted life year (QALY)**. QALYs represent years of life subsequent to a health care intervention that are weighted or adjusted for the quality of life experienced by the patient during those years (Torrance and Feeny 1989). QALYs provide a common unit for multiple purposes, including: estimating the overall burden of disease; comparing the relative impact of specific diseases, conditions, and health care interventions; and making economic comparisons, such as of the cost-effectiveness (in particular the cost-utility) of different health care interventions. Health economists have proposed setting priorities among alternative health care interventions by selecting among these so as to maximize the additional health gain in terms of QALYs. This is intended to optimize allocation of scarce resources and thereby maximize social welfare. Other units that are analogous to QALYs include **disability-adjusted life years (DALYs)** and **healthy-years equivalents** (HYEs). As a group, these types of measures are sometimes known as health-adjusted life years (HALYs) (Gold 2002; Johannesson 1993; Mehrez and Gafni 1993; World Development Report 1993).

The scale of quality of life used for QALYs can be based on general HRQL indexes or other methods of eliciting patient **utility** for certain states of life. This dimension is typically standardized to a scale ranging from 0.0 (death) to 1.0 (perfect health). A scale may allow for ratings below 0.0 for states of disability and distress that some patients consider to be worse than death (Patrick 1994). QALYs can be useful for making comparisons among alternative technologies because they are generic units that can reflect changes brought about by different health care interventions for the same or different health problems. **Box 8** shows how QALYs were used to compare the cost utilities of three alternative therapies for end-stage heart disease. **Box 9** lists the cost utility of different interventions for different health problems according to the amount of money that must be invested per QALY gained. The *CEA Registry* is a continuously updated, detailed set of standardized cost-utility analyses, including tables of cost-utility ratios for many types of health care interventions [www.hsph.harvard.edu/cearegistry].

Certain methodological aspects and the proposed use of QALYs or similar units in setting health care priorities remain controversial (Arnesen 2000; Gerard and Mooney 1993; Mason 1994; Nord 1994; Richardson 1994; Ubel 2000). Research on public perceptions of the value of health care programs indicates that health gain is not necessarily the only determinant of value, and that the rule of maximizing QALYs (or similar measures) per health expenditure to set priorities may be too restrictive, not reflecting public expectations regarding fairness or equity. For example, because people who are elderly or disabled may have a lower "ceiling" or potential for gain in QALYs or other measure of HRQL than other people would for the same health care expenditure, making resource allocation decisions based on cost-utility is viewed by some as inherently biased against the elderly and disabled.

Box 6. Domains of Selected General Health-Related Quality of Life Indexes

Sickness Impact Profile (Bergner 1981; de Bruin 1992)

• Body care and movement	• Emotional behavior
• Ambulation	• Alertness behavior
• Mobility	• Communication
• Sleep and rest	• Social interaction
• Home management	• Work
• Recreation and pastimes	• Eating

Nottingham Health Profile (Doll 1993; Jenkinson 1988)

• Physical mobility	• Energy
• Pain	• Social isolation
• Sleep	• Emotional reactions

Quality of Well-Being Scale (Kaplan 1988; Kaplan 1989)

• Symptom-problem complex	• Physical activity
• Mobility	• Social activity

Functional Independence Measure (Bunch 1994; Linacre 1994)

• Self-care	• Locomotion
• Sphincter control	• Communication
• Mobility	• Social cognition

Short Form (SF)-36 (McHorney 1994; Ware 1992)

• Physical functioning	• General mental health
• Role limitations due to physical problems	• Role limitations due to emotional problems
• Social functioning	• Vitality
• Bodily pain	• General health perceptions

EuroQol Descriptive System (Essink-Bot 1993; EuroQol Group 1990)

• Mobility	• Pain/discomfort
• Self-care	• Anxiety/depression
• Usual activities	

Katz Activities of Daily Living (Katz 1970; Lazaridis 1994)

• Bathing	• Mobility
• Dressing	• Continence
• Toileting	• Eating

Box 7. Domains of Selected Disease-Specific Health-Related Quality of Life Indexes

New York Heart Association Functional Classification (O'Brien 1993; van den Broek 1992)

Class I: Patients with cardiac disease but without resulting limitations of physical activity. Ordinary physical activity does not cause undue fatigue, palpitation, dyspnoea, or anginal pain.

Class II: Patients with cardiac disease resulting in slight limitation of physical activity. They are comfortable at rest. Ordinary physical activity results in fatigue, palpitation, dyspnoea or anginal pain.

Class III: Patients with cardiac disease resulting in marked limitation of physical activity. They are comfortable at rest. Less than ordinary physical activity causes fatigue, palpitation, dyspnoea or anginal pain.

Class IV: Patients with cardiac disease resulting in inability to carry on any physical activity without discomfort. Symptoms of cardiac insufficiency or of anginal syndrome may be present even at rest. If any physical activity is undertaken, discomfort is increased.

Arthritis Impact Measurement Scales (Kazis 1989; Meenan 1992)

• Mobility	• Social activities
• Walking and bending	• Support from family and friends
• Hand and finger function	• Arthritis pain
• Arm function	• Work
• Self care	• Level of tension
• Household tasks	• Mood

Visual Functioning (VF)-14 Index (Steinberg 1994)

- reading small print, such as labels on medicine bottles, a telephone book, or food labels
- reading a newspaper or book
- reading a large-print book or newspaper or the numbers on a telephone
- recognizing people when they are close to you
- seeing steps, stairs, or curbs
- reading traffic, street, or store signs
- doing fine handwork such as sewing, knitting, crocheting, or carpentry
- writing checks or filling out forms
- playing games such as bingo, dominos, card games, or mahjong
- taking part in sports such as bowling, handball, tennis, or golf
- cooking
- watching television
- daytime driving
- nighttime driving

Box 8. Cost-Utilities for Alternative Therapies for End-Stage Heart Disease

Therapy	Life years gained (yr)	Mean utility	QALY (yr)	Aggregate cost ($)	Cost per QALY ($/yr)
Conventional medical treatment	0.50	0.06	0.03	$28,500	$950,000
Heart transplantation	11.30	0.75	8.45	$298,200	$35,290
Total artificial heart	4.42	0.65	2.88	$327,600	$113,750

Notes: Costs and outcomes discounted at three percent per year; 20-year horizon. Mean utilities derived using time-tradeoff method on scale for which 1.0 was well, 0.0 was death, and states worse than death were valued between 0.0 and -1.0.

This table indicates that, although the cost of conventional medical treatment is the lowest, its cost per QALY is the highest, as the life-years gained and the patient utility of those years are low compared to the alternatives. The costs of heart transplantation and total artificial heart are of similar magnitude, but the cost per QALY is much lower for heart transplantation, as the life-years gained and the patient utility of those years are higher compared to the total artificial heart.

Source: Hogness 1991.

Box 9. Cost per QALY for Selected Health Care Technologies

	Cost per QALY (£ 1990)
Cholesterol testing and diet therapy (all 40-69 yrs)	220
Neurosurgery for head injury	240
General practitioner advice to stop smoking	270
Neurosurgery for subarachnoid hemorrhage	490
Antihypertensive therapy to prevent stroke (45-64 yrs)	940
Pacemaker implantation	1,100
Hip replacement	1,180
Valve replacement for aortic stenosis	1,140
Cholesterol testing and treatment	1,480
Coronary artery bypass graft surgery (left main disease, severe angina)	2,090
Kidney transplant	4,710
Breast cancer screening	5,780
Heart transplantation	7,840
Cholesterol testing and treatment (incremental) (all 25-39 yrs)	14,150
Home hemodialysis	17,260
Coronary artery bypass graft surgery (one-vessel disease, moderate angina)	18,830
Continuous ambulatory peritoneal dialysis	19,870
Hospital hemodialysis	21,970
Erythropoietin for dialysis anemia (with 10% reduction in mortality)	54,380
Neurosurgery for malignant intracranial tumors	107,780
Erythropoietin for dialysis anemia (with no increase in survival)	126,290

This table ranks selected procedures for a variety of health problems according to their cost utility, (i.e., the amount of money that must be spent on each procedure to gain one more QALY). There were some methodological differences in determining costs and QALYs among the studies from which these results were derived. Nonetheless, giving considerable latitude to these figures, the range in the magnitude of investment required to yield the next QALY for these treatments is great. This type of "bucks for the bang" (here, British pounds for the QALY) analysis helps to illustrate implicit choices made in allocating scarce health care resources, and suggests how decision makers might move toward reallocating those resources if societal gain in net health benefits (e.g., as measured using QALYs) is used as an allocation criterion.

Source: Maynard 1991.

Some work has been done recently to capture more dimensions of public preference and to better account for the value attributed to different health care interventions (Dolan 2001; Schwappach 2002). HRQL measures and QALYs continue to be used in HTA while substantial work continues in reviewing, refining and validating them.

Performance of Diagnostic Technologies

The relationships between most preventive, therapeutic, and rehabilitative technologies and health outcomes can be assessed as direct cause and effect relationships. The relationship between the use of diagnostic and screening technologies and health outcomes is typically indirect, as these technologies provide information that may be used to inform providers concerning the use of interventions that may in turn affect health outcomes.

Many tests and other technologies used for diagnosis are also used for screening, and most of the concepts discussed here for diagnostic technologies pertain as well to screening technologies. A basic difference between screening and diagnosis is that diagnosis is done in symptomatic patients and screening is typically done in asymptomatic patient groups. For a given test used for either screening or diagnosis, this difference has a great effect on the probability that a patient has a disease or other health condition.

The immediate purpose of a diagnostic test is to provide information about the presence (and, less often, the extent) of a disease or other health condition. That is, the diagnostic test should be able to discriminate between patients who have a particular disease and those who do not have the disease (or discriminate among different extents of disease in a given patient).

The technical performance of a diagnostic test depends on a number of factors. Among these are the **precision** and **accuracy** of the test, the observer variation in reading the test data, and the relationship between the disease of interest and the cutoff level of the marker or surrogate used in the diagnostic test to determine the presence or absence of that disease. These factors contribute to the ability of a diagnostic test to detect a disease when it is present and to not detect a disease when it is not present.

The marker for a disease or condition is typically defined as a certain cutoff level of a variable such as blood pressure (e.g., for hypertension), glucose level (e.g., for diabetes), or prostate specific antigen level (e.g., for prostate cancer). Disease markers have distributions in non-diseased as well as in diseased populations. For most diseases, these distributions overlap, so that a single cutoff level does not clearly separate non-diseased from diseased people. For instance, in the case of hypertension, a usual marker for the disease is diastolic blood pressure, the cutoff level of which is often set at 95mm Hg. In fact, some people whose diastolic blood pressure is above 95mm will not be hypertensive (false positive result), and some people with diastolic blood pressure below 95mm will be hypertensive (false negative result). Lowering the cutoff to 90mm will decrease the number of false positives, but increase the number of false negatives.

A diagnostic test can have four basic types of outcomes, as shown in **Box 10.** A **true positive** diagnostic test result is one that detects a marker when the disease is present. A **true negative** test result is one that does not detect the marker when the disease is absent. A **false positive** test result is one that detects a marker when the disease is absent. A **false negative** test result is one that does not detect a marker when the disease is present.

Box 10. Possible Outcomes of Diagnostic Tests

Test Result	Disease Status	
	Present	Absent
Positive (+)	True +	False +
Negative (-)	False -	True -

Operating characteristics of diagnostic tests and procedures are measures of the technical performance of these technologies. These characteristics are based on the probabilities of the four possible types of outcomes of a diagnostic test. The two most commonly used operating characteristics of diagnostic tests are sensitivity and specificity. **Sensitivity** measures the ability of a test to detect disease when it is present. **Specificity** measures the ability of a test to correctly exclude disease in a non-diseased person. One graphical way of depicting these operating characteristics for a given diagnostic test is with a **receiver operating characteristic (ROC) curve**, which plots the relationship between the true positive ratio (sensitivity) and false positive ratio (1 - specificity) as a function of the cutoff level of a disease (or condition) marker. ROC curves help to demonstrate how raising or lowering the cutoff point for defining a positive test result affects tradeoffs between correctly identifying people with a disease (true positives) and incorrectly labeling a person as positive who does not have the condition (false positives).

Taken alone, sensitivity and specificity do not reveal the probability that a given patient really has a disease if the test is positive, or the probability that a given patient does not have the disease if the test is negative. These probabilities are captured by two other operating characteristics. **Predictive value positive** is the proportion of those patients with a positive test result who actually have the disease. **Predictive value negative** is the proportion of patients with a negative test result who actually do not have the disease. (See **Box 11.**) Unlike sensitivity and specificity, predictive value positive and predictive value negative are not constant performance characteristics of a diagnostic test; they change with the prevalence of the disease in the population of interest. For example, if a disease is very rare in the population, even tests with high sensitivity and high specificity can have low predictive value positive, generating more false-positive than false negative results.

Beyond technical performance of diagnostic technologies, the effect of diagnostic technologies on health outcomes or health-related quality of life is less obvious than for other types of technologies. As health care decisionmakers increasingly demand to know how health care interventions affect health care outcomes, diagnostic technologies will have to demonstrate their efficacy/effectiveness accordingly.

The efficacy (or effectiveness) of a diagnostic technology can be determined along a chain of inquiry that leads from technical capacity of a technology to changes in patient health outcomes to cost effectiveness, as follows.

1. *Technical capacity*. Does the technology perform reliably and deliver accurate information?
2. *Diagnostic accuracy*. Does the technology contribute to making an accurate diagnosis?

Box 11. Operating Characteristics of Diagnostic Tests

Characteristic	Formula	Definition
Sensitivity	$\frac{\text{True Positives}}{\text{True positives + False negatives}}$	Proportion of people with condition who test positive
Specificity	$\frac{\text{True Negatives}}{\text{True negatives + False positives}}$	Proportion of people without condition who test negative
Predictive value positive	$\frac{\text{True Positives}}{\text{True positives + False positives}}$	Proportion of people with positive test who have condition
Predictive value negative	$\frac{\text{True Negatives}}{\text{True negatives + False negatives}}$	Proportion of people with negative test who do not have condition

3. *Diagnostic impact*. Do the diagnostic results influence use of other diagnostic technologies, e.g., does it replace other diagnostic technologies?
4. *Therapeutic impact*. Do the diagnostic findings influence the selection and delivery of treatment?
5. *Patient outcome*. Does use of the diagnostic technology contribute to improved health of the patient?
6. *Cost effectiveness*. Does use of the diagnostic technology improve the cost effectiveness of health care compared to alternative interventions?

If a diagnostic technology is not efficacious at any step along this chain, then it is not likely to be efficacious at any later step. Efficacy at a given step does not imply efficacy at a later step (Feeny 1986; Fineberg 1977; Institute of Medicine 1985). **Box 12** shows a hierarchy of studies for assessing diagnostic imaging technologies that is consistent with the chain of inquiry noted above. Some groups have developed standards for reporting studies of the accuracy of diagnostic tests (Bossuyt 2003).

For diagnostic technologies that are still prototypes or in other early stages of development, there are limited data upon which to base answers to questions such as these. Even so, investigators and advocates of diagnostic technologies should be prepared to describe, at least qualitatively, the ways in which the technology might affect diagnostic accuracy, diagnostic impact, therapeutic impact, patient outcomes and cost effectiveness; how these effects might be measured; approximately what levels of performance would be needed to successfully implement the technology; and how further investigations should be conducted to make these determinations.

Types of Organizations That Conduct HTA

The types of organizations that undertake some form of HTA include:

- Regulatory agencies
- Government and private sector payers
- Managed care organizations

Box 12. Hierarchical Model of Efficacy for Diagnostic Imaging: Typical Measures of Analysis

Level 1. Technical efficacy

Resolution of line pairs
Modulation transfer function change
Gray-scale range
Amount of mottle
Sharpness

Level 2. Diagnostic accuracy efficacy

Yield of abnormal or normal diagnoses in a case series
Diagnostic accuracy (% correct diagnoses in case series)
Sensitivity and specificity in a defined clinical problem setting
Measures of area under the ROC curve

Level 3. Diagnostic thinking efficacy

Number (%) of cases in a series in which image judged "helpful" to making the diagnosis
Entropy change in differential diagnosis probability distribution
Difference in clinicians' subjectively estimated diagnosis probabilities pre- to post-test information
Empirical subjective log-likelihood ratio for test positive and negative in a case series

Level 4. Therapeutic efficacy

Number (%) of times image judged helpful in planning management of patient in a case series
% of times medical procedure avoided due to image information
Number (%) of times therapy planned before imaging changed after imaging information obtained (retrospectively inferred from clinical records)
Number (%) of times clinicians' prospectively stated therapeutic choices changed after information obtained

Level 5. Patient outcome efficacy

% of patients improved with test compared with/without test
Morbidity (or procedures) avoided after having image information
Change in quality-adjusted life expectancy
Expected value of test information in quality-adjusted life years (QALYs)
Cost per QALY saved with imaging information
Patient utility assessment; e.g., Markov modeling; time trade-off

Level 6. Societal efficacy

Benefit-cost analysis from societal viewpoint
Cost-effectiveness analysis from societal viewpoint

Source: Thornbury 1992.

- Health professions organizations
- Standards setting organizations
- Hospitals and health care networks
- Group purchasing organizations
- Patient and consumer organizations
- Government policy research agencies
- Private sector assessment/policy research organizations
- Academic health centers
- Biomedical research agencies
- Health product companies
- Venture capital groups and other investors

The purposes, scope, methods, and other characteristics of HTAs that are conducted or sponsored by these organizations vary widely. Examples of these organizations are noted in this document. As in other fields, professional societies and organizational consortia exist in HTA. At the international level, HTA International (HTAi) [www.htai.org] has members from HTA agencies, academic institutions, health professions, hospitals and other health care providers, payers, industry, and others from more than 40 countries. The International Network of Agencies for Health Technology Assessment (INAHTA) [www.inahta.org] is a network of about 40 organizations (including government agencies and non-profit private sector organizations) that generate a shared HTA report database and engage in related collaborative activities. Examples of other professional organizations whose interests include areas related to HTA include:

- AcademyHealth [www.academyhealth.org]
- Cochrane Collaboration [www.cochrane.org]
- International Health Economics Association [www.health economics.org]
- International Society for Pharmacoeconomics and Outcomes Research [www.ispor.org]
- Society for Medical Decision Making [www.smdm.org]

Expertise for Conducting HTA

Given the variety of impacts addressed and the range of methods that may be used in an assessment, multiple types of experts are needed in HTA. Depending upon the topic and scope of assessment, these may include a selection of the following:

- Physicians, nurses, dentists, and other clinicians
- Managers of hospitals, clinics, nursing homes, and other health care institutions
- Radiology technicians, laboratory technicians and other allied health professionals
- Biomedical and clinical engineers

- Pharmacologists
- Patients and patient affairs representatives
- Epidemiologists
- Biostatisticians
- Economists
- Lawyers
- Social scientists
- Ethicists
- Decision scientists
- Computer scientists/programmers
- Librarians/information specialists

Certain individuals have expertise in more than one area. The set of participants in an assessment depends upon its purpose, available resources and other factors. For example, the standing members of a hospital technology assessment committee might include: the chief executive officer, chief financial officer, physician chief of staff, director of nursing, director of planning, materials manager and director of biomedical engineering (Sadock 1997; Taylor 1994). Physician specialists and marketing, legal, patient affairs and additional analytical support staff could be involved as appropriate.

Ten Basic Steps of HTA

There is great variation in the scope, selection of methods and level of detail in the practice of HTA. Nevertheless, most HTA activity involves some form of the following basic steps.

1. Identify assessment topics
2. Specify the assessment problem
3. Determine locus of assessment
4. Retrieve evidence
5. Collect new primary data (as appropriate)
6. Appraise/interpret evidence
7. Integrate/synthesize evidence
8. Formulate findings and recommendations
9. Disseminate findings and recommendations
10. Monitor impact

Not all assessment programs conduct all of these steps, and they are not necessarily conducted in a linear manner. Many HTA programs rely largely on integrative methods of reviewing and synthesizing data from existing primary data studies (reported in journal articles or from epidemiological or administrative data sets), and do not collect primary data. Some assessment efforts involve multiple

cycles of retrieving/collecting, interpreting, and integrating evidence before completing an assessment. For example, to gain regulatory approval (e.g., by the US FDA) to market a new drugs, pharmaceutical companies typically sponsor several iterations of new data collection: preclinical testing in the laboratory and in animals and phase I, II, and III studies in humans; additional phase IV post marketing studies may be a condition of approval. The steps of appraising and integrating evidence may be done iteratively, such as when a group of primarily data studies are appraised individually for quality, then are integrated into a body of evidence, which in turn is appraised for its overall quality. Depending upon the circumstances of an HTA, the dissemination of findings and recommendations and monitoring of impact may not be parts of the HTA itself, although they may be important responsibilities of the sponsoring program or parent organization.

Another framework for HTA is offered by the European Collaboration for Health Technology Assessment (Busse 2002), as follows.

- Submission of an assessment request/identification of an assessment need
- Prioritization
- Commissioning
- Conducting the assessment
    - Definition of policy question(s)
    - Elaboration of HTA protocol
    - Collecting background information/determination of the status of the technology
    - Definition of the research questions
    - Sources of data, appraisal of evidence, and synthesis of evidence for each of:
        - Safety
        - Efficacy/effectiveness
        - Psychological, social, ethical
        - Organizational, professional
        - Economic
    - Draft elaboration of discussion, conclusions, and recommendations
    - External review
    - Publishing of final HTA report and summary report
- Dissemination
- Use of HTA
- Update of the HTA

As indicated by various chapter and section headings, all ten of the basic steps of HTA listed above are described in this document.

III. PRIMARY DATA AND INTEGRATIVE METHODS

HTA embraces a diverse group of methods that can be grouped into two broad categories. Primary data methods involve collection of original data, ranging from more scientifically rigorous approaches such as randomized controlled trials to less rigorous ones such as case studies. Integrative methods (also known as "secondary" or "synthesis" methods) involve combining data or information from existing sources, including from primary data studies. These can range from quantitative, structured approaches such as meta-analyses or systematic literature reviews to informal, unstructured literature reviews. Cost analysis methods, which can involve one or both of primary data methods and integrative methods, are discussed separately a following section.

Most HTA programs use integrative approaches, with particular attention to formulating findings that are based on distinguishing between stronger and weaker evidence drawn from available primary data studies. Some HTA programs do collect primary data, or are part of larger organizations that collect primary data. It is not always possible to conduct, or base an assessment on, the most rigorous types of studies. Indeed, policies often must be made in the absence, or before completion, of definitive studies. There is no standard methodological approach for conducting HTA. Given their varying assessment orientations, resource constraints and other factors, assessment programs tend to rely on different combinations of methods. Even so, the general trend in HTA is to call for and emphasize the more rigorous methods.

Types of Methodological Validity

Validity refers to whether what we are measuring is what we intend to measure. Methodological designs vary in their ability to produce valid findings. Understanding different aspects of validity helps in comparing alternative methodological designs and interpreting the results of studies using those designs. Although these concepts are usually addressed in reference to primary data methods, they apply as well to integrative methods.

Internal validity refers to the extent to which the findings of a study accurately represent the causal relationship between an intervention and an outcome in the particular circumstances of an investigation. This includes the extent to which a study minimizes any systematic or non-random error in the data.

External validity refers to the extent to which the findings obtained from an investigation conducted under particular circumstances can be generalized to other circumstances. To the extent that the circumstances of a particular investigation (e.g., patient characteristics or the manner of delivering a treatment) differ from the circumstances of interest, the external validity of the findings of that investigation may be questioned.

Face validity is the ability of a measure to represent reasonably a construct (i.e., a concept or domain of interest) as judged by someone with expertise in the health problem and interventions of interest. *Content validity* refers to the degree to which a measure covers that range of meanings or dimensions of a construct. As noted above, an outcome measure is often used as a marker or surrogate for a disease of interest. For example, how well do changes in prostate specific antigen (PSA) levels represent or predict the risk of prostate cancer? How well does performance on an exercise treadmill represent cardiovascular fitness?

Construct validity is the ability of a measure to correlate with other accepted measures of the construct of interest, and to discriminate between groups known to differ according to the variable. *Convergent validity* refers to the extent to which two different measures that are intended to measure the same construct do indeed yield similar results. *Discriminant validity,* opposite convergent validity, concerns whether different measures that are intended to measure different constructs do indeed fail to be positively associated with each other. *Concurrent validity* refers to the ability of a measure to accurately differentiate between different groups at the time the measure is applied, or the correlation of one measure with another at the same point in time. *Predictive validity* refers to the ability to use differences in a measure to predict future events or outcomes.

Primary Data Methods

The considerable and diverse array of primary data methods includes, e.g., true experiments such as randomized controlled trials (RCTs) and other controlled trials; other prospective but uncontrolled trials; observational studies such as case-control, cross-sectional studies, and surveillance studies; and simpler designs such as case series and single case reports or anecdotes. These methods can be described and categorized in terms of multiple attributes or dimensions, such as whether they are prospective or retrospective, interventional or observational, controlled or uncontrolled, and other attributes noted below. Some of these methods have alternative names, and many studies employ nearly limitless combinations of these attributes.

Some Fundamental Attributes

Prospective studies are planned and implemented by investigators using real-time data collection. These typically involve identification of one or more patient groups, collection of baseline data, delivering one or more interventions, collecting follow-up data, and comparing baseline to follow-up data for the patient groups. In retrospective studies, investigators collect samples of data from past events (interventions and outcomes) involving one or more patient groups. In an **interventional study**, investigators prospectively deliver, manipulate, or manage the intervention(s) of interest. In an **observational study** investigators only monitor or follow an intervention or exposure (e.g., of a risk factor) of interest, but do not themselves intervene in the delivery the intervention or exposure.

Many studies use separate **control groups** of patients as a basis of comparison to the one or more groups receiving an intervention of interest. Some studies do not use control groups; these uncontrolled studies rely on comparing patient measures before and after an intervention to determine whether the intervention had an effect. Control groups of patients are constituted in ways to ensure as much similarity as possible to patients in intervention groups. Except for the intervention of interest, the treatment and management of control groups and intervention groups are as similar as possible. By comparing changes in intervention groups to changes in control groups, investigators seek to isolate the effect of an intervention on patient outcomes from any effects on patient outcomes from extraneous factors. While most controlled studies used contemporaneous controls alongside (i.e., identified and followed at the same time as) intervention groups, investigators sometimes use **historical control** groups. In a **crossover design** study, patients start in one group (intervention or control) and then are switched to the other, thereby acting as their own controls.

Various means are used to ensure that intervention and control groups comprise patients with similar characteristics, so that "baseline" (initial) differences in these groups will not affect (confound) the relative changes in patient outcomes between the groups. Although such means as alternate assignment or using birthdays or identification numbers are used to assign patients to intervention and control groups, assignment based on **randomization** is preferred because it minimizes opportunities for bias to affect the composition of these groups at baseline. Knowledge of assignment of patients to one group or another, e.g., to a group receiving a new intervention or a group receiving standard care, can itself affect outcomes as experienced by patients and/or assessed by investigators. Therefore, some studies employ **blinding** of patients, and sometimes of investigators and data analysts, to knowledge of patient assignment to intervention and control groups in an effort to eliminate the confounding effects of such knowledge.

Alternative Methods Offer Tradeoffs in Validity

Although primary study investigators and assessors would prefer to have methods that are both internally and externally valid, they often find that study design attributes that increase one type of validity jeopardize the other. A well designed and conducted RCT is widely considered to be the best approach for ensuring internal validity, as it gives investigators the most control over factors that could confound the causal relationship between the a technology and health outcomes. However, for the reasons that a good RCT has high internal validity, its external validity may be limited.

Most RCTs are designed to investigate the effects of an intervention in specific types of patients so that the relationship between the intervention and outcomes is less likely to be confounded by patient variations. However, findings from an RCT involving a narrowly defined patient group may not be applicable for the same intervention given to other types of patients. Patients allowed to enroll in RCTs are often subject to strict inclusion and exclusion criteria pertaining, e.g., to age, risk factors, and previous and current treatments. This is done for various reasons, including to avoid confounding the treatment effect of the intervention in question by previous or current other treatments, and to limit the extent to which variations in patients' response to a treatment might dilute the treatment effect across the enrolled patient population. As a result, the patient population in an RCT may be less likely to be representative of the desired or potential target population in practice. As noted above, RCTs often involve special protocols of care and testing that may not be characteristic of general care, and are often conducted in university medical centers or other special settings that may not represent the general or routine settings in which most health care is provided.

Findings of some large **observational studies** (e.g., from large cross-sectional studies or registries) have external validity to the extent that they can provide insights into the types of outcomes that are experienced by different patient groups in different circumstances. However, these less rigorous designs are more subject to certain forms of **bias** that threaten internal validity, diminishing the certainty with which particular outcomes can be attributed to an intervention. Interesting or promising findings from weaker studies can raise hypotheses that can be tested using stronger studies. The use of "large, simple trials" (discussed below) is an attempt to combine the strengths of RCTs and observational studies.

RCTs Not Best Design for All Questions

While RCTs are the "gold standard" of internal validity for causal relationships, they are not necessarily the best method for answering all questions of relevance to an HTA. As noted by Eisenberg (1999):

"Those who conduct technology assessments should be as innovative in their evaluations as the technologies themselves .... The randomized trial is unlikely to be replaced, but it should be complemented by other designs that address questions about technology from different perspectives."

Other types of studies may be preferred to RCTs for different questions. For example, a good way to describe the prognosis for a given disease or condition may be a set of follow-up studies of patient cohorts at uniform points in the clinical course of a disease. Case control studies are often used to identify risk factors for diseases, disorders, and adverse events. The accuracy of a diagnostic test (as opposed to its ultimate effect on health outcomes) may be determined by a cross-sectional study of patients suspected of having a disease or disorder. Non-randomized trials or case series may be preferred for determining the effectiveness of interventions for otherwise fatal conditions, i.e., where little or nothing is to be gained by comparison to placebos or known ineffective treatments. Surveillance and registries are used to determine the incidence of rare, serious adverse events that may be associated with an interventions. For incrementally modified technologies posing no known additional risk, registries may be appropriate for determining safety and effectiveness.

Collecting New Primary Data

It is beyond the scope of this document to describe the planning, design, and conduct of clinical trials, observational studies, and other investigations for collecting new primary data. There is a considerable and evolving literature on these subjects (Chow 1998; Spilker 1991; Spilker 1995; Rothenberg 2003). Also, there is a literature on priority setting and efficient resource allocation for clinical trials, and cost-effective design of clinical trials (Detsky 1990; Thornquist 1993).

As noted above, compiling evidence for an assessment may entail collection of new primary data. An assessment program may determine that existing evidence is insufficient for meeting the desired policy needs, and that new studies are needed to generate data for particular aspects of the assessment. Once available, the new data can be interpreted and incorporated into the existing body of evidence.

In the US, major units of the NIH such as the National Cancer Institute and the National Heart, Lung and Blood Institute sponsor and conduct biomedical research, including clinical trials. Elsewhere at NIH, the Office of Medical Applications of Research coordinates the NIH Consensus Development Program, but does not collect primary clinical data, although it occasionally surveys physician specialists and other groups for which the NIH assessment reports are targeted. The Veterans Health Administration (VHA) Cooperative Studies Program [www.va.gov/resdev/csp.cfm] is responsible for the planning and conduct of large multicenter clinical trials within the VHA, including approximately 60 cooperative studies at any one time. The FDA does not typically conduct primary studies related to the marketing of new drugs and devices; rather, the FDA reviews primary data from studies sponsored or conducted by the companies that make these technologies.

The ability of most assessment programs to undertake new primary data collection, particularly clinical trials, is limited by such factors as programs' financial constraints, time constraints, responsibilities that do not include conducting or sponsoring clinical studies, and other aspects of the roles or missions of the programs. An HTA program may decide not to undertake and assessment if insufficient data are available. Whether or not an assessment involves collection of new primary data, the assessment reports

should note what new primary studies should be undertaken to address gaps in the current body of evidence, or to meet anticipated assessment needs.

Health professional organizations and societies, e.g., the American College of Physicians and American College of Cardiology, work almost exclusively from existing data and do not do clinical research. Third-party payers generally do not sponsor clinical studies, but increasingly analyze claims data and other administrative data. Payers have supported trials of new technologies indirectly by paying for care associated with clinical studies of those technologies, or by paying unintentionally for uncovered new procedures that were coded as covered procedures. As noted above, some payers are providing conditional coverage for certain investigational technologies in selected settings in order to compile data that can be used to make more informed coverage decisions. One recent, controversial example, is the multicenter RCT of lung-volume reduction surgery, the National Emphysema Treatment Trial (NETT), funded by the NHLBI and the Centers for Medicare and Medicaid Services (CMS, which administers the US Medicare program) (Fishman 2003; Ramsey 2003).

Primary Data Collection Trends Relevant to HTA

Primary data collection methods have evolved in certain important ways that affect the body of evidence used in HTA. Among these, investigators have made progress in trying to combine some of the desirable attributes of RCTs and observational studies. For example, while retaining the methodological strengths of prospective, randomized design, "**large, simple trials**" use large numbers of patients, more flexible patient entry criteria and multiple study sites to improve external validity and gain effectiveness data. Also, fewer types of data may be collected for each patient, easing participation by patients and clinicians (Buring 1994; Ellenberg 1992; Peto 1993; Yusuf 1990). Examples of these approaches include large, simple trials supported by NIH; certain large, multicenter RCTs coordinated by the VA Cooperative Studies Program, and "firm trials" involving random assignment of patients and providers to alternative teams to evaluate organizational and administrative interventions (Cebul 1991).

Clinical trials conducted for the purposes of advancing biomedical research or for achieving market clearance by regulatory bodies approval do not necessarily address clinical choices or policy decisions (e.g., coverage policies). The call for "pragmatic" or "practical" clinical trials (PCTs) is intended to meet these needs more directly. Among their main attributes, PCTs (1) select clinically relevant alternative interventions to compare, (2) include a diverse population of study participants, (3) recruit participants from heterogeneous practice settings, and (4) collect data on a broad range of health outcomes. PCTs will require that clinical and health policy decision makers become more involved in priority setting, research design, funding, and other aspects of clinical research (Tunis 2003).

Biomedical research organizations such as NIH and regulatory agencies such as FDA permit certain mid-trial changes in clinical trial protocols such as drug dosage modifications and patient cross-overs to alternative treatment groups to reflect the most recent scientific findings. Selected use of **surrogate endpoints**, especially biological markers, is employed where these are known to be highly correlated with “hard endpoints” such as morbidity and mortality that may not occur until months or years later. For example, a long-standing surrogate marker for stroke risk is hypertension, although understanding continues to evolve of the respective and joint roles of systolic and diastolic pressures in predicting stroke (Basile 2002). Trials of new drugs for HIV/AIDS use such biological markers as virological (e.g., plasma HIV RNA) and immunological (e.g., CD4+ cell counts) levels (Lalezari 2003).

Streamlining or combining clinical trial phases and "parallel track" availability of technologies to patients outside of ongoing formal RCTs are intended to speed regulatory approval and make technologies available to patients who are ineligible for RCT protocols but have exhausted other treatments. For example, the FDA provides several types of access to investigational treatments. "Emergency use" is allowed in situations where there is a need for an investigational technology in a manner that is not consistent with the approved investigational protocol or by a physician who is not part of the clinical trial, and may occur before FDA approval of the investigational plan (e.g., IND for drugs or IDE for devices). "Compassionate" use allows access for patients who do not meet the requirements for inclusion in an ongoing clinical trial but for whom the treating physician believes the technology may provide a benefit; this usually applies to patients with a serious disease or condition for whom there is no viable alternative treatment. "Treatment use" refers to instances where data collected during a trial indicates that a technology is effective, so that during the trial or prior to full FDA review and approval for marketing, the technology may be provided to patients not in the trial, subject to the other requirements of the trial (e.g., under an IND for drugs or IDE for devices). "Continued access" allows continued enrollment of patients after the trial has been completed, in order to enable access to the investigational technology while the marketing application is being prepared by the sponsor or reviewed by the FDA. Although many of these adaptations were originally instituted for RCTs involving new drug treatments for HIV/AIDS, cancer, and other life-threatening conditions (Merigan 1990), their use in trials of treatments for other conditions is increasing.

Another important type of development in primary data collection is the incorporation of contemporaneous cost data collection in prospective clinical trials. Health care product companies increasingly are using such data in product promotion and to help secure favorable payment decisions (Anis 1998; Henry 1999). The generation of health and economic data are increasingly influenced by technical guidance for data submissions provided by national and regional HTA agencies, particularly in Canada, Europe, and Australia (Hill 2000; Hjelmgren 2001; Taylor 2002).

Integrative Methods

Having considered the merits of individual studies, an assessment group must begin to integrate, synthesize, or consolidate the available findings. For many topics in HTA, there is no single definitive primary study, e.g., that settles whether one technology is better than another for a particular clinical situation. Even where definitive primary studies exist, findings from different types of studies must be combined or considered in broader social and economic contexts in order to formulate policies.

Methods used to combine or integrate data include the following:

- Meta-analysis
- Modeling (e.g., decision trees, Markov models)
- Group judgment ("consensus development")
- Systematic literature review
- Unstructured literature review
- Expert opinion

The biases inherent in traditional means of consolidating literature (i.e., non-quantitative or unstructured literature reviews and editorials) are well recognized, and greater emphasis is given to more structured,

quantified and better documented methods. The body of knowledge concerning how to strengthen and apply these integrative methods has grown substantially in recent years. Considerable work has been done to improve the validity of decision analysis and meta-analysis in particular (Eckman 1992; Eddy 1992; Lau 1992). Experience with the NIH Consensus Development Program, the panels on appropriateness of selected medical and surgical procedures conducted by the RAND Corporation, the clinical practice guidelines activities sponsored until the mid-1990s by the former AHCPR (renamed as AHRQ), and others continues to add to the body of knowledge concerning group judgment processes.

Three major types of integrative methods−meta-analysis, decision analysis, and consensus development−are described below.

Meta-analysis

Meta-analysis refers to a group of statistical techniques for combining results of multiple studies to obtain a quantitative estimate of the overall effect of a particular technology (or variable) on a defined outcome. This combination may produce a stronger conclusion than can be provided by any individual study (Laird 1990; Normand 1999; Thacker 1988). The purposes of meta-analysis are to:

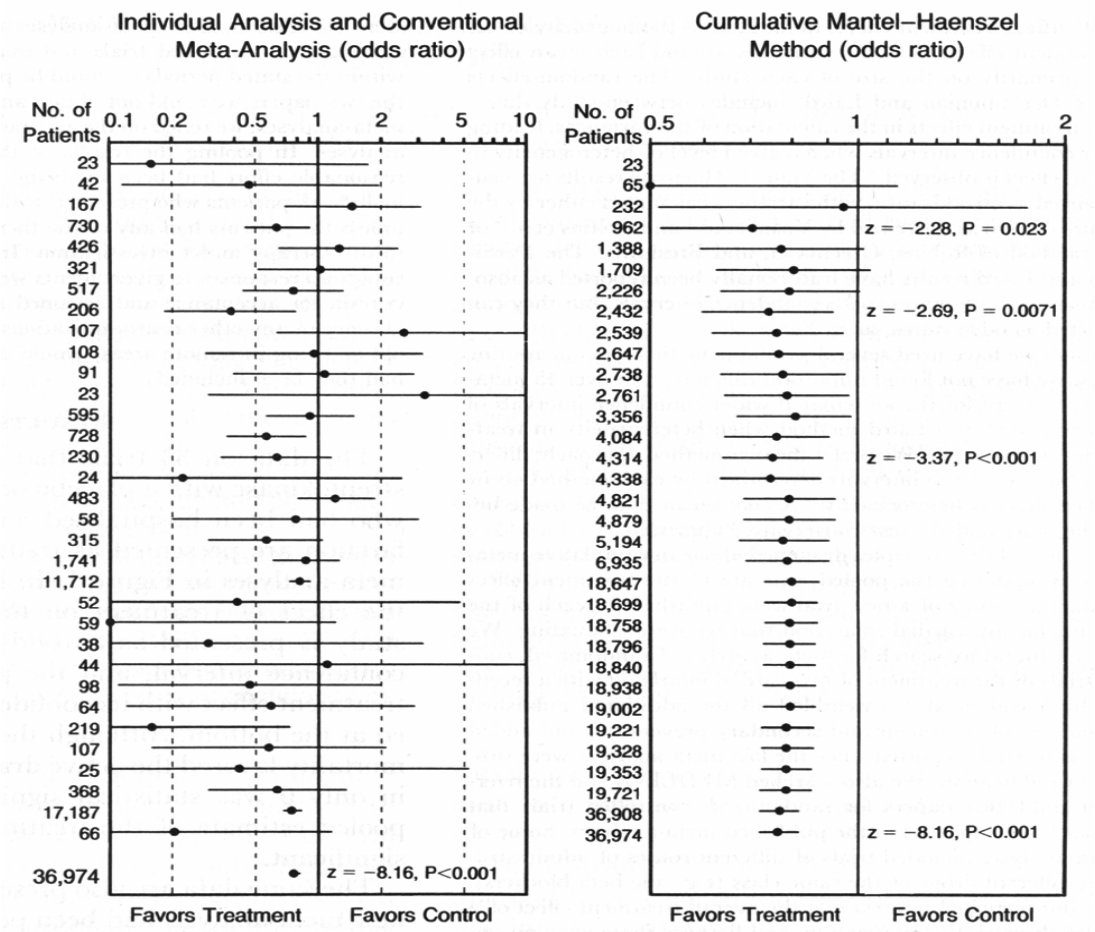
- Encourage systematic organization of evidence
- Increase statistical power for primary end points
- Increase general applicability (external validity) of findings
- Resolve uncertainty when reports disagree
- Assess the amount of variability among studies
- Provide quantitative estimates of effects (e.g., odds ratios or effect sizes)
- Identify study characteristics associated with particularly effective treatments
- Call attention to strengths and weaknesses of a body of research in a particular area
- Identify needs for new primary data collection

Meta-analysis typically is used for topics that have no definitive studies, including topics for which non-definitive studies are in some disagreement. Evidence collected for assessments often includes studies with insufficient statistical power (e.g., because of small sample sizes) to detect any true treatment effects. By combining the results of multiple studies, a meta-analysis may have sufficient power to detect a true treatment effect if one exists, or at least narrow the confidence interval about the mean treatment effect. **Box 13** shows an example of a meta-analysis of thrombolytic therapy.

Although meta-analysis has been applied primarily for treatments, meta-analytic techniques are being extended to diagnostic technologies. As in other applications of meta-analysis, the usefulness of these techniques for diagnostic test accuracy is subject to publication bias and the quality of primary studies of diagnostic test accuracy (Deeks 2001; Hasselblad 1995; Irwig 1994; Littenberg 1993). Although meta-analysis is often applied to RCTs, it may be used for observational studies as well (Stroup 2000).

Box 13. Meta-Analysis: Clinical Trials of Intravenous Streptokinase for Acute Myocardial Infarction

Lau et al. conducted two types of meta-analysis of 33 clinical trials of the effect on mortality of using the thrombolytic (i.e., to dissolve blood clots) drug streptokinase for treating myocardial infarction.



The conventional meta-analysis at left depicts observed treatment effects (odds ratios) and confidence intervals of the 33 individual studies, most of which involved few patients. Although most trials favored streptokinase, the 95 percent confidence intervals of most trials included odds ratios of 1.0 (indicating no difference between treatment with streptokinase and the control intervention). Several studies favored the control treatment, although all of their confidence intervals included odds ratios of 1.0. As shown at the bottom, this meta-analysis pooled the data from all 33 studies (involving a total of nearly 37,000 patients) and detected an overall treatment effect favoring streptokinase, with a narrow 95 percent confidence interval that fell below the 1.0 odds ratio, and *P* less than 0.001. (*P* values less than 0.05 or 0.01 are generally accepted as statistically significant.)

The chart at right depicts a "cumulative" meta-analysis in which a new meta-analysis is performed with the chronological addition of each trial. As early as 1971, when available studies might have appeared to be inconclusive and contradictory, a meta-analysis involving only four trials and 962 patients would have indicated a statistically significant treatment effect favoring streptokinase (note 95% confidence interval and $P<0.05$). By 1973, after eight trials and 2,432 patients, *P* would have been less than 0.01. By 1977, the *P* value would have been less than 0.001, after which the subsequent trials had little or no effect on the results establishing the efficacy of streptokinase in saving lives. This approach indicates that streptokinase could have been shown to be lifesaving two decades ago, long before FDA approval was sought and it was adopted into routine practice.

Source: Lau 1992.

The basic steps in meta-analysis are the following:

1. Specify the problem of interest.
2. Specify the criteria for inclusion of studies (e.g., type and quality).
3. Identify all studies that meet inclusion criteria.
4. Classify study characteristics and findings according to, e.g.: study characteristics (patient types, practice setting, etc.), methodological characteristics (e.g., sample sizes, measurement process), primary results and type of derived summary statistics.
5. Statistically combine study findings using common units (e.g., by averaging effect sizes); relate these to study characteristics; perform sensitivity analysis.
6. Present results.

Some of the particular techniques used in the statistical combination of study findings in meta-analysis are: pooling, effect size, variance weighting, Mantel-Haenszel, Peto, DerSimonian and Laird, and confidence profile method. The suitability of any of these techniques for a group of studies depends upon the comparability of circumstances of investigation, type of outcome variables used, assumptions about the uniformity of treatment effects and other factors (Eddy 1992; Laird 1990; Normand 1999).

The different techniques of meta-analysis have specific rules about whether or not to include certain types of studies and how to combine their results. Some meta-analytic techniques adjust the results of the individual studies to try to account for differences in study design and related biases to their internal and external validity. Special computational tools (e.g., computer software) may be required to make the appropriate adjustments for the various types of biases in a systematic way (Detsky 1992; Moher 1999; van Houwelingen 2002).

Meta-analysis can be limited by poor quality of data in primary studies, publication bias, biased selection of available studies, insufficiently comparable studies selected (or available) for a meta-analysis, and biased interpretation of findings. The quality of RCTs used in meta-analyses can bias results. The results of meta-analyses that are based on sets of RCTs with lower methodological quality tend to show greater treatment effects (i.e., greater efficacy of interventions) than those based on sets of RCTs of higher methodological quality (Moher 1998). However, it is not apparent that any individual quality measures are associated with the magnitude of treatment effects in meta-analyses of RCTs (Balk 2002). As is the case for RCTs, there are instruments for assessing the quality of meta-analyses and systematic reviews (Moher 1999), as shown in **Box 14**.

The shortcomings of meta-analyses, which are shared by unstructured literature reviews and other less rigorous synthesis methods, can be minimized by maintaining a systematic approach to meta-analysis. Performing meta-analyses in the context of systematic reviews, i.e., that have objective means of searching the literature and applying predetermined inclusion and exclusion criteria to the primary studies used, can diminish the impact of these shortcomings on the findings of meta-analyses (Egger 2001). Compared to the less rigorous methods of combining evidence, meta-analysis can be time-consuming and requires greater statistical and methodologic skills. However, meta-analysis is a much more explicit and accurate method.

Box 14. Instrument to Assess the Quality of Meta-Analysis Reports

Heading	Subheading	Descriptor	Reported Y/N	Page Number
Title		Identify the report as a meta-analysis [or systematic review] of RCTs		
Abstract		Use a structured format		
		Describe		
	Objectives	The clinical question explicitly		
	Data sources	The databases (ie, list) and other information sources		
	Review methods	The selection criteria (ie, population, intervention, outcome, and study design); methods for validity assessment, data abstraction, and study characteristics, and quantitative data synthesis in sufficient detail to permit replication		
	Results	Characteristics of the RCTs included and excluded; qualitative and quantitative findings (ie, point estimates and confidence intervals); and subgroup analyses		
	Conclusion	The main results		
		Describe		
Introduction		The explicit clinical problem, biological rationale for the intervention, and rationale for review		
Methods	**Searching**	The information sources, in detail (eg, databases, registers, personal files, expert informants, agencies, hand-searching), and restrictions (years considered, publication status, language of publication)		
	Selection	The inclusion and exclusion criteria (defining population, intervention, principal outcomes, and study design		
	Validity assessment	The criteria and process used (eg, masked conditions, quality assessment, and their findings)		
	Data abstraction	The process or processes used (eg, completed independently, in duplicate)		
	Study characteristics	The type of study design, participants' characteristics, details of intervention, outcome definitions, &c, and how clinical heterogeneity was assessed		
	Quantitative data synthesis	The principal measure of effect (eg, relative risk), method of combining results (statistical testing and confidence intervals), handling of missing data; how statistical heterogeneity was assessed, a rationale for any a-priori sensitivity and subgroup analyses; and any assessment of publication bias		
Results	Trial Flow	Provide a meta-analysis profile summarizing trial flow		
	Study characteristics	Present descriptive data for each trial (eg, age, sample size, intervention, dose, duration, follow-up period)		
	Quantitative data synthesis	Report agreement on the selection and validity assessment; present simple summary results (for each treatment group in each trial, for each primary outcome); present data needed to calculate effect sizes and confidence intervals in intention-to-treat analyses (eg, 2x2 tables of counts, means and SDs, proportions)		
Discussion		Summarise key findings; discuss clinical inferences based on internal and external validity; interpret the results in light of the totality of available evidence; describe potential biases in the review process (eg, publication bias); and suggest a future research agenda		

Source: Moher 1999.

Even though many assessments still tend to rely on overall subjective judgments and similar less rigorous approaches of integrating evidence, there is a clear trend toward learning about and using more meta-analytic approaches. An assessment group that uses the inclusion/exclusion rules and other stipulations of meta-analysis is likely to conduct a more thorough and credible assessment, even if the group decides not to perform the final statistical consolidation of the results of pertinent studies.

More advanced meta-analytic techniques are being applied to assessing health technologies, e.g., involving multivariate treatment effects, meta-regression, and Bayesian methods (van Houwelingen 2002). As meta-analysis and other structured literature syntheses are used more widely in evaluating health care interventions, methodological standards for conducting and reporting meta-analyses are rising (Egger, Davey Smith 2001, Moher 1999, Petitti 2001).

Modeling

Quantitative modeling is used to evaluate the clinical and economic effects of health care interventions. Models are used to answer "What if?" questions. That is, they are used to represent (or simulate) health care processes or decisions and their impacts under conditions of uncertainty, such as in the absence of actual data or when it is not possible to collect data on all potential conditions, decisions, and outcomes. For example, decision analytic modeling is used to represent the sequence of clinical decisions and its health and economic impacts. Economic modeling can be used to estimate the cost-effectiveness of alternative technologies for a given health problem.

By making informed adjustments or projections of existing primary data, modeling can help account for patient conditions, treatment effects, and costs that are not present in primary data. This may include bridging efficacy findings to estimates of effectiveness, and projecting future costs and outcomes.

Among the main types of techniques used in quantitative modeling are decision analysis (described below), Markov model process, Monte Carlo simulation, survival and hazard functions, and fuzzy logic (Tom 1997). A **Markov model** (or chain) is a way to represent and quantify changes from one state of health to another. A **Monte Carlo simulation** uses sampling from random number sequences to assign estimates to parameters with multiple possible values, e.g., certain patient characteristics (Caro 2002; Gazelle 2003).

Decision analysis uses available quantitative estimates to represent (model or simulate) the sequences of alternative strategies (e.g., of diagnosis and/or treatment) in terms of the probabilities that certain events and outcomes will occur and the values of the outcomes that would result from each strategy (Pauker 1987; Thornton 1992). Decision models often are shown in the form of "decision trees" with branching steps and outcomes with their associated probabilities and values. Various software programs may be used in designing and conducting decision analyses, accounting for differing complexity of the strategies, extent of sensitivity analysis, and other quantitative factors.

Decision models can be used in different ways. They can be used to predict the distribution of outcomes for patient populations and associated costs of care. They can be used as a tool to support development of clinical practice guidelines for specific health problems. For individual patients, decision models can be used to relate the likelihood of potential outcomes of alternative clinical strategies, and/or to identify the clinical strategy that has the greatest utility for a patient. Decision models are also used to set priorities for HTA (Sassi 2003).

The basic steps of decision analysis are:

1. Develop a model (e.g., a decision tree) that depicts the set of important choices (or decisions) and potential outcomes of these choices. For treatment choices, the outcomes may be health outcomes (health states); for diagnostic choices, the outcomes may be test results (e.g., positive or negative).
2. Assign estimates (based on available literature) of the probabilities (or magnitudes) of each potential outcome given its antecedent choices.
3. Assign estimates of the value of each outcome to reflect its utility or desirability (e.g., using a HRQL measure or QALYs).
4. Calculate the expected value of the outcomes associated with the particular choice(s) leading to those outcomes. This is typically done by multiplying the set of outcome probabilities by the value of each outcome.
5. Identify the choice(s) associated with the greatest expected value. Based on the assumptions of the decision model, this is the most desirable choice, as it provides the highest expected value given the probability and value of its outcomes.
6. Conduct a **sensitivity analysis** of the model to determine if plausible variations in the estimates of probabilities of outcomes or utilities change the relative desirability of the choices. (Sensitivity analysis is used because the estimates of key variables in the model may be based on limited data or simply expert conjecture.)

Box 15 shows an example of a decision tree for alternative therapies for managing recurrent angina following coronary artery bypass graft surgery. A limitation of modeling with decision trees is representing recurrent health states (e.g., recurrent complications or stages of a chronic disease). An alternative approach is to use state-transition models that use probabilities of moving from one state of health to another, including remaining in a given state or returning to it after intervening health states. Markov modeling is a commonly used type of state-transition modeling.

The assumptions and estimates of variables used in models should be validated against actual data as it becomes available, and the models should be modified accordingly. Modeling should incorporate sensitivity analyses to quantify the conditional relationships between model inputs and outputs.

Models and their results are only aids to decision-making, not statements of scientific, clinical, or economic fact. The report of any modeling study should carefully explain and document the assumptions, data sources, techniques, and software. Modelers should make clear that the findings of a model are conditional upon these components. The use of decision modeling in cost-effectiveness analysis in particular has advanced in recent years, with development of checklists and standards for these applications (Gold 1996; Soto 2002; Weinstein 2003).

Consensus Development

In various forms, group judgment or **consensus development** is used to set standards, make regulatory recommendations/decisions, make payment recommendations/policies, make technology acquisition decisions, formulate practice guidelines, define the state-of-the-art, and other purposes. "Consensus

Box 15. Decision Tree: Management of Angina After Coronary Artery Bypass Surgery

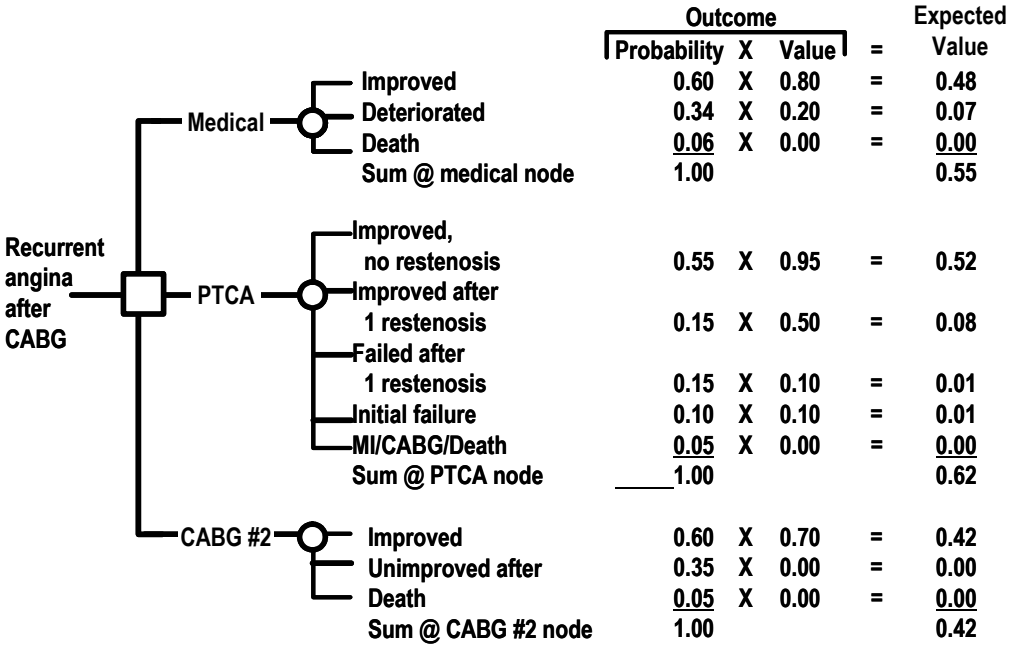


Figure 1. Conventional and Cumulative Meta-Analyses of 33 Trials of Intravenous Streptokinase for Acute Myocardial Infarction. The odds ratios and 95 percent confidence intervals for an effect of treatment on mortality are shown on a logarithmic scale. A bibliography of the published trial reports is available from the authors.

The decision tree shown above depicts the outcomes for three alternatives for managing recurrent angina among patients who have had previous coronary artery bypass graft surgery (CABG). Each of the three main treatment choices--medical management (drugs and lifestyle changes), percutaneous transluminal coronary angioplasty (PTCA) and repeat CABG--results in a set of potential outcomes. For each outcome there is a probability and a value (ranging from 0.0 to 1.0). For example, given medical management, there is a 60% probability of improvement, which is valued at 0.80. According to this model, the most desirable treatment choice is PTCA, as the expected value of its outcomes is 0.62, exceeding that of medical management (0.55) and repeat CABG (0.42).

Source: Mills 1991.

development" can refer to discrete group processes or techniques that contribute to an assessment, such as the nominal group technique or Delphi method; it also can refer to particular consensus development approaches (e.g., the consensus development conferences conducted by the NIH).

In contrast to the quantitative synthesis methods of meta-analysis and decision analysis, consensus development is generally qualitative in nature. It may be unstructured and informal, or it may involve formal group methods such as the nominal group technique and Delphi technique (Fink 1984; Gallagher 1993; Jairath 1994). Although these processes typically involve face-to-face interaction, some consensus development efforts combine remote, iterative interaction of panelists (as in the formal Delphi technique) with face-to-face meetings. Computer conferencing and related telecommunications approaches also are used. There is a modest but growing body of literature on consensus development methods in HTA. For example, one review examined the factors affecting the findings of these processes, including selection of topics and questions, selection of participants, choosing and preparing the scientific evidence, structuring the interaction among participants, and methods for synthesizing individual judgments. Findings and associated support regarding methods of synthesizing individual judgments into consensus are summarized in **Box 16**.

Box 16. Consensus Development: Findings and Associated Support Regarding Methods of Synthesizing Individual Judgments

- An implicit approach to aggregating individual judgments may be adequate for establishing broad policy guidelines. More explicit methods based on quantitative analysis are needed to develop detailed, specific guidelines. **[C]**
- The more demanding the definition of agreement, the more anodyne [bland or non-controversial] the results will be. If the requirement is too demanding, either no statements will qualify or those that do will be of little interest. **[C]**
- Differential weighting of individual participants' views produces unreliable results unless there is a clear empirical basis for calculating the weights. **[B]**
- The exclusion of individuals with extreme views (outliers) can have a marked effect on the content of guidelines. **[A]**
- There is no agreement as to the best method of mathematical aggregation. **[B]**
- Reports of consensus development exercises should include an indication of the distribution or dispersal of participants' judgments, not just the measure of central tendency. In general, the median and the inter-quartile range are more robust than the mean and standard deviation. **[A]**

The extent to which research support exists for any conclusion is indicated, although these should not necessarily be considered as a hierarchy: **A** = clear research evidence; **B** = supporting research evidence; **C** = experienced common-sense judgment.

Source: Murphy 1998.

Virtually all HTA efforts involve consensus development at some juncture, particularly to formulate findings and recommendations. Consensus development also can be used for ranking, such as to set

assessment priorities, and rating. For example, RAND has used a two-stage modified Delphi process (first stage, independent; second stage, panel meeting) in which expert panels rate the appropriateness of a procedure (e.g., tympanostomy tubes for children) for each of many possible patient indications on a scale of 1.0 (extremely inappropriate) to 9.0 (extremely appropriate) (Kleinman 1994).

The opinion of an expert committee concerning, e.g., the effectiveness of a particular intervention, does not in itself constitute strong evidence. Where the results of pertinent, rigorous scientific studies exist, these should take precedence. In the absence of strong evidence, and where practical guidance is needed, expert group opinion can be used to infer or extrapolate from the limited available evidence. Where many assessment efforts are deficient is not making clear where the evidence stops and where the expert group opinion begins.

Consensus development programs typically embrace most of the steps of HTA described here. In these programs, the consensus development conference usually spans at least three of the HTA steps: interpret evidence, integrate evidence, and formulate findings and recommendations. Increasingly, consensus development efforts start with presentations of previously compiled evidence reports.

Many current consensus development programs in the US and around the world are derived from the model of consensus development conference originated at the US NIH in 1977 as part of an effort to improve the translation of NIH biomedical research findings to clinical practice. NIH has modified and experimented with its process over the years. As of late 2003, NIH had held 120 consensus development conferences. (NIH also has conducted about 25 state-of-the-science conferences, using a similar format to its consensus development conferences.) Australia, Canada, Denmark, France, Israel, Japan, The Netherlands, Spain, Sweden and the UK are among the countries that use various forms of consensus development programs to evaluate health technologies. Consensus conferences also are held jointly by pairs of countries and by international health organizations. Although they originally may have been modeled after the US program, these programs have evolved to meet the needs of their respective national environments and sponsoring organizations (McGlynn 1990).

The variety in consensus development programs can be described and compared along several main types of characteristics, as follows.

- Context of the process: e.g., intended audience, topics and impacts addressed, topic selection
- Pre-panel preparation: e.g., responsibility for planning, evidence preparation, prior drafting of questions and/or recommendations
- Panel composition: e.g., panel size, selection, types of expertise/representation, characteristics of chairperson
- Consensus conference attributes: e.g., length of conference, public involvement, private panel sessions, definition of consensus, decision-making procedures (such as voting), process for handling disagreement, format and dissemination of final product

Among most programs, preparation for conferences takes approximately one year. Some programs prepare assessment questions and draft a consensus statement prior to the consensus conference; other programs do not. Most programs assemble compilations of evidence and share this in advance some form with the panelists; in some instances, this involves providing systematic literature reviews with specific review instructions to panelists weeks in advance of the conference. Programs usually provide for speakers to present the evidence during the consensus conference. Most programs select panels of 9-

18 members, including clinicians, scientists and analysts, and lay people, with varying attention to balancing panels for members' known positions on the assessment issues. In most instances, the conference is held over a two-to-three day period, although others have multiple meetings over longer periods of time. Programs generally provide for part or all of the conference to be held in a public forum. Some programs also conduct evaluations of their programs on such matters as impact of conference findings and panelist satisfaction with the process (Ferguson 2001; Thamer 1998).

In general, the advantages of consensus development processes are that they:

- Provide a focus for assembling experts on an assessment topic
- Provide a means for participation of lay people
- Are relatively inexpensive and less time-consuming compared to new primary data collection
- Provide a good way to call public attention to HTA
- Increase exposure of participants and the public to relevant evidence
- Can prompt face-to-face, evidence-based resolution of opposing viewpoints
- Can apply expert judgment in areas where data are insufficient or inconclusive

In general, the disadvantages of consensus development processes are that they:

- Do not generate new scientific evidence
- May appear to offer veracity to viewpoints that are not supported by evidence
- May over-emphasize or inhibit viewpoints depending upon the stature or personalities of the participants
- May be structured to force or give the appearance of group agreement when it does not exist
- Are difficult to validate

Various reports have made recommendations concerning how to strengthen consensus development programs in particular or in general (Goodman 1990; Institute of Medicine 1990; Olsen 1995). A synopsis of these recommendations is shown in **Box 17**.

Consensus development programs are not immune to the economic, political, and social forces that often serve as barriers or threats to evidence-based processes. Organizations that sponsor consensus development conferences may do so because they have certain expectations for the findings of these processes, and may find themselves at odds with evidence-based findings. Other stakeholders, including from industry, biomedical research institutions, health professions, patient groups, and politicians seeking to align themselves with certain groups, may seek to pressure consensus development panelists or even denounce a panel's findings in order to render desired results, the evidence notwithstanding. An infamous instance of such political reaction occurred in connection with the NIH Consensus Development Conference on Breast Cancer Screening for Women Ages 40 to 49, sponsored by the National Cancer Institute in 1997 (Fletcher 1997).

Box 17. Strategies for Better Consensus Development Programs

1. Programs, or their sponsoring organizations, should have the ability to disseminate and/or implement their consensus findings and recommendations.
2. For each assessment, programs and/or panels should identify the intended audiences and means for achieving intended impact of consensus reports.
3. Programs should describe their scope of interest and/or responsibility, including their purposes, topics, and technological properties or impacts of concerns for the program in general and for specific assessments.
4. Programs should conduct assessments and provide reports in a timely fashion, including the timeliness of assessments relative to the selected topics and the timeframe for planning, conducting, and reporting of assessments.
5. Programs should document the procedures and criteria for selecting conference topics and panel members.
6. The topic and scope of each assessment should be specific and manageable, i.e., commensurate with the available evidence, time, and other resources.
7. Panelists should represent the relevant health professions, methodologists such as epidemiologists and biostatisticians, economists, administrators, patient or other consumer representatives, and others who can provide relevant perspectives. Chairpersons should be recognized as objective with regard to consensus topics and skilled in group processes.
8. Programs should compile the available evidence concerning the assessment topics, and provide a systematic compilation or synthesis of this to panelists prior to the conference.
9. Programs should provide basic guidance concerning the interpretation of evidence, to help ensure that all panelists can be involved in this activity, regardless of their formal expertise in this area.
10. The consensus development processes should be structured and documented, including, e.g., advance identification of key questions/issues, operational definition of consensus, systematically organized evidence, opportunity for equitable participation of panelists, and duration and spacing of sessions to facilitate panelists' full and alert participation.
11. Consensus reports should include at least: description of the consensus process used, notations regarding the strength of agreement or assurance of the panel's findings, description of the reasoning used by the panel and the evidential basis for its findings, recommendations for research needed to address unresolved issues and otherwise advance understanding of the topic.
12. Programs should monitor new developments that might justify reassessments.
13. Programs should provide for periodic, independent evaluation of the program and its impacts.

Adapted from: Goodman 1990.

IV. COST ANALYSIS METHODS

Studies of costs and related economic implications comprise a major group of methods used in HTA. These studies can involve attributes of either or both of primary data collection and integrative methods. That is, cost data can be collected as part of RCTs and other clinical studies, as well as administrative databases used in health care payment. Cost data from one or more such sources often are combined with data from primary clinical studies, epidemiological studies, and other sources to conduct cost-effectiveness analyses and other cost studies that involve weighing health and economic impacts of health technology.

Interest in cost analyses has accompanied concerns about rising health care costs, pressures on health care policymakers to allocate resources, and the need for health product makers and other technology advocates to demonstrate the economic benefits of their technologies. This interest is reflected in a considerable increase in the number of reports of cost analyses in the literature and further refinement of methods.

Main Types of Cost Analysis

There is a variety of approaches to cost analysis, the suitability of any of which depends upon the purpose of an assessment and the availability of data and other resources. It is rarely possible or necessary to identify and quantify all costs and all benefits (or outcomes), and the units used to quantify these may differ.

Main types of cost analysis include the following.

- *Cost-of-illness analysis:* a determination of the economic impact of an illness or condition (typically on a given population, region, or country) e.g., of smoking, arthritis or bedsores, including associated treatment costs
- *Cost-minimization analysis:* a determination of the least costly among alternative interventions that are assumed to produce equivalent outcomes
- *Cost-effectiveness analysis (CEA):* a comparison of costs in monetary units with outcomes in quantitative non-monetary units, e.g., reduced mortality or morbidity
- *Cost-utility analysis (CUA):* a form of cost-effectiveness analysis that compares costs in monetary units with outcomes in terms of their utility, usually to the patient, measured, e.g., in QALYs
- *Cost-consequence analysis:* a form of cost-effectiveness analysis that presents costs and outcomes in discrete categories, without aggregating or weighting them
- *Cost-benefit analysis (CBA):* compares costs and benefits, both of which are quantified in common monetary units.

Box 18 contrasts the valuation of costs and outcomes among these alternative economic analyses.

Box 18. Different Types of Economic Analysis

	Valuation of costs		Valuation of outcomes
Cost of Illness	$	vs.	None
Cost Minimization	$	vs.	Assume same
Cost Effectiveness	$	÷	Natural units
Cost Utility	$	÷	Utiles (e.g., QALYs)
Cost Benefit	$	÷ or -	$

Cost-minimization analysis, CEA and CUA necessarily involve comparisons of alternative interventions. A technology cannot be simply cost effective, though it may be cost effective compared to something else. Although CBA typically involves comparisons of alternative technologies, this is not necessary.

Because it measures costs and outcomes in monetary (not disease-specific) terms, CBA enables comparison of disparate technologies, e.g., coronary artery bypass graft surgery and screening for breast cancer. A drawback of CBA is the difficulty of assigning monetary values to all pertinent outcomes, including changes in the length or quality of human life. CEA avoids this limitation by using more direct or natural units of outcomes such as lives saved or strokes averted. As such, CEA can only compare technologies whose outcomes are measured in the same units. In CUA, estimates of utility are assigned to health outcomes, enabling comparisons of disparate technologies.

Two basic approaches for cost-benefit analysis (CBA) are ratio approach and the net benefit approach. The ratio approach indicates the amount of benefits (or outcomes) that can be realized per unit expenditure on a technology vs. a comparator. In the ratio approach, a technology is cost beneficial vs. a comparator if the ratio of the change in costs to the change in benefits is less than one. The net benefits approach indicates the absolute amount of money saved or lost due to a use of a technology vs. a comparator. In the net benefits formulation, a technology is cost-beneficial vs. a comparator if the net change in benefits exceeds the net change in costs. The choice between a net benefits approach or a benefit/cost approach for a CBA can affect findings. The approach selected may depend upon such factors as whether costs must be limited to a certain level, whether the intent is to maximize the absolute level of benefits, whether the intent is to minimize the cost/benefit ratio regardless of the absolute level of costs, etc. Indeed, under certain circumstances these two basic approaches may yield different preferences among alternative technologies.

Box 19 shows basic formulas for determining CEA, CUA, and CBA.

Box 19. Basic Formulas for CEA, CUA, and CBA

Int: Intervention; Comp: Comparator

Cost-Effectiveness Ratio:

$$\text{CE Ratio} = \frac{\$\text{Cost}_{\text{Int}} - \$\text{Cost}_{\text{Comp}}}{\$\text{Effect}_{\text{Int}} - \$\text{Effect}_{\text{Comp}}}$$

For example: "$45,000 per life-year saved" or "$10,000 per lung cancer case averted"

Cost-Utility Ratio:

$$\text{CU Ratio} = \frac{\$\text{Cost}_{\text{Int}} - \$\text{Cost}_{\text{Comp}}}{\$\text{Utile}_{\text{Int}} - \$\text{Utile}_{\text{Comp}}}$$

Utiles, units of utility or preference, are often measured in QALYs. So, for example: "$45,000 per life-year saved" or "$10,000 per lung cancer case averted"

Cost-Benefit, Ratio Approach:

$$\text{CE Ratio} = \frac{\$\text{Cost}_{\text{Int}} - \$\text{Cost}_{\text{Comp}}}{\$\text{Benefit}_{\text{Int}} - \$\text{Benefit}_{\text{Comp}}}$$

For example: "Cost-benefit ratio of 1.5"

Cost-Benefit, Net Benefit Approach:

$$\text{CB Net} = (\$\text{Cost}_{\text{Int}} - \$\text{Cost}_{\text{Comp}}) - (\$\text{Benefit}_{\text{Int}} - \$\text{Benefit}_{\text{Comp}})$$

For example: "Net cost of $5,000."

Quadrants of Cost-Effectiveness

A basic approach to portraying a cost-effectiveness (or cost-utility) comparison of a new intervention to a standard of care is to consider the cost and effectiveness of a new intervention in the space of four fields as shown in **Box 20**, starting with the upper figure. The level of costs and the level of effectiveness for the standard of care are indicated by the "X" in the middle of the figure. A new intervention may have higher or lower costs, and higher or lower effectiveness, such that its plot may fall into one of the four quadrants surrounding the costs and effectiveness of the standard of care. If it is known that the plot of the new intervention falls into either of two of the quadrants, i.e., where the new intervention has higher costs and lower effectiveness (indicating that it should be rejected), or it has lower costs and higher effectiveness (indicating that it should be adopted), then no further analysis may be required. If it is known that the plot of the new intervention falls into either of the other two quadrants, i.e., where the new intervention has higher costs and higher effectiveness, or it has lower costs and lower effectiveness, then further analysis weighing the marginal costs and effectiveness of the new intervention compared to the standard of care may be required.

Within either of the two quadrants that entail weighing tradeoffs of costs and effectiveness, it may be apparent that the marginal tradeoff of costs and outcomes is so high or low as to suggest rejection or adoption. As shown in the lower figure of **Box 20**, this arises when the new intervention yields only very low marginal gain in effectiveness at a very high marginal cost (reject), or yields very high marginal improvements in effectiveness at a very low marginal cost (adopt).

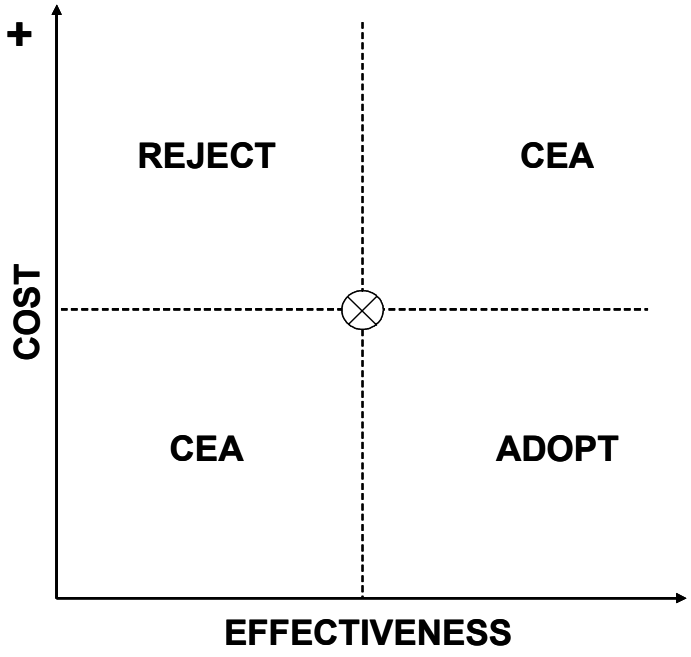
Key Attributes of Cost Analyses

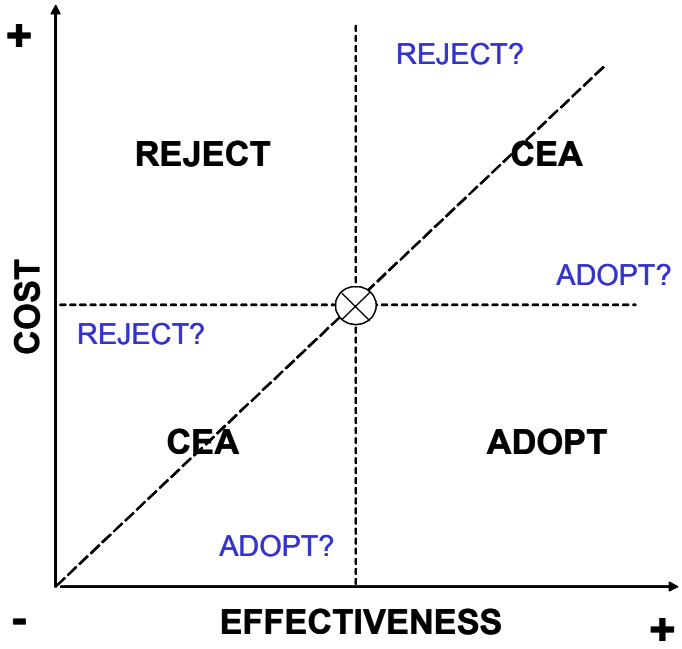
The approaches to accounting for costs and outcomes in cost analyses can vary in a number of important respects, some of which are addressed briefly below. These should be carefully considered by assessors, as well as the policymakers who intend to make use of assessment findings. Given the different ways in which costs and outcomes may be determined, all studies should make clear their methodology in these respects (Byford 1998; Drummond 1997; Gold 1996).

Comparator. Any cost analysis of one intervention versus another must be specific about the comparator. This may be standard of care (current best practice), minimum practice, or no intervention. Some analyses that declare the superiority of a new intervention may have used a comparator that is no longer in practice or is considered sub-standard care or that is not appropriate for the patient population of interest.

Perspective. The perspective of a cost analysis refers to the standpoint at which costs and outcomes (or consequences or benefits) are realized. For instance, the perspective of an analysis may be that of society overall, a third-party payer, a physician, a hospital, or a patient. Clearly, costs and outcomes are not realized in the same way from each of these perspectives. Many analysts favor using the broad perspective of society and identifying all costs and all outcomes accordingly. However, "society" as such may not be the decisionmaker, and what is cost effective from that perspective may not be what is cost effective from the standpoint of a ministry of health, third-party payer, hospital manager, patient, or other decisionmaker. It is possible that this perspective may resemble that of a national or regional

Box 20. Quadrants of Cost-Effectiveness





government, if indeed that government experiences (or is responsible for representing the perspectives of those that experience) all of the costs and outcomes that are included in a societal perspective.

Direct Costs. Depending upon the perspective taken, cost analyses should identify two types of **direct costs**. Direct costs represent the value of all goods, services, and other resources consumed in providing health care or dealing with side effects or other current and future consequences of health care. Two types of direct costs are direct health care costs and direct non-health care costs.

Direct health care costs include costs of physician services, hospital services, drugs, etc. involved in delivery of health care. Direct non-health care costs are incurred in connection with health care, such as for care provided by family members and transportation to and from the site of care. In quantifying direct health care costs, many analyses use readily available hospital or physician *charges* (i.e., price lists) rather than true *costs,* whose determination may require special analyses of resource consumption. However, charges (as well as actual payments) tend to reflect provider cost shifting and other factors that decrease the validity of using charges to represent the true costs of providing care.

Indirect Costs. Analyses should account for **indirect costs**, sometimes known as "productivity losses." These include the costs of lost work due to absenteeism or early retirement, impaired productivity at work, and lost or impaired leisure activity. Indirect costs also include the costs of premature mortality. **Intangible costs** of pain, suffering, and grief are real, yet very difficult to measure and are often omitted from cost analyses.

Time Horizon. Interpretation of cost analyses must consider that the time horizon (or time-frame) of a study is likely to affect the findings regarding the relative magnitudes of costs and outcomes of a health care intervention. Costs and outcomes usually do not accrue in steady streams over time. Comparisons of costs and outcomes after one year may yield much different findings than comparisons made after 5, 10, or 25 years. The meaningful time horizons for assessing the cost horizons of each of emergency appendectomies, cholesterol-lowering in high-risk adults, and smoking cessation in teenagers are likely to be quite different. For example, an analysis conducted for the Medicare program in the US to determine cost and time tradeoffs of hemodialysis and kidney transplantation showed that the annualized expenditure by the Medicare End-Stage Renal Disease Program for a dialysis patient was $32,000. Although patients with functioning transplanted kidneys required a first-year expenditure of $56,000, they cost Medicare only an average of $6,400 in succeeding years. On average, estimated cumulative dialysis and transplantation costs reach a break-even point in about three years, after which transplantation provides a net financial gain compared to dialysis (Rettig 1991).

Time horizons should be long enough to capture streams of health and economic outcomes (including significant intended and unintended ones). These could encompass a disease episode, patient life, or even multiple generations of life (such as for interventions in women of child-bearing age or interventions that may cause heritable genetic changes). Quantitative modeling approaches may be needed to estimate costs and outcomes that are beyond those of available data. Of course, the higher the discount rate used in an analysis, the less important are future outcomes and costs.

Average Costs vs. Marginal Costs. Assessments should make clear whether *average costs* or **marginal costs** are being used in the analysis. Whereas average cost analysis considers the total (or absolute) costs and outcomes of an intervention, marginal cost analysis considers how outcomes change with changes in costs (e.g., relative to a comparator), which may provide more information about how to

use resources efficiently. Marginal cost analysis may reveal that, beyond a certain level of spending, the additional benefits are no longer worth the additional costs. For example, as shown in **Box 21**, the average cost per desired outcome of an iterative screening test may appear to be quite acceptable (e.g., $2,451 per case of colorectal cancer detected assuming a total of six tests per person), whereas marginal cost analysis demonstrates that the cost of adding the last test (i.e., the additional cost of the sixth test per person) to detect another case of cancer would be astronomical.

Box 21. Average Cost Analysis vs. Marginal Cost Analysis

The importance of determining marginal costs is apparent in the analysis by Neuhauser and Lewicki of a proposed protocol of sequential stool guaiac testing for colon cancer. Here, average cost figures obscure a steep rise in marginal costs of testing because the high detection rate from the initial tests is averaged over subsequent tests that contribute little to the detection rate. This type of analysis helps to demonstrate how it is possible to spend steeply increasing health care resources for diminishing returns in health benefits.

Cancer screening and detection costs with sequential guaiac tests

No. of tests	No. of cancers detected	Additional cancers detected	Total cost ($) of diagnosis	Additional ($) cost of diagnosis	Average cost ($) per cancer detected	Marginal cost ($) per cancer detected
1	65.9469	65.9469	77,511	77,511	1,175	1,175
2	71.4424	5.4956	107,690	30,179	1,507	5,492
3	71.9004	0.4580	130,199	22,509	1,810	49,150
4	71.9385	0.0382	148,116	17,917	2,059	469,534
5	71.9417	0.0032	163,141	15,024	2,268	4,724,695
6	71.9420	0.0003	176,331	13,190	2,451	47,107,214

This analysis assumed that there were 72 true cancer cases per 10,000 population. The testing protocol provided six stool guaiac tests per person to detect colon cancer. If any one of the six tests was positive, a barium-enema test was performed, which was assumed to yield no false-positive and no false-negative results. Other assumptions: the true-positive cancer detection rate of any single guaiac test was 91.667%; the false-positive rate of any single guaiac test was 36.508%; the cost of the first stool guaiac test was $4 and each subsequent guaiac test was $1; the cost of a barium-enema was $100. The marginal cost per case detected depends on the population screened and the sensitivity of the test used.

Source: Neuhauser 1975.

Discounting. Cost analyses should account for the effect of the passage of time on the value of costs and outcomes. Costs and outcomes that occur in the future usually have less present value than costs and outcomes realized today. Discounting reflects the time preference for benefits earlier rather than

later; it also reflects the opportunity costs of capital, i.e., whatever returns on investment that could have been gained if resources had been invested elsewhere. Thus, costs and outcomes should be *discounted* relative to their present value (e.g., at a rate of five percent per year). Discounting allows comparisons involving costs and benefits that flow differently over time. It is less relevant for "pay as you go" benefits, such as if all costs and benefits are realized together within one year. It is more relevant in instances where these do not occur in parallel, such as when most costs are realized early and most benefits are realized in later years. Discount rates used in cost analyses are typically based on interest rates of government bonds or the market interest rates for the cost of capital whose maturity is about the same as the duration of the effective time horizon of the health care intervention of program being evaluated. **Box 22** shows the basic formula for calculating present values for a given discount rate, as well as how the present value of a cost or benefit that is discounted at selected rates is affected over time.

Cost analyses should also correct for the effects of *inflation* (which is different from the time preference accounted for by discounting), such as when costs or cost-effectiveness for one year are compared to another year.

Sensitivity Analysis. Any estimate of costs, outcomes, and other variables used in a cost analysis is subject to some uncertainty. Therefore, **sensitivity analysis** should be performed to determine if plausible variations in the estimates of certain variables thought to be subject to significant uncertainty affect the results of the cost analysis. A sensitivity analysis may reveal, for example, that including indirect costs, or assuming the use of generic as opposed to brand name drugs in a medical therapy, or using a plausible higher discount rate in an analysis changes the cost-effectiveness of one intervention compared to another.

Collecting Cost Data Alongside Clinical Studies

The validity of a cost-related study depends upon the sources of the data for costs and outcomes. Increased attention is being given to collection of cost data in more rigorous, prospective studies, particularly RCTs. The closer integration of economic and clinical studies raises important methodological issues. In order to promote more rational diffusion of new technologies, it would be desirable to generate reliable cost and outcomes data during the early part of a technology's lifecycle, such as during RCTs required prior to marketing approval. An RCT would be expected to yield the most reliable data concerning efficacy of an intervention; however, the care given in an RCT and the costs of providing it may be atypical compared to more general settings. For example, RCTs may involve more extensive and frequent laboratory tests and other patient monitoring, and may occur more often in academic medical centers whose costs tend to be higher than in community health care institutions. Other aspects of trial design, sample size, choice of outcome measures, identification and tabulation of costs, burden on investigators of data collection and related matters affect the usefulness of clinical trial data for meaningful economic studies (Briggs 2003; Drummond 1991; Graves 2002; Poe 1995). Also, the growth of multinational clinical trials of drugs and other technologies raises challenges of estimating country-specific treatment effects and cost-effectiveness, given differences in epidemiological factors, health care delivery models, resource use, and other factors (Willke 1998).

Box 22. Discount Rate Calculation and Use in Determining Present Value of Future Costs and Benefits

Discount rate calculation: compiling the discounted stream of costs (or benefits) over time

$$P = \sum_{n=1}^{n} \frac{Fn}{(1+r)^n}$$

P = present value

F = future cost (or benefits) at year n

r = annual discount rate

Present value (P) of future cost (F) occurring at year n at selected annual discount rate (r)

	Discount Rate		
Year	3%	5%	10%
1	0.97	0.95	0.91
5	0.86	0.78	0.62
25	0.48	0.30	0.09
50	0.23	0.09	0.009

For example, the present value of a cost (or benefit) of $1,000 occurring:

- 1 year in the future, using 10% discount rate, is $910
- 5 years in the future, using 3% discount rate, is $860
- 50 yrs in the future, using 5% discount rate, is $90

In practice, there is wide variation in economic study methodologies (Elixhauser 1998; Nixon 2000). Although some variation is unavoidable, many differences in perspective, accounting for direct and indirect costs, time frames, discounting and other aspects are often arbitrary, result from lack of expertise, and may reflect biases on the part of investigators or sponsors. This diminishes comparability and transferability of study results as well as credibility of findings. National and international groups have developed and revised voluntary standards for conducting and reporting economic studies of health care technologies (Drummond 1996; Glennie 1999; Gold 1996; Taylor 2002). A recent review of 25 guidelines from North America, Europe, and Australia found a general trend toward harmonization in most methodological aspects, although there were more differences in such dimensions as choice of economic perspective, resources, and costs to be included in analysis (Hjelmgren 2001).

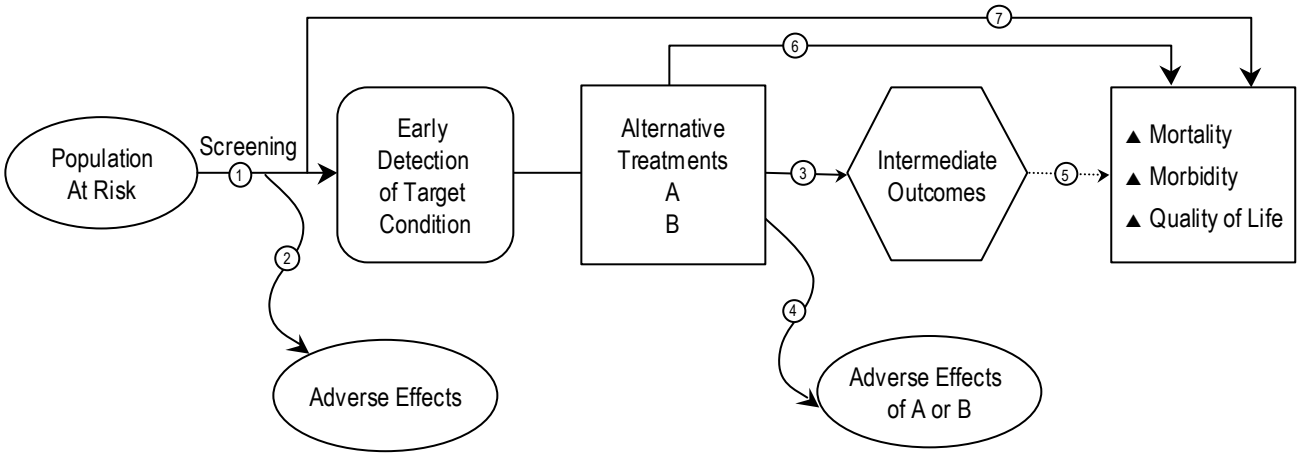
V. APPRAISING THE EVIDENCE

A challenge for any HTA is to derive substantial findings from scientific evidence drawn from different types of studies of varying quality. Assessors should use a systematic approach to critically appraise the quality of the available studies.

Interpreting evidence requires knowledge of investigative methods and statistics. Assessment groups should include members who are knowledgeable in these areas. Some assessment programs assign content experts and evidence evaluation experts to prepare background papers that present and appraise the available evidence for use by assessment groups. Notwithstanding the expertise required to thoroughly and accurately assess evidence, even a basic understanding of fundamental evidence principles can help decision makers to appreciate the importance to health practice and policy of distinguishing between stronger and weaker evidence.

As suggested by the causal pathway in **Box 23**, assessors can interpret evidence at multiple levels. Evidence can be interpreted at the level of an individual study, e.g., an RCT pertaining to a particular intervention and outcome. It also can be interpreted at the level of a body of evidence (e.g., set of clinical studies) pertaining to the intervention and outcome. In some instances, evidence can be interpreted for a broader body of evidence for a linked set of interventions as a whole, such as for a screening test linked to results that are linked to one or more treatments with intermediate and long-term outcomes (Harris 2001). For example, the main criteria for judging evidence quality at each of these levels by the US Preventive Services Task Force are shown in **Box 24**.

Box 23. A General Causal Pathway: Screening Procedure and Alternative Treatments



1. Is screening test accurate for target condition?
2. Does screening result in adverse effects?
3. Do treatments change intermediate outcomes?
4. Do treatments result in adverse effects?
5. Are changes in intermediate outcomes associated with changes in health outcomes?
6. Does treatment improve health outcomes?
7. Is there direct evidence that screening improves health outcomes?

Source: Adapted from Harris 2001.

Box 24. Evaluating Evidence Quality at Three Levels

Level of Evidence	Criteria for Judging Quality
Individual study	- Internal validity[a] - External validity[b]
Linkage in analytic framework	- Aggregate internal validity[a] - Aggregate external validity[b] - Coherence/consistency
Entire preventive service	- Quality of the evidence from Stratum 2 for each linkage in the analytic framework - Degree to which there is a complete chain of linkages supported by adequate evidence to connect the preventive service to health outcomes - Degree to which the complete chain of linkages "fit" together[c] - Degree to which the evidence connecting the preventive service and health outcomes is "direct"[d]

[a] Internal validity is the degree to which the study(ies) provides valid evidence for the population and setting in which it was conducted.

[b] External validity is the extent to which the evidence is relevant and generalizable to the population and conditions of typical primary care practice.

[c] "Fit" refers to the degree to which the linkages refer to the same population and conditions. For example, if studies of a screening linkage identify people who are different from those involved in studies of the treatment linkage, the linkages are not supported by evidence that "fits" together.

[d] "Directness" of evidence is inversely proportional to the number of bodies of evidence required to make the connection between the preventive service and health outcomes. Evidence is direct when a single body of evidence makes the connection, and more indirect if two or more bodies of evidence are required.

Source: Harris 2001.

Appraising Individual Studies

Certain attributes of primary studies produce better evidence than others. In general, the following attributes of primary studies can be used to distinguish between stronger and weaker evidence for internal validity (i.e., for accurately representing the causal relationship between an intervention and an outcome in the particular circumstances of a study).

- Prospective studies are superior to retrospective studies.
- Experimental study designs are superior to observational study designs.
- Controlled studies are superior to uncontrolled ones.
- Contemporaneous (occurring at the same time) control groups are superior to historical control groups.

- Internal control groups (i.e., managed within the study) are superior to studies with external control groups.
- Randomized studies are superior to nonrandomized ones.
- Large studies (i.e., involving enough patients to detect with acceptable confidence levels any true treatment effects) are superior to small studies.
- Blinded studies (in which patients, and clinicians and data analysts where possible, do not know which intervention is being used) are superior to unblinded studies.
- Studies that clearly define patient populations, interventions, and outcome measures are superior to those that do not clearly define these parameters.

Basic types of methods for generating new data on the effects of health care technology in humans include the following.

- Large randomized controlled trial (RCT)
- Small RCT
- Nonrandomized trial with contemporaneous controls
- Nonrandomized trial with **historical control**s
- **Cohort study**
- **Case-control study**
- **Cross-sectional study**
- Surveillance (e.g., using **databases, registers**, or surveys)
- **Series** of consecutive cases
- Single case report (anecdote)

Consistent with the attributes of stronger evidence noted above, these methods are listed in rough order of most to least scientifically rigorous for internal validity. This ordering of methods assumes that each study is properly designed and conducted. This list is representative; there are other variations of these study designs and some investigators use different terminology for certain methods. The demand for studies of higher methodological rigor is increasing among health care technology regulators, payers, providers and other policymakers.

It is not only the basic type of a study design (e.g., RCT or case-control study) that affects the quality of the evidence, but the way in which the study was designed and conducted. There are systematic ways to evaluate the quality of individual studies. In particular, there are numerous approaches for assessing studies of health care interventions, particularly RCTs (Schulz 1995, Jadad 1996). Such approaches typically use one of three main approaches: component, checklist, and scale assessment (Moher, Jadad 1996), for example, as shown in **Box 25** and **Box 26**. Available research indicates that the more complex scales do not seem to produce more reliable assessments of the validity or "quality" of a study (Juni 1999).

Box 25. Basic Checklist for Reviewing Reports of Randomized Controlled Trials

Did the trial:

Yes	No	
		1. Specify outcome measures (endpoints) prior to the trial?
__	__	2. Provide patient inclusion/exclusion criteria?
__	__	3. Specify a-level for defining statistical significance?
__	__	4. Specify b-level (power) to detect a treatment effect of a given meaningful magnitude?
__	__	5. Make a prior estimate of required sample size (to satisfy levels of a and b)?
__	__	6. Use a proper method for random allocation of patients to treatment and control groups?
__	__	7. Use blinding (where possible):
__	__	a. in the randomization process?
__	__	b. for patients regarding their treatment?
__	__	c. for observers/care givers regarding treatment?
__	__	d. in collecting outcome data?
__	__	8. State the numbers of patients assigned to the respective treatment and control groups?
__	__	9. Clearly describe treatment and control (including placebo where applicable)?
__	__	10. Account for patient compliance with treatments/regimens?
__	__	11. Account for all events used as primary outcomes?
__	__	12. Account for patient withdrawals/losses to follow-up?
__	__	13. Analyze patient withdrawals/losses to follow-up
__	__	a. by intention-to-treat?
__	__	b. by treatment actually received?
__	__	14. Account for treatment complications/side effects?
__	__	15. Provide test statistics (e.g., *F, t, Z, chi-square*) and *P* values for endpoints?
__	__	16. Provide confidence intervals or confidence distributions?
__	__	17. Discuss whether power was sufficient for negative trials?
__	__	18. Interpret retrospective analyses (post hoc examination of subgroups and additional endpoints not identified prior to trial) appropriately?

Source: Goodman 1993.

Box 26. Jadad Instrument to Assess the Quality of RCT Reports

This is not the same as being asked to review a paper. It should not take more than 10 minutes to score a report and there are no right or wrong answers.

Please read the article and try to answer the following questions (see attached instructions):

1) Was the study described as randomized (this includes the use of words such as randomly, random, and randomization)?
2) Was the study described as double blind?
3) Was there a description of withdrawals and dropouts?

Scoring the items:
Either give a score of 1 point for each "yes" or 0 points for each "no." There are no in-between marks.

Give 1 additional point if: For question 1, the method to generate the sequence of randomization was described **and** it was **appropriate** (table of random numbers, computer generated, etc.)

and/or: If for question 2, the method of double blinding was described and it was appropriate (identical placebo, active placebo, dummy, etc.)

Deduct 1 point if: For question 1, the method to generate the sequence of randomization was described and it was inappropriate (patients were allocated alternately, or according to date of birth, hospital number, etc.)

and/or: for question 2, the study was described as double blind but the method of blinding was inappropriate (e.g., comparison of tablet vs. injection with no double dummy)

Guidelines for Assessment

1. **Randomization:** A method to generate the sequence of randomization will be regarded as appropriate if it allowed each study participant to have the same chance of receiving each intervention and the investigators could not predict which treatment was next. Methods of allocation using date of birth, date of admission, hospital numbers, or alternation should not be regarded as appropriate.

2. **Double blinding:** A study must be regarded as double blind if the word "double blind" is used. The method will be regarded as appropriate if it is stated that neither the person doing the assessments nor the study participant could identify the intervention being assessed, or if in the absence of such a statement the use of active placebos, identical placebos, or dummies is mentioned.

3. **Withdrawals and dropouts:** Participants who were included in the study but did not complete the observation period or who were not included in the analysis must be described. The number **and** the reasons for withdrawal in each group must be stated. If there were no withdrawals, it should be stated in the article. If there is no statement on withdrawals, this item must be given no points.

Source: Jadad 1996.

The criteria used for assessing quality of studies vary by type of design. For example, the internal validity of an RCT depends on such methodological criteria as: method of randomization, accounting for withdrawals and dropouts, and blinding/masking of outcomes assessment. The internal validity of systematic reviews (discussed below) depends on such methodological criteria as: time period covered by the review, comprehensiveness of the sources and search strategy used, relevance of included studies to the review topic, and application of a standard appraisal of included studies.

The ability of analysts to determine the internal and external validity of a published study and to otherwise interpret its quality depends on how thoroughly and clearly the information about its study's design, conduct, statistical analysis, and other aspects are reported. The inadequate quality of a high proportion of published reports of RCTs, even in leading journals, has been well documented (Freiman 1978; Moher 1994). Several national and international groups of researchers and medical journal editors have developed standards for reporting of RCTs and other studies (Moher 2001; International Committee of Medical Journal Editors 1997). The trend of more journals to require structured abstracts has assisted analysts in identifying and screening reports of RCTs and other studies.

Many primary studies of health care technologies involve small, non-randomized series of consecutive cases or single case reports, and therefore have methodological limitations that make it difficult to establish the efficacy (or other attributes) of the technologies with sound scientific validity. To some extent, these methodological shortcomings are unavoidable given the nature of the technologies being evaluated, or are otherwise beyond the control of the investigators. In the instance of determining the efficacy of a new drug, the methodological standard is a large, prospective, double-blind, placebo-controlled RCT. These methodological attributes increase the chances of detecting any real treatment effect of the new drug, control for patient characteristics that might influence any treatment effect, and reduce opportunities for investigator or patient bias to affect results.

Although their contributions to methodological validity are generally well recognized, it is not possible to apply all of these attributes for trials of certain types of technologies or for certain clinical indications or settings. Further, these attributes are controversial in certain instances. Patient and/or investigator blinding is impractical or impossible for many medical devices and most surgical procedures. For clinical trials of technologies for rare diseases (e.g., "orphan drugs" and devices), it may be difficult to recruit numbers of patients large enough to detect convincing treatment effects.

Among the various areas of methodological controversy in clinical trials is the appropriate use of **placebo** controls. Issues include: (1) appropriateness of using a placebo in a trial of a new therapy when a therapy judged to be effective already exists, (2) statistical requirements for discerning what may be smaller differences in outcomes between a new therapy and an existing one compared to differences in outcomes between a new therapy and a placebo, and (3) concerns about comparing a new treatment to an existing therapy that, except during the trial itself, may be unavailable in a given setting (e.g., a developing country) because of its cost or other economic or social constraints (Rothman 1994; Varmus 1997). As in other health technologies, surgical procedures can be subject to the **placebo effect**. In recent years, following previous missteps that raised profound ethical concerns, guidance has emerged for using "sham" procedures as placebos in RCTs of surgical procedures (Horng 2003). Some instances of patient blinding have been most revealing about the placebo effect in surgery, including arthroscopic knee surgery (Moseley 2002), percutaneous myocardial laser revascularization (Stone 2002), and neurotransplantation surgery (Boer 2002).

Notwithstanding the limitations inherent in clinical study of many technologies, the methodological rigor used in many primary studies falls short of what it could be. Clinicians, patients, payers, hospital managers, national policymakers, and others who make technology-related decisions and policies are becoming more sophisticated in demanding and interpreting the strength of scientifically-based findings.

Decide How to Use Studies

Most assessment groups have decided that it is not appropriate to consider all studies equally important, and that studies of higher quality should influence their findings more than studies of lesser quality. Experts in evidence interpretation do not agree on the proper approach for deciding how to use studies of differing quality. According to some experts, the results of studies that do not have randomized controls are subject to such great bias that they should not be included for determining the effects of an intervention. Others say that studies from nonrandomized prospective studies, observational studies, and other weaker designs should be used, but given less weight or adjusted for their biases.

There are several basic approaches to deciding how to use the individual studies in an assessment. These are: use all studies as reported; decide whether to include or exclude each study as reported; weight studies according to their relative quality; and make adjustments to the results of studies to compensate for their biases. Each approach has advantages and disadvantages, as well as differing technical requirements. As noted below with regard to establishing search strategies, the approaches to determining what types of studies to be used in an assessment should be determined prospectively as much as possible, so as to avoid injecting selection bias into study selection. Therefore, to the extent that assessors decide to use only certain types of studies (e.g., RCTs and systematic reviews) or not to use certain types of studies (e.g., case studies, case series, and other weaker designs), they should set their inclusion and exclusion criteria prospectively and design their literature search strategies accordingly. Assessment reports should document the criteria or procedures by which they chose to make use of study results for use in the assessment.

Appraising a Body of Evidence

As described above, certain attributes of primary study designs produce better evidence than others. A useful step in appraising evidence is to classify it by basic design type and other study characteristics.

Evidence tables provide a useful way to summarize and display important qualities about multiple individual studies pertaining to a given question. The information summarized in evidence tables may include attributes of study design (e.g., **randomization**, control, blinding, patient characteristics (e.g., number, age, gender), patient outcomes (e.g., mortality, morbidity, HRQL) and derived summary statistics (e.g., ***P* values, confidence intervals**). The tabular format enables reviewers to compare systematically the key attributes of studies and to provide an overall picture of the amount and quality of the available evidence. **Box 27** is an evidence table of selected study characteristics and outcomes of double-blind placebo-controlled RCTs of aspirin for patients after myocardial infarction.

"Grading" a body of evidence according to its methodological rigor is a standard part of HTA. It can take various forms, each of which involves structured, critical appraisal of the evidence against formal criteria (RTI International-University of North Carolina 2002). **Box 28** shows an evidence hierarchy that ranks study types from "well-designed randomized controlled trials" at the top through "opinions of respected authorities based on clinical experience" and similar types of expert views at the bottom. **Box**

Box 27. Evidence Table: Double-Blind Placebo-Controlled RCTs of Aspirin in Patients After Myocardial Infarction

Trial, year	No. patients randomized	Age range (mean)	Male %	Months from qualif. event to trial entry	Daily dose ASA[1] (mg)	Follow-up (years)	Average Mortality (%)[2]	Sum. stat.[3]	Cardiac death (%)	Sum. stat.	Nonfatal MI (%)	Sum. stat.
AMIS 1980	ASA: 2,267 plac:2,257	30-69 (54.8)	89	2 - 60	1,000	3.2	10.8 9.7	Z=1.27	8.7 8.0	Z=0.82	7.7 9.5	Z=-2.11
Breddin 1980	ASA: 317 plac: 309	45-70	78	1 - 1.4	1,500	2.0	8.5 10.4	Z=-0.79	1.6 3.2		3.5 4.9	
CDPR 1976	ASA: 758	62%≥55yrs 61%≥55yrs	100	74% ≥ 60 77% ≥ 60	972	1.8	5.8 8.3	Z=-1.90	5.4 7.8	Z=-1.87	3.7 4.2	Z=-0.46
Elwood 1974	ASA: 615 plac: 624	57%≥55yrs 54%≥55yrs (55.0)	100	76% ≤ 3	300	1.0	7.6 9.8	not sig.	-		-	
Elwood 1979	ASA: 832 plac: 850	(56.0)	85	50% ≤ 0.25	900	1.0	12.3 14.8	not sig. at P<0.05	-		-	
PARIS 1980	ASA: 810 plac: 406	30-74	87	2 - 60	972	3.4	10.5 12.8	Z=-1.21	8.0 10.1	Z=-1.24	6.9 9.9	not sig.

[1] ASA: aspirin (acetylsalicylic acid); plac: placebo

[2] Percent of mortality, cardiac death, and nonfatal myocardial infarction based on number of patients randomized.

[3] Sum. stat.: summary statistic. Z is a statistical test that can be used to determine whether the difference in proportions or means between a treatment group and a control group are statistically significant. For a two-tailed test, Z values of $\pm$1.96 and $\pm$2.58 are approximately equivalent to P values of 0.05 and 0.01.

Sources: Aspirin Myocardial Infarction Study Research Group 1980; Breddin 1980; The Coronary Drug Project Research Group 1976; Elwood 1979; Elwood 1974; Elwood 1983; The Persantine-Aspirin Reinfarction Study Research Group 1980.

Box 28. UK NHS Centre for Reviews and Dissemination: Hierarchy of Evidence

Level	Description
I	Well-designed randomized controlled trials
II-1a	Well-designed controlled trial with pseudo-randomization
II-1b	Well-designed controlled trials with no randomization
II-2a	Well-designed cohort (prospective) study with concurrent controls
II-2b	Well-designed cohort (prospective) study with historical controls
II-2c	Well-designed cohort (retrospective) study with concurrent controls
II-3	Well-designed case-control (retrospective) study
III	Large differences from comparisons between times and/or places with and without intervention
	(in some circumstances these may be equivalent to level II or I)
IV	Opinions of respected authorities based on clinical experience; descriptive studies; reports of expert committees

Source: NHS Centre for Reviews and Dissemination 1996.

29 shows a basic evidence-grading scheme that has been used by the US Preventive Services Task Force. This scheme grades evidence in a manner that favors certain attributes of stronger studies for primary data, beginning with properly-designed RCTs. In order to better address how well studies are conducted, the task force augmented this hierarchy with a three-category rating of the internal validity of each study, shown in **Box 30**.

Another type of evidence table, shown in **Box 31**, has a count of articles published during a given time period, arranged by type of study, about the use of percutaneous transluminal coronary angioplasty. Rather than showing details about individual studies, this evidence table shows that the distribution of types of studies in an apparently large body of evidence included a relatively small number of RCTs, and a large number of less rigorous observational studies.

Assessment groups can classify studies in evidence tables to gain an understanding of the distribution of evidence by type, and apply evidence hierarchies such as those shown above to summarize a body of evidence. However, more information may be needed to characterize the evidence in a useful way. For example, more detailed grading schemes can be used to account for instances where two or more well-designed studies have conflicting (heterogeneous) results. **Box 32** distinguishes between groups of studies with homogeneous and heterogeneous results. This hierarchy also recognizes as stronger evidence studies with low probabilities of false positive error (α) and false negative error (β). This hierarchy also distinguishes between bodies of evidence depending on whether high-quality overviews (i.e., systematic reviews or meta-analyses) are available.

Box 29. US Preventive Services Task Force: Hierarchy of Research Design

I: Evidence obtained from at least one properly-designed randomized controlled trial.

II-1: Evidence obtained from well designed controlled trials without randomization.

II-2: Evidence obtained from well designed cohort or case-controlled analytic studies, preferably from more than one center or research group.

II-3: Evidence obtained from multiple time series with or without the intervention. Dramatic results in uncontrolled experiments (such as the results of the introduction of penicillin treatment in the 1940s) could also be regarded as this type of evidence.

III: Opinions of respected authorities, based on clinical experience, descriptive studies, or reports of expert committees.

Source: Harris 2001.

Box 30. US Preventive Services Task Force: Grades for Strength of Overall Evidence

Grade	Definition
Good:	Evidence includes consistent results from well-designed, well-conducted studies in representative populations that directly assess effects on health outcomes
Fair:	Evidence is sufficient to determine effects on health outcomes, but the strength of the evidence is limited by the number, quality, or consistency of the individual studies; generalizability to routine practices; or indirect nature of the evidence on health outcomes
Poor:	Evidence is insufficient to assess the effects on health outcomes because of limited number or power of studies, important flaws in their design or conduct, gaps in the chain of evidence, or lack of information on important health outcomes.

Source: U.S. Preventive Services Task Force 2002.

Box 31. Distribution of Research Articles on PTCA by Year of Publication and Method Used to Collect or Review Data

Article Class	1980	‘81	‘82	‘83	‘84	‘85	‘86	‘87	‘88	‘89	‘90	Total
Prospective RCT	0	0	0	0	1	1	2	4	2	1	2	13
Prospective non-RCT	0	0	1	3	4	5	5	6	11	8	3	46
Prospective registry	0	0	2	4	13	2	1	1	3	4	7	37
Case-control & adjusted cohort	0	0	1	2	0	0	2	2	4	5	2	18
Observational	1	1	1	3	12	12	12	27	25	29	8	131
Survey	0	0	0	0	0	0	0	0	1	0	1	2
Editorial	0	0	0	1	2	3	1	4	2	4	5	22
Review	0	0	0	2	3	4	4	5	16	14	11	59
Cross-sectional	0	0	0	0	0	0	0	2	1	0	0	3
Decision analysis	0	0	0	0	0	0	0	0	0	0	1	1
Total	1	1	5	15	35	27	27	51	65	65	40	332

Articles were retrieved using *MEDLINE* searches.

Source: Hilborne 1991.

Box 32. Levels of Evidence and Grades of Recommendations

If No Overview Available	If High-Quality Overview Available	
Level of Evidence	Level of Evidence	**Grade of Recommendation**
I: Randomized trials with low false-positive (α) and low false negative (β) errors.	Lower limits of CI for treatment effect *exceeds* clinically significant benefit and: I+: Individual study results homogeneous I-: Individual study results heterogeneous	A
II: Randomized trials with high false-positive (α) and high false negative (β) errors.	Lower limit of CI for treatment effects falls below clinically significant benefit and: II+: Individual study results homogeneous II-: Individual study results heterogeneous	B
III: Nonrandomized concurrent cohort studies		C
IV: Nonrandomized historical cohort studies		
V: Case series		

Source: Cook 1992.

The more comprehensive evidence hierarchy from the UK NHS Centre for Evidence Based Medicine, shown in **Box 33**, provides levels of evidence (1a-c, 2a-c, etc.) to accompany findings based on evidence derived from various study designs and applications in prevention, therapy, diagnosis, economic analysis, etc.

Of course, HTAs may involve multiple questions about the use of a technology, e.g., pertaining to particular patient populations or health care settings. Therefore, the evidence and recommendations applying to each question may be evaluated separately or at different levels, as suggested in the causal pathway shown in **Box 23**.

Link Recommendations to Evidence

Findings and recommendations should be linked explicitly to the quality of the evidence. The process of interpreting and integrating the evidence helps assessment groups to determine the adequacy of the evidence for addressing aspects of their assessment problems (Hayward 1995).

An example of linking recommendations to evidence is incorporated into the evidence appraisal scheme cited above in **Box 32**, which assigns three grade levels to recommendations based on the evidence. Accompanying the grades for evidence (as shown in **Box 30**), the US Preventive Services Task Force provides grades for recommendations based on the evidence. This approach, shown in **Box 34**, reflects two dimensions: the direction of the recommendation (e.g., for or against providing a preventive service) and the strength of the recommendation, tied to the grade of evidence (e.g., a strong recommendation if there is good evidence). Finally, the comprehensive evidence hierarchy shown in **Box 33** also includes grades of recommendation that are linked to levels of evidence, including levels that account for evidence homogeneity and heterogeneity.

Even for those aspects of an assessment problem for which there is little useful evidence, an assessment group may have to provide some type of findings or recommendations. This may involve making inferences from the limited evidence, extrapolations of evidence from one circumstance to another, theory, or other subjective judgments. Whether a recommendation about using a technology in particular circumstances is positive, negative, or equivocal (neutral), users of the assessment should understand the basis of that recommendation and with what level of confidence it was made. Unfortunately, the recommendations made in many assessment reports do not reflect the relative strength of the evidence upon which they are based. In these instances, readers may have the mistaken impression that all of the recommendations in an assessment report are equally valid or authoritative.

Approaches for linking the quality of available evidence to the strength and direction of findings and recommendations are being improved and new ones are being developed (Harbour 2001). Using evidence this way enables readers to better understand the reasoning behind the assessment findings and recommendations. It also provides readers with a more substantive basis upon which to challenge the assessment as appropriate. Further, it helps assessment programs and policymakers to determine if a reassessment is needed as relevant new evidence becomes available.

Box 33. Oxford Centre for Evidence-based Medicine Levels of Evidence (May 2001)

Level	Therapy/Prevention, Aetiology/Harm	Prognosis	Diagnosis	Differential diagnosis/symptom prevalence study	Economic and decision analyses
1a	SR (with homogeneity*) of RCTs	SR (with homogeneity*) of inception cohort studies; CDR† validated in different populations	SR (with homogeneity*) of Level 1 diagnostic studies; CDR† with 1b studies from different clinical centres	SR (with homogeneity*) of prospective cohort studies	SR (with homogeneity*) of Level 1 economic studies
1b	Individual RCT (with narrow Confidence Interval‡)	Individual inception cohort study with ≥ 80% follow-up; CDR† validated in a single population	Validating** cohort study with good††† reference standards; or CDR† tested within one clinical centre	Prospective cohort study with good follow-up****	Analysis based on clinically sensible costs or alternatives; systematic review(s) of the evidence; and including multi-way sensitivity analyses
1c	All or none§	All or none case-series	Absolute SpPins and SnNouts††	All or none case-series	Absolute better-value or worse-value analyses ††††
2a	SR (with homogeneity*) of cohort studies	SR (with homogeneity*) of either retrospective cohort studies or untreated control groups in RCTs	SR (with homogeneity*) of Level >2 diagnostic studies	SR (with homogeneity*) of 2b and better studies	SR (with homogeneity*) of Level >2 economic studies
2b	Individual cohort study (including low quality RCT; e.g., <80% follow-up)	Retrospective cohort study or follow-up of untreated control patients in an RCT; Derivation of CDR† or validated on split-sample§§§ only	Exploratory** cohort study with good†††reference standards; CDR† after derivation, or validated only on split-sample§§§ or databases	Retrospective cohort study, or poor follow-up	Analysis based on clinically sensible costs or alternatives; limited review(s) of the evidence, or single studies; and including multi-way sensitivity analyses
2c	"Outcomes" Research; Ecological studies	"Outcomes" Research		Ecological studies	Audit or outcomes research
3a	SR (with homogeneity*) of case-control studies		SR (with homogeneity*) of 3b and better studies	SR (with homogeneity*) of 3b and better studies	SR (with homogeneity*) of 3b and better studies
3b	Individual Case-Control Study		Non-consecutive study; or without consistently applied reference standards	Non-consecutive cohort study, or very limited population	Analysis based on limited alternatives or costs, poor quality estimates of data, but including sensitivity analyses incorporating clinically sensible variations.
4	Case-series (and poor quality cohort and case-control studies§§)	Case-series (and poor quality prognostic cohort studies***)	Case-control study, poor or non-independent reference standard	Case-series or superseded reference standards	Analysis with no sensitivity analysis
5	Expert opinion without explicit critical appraisal, or based on physiology, bench research or "first principles"	Expert opinion without explicit critical appraisal, or based on physiology, bench research or "first principles"	Expert opinion without explicit critical appraisal, or based on physiology, bench research or "first principles"	Expert opinion without explicit critical appraisal, or based on physiology, bench research or "first principles"	Expert opinion without explicit critical appraisal, or based on economic theory or "first principles"

Box 33 (cont.). Oxford Centre for Evidence-Based Medicine Levels of Evidence (May 2001)

Notes

Users can add a minus-sign "-" to denote the level of that fails to provide a conclusive answer because of:

- EITHER a single result with a wide Confidence Interval (such that, for example, an ARR in an RCT is not statistically significant but whose confidence intervals fail to exclude clinically important benefit or harm)
- OR a Systematic Review with troublesome (and statistically significant) heterogeneity.
- Such evidence is inconclusive, and therefore can only generate Grade D recommendations.

*	By homogeneity we mean a systematic review that is free of worrisome variations (heterogeneity) in the directions and degrees of results between individual studies. Not all systematic reviews with statistically significant heterogeneity need be worrisome, and not all worrisome heterogeneity need be statistically significant. As noted above, studies displaying worrisome heterogeneity should be tagged with a "-" at the end of their designated level.
†	Clinical Decision Rule. (These are algorithms or scoring systems which lead to a prognostic estimation or a diagnostic category.)
‡	See note #2 for advice on how to understand, rate and use trials or other studies with wide confidence intervals.
§	Met when all patients died before the Rx became available, but some now survive on it; or when some patients died before the Rx became available, but none now die on it.
§§	By poor quality cohort study we mean one that failed to clearly define comparison groups and/or failed to measure exposures and outcomes in the same (preferably blinded), objective way in both exposed and non-exposed individuals and/or failed to identify or appropriately control known confounders and/or failed to carry out a sufficiently long and complete follow-up of patients. By poor quality case-control study we mean one that failed to clearly define comparison groups and/or failed to measure exposures and outcomes in the same (preferably blinded), objective way in both cases and controls and/or failed to identify or appropriately control known confounders.
§§§	Split-sample validation is achieved by collecting all the information in a single tranche, then artificially dividing this into "derivation" and "validation" samples.
††	An "Absolute SpPin" is a diagnostic finding whose Specificity is so high that a Positive result rules-in the diagnosis. An "Absolute SnNout" is a diagnostic finding whose Sensitivity is so high that a Negative result rules-out the diagnosis.
‡‡	Good, better, bad and worse refer to the comparisons between treatments in terms of their clinical risks and benefits.
†††	Good reference standards are independent of the test, and applied blindly or objectively to applied to all patients. Poor reference standards are haphazardly applied, but still independent of the test. Use of a non-independent reference standard (where the 'test' is included in the 'reference', or where the 'testing' affects the 'reference') implies a level 4 study.
††††	Better-value treatments are clearly as good but cheaper, or better at the same or reduced cost. Worse-value treatments are as good and more expensive, or worse and the equally or more expensive.
**	Validating studies test the quality of a specific diagnostic test, based on prior evidence. An exploratory study collects information and trawls the data (e.g. using a regression analysis) to find which factors are 'significant'.
***	By poor quality prognostic cohort study we mean one in which sampling was biased in favour of patients who already had the target outcome, or the measurement of outcomes was accomplished in <80% of study patients, or outcomes were determined in an unblinded, non-objective way, or there was no correction for confounding factors.
****	Good follow-up in a differential diagnosis study is >80%, with adequate time for alternative diagnoses to emerge (eg 1-6 months acute, 1 - 5 years chronic)

Grades of Recommendation

A	consistent level 1 studies
B	consistent level 2 or 3 studies ***or*** extrapolations from level 1 studies
C	level 4 studies ***or*** extrapolations from level 2 or 3 studies
D	level 5 evidence ***or*** troublingly inconsistent or inconclusive studies of any level

Source: Center for Evidence-Based Medicine 2003.

Box 34. US Preventive Services Task Force: Grades for Strength of Recommendations

Grade	Recommendation
A	The USPSTF strongly recommends that clinicians routinely provide [the service] to eligible patients. *The USPSTF found good evidence that [the service] improves important health outcomes and concludes that benefits substantially outweigh harms.*
B	The USPSTF recommends that clinicians routinely provide [the service] to eligible patients. *The USPSTF found at least fair evidence that [the service] improves important health outcomes and concludes that benefits outweigh harms.*
C	The USPSTF makes no recommendation for or against routine provision of [the service]. *The USPSTF found at least fair evidence that [the service] can improve health outcomes but concludes that the balance of benefits and harms is too close to justify a general recommendation.*
D	The USPSTF recommends against routinely providing [the service] to asymptomatic patients. The USPSTF found at least fair evidence that [the service] is ineffective or that harms outweigh benefits.
I	The USPSTF concludes that the evidence is insufficient to recommend for or against routinely providing [the service]. *Evidence that [the service] is effective is lacking,, of poor quality, or conflicting, and the balance of benefits and harms cannot be determined.*

Source: U.S. Preventive Services Task Force 2002.

Assessment organizations and others that review evidence are increasingly providing guidance to technology sponsors and other stakeholders for preparing dossiers and other submissions of clinical and economic evidence. For example, the UK National Institute for Clinical Excellence (NICE) provides guidance to technology manufacturers and sponsors for preparing submissions of evidence to inform NICE technology appraisals (National Institute for Clinical Excellence 2001). The Academy of Managed Care Pharmacy (AMCP) provides a recommended format for submission of clinical and economic data in support of formulary consideration by pharmacy and therapeutics committees of health plans in the US (Academy of Managed Care Pharmacy 2002).

VI. DETERMINING TOPICS FOR HTA

Organizations that conduct or sponsor HTAs have only limited resources for this activity. With the great variety of potential assessment topics, HTA organizations need some practical means of determining what to assess. This section considers how assessment programs identify candidate assessment topics and set priorities among these.

Identify Candidate Topics

To a large extent, assessment topics are determined or bounded, by the mission or purpose of an organization. For example, the US FDA [http://www.fda.gov] is systematically required to assess all new drugs and to assess health devices according to specific provisions made for particular classes of devices. For a new drug, a company normally files an **Investigational New Drug Application (IND)** with the FDA for permission to begin testing the drug in people; later, following successful completion of necessary clinical trials, the company files a **New Drug Application (NDA)** to seek FDA approval to market the drug. For certain medical devices (i.e., new "Class III" devices that sustain or support life, are implanted in the body, or present a potential risk of illness or injury), the **Investigational Device Exemption (IDE)** and **Premarketing Approval (PMA) Application** are analogous to the IND and NDA, respectively. The FDA is notified about many other devices when a company files a "510(k)" application seeking market approval based on a device's "substantial equivalence" to another device that has already received FDA marketing approval.

Third-party payers generally assess technologies on a reactive basis; a new medical or surgical procedure that is not recognized by payers as being standard or established may become a candidate for assessment. For the US Centers for Medicare and Medicaid Services (CMS), assessment topics arise in the form of requests for national coverage policy determinations that cannot be resolved at the local level or that are recognized to be of national interest. These requests typically originate with Medicare contractors that administer the program in their respective regions, Medicare beneficiaries (patients), physicians, health product companies, health professional associations, and government entities. CMS may request assistance in the form of evidence reports or other assessments by a sister agency, AHRQ.

For the Evidence-based Practice Centers program, also administered by AHRQ, the agency solicits topic nominations for evidence reports and technology assessments in a public notice in the US *Federal Register*. Topics have been nominated by a variety of other government agencies, payers, health systems and networks, health professions associations, employer and consumer groups, disease-based organizations, and others. In selecting topics, AHRQ considers not only the information about the topic itself, but the plans of the nominating organization to make use of the findings of the assessment. Information required in these nominations is shown in **Box 35**.

The American College of Physicians (ACP) Clinical Efficacy Assessment Program (CEAP), which develops clinical practice guidelines, determines its guideline topics based upon evidence reports developed by the AHRQ Evidence-based Practice Centers (EPC) program. (Topics of the EPC program are nominated by outside groups, including ACP.) The topics undertaken by ECRI's technology assessment service are identified by request of the service's subscribers, including payers, providers, and others. For the Cochrane Collaboration, potential topics generally arise from members of the review

groups, who are encouraged to investigate topics of interest to them, subject to the agreement of their review groups (Clarke 2003).

Box 35. Evidence-based Practice Centers Topic Nominations

Topic nominations for the AHRQ EPC program should include:

- Defined condition and target population
- Three to five very focused questions to be answered
- Incidence or prevalence, and indication of disease burden (e.g., mortality, morbidity, functional impairment) in the US general population or in subpopulations (e.g., Medicare and Medicaid populations)
- Costs associated with the conditions, including average reimbursed amounts for diagnostic and therapeutic interventions
- Impact potential of the evidence report or technology assessment to decrease health care costs or to improve health status or clinical outcomes
- Availability of scientific data and bibliographies of studies on the topic
- References to significant differences in practice patterns and/or results; alternative therapies or controversies
- Plans of the nominating organization to incorporate the report into its managerial or policy decision making (e.g., practice guidelines, coverage policies)
- Plans of the nominating organization for dissemination of these derivative products to its membership
- Process by which the nominating organization will measure members' use of the derivative products
- Process by which the nominating organization will measure the impact of such use on clinical practice

Source: Agency for Healthcare Research and Quality 2003.

Horizon Scanning

The demand for scanning of multiple types of sources for information about new health care interventions has prompted the development of "early warning" or "horizon scanning" functions in the US, Europe, and elsewhere (Douw 2003). Horizon scanning functions are intended to serve multiple purposes, including to:

- Identify potential topics for HTA and information for setting priorities among these
- Clarify expectations for the uses or indications of a technology
- Increase public awareness about new technologies

- Estimate the expected health and economic impacts
- Identify critical thresholds of effectiveness improvements in relation to additional costs, e.g., to demonstrate the cost-effectiveness of a new intervention
- Anticipate potential social, ethical, or legal implications of a technology (Harper 1998; Stevens 1998; Carlsson 1998).

Among the organizations with horizon scanning functions are:

- Canadian Coordinating Office on Health Technology Assessment (CCOHTA) Emerging Technologies Program (CETAP) [www.ccohta.ca]
- ECRI: TARGET™ (TA Resource Guide for Emerging Technologies) and Health Technology Forecast [www.ecri.org]
- National Horizon Scanning Centre (NHSC), UK [www.publichealth.bham.ac.uk/horizon/]
- EuroScan (secretariat at NHSC, UK) [www.publichealth.bham.ac.uk/euroscan/]

For example, CETAP draws its information from the Internet, published literature, CCOHTA committee members, and other experts. The products of CETAP include short *Alerts* that address very early technologies, and as more evidence becomes available, CCOHTA publishes more in-depth, peer-reviewed *Issues in Emerging Health Technologies* bulletins. The purposes of EuroScan (European Information Network on New and Changing Health Technologies), a collaborative network of more than a dozen HTA agencies, are to: evaluate and exchange information on new and changing technologies, develop information sources, develop applied methods for early assessment, and disseminate information on early identification and assessment activities.

As shown in **Box 36**, a considerable variety of online databases, newsletters, and other sources provide streams of information pertaining to new and emerging health care interventions. Certainly, an important set of sources for identifying new topics are bibliographic databases such as MEDLINE (accessible, e.g., via PubMed) and EMBASE. The Cochrane Collaboration protocols are publicly available, detailed descriptions of systematic reviews currently underway by Cochrane, which include detailed descriptions of the rationale for the review, information sources, and search strategies.

Although the major thrust of horizon scanning has been to identify "rising" technologies that eventually may merit assessment, horizon scanning may turn to the other direction to identify "setting" technologies that may be outmoded or superseded by newer ones. In either case, horizon scanning provides an important input into setting assessment priorities.

Setting Assessment Priorities

Some assessment programs have explicit procedures for setting priorities; others set priorities only in an informal or vague way. Given very limited resources for assessment and increasing accountability of assessment programs to their parent organizations and others who use or are affected by their assessments, it is important to articulate how assessment topics are chosen.

Box 36. Information Sources for New and Emerging Health Care Interventions

- Trade journals (e.g., *F-D-C Reports: The Pink Sheet, NDA Pipeline, The Gray Sheet; In Vivo; Adis International; Biomedical Instrumentation and Technology; R&Directions)*
- General news *(PR Newswire, Reuters Health, New York Times)*
- Health professions and industry newsletters (e.g., *Medscape, Medicine & Health, American Health Line, CCH Health & Medicine)*
- Conferences (and proceedings) of medical specialty societies and health industry groups
- General medical journals and specialty medical journals
- Technology company web sites
- Publicly available market research reports (IHS Health Group)
- FDA announcements of market approvals of new pharmaceuticals (e.g., NDAs, NDA supplements), biotechnologies (e.g., BLAs), and devices (e.g., PMAs, PMA supplements, and 510[k]s)*
- Adverse event/alert announcements (from FDA, USP, NIH Clinical Alerts and Advisories, etc.)
- New Medicines in Development (disease- and population-specific series from PhRMA, including clinical trial status)
- Databases of ongoing research, e.g., Clinicaltrials.gov and HSRProj (Health Services Research Projects in Progress) from NLM
- Reports and other sources of information on significant variations in practice, utilization, or payment policies (e.g., *The Dartmouth Atlas,* LMRP.NET)
- Special reports on health care trends and futures (e.g., *Health and Health Care 2010* (Institute for the Future 2000); *Health Technology Forecast* (ECRI 2002)
- Priority lists and forthcoming assessments from public and non-profit evaluation/assessment organizations (e.g., INAHTA member organizations)
- Cochrane Collaboration protocols

*NDA: New Drug Application approvals; BLA: Biologics License Application approvals; PMA: Premarket Approval Application approvals; 510(k): substantially equivalent device application approvals.

Most assessment programs have criteria for topic selection, although these criteria are not always explicit. Is it most important to focus on costly health problems and technologies? What about health problems that affect large numbers of people, or health problems that are life-threatening? What about technologies that cause great public controversy? Should an assessment be undertaken if it is unlikely that its findings will change current practice? Examples of selection criteria that are used in setting assessment priorities are:

- High individual burden of morbidity, mortality, or disability
- High population burden of morbidity, mortality, or disability
- High unit cost of a technology or health problem
- High aggregate cost of a technology or health problem
- Substantial variations in practice
- Available findings not well disseminated or adopted by practitioners
- Need to make regulatory decision
- Need to make a health program implementation decision (e.g., for initiating a major immunization program)
- Need to make payment decision (e.g., provide coverage or include in health benefits)
- Scientific controversy or great interest among health professionals
- Public or political demand
- Sufficient research findings available upon which to base assessment
- Timing of assessment relative to available evidence (e.g., recent or anticipated pivotal scientific findings)
- Potential for the findings of an assessment to be adopted in practice
- Potential for change in practice to affect patient outcomes or costs
- Feasibility given resource constraints (funding, time, etc.) of the assessment program

The timing for undertaking an assessment may be sensitive to the availability of evidence. For example, the results of a recently completed RCT or meta-analysis may challenge standard practice, and prompt an HTA to consolidate these results with other available evidence for informing clinical or payment decisions. Or, an assessment may be delayed pending the results of an ongoing study that has the potential to shift the weight of the body of evidence on that topic.

A systematic priority-setting process could include the following steps (Donaldson and Sox 1992; Lara and Goodman 1990).

1. Select criteria to be used in priority setting.
2. Assign relative weights to the criteria.
3. Identify candidate topics for assessment (e.g., as described above).

4. If the list of candidate topics is large, reduce it by eliminating those topics that would clearly not rank highly according to the priority setting criteria.
5. Obtain data for rating the topics according to the criteria.
6. For each topic, assign a score for each criterion.
7. Calculate a priority score for each topic.
8. Rank the topics according to their priority scores.
9. Review the priority topics to ensure that assessment of these would be consistent with the organizational purpose.

Processes for ranking assessment priorities range from being highly subjective (e.g., informal opinion of a small group of experts) to quantitative (e.g., using a mathematical formula) (Donaldson 1992; Eddy 1989; Phelps 1992). **Box 37** shows a quantitative model for priority setting. The Cochrane Collaboration uses a more decentralized approach. Starting with topics suggested by their review group members, many Cochrane Collaboration review groups set priorities by considering burden of disease and other criteria, as well as input from discussions with key stakeholders and suggestions from consumers. These priorities are then offered to potential reviewers who might be interested in preparing and maintaining relevant reviews in these areas (Clarke 2003).

Of course, there is no single correct way to set priorities. The great diversity of potential assessment topics, the urgency of some policymaking needs, and other factors may diminish the practical benefits of using highly systematic and quantitative approaches. On the other hand, ad hoc, inconsistent, or non-transparent processes are subject to challenges and skepticism of policymakers and other observers who are affected by HTA findings. Certainly, there is a gap between theory and application of priority setting. Many of the priority setting models are designed to support resource allocation that maximizes health gains, i.e., identify health interventions which, if properly assessed and appropriately used, could result in substantial health improvements at reasonable costs. However, some potential weaknesses of these approaches are that they tend to set priorities among interventions rather than the assessments that should be conducted, they do not address priority setting in the context of a research portfolio, and they do not adopt an incremental perspective (i.e., consideration of the net difference that conducting an assessment might accomplish) (Sassi 2003).

Reviewing the process by which an assessment program sets its priorities, including the implicit and explicit criteria it uses in determining whether or not to undertake an assessment, can help to ensure that the HTA program is fulfilling its purposes effectively and efficiently.

Specify the Assessment Problem

One of the most important aspects of an HTA is to specify clearly the problem(s) or question(s) to be addressed; this will affect all subsequent aspects of the assessment. An assessment group should have an explicit understanding of the purpose of the assessment and who the intended users of the assessment are to be. This understanding might not be established at the outset of the assessment; it may take more probing, discussion and clarification.

Box 37. A Quantitative Model for Priority Setting

A 1992 report by the Institute of Medicine provided recommendations for priority setting to the Agency for Health Care Policy and Research (now AHRQ). Seven criteria were identified:

- Prevalence of a health condition
- Burden of illness
- Cost
- Variation in rates of use
- Potential of results to change health outcomes
- Potential of results to change costs
- Potential of results to inform ethical, legal, or social issues

The report offered the following formula for calculating a priority score for each candidate topic.

$$\text{Priority Score} = W_1 \ln S_1 + W_2 \ln S_2 + \ldots W_7 \ln S_7$$

where:

W is the relative weight of each of seven priority-setting criteria

S is the score of a given candidate topic for a criterion

ln is the natural logarithm of the criterion scores.

Candidate topics would then be ranked according to their priority score.

Source: Donaldson 1992.

The intended users or target groups of an assessment should affect its content, presentation, and dissemination of results. Clinicians, patients, politicians, researchers, hospital managers, company executives, and others have different interests and levels of expertise. They tend to have different concerns about the effects or impacts of health technologies (health outcomes, costs, social and political effects, etc.). They also have different needs regarding the scientific or technical level of reports, the presentation of evidence and findings, and the format (e.g., length and appearance) of reports.

When the assessment problem and intended users have been specified, they should be reviewed by the requesting agency or sponsors of the HTA. The review of the problem by the assessment program may have clarified or focused the problem in a way that differs from the original request. This clarification may prompt a reconsideration or restatement of the problem before the assessment proceeds.

Problem Elements

There is no single correct way to state an assessment problem. In general, an assessment problem could entail specifying at least the following elements: health care problem(s); patient population(s); technology(ies); practitioners or users; setting(s) of care; and properties (or impacts or health outcomes) to be assessed.

For example, a basic specification of one assessment problem would be:

- **Health care problem:** management of moderate hypertension
- **Patient population:** males and females, age ≥60 years, diastolic blood pressure 90-114 mm Hg, systolic blood pressure <240 mm Hg, no other serious health problems
- **Technologies:** specific types/classes of pharmacologic and nonpharmacologic treatments
- **Practitioners:** primary care providers
- **Setting of care:** outpatient care, self care
- **Properties, impacts, or outcomes:** safety (including side-effects), efficacy, effectiveness and cost-effectiveness (especially cost-utility)

Causal Pathways

A useful means of presenting an assessment problem is a "causal pathway," sometimes known as an "analytical framework." Causal pathways depict direct and indirect linkages between interventions and outcomes. Although often used to present clinical problems, they can be used as well for organizational, financing, and other types of interventions or programs in health care.

Causal pathways provide clarity and explicitness in defining the questions to be addressed in an HTA, and draw attention to pivotal linkages for which evidence may be lacking. They can be useful working tools to formulate or narrow the focus of an assessment problem. For a clinical problem, a causal pathway typically includes a patient population, one or more alternative interventions, intermediate outcomes (e.g., biological markers), health outcomes, and other elements as appropriate. In instances where a topic concerns a single intervention for narrowly defined indications and outcomes, these pathways can be relatively straightforward. However, given the considerable breadth and complexity of some HTA topics, which may cover multiple interventions for broadly defined health problem (e.g., screening, diagnosis, and treatment of osteoporosis in various population groups), causal pathways can become detailed. While the development of a perfectly representative causal pathway is not the objective of an HTA, these can be specified to a level of detail that is sufficient for the sponsor of an HTA and the group that will conduct the HTA concur on the assessment problem. In short, it helps to draw a picture.

An example of a general causal pathway for a screening procedure with alternative treatments is shown in **Box 23**. As suggested in this example, the evidence that is assembled and interpreted for an HTA may be organized according to an indirect relationship (e.g., between a screening test and an ultimate health outcome) as well as various intervening direct causal relationships (e.g., between a treatment indicated by the screening test and a biological marker, such as blood pressure or cholesterol level).

Reassessment and the Moving Target Problem

Health technologies are "moving targets" for assessment (Goodman 1996). As a technology matures, changes occur in the technology itself or other factors that can diminish the currency of an HTA report and its utility for health care policies. As such, HTA can be more of an iterative process than a one-time analysis. Some of the factors that would trigger a reassessment might include changes in the:

- Evidence pertaining to the safety, effectiveness, and other outcomes or impacts of using the technology (e.g., publication of significant new results of a major clinical trial or a new meta-analysis)
- Technology itself (modified techniques, models, formulations, delivery modes, etc.)
- Indications for use (different health problems, degree of severity, etc.)
- Populations in which it is used (different age groups, comorbidities, etc.)
- Protocols or care pathways of which the technology is a part that may alter the role or utility of the technology
- Care setting in which the technology is applied (inpatient, outpatient, physician office, home, long-term care)
- Provider of the technology (type of clinician, other caregiver, patient, etc.)
- Practice patterns (e.g., large practice variations)
- Alternative technology or standard of care to which the technology is compared
- Outcomes or impacts considered to be important (e.g., types of costs or quality of life)
- Resources available for health care or the use of a particular technology (i.e., raising or lowering the threshold for decisions to use the technology)
- Adoption or use of guidelines, payment policies, or other decisions that are based on the HTA report
- Interpretation of existing research findings (e.g., based on corrections or re-analyses).

There are numerous instances of moving targets that have prompted reassessments. For example, since the inception of percutaneous transluminal coronary angioplasty (PTCA, approved by the US FDA in 1980), its clinical role vis-à-vis coronary artery bypass graft surgery (CABG) has changed as the techniques and instrumentation for both technologies have evolved, their indications have expanded, and as competing, complementary, and derivative technologies have emerged (e.g., laser angioplasty, coronary artery stents, minimally-invasive and "beating-heart" CABG). The emergence of viable pharmacological therapy for osteoporosis (e.g., with bisphosphonates and selective estrogen receptor modulators) has increased the clinical utility of bone densitometry. Long rejected for its devastating teratogenic effects, thalidomide has reemerged for carefully managed use in a variety of approved and investigational uses in leprosy and other skin diseases, certain cancers, chronic graft-vs.-host disease, and other conditions (Combe 2001; Richardson 2002).

While HTA programs cannot avoid the moving target problem, they can manage and be responsive to it. **Box 38** lists approaches for managing the moving target problem.

Box 38. Managing the Moving Target Problem

- Recognize that HTA must have the capacity to revisit topics as needed, whether periodically or as prompted by important changes that have transpired since preparation of the original HTA report.
- Document in HTA reports the information sources, assumptions, and processes used. This information baseline will better enable HTA programs and other interested groups to recognize when it is time for reassessment.
- In the manner of a sensitivity analysis, indicate in HTA reports what magnitudes of change in key variables (e.g., accuracy of a diagnostic test, effectiveness of a drug, patient compliance, costs) would result in a significant change in the report findings.
- Note in HTA reports any known ongoing research, work on next-generation technologies, population trends, and other developments that might prompt the need for reassessment.
- Have or subscribe to a scanning or monitoring function to help detect significant changes in technologies and other developments that might trigger a reassessment.
- Recognize that, as the number of technology decision makers increases and evidence-based methods diffuse, multiple assessments are generated at different times from different perspectives. This may diminish the need for clinicians, payers, and other decision makers to rely on a single, definitive assessment on a particular topic.

Aside from changes in technologies and their applications, even new interpretations of, or corrections in, existing evidence can prompt a new assessment. This was highlighted by a 2001 report of a Cochrane Center that prompted the widespread re-examination of screening mammography guidelines by government and clinical groups. The report challenged the validity of evidence indicating that screening for breast cancer reduces mortality, and suggested that breast cancer mortality is a misleading outcome measure (Olsen 2001).

Some research has been conducted on the need to reassess a particular application of HTA findings, i.e., clinical practice guidelines. For example, for a study of the validity of 17 guidelines developed in the 1990s by AHCPR (now AHRQ), investigators developed criteria defining when a guideline needs to be updated, surveyed members of the panels that prepared the respective guidelines, and searched the literature for relevant new evidence published since the appearance of the guidelines. Using a "survival analysis," the investigators determined that about half of the guidelines were outdated in 5.8 years, and that at least 10% of the guidelines were no longer valid by 3.6 years. They recommended that, as a general rule, guidelines should be reexamined for validity every three years (Shekelle, Ortiz 2001). Others counter that the factors that might prompt a reassessment do not arise predictably or at regular intervals (Brownman 2001). Some investigators have proposed models for determining whether a guideline or other evidence-based report should be reassessed (Shekelle, Eccles 2001).

Changes in the volume or nature of publications may trigger the need for an initial assessment or reassessment. A "spike" (sharp increase) in publications on a topic, such as in the number of research reports or commentary, may signal trends that would merit attention for assessment. However, in order to determine whether such publication events are reliable indicators of technology emergence or moving

targets requiring assessment, further bibliometric research should be conducted to determine whether actual emergence of new technologies or substantial changes in them or their use has been correlated with such publication events or trends (Mowatt 1997).

Not all changes require conducting a reassessment, or that a reassessment should entail a full HTA. A reassessment may require updating only certain aspects of an original report. In some instances, current clinical practices or policies may be recognized as being optimal relative to available evidence, so that a new assessment would have little potential for impact; or the set of clinical alternatives and questions have evolved so much since the original assessment that it would not be relevant to update it.

In some instances, an HTA program may recognize that it should withdraw an existing assessment because to maintain it could be misleading to users and perhaps even have adverse health consequences. This may arise, for example, when an important flaw is identified in a pivotal study in the evidence base underlying the assessment, when new research findings appear to refute or contradict the original research base, or when the assumptions used in the assessment are determined to be flawed. The determination to maintain or withdraw the existing assessment while a reassessment is conducted, to withdraw the existing assessment and not conduct a reassessment, or to take other actions, depends on the risks and benefits of these alternative actions for patient health, and any relevant legal implications for the assessment program or users of its assessment reports.

Once an HTA program determines that a report topic is a candidate for being updated, the program should determine the need to undertake a reassessment in light of its other priorities. Assessment programs may consider that candidates for reassessment should be entered into the topic priority-setting process, subject to the same or similar criteria for selecting HTA topics.

VII. RETRIEVING EVIDENCE FOR HTA

One of the great challenges in HTA is to assemble the evidence−the data, literature and other information−that is relevant to a particular assessment. For very new technologies, this information may be sparse and difficult to find; for many technologies, it can be profuse, scattered and of widely varying quality. Literature searching and related evidence retrieval are integral to successful HTA, and the time and resources required for these activities should be carefully considered in planning any HTA (Auston 1994; Goodman 1993).

Types of Sources

Available information sources cover different, though often overlapping, sectors of health care information. Although some are devoted to health care topics, others cover the sciences more broadly. Multiple sources should be searched to increase the likelihood of retrieving relevant reports. The variety of types of sources that may be useful for HTA include:

- Computer databases of published literature
- Computer databases of clinical and administrative data
- Printed indexes and directories
- Government reports and monographs
- Policy and research institute reports
- Professional association reports and guidelines
- Market research reports
- Company reports and press releases
- Reference lists in available studies and reviews
- Special inventories/registers of reports
- Health newsletters and newspapers
- Colleagues and investigators

Of course, the Internet is an extraordinarily broad and readily accessible medium that provides access to many of these information sources.

There are hundreds of publicly available computer databases for health care and biomedical literature. Among these are various general types. For example, **bibliographic databases** have indexed citations for journal articles and other publications. **Factual databases** provide information in the form of guidelines for diagnosis and treatment, patient **indications**, and **contraindications**, and other authoritative information. Referral databases provide information about organizations, services and other information sources.

The National Information Center on Health Services Research & Health Care Technology (NICHSR) [www.nlm.nih.gov/nichsr/nichsr.html] of the US National Library of Medicine (NLM) provides an extensive, organized set of the many, evolving databases, publications, outreach and training, and other information resources for HTA. One online source, *Etext on Health Technology Assessment (HTA) Information Resources* [www.nlm.nih.gov/nichsr/ehta/], is a comprehensive textbook on sources of HTA information and searching approaches compiled by information specialists and researchers from around the world (National Library of Medicine 2003). Various other useful compendia of HTA information resources have been prepared (Busse 2002; Glanville 2003; Chan 2003). Some of the main bibliographic and factual databases useful in HTA are listed in **Box 39**.

The most widely used of these resources for HTA are the large bibliographic databases, particularly MEDLINE, produced by NLM, and EMBASE, produced by Elsevier. MEDLINE can be accessed at the NLM website using PubMed, which also includes new in-process citations (with basic citation information and abstracts before being indexed with MeSH terms and added to MEDLINE), citations from various life science journals, and certain other entries. In addition, there are many specialized or more focused databases in such areas as AIDS, bioethics, cancer treatment, pharmaceutical research and development, ongoing clinical trials (e.g., ClinicalTrials.gov of NLM), and practice guidelines (e.g., National Guideline Clearinghouse of AHRQ).

The Cochrane Collaboration [www.cochrane.org] is an international organization that prepares, maintains and disseminates systematic reviews of RCTs (and other evidence when appropriate) of treatments for many clinical conditions. More than 1,500 systematic reviews have been produced by nearly 50 Cochrane review groups in such areas as acute respiratory infections, breast cancer, diabetes, hypertension, infectious diseases, and pregnancy and childbirth. The Cochrane Collaboration produces the Cochrane Library, which includes databases and registers produced by the Cochrane Collaboration as well as some produced by other organizations. The Database of Abstracts of Reviews and Dissemination (DARE) [agatha.york.ac.uk/darehp.htm] and the NHS Economic Evaluation Database are produced by the NHS Centre for Reviews and Dissemination (NHSCRD). The HTA Database is produced by the International Network of Agencies for Health Technology Assessment (INAHTA) [www.inahta.org], in collaboration with the NHSCRD.

The selection of sources for literature searches should depend on the purpose of the HTA inquiry and pertinent time and resource constraints. Most searches are likely to involve MEDLINE or another large database of biomedical literature (Suarez-Almazor 2000; Topfer 1999). However, the selection of other databases may differ by purpose, e.g., horizon scanning, ascertaining regulatory or payment status of technologies, comprehensive systematic review, or identifying literature in particular clinical areas.

Gray Literature

Much valuable information is available beyond the traditional published sources. This "gray" or "fugitive" literature is found in industry and government monographs, regulatory documents, professional association reports and guidelines, market research reports, policy and research institute studies, spot publications of special panels and commissions, conference proceedings, and other sources. Many of these can be found via the Internet. Although the **gray literature** can be timely and cover aspects of technologies that are not addressed in mainstream sources, it is usually not subject to **peer review**, and must be scrutinized accordingly.

Box 39. Selected Bibliographic and Factual Databases for HTA

Some Core Sources

- *MEDLINE:* citations for biomedical journal articles
- *EMBASE:* citations for biomedical journal articles (Elsevier)
- *Cochrane Database of Systematic Reviews:* systematic reviews of controlled trials on hundreds of clinical topics
- *Cochrane Controlled Trials Register:* bibliography of controlled trials including sources outside peer-reviewed journal literature
- *Database of Abstracts of Reviews of Effectiveness (DARE):* structured abstracts of systematic reviews from around the world, critically appraised by NHS Centre for Reviews and Dissemination
- *NHS Economic Evaluation Database:* abstracts and other information about published economic evaluations of health care interventions
- *Health Technology Assessment Database:* records of ongoing projects of members of INAHTA and completed HTAs by INAHTA members and other organizations
- *National Guideline Clearinghouse:* evidence-based clinical practice guidelines (AHRQ)

Additional Sources

- *Other NLM/NIH sources:*
  - *ClinicalTrials.gov:* current information about current clinical research studies in health services research and behavioral and social sciences
  - *DIRLINE:* directory of organizations
  - *HSRProj:* ongoing health services research projects
  - *HSRR (Health Services/Sciences Research Resources):* research datasets and instruments/indices.
  - *HSTAT:* full text of US clinical practice guidelines, consensus development reports, technology assessment reports, etc.
  - *PDQ:* cancer treatment, supportive care, screening, prevention, clinical trials
  - Other specialized databases such as *AIDSLINE, Bioethics,* and *HealthSTAR* have been incorporated into *MEDLINE,* accessed, e.g., via *PubMed*
- *ACP Journal Club:* selected studies and systematic reviews for immediate attention of clinicians, with "value added" abstracts and commentary
- *AltHealthWatch:* information resources on alternative medicine
- *Bandolier:* journal of evidence summaries
- *Best Evidence* (ACP Journal Club plus *Evidence Based Medicine)*
- *BIOSIS Previews:* citations of life sciences literature (BIOSIS)
- *CEA Registry:* database of standardized cost-utility analyses (Harvard School of Public Health)
- *CINAHL*: citations for nursing and allied health literature (Cinahl Information Systems)
- *CDC Wonder*: gateway to reports and data of the US Centers for Disease Control and Prevention (CDC)
- *Cochrane Methodology Register:* bibliography of articles and books on the science of research synthesis
- *Cochrane Database of Methodology Reviews:* full text of systematic reviews of empirical methodological studies
- *HDA Evidence Base:* summaries of systematic reviews of effectiveness, literature reviews, meta-analyses, expert group reports, and other review-level information (NHS Health Development Agency, UK)
- *MANTIS:* bibliographic database on manual, alternative, and natural therapies
- *Netting the Evidence:* (ScHARR, University of Sheffield, UK)
- *PsycINFO:* citations of psychological literature (American Psychological Association)
- SciSearch*:* citations for scientific journal articles (Institute for Scientific Information)

Publication Bias

Various forms of **bias** can affect the validity of HTA. One reason for careful planning and conduct of search strategies for HTA is minimize, or at least recognize, the effects of **publication bias**. Studies of the composition of the biomedical research literature have found imbalances in the publication of legitimate studies (Chalmers 1990). For instance, positive studies (that find statistically significant treatment effects) are more likely than negative studies (that find no treatment effects) to be published in peer-reviewed journals (Dickersin 1993; Dickersin 1997). A study sponsored by a health product company or other group with an interest in the results may be less likely to be submitted for publication if the findings are not favorable to the interests of that group. RCTs conducted for market approval (e.g., by the US FDA) often are not published (MacLean 2003). Some research indicates that, among published studies of health technologies, smaller studies tend to report positive results more frequently (Agema 2002). Positive studies are more likely to be published in English-language journals, be reported in multiple publications, and be cited in other articles (Easterbrook 1991, Gøtzsche 1989). These multiple appearances and citations increase the likelihood of being identified in literature searches and included in meta-analyses and other systematic reviews, which may introduce bias into the results of these syntheses as well (Sterne 2001). The prevalence of unpublished studies may vary by specialty; for example, oncology appears to have a high prevalence of unpublished studies.

One detailed analysis of the characteristics of clinical trials used in systematic reviews indicated that, compared to other clinical areas, trials in the fields of psychiatry, rheumatology, and orthopedics tend more often to be published in non-English languages and appear in sources not indexed in MEDLINE (Egger 2003). **Time lag bias** occurs when the time from completion of a clinical trial to its publication is affected by the direction (positive vs. negative findings) and strength (statistical significance) of the trial results (Ioannidis 1998).

Certainly, bias in selection of studies used in HTA may arise to the extent that the literature search does not include studies that appear in languages other than English (**language bias**), are not indexed in MEDLINE or other major bibliographic databases, are unpublished, or are of lesser methodological quality. While the validity of an HTA is likely linked to the effort to include an unbiased sample of relevant studies, the size and direction of this relationship varies. There is a growing literature on the extent to which more or less restrictive inclusion criteria for meta-analyses affect their results. For example, some research indicates that systematic reviews limited to the English language literature that is accessible via the major bibliographic databases produces similar or same results to those based on less restricted reviews (Egger 2003). Lowering the standard of methodological quality for inclusion of published studies in an HTA may bias the findings if these studies tend to report positive findings more often that higher-quality studies.

In planning a literature search, assessors should weigh the anticipated quality of a search with time and resource constraints. Efforts to recognize and minimize bias may be further subject to such factors as the availability of studies by language and for particular clinical areas, and their accessibility via bibliographic databases.

Help for Searchers

Given the great number of databases and the variety in their scope, means of access, controlled vocabularies and search commands, it is advisable to consult health information specialists. These experts can be especially helpful when planning which databases to search, inclusion and exclusion criteria, and other aspects of literature searches. An expanding network of HTA information specialists who work with HTA agencies and other evidence-based medicine organizations around the world have formed the HTAi Information Resources Group, which is extending the capabilities, expertise, and collaboration in the field. Improved indexing, text word searching, user-friendly interfaces, more powerful personal computers and other advances in medical informatics are helping non-expert searchers to retrieve valuable information more effectively and efficiently. Indeed, the enhanced ability of all types of assessors to probe these databases provides a more immediate, hands-on understanding of the scope and quality of literature on any given topic.

During the last decade, the NLM has undertaken to improve its **MeSH** (*Me*dical *S*ubject *H*eadings) **controlled vocabulary** (used to index and search literature in MEDLINE and other NLM databases) in the related fields of HTA and health services research. In cooperation with the Cochrane Collaboration and others, NLM has improved the indexing of citations in MEDLINE and other databases to improve identification of RCTs (Dickersin 1994). Most bibliographic and factual databases have user-friendly tutorials, search engines, and other searching tools that are increasingly standard and familiar to expert and non-expert searchers alike. There is a growing number of resources for supporting searching strategies for HTA (Goodman 1993, Sackett 1997). A new resource from the NLM NICHSR, *Etext on Health Technology Assessment (HTA) Information Resources* [www.nlm.nih.gov/nichsr/ehta/], provides extensive guidance and resources for searching in HTA (National Library of Medicine 2003). Particularly instructive and useful for clinicians is the series of articles published in the *Journal of the American Medical Association: Users' Guides to the Medical Literature,* from the Evidence-Based Medicine Working Group (Hunt 2000).

The search for pertinent existing evidence is normally one of the first major tasks of an assessment, and should be planned accordingly. Costs associated with evidence searches can be significant, coming in the form of staff time and acquisition of literature, data tapes, and other documentation. Although access to MEDLINE (e.g., via PubMed) and other public-source databases is generally free of inexpensive, using some specialized scientific and business databases can be more costly. Database vendors offer a variety of packages of databases and pricing algorithms for these. HTA programs of such organizations as ECRI, the Blue Cross and Blue Shield Association, and Hayes sell their reports on a subscription basis. Some market research monographs and other reports oriented for health product companies, investors and other business interests are priced in the thousands of dollars.

VIII. DISSEMINATING FINDINGS AND RECOMMENDATIONS

To the analysts and other experts who have participated in an HTA, the importance of its findings and recommendations may be self-evident. Dissemination of these findings and recommendations, whether for internal use in the same organization or into the national or international health information mainstream, often is considered as an administrative afterthought.

Worthy HTA messages get lost because of misidentified and misunderstood audiences, poor packaging, wrong transmission media, bad timing, and other factors. Although there is some convergence on the format and content of information to be included in an HTA report, much research is needed regarding how to optimize the dissemination of HTA findings and recommendations (Goldberg 1994; Mittman and Siu 1992; Mittman and Tonesk 1992; Busse 2002).

Competing for Attention

Dissemination efforts must compete with the burgeoning flow of health-related information being transmitted across diverse channels using increasingly sophisticated means. Advanced communications technologies provide alternative means to transmit more data where and when it can influence decision makers. Marketing, long practiced effectively by health care product companies, offers an evolving, continually researched variety of techniques that are being adapted throughout the health care sector. As the ground shifts in health care organization, delivery and financing, the cast of decision makers constituting the potential users of HTA changes.

There is considerable current controversy regarding various policies and practices of disseminating information about health technologies, particularly by pharmaceutical and other health technology companies. One area is the use of direct-to-consumer advertising by pharmaceutical and other health technology companies, including whether this is to be permitted at all and, if so, what requirements should pertain to the content and format of the message. In particular, while there is strong evidence that these messages increase awareness of prescription drugs, they is far less evidence that they are effective in educating patients about medications for their conditions (Lyles 2002). A second area of controversy concerns whether health technology companies can distribute published and unpublished reports of clinical trials of their products for indications that have not been cleared for marketing by the appropriate authority, e.g., by the US FDA (Stryer 1996). A third area of controversy concerns the conditions under which pharmaceutical and other health technology companies can make claims in their marketing information about the cost-effectiveness of their products, what the rigor of supporting evidence should be, and which agencies should have regulatory oversight for such economic claims (Neumann 2000).

Dissemination Dimensions

Approaches for disseminating reports of HTAs can be described along three dimensions: target groups (intended audiences), media, and implementation techniques or strategies, as shown in **Box 40**.

The results of the same HTA may be packaged for dissemination in different formats, e.g., for patients, clinicians, payers, and researchers or policy analysts. Reaching the same decisionmaker may require repeated messages and/or multiple media. The style in which an assessment report is written (e.g., an

Box 40. Approaches for HTA Report Dissemination

Target groups

- Clinicians (individuals, specialty/professional organizations)
- Patients/consumers (individuals, organizations)
- Provider organizations (hospitals, clinics, managed care organizations);
- Third party payers (government, private sector)
- Quality assurance and utilization review organizations
- Government policymakers (international, national, state, local)
- Biomedical researchers
- Health care product companies
- News professionals (popular and scientific/professional journalists and editors)
- Educational institutions (schools, continuing professional education programs)

Media

- Printed: direct mail, newspapers and popular journals, scientific/professional journals and newsletters, posters, pocket cards
- Electronic: internet, television, radio, video disks, computer databases (online and disk)
- Word of mouth: informal consultation, formal lectures and presentations, focus groups

Implementation techniques or strategies

- Patient-oriented: mass media campaigns, community based campaigns, interaction with clinicians (including shared decision procedures, interactive video disk), modify insurance coverage (more or less generous benefits, change copayments)
- Clinician-oriented: conferences and workshops; continuing professional education; professional curriculum development; opinion leaders; one-on-one educational visits ("academic detailing"); coverage/reimbursement policy; precertification; mandatory second opinion; drug formulary restrictions; feedback (e.g., on laboratory test ordering relative to criteria/guidelines); reminder systems (e.g., as part of computer-based patient record systems); medical audit/peer review; criteria for board certification/recertification, state licensure, Medicare PRO action, specialty designation, professional/specialty society membership; public availability of performance data (e.g., adjusted mortality rates for certain procedures); defense against sanctions and malpractice action
- Institution-oriented: accreditation, standards (e.g., hospital infection control, clinical laboratories), benchmarking, public availability of performance data

academic, scholarly tone versus a practical, concrete tone) may affect the receptiveness of researchers, practitioners and others (Kahan 1988).

Dissemination Plan

Dissemination should be planned at the outset of an assessment along with other assessment phases or activities. The costs, time and other resources needed for dissemination should be budgeted accordingly. This does not mean that dissemination plans should be rigid; the nature of the findings and recommendations themselves may affect the choice of target groups and the types of messages to be delivered. Dissemination should be designed to influence behavior of decision makers. This is not always straightforward, as research findings concerning what works for HTA dissemination strategies do not point to any universally successful approaches.

Mediating Access

There are many approaches to controlling or enhancing access to assessment reports. As noted above, some assessment programs provide their assessments only to paid subscribers or member organizations, or charge fees intended to help recoup the cost of the assessment or provide a profit. While some assessments are public documents made available at no cost via the internet or in public libraries, others are held as proprietary (e.g., company assessments of new products). Access to assessment literature is also mediated by the capacity of bibliographic organizations (e.g., the NLM and commercial database vendors) to index and abstract the literature, and the availability of such information via online databases and other information services. The wording used by assessment report authors for titles and abstracts can influence the indexing that serves as a key to accessing these reports.

IX. MONITORING IMPACT OF HTA

The impacts of HTAs, from market research reports to RCT reports to expert panel statements, are variable and inconsistently understood. Whereas some HTA reports are translated directly into policies with clear and quantifiable impacts, the findings of some "definitive" RCTs and authoritative, well-documented assessment reports go unheeded or are not readily adopted into general practice (Banta 1993; Ferguson, Dubinsky 1993; Henshall 2002; Institute of Medicine 1985).

As is the case for the technologies that are the subjects of HTA, the reports of HTAs can have intended, direct impacts as well as unintended, indirect ones. Some of the ways in which a HTA report can make an impact (Banta 1993) are:

- Affect corporate investment decisions
- Modify R&D priorities/spending levels
- Change regulatory policy
- Modify marketing of a technology
- Change third-party payment policy
- Affect acquisition or adoption of a new technology
- Change the rate of use of a technology
- Change clinician behavior
- Change patient behavior
- Change the organization or delivery of care
- Reallocate national or regional health care resources

Attributing Impact to HTA Reports

The impact of a HTA depends upon the target groups' legal, contractual, or administrative obligation to comply with it (Anderson 1993; Ferguson, Dubinsky 1993; Gold 1993). FDA market approvals of new drugs and devices are translated directly into binding policy. Most of the HTAs conducted by AHRQ are requested by CMS for use in the Medicare program, although CMS is not obligated to comply with AHRQ findings. The impacts of NIH consensus development conference statements, which are not statements of government policy, are inconsistent and difficult to measure. The ability of NIH statements to change behavior seems to depend upon a variety of factors intrinsic to particular topics, the consensus development process and a multitude of contextual factors (Ferguson 1993; Ferguson 2001).

The task of measuring the impact of HTA can range from elementary to infeasible. Even if an intended change does occur, it may be difficult or impossible to attribute this change to the HTA. A national-level assessment that recommends increased use of a particular intervention for a given clinical problem may be followed by a documented change in behavior consistent with that recommendation. However,

the recommendation may be made at a time when the desired behavior change is already underway, when third-party payment policy is shifting in favor of the technology, during a strong marketing effort by industry, or close to the time of announcement of the results of a convincing clinical trial. Given widespread and nearly instant communications in health care, it may be difficult to control for factors other than a particular HTA report that might influence behavior change.

As is the case for attributing changes in patient outcomes to a technological intervention, the ability to demonstrate that the results of an HTA have an impact depends upon the conditions under which the assessment results were made known and the methodological approach used to determine the impact. Evaluations of the impact of an assessment often are unavoidably observational in nature; however, under some circumstances, quasi-experimental or experimental evaluations are used (Goldberg 1994). To the extent that impact evaluations are prospective, involve pre- and post-dissemination data collection, and involve directed dissemination to clearly identified groups with well-matched controls (or at least retrospective adjustment for reported exposure to dissemination), they are more likely to detect a causal connection between an HTA report and behavior change. Even so, generalizing from one experience to others may be impractical, as it is difficult to describe and replicate the conditions of a particular HTA report dissemination.

Factors Mediating Impact

The factors that can affect the impact of HTA reports are many. Beyond the particular dissemination techniques used, characteristics of the target groups, the environment and the HTAs themselves can mediate the impact (Goldberg 1994; Mittman and Siu 1992; Mittman and Tonesk 1992). Examples are shown in **Box 41**. Knowledge about these factors can be used prospectively. As noted above, assessment programs should consider how to properly target and modify their dissemination strategies to achieve the desired impact given particular characteristics of organizations, clinicians, environments, etc.

Systematic attempts to document the dissemination processes and impacts of HTA programs are infrequent (Banta 1993; Goodman 1988; Institute of Medicine 1985), though a few, notably the NIH Consensus Development Program (Ferguson 1993), have been studied in detail. Like other interventions in health care, HTA programs may be expected to demonstrate their own cost-effectiveness, i.e., that the health and/or economic benefits resulting from an HTA program outweigh the cost of the program itself.

Box 41. Examples of Factors That Can Affect Impact of HTA Reports

Target provider organization characteristics

- Hospitals: general versus specialized, size, teaching status, patient mix, for-profit vs. non-profit, distribution of payment sources (e.g., fee-for-service vs. capitation), ownership status, financial status, accreditation
- Physicians' offices: group practice vs. solo practice, hospital affiliation, teaching affiliation, board certification, distribution of payment sources

Target clinician characteristics

- Type of clinician: physician, nurse, dentist, etc.
- Specialty
- Training
- Professional activities/affiliations
- Institutional affiliations (e.g., community hospital, university hospital)
- Familiarity with and access to recent literature

Environmental characteristics

- Urban, suburban, rural
- Competitive environment
- Economic status
- Third-party payment status (e.g., percentage of patients in HMOs, private insurance, etc.)
- State and local laws, regulations
- Malpractice activity

Characteristics of HTA findings/recommendations

- Type: research findings, practice guidelines, standards (e.g., equipment acquisition, use, maintenance), appropriateness criteria
- Format: printed, word-of-mouth, electronic, etc.
- Frequency of message
- Required level of compliance (ranging from mandatory to optional)
- Locus of decision: general practitioner/primary care physician only, physician specialist only, multiple clinicians, physician with patient input, patient only
- Perceived inappropriate rigidity (allowance for discretion for differing circumstances)
- Cost of relevant procedure/management of condition
- Payment issue(s) addressed: coverage status, payment level
- Reputation of sponsoring organization, analysts, expert panel
- Overall strength of evidentiary base (e.g., existence of "definitive" clinical trial)
- Credibility/rigor of assessment process
- Existence or potential for malpractice action
- Timeliness of dissemination, especially compared to degree of uncertainty, most recent research findings, or current levels/change rates of utilization of procedure
- Existence and nature of other HTA findings on same topic.

Sources: Goldberg 1994; Mittman 1992; others.

X. SELECTED ISSUES IN HTA

Locus of Assessment: "Make or Buy?"

The nature of an assessment problem will affect the determination of the most appropriate organization to conduct it. Certainly, a comprehensive HTA addressing multiple attributes of a technology can be very resource intensive, requiring considerable and diverse expertise, data sources, and other resources.

Some health care organizations, such as some ministries of health and national health services, major insurance companies, health plans, and integrated health systems, have their own internal HTA programs. For example, in a large hospital or health plan, this might include a core staff and a multidisciplinary HTA committee representing major clinical departments, nursing, pharmacy, allied health, biomedical engineering. This committee might interact with other committees such as pharmacy & therapeutics ("P&T"), strategic planning, and capital planning committees (Kaden 2002; University HealthSystem Consortium 1996).

Other organizations rely on assessment reports acquired from organizations that have devoted functions or otherwise specialize in HTA. For example, the US, the CMS requests HTAs from AHRQ to inform Medicare coverage decisions by CMS. Similarly, in support of its technology appraisals and clinical guidelines, the National Institute for Clinical Excellence (NICE) [www.nice.org.uk/] requests HTAs from the National Coordinating Centre for HTA (NCCHTA), which coordinates the NHS R&D Division HTA Programme [www.hta.nhsweb.nhs.uk/].

Other vendors for HTAs in the US and around the world include, e.g., Blue Cross and Blue Shield Association Technology Evaluation Center [www.bluecares.com/guide/treattech.html], Cochrane Collaboration, ECRI [www.ecri.org], Hayes Inc. [www.hayesinc.com], Institute for Clinical Systems Improvement [www.icsi.com], MetaWorks Inc. [www.metawork.com], and University HealthSystem Consortium [www.uhc.edu]. Depending upon the producing HTA organization, these HTA reports may be available at no cost, for members only, on a subscription basis, or for a specific price per report.

Health care decision makers can "make or buy" HTAs. Determining the responsibility for sponsoring or conducting an assessment depends upon the nature of the problem, financial resources available, expertise of available personnel, time constraints, and other factors. For any assessment problem, an organization must determine the extent to which it will devote its resources to conducting the assessment itself or purchasing it from other sources. Some health care organizations commission selected components of an HTA, such as evidence retrieval and synthesis, and perform the other steps in-house.

One of the potential advantages of requesting or commissioning an outside group to conduct HTAs is to gain an independent, outside view where a requesting agency might have a perceived conflict of interest. Thus, a major health care payer might seek an HTA from an outside group to inform its coverage decision about a costly new technology in order to diminish perceptions of a potential bias against making a decision not to cover the technology.

Factors that influence the "make or buy" decision include the following (Goodman, Snider 1996).

- Is an existing assessment available? If an existing assessment is available, does it address the specific assessment problem of interest, including the technology or intervention, patient population, and impacts of interest? Does it have a compatible perspective? Is the assessment still current? Is the methodology used sufficiently credible? Is the report worth its price?
- If an existing assessment needs to be updated or is not available, do people in the organization have the time and expertise to perform the required data collection and analyses? If a synthesis of existing information is needed, does the organization have database searching capabilities and staff to review and interpret the literature? If new data are needed, does the organization have the requisite resources and expertise?
- What methodology will be used? If, for example, a consensus development approach is preferred, does that consensus need to incorporate and reflect the opinions of the organization's own clinicians? Will local clinicians accept the results and report recommendations if they do not participate in the assessment?

Quality of Care and HTA

The relationship between HTA and quality of care is often poorly understood. Although a thorough discussion of this subject is not possible here, the following are some definitions and fundamental relationships concerning these concepts.

Quality of care is a measure or indicator of the degree to which health care is expected to increase the likelihood of desired health outcomes and is consistent with standards of health care. HTA and *quality assurance* are distinct yet interdependent processes that contribute to quality of care.

HTA generates findings that add to our knowledge about the relationship between health care interventions and health care outcomes. This knowledge can be used to develop and revise a range of standards and guidelines for improving health care quality, including practice guidelines, manufacturing standards, clinical laboratory standards, adverse event reporting, architecture and facility design standards, and other criteria, practices, and policies regarding the performance of health care.

The purpose of quality assurance activities is to ensure that the best available knowledge concerning the use of health care to improve health outcomes is properly used. It involves the implementation of health care standards, including activities to correct, reduce variations in, or otherwise improve health care practices relative to these standards. Continuous quality improvement (CQI) and total quality management (TQM) (Gann 1994; Wakefield 1993) are among the contemporary systematic approaches to quality assurance that are being adapted for hospitals and other health care institutions. Such approaches include, for example, the identification of *best practices* and the use of **benchmarking** to develop improved **clinical pathways** or **disease management** for medical and surgical procedures, administrative operations, etc. (Kim 2003; Kwan 2002; Pilnick 2001). For example, CQI has been evaluated in a recent multicenter RCT as a means improve the adoption of two process of care measures for CABG: preoperative β–blockade therapy and internal mammary artery grafting (Ferguson 2003). Notably, in this RCT, the intervention being tested was not those two health care interventions, but CQI.

Quality assurance involves a measurement and monitoring function, (i.e., *quality assessment*). **Quality assessment** is, primarily, a means for determining how well health care is delivered in comparison with

applicable standards or acceptable bounds of care. These standards or bounds may be grouped according to the structure of care (institutional, professional and physical characteristics), the process of care (content or nature of the health care delivered) and the outcomes of care (health status and well-being of patients) (Donabedian 1988). Increasingly, quality assurance involves studies of effectiveness data, including health outcomes and the determinants of those outcomes from the perspectives of clinicians, patients, administrators, and policymakers (McDonald 2000). In detecting these differences between how well health care is delivered and applicable standards, quality assessment can also call attention to the need for further HTA or other investigations. In recent years, there has been further development and overlap of the fields of HTA and quality assurance, along with outcomes research, clinical epidemiology, and evidence-based medicine.

In summary, HTA contributes knowledge used to set standards for health care, and quality assurance is used to determine the extent to which health care providers adhere to these standards (Lohr 1990; Lohr and Rettig 1988). Indeed, major reorganization of health care systems may be required to ensure that stronger evidence is generated systematically for setting standards of care, and that standards of care are broadly implemented (Institute of Medicine, 2001).

Outcomes Research and HTA

In principle, outcomes research concerns any inquiry into the health benefits of using a technology for a particular problem under general or routine conditions. In practice, the term **outcomes research** has been used interchangeably with the term **effectiveness research** since the late 1980s to refer to a constellation of methods and characteristics that overlap considerably with HTA. It has received increased attention in the US, particularly in the form of research funded by the AHRQ [www.ahrq.gov] (formerly the Agency for Health Care Policy and Research). The attention given to outcomes or effectiveness research by government and, increasingly, the private sector (Mendelson 1998) reflects greater demand for data on patient and provider experience with technologies beyond what can be learned from the limited number of carefully circumscribed efficacy trials, e.g., premarketing clinical trials for new drugs and devices (McDonald 2000).

Outcomes/effectiveness research has emphasized health problem-oriented assessments of care delivered in general or routine settings; interdisciplinary teams; a wide range of patient outcomes including mortality, morbidity, adverse events and HRQL measures; the use of nonexperimental data (e.g., from epidemiological studies and administrative data sets); variations in practice patterns and their relationship to patient outcomes; and patient roles in clinical decision-making. The scope of outcomes/effectiveness research has expanded in recent years to include collection of experimental data on effectiveness, e.g., from large, simple trials conducted in general practice settings.

Decentralization of HTA

Although technology assessment originated as a primarily centralized function conducted by federal government agencies or other national- or regional-level organizations, HTA has become a more decentralized activity conducted by a great variety of organizations in the public and private sectors that make technology-related policy decisions (Goodman 1998; Rettig 1997). As noted above, an HTA done from a particular perspective may not serve the technology-related policymaking needs of other perspectives. Even for the same technology or clinical problem, there can be widely different assessment needs of politicians, regulatory agencies, health technology companies, hospitals, payers,

physicians, and others. These needs are heightened with increased economic responsibilities and pressures on these different parties.

The growth in decentralized HTA activity has arisen less from a reduction in the level of centralized activity than expansion of HTA programs for particular decision-making needs. In the US, there remain multiple government centers with ongoing HTA responsibilities to fulfill particular purposes, e.g., drug and device regulation at the FDA, NIH consensus development conferences, Medicare coverage policies by the CMS, and the technology assessment program of AHRQ. There has been considerable expansion in activities elsewhere, particularly in the private sector, as well as greater reliance by centralized sources on HTA inputs from outside sources. Increasingly, large health care providers and major health care product companies are establishing units devoted to "technology assessment," "pharmacoeconomics," "clinical effectiveness," "health outcomes research," and related areas. More health plans (including various managed care organizations and insurance companies) have established formal programs to assess new procedures and other technologies in support of payment decisions. The number and magnitude of private firms and university centers involved in HTA is increasing. HTA committees (with various names) are now common among medical specialty and subspecialty societies. Hospital networks, managed care organizations and other large health care providers in the private sector have HTA programs to support acquisition and management of pharmaceuticals (e.g., P&T committees), equipment and other technologies and other technology-related needs throughout their systems (Kaden 2002; University HealthSystem Consortium 1996).

Aside from the growth of HTA in the private sector, even HTA conducted by government agencies is drawing upon more decentralized resources. In the US, the FDA has long relied on advisory panels comprising outside experts to examine clinical trial findings and other evidence to provide recommendations regarding market approval of new drugs, biologicals, and medical devices. CMS has a large Medicare Coverage Advisory Committee (MCAC) [www.cms.gov/mcac/default.asp], arranged into various panels, that provides recommendations for national coverage policies on new technologies and other interventions, based on review of the clinical literature, consultations with experts, and other data. AHRQ's Evidence-based Practice Centers (EPC) program [www.ahrq.gov/clinic/epcix.htm] has established contracts with 13 EPCs, mostly academic health centers and other institutions, including three in Canada, which generate "evidence reports" and technology assessments in support of clinical practice guidelines, coverage policies, and other practices and policies. Indeed, some EPC reports are conducted at the request, via AHRQ, of the NIH Consensus Development Program, CMS, and other government agencies; other requests are made by other organizations in the private sector, such as health professional organizations. In this manner, AHRQ provides a portal for decentralized HTA, via the 13 EPCs, on behalf of government and non-government organizations. AHRQ also administers the US Preventive Services Task Force [www.ahrq.gov/clinic/uspstfix.htm], an independent panel of experts in primary care and prevention that systematically reviews evidence of effectiveness and develops recommendations for a broad range of clinical preventive services.

The Cochrane Collaboration [www.cochrane.org], another highly decentralized, successful model, involves 50 workgroups of volunteer experts around the world, coordinated through about 14 centers based in 12 countries, who conduct systematic reviews of a diverse variety of health care interventions.

Decentralization of HTA and related functions widens the expertise available to HTA and brings broader perspectives to the process and diminishes or balances potential conflicts of interest. Together,

these generally add to the credibility of HTA processes and findings, and lessen any charges that assessments reflect narrow or self-serving interests of a particular agencies or organizations.

Tracking changes in the locus and magnitude of HTA is confounded by a broadening connotation of the term. Rather than referring only to the comprehensive inquiries involving broad societal impacts envisioned for the field in the 1960s, HTA is now used to refer to almost any evaluation or analysis pertaining to health care technology. Much of the expansion of HTA concerns meeting focused, immediate needs such as a coverage decision for a particular procedure, determination of the cost-effectiveness of a new device, or an equipment purchase decision. Another shift in locus concerns professional responsibility. Whereas technology-related decision-making in health care organizations was largely the responsibility of physicians, it is increasingly shared or redistributed among a wider spectrum of managers and other professionals.

Certain changes in the health care market are prompting greater balance between centralized and decentralized HTA. Hospital networks, large managed care systems and other large systems such as the Department of Veterans Affairs (VA) continually seek to build economies of scale and buying leverage for health care products, ranging from surgical gloves to hip joint implants. With HTA units that are centralized yet responsive to needs of individual facilities, these large organizations can consolidate their HTA efforts and support system-wide acquisition of drugs, equipment, and services.

As health care providers and payers realize the resource requirements for conducting well-designed evaluations of health care technologies, they weigh the tradeoffs of conducting their own assessments versus subscribing to assessment report series from outside assessment groups. Clearly, assessment requirements vary widely depending on the type of technology involved. Acquisition of commodity products such as most types of syringes and surgical gloves is largely based on price, whereas acquisition of the latest drug-coated coronary artery stent requires a more considered evaluation of safety, effectiveness, cost, and other attributes. Nearly all hospitals and health care networks in the US rely on group purchasing organizations (GPOs) that use economies of scale to acquire most of their health care products. These GPOs, particularly the larger ones, have their own technology evaluation or clinical review committees that examine available evidence on technologies such as implantable cardiac defibrillators and MRI units, whose acquisition is a matter of factors other than price alone. In turn, many GPOs also subscribe to technology assessment report services (Lewin Group 2002).

Barriers to HTA

Although the general trend in health care is toward wider and improved HTA, several countervailing forces to HTA remain. Foremost, particularly in the US and other wealthy countries, has been a "**technological imperative**" comprising an abiding fascination with technology, the expectation that new is better, and the inclination to use a technology that has potential for some benefit, however marginal or even poorly substantiated (Deyo 2002). Some argue that the increased potential of technology only raises the imperative for HTA (Hoffman 2002). Another countervailing factor is the sway of prestigious proponents or a "champion" of a technology in the absence of credible evidence. A third impediment is the inertia of medical practice, e.g., in the form of reluctance to change long-standing practice routines, conservative payment policies, and quickly outdated education. This is complemented by lack of opportunities for, or encouragement of, scientific inquiry and skepticism in clinical education.

Ever more effective marketing and promotions, including short courses sponsored by medical product companies to train physicians in using these products, can divert attention from key concerns of HTA. Another obstacle is the limited level of investment, by government and industry sources in HTA and related evaluations of what works in health care. Although some assessment programs and certain HTA findings are nationally or internationally recognized, the resources allocated for HTA in the US are virtually lost in the rounding error of national health care expenditures. Finally, the impression persists in some quarters that the goal of HTA is to limit the innovation and diffusion of health care technology.

Political processes can circumvent or threaten evidence-based processes (Fletcher 1997). One of the higher-profile applications of HTA is in determining covered services for health programs that are provided or funded by governments as well by the private sector. While most of these health programs have HTA processes that support benefits determinations, they are also subject to legislation (laws) in their respective countries, states, provinces, and other jurisdictions. Legislative bodies at these levels can mandate, or require, that health programs provide certain services. For example, in the US, the Congress has mandated that the Medicare program (for the elderly and disabled) provide certain services (e.g., screening procedures) that are not included in the benefit categories under the original Medicare statute. State legislatures have mandated that their Medicaid programs (for people with low incomes) as well as private sector health plans operating in their states, provide certain services. Recent examples of mandated services include autologous bone marrow transplant with high-dose chemotherapy (ABMT-HDC) for advanced breast cancer, bone densitometry screening for osteoporosis, screening mammography, prostate cancer screening, and treatment for temporomandibular joint disorder. Such mandates, including the ones noted here, may or may not be based upon the types of evidence-based methods used in HTA processes. As is the case for other industries, these mandates can be affected by political influence brought, e.g., by "lobbying" or "pressure groups" representing patient advocate organizations, physician groups, health product makers, and others (Deyo 1997; Sheingold 1998).

In some instances, legislative mandates arise through frustration with slowed or delayed HTA processes. A notable instance was the mandate by the US Congress for Medicare coverage of dual energy x-ray absorption (DEXA) for bone mineral density measurement, which had been subject to an assessment involving two federal agencies over a seven-year period (Lewin Group 2000). However, these mandates often circumvent evidence-based coverage policy, by providing an alternative, political route to coverage of technologies. The apparently direct process of mandating coverage of a technology, rather than subjecting it to well-founded HTA, can mask more complex clinical consequences. In the 1990s, many health plans reluctantly agreed to cover HDC-ABMT in response to state legislative mandates brought about by intensive political pressure, and the threat of litigation (legal action in courts). It was not until 1999, after tens of thousands of women were subjected to the procedure, that results of five well-conducted RCTs demonstrated that the procedure conferred no benefit over standard-dose treatment for breast cancer, and caused unnecessary suffering in some women (Berger 1999; Mello 2001; Sharf 2001).

Aside from barriers to conducting HTA are barriers to implementing its findings and recommendations, particularly by decision makers and policymakers for whom HTA reports are intended. Among these are: lack of access to HTA reports, complex and technical formats of HTA reports, questionable data quality, absence of real-world applications, and narrow focus (Henshall 2002).

HTA and Underused Technologies

When used properly, HTA can reduce or eliminate the use of technologies that are not safe and effective, or whose cost is too high relative to their benefits. As discussed above, HTA can also be used to remove technologies from the market that are harmful or ineffective. Less attention is given to the ability of HTA to identify technologies that are underused, and to help determine why they are underused (Asch 2000; McNeil 2001). Underuse is prevalent in preventive, acute, and chronic care (McGlynn 2003) and contributes to tens of thousands of deaths and billions of dollars of losses to the economy and unnecessary health care costs (National Committee for Quality Assurance 2003).

For example, there is overwhelming evidence that smoking cessation interventions, including nicotine replacement therapy, the antidepressant bupropion, and counseling, are safe, effective, and cost effective (Anderson 2002; Foulds 2002; Jorenby 1999; Woolacott 2002). However, in Europe, North America, and elsewhere, these interventions are used far less than is indicated. Underuse is attributed to various reasons, including: lack of insurance coverage, concerns about short-term costs without regard to cost-effectiveness in the short-term (e.g., for pregnant women and infants) and the long-term; lack of smoker awareness of effective interventions; insufficient demand by patients, physicians, and the tobacco-control community; and the influence of the tobacco industry on policymaking (Schauffler 2001).

Box 42 shows examples of health care technologies for which good evidence exists of effectiveness or cost-effectiveness, but that are used significantly less than is indicated, even where they are affordable. Although this list applies primarily to the US, many of these technologies are underused elsewhere in North America, Western Europe, and other of the wealthier countries. The reasons that worthy technologies are underused are diverse, and include the following.

- Lack of awareness on the part of patients, physicians, and others
- Inadequate information dissemination
- Limited coverage and reimbursement
- Concerns about short-term cost without regard for cost savings and cost-effectiveness in the short- and long-terms
- Inappropriate or unsubstantiated concerns about improper use (e.g., pain therapy)
- Inconvenience and misperceptions on the part of clinicians or patients
- Clinical inertia
- Insufficient supply (e.g., organs for transplantation)
- Disproportionate concerns about adverse effects (e.g., warfarin to reduce risk of stroke)
- Concerns about patient **compliance** (e.g., polypharmacy for HIV/AIDS)
- Fear of stigma (e.g., treatment of depression)
- Professional conflicts and "turf battles" on the part of physician specialists, provider institutions, industry, and others

Box 42. Underused Health Care Technologies (US)

- ACE inhibitors for treatment of heart failure
- ACE inhibitors for prevention of renal deterioration in insulin-dependent diabetics
- Ambulation aids (canes, crutches, walkers)
- Antibiotics for gastrointestinal ulcers
- Beta blockers for survivors of acute myocardial infarction
- Cholesterol-lowering drugs for patients at risk of coronary artery disease
- Cochlear implants for severe-to-profound deafness
- Colorectal cancer screening
- Corticosteroid inhalants for treating asthma
- Corticosteroid therapy for fetuses at risk of preterm delivery
- Depression diagnosis and treatment
- Diabetic retinopathy screening
- Hepatitis B virus vaccination of infants
- Implantable cardioverter-defibrillators for survivors of cardiac arrest
- Incontinence diagnosis and treatment
- Intraocular pressure screening for glaucoma
- Oral rehydration therapy for dehydrated children
- Organ transplantation
- Pain management
- Polypharmacy (with protease inhibitors) for HIV/AIDS
- Pneumococcal vaccine for high risk patients
- Prenatal care
- Smoking cessation interventions
- Thrombolytic therapy for acute myocardial infarction
- Thrombolytic therapy for ischemic stroke
- Warfarin to prevent strokes due to atrial fibrillation

Conflict of Interest

HTA should consider the potential for conflict of interest on multiple levels. One is on the part of investigators who conducted and reported on the clinical trials and other studies that comprise the body of evidence under review. A second is on the part of sponsors of the primary research, e.g., technology companies, who have varying degrees of control over what research is conducted, selection of intervention and control treatments, selection of endpoints and follow-up periods, and whether research results are submitted for publication. Another is on the part of the health technology assessors themselves, including analysts, panel members, or other experts involved in reviewing the evidence and making findings and recommendations.

Interpreting the literature for an assessment should include consideration of the existence of potential conflicts of interest that may have affected the conduct of a study or presentation of results. For study investigators, conflicts of interest may arise from having a financial interest (e.g., through salary support, ongoing consultancy, owning stock, owning a related patent) in a health care company (or one of its competitors) that may be affected by the results of a study or being an innovator of a technology under study. A systematic review of research on financial conflicts of interest among biomedical researchers found that approximately one-fourth of investigators have industry affiliations, and two-thirds of academic institutions hold equity in start-up companies that sponsor research performed at the same institutions. Industry sponsorship of research also was associated with restrictions on publication and data sharing (Bekelman 2003). Clinical trials and cost-effectiveness analyses that are sponsored by industry yield positive results more often that studies that are funded or conducted by others (Chopra 2003; Friedberg 1999). Among the reasons suggested for this discrepancy are that industry's publication restrictions tend to withhold studies with negative results. Another is that industry is more likely to sponsor studies (particularly RCTs) in which the results are likely to be positive, i.e., where there is an expectation that one intervention (e.g., a new drug) is superior to the alternative intervention. In the case of RCTs, this latter tendency would undermine the principle of "**equipoise**" for enrolling patients in an RCT (Djulbegovic 2000).

Peer-reviewed journals increasingly require disclosure of information pertaining to financial interests of investigators and the source of funding of studies (International Committee of Medical Journal Writers 1993; Kassirer 1993; Lo 2000). Some journals have particular requirements regarding protection against conflict of interest for economic analyses that have been subject to considerable controversy (Kassirer 1994; Steinberg 1995). Information about investigators, sponsorship of a study, or other factors that suggests the potential for conflict of interest should be considered when interpreting the evidence. Studies that are subject to potential conflicts of interest may have to be discounted or dropped from the body of evidence accordingly.

HTA programs should take active measures to protect against potential conflicts of interest among assessment teams and panelists (Fye 2003; Phillips 1994). A conflict of interest may be any financial or other interest that conflicts with one's service on an assessment group because it could impair that person's objectivity or could create an unfair advantage. Conflict of interest is not the same as bias among assessment teams and panelists, which may entail views or inclinations that are intellectually motivated or that would be expected to arise from having a given organizational or professional affiliation. HTA programs should take active measures to minimize or balance bias among assessment teams and panel members.

The following recommendations for managing conflict of interest in practice guidelines development (Choudhry 2002) may be relevant as well to panels involved in HTA and related evidence-based activities.

- A formal process should exist to disclose potential conflict of interest before the guideline development begins.
- All members of the guideline group should be involved in a discussion of conflicts of interest and how significant relationships will be managed.
- Participants who have relationships with industry, government agencies, health care organizations or specialty societies need not necessarily be excluded, but the group has to decide among itself a threshold for exclusion.
- There must be complete disclosure to readers of the practice guidelines of financial and/or other relationships with industry, government agencies, health care organizations and specialty societies.

APPENDIX. SUGGESTED READINGS IN HTA

Origins and Evolution of Technology Assessment

Arnstein SA. Technology assessment: opportunities and obstacles. *IEEE Transactions on Systems, Man, and Cybernetics*. 1977;SMC-7:571-82.

Brooks H, Bowers R. The assessment of technology. *Science*. 1970;222(2):13-20.

Coates JF. Technology assessment: the benefits ... the costs ... the consequences. *The Futurist*. 1971;December:225-31.

National Academy of Engineering, Committee on Public Engineering Policy. *A Study of Technology Assessment.* Washington, DC: US Government Printing Office; 1969.

National Research Council, Committee on the Life Sciences and Social Policy. *Assessing Biomedical Technologies: An Inquiry into the Nature of the Process*. Washington, DC: National Academy of Sciences; 1975.

Porter AL. Technology assessment. *Impact Assessment*. 1995;13(2):135-51.

Rettig RA. *Health Care in Transition: Technology Assessment in the Private Sector*. Santa Monica, Ca: RAND; 1997.

US Congress, Office of Technology Assessment. *Technology Assessment in Business and Government. Summary and Analysis.* Washington, DC: US Government Printing Office; 1977.

US Congress, Office of Technology Assessment. *Assessing the Efficacy and Safety of Medical Technologies.* Washington, DC: US Government Printing Office; 1978.

US Congress, Office of Technology Assessment. *Development of Medical Technology: Opportunities for Assessment.* Washington, DC: US Government Printing Office, 1976.

US Congress, House of Representatives. Committee on Science and Astronautics. Technology Assessment. *Statement of Emilio Q. Daddario, Chairman, Subcommittee on Science Research and Development.* 90th Cong., 1st sess., Washington, DC; 1967.

Overview of HTA

Banta HD, Luce BR. *Health Care Technology and its Assessment: An International Perspective*. New York: Oxford University Press; 1993.

Busse R, Orvain J, Velasco M, et al. Best practice in undertaking and reporting health technology assessments. *Int J Technol Assess Health Care*. 2002;18:361-422.

Cochrane AL. *Effectiveness and Efficiency: Random Reflections on Health Services.* London, England: Nuffield Provincial Hospitals Trust; 1972.

Eddy DM. *A Manual for Assessing Health Practices & Designing Practice Policies: The Explicit Approach.* Philadelphia, Pa: American College of Physicians; 1992.

Eisenberg JM. Ten lessons for evidence-based technology assessment. *JAMA*. 1999;17:1865-9.

Feeny D, Guyatt G, Tugwell P, eds. *Health Care Technology: Effectiveness, Efficiency, and Public Policy.* Montreal, Quebec: Institute for Research on Public Policy; 1986.

Franklin C. Basic concepts and fundamental issues in technology assessment. *Intensive Care Medicine*. 1993;19(2):117-21.

Frazier HS, Mosteller F, eds. *Medicine Worth Paying for: Assessing Medical Innovations.* Cambridge, Ma: Harvard University Press; 1995.

Goodman CS. Healthcare technology assessment: methods, framework, and role in policy making. *Am J Manag Care*. 1998;4:SP200-14.

Institute of Medicine. *Assessing Medical Technologies*. Washington, DC: National Academy Press; 1985.

Joint Report of the Council on Scientific Affairs and the Council on Medical Service. Technology assessment in medicine. *Arch Intern Med*. 1992;152(1):46-50.

Littenberg B. Technology assessment in medicine. *Academic Medicine*. 1992;67(7):424-8.

Porter AL, Rossini FA, Carpenter SR, Roper AT. *A Guidebook for Technology Assessment and Impact Analysis.* New York, NY: North Holland, 1980.

Power EJ, Tunis SR, Wagner JL. Technology assessment and public health. *Annual Review of Public Health.* 1994;15:561-79.

US Congress, Office of Technology Assessment. *Strategies for Medical Technology Assessment.* Washington, DC: US Government Printing Office; 1982.

US Congress, Office of Technology Assessment. *Identifying Health Technologies That Work: Searching for Evidence.* OTA-H-608. Washington, DC: US Government Printing Office; 1994.

Priority Setting

Chase D, Milne R, Stein K, Stevens A. What are the relative merits of the sources used to identify potential research priorities for the NHS HTA programme? *Int J Technol Assess Health Care.* 2000;16(3):743-50.

Donaldson MS, Sox HC, Jr, eds. *Setting Priorities for Health Technology Assessment: A Model Process*. Washington, DC: National Academy Press; 1992.

Eddy DM. Selecting technologies for assessment. *Int J Technol Assess Health Care*. 1989;5(4):485-501.

Hadorn DC. The problem of discrimination in health care priority setting. *JAMA*. 1992;268(11):1454-9.

Henshall C, Oortwijn W, Stevens A, Granados A, Banta D. Priority setting for health technology assessment. Theoretical considerations and practical approaches. Priority setting subgroup of the EUR-ASSESS project. *Int J Technol Assess Health Care*. 1997;13(2):144-85.

Lara ME, Goodman C, eds. *National Priorities for the Assessment of Clinical Conditions and Medical Technologies: Report of a Pilot Study.* Washington, DC: National Academy Press; 1990.

Oortwijn WJ, Vondeling H, van Barneveld T, et al. Priority setting for health technology assessment in the Netherlands: principles and practice. *Health Policy*.2002;62(3):227-42.

Phelps CE, Parente ST. Priority setting in medical technology and medical practice assessment. *Medical Care*. 1990;28(8):703-23.

Phelps CE, Mooney C. Correction and update on 'priority setting in medical technology assessment.' *Medical Care*. 1992;30(8):744-51.

Sassi F. Setting priorities for the evaluation of health interventions: when theory does not meet practice. *Health Policy*. 2003;63:141-54.

Shekelle P, Eccles MP, Grimshaw JM, Woolf SH. When should clinical guidelines be updated? *BMJ*. 2001;323:155-7.

Randomized Controlled Trials

Buring JE, Jonas MA, Hennekens CH. Large and simple randomized trials. In *Tools for Evaluating Health Technologies: Five Background Papers,* US Congress, Office of Technology Assessment. BP-H-142. Washington, DC: US Government Printing Office; 1995:67-91.

Chalmers TC, Celano P, Sacks HS, Smith H, Jr. Bias in treatment assignment in controlled clinical trials. *N Engl J Med*. 1983;309:1358-61.

Chow S, Liu J. *Design and Analysis of Clinical Trials.* New York, NY: John Wiley & Sons; 1998.

Ellenberg SS. Do large, simple trials have a place in the evaluation of aids therapies? *Oncology*. 1992;6(4):55-9,63.

Freiman, JA, Chalmers TC, Smith H, Jr., Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized controlled trial. Survey of 71 "negative" trials. *N Engl J Med*. 1978;299(13):690-4.

Gurwitz JH, Col NF, Avorn J. The exclusion of the elderly and women from clinical trials in acute myocardial infarction. *JAMA*. 1992;268(11):1417-22.

Juni P, Altman DG, Egger M. Systematic reviews in health care: assessing the quality of controlled clinical trials. *BMJ*. 2001;323(7303):42-6.

Kraemer HC, Pruyn JP. The evaluation of different approaches to randomized clinical trials. *Arch Gen Psych*. 1990;47(12):1163-9.

Lubsen J, Tijssen JG. Large trials with simple protocols: indications and contraindications. *Control Clin Trials*. 1989;10(suppl. 4):151S-60S.

Merigan TC. You can teach an old dog new tricks: how AIDS trials are pioneering new strategies. *N Engl J Med*. 1990;323(19):1341-3.

Merkatz RB, Temple R, Sobel S, Feiden K, Kessler D. Women in clinical trials of new drugs: a change in Food and Drug Administration Policy. *N Engl J Med*. 1993;329(4):292-6.

Moher D, Cook DJ, Eastwood S, Olkin I, et al. Improving the quality of reports of meta-analyses of randomized controlled trials: The QUOROM Statement. Quality of reporting of meta-analyses. *Lancet*. 1999;354:1896-900.

Moher D, Schulz KF, Altman D, et al. The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-group randomized trials. *JAMA*. 2001;285(15):1987-91.

Peto R, Collins R, Gray R. Large-scale randomized evidence: large, simple trials and overviews of trials. *Journal of Clinical Epidemiology*. 1995;48(1):23-40.

Pocock SJ, Hughes MD, Lee RJ. Statistical problems in the reporting of clinical trials. *N Engl J Med*. 1987;317(7):426-32.

Rothenberg, Carbone DP, Johnson DH. Improving the evaluation of new cancer treatments: challenges and opportunities. *Nat Rev Cancer.* 2003;3:303-9.

Spilker B. *Guide to Clinical Trials*. New York, NY: Raven Press; 1991.

Yusuf S, Held P, Teo KK, Toretsky ER. Selection of patients for randomized controlled trials: implications of wide or narrow eligibility criteria. *Statistics in Medicine*. 1990;9(1-2):73-83.

HTA of Diagnostic and Screening Technologies

Begg CB. Biases in the assessment of diagnostic tests. *Stat Med*. 1987;6(4):411-23.

Boyko EJ. Ruling out or ruling in disease with the most sensitive or specific diagnostic test: short cut or wrong turn? *Med Dec Making*. 1994;14(2):175-9.

Canadian Task Force on the Periodic Health Examination. The periodic health examination: 2. 1987 update. *CAMJ*. 1988;138(7):617-26.

Deyo RA. Cascade effects of medical technology. *Annu Rev Public Health*. 2002;23:23-44.

Goin JE, Hermann GA. The clinical efficacy of diagnostic imaging evaluation studies: Problems, paradigms, and prescription. *Investigative Radiology*. 1991;26(5):507-11.

Kassirer JP. Our stubborn quest for diagnostic certainty. a cause of excessive testing. *N Engl J Med*. 1989;321(18):1272-3.

Mushlin AI, Mooney C, Grow V, Phelps CE. The value of diagnostic information to patients with suspected multiple sclerosis. *Archives of Neurology*. 1994;51(1):67-72.

Riegelman RK, Hirsch RP, eds. *Studying a Study and Testing a Test. How to Read the Medical Literature.* Second Edition. Boston, Ma: Little, Brown; 1989.

Sox H, Stern S, Owens D, Abrams HL. *Assessment of Diagnostic Technology in Health Care: Rationale, Methods, Problems, and Directions.* Institute of Medicine. Washington, DC: National Academy Press; 1989.

Thornbury JR. Clinical efficacy of diagnostic imaging: love it or leave it. *AJR.* 1994;162(1):1-8.

Thornbury JR, Fryback DG. Technology assessment − an American view. *Eur J Radiol*. 1992;14(2):147-56.

US Preventive Services Task Force. *Guide to Clinical Preventive Services: Second edition*. Baltimore, Md: Williams & Wilkins; 1996.

Wells PN, Garrett JA, Jackson PC. Assessment criteria for diagnostic imaging technologies. *Medical Progress Through Technology*. 1991;17(2):93-101.

Evidence in HTA

Bailar JC, Mosteller F, eds. *Medical Uses of Statistics*. Boston, Ma: NEJM Books; 1992.

Chalmers I. Under-reporting research is scientific misconduct. *JAMA*. 1990;263(10):1405-8.

Chalmers TC, Frank CS, Reitman D. Minimizing the three stages of publication bias. *JAMA*. 1990;263(10):1392-5.

Cook DJ, Guyatt GH, Laupacis A, Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. *Chest*. 1992;102(suppl. 4):305S-11S.

Daly LE, Bourke GJ, McGilvray J. *Interpretation and Uses of Medical Statistics*. Oxford, England: Blackwell Scientific Publications; 1991.

Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. *Lancet*. 1991;337:867-72.

Evidence-Based Medicine Working Group. Evidence-based medicine: a new approach to teaching the practice of medicine. *JAMA*. 1992;268(17):2420-5.

Flood AB. Peaks and pits of using large data bases to measure quality of care. *Int J Technol Assess Health Care*. 1990;6(2):253-62.

Greer N, Mosser G, Logan G, Halaas GW. A practical approach to evidence grading. *Jt Comm J Qual Improv.* 2000;26(12):700-12.

Guyatt GH, Haynes RB, Jaeschke RZ, et al. Users' guide to the medical literature, XXV: Evidence-based medicine: principles for applying the users' guides to patient care. Evidence-Based Medicine Working Group. *JAMA*. 2000;284:1290-6.

Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature. II. How to use an article about a therapy or prevention. A. Are the results of the study valid? Evidence-Based Medicine Working Group. *JAMA*. 1993;270(21):2598-601.

Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature. II. How to use an article about a therapy or prevention. B. What were the results and will they help me in caring for my patients? Evidence-Based Medicine Working Group. *JAMA*. 1994;271(1):59-63.

Harris RP, Helfand M, Woolf SH, et al. Current methods of the US Preventive Services Task Force. A review of the process. *Am J Prev Med*. 2001;20(3S):21-35.

Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? Evidence-Based Medicine Working Group. *JAMA*. 1994;271(5):389-91.

Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? Evidence-Based Medicine Working Group. *JAMA*. 1994;271(9):703-7.

Khan KS, ter Riet G, Glanville J, Sowden AJ, Kleijnen J. Undertaking Systematic Reviews of Research on Effectiveness. CRD's Guidance for Carrying Out or Commissioning Reviews: York, England: University of York, NHS Centre for Reviews and Dissemination; 2000.

Oxman AD, Cook DJ, Guyatt GH. Users' guides to the medical literature. VI. How to use an overview. Evidence-Based Medicine Working Group. *JAMA*. 1994;272(17):1367-71.

Oxman AD, Sackett DL, Guyatt GH. Users' guides to the medical literature. I. How to get started. Evidence-Based Medicine Working Group. *JAMA*. 1993;270(17):2093-5.

RTI International-University of North Carolina Evidence-based Practice Center. *Systems to Rate the Strength of Scientific Evidence.* Evidence Report/Technology Assessment Number 47. AHRQ Publication No. 02-E016, Rockville, Md: Agency for Healthcare Research and Quality, April 2002.

Sackett DL, Richardson WS, Rosenberg W, Haynes RB. *Evidence-Based Medicine*. New York, NY: Churchill Livingstone; 1997.

Cost-Effectiveness and Related Economic Analyses

Doubilet P, Weinstein MC, McNeil BJ. Use and misuse of the term "cost effective" in medicine. *N Engl J Med*. 1986;314(4):253-6.

Drummond MF, Davies L. Economic analysis alongside clinical trials: revisiting the methodological issues. *Int J Technol Assess Health Care*. 1991;7(4):561-73.

Drummond MF, O'Brien B, Stoddart GL, Torrance GW. *Methods for the Economic Evaluation of Health Care Programmes.* Second Edition. Oxford, England: Oxford University Press; 1997.

Eddy DM. Clinical decision making: from theory to practice. cost-effectiveness analysis. Is it up to the task? *JAMA*. 1992;267(24):3342-8.

Eisenberg JM. Clinical economics: A guide to the economic analysis of clinical practices. *JAMA*. 1989;262(20):2879-86.

Gold MR, Siegel JE, Russell LB, Weinstein MC. *Cost-Effectiveness in Health and Medicine.* New York, NY: Oxford University Press; 1996.

Laupacis A, Feeny D, Detsky AS, Tugwell PX. How attractive does a new technology have to be to warrant adoption and utilization? Tentative guidelines for using clinical and economic evaluations. *CAMJ*. 1992;146(4):473-81.

Poe NR, Griffiths RI. Clinical-economic trials. In *Tools for Evaluating Health Technologies: Five Background Papers*. US Congress, Office of Technology Assessment, 125-49. BP-H-142. Washington, DC: US Government Printing Office; 1995.

Russell LB. Opportunity costs in modern medicine. *Health Aff*. 1992;11(2):162-9.

Warner KE, Luce BR. *Cost-Benefit and Cost-Effectiveness Analysis in Health Care*. Ann Arbor, Mich: Health Administration Press; 1982.

Weinstein MC. Principles of cost-effective resource allocation in health care organizations. *Int J Technol Assess Health Care*. 1990;6(1):93-103.

Meta-Analysis

Chalmers TC, Lau J. Meta-analytic stimulus for changes in clinical trials. *Statistical Methods in Medical Research*. 1993;2(2):161-72.

Cook TD, Cooper H, Cordray DS, Hartmann H, Hedges LV, Light RJ, Louis TA, Mosteller F. *Meta-Analysis for Explanation: A Casebook.* New York, NY: Russell Sage Foundation; 1992.

Deeks JJ. Systematic reviews in health care: systematic reviews of evaluations of diagnostic and screening tests. *BMJ* 2001;323:157-62.

Eddy DM. *A Manual for Assessing Health Practices & Designing Practice Policies: The Explicit Approach.* Philadelphia, Pa: American College of Physicians; 1992.

Egger M, Davey Smith G, Altman DG, eds. *Systematic Reviews in Health Care: Meta-Analysis in Context*. 2nd ed. London, England: BMJ Books; 2001.

Hasselblad V, McCrory DC. Meta-analytic tools for medical decision making: a practical guide. *Medical Decision Making*. 1995;15(1):81-96.

Jones DR. Meta-analysis of observational epidemiological studies: a review. *J Royal Soc Med*. 1992;85(3):165-8.

Laird NM, Mosteller F. Some statistical methods for combining experimental results. *Int J Technol Assess Health Care*. 1990;6(1):5-30.

Longnecker MP. Meta-analysis. In *Tools for Evaluating Health Technologies: Five Background Papers*. US Congress, Office of Technology Assessment, 93-123. BP-H-142. Washington, DC: US Government Printing Office; 1995.

Moher D, Cook DJ, Eastwood S, Olkin I, et al. Improving the quality of reports of meta-analyses of randomized controlled trials: the QUOROM statement. Quality of reporting of meta-analyses. *Lancet*. 1999;354:1896-900.

Normand SL. Meta-analysis: formulating, evaluating, combining, and reporting. *Stat Med*. 1999;18:321-59.

Petitti DB. Approaches to heterogeneity in meta-analysis. *Stat Med*. 2001;20:3625-33.

Petitti DB. *Meta-Analysis, Decision Analysis, and Cost-Effectiveness Analysis. Methods for Quantitative Synthesis in Medicine*. New York, NY: Oxford University Press; 1994.

Thacker SB. Meta-analysis: A quantitative approach to research integration. *JAMA*. 1988;259(11):1685-9.

van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Stat Med*. 2002;21:589-624.

Decision Analysis

Eckman MH, Levine HJ, Pauker SG. Decision analytic and cost-effectiveness issues concerning anticoagulant prophylaxis in heart disease. *Chest*. 1992;102(suppl. 4):538-49S.

Fleming C, Wasson JH, Albertsen PC, Barry MJ, Wennberg JE. A decision analysis of alternative treatment strategies for clinically localized prostate cancer. *JAMA*. 1993;269(20):2650-8.

Hagen MD. Decision analysis: a review. *Family Medicine*. 1992;24(5):349-54.

McNeil BJ, Pauker SG. Decision analysis for public health: principles and illustrations. *Annu Rev Pub Health*. 1984;5:135-61.

Pauker SG, Kassirer JP. Medical progress: decision analysis. *N Engl J Med*. 1987;316(5):250-8.

Soto J. Health economic evaluations using decision analytic modeling. Principles and practices--utilization of a checklist to their development and appraisal. *Int J Technol Assess Health Care*. 2002;18:94-111.

Thornton JG, Lilford RJ, Johnston N. Decision analysis in medicine. *BMJ*. 1992;304(6834):1099-103.

Weinstein MC, O'Brien B, Hornberger J, et al. Principles of good practice for decision analytic modeling in health-care evaluation: report of the ISPOR task force on good research practices−modeling studies. *Value Health*. 2003;6(1):9-17.

Zalkind DL, Shachtman RH. A decision analysis approach to the swine influenza vaccination decision for an individual. *Med Care*. 1980;18(1):59-72.

Consensus Development

Andreasen PB. Consensus conferences in different countries: aims and perspectives. *Int J Technol Assess Health Care*. 1988;4(2):305-8.

Fink A, Kosecoff J, Chassin M, Brook RH. Consensus methods: characteristics and guidelines for use. *Am J Pub Health*. 1984;74(9):979-83.

Ferguson JH, Sherman CR. Panelists' views of 68 NIH consensus conferences. *Int J Technol Assess Health Care*. 2001;17(4):542-58.

Goodman C, Baratz SR, eds. *Improving Consensus Development for Health Technology Assessment: An International Perspective*. Washington, DC: National Academy Press, 1990.

Greer AL. The two cultures of biomedicine: can there be a consensus? *JAMA*. 1987;258(19):2739-40.

Kosecoff J, Kanouse DE, Rogers WH, McCloskey L, Winslow CM, Brook RH. Effects of the National Institutes of Health Consensus Development Program on physician practice. *JAMA*. 1987;258(19):2708-13.

Murphy MK, Black NA, Lamping DL, et al. Consensus development methods, and their use in clinical guideline development. *Health Technol Assessment*. 1998;2(3):i-iv,1-88.

Perry S. The NIH Consensus Development Program. A decade later. *N Engl J Med*. 1987;317(8):485-488.

Veatch RM. Consensus of expertise: the role of consensus of experts in formulating public policy and estimating facts. *J Med Philos*. 1991;16(4):427-45.

Wortman PM, Vinokur A, Sechrest L. Do consensus conferences work? A process evaluation of the NIH Consensus Development Program. *J Health Politics, Policy Law*. 1998;13(3):469-98.

Health-Related Quality of Life Measures

Fitzpatrick R, Fletcher A, Gore S, Jones D, Spiegelhalter D, Cox D. Quality of life measures in health care. I: applications and issues in assessment. *BMJ*. 1992;305(6861):1074-7.

Fletcher A, Gore S, Jones D, Fitzpatrick R, Spiegelhalter D, Cox D. Quality of life measures in health care. II: design, analysis, and interpretation. *BMJ*. 1992;305(6862):1145-8.

Gill TM, Feinstein AR. A critical appraisal of the quality of quality-of-life measurements. *JAMA*. 1994;272(8):619-26.

Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life. *Ann Intern Med*. 1993;118(8):622-9.

Mason JM. Cost-per QALY league tables: their roles in pharmacoeconomic analysis. *Pharmacoeconomics*. 1994;5:472-81.

Mehrez A, Gafni A. Healthy-years equivalents versus quality-adjusted life years: in pursuit of progress. *Med Dec Making*. 1993;13(4):287-92.

Mosteller F, Falotico-Taylor J, eds. *Quality of Life and Technology Assessment*. Washington, DC: National Academy Press, 1989.

Patrick DL, Erickson P. *Health Status and Health Policy. Quality of Life in Health Care Evaluation and Resource Allocation*. New York, NY: Oxford University Press; 1993.

Spiegelhalter D, Gore S, Fitzpatrick R, Fletcher AE, Jones DR, Cox DR. Quality of life measures in health care. III: Resource allocation. *BMJ*. 1992;305(6863):1205-9.

Ubel PA, Nord E, Gold M, et al. Improving value measurement in cost-effectiveness analysis. *Med Care.* 2000;38:892-901.

Ware JE, Jr. The status of health assessment 1994. *Ann Rev Pub Health*. 1995;15:327-54.

Williams A. Is the QALY a technical solution to a political problem? Of course not. *Int J Health Services.* 1991;21(2):365-9.

Ethical, Legal, Social, and Political Issues

Anderson GF, Hall MA, Steinberg EP. Medical technology assessment and practice guidelines: their day in court. *Am J Pub Health*. 1993;83(11):1635-9.

Banta HD, Andreasen PB. The political dimension in health care technology assessment programs. *Int J Technol Assess Health Care*. 1990;6(1):115-23.

Beauchamp TL, Childress JF. *Principles of Biomedical Ethics*. 4th ed. New York, NY: Oxford University Press; 1993.

Diamond GA. The importance of subjective values to medical technology. *J Invasive* Cardiol. 1993;5(1):17-22.

Ferguson JH, Dubinsky M, Kirsch PJ. Court-ordered reimbursement for unproven medical technology. *JAMA*. 1993;269(16):2116-21.

Gold JA, Zaremski MJ, Lev ER, Shefrin DH. Daubert v. Merrell Dow. The Supreme Court tackles scientific evidence in the courtroom. *JAMA*. 1993;270(24):2964-7.

Heitman E. Ethical issues in technology assessment. Conceptual categories and procedural considerations. *Int J Technol Assess Health Care*. 1998;14(3):544-66.

Lantos J. Ethics, randomization, and technology assessment. *Cancer*. 1994;74(suppl. 9):2653-6.

Passamani E. Clinical trials: are they ethical? *N Engl J Med*. 1991;324(22):1589-92.

Sibbald WJ, Inman KJ. Problems in assessing the technology of critical care medicine. *Int J Technol Assess Health Care*. 1992;8(3):419-43.

Tangri SS, Levin TE. New medical technologies: economic, ethical, and political considerations for policy. *J Soc Issues*. 1993;49(2):1-211.

Van der Wilt GJ, Reuzel R, Banta HD. The ethics of assessing health technologies. *Theor Med Bioeth.* 2000;21(1):103-15.

HTA and Quality of Care

Brook RH. Using scientific information to improve quality of health care. *Ann N Y Acad Sci*. 1993;703:74-84.

Donabedian A. Quality assessment and assurance: unity of purpose, diversity of means. *Inquiry*. 1988;25(1):173-92.

Donabedian A. The quality of care: how can it be assessed? *JAMA*. 1988;260(12):1743-8.

Lohr KN, ed. *Medicare: A Strategy for Quality Assurance*. Volume I. Institute of Medicine. Washington, DC: National Academy Press; 1990.

Lohr KN, Rettig RA, eds. *Quality of Care and Technology Assessment. Report of a Forum of the Council on Health Care Technology*. Washington, DC: National Academy Press; 1988.

McNeil BJ. Shattuck Lecture − Hidden barriers to improvement in the quality of care. *N Engl J Med*. 2001;345(22):1612-20.

McDonald IG. Quality assurance and technology assessment: pieces of a larger puzzle. *J Qual Clin Pract*. 2000;20(2-3):87-94.

Searching the HTA Literature

Auston I, Cahn MA, Selden CR. Literature search methods for the development of clinical practice guidelines. In *Clinical Practice Guideline Development: Methodology Perspectives*, eds., McCormick KA, Moore SR, Siegel RA. Rockville, Md: Agency for Health Care Policy Research; 1994.

Boynton J, Glanville J, McDaid D, et al. Identifying systematic reviews in MEDLINE: developing an objective approach to search strategy design. *J Inform Sci*. 1998;24:137-57.

Chan L, Topfer L. *Health Technology Assessment on the Net: A Guide to Internet Sources of Information.* Edmonton, Canada. Alberta Heritage Foundation for Medical Research, June 2003.

Counsell C. Formulating questions and locating primary studies for inclusion in systematic reviews. *Ann Intern Med*. 1997;127:380-7.

Glanville J, Lefebvre C. Identifying systematic reviews: key resources. *Evidence-Based Medicine*. 2000;3:68-9.

Glanville J, Wilson P, Richardson R. Accessing online evidence: a guide to key sources of research information on clinical and cost effectiveness. *Qual Saf Health Care*. 2003;12:229-31.

Goodman C. Literature Searching and Evidence Interpretation for Assessing Health Care Practices. Stockholm, Sweden: *Swedish Council on Technology Assessment in Health Care;* 1993.

Hopewell S, Clarke M, Lusher A, Lefebvre C, et al. A comparison of handsearching versus MEDLINE searching to identify reports of randomized controlled trials. *Stat Med*. 2002;21(11):1625-34.

Jadad AR, McQuay HJ. A high-yield strategy to identify randomized controlled trials for systematic reviews. *Online Journal of Current Clinical Trials*. 1993;Doc. No. 33.

National Library of Medicine, National Information Center on Health Services Research & Health Care Technology. *Etext on Health Technology Assessment (HTA) Information Resources.* http://www.nlm.nih.gov/nichsr/ehta. Accessed June 1, 2003.

NHS Centre for Reviews and Dissemination. Accessing the evidence on clinical effectiveness. *Effectiveness Matters*. 2001;5(1).

Suarez-Almazor ME, Belseck E, Homik J, et al. Identifying clinical trials in the medical literature with electronic databases: MEDLINE alone is not enough. *Control Clin Trials*. 2000;21:476-87.

Topfer LA, Parada A, Menon D, et al. Comparison of literature searches on quality and costs for health technology assessment using the MEDLINE and EMBASE databases. *Int J Technol Assess Health Care*. 1999;15(2):297-303.

GLOSSARY

Absolute risk reduction: a measure of treatment effect that compares the probability (or mean) of a type of outcome in the control group with that of a treatment group, [i.e.: $P_c - P_t$ (or $\mu_c - \mu_t$)]. For instance, if the results of a trial were that the probability of death in a control group was 25% and the probability of death in a treatment group was 10%, the absolute risk reduction would be (0.25 - 0.10) = 0.15. (See also **number needed to treat, odds ratio,** and **relative risk reduction.**)

Accuracy: the degree to which a measurement (e.g., the mean estimate of a treatment effect) is true or correct. An estimate can be accurate, yet not be precise, if it is based upon an unbiased method that provides observations having great variation (i.e., not close in magnitude to each other). (Contrast with **precision.**)

Alpha (α): the probability of a Type I (false-positive) error. In hypothesis testing, the α-level is the threshold for defining statistical significance. For instance, setting α at a level of 0.05 implies that investigators accept that there is a 5% chance of concluding incorrectly that an intervention is effective when it has no true effect. The α-level is commonly set at 0.01 or 0.05 or 0.10.

Benchmarking: a quality assurance process in which an organization sets goals and measures its performance in comparison to those of the products, services, and practices of other organizations that are recognized as leaders.

Beta (β): the probability of a Type II (false-negative) error. In hypothesis testing, β is the probability of concluding incorrectly that an intervention is not effective when it has true effect. (1-β) is the **power** to detect an effect of an intervention if one truly exists.

Bias: in general, any factor that distorts the true nature of an event or observation. In clinical investigations, a bias is any systematic factor other than the intervention of interest that affects the magnitude of (i.e., tends to increase or decrease) an observed difference in the outcomes of a treatment group and a control group. Bias diminishes the accuracy (though not necessarily the precision) of an observation. Randomization is a technique used to decrease this form of bias. Bias also refers to a prejudiced or partial viewpoint that would affect someone's interpretation of a problem. Double blinding is a technique used to decrease this type of bias.

Bibliographic database: an indexed computer or printed source of citations of journal articles and other reports in the literature. Bibliographic citations typically include author, title, source, abstract, and/or related information (including full text in some cases). Examples are *MEDLINE* and *EMBASE*.

Blinding: also known as "masking," the knowledge of patients and/or investigators about whether individual patients are receiving the investigational intervention(s) or the control (or standard) intervention(s) in a clinical trial. Blinding is intended to eliminate the possibility that knowledge of which intervention is being received will affect patient outcomes or investigator behaviors that may affect outcomes. Blinding is not always practical (e.g. when comparing surgery to drug treatment), but it should be used whenever it is possible and compatible with optimal patient care. A **single-blinded** trial is one in which this knowledge is withheld only from patients; a **double-blinded** trial is one in

which the knowledge is also withheld from investigators; and a triple-blinded trial is one in which the knowledge is also withheld from the statisticians or other analysts of trial data.

Case-control study: a retrospective observational study designed to determine the relationship between a particular outcome of interest (e.g., disease or condition) and a potential cause (e.g., an intervention, risk factor, or exposure). Investigators identify a group of patients with a specified outcome (cases) and a group of patients without the specified outcome (controls). Investigators then compare the histories of the cases and the controls to determine the rate or level at which each group experienced a potential cause. As such, this study design leads from outcome (disease or condition) to cause (intervention, risk factor, or exposure).

Case series: see **series**.

Case study: an uncontrolled (prospective or retrospective) observational study involving an intervention and outcome in a single patient. (Also known as a single case report or anecdote.)

Causal pathway: also known as an "analytical framework," a depiction (e.g., in a schematic) of direct and indirect linkages between interventions and outcomes. For a clinical problem, a causal pathway typically includes a patient population, one or more alternative interventions (e.g., screening, diagnosis, and/or treatment), intermediate outcomes (e.g., biological markers), and health outcomes. Causal pathways are intended to provide clarity and explicitness in defining the questions to be addressed in an assessment; they are useful in identifying pivotal linkages for which evidence may be lacking.

Citation: the record of an article, book, or other report in a bibliographic database that includes summary descriptive information, e.g., authors, title, abstract, source, and indexing terms.

Clinical pathway: a multidisciplinary set of daily prescriptions and outcome targets for managing the overall care of a specific type of patient, e.g., from pre-admission to post-discharge for patients receiving inpatient care. Clinical pathways often are intended to maintain or improve quality of care and decrease costs for patients in particular diagnosis-related groups.

Clinical practice guidelines: a systematically developed statement to assist practitioner and patient decisions about appropriate health care for one or more specific clinical circumstances. The development of clinical practice guidelines can be considered to be a particular type of HTA; or, it can be considered to be one of the types of policymaking that is informed or supported by HTA.

Clinical significance: a conclusion that an intervention has an effect that is of practical meaning to patients and health care providers. Even though an intervention is found to have a statistically significant effect, this effect might not be clinically significant. In a trial with a large number of patients, a small difference between treatment and control groups may be statistically significant but clinically unimportant. In a trial with few patients, an important clinical difference may be observed that does not achieve statistical significance. (A larger trial may be needed to confirm that this is a statistically significant difference.)

Cohort study: an observational study in which outcomes in a group of patients that received an intervention are compared with outcomes in a similar group i.e., the cohort, either contemporary or

historical, of patients that did not receive the intervention. In an adjusted- (or matched-) cohort study, investigators identify (or make statistical adjustments to provide) a cohort group that has characteristics (e.g., age, gender, disease severity) that are as similar as possible to the group that experienced the intervention.

Compliance: a measure of the extent to which patients undergo an assigned treatment or regimen, e.g., taking drugs, undergoing a medical or surgical procedure, doing an exercise regimen, or abstaining from smoking.

Concealment of allocation: the process used to assign patients to alternative groups in an RCT in a manner that prevents foreknowledge (by the person managing the allocation as well as the patients) of this assignment. Medical record numbers, personal identification numbers, or birthdays are not adequate for concealment of allocation. Certain centralized randomization schemes and sequentially numbered sealed, opaque envelopes are among adequate methods of allocation concealment.

Concurrent nonrandomized control: a control group that is observed by investigators at the same time as the treatment group, but that was not established using random assignment of patients to control and treatment groups. Differences in the composition of the treatment and control groups may result.

Confidence interval: depicts the range of uncertainty about an estimate of a treatment effect. It is calculated from the observed differences in outcomes of the treatment and control groups and the sample size of a study. The confidence interval (CI) is the range of values above and below the point estimate that is likely to include the true value of the treatment effect. The use of CIs assumes that a study provides one sample of observations out of many possible samples that would be derived if the study were repeated many times. Investigators typically use CIs of 90%, 95%, or 99%. For instance, a 95% CI indicates that there is a 95% probability that the CI calculated from a particular study includes the true value of a treatment effect. If the interval includes a null treatment effect (usually 0.0, but 1.0 if the treatment effect is calculated as an odds ratio or relative risk), the null hypothesis of no true treatment effect cannot be rejected.

Consensus development: various forms of group judgment in which a group (or panel) of experts interacts in assessing an intervention and formulating findings by vote or other process of reaching general agreement. These process may be informal or formal, involving such techniques as the nominal group and Delphi techniques.

Contraindication: a clinical symptom or circumstance indicating that the use of an otherwise advisable intervention would be inappropriate.

Control group: a group of patients that serves as the basis of comparison when assessing the effects of the intervention of interest that is given to the patients in the treatment group. Depending upon the circumstances of the trial, a control group may receive no treatment, a "usual" or "standard" treatment, or a placebo. To make the comparison valid, the composition of the control group should resemble that of the treatment group as closely as possible. (See also **historical control** and **concurrent nonrandomized control**.)

Controlled clinical trial: a prospective experiment in which investigators compare outcomes of a group of patients receiving an intervention to a group of similar patients not receiving the intervention. Not all clinical trials are RCTs, though all RCTs are clinical trials.

Controlled vocabulary: a system of terms, involving, e.g., definitions, hierarchical structure, and cross-references, that is used to index and retrieve a body of literature in a bibliographic, factual, or other database. An example is the *MeSH* controlled vocabulary used in *MEDLINE* and other *MEDLARS* databases of the NLM.

Cost-benefit analysis: a comparison of alternative interventions in which costs and outcomes are quantified in common monetary units.

Cost-consequence analysis: A form of cost-effectiveness analysis in which the components of incremental costs (of therapies, hospitalization, etc.) and consequences (health outcomes, adverse effects, etc.) of alternative interventions or programs are computed and displayed, without aggregating these results (e.g., into a cost-effectiveness ratio).

Cost-effectiveness analysis: a comparison of alternative interventions in which costs are measured in monetary units and outcomes are measured in non-monetary units, e.g., reduced mortality or morbidity.

Cost-minimization analysis: a determination of the least costly among alternative interventions that are assumed to produce equivalent outcomes.

Cost-utility analysis: a form of cost-effectiveness analysis of alternative interventions in which costs are measured in monetary units and outcomes are measured in terms of their utility, usually to the patient, e.g., using QALYs.

Cost-of-illness analysis: a determination of the economic impact of an disease or health condition, including treatment costs; this form of study does not address benefits/outcomes.

Crossover bias: occurs when some patients who are assigned to the treatment group in a clinical study do not receive the intervention or receive another intervention, or when some patients in the control group receive the intervention (e.g., outside the trial). If these crossover patients are analyzed with their original groups, this type of bias can "dilute" (diminish) the observed treatment effect.

Crossover design: a clinical trial design in which patients receive, in sequence, the treatment (or the control), and then, after a specified time, switch to the control (or treatment). In this design, patients serve as their own controls, and randomization may be used to determine the order in which a patient receives the treatment and control.

Cross-sectional study: a (prospective or retrospective) observational study in which a group is chosen (sometimes as a random sample) from a certain larger population, and the exposures of people in the group to an intervention and outcomes of interest are determined.

Database (or register): any of a wide variety of repositories (often computerized) for observations and related information about a group of patients (e.g., adult males living in Göteborg) or a disease (e.g.,

hypertension) or an intervention (e.g., antihypertensive drug therapy) or other events or characteristics. Depending upon criteria for inclusion in the database, the observations may have controls. Although these can be useful, a variety of confounding factors (e.g., no randomization and possible selection bias in the process by which patients or events are recorded) make them relatively weak methods for determining causal relationships between an intervention and an outcome.

Decision analysis: an approach to decision making under conditions of uncertainty that involves modeling of the sequences or pathways of multiple possible strategies (e.g., of diagnosis and treatment for a particular clinical problem) to determine which is optimal. It is based upon available estimates (drawn from the literature or from experts) of the probabilities that certain events and outcomes will occur and the values of the outcomes that would result from each strategy. A decision tree is a graphical representation of the alternate pathways.

Delphi technique: an iterative group judgment technique in which a central source forwards surveys or questionnaires to isolated, anonymous (to each other) participants whose responses are collated/summarized and recirculated to the participants in multiple rounds for further modification/critique, producing a final group response (sometimes statistical).

Direct costs: the fixed and variable costs of all resources (goods, services, etc.) consumed in the provision of an intervention as well as any consequences of the intervention such as adverse effects or goods or services induced by the intervention. Includes direct medical costs and direct nonmedical costs such as transportation or child care.

Disability-adjusted life years (DALYs): a unit of health care status that adjusts age-specific life expectancy by the loss of health and years of life due to disability from disease or injury. DALYs are often used to measure the global burden of disease.

Discounting: the process used in cost analyses to reduce mathematically future costs and/or benefits/outcomes to their present value. These adjustments reflect that given levels of costs and benefits occurring in the future usually have less value in the present than the same levels of costs and benefits realized in the present.

Discount rate: the interest rate used to discount or calculate future costs and benefits so as to arrive at their present values, e.g., 3% or 5%. This is also known as the opportunity cost of capital investment. Discount rates are usually based on government bonds or market interest rates for cost of capital whose maturity is about same as the time period during which the intervention or program being evaluated. For example, the discount rate used by the US federal government is based on the Treasury Department cost of borrowing funds and will vary, depending on the period of analysis.

Disease management: a systematic process of managing care of patients with specific diseases or conditions (particularly chronic conditions) across the spectrum of outpatient, inpatient, and ancillary services. The purposes of disease management may include: reduce acute episodes, reduce hospitalizations, reduce variations in care, improve health outcomes, and reduce costs. Disease management may involve continuous quality improvement or other management paradigms. It may involve a cyclical process of following practice protocols, measuring the resulting outcomes, feeding those results back to clinicians, and revising protocols as appropriate.

Dissemination: any process by which information is transmitted (made available or accessible) to intended audiences or target groups.

Effect size: same as **treatment effect**. Also, a dimensionless measure of treatment effect that is typically used for continuous variables and is usually defined as the difference in mean outcomes of the treatment and control group divided by the standard deviation of the outcomes of the control group. One type of meta-analysis involves averaging the effect sizes from multiple studies.

Effectiveness: the benefit (e.g., to health outcomes) of using a technology for a particular problem under general or routine conditions, for example, by a physician in a community hospital or by a patient at home.

Effectiveness research: see **outcomes research.**

Efficacy: the benefit of using a technology for a particular problem under ideal conditions, for example, in a laboratory setting, within the protocol of a carefully managed randomized controlled trial, or at a "center of excellence."

Endpoint: a measure or indicator chosen for determining an effect of an intervention.

Equipoise: a state of uncertainty regarding whether alternative health care interventions will confer more favorable outcomes, including balance of benefits and harms. Under the principle of equipoise, a patient should be enrolled in a randomized contolled trial only if there is substantial uncertainty, (an expectation for equal likelihood) about which intervention will benefit the patient most.

Evidence-based medicine: the use of current best evidence from scientific and medical research to make decisions about the care of individual patients. It involves formulating questions relevant to the care of particular patients, searching the scientific and medical literature, identifying and evaluating relevant research results, and applying the findings to patients.

Evidence table: a summary display of selected characteristics (e.g., of methodological design, patients, outcomes) of studies of a particular intervention or health problem.

External validity: the extent to which the findings obtained from an investigation conducted under particular circumstances can be generalized to other circumstances. To the extent that the circumstances of a particular investigation (e.g., patient characteristics or the manner of delivering a treatment) differ from the circumstances of interest, the external validity of the findings of that investigation may be questioned.

Factual database: an indexed computer or printed source that provides reference or authoritative information, e.g., in the form of guidelines for diagnosis and treatment, patient indications, or adverse effects.

False negative error: occurs when the statistical analysis of a trial detects no difference in outcomes between a treatment group and a control group when in fact a true difference exists. This is also known as a **Type II error**. The probability of making a Type II error is known as β (beta).

False positive error: occurs when the statistical analysis of a trial detects a difference in outcomes between a treatment group and a control group when in fact there is no difference. This is also known as a **Type I error**. The probability of a Type I error is known as α (alpha).

Follow-up: the ability of investigators to observe and collect data on all patients who were enrolled in a trial for its full duration. To the extent that data on patient events relevant to the trial are lost, e.g., among patients who move away or otherwise withdraw from the trial, the results may be affected, especially if there are systematic reasons why certain types of patients withdraw. Investigators should report on the number and type of patients who could not be evaluated, so that the possibility of bias may be considered.

Gray literature: research reports that are not found in traditional peer-reviewed publications, for example: government agency monographs, symposium proceedings, and unpublished company reports.

Health-related quality of life (HRQL) measures: patient outcome measures that extend beyond traditional measures of mortality and morbidity, to include such dimensions as physiology, function, social activity, cognition, emotion, sleep and rest, energy and vitality, health perception, and general life satisfaction. (Some of these are also known as health status, functional status, or quality of life measures.)

Health technology assessment (HTA): the systematic evaluation of properties, effects, and/or impacts of health care technology. It may address the direct, intended consequences of technologies as well as their indirect, unintended consequences. Its main purpose is to inform technology-related policymaking in health care. HTA is conducted by interdisciplinary groups using explicit analytical frameworks drawing from a variety of methods.

Health services research: a field of inquiry that examines the impact of the organization, financing and management of health care services on the delivery, quality, cost, access to and outcomes of such services.

Healthy-years equivalents (HYEs): the number of years of perfect health that are considered equivalent to (i.e., have the same utility as) the remaining years of life in their respective health states.

Historical control: a control group that is chosen from a group of patients who were observed at some previous time. The use of historical controls raises concerns about valid comparisons because they are likely to differ from the current treatment group in their composition, diagnosis, disease severity, determination of outcomes, and/or other important ways that would confound the treatment effect. It may be feasible to use historical controls in special instances where the outcomes of a standard treatment (or no treatment) are well known and vary little for a given patient population.

Hypothesis testing: a means of interpreting the results of a clinical trial that involves determining the probability that an observed treatment effect could have occurred due to chance alone if a specified hypothesis were true. The specified hypothesis is normally a **null hypothesis,** made prior to the trial, that the intervention of interest has no true effect. Hypothesis testing is used to determine if the null hypothesis can or cannot be rejected.

Incidence: the rate of occurrence of new cases of a disease or condition in a population at risk during a given period of time, usually one year.

Indication: a clinical symptom or circumstance indicating that the use of a particular intervention would be appropriate.

Indirect costs: the cost of time lost from work and decreased productivity due to disease, disability, or death. (In cost accounting, it refers to the overhead or fixed costs of producing goods or services.)

Intangible costs: the cost of pain and suffering resulting from a disease, condition, or intervention.

Intention to treat analysis: a type of analysis of clinical trial data in which all patients are included in the analysis based on their original assignment to intervention or control groups, regardless of whether patients failed to fully participate in the trial for any reason, including whether they actually received their allocated treatment, dropped out of the trial, or crossed over to another group.

Internal validity: the extent to which the findings of a study accurately represent the causal relationship between an intervention and an outcome in the particular circumstances of that study. The internal validity of a trial can be suspect when certain types of biases in the design or conduct of a trial could have affected outcomes, thereby obscuring the true direction, magnitude, or certainty of the treatment effect.

Investigational Device Exemption (IDE): a regulatory category and process in which the US Food and Drug Administration (FDA) allows specified use of an unapproved health device in controlled settings for purposes of collecting data on safety and efficacy/effectiveness; this information may be used subsequently in a premarketing approval application.

Investigational New Drug Application (IND): an application submitted by a sponsor to the US FDA prior to human testing of an unapproved drug or of a previously approved drug for an unapproved use.

Language bias: a form of bias that may affect the findings of a systematic review or other literature synthesis that arises when research reports are not identified or are excluded based on the language in which they are published.

Large, simple trials: prospective, randomized controlled trials that use large numbers of patients, broad patient inclusion criteria, multiple study sites, minimal data requirements, and electronic registries; their purposes include detecting small and moderate treatment effects, gaining effectiveness data, and improving external validity.

Literature review: a summary and interpretation of research findings reported in the literature. May include unstructured qualitative reviews by single authors as well as various systematic and quantitative procedures such as meta-analysis. (Also known as overview.)

Marginal benefit: the additional benefit (e.g., in units of health outcome) produced by an additional resource use (e.g., another health care intervention).

Marginal cost: the additional cost required to produce an additional unit of benefit (e.g., unit of health outcome).

Markov model: A type of quantitative modeling that involves a specified set of mutually exclusive and exhaustive states (e.g., of a given health status), and for which there are transition probabilities of moving from one state to another (including of remaining in the same state). Typically, states have a uniform time period, and transition probabilities remain constant over time.

Meta-analysis: systematic methods that use statistical techniques for combining results from different studies to obtain a quantitative estimate of the overall effect of a particular intervention or variable on a defined outcome. This combination may produce a stronger conclusion than can be provided by any individual study. (Also known as data synthesis or quantitative overview.)

Monte Carlo simulation: a technique used in computer simulations that uses sampling from a random number sequence to simulate characteristics or events or outcomes with multiple possible values. For example, this can be used to represent or model many individual patients in a population with ranges of values for certain health characteristics or outcomes. In some cases, the random components are added to the values of a known input variable for the purpose of determining the effects of fluctuations of this variable on the values of the output variable.

Moving target problem: changes in health care that can render the findings of HTAs out of date, sometimes before their results can be implemented. Included are changes in the focal technology, changes in the alternative or complementary technologies i.e., that are used for managing a given health problem, emergence of new competing technologies, and changes in the application of the technology (e.g., to different patient populations or to different health problems).

N of 1 trial: a clinical trial in which a single patient is the total population for the trial, including a single case study. An N of 1 trial in which random allocation is used to determine the order in which an experimental and a control intervention are given to a patient is an N of 1 RCT.

Negative predictive value: see **predictive value negative**.

New Drug Application (NDA): an application submitted by a sponsor to the FDA for approval to market a new drug (a new, nonbiological molecular entity) for human use in US interstate commerce.

Nonrandomized controlled trial: a controlled clinical trial that assigns patients to intervention and control groups using a method that does not involve randomization, e.g., at the convenience of the investigators or some other technique such as alternate assignment.

Nominal group technique: a face-to-face group judgment technique in which participants generate silently, in writing, responses to a given question/problem; responses are collected and posted, but not identified by author, for all to see; responses are openly clarified, often in a round-robin format; further iterations may follow; and a final set of responses is established by voting/ranking.

Null hypothesis: in hypothesis testing, the hypothesis that an intervention has no effect, i.e., that there is no true difference in outcomes between a treatment group and a control group. Typically, if statistical

tests indicate that the *P* value is at or above the specified a-level (e.g., 0.01 or 0.05), then any observed treatment effect is not statistically significant, and the null hypothesis cannot be rejected. If the P value is less than the specified a-level, then the treatment effect is statistically significant, and the null hypothesis is rejected. If a confidence interval (e.g., of 95% or 99%) includes zero treatment effect, then the null hypothesis cannot be rejected.

Number needed to treat: a measure of treatment effect that provides the number of patients who need to be treated to prevent one outcome event. It is the inverse of absolute risk reduction (1 ÷ absolute risk reduction); i.e., $1.0 \div (P_c - P_t)$. For instance, if the results of a trial were that the probability of death in a control group was 25% and the probability of death in a treatment group was 10%, the number needed to treat would be $1.0 \div (0.25 - 0.10) = 6.7$ patients. (See also **absolute risk reduction, relative risk reduction,** and **odds ratio.**)

Observational study: a study in which the investigators do not manipulate the use of, or deliver, an intervention (e.g., do not assign patients to treatment and control groups), but only observe patients who are (and sometimes patients who are not as a basis of comparison) exposed to the intervention, and interpret the outcomes. These studies are more subject to selection bias than experimental studies such as randomized controlled trials.

Odds ratio: a measure of treatment effect that compares the probability of a type of outcome in the treatment group with the outcome of a control group, i.e., $[P_t \div (1 - P_t)]$ $[P_c \div (1 - P_c)]$. For instance, if the results of a trial were that the probability of death in a control group was 25% and the probability of death in a treatment group was 10%, the odds ratio of survival would be $[0.10 \div (1.0 - 0.10)] \div [(0.25 \div (1.0 - 0.25)] = 0.33$. (See also **absolute risk reduction, number needed to treat,** and **relative risk.**)

Outcomes research: evaluates the impact of health care on the health outcomes of patients and populations. It may also include evaluation of economic impacts linked to health outcomes, such as cost effectiveness and cost utility. Outcomes research emphasizes health problem- (or disease-) oriented evaluations of care delivered in general, real-world settings; multidisciplinary teams; and a wide range of outcomes, including mortality, morbidity, functional status, mental well-being, and other aspects of health-related quality of life. It may entail any in a range of primary data collection methods and synthesis methods that combine data from primary studies.

***P* value:** in hypothesis testing, the probability that an observed difference between the intervention and control groups is due to chance alone if the null hypothesis is true. If *P* is less than the α-level (typically 0.01 or 0.05) chosen prior to the study, then the null hypothesis is rejected.

Parallel group (or independent group) trial: a trial that compares two contemporaneous groups of patients, one of which receives the treatment of interest and one of which is a control group (e.g., a randomized controlled trial). (Some parallel trials have more than one treatment group; others compare two treatment groups, each acting as a control for the other.)

Patient selection bias: a bias that occurs when patients assigned to the treatment group differ from patients assigned to the control group in ways that can affect outcomes, e.g., age or disease severity. If the two groups are constituted differently, it is difficult to attribute observed differences in their

outcomes to the intervention alone. Random assignment of patients to the treatment and control groups minimizes opportunities for this bias.

Peer review: the process by which manuscripts submitted to health, biomedical, and other scientifically oriented journals and other publications are evaluated by experts in appropriate fields (usually anonymous to the authors) to determine if the manuscripts are of adequate quality for publication.

Phase I, II, III, and IV studies: phases of clinical trials of new technologies (usually drugs) in the development and approval process required by the FDA (or other regulatory agencies). Phase I trials typically involve approximately 20-80 healthy volunteers to determine a drug's safety, safe dosage range, absorption, metabolic activity, excretion, and the duration of activity. Phase II trials are controlled trials in approximately 100-300 volunteer patients (with disease) to determine the drug's efficacy and adverse reactions (sometimes divided into Phase IIa pilot trials and Phase IIb well-controlled trials). Phase III trials are larger controlled trials in approximately 1,000-3,000 patients to verify efficacy and monitor adverse reactions during longer-term use (sometimes divided into Phase IIIa trials conducted before regulatory submission and Phase IIIb trials conducted after regulatory submission but before approval). Phase IV trials are postmarketing studies to monitor long-term effects and provide additional information on safety and efficacy, including for different regimens patient groups.

Placebo: an inactive substance or treatment given to satisfy a patient's expectation for treatment. In some controlled trials (particularly of drug treatments) placebos that are made to be indistinguishable by patients (and providers when possible) from the true intervention are given to the control group to be used as a comparative basis for determining the effect of the investigational treatment.

Placebo effect: the effect on patient outcomes (improved or worsened) that may occur due to the expectation by a patient (or provider) that a particular intervention will have an effect. The placebo effect is independent of the true effect (pharmacological, surgical, etc.) of a particular intervention. To control for this, the control group in a trial may receive a placebo.

Positive predictive value: see **predictive value positive**.

Power: the probability of detecting a treatment effect of a given magnitude when a treatment effect of at least that magnitude truly exists. For a true treatment effect of a given magnitude, power is the probability of avoiding Type II error, and is generally defined as $(1 - \beta)$.

Precision: the degree to which a measurement (e.g., the mean estimate of a treatment effect) is derived from a set of observations having small variation (i.e., close in magnitude to each other). A narrow confidence interval indicates a more precise estimate of effect than a wide confidence interval. A precise estimate is not necessarily an accurate one. (Contrast with **accuracy.**)

Predictive value negative: an operating characteristic of a diagnostic test; predictive value negative is the proportion of persons with a negative test who truly do not have the disease, determined as: [true negatives ÷ (true negatives + false negatives)]. It varies with the prevalence of the disease in the population of interest. (Contrast with **predictive value negative.**)

Predictive value positive: an operating characteristic of a diagnostic test; predictive value positive is the proportion of persons with a positive test who truly have the disease, determined as: [true positives ÷ (true positives + false positives)]. It varies with the prevalence of the disease in the population of interest. (Contrast with **predictive value negative.**)

Premarketing Approval (PMA) Application: an application made by the sponsor of a health device to the FDA for approval to market the device in US interstate commerce. The application includes information documenting the safety and efficacy/effectiveness of the device.

Prevalence: the number of people in a population with a specific disease or condition at a given time, usually expressed as a ratio of the number of affected people to the total population.

Primary study: an investigation that collects original (primary) data from patients, e.g., randomized controlled trials, observational studies, series of cases, etc. (Contrast with **synthetic/integrative study**).

Probability distribution: portrays the relative likelihood that a range of values is the true value of a treatment effect. This distribution often appears in the form of a bell-shaped curve. An estimate of the most likely true value of the treatment effect is the value at the highest point of the distribution. The area under the curve between any two points along the range gives the probability that the true value of the treatment effect lies between those two points. Thus, a probability distribution can be used to determine an interval that has a designated probability (e.g., 95%) of including the true value of the treatment effect.

Prospective study: a study in which the investigators plan and manage the intervention of interest in selected groups of patients. As such, investigators do not know what the outcomes will be when they undertake the study. (Contrast with **retrospective study**.)

Publication bias: unrepresentative publication of research reports that is not due to the quality of the research but to other characteristics, e.g., tendencies of investigators to submit, and publishers to accept, positive research reports (i.e., ones with results showing a beneficial treatment effect of a new intervention).

Quality-adjusted life year (QALY): a unit of health care outcomes that adjusts gains (or losses) in years of life subsequent to a health care intervention by the quality of life during those years. QALYs can provide a common unit for comparing cost-utility across different interventions and health problems. Analogous units include disability-adjusted life years (DALYs) and healthy-years equivalents (HYEs).

Quality assessment: a measurement and monitoring function of quality assurance for determining how well health care is delivered in comparison with applicable standards or acceptable bounds of care.

Quality assurance: activities intended to ensure that the best available knowledge concerning the use of health care to improve health outcomes is properly implemented. This involves the implementation of health care standards, including quality assessment and activities to correct, reduce variations in, or otherwise improve health care practices relative to these standards.

Quality of care: the degree to which health care is expected to increase the likelihood of desired health outcomes and is consistent with standards of health care. (See also **quality assessment** and **quality assurance**.)

Random variation (or random error): the tendency for the estimated magnitude of a parameter (e.g., based upon the average of a sample of observations of a treatment effect) to deviate randomly from the true magnitude of that parameter. Random variation is independent of the effects of systematic biases. In general, the larger the sample size is, the lower the random variation is of the estimate of a parameter. As random variation decreases, precision increases.

Randomization: a technique of assigning patients to treatment and control groups that is based only on chance distribution. It is used to diminish patient selection bias in clinical trials. Proper randomization of patients is an indifferent yet objective technique that tends to neutralize patient prognostic factors by spreading them evenly among treatment and control groups. Randomized assignment is often based on computer-generated tables of random numbers.

Randomized controlled trial (RCT): a prospective experiment in which investigators randomly assign an eligible sample of patients to one or more treatment groups and a control group and follow patients' outcomes. (Also known as **randomized clinical trial**.)

Receiver operating characteristic (ROC) curve: a graphical depiction of the relationship between the true positive ratio (sensitivity) and false positive ratio (1 - specificity) as a function of the cutoff level of a disease (or condition) marker. ROC curves help to demonstrate how raising or lowering the cutoff point for defining a positive test result affects tradeoffs between correctly identifying people with a disease (true positives) and incorrectly labeling a person as positive who does not have the condition (false positives).

Register: see **database**.

Reliability: the extent to which an observation that is repeated in the same, stable population yields the same result (i.e., test-retest reliability). Also, the ability of a single observation to distinguish consistently among individuals in a population.

Relative risk reduction: a type of measure of treatment effect that compares the probability of a type of outcome in the treatment group with that of a control group, i.e.: $(P_c - P_t) \div P_c$. For instance, if the results of a trial show that the probability of death in a control group was 25% and the probability of death in a control group was 10%, the relative risk reduction would be: $(0.25 - 0.10) \div 0.25 = 0.6$. (See also **absolute risk reduction, number needed to treat,** and **odds ratio**.)

Retrospective study: a study in which investigators select groups of patients that have already been treated and analyze data from the events experienced by these patients. These studies are subject to bias because investigators can select patient groups with known outcomes. (Contrast with **prospective study**.)

Safety: a judgment of the acceptability of risk (a measure of the probability of an adverse outcome and its severity) associated with using a technology in a given situation, e.g., for a patient with a particular health problem, by a clinician with certain training, or in a specified treatment setting.

Sample size: the number of patients studied in a trial, including the treatment and control groups, where applicable. In general, a larger sample size decreases the probability of making a false-positive error (α) and increases the power of a trial, i.e., decreases the probability of making a false-negative error (β). Large sample sizes decrease the effect of random variation on the estimate of a treatment effect.

Sensitivity: an operating characteristic of a diagnostic test that measures the ability of a test to detect a disease (or condition) when it is truly present. Sensitivity is the proportion of all diseased patients for whom there is a positive test, determined as: [true positives ÷ (true positives + false negatives)]. (Contrast with **specificity.**)

Sensitivity analysis: a means to determine the robustness of a mathematical model or analysis (such as a cost-effectiveness analysis or decision analysis) that tests a plausible range of estimates of key independent variables (e.g., costs, outcomes, probabilities of events) to determine if such variations make meaningful changes the results of the analysis. Sensitivity analysis also can be performed for other types of study; e.g., clinical trials analysis (to see if inclusion/exclusion of certain data changes results) and meta-analysis (to see if inclusion/exclusion of certain studies changes results).

Series: an uncontrolled study (prospective or retrospective) of a series (succession) of consecutive patients who receive a particular intervention and are followed to observe their outcomes. (Also known as case series or clinical series or series of consecutive cases.)

Specificity: an operating characteristic of a diagnostic test that measures the ability of a test to exclude the presence of a disease (or condition) when it is truly not present. Specificity is the proportion of non-diseased patients for whom there is a negative test, expressed as: [true negatives ÷ (true negatives + false positives)]. (Contrast with **sensitivity.**)

Statistical power: see **power**.

Statistical significance: a conclusion that an intervention has a true effect, based upon observed differences in outcomes between the treatment and control groups that are sufficiently large so that these differences are unlikely to have occurred due to chance, as determined by a statistical test. Statistical significance indicates the probability that the observed difference was due to chance if the null hypothesis is true; it does not provide information about the magnitude of a treatment effect. (Statistical significance is necessary but not sufficient for **clinical significance**.)

Statistical test: a mathematical formula (or function) that is used to determine if the difference in outcomes of a treatment and control group are great enough to conclude that the difference is statistically significant. Statistical tests generate a value that is associated with a particular *P* value. Among the variety of common statistical tests are: *F, t, Z,* and *chi-square.* The choice of a test depends upon the conditions of a study, e.g., what type of outcome variable used, whether or not the patients

were randomly selected from a larger population, and whether it can be assumed that the outcome values of the population have a normal distribution or other type of distribution.

Surrogate endpoint: an outcome measure that is used in place of a primary endpoint (outcome). Examples are decrease in blood pressure as a predictor of decrease in strokes and heart attacks in hypertensive patients, and increase in T-cell (a type of white blood cell) counts as an indicator of improved survival of AIDS patients. Use of a surrogate endpoint assumes that it is a reliable predictor of the primary endpoint(s) of interest.

Synthetic (or integrative) study: a study that does not generate primary data but that involves the qualitative or quantitative consolidation of findings from multiple primary studies. Examples are literature review, meta-analysis, decision analysis, and consensus development.

Systematic review: a form of structure literature review that addresses a question that is formulated to be answered by analysis of evidence, and involves objective means of searching the literature, applying predetermined inclusion and exclusion criteria to this literature, critically appraising the relevant literature, and extraction and synthesis of data from evidence base to formulate findings.

Technological imperative: the inclination to use a technology that has potential for some benefit, however marginal or unsubstantiated, based on an abiding fascination with technology, the expectation that new is better, and financial and other professional incentives.

Technology: the application of scientific or other organized knowledge--including any tool, technique, product, process, method, organization or system--to practical tasks. In health care, technology includes drugs; diagnostics, indicators and reagents; devices, equipment and supplies; medical and surgical procedures; support systems; and organizational and managerial systems used in prevention, screening, diagnosis, treatment and rehabilitation.

Time lag bias: a form of bias that may affect identification of studies to be included in a systematic review; occurs when the time from completion of a study to its publication is affected by the direction (positive vs. negative findings) and strength (statistical significance) of its results.

Treatment effect: the effect of a treatment (intervention) on outcomes, i.e., attributable only to the effect of the intervention. Investigators seek to estimate the true treatment effect using the difference between the observed outcomes of a treatment group and a control group. (See **effect size**.)

Type I error: same as **false-positive error**.

Type II error: same as **false-negative error.**

Utility: the relative desirability or preference (usually from the perspective of a patient) for a specific health outcome or level of health status.

Validity: The extent to which a measure accurately reflects the concept that it is intended to measure. See **internal validity** and **external validity**.

REFERENCES

Academy of Managed Care Pharmacy. *Format for Formulary Submissions. Version 2.0.* Alexandria, VA, October 2002.

Agema WR, Jukema JW, Zwinderman AH, van der Wall EE. A meta-analysis of the angiotensin-converting enzyme gene polymorphism and restenosis after percutaneous transluminal coronary revascularization: evidence for publication bias. *Am Heart J* 2002;144(5):760-8.

Agency for Healthcare Research and Quality. Nominations of topics for evidence-based practice centers (EPCs). *Federal Register* 68(18), January 28, 2003:4213-6.

Anderson, GF, MA Hall, EP Steinberg. Medical technology assessment and practice guidelines: their day in court. *Am J Pub Health*. 1993;83(3):1635-9.

Anderson JE, Jorenby DE, Scott WJ, Fiore MC. Treating tobacco use and dependence: an evidence-based clinical practice guideline for tobacco cessation. *Chest.* 2002;121(3):932-41.

Andersson I, Aspegren K, Janzon L, Landberg T, Lindholm K, Linell F, Ljungberg O, Ronstam J, and Sigfusson B. Mammographic Screening and mortality from breast cancer: The Malmö Mammographic Screening Trial. *BMJ*. 1988; 297(6654):943-8.

Anis AH, Rahman T, Schechter MT. Using pharmacoeconomic analysis to make drug insurance coverage decisions. *Pharmacoeconomics.* 1998;13(1 Pt 2):119-26.

Anonymous. Panel questions rules on trauma research. *The New York Times*; April 24, 1994:10.

Arnesen T, Nord E. The value of DALY life: problems with ethics and validity of disability adjusted life years. *BMJ.* 2000;320:1398.

Asch SM, Sloss EM, Hogan C, Brook RH, Kravitz RL. Measuring underuse and necessary care among elderly Medicare beneficiaries using inpatient and outpatient claims. *JAMA.* 2000;284:2325-33.

Aspirin Myocardial Infarction Study Research Group. A randomized, controlled trial of aspirin in persons recovered from myocardial infarction. *JAMA.* 1980;243(7): 661-9.

Auston I, Cahn MA, Selden CR. Literature Search Methods for the Development of Clinical Practice Guidelines. In *Clinical Practice Guideline Development: Methodology Perspectives*, McCormick KA, Moore SR, Siegel RA, 123-7. Rockville Md: Agency for Health Care Policy and Research, 1994.

Baidas S, Tfayli A, Bhargava P. Thalidomide: an old drug with new clinical applications. *Cancer Invest.* 2002;20(5-6):835-48.

Balk EM, Bonis PAL, Moskowitz H, et al. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. *JAMA.* 2002;287:2973-82.

Banta HD, Luce BR. *Health Care Technology and Its Assessment: An International Perspective*. New York, NY: Oxford University Press; 1993.

Basile J. Hypertension in the elderly: a review of the importance of systolic blood pressure elevation. *J Clin Hypertens*. 2002;4(2):108-12.

Beebe B, Rosenfeld AB, Collins N. An approach to decisions about coverage of investigational treatments. *HMO Practice*. 1997;11(2):65-7.

Bekelman JE, Li Y, Gross CP. Scope and impact of financial conflicts of interest in biomedical research: a systematic review. *JAMA*. 2003;289:454-65.

Berger A. High dose chemotherapy offers little benefit in breast cancer. *BMJ*. 1999 May 29;318(7196):1440.

Berger RL, Celli BR, Meneghetti AL, Bagley PH, et al. Limitations of randomized clinical trials for evaluating emerging operations: the case of lung volume reduction surgery. *Ann Thorac Surg*. 2001 Aug;72(2):649-57.

Bergner M, Bobbitt RA, Carter WB, Gilson BS. The Sickness Impact Profile: development and final revision of a health status measure. *Med Care*. 1981;19(8):787-805.

Bloom BS, Fendrick AM. The tension between cost containment and the underutilization of effective health services. *Int J Technol Assess Health Care*. 1996;12:1-8.

Boer GJ, Widner H. Clinical neurotransplantation: core assessment protocol rather than sham surgery as control. *Brain Res Bull*. 2002;58:547-53.

Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Clin Chem.* 2003;49(1):1-6.

Brenner M, Jones B, Daneschvar HL, Triff S. New National Emphysema Treatment Trial paradigm of health care financing administration-sponsored clinical research trials: advances and dilemmas. *J Investig Med.* 2002;Mar;50(2):95-100.

Briggs A, Clark T, Wolstenholme J, Clarke P. Missing… presumed at random: cost-analysis of incomplete data. *Health Econ.* 2003;12:377-92.

Brooks H, Bowers R. The assessment of technology. *Science*.1970;222(2):13-20.

Brownman GP. Development and aftercare of clinical guidelines: the balance between rigor and pragmatism. *JAMA*. 2001;286:1509-11.

Bunch WH, Dvonch VM. The value of functional independence measure scores. *Am J Phys Med Rehab*. 1994;73(1):40-3.

Buring JE, Jonas MA, Hennekens CH. Large and simple randomized trials. In *Tools for Evaluating Health Technologies: Five Background Papers*. US Congress, Office of Technology Assessment, 1995;167-91. BP-H-142. Washington, DC: US Government Printing Office; 1994.

Busse R, Orvain J, Velasco M, et al. Best practice in undertaking and reporting health technology assessments. *Int J Technol Assess Health Care*. 2002;18:361-422.

Buxton MJ. Problems in the economic appraisal of new health technology: the evaluation of heart transplants in the UK. In *Economic Appraisal of Health Technology in the European Community*. Drummond MF, ed. Oxford, England. Oxford Medical Publications, 1987.

Byford S, Palmer S. Common errors and controversies in pharmacoeconomic analyses. *Pharmacoeconomics.* 1998;13:659-66.

Carlsson P, Jørgensen T. Scanning the horizon for emerging health technologies. *Int J Technol Assess Health Care*. 1998;14:695-704.

Caro JJ, Caro G, Getsios D, et al. The migraine ACE model: evaluating the impact on time lost and medical resource use. *Headache*. 2002;40:282-91.

Cebul RD Randomized, controlled trials using the metro firm system. *Med Care*. 1991;29(suppl. 7):JS9-18.

Center for Evidence-Based Medicine. http://www.cebm.net/levels_of_evidence.asp#levels. Accessed June 1, 2003.

Chalmers TC, Frank CS, Reitman DR. Minimizing the three stages of publication bias. *JAMA*. 1990;263(10):1392-5.

Chan L, Topfer L. *Health Technology Assessment on the Net: A Guide to Internet Sources of Information.* Edmonton, Canada: Alberta Heritage Foundation for Medical Research; June 2003.

Chopra SS. Industry funding of clinical trials: Benefit or bias? *JAMA*. 2003;290:113.

Choudhdry NK, Stelfox HT, Detsky AS. Relationships between authors of clinical practice guidelines and the pharmaceutical industry. *Lancet*. 2002;287:612-7.

Chow S, Liu J. *Design and Analysis of Clinical Trials.* New York, NY: John Wiley & Sons; 1998.

Clarke M. Co-Chair, Cochrane Collaboration Steering Group. Personal communication; July 2, 2003.

Coates & Jarratt, Inc. *Course Workbook: Technology Assessment. Anticipating the Consequences of Technological Choices*. 1992. Washington, DC.

Coleman RE, Briner WH, Siegel BA. Clinical PET scanning. A short-lived orphan. *Int J Technol Assess Health Care*. 1992;8(4):610-22.

Combe B. Thalidomide: new indications? *Joint Bone Spine*. 2001;68:582-7.

Cook DJ, Guyatt GH, Laupacis A, Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. *Chest*. 1992;102(suppl. 4):305S-11S.

Cooper JD. Paying the piper: the NETT strikes a sour note. National Emphysema Treatment Trial. *Ann Thorac Surg*. 2001;Aug;72(2):330-3.

Coplen SE, Antman EM, Berlin JA, Hewitt P, Chalmers TC. Efficacy and safety of quinidine therapy for maintenance of sinus rhythm after cardioversion. A meta-analysis of randomized control trials. *Circulation*. 1990;82:1106-16.

Cutler DM, McClellan M. Is technological change in medicine worth it? *Health Aff*. 2001;20(5):11-29.

de Bruin AF, de Witte LP, Stevens F, Diederiks JP. Sickness Impact Profile: the state of the art of a generic functional status measure. *Soc Sci Med*. 1992;35(8):1003-14.

Deeks JJ. Systematic reviews in health care: systematic reviews of evaluations of diagnostic and screening tests. *BMJ*. 2001;323:157-62.

Detsky AS. Using cost-effectiveness analysis to improve the efficiency of allocating funds to clinical trials. *Stat Med*. 1990;9(1)-2:173-84.

Detsky AS, Naylor CD, O'Rourke K, McGeer AJ, L'Abbe KA. Incorporating variations in the quality of individual randomized trials into meta-analysis. *J Clin Epid*. 1992;45(3):255-65.

Deyo RA. Cascade effects of medical technology. *Annu Rev Public Health*. 2002;23:23-44.

Deyo RA, Psaty BM, Simon G, Wagner EH, Omenn GS. The messenger under attack − intimidation of researchers by special-interest groups. *N Engl J Med*. 1997;336(16):1176-80.

Dickersin K. How important is publication bias? *AIDS Educ Prev*. 1997;9(suppl 1):15-21.

Dickersin K, Min YI. Publication bias: the problem that won't go away. *Ann N Y Acad Sci*. 1993;703,:135-46.

Dickersin K, Scherer R, Lefebvre C. Identifying relevant studies for systematic reviews. *BMJ*. 1994;309(6964):1286-91.

Djulbegovic B, Lacevic M, Cantor A, et al. The uncertainty principle and industry-sponsored research. *Lancet* 2000;356:635-8.

Dolan P. Utilitarianism and the measurement and aggregation of quality-adjusted life years. *Health Care Anal*. 2001;9(1):65-76.

Doll H.A, Black NA, Flood AB, McPherson K. Criterion validation of the Nottingham health profile: patient views of surgery for benign prostatic hypertrophy. *Soc Sci Med*. 1993;37(1):115-22.

Donabedian A. Quality assessment and assurance: unity of purpose, diversity of means. *Inquiry*. 1988;25(1):173-92.

Donaldson MS, Sox HC, Jr, eds. *Setting Priorities for Health Technology Assessment: A Model Process*. Washington, DC: National Academy Press; 1992. Reprinted with permission from the National Academy of Sciences, courtesy of the National Academies Press, Washington, DC.

Dougherty CJ. Ethical perspectives on prospective payment. *Hastings Center Report*. 1989;19(January/February):5-11.

Douw K, Vondeling H, Eskildensen D, Simpson S. Use of the Internet in scanning the horizon for new and emerging health technologies; a survey involved in horizon scanning. *J Med Internet Res.* 2003;5(1):e6.

Drummond MF, Jefferson TO. Guidelines for authors and peer reviewers of economic submissions to the BMJ. The BMJ Economic Evaluation Working Party. *BMJ*. 1996;313:275-83.

Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. *Lancet*. 1991;337:867-72.

Eckman MH, Levine HJ, Pauker SG. Decision analytic and cost-effectiveness issues concerning anticoagulant prophylaxis in heart disease. *Chest*. 1992;102, (suppl. 4):538-49S.

Eddy DM. Selecting technologies for assessment. *Int J Technol Assess Health Care*.1989;5(4):485-501.

Eddy DM. *A Manual for Assessing Health Practices & Designing Practice Policies: The Explicit Approach*. Philadelphia, Pa: American College of Physicians, 1992.

Egger M, Davey Smith G, Altman DG, eds. *Systematic Reviews in Health Care: Meta-analysis in Context. 2nd ed.* London, England: BMJ Books; 2001.

Egger M, Davey Smith G, Sterne JA. Uses and abuses of meta-analysis. *Clin Med*. 2001;1:478-84.

Eisenberg JM. Ten lessons for evidence-based technology assessment. *JAMA*. 1999;17:1865-9.

Elixhauser A, Halpern M, Schmier J, Luce BR. Health care CBA and CEA from 1991 to 1996: an updated bibliography. *Med Care*. 1998;36(suppl. 5):MS1-9,MS18-147.

Ellenberg SS. Do large, simple trials have a place in the evaluation of aids therapies? *Oncology*. 1992;6:55-9,63.

Elwood PC. British studies of aspirin and myocardial infarction. *Am J Med*.1983;74(6A):50-4.

Elwood PC, Cochrane AL, Burr ML, et al. A randomized controlled trial of acetyl salicylic acid in the secondary prevention of mortality from myocardial infarction. *BMJ*. 1974;1(905):436-40.

Elwood PC, Sweetnam PM. Aspirin and secondary mortality after myocardial infarction. *Lancet*. 1979;2(8156-8157):1313-5.

Enkin M, Keirse MJNC, Renfrew M, Neilson J. *A Guide to Effective Care in Pregnancy and Childbirth. Second Edition*. New York, NY: Oxford University Press; 1995.

Essink-Bot ML, Stouthard ME, Bonsel GJ. Generalizability of valuations on health states collected with the EuroQol© questionnaire. *Health Economics*. 1993;2(3):237-46.

EuroQol© Group. "EuroQol© − A new facility for the measurement of health related quality of life. *Health Policy*.1990;16(3):199-208.

Feeny D, Guyatt G, Tugwell P, eds. *Health care Technology: Effectiveness, Efficiency, and Public Policy*. Montreal, Canada: Institute for Research on Public Policy; 1986.

ECRI. *Health Technology Forecast.* Plymouth Meeting, Pa: ECRI; 2002.

Egger M, Juni P, Bartlett C, Holenstein F, Sterne J. How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study. *Health Technol Assess*. 2003;7(1):1-76.

Ferguson JH. NIH consensus conferences: dissemination and impact. *Ann N Y Acad Sci*. 1993;703:180-98.

Ferguson JH, Dubinsky M, Kirsch PJ. Court-ordered reimbursement for unproven medical technology. *JAMA*. 1993;269(16):2116-21.

Ferguson JH, Sherman CR. Panelists' views of 68 NIH consensus conference. *Int J Technol Assess Health Care.* 2001;17(4):542-58.

Ferguson TB Jr, Peterson ED, Coombs LP, et al. Use of continuous quality improvement to increase use of process measures in patients undergoing coronary artery bypass graft surgery: A randomized controlled trial. *JAMA.* 2003;290:49-56.

Fineberg HV, Bauman R, Sosman M. Computerized cranial tomography: effect on diagnostic and therapeutic plans. *JAMA*. 1977;238(3):224-30.

Fink A, Kosecoff J, Chassin M, Brook RH. Consensus methods: characteristics and guidelines for use. *Am J Pub Health*. 1984;74(9):979-83.

Fishman A, Martinez F, Naunheim K, et al; National Emphysema Treatment Trial Research Group. A randomized trial comparing lung-volume-reduction surgery with medical therapy for severe emphysema. *N Engl J Med*. 2003 May 22;348(21):2059-73.

Fletcher SW. Whither scientific deliberation in health policy recommendations? *N Engl J Med.* 1997;16:1180-3.

Fletcher SW, Colditz GA. Failure of estrogen plus progestin therapy for prevention. *JAMA*. 2002;288:366-8.

Forrow L, Taylor WC, Arnold RM. Absolutely relative: how research results are summarized can affect treatment decisions. *Am J Med*. 1992;92(2):121-94.

Foulds J. Effectiveness of smoking cessation initiatives. Smoking cessation services show good return on investment. *BMJ*. 2002 Mar 9;324(7337):608-9.

Friedberg M, Saffran B, Stinson TJ, et al. Evaluation of conflict of interest in economic analyses of new drugs used in oncology. *JAMA*. 1999;282:1453-7.

Freiman JA, Chalmers TC, Smith H, Jr., Kuebler RR. The importance of beta, the Type II error and sample size in the design and interpretation of the randomized controlled trial. survey of 71 "negative" trials. *N Engl J Med*. 1978;299(13):690-4.

Fye WB. The power of clinical trials and guidelines,and the challenge of conflicts of interest. *J Am Coll Cardiol*. 2003;41:1237-42.

Gallagher M, Hares T, Spencer J, Bradshaw C, Webb I. The nominal group technique: a research tool for general practice? *Family Practice*. 1993;10(1):76-81.

Gann MJ, Restuccia JD. Total quality management in health care: a view of current and potential research. *Med Care Rev*. 1994;51(4):467-500.

Gazelle GS, Hunink MG, Kuntz KM, et al. Cost-effectiveness of hepatic metastasectomy in patients with metastatic colorectal carcinoma: a state-transition Monte Carlo decision analysis. *Ann Surg*. 2003;237:544-55.

Gelijns A, Rosenberg N. The dynamics of technological change in medicine. *Health Aff*. 1994;13(3):28-46.

Gerard K, Mooney G. QALY league tables: handle with care. *Health Economics*. 1993;2(1):59-64.

Gibson JL, Martin DK, Singer PA. Priority setting for new technologies in medicine: a transdisciplinary study. *BMC Health Serv Res*. 2002;2(1):14

Glennie JL, Torrance GW, Baladi JF, et al. The revised Canadian guidelines for the economic evaluation of pharmaceuticals. *Pharmacoeconomics*. 1999;15(5):459-68.

Gold JA, Zaremski MJ, Lev ER, Shefrin DH. Daubert v. Merrell Dow. The Supreme Court tackles scientific evidence in the courtroom. *JAMA*. 1993;270(24):2964-7.

Gold MR, Stevenson D, Fryback DG. HALYS and QALYS and DALYS, oh my: similarities and differences in summary measures of population health. *Annu Rev Public Health*. 2002;23:115-34.

Gold MR, Siegel JE, Russell LB, Weinstein MC. *Cost-Effectiveness in Health and Medicine.* New York, NY: Oxford University Press; 1996.

Goldberg HI, Cummings MA, Steinberg EP, et al. Deliberations on the dissemination of PORT products: translating research findings into improved patient outcomes. *Med Care*. 1994;32(suppl. 7):JS90-110.

Goodman C, ed. *Medical Technology Assessment Directory: A Pilot Reference to Organizations, Assessments, and Information Resources*. Washington, DC: Institute of Medicine; 1988.

Goodman C. It's time to rethink health care technology assessment. *Int J Technol Assess Health Care*. 1992;8(2):335-58.

Goodman C. *Literature Searching and Evidence Interpretation for Assessing Health Care Practices*. Stockholm, Sweden: Swedish Council on Technology Assessment in Health Care; 1993.

Goodman C. The moving target problem and other lessons from percutaneous transluminal coronary angioplasty. In: A Szczepura, Kankaanpää J. *Assessment of Health Care Technologies: Case Studies, Key Concepts and Strategic Issues.* New York, NY: John Wiley & Sons; 1996:29-65.

Goodman C, Baratz SR, eds. *Improving Consensus Development for Health Technology Assessment: An International Perspective*. Washington, DC: National Academy Press; 1990.

Goodman C, Snider G, Flynn K. *Health Care Technology Assessment in VA*. Boston, Mass: Management Decision and Research Center. Washington, DC: Health Services Research and Development Service; 1996.

Goodman CS. Healthcare technology assessment: methods, framework, and role in policy making. *Am J Manag Care*. 1998;4:SP200-14.

Goodman CS. Technology assessment in healthcare: a means for pursuing the goals of biomedical engineering. *Med Biolog Eng Comput*. 1993;31(1):HTA3-10.

Gøtsche PC. Multiple publication of reports of drug trials. *Eur J Clin Pharmacol*. 1989;36:429-32.

Graves N, Walker D, Raine R, et al. Cost data for individual patients included in clinical studies: no amount of statistical analysis can compensate for inadequate costing methods. *Health Economics.* 2002;11:735-9.

Gregorian RS, Golden KA, Bahce A, Goodman C, Kwong WJ, Khan ZM. Antidepressant-induced sexual dysfunction. *Ann Pharmacother.* 2002;36:1577-89.

Grimes DA. Technology follies: the uncritical acceptance of medical innovation. *JAMA*. 1993;269(23):3030-3.

Guyatt GH, Haynes RB, Jaeschke RZ, et al. Users' guide to the medical literature, XXV: Evidence-based medicine: principles for applying the users' guides to patient care. Evidence-Based Medicine Working Group. *JAMA*. 2000;284:1290-6.

Harbour R, Miller J. A new system for grading recommendations in evidence based guidelines. *BMJ.* 2001;323:334-6.

Harper G, Townsend J, Buxton M. The preliminary economic evaluation of the health technologies for the prioritization of health technology assessments. *Int J Technol Assess Health Care*. 1998;14:652-62.

Harris RP, Helfand M, Woolf SH, et al. Current methods of the US Preventive Services Task Force. A review of the process. *Am J Prev Med*. 2001;20(3S):21-35.

Hasselblad V, Hedges LV. Meta-analysis of screening and diagnostic tests. *Psychol Bull*. 1995;117(1): 167-78.

Hayward RSA., Wilson MC, Tunis SR, Bass EB, Guyatt G. Users' guides to the medical literature. VIII. How to use clinical practice guidelines. A. Are the recommendations valid? *JAMA.* 1995;274(7):570-4.

Heitman E. Ethical issues in technology assessment. Conceptual categories and procedural considerations. Int *J Technol Assess Health Care.* 1998;14(3):544-66.

Henry D, Lopert R. Pharmacoeconomics and policy decisions: The Australian health care system. *Clin Ther.* 1999;21:909-15.

Henshall C, Koch P, von Below GC, Boer A, et al. Health technology assessment in policy and practice. *Int J Technol Assess Health Care.* 2002;18(2):447-55.

Hilborne LH, Leape LL, Kahan JP, Park RE, Kamberg CJ, Brook RH. *Percutaneous Transluminal Coronary Angioplasty: A Literature Review of Ratings of Appropriateness and Necessity*. Santa Monica, Ca: RAND; 1991.

Hill SR, Mitchell AS, Henry DA. Problems with the interpretation of pharmacoeconomic analyses: a review of submissions to the Australian Pharmaceutical Benefits Scheme. *JAMA.* 2000;283(16):2116-21.

Hjelmgren J, Berggren F, Andersson F. Health economic guidelines--similarities, differences and some implications. *Value Health.* 2001;4(3):225-50.

Hoffman B. Is there a technological imperative in health care? *Int J Technol Assess Health Care.* 2002;18(3):675-89.

Hogness JR,Van Antwerp M. *The Artificial Heart: Prototypes, Policies, and Patients*. Washington, DC: National Academy Press; 1991. Reprinted with permission from the National Academy of Sciences, courtesy of the National Academies Press, Washington, DC.

Horng S, Miller FG. Ethical framework for the use of sham procedures in clinical trials. *Crit Care Med*. 2003;31(suppl. 3):S126-30.

Hunt DL, Jaeschke R, McKibbon KA. Users' guides to the medical literature: XXI. Using electronic health information resources in evidence-based practice. Evidence-Based Medicine Working Group. *JAMA* 2000;283(14):1875-9.

International Network of Agencies for Health Technology Assessment (INAHTA). http://www.inahta.org. Accessed June 1, 2003.

Institute for the Future. *Health & Health Care 2010. The Forecast, The challenge.* San Francisco, Ca: Jossey-Bass Publishers; 2000.

Institute of Medicine. *Assessing Medical Technologies*. Washington, DC: National Academy Press; 1985.

Institute of Medicine. *Assessment of Diagnostic Technology in Health Care. Rationale,Methods, Problems, and Directions.* Washington, DC: National Academy Press; 1989. Reprinted with permission from the National Academy of Sciences, courtesy of the National Academies Press, Washington, DC.

Institute of Medicine. *Consensus Development at the NIH: Improving the Program*. Washington, DC: National Academy Press; 1990.

Institute of Medicine, Committee on Quality of Health Care in America. *Crossing the Quality Chasm: A New Health System for the 21st Century.* Washington, DC: National Academy Press; 2001.

International Committee of Medical Journal Writers. Conflict of interest. *Lancet*. 1993;341(8847):742-3.

International Committee of Medical Journal Writers. Uniform requirements for manuscripts submitted to biomedical journals. *N Engl J Med*. 1997;336:309-15.

Ioannidis JPA. Effects of statistical significance of results on the time to completion and publication of randomized efficacy trials. *JAMA*. 1998;279:281-6.

Irwig L, Tosteson AN, Gatsonis C, Lau J, Colditz G, Chalmers TC, Mosteller F. Guidelines for meta-analyses evaluating diagnostic tests. *Ann Intern Med*. 1994;120(8):667-76.

Jadad AR, Moore RA, Carrol D, et al. Asssessing the quality of reports of randomized clinical trials: Is blinding necessary? *Controlled Clin Trials*. 1996;17:1-12.

Jairath N, Weinstein J. The Delphi methodology (part two): a useful administrative approach. *Can J Nurs Admin*. 1994;7(4):7-20.

Jenkinson C, Fitzpatrick R, Argyle M. The Nottingham Health Profile: an analysis of its sensitivity to differentiating illness groups. *Soc Sci Med*. 1988;27(12):1411-4.

Johannesson M, Pliskin JS, Weinstein MC. Are healthy-years equivalents an improvement over quality-adjusted life years? *Med Dec Making*. 1993;13(4):281-6.

Jorenby DE, Leischow SJ, Nides MA, Rennard SI, et al. A controlled trial of sustained-release bupropion, a nicotine patch, or both for smoking cessation. *N Engl J Med*. 1999 Mar 4;340(9):685-91.

Juni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA*. 1999;182:1054-60.

Kaden RJ, Vaul JH, Palazola PA. Negotiating payment for new technology purchases. *Healthc Financ Manage.* 2002;56(12):44-8.

Kahan JP, Bernstein SJ, Leape LL, Hilborne LH, Park RE, Parker L, Kamberg CJ, Brook RH. Measuring the necessity of medical procedures. *Med Care*. 1994;32(4):357-65.

Kahan JP, Kanouse DE, Winkler JD. Stylistic variations in National Institutes of Health consensus statements, 1979-1983. *Int J Technol Assess Health Care*. 1988;4(2):289-304.

Kaplan RM, Anderson JP. A general health policy model: update and applications. *Health Services Research*. 1988;23,:203-35.

Kaplan RM, Anderson JP, Wu AW, Mathews WC, Kozin F, Orenstein D. The Quality of Well-being Scale: applications in AIDS, cystic fibrosis, and arthritis. *Med Care*. 1989;27(suppl. 3):S27-43.

Kassirer JP, Angell M. Financial conflicts of interest in biomedical research. *N Engl J Med*. 1993;329:570-1.

Kassirer JP, Angell M. The journal's policy on cost-effectiveness analyses. *N Engl J Med*. 1994;331:669-70.

Katz S, Downs TD, Cash HR, Grotz RC. Progress in the development of the index of ADL. *Gerontologist*. 1970;10(part I):20-30.

Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. *Med Care.* 1989;27(suppl. 3):S178-189.

Kim S, Losina E, Solomon DH, Wright J, Katz JN. Effectiveness of clinical pathways for total knee and total hip arthroplasty: literature review. *J Arthroplasty*. 2003;18(1):69-74.

Kleinman LC, Kosecoff J, Dubois RW, Brook RH. The medical appropriateness of tympanostomy tubes proposed for children younger than 16 years in the United States. *JAMA*. 1994;271(16):1250-5.

Kwan J, Sandercock P. In-hospital care pathways for stroke: a Cochrane systematic review. *Stroke.* 2003;34(2):587-8.

Laird NM, Mosteller F. Some statistical methods for combining experimental results. *Int J Technol Assess Health Care.* 1990;6(1):5-30.

Lalezari JP, Henry K, O'Hearn M, et al. Enfuvirtide, an HIV-1 fusion inhibitor, for drug-resistant HIV infection in North and South America. *N Engl J Med*. 2003;348:2175-85.

Lara ME, Goodman C, eds. *National Priorities for the Assessment of Clinical Conditions and Medical Technologies*. Washington, DC: National Academy Press; 1990.

Lau J, Antman EM, Jiminez-Silva J, Kupelnick B, Mosteller F, Chalmers TC. Cumulative meta-analysis of therapeutic trials for myocardial infarction. *N Engl J Med*. 1992;327(4):248-54. Copyright © 1992 Massachusetts Medical Society. All rights reserved.

Lazaridis EN, Rudberg MA, Furner MA, Cassel CK. Do activities of daily living have a hierarchical structure? An analysis using the longitudinal study of aging. *J Gerontol*. 1994;49(2):M47-51.

The Lewin Group. *Outlook for Medical Technology Innovation. Report 2: The Medicare Payment Process and Patient Access to Technology*. Washington, DC: AdvaMed; 2000.

The Lewin Group. *Outlook for Medical Technology Innovation. Report 4: The Impact of Regulation and Market Dynamics on Innovation*. Washington, DC: AdvaMed; 2001.

The Lewin Group. *The Clinical Review Process Conducted by Group Purchasing Organizations and Health Systems.* Prepared for the Health Industry Group Purchasing Association, April 2002.

Linacre JM, Heinemann AW, Wright BD, Granger CV, Hamilton BB. The structure and stability of the Functional Independence Measure. *Arch Phys Med Rehab*. 1994;75(2):127-32.

Littenberg B, Moses LE. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. *Med Dec Making*. 1993;13(4):313-21.

Lo B, Wolf LE, Berkeley A. Conflict-of-interest policies for investigators in clinical trials. *N Engl J Med.* 2000;343(22):1616-20.

Lohr KN, ed. Institute of Medicine. *Medicare: a Strategy for Quality Assurance. Volume I.* Washington, DC. National Academy Press; 1990.

Lohr KN, Rettig RA, eds. *Quality of Care and Technology Assessment. Report of a Forum of the Council on Health Care Technology*. Washington, DC: National Academy Press; 1988.

Lyles A. Direct marketing of pharmaceuticals to consumers. *Annu Rev Public Health.* 2002;23:73-91.

MacLean CH, Morton SC, Ofman JJ, Roth EA, Shekelle PG. How useful are unpublished data from the Food and Drug Administration in meta-analysis? *J Clin Epidemiol.* 2003 Jan;56(1):44-51.

Mason JM. Cost-per QALY league tables: Their roles in pharmacoeconomic analysis. *Pharmacoeconomics.* 1994;5:472-81.

Massa T. An industry perspective: challenges in the development and regulation of drug-device combination products. In Hanna K, Manning FJ, Bouxsein P, Pope A, eds. *Innovation and Invention in Medical Devices. Workshop Summary. Institute of Medicine.* Washington, DC: National Academy Press; 2002:16-20.

Maynard A. Developing the health care market. *Econ J*. 1991;101:1277-86.

McDonald IG. Quality assurance and technology assessment: pieces of a larger puzzle. *J Qual Clin Pract.* 2000;20(2-3):87-94.

McGivney WT. Proposal for assuring technology competency and leadership in medicine. *JNCI*. 1992;84(10):742-5.

McGlynn EA, Asch SM, Adams J, Keesey J, Hicks J, DeCristofaro A, Kerr EA. The quality of health care delivered to adults in the United States. *N Engl J Med.* 2003 Jun 26;348(26):2635-45.

McGlynn EA, Kosecoff J, Brook RH. Format and conduct of consensus development conferences. Multi-nation comparison. *Int J Technol Assess Health Care*. 1990;6(3):450-69.

McHorney CA, Ware JE, Jr, Lu JF, Sherbourne CD. 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*. 1994;32(1):40-66.

McKinlay JB. From "promising report" to "standard procedure": seven stages in the career of a medical innovation. *Milbank Memorial Fund Quarterly*. 1981;59(3):374-411.

McNeil BJ. Shattuck Lecture −Hidden barriers to improvement in the quality of care. *N Engl J Med.* 2001;345:1612-20.

Medical Technology Leadership Forum. *MTLF Summit: Conditional Coverage of Investigational Technologies*. Prepared by The Lewin Group. Washington, DC; October 1999.

Medicare Payment Advisory Commission. Accounting for new technology in hospital prospective payment systems. In *Report to the Congress: Medicare Payment Policy.* Washington, DC: Medicare Payment Advisory Commission; 2001;33-45.

Meenan RF, Mason JH, Anderson JJ, Guccione AA, Kazis LE. AIMS2. The content and properties of revised and expanded arthritis impact measurement scales health status questionnaire. *Arthritis and Rheumatism*. 1992;35(1):1-10.

Mehrez A, Gafni A. Healthy-years equivalents versus quality-adjusted life years: In pursuit of progress. *Med Dec Making*. 1993;13(4):287-92.

Mello MM, Brennan TA. The controversy over high-dose chemotherapy with autologous bone marrow transplant for breast cancer. *Health Aff.* 2001;20(5):101-17.

Mendelson DN, Goodman CS, Ahn R, Rubin RJ. Outcomes and effectiveness research in the private sector and role of the Agency for Health Care Policy and Research. *Health Aff*. 1998;17(5):75-90.

Merigan TC. You can teach an old dog new tricks: how AIDS trials are pioneering new strategies. *N Engl J Med* 323. 1990(19):1341-3.

Mills RM, Jr, M. Kalan J. Developing a rational management strategy for angina pectoris after coronary bypass surgery: a clinical decision analysis. *Clin Cardiol*. 1991;14(3):191-7.

Miranda B, Matesanz R. International issues in transplantation. Setting the scene and flagging the most urgent and controversial issues. *Ann N Y Acad Sci.* 1998;862:129-43.

Mittman BS, Siu AL. Changing provider behavior: applying research on outcomes and effectiveness in health care. In *Improving Health Policy and Management: Nine Critical Research Issues for the 1990s*. Shortell SM, Reinhardt UE eds. 195-226. Ann Arbor, Mich: Health Administration Press; 1992.

Mittman BS, Tonesk X, Jacobson PD. Implementing clinical practice guidelines: social influence strategies and practitioner behaviour change. *Qual Rev Bull* 1992;18:413-21.

Moher D, Cook DJ, Eastwood S, Olkin I, et al. Improving the quality of reports of meta-analyses of randomized controlled trials: the QUOROM statement. Quality of reporting of meta-analyses. *Lancet*. 1999;354:1896-900.

Moher D, Pham B, Jones A, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in nmeta-analyses? *Lancet* 1998;352:609-13.

Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. *JAMA*. 1994;272:122-4.

Moher D, Jadad AR, Tugwell R. Assessing the quality of randomized controlled trials. Current issues and future directions. *Int J Technol Assess Health Care*. 1996;12(2):195-208.

Moher D, Schulz KF, Altman D, et al. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *JAMA*. 2001;285(15):1987-91.

Moseley JB, O'Malley K, Petersen NJ, et al. A controlled trial of arthroscopic surgery for osteoarthritis of the knee. *N Engl J Med*. 2002;347:81-8.

Mowatt G, Bower DJ, Brebner JA, Cairns JA, Grant AM, McKee L. When and how to assess fast-changing technologies: a comparative study of medical applications of four generic technologies. *Health Technology Assessment.* 1997;1(14).

Murphy MK, Black NA, Lamping DL, et al. Consensus development methods, and their use in clinical guideline development. *Health Technol Assessment*. 1998;2(3):i-iv, 1-88.

National Academy of Engineering, Committee on Public Engineering policy. *A Study of Technology Assessment.* Washington, DC: US Government Printing Office; 1969.

National Committee for Quality Assurance. *The State of Health Care Quality: 2003.* Washington, DC, 2003.

National Institute for Clinical Evidence. *Guidance for Manufacturers and Sponsors. Technology Appraisals Process Series No. 5.* London, June 2001.

National Library of Medicine, National Information Center on Health Services Research & Health Care Technology. *Etext on Health Technology Assessment (HTA) Information Resources.* http://www.nlm.nih.gov/nichsr/ehta. Accessed June 1, 2003.

National Research Council, Committee on the Life Sciences and Social Policy. *Assessing Biomedical Technologies: An Inquiry into the Nature of the Process.* Washington, DC: National Academy of Sciences; 1975.

Neuhauser D, Lewicki AM. What do we gain from the sixth stool guaiac? *N Engl J Med*. 1975;293:226-8. Copyright © 1975 Massachusetts Medical Society. All rights reserved.

Neumann PJ, Claxton K, Weinstein MC. The FDA's regulation of health economic information. *Health Aff.* 2000;19(5):129-37.

Newcomer LN. Defining experimental therapy − a third party payer's dilemma. *N Engl J Med* 323. 1990(24):1702-4.

Newhouse JP. Medical care costs: how much welfare loss? *J Econ Perspect*. 1992;6(3):3-21.

NHS Centre for Reviews and Dissemination. *Undertaking Systematic Reviews of Research on Effectiveness. CRD Guidelines for Those Carrying Out or Commissioning Reviews. CRD Report 4.* University of York, UK; 1996.

Nixon J, Stoykova B, Glanville J, Christie J, Drummond M, Kleijnen J. The U.K. NHS Economic Evaluation Database. Economic issues in evaluations of health technology. *Int J Technol Assess Health Care.* 2000;731-42.

Nord E. The QALY − a measure of social value rather than individual utility? *Health Econ*. 1994;3(2):89-93.

Normand SL. Meta-analysis: formulating, evaluating, combining, and reporting. *Stat Med*. 1999;18:321-59.

O'Brien BJ, Buxton MJ, Patterson, DL. Relationship between functional status and health-related quality-of-life after myocardial infarction. *Med Care*. 1993;31(10):950-5.

Olson CM. Consensus statements: applying structure. *JAMA.* 1995;273(1):72-3.

Olson O, Gøtzsche PC. Cochrane review on screening for breast cancer with mammography. *Lancet*. 2001;358:1340-2.

Oxman AD, Sackett DL, Guyatt GH. Users' guides to the medical literature. I. How to get started. *JAMA*. 1993;270(17):2093-5.

Passamani E. Clinical trials: are they ethical? *N Engl J Med*. 1991;324(22):1589-92.

Patrick DL, Deyo RA. Generic and disease-specific measures in assessing health status and quality of life. *Mel Care*. 1989;27(suppl. 3):S217-32.

Patrick DL, Starks HE, Cain KC, Uhlmann RF, Pearlman RA. Measuring preferences for health states worse than death. *Med Dec Making*. 1994;14(1):9-18.

Pauker SG, Kassirer, JP. Decision analysis. *N Engl J Med*. 1987;316(5):250-8.

Petitti DB. Approaches to heterogeneity in meta-analysis. *Stat Med*. 2001;20:3625-33.

Peto R, Collins R, Gray R. Large-scale randomized evidence: large, simple trials and overviews of trials. *Ann N Y Acad Sci*. 1993;703:314-40.

Phelps CE, Mooney C. Correction and update on 'priority setting in medical technology assessment.' *Medical Care*. 1992;30(8):744-51.

Phelps CE, Mushlin AI. On the (near) equivalence of cost-effectiveness and cost-benefit analyses. *Int J Technol Assess Health Care*. 1991;7:12-21.

Phillips WR. Clinical policies: making conflicts of interest explicit. Task force on clinical policies for patient care. American Academy of Family Physicians. *JAMA*. 1994;272(19):1479.

Pilnick A, Dingwall R, Starkey K. Disease management: definitions, difficulties and future directions. *Bull World Health Organ.* 2001;79(8):755-63.

Poe NR, Griffiths RI. Clinical-economic Trials. In *Tools for Evaluating Health Technologies: Five Background Papers.* US Congress, Office of Technology Assessment, BP-H-142, 125-49. Washington, DC: US Government Printing Office; 1995.

Ramsey SD, Berry K, Etzioni R, Kaplan RM, Sullivan SD, Wood DE; National Emphysema Treatment Trial Research Group. Cost effectiveness of lung-volume-reduction surgery for patients with severe emphysema. *N Engl J Med*. 2003;348(21):2092-102.

Reiser SJ. Criteria for standard versus experimental therapy. *Health Aff.* 1994;13(3):127-36.

Relman AS. Financial incentives in clinical investigation. *N Engl J Med*. 1989;320:933-34.

Rettig RA. *Health Care in Transition: Technology Assessment in the Private Sector.* Santa Monica, Ca: RAND; 1997.

Rettig RA, Levinsky NG, eds. *Kidney Failure and the Federal Government*. Washington, DC: National Academy Press; 1991.

Reuzel RP, van der Wilt GJ, ten Have HA, de Vries Robbe PF. Interactive technology assessment and wide reflective equilibrium. *J Med Philos.* 2001;26(3):245-61.

Richardson P, Hideshima T, Anderson K. Thalidomide: emerging role in cancer medicine. *Annu Rev Med*. 2002;53;629-57.

Rothenberg, Carbone DP, Johnson DH. Improving the evaluation of new cancer treatments: challenges and opportunities. *Nat Rev Cancer.* 2003;3:303-9.

Rothman KJ, Michels KB. The continuing unethical use of placebo controls. *N Engl J Med*. 1994;331:394-7.

Rossouw JE, Anderson GL, Prentice RL, et al. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results From the Women's Health Initiative randomized controlled trial. *JAMA*. 2002;288:321-33.

RTI International-University of North Carolina Evidence-based Practice Center. *Systems to Rate the Strength of Scientific Evidence.* Evidence Report/Technology Assessment Number 47. AHRQ Publication No. 02-E016, Rockville, Md: Agency for Healthcare Research and Quality, April 2002.

Sassi, F. Setting priorities for the evaluation of health interventions: when theory does not meet practice. *Health Policy*. 2003;63:141-54.

Sackett DL, Richardson WS, Rosenberg W, Haynes RB. *Evidence-Based Medicine*. New York, NY: Churchill Livingstone, 1997.

Sadock J, Tolman ED. Capital equipment budgeting: changing an art to a science. *J Healthc Resour Manag.* 1997;15(8):16-21.

Schauffler HH, Barker DC, Orleans CT. Medicaid coverage for tobacco-dependence treatments. *Health Aff*. 2001;20(1):298-303.

Schulz KF, Chalmers I, Hayes RJ, Altman DG Empirical evidence of bias: dimensions of methodological quality associated with effects in controlled trials. *JAMA*. 1995;273:408-12.

Schwappach DL. Resource allocation, social values and the QALY: a review of the debate and empirical evidence. *Health Expect*. 2002;5(3):210-22.

Sharf BF. Out of the closet and into the legislature: breast cancer stories. *Health Aff.* 2001;20(1):213-8.

Sheingold, SH. Technology assessment, coverage decisions, and conflict: the role of guidelines. *Am J Manag Care*. 1998;4:SP117-25.

Shekelle P, Eccles MP, Grimshaw JM, Woolf SH. When should clinical guidelines be updated? *BMJ*.

2001;323:155-7.

Shekelle PG, Ortiz E, Rhodes S, Morton SC, Eccles MP, Grimshaw JM, Woolf SH. Validity of the Agency for Healthcare Research and Quality clinical practice guidelines: how quickly do guidelines become outdated? *JAMA*. 2001;286:1461-7.

Singer SJ, Bergthold LA. Prospects for improved decision making about medical necessity. *Health Aff.* 2001;20(1):200-6.

Soto J. Health economic evaluations using decision analytic modeling. Principles and practices--utilization of a checklist to their development and appraisal. *Int J Technol Assess Health Care*. 2002;18:94-111.

Soumerai SB, Lipton HL. Computer-based drug-utilization review − risk, benefit, or boondoggle? *N Engl J Med*. 1995;332:1641-5.

Spilker B. *Guide to Clinical Trials.* New York, NY: Raven Press, 1991.

Spilker B. *Quality of Life and Pharmacoeconomics in Clinical Trials.* Philadelphia, Pa: Lippincott Williams & Wilkins, 1995.

Steinberg EP. Cost-effectiveness analyses. *N Engl J Med*. 1995;332:123.

Steinberg EP, Tielsch JM, Schein OD, Javitt JC, et al. The VF-14. An index of functional impairment in patients with cataract. *Arch Ophthal*. 1994;112(5):630-8.

Sterne JA, Egger M, Smith GD. Systematic reviews in health care: investigating and dealing with publication and other biases in meta-analysis. *BMJ*. 2001;323:101-5.

Stevens A, Robert G, Packer C. Early warning of new health care technologies in the United Kingdom. *Int J Technol Assess Health Care*. 1998;14:680-6.

Stone GW, Teirstein PS, Rubenstein R, et al. A prospective, multicenter, randomized trial of percutaneous transmyocardial laser revascularization in patients with nonrecanalizable chronic total occlusions. *J Am Coll Cardiol*. 2002;39:1581-7.

Stroup DF, Berlin JA, Morton SC, et al. Meta-analysis of observational studies in epidemiology. A proposal for reporting. *JAMA.* 2000;283:2008-12.

Stryer D, Bero LA. Characteristics of materials distributed by drug companies. An evaluation of appropriateness. *J Gen Intern Med.* 1996;11(10):575-83.

Suarez-Almazor ME, Belseck E, Homik J, et al. Identifying clinical trials in the medical literature with electronic databases: MEDLINE alone is not enough. *Control Clin Trials.* 2000;21:476-87.

Taylor KS. Hospital tech assessment teams field requests − and politics. *Hosp Health Netw*. 1994;68(16):58.

Taylor RS, Hutton J, Culyer AJ. Developing the revised NICE appraisal technical guidance to manufacturers and sponsors: opportunity or threat? *Pharmacoeconomics.* 2002;20:1031-8.

Thacker SB. Meta-analysis: A quantitative approach to research integration. *JAMA.* 1988; 259(11):1685-9.

Thamer M, Ray NF, Henderson SC, et al. Influence of the NIH Consensus Conference on Helicobacter pylori on physician prescribing among a Medicaid population. *Med Care*. 1998;36(5):646-60.

The Coronary Drug Project Research Group. Aspirin in coronary heart disease. *J Chron Dis*. 1976;29(10):625-42.

The Ischemic Optic Neuropathy Decompression Trial Research Group. Optic nerve decompression surgery for nonarteritic anterior ischemic optic neuropathy (NAION) is not effective and may be harmful. *JAMA.* 1995;273(8):625-32.

The Nuremberg Code, from *Trials of War Criminals Before the Nuremberg Military Tribunals under Control Council Law.* 2(10):181-82. Washington, DC: US Government Printing Office; 1949.

The Persantine-Aspirin Reinfarction Study Research Group. Persantine and aspirin in coronary heart disease. *Circulation*. 1980;62(3):449-61.

Thornbury JR, Fryback DG. Technology assessment − An American view. *Eur J Radiol*. 1992;14(2):147-56.

Thornquist MD, Urban N, Tseng A, Edelstein C, Lund B, Omenn GS. Research cost analyses to aid in decision making in the conduct of a large prevention trial, CARET. Carotene and Retinol Efficacy Trial. *Control Clin Trials*. 1993;14(4):325-39.

Thornton JG, Lilford RJ, Johnston N. Decision analysis in medicine. *BMJ*. 1992;304(6834):1099-103.

Tom E, Schulman KA. Mathematical models in decision analysis. *Infect Control Hosp Epidemiol*. 1997 Jan;18(1):65-73.

Topfer L, Parado A, Menon D, et al. Comparison of literature searches on quality and costs for health technology assessment using MEDLINE and EMBASE databases. *Int J Technol Assess Health Care.* 1999;15(2):297-303.

Torrance GW, Feeny D. Utilities and quality-adjusted life years. *Int J Technol Assess Health Care*. 1989;5(4):559-75.

Tunis SR, Stryer DB, Clancy CM. Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. *JAMA.* 2003;290(12):1624-32.

Tymstra T. The imperative character of medical technology and the meaning of 'anticipated decision regret.' *Int J Technol Assess Health Care*. 1989;5:207-13.

Ubel PA, Nord E, Gold M, et al. Improving value measurement in cost-effectiveness analysis. *Med Care.* 2000;38:892-901.

UK National Health Service R&D Health Technology Assessment Programme, 2003. http://www.hta.nhsweb.nhs.uk/abouthta.htm. Accessed June 1, 2003.

University HealthSystem Consortium. *Clinical Practice Advancement Center. Technology Report: Implementation of Hospital-Based Technology Assessment.* Oak Brook Ill.: UHC Services Corporation; 1996.

US Congress, House of Representatives. Committee on Science and Astronautics. Technology Assessment. *Statement of Emilio Q. Daddario, Chairman, Subcommittee on Science Research and Development.* 90th Cong., 1st sess., Washington, DC; 1967.

US Congress, Office of Technology Assessment. *Technology Assessment in Business and Government. Summary and Analysis.* Washington, DC: US Government Printing Office; 1977.

US Congress, Office of Technology Assessment. *Strategies for Medical Technology Assessment.* Washington, DC: US Government Printing Office; 1983.

US Congress, Office of Technology Assessment. *Protecting Privacy in Computerized Medical Information*. Washington, DC: US Government Printing Office; 1994.

US Department of Health and Human Services, Agency for Health Care Policy and Research. *Extracranial-Intracranial Bypass to Reduce the Risk of Ischemic Stroke*. Health Technology Assessment Reports. No. 6. Rockville, Md; 1990.

US Department of Health and Human Services, Agency for Health Care Policy and Research. *Intermittent Positive Pressure Breathing: Old Technologies Rarely Die*. Rockville, Md; 1993.

US Preventive Services Task Force. Screening for prostate cancer: recommendations and rationale. *Ann Intern Med*. 2002;137:915-6.

van den Broek SA, van Veldhuisen DJ, de Graeff PA, et al. Comparison between New York Heart Association classification and peak oxygen consumption in the assessment of functional status and prognosis in patients with mild to moderate chronic congestive heart failure. *Am J Cardiol*. 1992;70(3):359-63.

Van der Wilt GJ, Reuzel R, Banta HD. The ethics of assessing health technologies. *Theor Med Bioeth.* 2000;21(1):103-15.

van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Stat Med*. 2002;21:589-624.

Varmus H, Satcher D. Ethical complexities of conducting research in developing countries. *N Engl J Med*. 1997;337(14):1003-5.

Wakefield DS, Wakefield BJ. Overcoming barriers to implementation of TQM/CQI in hospitals: myths and realities. *QRB. Quality Review Bulletin*. 1993;19(3):83-8.

Ware JE Jr., Sherbourne CD. The MOS 36-item Short Form Health Survey (SF-36): I. conceptual framework and item selection. *Med Care*. 1992;30(6):473-83.

Weinstein MC, O'Brien B, Hornberger J, et al. Principles of good practice for decision analytic modeling in health-care evaluation: report of the ISPOR Task Force on Good Research Practices − Modeling Studies. *Value Health*. 2003;6(1):9-17.

Willke RJ, Glick HA, Polsky D, Schulman K. Estimating country-specific cost-effectiveness from multinational clinical trials. *Health Economics.* 1998;7:481-93.

Wood DE, DeCamp MM. The National Emphysema Treatment Trial: a paradigm for future surgical trials. *Ann Thorac Surg*. 2001;72(2):327-9.

Woolacott NF, Jones L, Forbes CA, et al. The clinical effectiveness and cost-effectiveness of bupropion and nicotine replacement therapy for smoking cessation: a systematic review and economic evaluation. *Health Technol Assess*. 2002;6(16):1-245.

World Development Report 1993: Investing in Health. The World Bank. New York, NY: Oxford University Press, 1993.

Yoshida EM. Selecting candidates for liver transplantation: a medical ethics perspective on the microallocation of a scarce and rationed resource. *Can J Gastroenterol.* 1998;12(3):209-15.

Yusuf S, Held P, Teo KK, Toretsky ER. Selection of patients for randomized controlled trials: implications of wide or narrow eligibility criteria. *Stat Med*. 1990;9(1-2):73-83.