HTA 101
INTRODUCTION TO HEALTH TECHNOLOGY ASSESSMENT

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

2014

Suggested citation: Goodman CS. HTA 101: Introduction to Health Technology Assessment. Bethesda, MD: National Library of Medicine (US); 2014.

TABLE OF CONTENTS

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 third version follows two done in 1998 and 2004 at the request of the National Information Center on Health Services Research and Health Care Technology (NICHSR) of the US National Library of Medicine (NLM).

The core material for the 1998 version was 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 US Department of Veterans Affairs Health Services Research & Development Service and its Management Decision and Research Center, and the Association for Health Services Research, since then incorporated into AcademyHealth.

HTA 101 draws from the work of the many colleagues and thousands of authors whose publications are cited in the references. In particular, I acknowledge the influence of David Banta, Robert Brook, the late Thomas Chalmers, David Eddy, the late John Eisenberg, Egon Jonsson, and the late Fred Mosteller on my understanding of the field and appreciation for the importance of involving others in it. During her long tenure at NICHSR, Ione Auston contributed to this work directly as well as indirectly through her efforts to strengthen and encourage sharing and coordination of HTA information resources in the US and globally. Additional thanks go to the hundreds of people from around the world who have attended and provided feedback on the HTA 101 short courses I have given at annual meetings of Health Technology Assessment International (HTAi) and, before those, the International Society of Technology Assessment in Health Care, since the 1990s.

As were the earlier versions 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 expert guidance and support of Ione Auston, Catherine Selden, and Patricia Gallagher, the NICHSR project officers for these efforts. Thanks go as well to Debbie Faulk for formatting this document.

Clifford S. Goodman
The Lewin Group
May 2014

ABOUT THE AUTHOR

Clifford S. Goodman, PhD, is a Senior Vice President and Principal at The Lewin Group. He has 30 years of experience in such areas as health technology assessment (HTA), evidence-based health care, comparative effectiveness research, health economics, and studies pertaining to health care innovation, regulation, payment, and access. He directs studies and projects for an international range of government agencies; pharmaceutical, biotechnology, and medical device companies; health care provider institutions; and professional, industry, and patient advocacy groups. His work on databases in HTA and health services research contributed to the development of the HealthSTAR (later incorporated into MEDLINE) and HSRProj databases of the National Library of Medicine. He has testified to the US Congress on issues pertaining to Medicare coverage of health care technology. Dr. Goodman served as Chair (2009-12) of the Medicare Evidence Development & Coverage Advisory Committee (MEDCAC) for the US Centers for Medicare and Medicaid Services. He served as President (2011-13) of the professional society, Health Technology Assessment International (HTAi), and is a Fellow of the American Institute for Medical and Biological Engineering (AIMBE).

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 HTA 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 (SBU) in Stockholm. 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.

The Lewin Group (www.lewin.com) is a national health care and human services consulting firm based in Falls Church, Virginia, near Washington, DC. It has delivered objective analyses and strategic counsel to public agencies, nonprofit organizations, industry associations and private companies across the United States for more than 40 years. The Lewin Group does not advocate for any policy, program or legislation. The Lewin Group is an Optum company. Optum is an analytics, technology and consulting services firm. Optum is a wholly-owned subsidiary of UnitedHealth Group, a diversified health and well-being company. Neither Optum nor UnitedHealth Group or its subsidiaries review the work products of The Lewin Group. The Lewin Group operates with editorial independence and provides its clients with the expert and impartial health care and human services policy research and consulting services.

I. INTRODUCTION

Technological innovation has yielded truly remarkable advances in health care during the last five decades. In recent years, breakthroughs in a variety of areas have helped to improve health care delivery and patient outcomes, including antivirals, anticlotting drugs, antidiabetic drugs, antihypertensive drugs, antirheumatic drugs, vaccines, pharmacogenomics and targeted cancer therapies, cardiac rhythm management, diagnostic imaging, minimally invasive surgery, joint replacement, pain management, infection control, and health information technology.

The proliferation of health care technology and its expanding uses have contributed to burgeoning health care costs, and the former has been cited as "culprit" for the latter. However, this relationship is variable, complex, and evolving (Cutler 2001; Cutler 2011; Goyen 2009; Medicare Payment Advisory Commission 2001; Newhouse 1992; Smith 2000). In the US, the Congressional Budget Office concluded that "roughly half of the increase in health care spending during the past several decades was associated with the expanded capabilities of medicine brought about by technological advances" (US Congressional Budget Office 2008).

Few patients or clinicians are willing to forego access to state-of-the-art health care technology. In the wealthier countries and those with growing economies, 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, provider competition, effective marketing of technologies, and consumer awareness. **Box I-1** shows some of the factors that influence demand for health technology**.**

Box I-1. Factors That Reinforce the Market for Health Technology

- Advances in science and engineering
- Intellectual property, especially patent protection
- Aging populations
- Increasing prevalence of chronic diseases
- Emerging pathogens and other disease threats
- Third-party payment, especially fee-for-service payment
- Financial incentives of technology companies, clinicians, hospitals, and others
- Public demand driven by direct-to-consumer advertising, mass media reports, social media, and consumer awareness and advocacy
- Off-label use of drugs, biologics, and devices
- "Cascade" effects of unnecessary tests, unexpected results, or patient or physician anxiety
- Clinician specialty training at academic medical centers
- Provider competition to offer state-of-the-art technology
- Malpractice avoidance
- Strong or growing economies

In this era of increasing cost pressures, restructuring of health care delivery and payment, and heightened consumer demand—yet continued inadequate access to care for many millions of people—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 useful innovation. The development, adoption, and diffusion of technology are increasingly influenced 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, to ensure its appropriate 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 HTA, in particular from the for-profit and not-for-profit private sectors as well as from government agencies, is pushing the field to evolve more systematic and transparent assessment processes and reporting to diverse users. 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.

A. 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, weather modification, 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 (Banta 2009; Brooks and Bowers 1970; Kunkle 1995; Margolis 2003).

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 understanding the future business environment, and for producing options for decision makers.

TA methodology drew upon a variety of analytical, evaluative, and planning techniques. Among these were systems analysis, cost-benefit analysis, consensus development 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). **Box I-2** shows various definitions of TA.

Box I-2. Some Definitions of Technology Assessment

[Technology assessment is] the systematic study of the effects on society, that may occur when a technology is introduced, extended, or modified, with emphasis on the impacts that are unintended, indirect, or delayed (Coates 1976).

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 ultimately comprises a systems approach to the management of technology reaching beyond technology and industrial aspects into society and environmental domains. Initially, it deals with assessment of effects, consequences, and risks of a technology, but also is a forecasting function looking into the projection of opportunities and skill development as an input into strategic planning. In this respect, it also has a component both for monitoring and scrutinizing information gathering. Ultimately, TA is a policy and consensus building process as well (UN Branch for Science and Technology for Development 1991).

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).

Technology Assessment is a concept, which embraces different forms of policy analysis on the relation between science and technology on the one hand, and policy, society and the individual on the other hand. Technology Assessment typically includes policy analysis approaches such as foresight; economic analysis; systems analysis; strategic analysis etc. … Technology Assessment has three dimensions: the cognitive dimension – creating overview on knowledge, relevant to policy-making; the normative dimension – establishing dialogue in order to support opinion making; the pragmatic dimension – establish processes that help decisions to be made. And TA has three objects: the issue or technology; the social aspects; the policy aspects (European Parliamentary Technology Assessment 2013).

B. 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 some 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 (Drug bioequivalence 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, ultrasonography for fetal sex selection, 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. Indeed, the scope of HTA reports has been diversified in recent years by the use of "horizon scanning" and the demand for "rapid HTAs," which are described later in this document.

Various definitions of HTA are shown in **Box I-3**.

Box I-3. Some Definitions of Health Technology Assessment

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).

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 (US Congress, Office of Technology Assessment 1994).

Health Technology Assessment asks important questions about these technologies [drugs, devices, procedures, settings of care, screening] such as: When is counselling better than drug treatment for depression? What is the best operation for aortic aneurysms? Should we screen for human papilloma virus when doing cervical smears? Should aspirin be used for the primary prevention of cardiovascular disease? It answers these questions by investigating four main factors: whether the technology works, for whom, at what cost, how it compares with the alternatives (UK NHS National Institute for Health Research Health Technology Assessment Programme 2013).

HTA is a field of scientific research to inform policy and clinical decision making around the introduction and diffusion of health technologies…. HTA is a multidisciplinary field that addresses the health impacts of technology, considering its specific healthcare context as well as available alternatives. Contextual factors addressed by HTA include economic, organizational, social, and ethical impacts. The scope and methods of HTA may be adapted to respond to the policy needs of a particular health system (Health Technology Assessment International 2013).

Health technology assessment (HTA) is a multidisciplinary process that summarises information about the medical, social, economic and ethical issues related to the use of a health technology in a systematic, transparent, unbiased, robust manner. Its aim is to inform the formulation of safe, effective, health policies that are patient focused and seek to achieve best value. Despite its policy goals, HTA must always be firmly rooted in research and the scientific method (European network for Health Technology Assessment 2013).

References for Chapter I

Banta D. What is technology assessment? Int J Technol Assess Health Care. 2009;25 Suppl 1:7-9.

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

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

Coates JF. 1976. Technology assessment–A tool kit. Chemtech. 1976;372-83.

Coates & Jarratt, Inc. Course Workbook: Technology Assessment. Anticipating the Consequences of Technological Choices. 1992. Washington, DC.

Cutler DM, Ly DP. The (paper) work of medicine: understanding international medical costs. J Econ Perspect. 2011;25(2):3-25.

Cutler DM, McClellan M. Is technological change in medicine worth it? Health Aff (Millwood). 2001;20(5):11-29.

Drug Bioequivalence. Recommendations from the Drug Bioequivalence Study Panel to the Office of Technology Assessment, Congress of the United States. J Pharmacokinet Biopharm. 1974(2):433-66.

European network for Health Technology Assessment. Common Questions. What is Health Technology Assessment (HTA). Accessed Aug. 1, 2013 at: http://www.eunethta.eu/about-us/faq#t287n73.

European Parliamentary Technology Assessment. What is TA? 2011. Accessed Aug. 1, 2013 at: http://eptanetwork.org/what.php.

Goyen M, Debatin JF. Healthcare costs for new technologies. Eur J Nucl Med Mol Imaging. 2009;36 Suppl 1:S139-43.

Health Technology Assessment International. What is HTA? Accessed Aug. 1, 2013 at: http://www.htai.org/index.php?id=428.

Institute of Medicine. Assessing Medical Technologies. Washington, DC: National Academy Press; 1985. http://www.nap.edu/openbook.php?record_id=607.

Kunkle G. New challenges or the past revisited? The Office of Technology Assessment in historical context. Technology in Society 1995;17(2):175-96.

Margolis RM, Guston DH. The origins, accomplishments, and demise of the Office of Technology Assessment. In Morgan MG, Peha JM, eds. Science and Technology Advice for Congress. Washington, DC: Resources for the Future; 2003; 53-76.

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. http://www.medpac.gov/publications/congressional_reports/Mar01%20Ch3.pdf.

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.

Newhouse JP. Medical care costs: how much welfare loss? J Econ Perspect. 1992;6(3):3-21. http://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.6.3.3.

Smith SD, Heffler SK, Freeland MS. The impact of technological change on health care cost spending: an evaluation of the literature. Washington, DC: Health Care Financing Administration, July 2000. http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/downloads/tech_2000_0810.pdf.

UK NHS National Institute for Health Research Health Technology Assessment Programme. About the HTA Programme. Accessed Aug. 1, 2013 at: http://www.hta.ac.uk/about/index.shtml.

UN Branch for Science and Technology for Development. United Nations Workshop on Technology Assessment for Developing Countries. Hosted by the Office of Technology Assessment. Washington, DC: 1991.

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. Protecting Privacy in Computerized Medical Information. Washington, DC: US Government Printing Office; 1993. http://govinfo.library.unt.edu/ota/Ota_1/DATA/1993/9342.PDF.

US Congress, Office of Technology Assessment. Technology Assessment in Business and Government. Summary and Analysis. Washington, DC: US Government Printing Office; 1977. http://govinfo.library.unt.edu/ota/Ota_5/DATA/1977/7711.PDF.

US Congressional Budget Office. Technological Change and the Growth of Health Care Spending. Pub. No. 2764. Washington DC: Congress of the United States; January 2008. http://www.cbo.gov/sites/default/files/cbofiles/ftpdocs/89xx/doc8947/01-31-techhealth.pdf.

II. FUNDAMENTAL CONCEPTS

A. Health Technology

Technology is the practical application of knowledge. Health technology is the practical application of knowledge to improve or maintain individual and population health. Three ways to describe health technology include its physical nature, its purpose, and its stage of diffusion.

1. Physical Nature

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

- *Drugs:* e.g., aspirin, beta-blockers, antibiotics, cancer chemotherapy
- *Biologics:* e.g., vaccines, blood products, cellular and gene therapies
- *Devices, equipment and supplies:* e.g., cardiac pacemaker, magnetic resonance imaging (MRI) scanner, surgical gloves, diagnostic test kits, mosquito netting
- *Medical and surgical procedures:* e.g., acupuncture, nutrition counseling, psychotherapy, coronary angiography, gall bladder removal, bariatric surgery, cesarean section
- *Public health programs:* e.g., water purification system, immunization program, smoking prevention program
- *Support systems:* e.g., clinical laboratory, blood bank, electronic health record system, telemedicine systems, drug formulary,
- *Organizational and managerial systems:* e.g., medication adherence program, prospective payment using diagnosis-related groups, alternative health care delivery configurations

Certainly, these categories are interdependent; for example, vaccines are biologics that are used in immunization programs, and screening tests for pathogens in donated blood are used by blood banks.

2. 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, screening 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*: intended to improve or maintain health status or avoid further deterioration (e.g., antiviral therapy, coronary artery bypass graft surgery, psychotherapy)

- *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)
- *Palliation*: improve the quality of life of patients, particularly for relief of pain, symptoms, discomfort, and stress of serious illness, as well as psychological, social, and spiritual problems. (Although often provided for progressive, incurable disease, palliation can be provided at any point in illness and with treatment, e.g., patient-controlled analgesia, medication for depression or insomnia, caregiver support.)

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 who has a positive test result for a particular disease or condition truly has that disease or condition is greatly affected by whether the test was used for screening asymptomatic patients or diagnosing symptomatic patients. See discussion of "predictive value positive," below.) Some technologies are used for diagnosis as well as treatment, e.g., coronary angiography to diagnose heart disease and to guide percutaneous coronary interventions. Implantable cardioverter defibrillators detect potentially life-threatening heart arrhythmias and deliver electrical pulses to restore normal heart rhythm. Electronic health record systems can support all of these technological purposes or applications.

Certain "hybrid" or "combination" technologies combine characteristics of drugs, devices or other major categories of technology (Goodman 1993; Lewin Group 2001; Lauritsen 2009). 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., antibiotic bone cement, drug patches, drug inhalers, implantable drug pumps, and drug-eluting coronary artery stents); spermicidal condoms; and bioartificial organs that combine natural tissues and artificial components. Examples of hybrid technologies that have complicated regulatory approval and coverage decisions are positron-emission tomography (PET, used with radiopharmaceuticals) (Coleman 1992), metered-dose inhalers (Massa 2002), and certain targeted drugs that are developed in combination with pharmacogenomic tests that are predictive of patient response to those therapies. These pharmacogenomic test-drug combinations may require clinical trials demonstrating the clinical utility of the tests as well as the safety and efficacy of the accompanying drug (US Food and Drug Administration 2007; Hudson 2011).

3. 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 clinicians 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 cancers (Rettig 2007). Many technologies undergo multiple incremental innovations after their initial acceptance into general practice (Gelijns 1994; Reiser 1994). 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 50 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 (Breitkreutz 2008; Zhou 2013).

B. Health Technology Assessment

Health technology assessment (HTA) 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.

1. Purposes of HTA

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

- Regulatory agencies about whether to permit the commercial use (e.g., marketing) of a drug, device or other regulated technology
- Payers (health care authorities, health plans, drug formularies, employers, etc.) about technology coverage (whether or not to pay), coding (assigning proper codes to enable reimbursement), 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, performance, appropriate use, and other aspects of health care technologies
- Government health department officials about undertaking public health programs (e.g., immunization, 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 technology 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
- Research agencies about evidence gaps and unmet health needs

Many of the types of organizations noted above, including government and commercial payers, hospital networks, health professional organizations, and others, have their own HTA units or functions. Many HTA agencies are affiliated with national or regional governments or consortia of multiple organizations. Further, there are independent not-for-profit and for-profit HTA organizations.

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, in the US, the Joint Commission (formerly 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. The National Quality Forum (NQF) endorses national evidence-based consensus standards for measuring and reporting across a broad range of health care interventions.

Health professional associations (e.g., American College of Cardiology, American College of Physicians, American College of Radiology) and special panels (e.g., the US Preventive Services Task Force, the joint Department of Veterans Affairs/Department of Defense Clinical Practice Guidelines program) develop **clinical practice guidelines,** standards, and other statements regarding the appropriate use of technologies (see, e.g., Institute of Medicine 2011). The Guidelines International Network (G-I-N) of organizations and individual members from more than 40 countries supports evidence-based guideline development, adaptation, dissemination, and implementation toward reducing inappropriate practice variation throughout the world. The National Guideline Clearinghouse (NGC, sponsored by the US Agency for Healthcare Research and Quality), is a searchable database of evidence-based clinical practice guidelines. Among the criteria for a new guideline to be included in NGC effective June 2014 is that it be based on a carefully documented systematic review of the evidence, including a detailed search strategy and description of study selection.

Standards-setting organizations such as the American National Standards Institute (ANSI) 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. For example, ANSI has developed standards and specifications for electronic information sharing and interoperability in such areas as laboratory results reporting, medication management, personalized health care, immunizations, and neonatal screening (Kuperman 2010).

As noted above, HTA can be used to support decision making 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 (Glasziou 2011; Straus 2011). This prompted the appearance of many useful resources, including:

- *Evidence-Based Medicine* (Sackett 1997), a guide to the field, recently updated (Straus 2011)
- *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, originally published in the *Journal of the American Medical Association,* starting in the 1990s and more recently assembled and updated (Guyatt 2008)
- Centre for Evidence-Based Medicine

2. Basic HTA Orientations

The impetus for an HTA is not necessarily a particular 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 other impacts of cochlear implants, cervical cancer screening, PET scanners, or widespread adoption of electronic health record systems.
- *Problem-oriented* assessments focus on solutions or strategies for managing a particular disease, condition, or other problem for which alternative or complementary technologies might be used. For example, clinicians and other providers concerned with the problem of diagnosis of dementia may call for HTA to inform 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 PET scanner, considering the facilities, personnel, and other resources needed to install and operate a PET scanner; the hospital's financial status; local market potential for PET services; competitive factors; etc.

These basic assessment orientations can overlap and complement one another. Certainly, all three types could draw on 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 may compare the effectiveness, safety, and other impacts of alternative technologies for a given problem, e.g., alternative treatments for atrial fibrillation (e.g., drug therapy, surgery, or catheter ablation), and may draw on technology-oriented assessments of one or more of those alternatives as well as any direct ("head-to-head") comparisons of them. 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 appropriate for that hospital. Thus, many HTAs will blend aspects of all three basic orientations.

C. Properties and Impacts Assessed

What does HTA assess? 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

The properties, impacts, and other attributes assessed in HTA pertain across the 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.

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, or in a specified treatment setting.

Efficacy and effectiveness both refer to how well a technology works, i.e., accomplishes its intended purpose, usually based on changes in one or more specified health outcomes or "endpoints" as described below. A technology that works under carefully managed conditions does not always work as well under more heterogeneous or less controlled conditions. 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 RCT, 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. Whereas efficacy answers the question, "Can it work?" (in the best conditions), effectiveness answers the question "Does it work?" (in real-world conditions).

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. As these are observational studies, their **validity** for establishing causal connections between interventions and patient outcomes is limited compared to experimental studies, particularly RCTs. Even so, observational studies can be used to generate hypotheses for experimental trials, and they can provide evidence about effectiveness that can complement other evidence about efficacy, suggesting whether findings under ideal conditions may be extended to routine practice. As discussed below, some different types of trials are designed to incorporate varied groups of patients and settings.

Box II-1 shows certain distinctions in efficacy and effectiveness for diagnostic tests. Whereas the relationship between a preventive, therapeutic, or rehabilitative technology and patient outcomes is often direct (though not always easy to measure), the relationship between a technology used for diagnosis or screening and patient outcomes is usually indirect. Also, diagnostic and screening procedures can have their own short-term and long-term adverse health effects, e.g., arising from biopsies, certain radiological procedures, or genetic testing for certain disorders.

Box II-1. 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.

Economic attributes or impacts of health technologies can be microeconomic and macroeconomic. 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 in chapter V, Economic Analysis Methods.) Health technology can have or contribute to a broad range of macroeconomic impacts. These include impacts on: a nation's gross domestic product, national health care costs, and resource allocation across health care and other industrial sectors, and international trade. Health technology can also be a factor in national and global patterns of investment, innovation, competitiveness, technology transfer, and employment (e.g., workforce size and mobility). 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 that affect technological innovation, adoption, diffusion, and use.

Ethical, legal, and social considerations arise in HTA in the form of normative concepts (e.g., valuation of human life); choices about how and when to use technologies; research and the advancement of knowledge; resource allocation; and the integrity of HTA processes themselves (Heitman 1998). Indeed, the origins of technology assessment called for the field to support policymakers' broader considerations of technological impacts, such as the "social, economic, and legal implications of any course of action" (US Congress, House of Representatives 1967) and the "short- and long-term social consequences (for example, societal, economic, ethical, legal) of the application of technology" (Banta 1993). More recently, for example, an integral component of the Human Genome Project of the US National Institutes of Health is the Ethical, Legal and Social Implications (ELSI) Research Program (Green 2011). One recently proposed broader framework, "HELPCESS," includes consideration of: humanitarian, ethical, legal, public relationships, cultural, economic, safety/security, and social implications (Yang 2013).

Whether in health care or other sectors, technological innovation can challenge certain ethical, religious, cultural, and legal norms. Current examples include genetic testing, use of stem cells to grow new tissues, allocation of scarce organs for transplantation, and life-support systems for critically ill patients. For example, the slowly increasing supply of donated kidneys, livers, hearts, lungs, and other solid organs for transplantation continues to fall behind the expanding need for them, raising ethical, social, and political concerns about allocation of scarce, life-saving resources (Huesch 2012; 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 (Moss 2011; Rettig 1991). Even so, these concerns continue to prompt innovations to

overcome organ shortages (Lechler 2005), such as techniques for improving transplantation success rates with organs from marginal donors, organs from living donors, paired and longer chain donation, xenotransplantation (e.g., from pigs), stem cells to regenerate damaged tissues, and the longer-range goal of whole-organ tissue engineering (Soto-Gutierrez 2012).

Technologies that can diminish or strengthen patient dignity or autonomy include, e.g., end-of-life care, cancer chemotherapy, feeding devices, and assistive equipment for moving immobilized patients. Greater involvement of patients, citizens, and other stakeholders in health care decisions, technology design and development, and the HTA process itself is helping to address some concerns about the relationships between patients and health technology. Ethical questions also have led to improvements in informed consent procedures for patients involved in clinical trials.

Allocation of scarce resources to technologies that are expensive, misused, not uniformly accessible, or non-curative can raise broad concerns about equity and squandered opportunities to improve population health (Gibson 2002). The same technologies can pose various challenges in the context of different or evolving societal and cultural norms, economic conditions, and health care system delivery and financing configurations. Even old or "mainstream" technologies can raise concerns in changing social contexts, such as immunization, organ procurement for transplantation, or male circumcision (EUnetHTA 2008). In addition to technologies, certain actual or proposed uses of analytical methods can prompt such concerns; many observers object to using actual or implied cost per quality-adjusted life year (QALY) thresholds in coverage decisions (Nord 2010).

Methods for assessing ethical, legal, and social implications of health technology have been underdeveloped relative to other methods in HTA, although there has been increased attention in recent years to developing frameworks and other guidance for these analyses (Duthie 2011; Potter 2008). More work is needed for translating these implications into policy (Van der Wilt 2000), such as for involving 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). Some methods used in analysis of ethical issues in HTA, based on work assembled by the European network for Health Technology Assessment (EUnetHTA), are listed in **Box II-2**. Recent examination of alternative methods used in ethical analysis in HTA suggests that they can yield similar results, and that having a systematic and transparent approach to ethical analysis is more important than the choice of methods (Saarni 2011).

Box II-2. Methods Used for Ethical Analysis in HTA

Method	Description
Casuistry	Solves morally challenging situations by comparing them with relevant and similar cases where an undisputed solution exists
Coherence analysis	Tests the consistency of ethical argumentation, values or theories on different levels, with an ideal goal of a logically coherent set of arguments
Principlism	Approaches ethical problems by addressing basic ethical principles, rooted in society's common morality
Interactive, participatory HTA approaches	Involves different stakeholders in a real discourse, to reduce bias and improve the validity and applicability of the HTA
Social shaping of technology	Addresses the interaction between society and technology and emphasizes how to shape technology in the best ways to benefit people
Wide reflective equilibrium	Aims at a coherent conclusion by a process of reflective mutual adjustment among general principles and particular judgements

Source: Saarni et al. 2008.

As a form of objective scientific and social inquiry, HTA must be subject to ethical conduct, social responsibility, and cultural differences. Some aspects to be incorporated or otherwise addressed include: identifying and minimizing potential conflicts of interest on the part of assessment staff and expert advisors; accounting for social, demographic, economic, and other dimensions of representativeness and equity in HTA resource allocation and topic selection; and patient and other stakeholder input on topic selection, evidence questions, and relevant outcomes/endpoints.

The terms "appropriate" and "necessary" often are used to describe whether or not a technology should be used in particular circumstances. These are judgments that typically reflect considerations of one or more of the properties and impacts described above. 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, perhaps as summarized in an evidence-based clinical practice guideline. A technology may be considered necessary if it is likely to be effective and acceptably safe for particular patient indications, and if withholding it would be deleterious to the patient's health (Hilborne 1991; Kahan 1994; Singer 2001).

As described in chapter I, HTA inquires about the unintended consequences of health technologies as well an intended ones, which may involve some or all of the types of impacts assessed. Some unintended consequences include, or lead to, unanticipated uses of technologies. **Box II-3** lists some recent examples.

Box II-3. Recent Examples of Unintended Consequences of Health Technology

Technology	Intended or Original Uses	Unintended Consequences or Unanticipated Uses
Antibiotics (antibacterials)	Kill or inhibit growth of bacteria that cause infectious diseases	Overuse and improper use leading to multi-drug resistant bacterial strains[1]
Antiretroviral therapy (ART)	Treatment of HIV/AIDS	Return to risky sexual behaviors in some patient groups[2,3,4]
Aspirin	Relieve pain, fever, inflammation	Antiplatelet to prevent blood clots[5]
Bariatric surgery	Weight loss in obese patients	Cure or remission of type 2 diabetes in many of the obese patients[6]
Medical ultrasonography	Visualizing structures and blood flow in the body in real time	Fetal sex selection[7,8,9]
Prostate cancer screening with PSA test	Identify men with prostate cancer early enough to cure	Invasive testing, therapies, and adverse effects for men with slow-growing/low-risk cases that will never cause symptoms[10,11]
Sildenafil	Cardiovascular disorders, especially hypertension (used today for pulmonary arterial hypertension)	Treat male sexual dysfunction[12]

Sources:

[1]Hollis A, Ahmed Z. Preserving antibiotics, rationally. N Engl J Med. 2013;369(26):2474-6.

[2]Fu TC, et al. Changes in sexual and drug-related risk behavior following antiretroviral therapy initiation among HIV-infected injection drug users. AIDS. 2012;26(18):2383-91.

[3]Kembabazi A, et al. Disinhibition in risky sexual behavior in men, but not women, during four years of antiretroviral therapy in rural, southwestern Uganda. PLoS One. 2013;8(7):e69634.

[4]Tun W, et al. Increase in sexual risk behavior associated with immunologic response to highly active antiretroviral therapy among HIV-infected injection drug users. Clin Infect Dis. 2004;38(8):1167-74.

[5]Hackam DG, Eikelboom JW. Antithrombotic treatment for peripheral arterial disease. Heart. 2007;93(3):303-8.

[6]Brethauer SA, et al. Can diabetes be surgically cured? Long-term metabolic effects of bariatric surgery in obese patients with type 2 diabetes mellitus. Ann Surg. 2013;258(4):628-36.

[7]George SM. Millions of missing girls: from fetal sexing to high technology sex selection in India. Prenat Diagn. 2006 Jul;26(7):604-9.

[8]Nie JB. Non-medical sex-selective abortion in China: ethical and public policy issues in the context of 40 million missing females. Br Med Bull. 2011;98:7-20.

[9]Thiele AT, Leier B. Towards an ethical policy for the prevention of fetal sex selection in Canada. J Obstet Gynaecol Can. 2010 Jan;32(1):54-7.

[10]Hayes JH, Barry MJ. Screening for prostate cancer with the prostate-specific antigen test: a review of current evidence. JAMA. 2014;311(11):1143-9.

[11]Lin K, Lipsitz R, Miller T, Janakiraman S; U.S. Preventive Services Task Force. Benefits and harms of prostate-specific antigen screening for prostate cancer: an evidence update for the U.S. Preventive Services Task Force. Ann Intern Med. 2008;149(3):192-9.

[12]Kling J. From hypertension to angina to Viagra. Mod Drug Discov. 1998;1(2):31-8.

1. Measuring Health Outcomes

Health outcome variables are used to measure the safety, efficacy and effectiveness of health care technologies. Main categories of health outcomes are:

- Mortality (death rate)
- Morbidity (disease rate)
- Adverse health events (e.g., harmful side effects)
- Quality of life
- Functional status
- Patient satisfaction

For example, for a cancer treatment, the main outcome of interest may be five-year survival rate; for treatments of coronary artery disease, the main endpoints may be **incidence** of fatal and nonfatal acute myocardial infarction (heart attack) and recurrence of angina pectoris (chest pain due to poor oxygen supply to the heart). Although mortality, morbidity, and adverse events are usually the outcomes of greatest interest, the other types of outcomes are often important as well to patients and others. Many technologies affect patients, family members, providers, employers, and other interested parties in other important ways; this is particularly true for many chronic diseases. As such, there is increasing emphasis on quality of life, functional status, patient satisfaction, and related types of patient outcomes.

In a clinical trial and other studies comparing alternative treatments, the effect on health outcomes of one treatment relative to another (e.g., a new treatment vs. 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 II-4** shows how choice of treatment effect measures can give different impressions of study results.

Box II-4. 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 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 $\geq$55 had a 20% reduction in mortality from breast cancer." Although this statement of *relative risk reduction* is true, it is based on the reduction from an already low-probability event in the control group to an even lower one in the screened group. Calculation of other types of treatment effect measures provides important additional information. The table below shows the number of women aged $\geq$55 and breast cancer deaths in the screened group and control group, respectively. Based on these results, 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	No. of Patients	Deaths from breast cancer	Probability of death from breast cancer	Relative risk reduction[1]	Absolute reduction[2]	Odds ratio[3]	No. needed to screen[4]
Screened	13,107	35	$P_C = 0.0027$	20.6%	0.0007	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. Relative risk reduction: $(P_C - P_S) \div P_C$

2. Absolute risk reduction: $P_C - P_S$

3. Odds ratio: $[P_S \div (1 - P_S)] \div [P_C \div (1 - P_C)]$

4. Number needed to screen to prevent one breast cancer death: $1 \div (P_C - P_S)$

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

a. Biomarkers and Surrogate Endpoints

Certain health outcomes or clinical endpoints have particular roles in clinical trials, other research, and HTA, including biomarkers, intermediate endpoints, and surrogate endpoints.

A **biomarker** (or biological marker) is an objectively measured variable or trait that is used as an indicator of a normal biological process, a disease state, or effect of a treatment (Biomarkers Definitions Working Group 2001). It may be a physiological measurement (height, weight, blood pressure, etc.), blood component or other biochemical assay (red blood cell count, viral load, glycated hemoglobin [HbA1c] level, etc.), genetic data (presence of a specific genetic mutation), or measurement from an image (coronary artery stenosis, cancer metastases).

An **intermediate endpoint** is a non-ultimate endpoint (e.g., not mortality or morbidity) that may be associated with disease status or progression toward an ultimate endpoint such as mortality or morbidity. They include certain biomarkers (e.g., HbA1c in prediabetes or diabetes, bone density in osteoporosis, tumor progression in cancer) or disease symptoms (e.g., angina frequency in heart disease, measures of lung function in chronic obstructive pulmonary disease). Some intermediate endpoints can serve as surrogate endpoints.

A **surrogate endpoint** is a measure (typically a biomarker) that is used as a substitute for a clinical endpoint of interest, such as morbidity and mortality. They are used in clinical trials when it is impractical to measure the primary endpoint during the course of the trial, such as when observation of the clinical endpoint would require years of follow-up. A surrogate endpoint is assumed, based on scientific evidence, to be a valid and reliable predictor of a clinical endpoint of interest. As such, changes in a surrogate endpoint should be highly correlated with changes in the clinical endpoint. For example, a long-standing surrogate marker for risk of stroke is hypertension, although understanding continues to evolve of the respective and joint roles of systolic and diastolic pressures in predicting stroke in the general population and in high-risk populations (Malyszko 2013). RCTs of new drugs for HIV/AIDS use biological markers such as virological (e.g., plasma HIV RNA) levels (or "loads") and immunological (e.g., CD4+ cell counts) levels (Lalezari 2003) as surrogates for mortality and morbidity. Other examples of surrogate endpoints for clinical endpoints are negative cultures for cures of bacterial infections and decrease of intraocular pressure for loss of vision in glaucoma.

b. Quality of Life Measures

Quality of life (QoL) measures, or "health-related quality of life" measures or indexes, are increasingly used along with more traditional outcome measures to assess efficacy and effectiveness, providing a more complete picture of the ways in which health care affects patients. QoL measures capture such dimensions (or domains) as: physical function, social function, cognitive function, anxiety/distress, bodily pain, sleep/rest, energy/fatigue and general health perception. These measures may be generic (covering overall health) or disease-specific. They may provide a single aggregate score or yield a set of scores, each for a particular dimension. Some examples of widely used generic measures are:

- CAHPS (formerly Consumer Assessment of Healthcare Providers and Systems)
- EuroQol (EQ-5D)
- Health Utilities Index
- Nottingham Health Profile
- Quality of Well-Being Scale
- Short Form (12) Health Survey (SF-12)
- Short Form (36) Health Survey (SF-36)
- Sickness Impact Profile

Dimensions of selected generic QoL measures that have been used extensively and that are well validated for certain applications are shown in **Box II-5**. There is an expanding literature on the relative strengths and weaknesses of these generic QoL indexes, including how sensitive they are to changes in quality of life for people with particular diseases and disorders (Coons 2000; Feeny 2011; Fryback 2007; Kaplan 2011; Kaplan 1998; Post 2001; Saban 2008).

Box II-5. Domains of Selected General Health-Related Quality of Life Indexes

EuroQol EQ-5D (Rabin 2001)

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

Functional Independence Measure (Hsueh 2002; Linacre 1994)

• Self-care	• Communication
• Sphincter control	• Psychosocial
• Mobility	• Cognition

Nottingham Health Profile (Doll 1993; Jenkinson 1988)

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

Quality of Well-Being Scale (Frosch 2004; Kaplan 1989)

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

Short Form (SF)-36 (Martin 2011; Ware 1992)

• Physical functioning	• Mental health
• Role - physical	• Role - emotional
• Social functioning	• Vitality
• Bodily pain	• General health perceptions

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

Some of the diseases or conditions for which there are disease- (or condition-) specific measures are: angina, arthritis, asthma, epilepsy, heart disease, kidney disease, migraine, multiple sclerosis, urinary incontinence, and vision problems. See **Box II-6** for dimensions used in selected measures.

Box II-6. Domains of Selected Disease-Specific Health-Related Quality of Life Indexes

Adult Asthma Quality of Life Questionnaire (Juniper 2005; Juniper 1993)

• Activity limitations	• Exposure to environmental stimuli
• Emotional function	• Symptoms

Arthritis Impact Measurement Scales (AIMS2) (Söderlin 2004; Meenan 1992)

• Mobility	• Social activity
• 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

Urinary Incontinence-Specific Quality of Life Instrument (I-QOL) (Patrick 1999; Wagner 1996)

• Avoidance and limiting behavior	• Social embarrassment
• Psychosocial impacts	

Considerable advances have been made in the development and validation of generic and disease-specific measures since the 1980s. These 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, depression, arthritis) but may have different side effect profiles that affect patients' quality of life (Gregorian 2003).

c. Health-Adjusted Life Years: QALYs, DALYs, and More

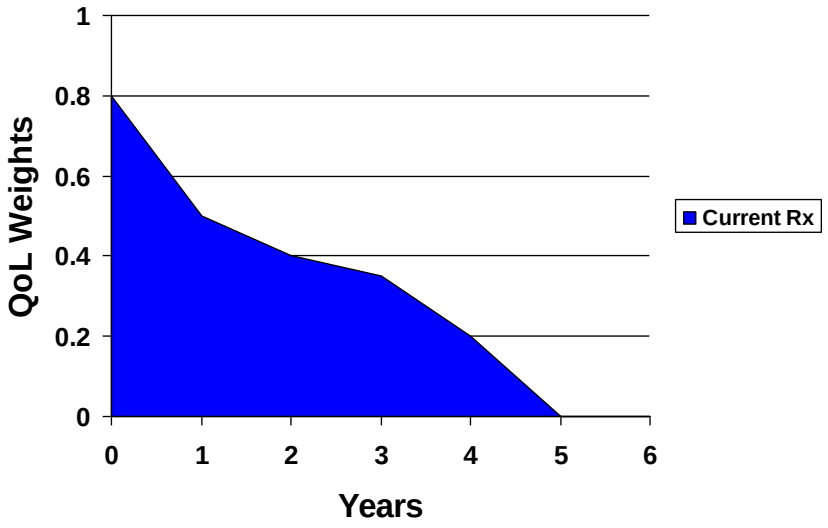
The category of measures known as health-adjusted life years (HALYs) recognizes that changes in an individual's health status or the burden of population health should reflect not only the dimension of life expectancy but a dimension of QoL or functional status. Three main types of HALYs are: **quality-adjusted life years (QALYs), disability-adjusted life years (DALYs),** and **healthy-years equivalents (HYEs).** One of the attributes of HALYs is that they are not specific to a particular disease or condition.

The QALY is a unit of health care outcome that combines gains (or losses) in length of life with quality of life. QALYs are usually used to 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 1989). QALYs provide a common unit for multiple purposes, including: estimating the overall burden of disease; comparing the relative impact on personal and population health of specific diseases or conditions, comparing the relative impact on personal and population health of specific technologies; and making economic comparisons, such as of the cost-effectiveness (in particular the cost-utility) of different health care interventions. Some health economists and policymakers 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 (Gold 2002; Johannesson 1993; Mullahy 2001). QALYs are used routinely in assessing the impact or value of technologies by some HTA organizations, e.g., the National Institute for Health and Care Excellence (NICE) in the UK. **Box II-7** illustrates the dual dimensions of QALYs, and how an intervention can result in a gain in QALYs.

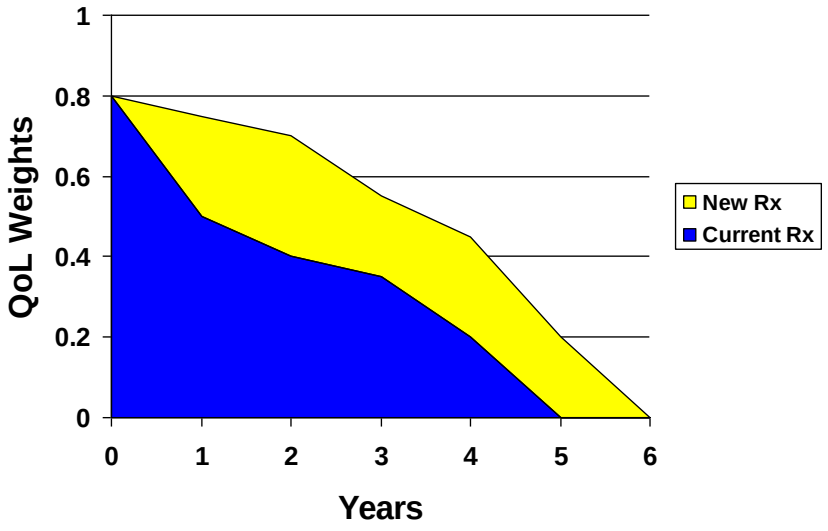
Box II-7. Gain in Quality-Adjusted Life Years from a New Intervention

QALY = Length of life X Quality Weight

Survival and Quality of Life with Current Treatment



Survival and Quality of Life with New Treatment
QALY Gain is Represented by the Area of Increased Survival and Quality of Life



Although HALYs arise from a common concept of adjusting duration of life by individuals' experience of quality of life, they differ in ways that have implications for their appropriate use, including for assessing cost-effectiveness. QALYs are used primarily to adjust a person's life expectancy by the levels of health-related quality of life that the person is predicted to experience during the remainder of life or some interval of it. DALYs are primarily used to measure population disease burden; they are a measure of something 'lost' rather than something 'gained.' The health-related quality of life weights used for QALYs are intended to represent quality of life levels experienced by individuals in particular health

states, whereas the disability weights used for DALYs represent levels of loss of functioning caused by mental or physical disability caused by disease or injury. Another key distinction is that the burden of disability in calculating DALYs depends on one's age. That is, DALYs incorporate an age-weighting function that assigns different weights to life years lived at different ages. Also, the origins of quality of life weights and disability weights are different (Sassi 2006; Fox-Rushby 2001).

The scale of quality of life used for QALYs can be based on general, multi-attribute QoL indexes or preference survey methods (Bleichrodt 1997; Doctor 2010; Weinstein 2010). The multi-attribute QoL indexes used for this purpose include, e.g., the SF-6D (based on the SF-36), EQ-5D, versions of the Health Utilities Index, and Quality of Well-Being Scale. The preference survey methods are used to elicit the **utility** or preferences of individuals (including patients, disabled persons, or others) for certain states of health or well-being, such as the standard gamble, time-tradeoff, or rating scale methods (e.g., a visual analog scale). Another preference survey method, the person trade-off, is used for eliciting preferences for the health states of a community or population, although the standard gamble, time tradeoff, and rating scales can be used at that level as well. This scale is typically standardized to a range of 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). Some work has been done to capture more dimensions of public preference and to better account for the value attributed to different health care interventions (Dolan 2001; Schwappach 2002). There is general agreement about the usefulness of the standard measures of health outcomes such as QALYs to enable comparisons of the impacts of technologies across diseases and populations, and standard approaches for valuing utilities for different health states. Among the areas of controversy are:

- whether the QALY captures the full range of health benefits,
- that the QALY does not account for social concerns for equity
- whether the QALY is the most appropriate generic preference-based measure of utility
- whether a QALY is the same regardless of who experiences it
- what the appropriate perspective is for valuing health states, e.g., from the perspective of patients with particular diseases or the general public (Whitehead 2010).

Regarding perspective, for example, the values of the general public may not account for adaptation of the patients to changes in health states, and patients' values may incorporate self-interest. Given this divergence, the appropriate perspective for health state valuations should depend on the context of the decisions or policies to be informed by the evaluation (Stamuli 2011; Oldridge 2008).

QoL measures and QALYs continue to be used in HTA while substantial work continues in reviewing, refining and validating them. As described in chapter V, Economic Analysis Methods, the QALY is often used as the unit of patient outcomes in cost-utility analyses.

2. Performance of Screening and Diagnostic Technologies

Screening and diagnostic tests provide information about the presence of a disease or other health condition. As such, they must be able to discriminate between patients who have a particular disease or condition and those who do not have it. Although the tests used for them are often the same, screening and diagnosis are distinct applications: screening is conducted in asymptomatic patients; diagnosis is conducted in symptomatic patients. As described below, whether a particular test is used for screening or it is used for diagnosis can have a great effect on the probability that the test result truly indicates

whether or not a patient has a given disease or other health condition. Although these tests are most often recognized as being used for screening and diagnosis, there are other, related uses of these tests across the spectrum of managing a disease or condition, as listed in **Box II-8**.

Box II-8. Uses of Tests for Asymptomatic and Symptomatic Patients

Asymptomatic patients (no known disease)

- Susceptibility: presence of a risk factor for a disease (e.g., a gene for a particular form of cancer)
- Presence of (hidden or occult) disease (e.g., Pap smear for cervical cancer)

Symptomatic patients (known or probable disease)

- Diagnosis: presence of a particular disease or condition (e.g., thyroid tests for suspected hyperthyroidism)
- Differential diagnosis: determine which disease or condition a patient has from among multiple possible alternatives (e.g., in a process of elimination using a series of tests to rule out particular diseases or conditions)
- Staging: extent or progression of a disease (e.g., imaging to determine stages of cancer)
- Prognosis: probability of progression of a disease or condition to a particular health outcome (e.g., a multi-gene test for survival of a particular type of cancer)
- Prediction: probability of a treatment to result in progression of a disease or condition to a particular health outcome (e.g., a genetic test for the responsiveness of colorectal cancer to a particular chemotherapy)
- Surveillance: periodic testing for recurrence or other change in disease or condition status
- Monitoring: response to treatment (e.g., response to anticoagulation therapy)

The technical performance of a test depends on multiple 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 designated cutoff level (threshold) of the variable (usually a biomarker) used to determine the presence or absence of that disease. These factors contribute to the ability of a test to detect a disease when it is present and to not detect a disease when it is not present.

A screening or diagnostic test can have four basic types of outcomes, as shown in **Box II-9.** A **true positive** 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 II-9. Possible Outcomes of a Screening or Diagnostic Test

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

Operating characteristics of tests and procedures are measures of their technical performance. These characteristics are based on the probabilities of the four possible types of outcomes of a test noted above. The two most commonly used operating characteristics of screening and diagnostic tests are sensitivity and specificity. **Sensitivity** measures the ability of a test to detect a particular disease

(e.g., a particular type of infection) or condition (a particular genotype) when it is present. **Specificity** measures the ability of a test to correctly exclude that disease or condition in a person who truly does not have that disease or condition. The sensitivity and specificity of a test are independent of the true prevalence of the disease or condition in the population being tested.

A 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) for all cutoff points of a disease or condition marker. For a perfect test, the area under the ROC curve would be 1.0; for a useless test (no better than a coin flip), the area under the ROC curve would be 0.5. ROC curves help to demonstrate how raising or lowering a cutoff point selected 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 disease (false positives).

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, shown in **Box II-10**. **Positive predictive value** is the proportion of those patients with a positive test result who actually have the disease. **Negative predictive value** is the proportion of patients with a negative test result who actually do not have the disease. Unlike sensitivity and specificity, the positive and negative predictive values of a test do depend on the true prevalence of the disease or condition in the population being tested. That is, the positive and negative predictive values of a test result are not constant performance characteristics of a test; they vary with the prevalence of the disease or condition 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.

Box II-10. 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
Positive predictive value	$\frac{\text{True Positives}}{\text{True positives + False positives}}$	Proportion of people with positive test who have condition
Negative predictive value	$\frac{\text{True Negatives}}{\text{True negatives + False negatives}}$	Proportion of people with negative test who do not have condition

a. Biomarkers and Cutoff Points in Disease Detection

The biomarker for certain diseases or conditions is typically defined as a certain cutoff level of one or more variables. Examples of variables used for biomarkers for particular diseases are systolic and diastolic blood pressure for hypertension, HbA1c level for type 2 diabetes, coronary calcium score for coronary artery disease, and high-sensitivity cardiac troponin T for acute myocardial infarction. The usefulness of such biomarkers in making a definitive finding about presence or absence of a disease or

condition varies; many are used in conjunction with information from other tests or patient risk factors. Biomarkers used to detect diseases 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 example, an HbA1c level of 6.5% may be designated as the cutoff point for diagnosing type 2 diabetes. In fact, some people whose HbA1c level is lower than 6.5% also have diabetes (as confirmed by other tests), and some people whose HbA1c level is higher than 6.5% do not have diabetes. Lowering the cutoff point to 6.0% or 5.5% will correctly identify more people who are diabetic, but it will also incorrectly identify more people as being diabetic who are not. For diabetes as well as other conditions, clinically useful cutoff points may vary among different population subgroups (e.g., by age or race/ethnicity).

A cutoff point that is set to detect more true positives will also yield more false positives; a cutoff point that is set to detect more true negatives will also yield more false negatives. There are various statistical approaches for determining "optimal" cutoff points, e.g., where the intent is to minimize total false positives and false negatives, with equal weight given to sensitivity and specificity (Perkins 2006). However, the selection of a cutoff point should consider the acceptable risks of false positives vs. false negatives. For example, if the penalty for a false negative test is high (e.g., in patients with a fatal disease for which there is an effective treatment), then the cutoff point is usually set to be highly sensitive to minimize false negatives. If the penalty for a false positive test is high (e.g., leading to confirmatory tests or treatments that are invasive, associated with adverse events, and expensive), then the cutoff point is usually set to be highly specific to minimize false positives. Given the different purposes of screening and diagnosis, and the associated penalties of false positives and false negatives, cutoff points may be set differently for screening and diagnosis of the same disease.

b. Tests and Health Outcomes

Beyond technical performance of screening and diagnostic tests, their effect on health outcomes or health-related quality of life is often less immediate or direct than for other types of technologies. The impacts of most preventive, therapeutic, and rehabilitative technologies on health outcomes can be assessed as direct cause-and-effect relationships between interventions and outcomes. However, the relationship between the use of screening and diagnostic tests and health outcomes is typically indirect, given intervening decisions or other steps between the test and health outcomes. Even highly accurate test results may be ignored or improperly interpreted by clinicians. Therapeutic decisions that are based on test results can have differential effects on patient outcomes. Also, the impact of those therapeutic decisions may be subject to other factors, such as patient adherence to a drug regimen. Even so, health care decision makers and policymakers increasingly seek direct or indirect evidence demonstrating that a test is likely to have an impact on clinical decisions and health care outcomes.

The effectiveness (or efficacy) of a diagnostic (or screening) 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 (where relevant to decision makers), 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?
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 effective at any step along this chain, then it is not likely to be effective at any subsequent step. Effectiveness at a given step does not imply effectiveness at a later step (Feeny 1986; Fineberg 1977; Institute of Medicine 1985). An often-cited hierarchy of studies for assessing diagnostic imaging technologies that is consistent with the chain of inquiry noted above is shown in **Box II-11**. A generic analytical framework of the types of evidence questions that could be asked about the impacts of a screening test is presented in **Box II-12**. Some groups have developed standards for assessing the quality of studies of the accuracy of screening and diagnostic tests, such as for conducting systematic reviews of the literature on those tests (Smidt 2006; Whiting 2011).

Box II-11. 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

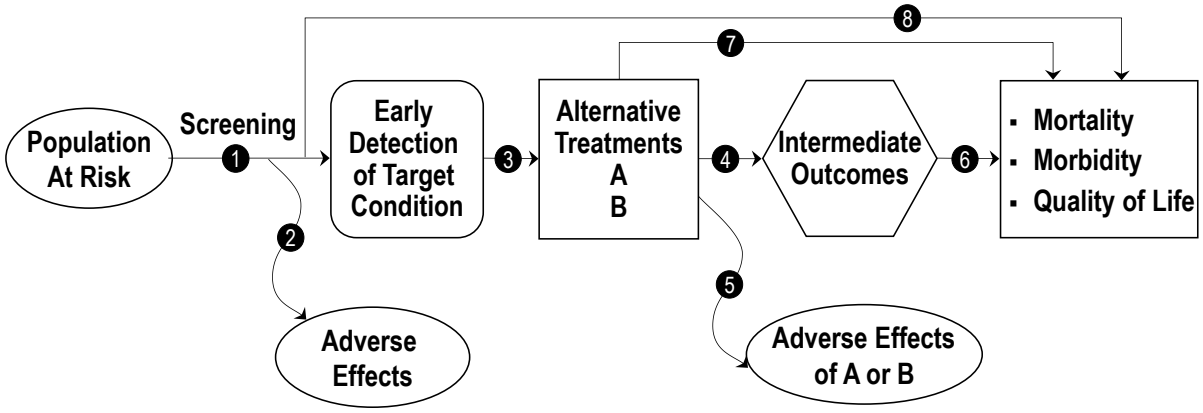
- % of patients improved with test compared with/without test
- Morbidity (or procedures) avoided after having image information
- Change in quality-adjusted life expectancy
- Expected value of test information in quality-adjusted life years (QALYs)
- Cost per QALY saved with imaging information
- Patient 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 JR, Fryback DG. Technology assessment − An American view. Eur J Radiol. 1992;14(2):147-56.

Box II-12. Example of Analytical Framework of Evidence Questions: Screening



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

Source: Adapted from: 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.

For diagnostic (or screening) technologies that are still prototypes or in other early stages of development, there may be limited data on which to base answers to such questions as these. Even so, investigators and advocates of diagnostic technologies should be prepared to describe, at least qualitatively, how the technology might affect diagnostic accuracy, diagnostic impact, therapeutic impact, patient outcomes and cost effectiveness (where appropriate); 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.

3. 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. Regulators, payers, clinicians, hospital managers, investors, 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 recent 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 II-13** shows examples of health care technologies found to be ineffective or harmful after being widely diffused.

Box II-13. Technologies Found to be Ineffective or Harmful for Some or All Indications After Diffusion

- Autologous bone marrow transplantation with high-dose chemotherapy for advanced breast cancer
- Antiarrhythmic drugs
- Bevacizumab for metastatic 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 preventing heart disease in healthy menopausal women
- Hydralazine for chronic heart failure
- Intermittent positive pressure breathing
- Mammary artery ligation for coronary artery disease
- Magnetic resonance imaging (routine) for low back pain in first 6 weeks
- Optic nerve decompression surgery for nonarteritic anterior ischemic optic neuropathy
- Oxygen supplementation for premature infants
- Prefrontal lobotomy for mental disturbances
- Prostate-specific antigen (PSA) screening for prostate cancer
- Quinidine for suppressing recurrences of atrial fibrillation
- Radiation therapy for acne
- Rofecoxib (COX-2 inhibitor) for anti-inflammation
- 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: Chou 2011; Coplen 1990; Enkin 2000; Feeny 1986; FDA Center for Drug Evaluation and Research 2010; Fletcher 2002; Grimes 1993; Mello 2001; The Ischemic Optic Neuropathy Decompression Trial Research Group 1995; Jüni 2004; Passamani 1991; Peters 2005; Rossouw 2002; Srinivas 2012; Toh 2010; US DHHS 1990, 1993; others.

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. By the time a HTA is conducted, reviewed, and disseminated, its findings may be outdated by changes in a technology, how it is used, its competing technologies (comparators) for a given health problem (indication), the health problems for which it is used, and other factors (Goodman 1996). See chapter VI, Determine Topics for

HTA, for further discussion of identification of candidate assessment topics, **horizon scanning**, setting assessment priorities, reassessment, and the moving target problem.

In recent years, the demand for HTA by health care decision makers has increasingly involved requests for faster responses to help inform emergent decisions. This has led to development of "rapid HTAs" that are more focused, less-comprehensive assessments designed to provide high-level responses to such decision maker requests within approximately four-to-eight weeks. See discussion of **rapid HTA** in chapter X, Selected Issues in HTA.

Among the factors affecting the timing of HTA is the sufficiency of evidence to undertake an HTA. One of the types of circumstances in which there are tradeoffs in "when to assess" is a coverage decision for a new technology (or new application of an existing technology) for which there is promising, yet non-definitive or otherwise limited, evidence. For some of these technologies, delaying any reimbursement until sufficient evidence is available for a definitive coverage decision could deny access for certain patients with unmet medical need who might benefit. Further, the absence of any reimbursement could slow the generation of evidence. In such instances, payers may provide for **coverage with evidence development** or other forms of **managed entry** of the technology in which reimbursement is made for particular indications or other well-defined uses of the technology in exchange for collection of additional evidence. See further discussion of managed entry in chapter X.

D. 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 include a selection of the following:

- Physicians, nurses, other clinicians
- Managers of hospitals, clinics, nursing homes, and other health care institutions
- Pharmacists and pharmacologists
- Laboratory technicians, radiology technicians, and other allied health professionals
- Biomedical and clinical engineers
- Patients and community representatives
- Epidemiologists
- Biostatisticians
- Economists
- Social scientists
- Decision scientists
- Ethicists
- Lawyers
- Computer scientists/programmers
- Librarians/information specialists

Of course, certain individuals have multiple types of expertise. The set of participants in an HTA depends on the scope and depth of the topic, 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). Certain clinical specialists, and marketing, legal, and analytical staff and patient or community representatives could be involved as appropriate.

E. Basic HTA Frameworks

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 or questions
3. Determine organizational locus or responsibility for assessment
4. Retrieve available relevant evidence
5. Generate or collect new evidence (as appropriate)
6. Appraise/interpret quality of the 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 (using systematic reviews and meta-analyses) based on existing relevant primary data studies (reported in journal articles or from epidemiological or administrative data sets). Some assessment efforts involve multiple cycles of retrieving/collecting, interpreting, and integrating evidence before completing an assessment. The steps of appraising and integrating evidence may be done iteratively, such as when individual primary data studies pertaining to a particular evidence question are appraised individually for quality and then are integrated into a body of evidence, which in turn is appraised for its overall quality, as described in chapter III and chapter IV. Depending on 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. As indicated by various chapter and section headings, all ten of the basic steps of HTA listed above are described in this document.

EUnetHTA has developed a "core model" for HTA to serve as a generic framework to enable international collaboration for producing and sharing the results of HTAs (EUnetHTA 2013). Core HTAs are intended to serve as a basis for local (i.e., a particular nation, region, or program) reports, and as such do not contain recommendations on technology use. The core model involves the following domains and production phases (EUnetHTA 2008; Lampe 2009):

EUnetHTA Core Model Domains

1. Health problem and current use of technology
2. Description and technical characteristics of technology
3. Safety
4. Clinical effectiveness
5. Costs and economic evaluation
6. Ethical analysis
7. Organizational aspects
8. Social aspects
9. Legal aspects

EUnetHTA Core Model Phases

1. Definition of the technology to be assessed
2. Definition of project type
3. Relevance of assessment elements
4. Translation of relevant issues into research questions
5. Compiling of a core HTA protocol
6. Research
7. Entering the results

HTA embraces a diverse group of methods. Two of the main types of HTA methods are primary data collection methods and secondary or integrative methods. **Primary data methods** (described in chapter III) involve collection of original data, such as clinical trials and observational studies. **Integrative methods**, or secondary or synthesis methods (chapter IV), involve combining data or information from existing sources, including from primary data studies. **Economic analysis methods** (chapter V) can involve one or both of primary data methods and integrative methods.

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. Given their varying assessment orientations, resource constraints and other factors, HTA programs tend to rely on different combinations of methods. Even so, the general trend in HTA is to call for and emphasize evidence based on the more rigorous and systematic methods.

References for Chapter II

Andersson I, Aspegren K, Janzon L, et al. Mammographic screening and mortality from breast cancer: The Malmö Mammographic Screening Trial. BMJ. 1988; 297(6654):943-8. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1834636.

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

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.

Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther. 2001;69(3):89-95.

Bleichrodt H, Johannesson M. Standard gamble, time trade-off and rating scale: experimental results on the ranking properties of QALYs. J Health Econ. 1997;16(2):155-75.

Breitkreutz I, Anderson KC. Thalidomide in multiple myeloma--clinical trials and aspects of drug metabolism and toxicity. Expert Opin Drug Metab Toxicol. 2008;4(7):973-85.

Brethauer SA, Aminian A, Romero-Talamás H, Batayyah E, et al. Can diabetes be surgically cured? Long-term metabolic effects of bariatric surgery in obese patients with type 2 diabetes mellitus. Ann Surg. 2013;258(4):628-36.

Buxton MJ. Problems in the economic appraisal of new health technology: the evaluation of heart transplants in the UK. In Drummond MF, ed. Economic Appraisal of Health Technology in the European Community. Oxford, England. Oxford Medical Publications, 1987.

Chou R, Croswell JM, Dana T, et al. Screening for Prostate Cancer: A Review of the Evidence for the U.S. Preventive Services Task Force. Ann Intern Med. 2011 Dec 6;155(11):762-71

Coleman RE, Briner WH, Siegel BA. Clinical PET scanning. A short-lived orphan. Int J Technol Assess Health Care. 1992;8(4):610-22.

Coons SJ, Rao S, Keininger DL, Hays RD. A comparative review of generic quality-of-life instruments. Pharmacoeconomics 2000;17(1):13-35.

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(4):1106-16.

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.

Doctor JN, Bleichrodt H, Lin HJ. Health utility bias: a systematic review and meta-analytic evaluation. Med Decis Making. 2010;30(1):58-67.

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.

Duthie K, Bond K. Improving ethics analysis in health technology assessment. Int J Technol Assess Health Care 2011;27(1):64-70.

Enkin M, Nelison J, Crowther C, Duley L, et al. A Guide to Effective Care in Pregnancy and Childbirth. 3rd ed. New York, NY: Oxford University Press; 2000.

EUnetHTA (European Network for Health Technology Assessment). HTA Core Model for Medical and Surgical Interventions Version 1.0 Work Package 4. The HTA Core Model. December 2008.

EUnetHTA. HTA Core Model® Online. National Institute for Health and Welfare. Helsinki; 2013. Accessed Dec. 30, 2013 at: https://meka.thl.fi/htacore/Default.aspx.

FDA Center for Drug Evaluation and Research. Memorandum to the File: BLA 125085 Avastin (bevacizumab). Regulatory Decision to Withdraw Avastin (bevacizumab) Firstline Metastatic Breast Cancer Indication December 15, 2010. Accessed Sept. 1, 2011 at: http://www.fda.gov/downloads/Drugs/DrugSafety/PostmarketDrugSafetyInformationforPatientsandProviders/UCM237171.pdf.

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

Feeny D, Spritzer K, Hays RD, Liu H, et al. Agreement about identifying patients who change over time: cautionary results in cataract and heart failure patients. Med Decis Making. 2012;32(2):273-86. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3749910.

Fineberg HV, Bauman R, Sosman M. Computerized cranial tomography: effect on diagnostic and therapeutic plans. JAMA. 1977;238(3):224-7.

Fletcher SW, Colditz GA. Failure of estrogen plus progestin therapy for prevention. JAMA. 2002;288(3):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.

Fox-Rushby JA, Hanson K. Calculating and presenting disability adjusted life years (DALYs) in cost-effectiveness analysis. Health Policy Plan. 2001;16(3):326-31. http://heapol.oxfordjournals.org/content/16/3/326.long.

Frosch DL, Kaplan RM, Ganiats TG, Groessl EJ, Sieber WJ, Weisman MH. Validity of self-administered quality of well-being scale in musculoskeletal disease. Arthritis Rheum. 2004;51(1):28-33. http://onlinelibrary.wiley.com/doi/10.1002/art.20071/abstract;jsessionid=03A0085CEC3880A9773571AA68E4FFBE.d04t02!.

Fryback DG, Dunham NC, Palta M, Hanmer J, et al. US norms for six generic health-related quality-of-life indexes from the National Health Measurement study. Med Care. 2007;45(12):1162-70. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2647803.

Fu TC, Westergaard RP, Lau B, Celentano DD, Vlahov D, Mehta SH, Kirk GD. Changes in sexual and drug-related risk behavior following antiretroviral therapy initiation among HIV-infected injection drug users. AIDS. 2012;26(18):2383-91. Accessed Dec. 1, 2014 at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3678983.

Gelijns A, Rosenberg N. The dynamics of technological change in medicine. Health Aff (Millwood). 1994;13(3):28-46. http://content.healthaffairs.org/content/13/3/28.long.

George SM. Millions of missing girls: from fetal sexing to high technology sex selection in India. Prenat Diagn. 2006 Jul;26(7):604-9.

Gibson JL, Martin DK, Singer PA. Priority setting for new technologies in medicine: a transdisciplinary study. BMC Health Serv Res. 2002;2(1):14. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC119858.

Glasziou P. What is EBM? Evid Based Med. 2011;16(5):129-30.

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.

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 CS. Technology assessment in healthcare: a means for pursuing the goals of biomedical engineering. Med Biolog Eng Comput. 1993;31(1):HTA3-10.

Green ED, Guyer MS; National Human Genome Research Institute. Charting a course for genomic medicine from base pairs to bedside. Nature. 2011;470(7333):204-13.

Gregorian RS, Golden KA, Bahce A, Goodman C, et al. Antidepressant-induced sexual dysfunction. Ann Pharmacother. 2002;36(10):1577-89.

Grimes DA. Technology follies: the uncritical acceptance of medical innovation. JAMA. 1993;269(23):3030-3.

Guyatt G, Rennie D, Meade MO, Cook DJ. Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice, 2nd Edition. New York: McGraw-Hill, 2008.

Hackam DG, Eikelboom JW. Antithrombotic treatment for peripheral arterial disease. Heart. 2007;93(3):303-8. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1861457.

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.

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

Hilborne LH, Leape LL, Kahan JP, Park RE, et al. Percutaneous Transluminal Coronary Angioplasty: A Literature Review of Ratings of Appropriateness and Necessity. Santa Monica, Ca: RAND; 1991. http://www.rand.org/pubs/joint_reports-health/JRA01.html.

Hollis A, Ahmed Z. Preserving antibiotics, rationally. N Engl J Med. 2013;369(26):2474-6. http://www.nejm.org/doi/full/10.1056/NEJMp1311479.

Hsueh IP, Lin JH, Jeng JS, Hsieh CL. Comparison of the psychometric characteristics of the functional independence measure, 5 item Barthel index, and 10 item Barthel index in patients with stroke. J Neurol Neurosurg Psychiatry. 2002;73(2):188-90.

Hudson KL. Genomics, health care, and society. N Engl J Med. 2011;365(11):1033-41.

Huesch MD. One and done? Equality of opportunity and repeated access to scarce, indivisible medical resources. BMC Med Ethics. 2012;13:11.

Institute of Medicine. Assessing Medical Technologies. Washington, DC: National Academy Press; 1985. http://www.nap.edu/openbook.php?record_id=607.

Institute of Medicine. Clinical Practice Guidelines We Can Trust. Washington, DC: National Academies Press; 2011.

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.

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.

Jüni P, Nartey L, Reichenbach S, et al. Risk of cardiovascular events and rofecoxib: cumulative meta-analysis. Lancet. 2004;364(9450):2021-9.

Juniper EF, Guyatt GH, Ferrie PJ, Griffith LE. Measuring quality of life in asthma. Am Rev Respir Dis. 1993;147(4):832-8.

Juniper EF, Svensson K, Mörk AC, Ståhl E. Modification of the asthma quality of life questionnaire (standardised) for patients 12 years and older. Health Qual Life Outcomes. 2005;3:58. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1262746.

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.

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(3 Suppl): S27-43.

Kaplan RM, Ganiats TG, Sieber WJ, Anderson JP. The Quality of Well-Being Scale: critical similarities and differences with SF-36. Int J Qual Health Care. 1998;10(6):509-20. http://intqhc.oxfordjournals.org/content/10/6/509.long.

Kaplan RM, Tally S, Hays RD, Feeny D, Ganiats TG, et al. Five preference-based indexes in cataract and heart failure patients were not equally responsive to change. J Clin Epidemiol 2011;64(5):497-506.

Kembabazi A, Bajunirwe F, Hunt PW, Martin JN, et al. Disinhibition in risky sexual behavior in men, but not women, during four years of antiretroviral therapy in rural, southwestern Uganda. PLoS One. 2013;8(7):e69634. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3716596.

Kling J. From hypertension to angina to Viagra. Mod. Drug Discov. 1998;1(2):31-8.

Kuperman GJ, Blair JS, Franck RA, Devaraj S, Low AF; NHIN Trial Implementations Core Services Content Working Group. Developing data content specifications for the nationwide health information network trial implementations. J Am Med Inform Assoc. 2010;17(1):6-12. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2995626.

The Lewin Group. Outlook for Medical Technology Innovation. Report 4: The Impact of Regulation and Market Dynamics on Innovation. Washington, DC: AdvaMed; 2001.

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(22):2175-85. http://www.nejm.org/doi/full/10.1056/NEJMoa035026.

Lampe K, Mäkelä M, Garrido MV, et al.; European network for Health Technology Assessment (EUnetHTA). The HTA core model: a novel method for producing and reporting health technology assessments. Int J Technol Assess Health Care. 2009;25 Suppl 2:9-20.

Lauritsen KJ, Nguyen T. Combination products regulation at the FDA. Clin Pharmacol Ther. 2009;85(5):468-70.

Lechler RI, Sykes M, Thomson AW, Turka LA. Organ transplantation--how much of the promise has been realized? Nat Med. 2005;11(6):605-13.

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.

Malyszko J, Muntner P, Rysz J, Banach M. Blood pressure levels and stroke: J-curve phenomenon? Curr Hypertens Rep. 2013;15(6):575-81.

Martin ML, Patrick DL, Gandra SR, Bennett AV, et al. Content validation of two SF-36 subscales for use in type 2 diabetes and non-dialysis chronic kidney disease-related anemia. Qual Life Res 2011;20(6):889-901.

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; 2001:16-20. http://www.nap.edu/openbook.php?record_id=10225&page=1.

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.

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 Rheum. 1992;35(1):1-10.

Mello MM, Brennan TA. The controversy over high-dose chemotherapy with autologous bone marrow transplant for breast cancer. Health Aff (Millwood). 2001;20(5):101-17. http://content.healthaffairs.org/content/20/5/101.long.

Moss AH. Ethical principles and processes guiding dialysis decision-making. Clin J Am Soc Nephrol. 2011;6(9):2313-7. http://cjasn.asnjournals.org/content/6/9/2313.long.

Mowatt G, Bower DJ, Brebner JA, Cairns JA, et al. When and how to assess fast-changing technologies: a comparative study of medical applications of four generic technologies. Health Technol Assess. 1997;1(14). http://www.journalslibrary.nihr.ac.uk/__data/assets/pdf_file/0006/64824/FullReport-hta1140.pdf.

Newcomer LN. Defining experimental therapy – a third party payer's dilemma. N Engl J Med. 1990;323(24):1702-4.

Nie JB. Non-medical sex-selective abortion in China: ethical and public policy issues in the context of 40 million missing females. Br Med Bull. 2011;98:7-20. http://bmb.oxfordjournals.org/content/98/1/7.long.

Nord E, Enge AU, Gundersen V. QALYs: is the value of treatment proportional to the size of the health gain? Health Econ. 2010;19(5):596-607.

Oldridge N, Furlong W, Perkins A, Feeny D, Torrance GW. Community or patient preferences for cost-effectiveness of cardiac rehabilitation: does it matter? Eur J Cardiovasc Prev Rehabil. 2008;15(5):608-15.

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

Patrick DL, Martin ML, Bushnell DM, Marquis P, Andrejasich CM, Buesching DP: Cultural adaptation of a quality-of-life measure for urinary incontinence. Eur Urol. 1999;36(5):427-35.

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.

Perkins NJ, Schisterman EF. The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. Am J Epidemiol. 2006;163(7):670-5. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1444894.

Peters WP, Rosner GL, Vredenburgh JJ, et al. Prospective, randomized comparison of high-dose chemotherapy with stem-cell support versus intermediate-dose chemotherapy after surgery and adjuvant chemotherapy in women with high-risk primary breast cancer: a report of CALGB 9082, SWOG 9114, and NCIC MA-13. J Clin Oncol. 2005;23(10):2191-200.

Post MW, Gerritsen J, Diederikst JP, DeWittet LP. Measuring health status of people who are wheelchair-dependent: validity of the Sickness Impact Profile 68 and the Nottingham Health Profile. Disabil Rehabil. 2001;23(6):245-53.

Potter BK, Avard D, Graham ID, et al. Guidance for considering ethical, legal, and social issues in health technology assessment: application to genetic screening. Int J Technol Assess Health Care. 2008;24(4):412-22.

Rabin R, de Charro F. EQ-5D: a measure of health status from the EuroQol Group. Ann Med. 2001;33(5):337-43.

Reiser SJ. Criteria for standard versus experimental therapy. Health Aff (Millwood). 1994;13(3):127-36. http://content.healthaffairs.org/content/13/3/127.long.

Rettig RA, Jacobson PD, Farquhar CM, Aubry WM. False Hope: Bone Marrow Transplantation for Breast Cancer. New York: Oxford University Press; 2007.

Rettig RA, Levinsky NG, eds. Kidney Failure and the Federal Government. Washington, DC: National Academy Press; 1991. http://www.nap.edu/openbook.php?record_id=1818&page=1.

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.

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(3):321-33.

Saarni SI, Braunack-Mayer A, Hofmann B, van der Wilt GJ. Different methods for ethical analysis in health technology assessment: An empirical study. Int J Technol Assess Health Care 2011;27(4):305-12.

Saban KL, Stroupe KT, Bryant FB, Reda DJ, et al. Comparison of health-related quality of life measures for chronic renal failure: quality of well-being scale, short-form-6D, and the kidney disease quality of life instrument. Qual Life Res. 2008;17(8):1103-15.

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.

Sassi F. Setting priorities for the evaluation of health interventions: when theory does not meet practice. Health Policy. 2003;63(2):141-54.

Schwappach DL. Resource allocation, social values and the QALY: a review of the debate and empirical evidence. Health Expect. 2002;5(3):210-22.

Singer SJ, Bergthold LA. Prospects for improved decision making about medical necessity. Health Aff (Millwood). 2001;20(1):200-6. http://content.healthaffairs.org/content/20/1/200.long.

Smidt N, Rutjes AW, van der Windt DA, Ostelo RW, et al. Reproducibility of the STARD checklist: an instrument to assess the quality of reporting of diagnostic accuracy studies. BMC Med Res Methodol. 2006;6:12. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1522016.

Söderlin MK, Kautiainen H, Skogh T, Leirisalo-Repo M. Quality of life and economic burden of illness in very early arthritis. A population based study in southern Sweden. J Rheumatol 2004;31(9):1717-22.

Soto-Gutierrez A, Wertheim JA, Ott HC, Gilbert TW. Perspectives on whole-organ assembly: moving toward transplantation on demand. J Clin Invest. 2012;122(11):3817-23. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3484436.

Srinivas SV, Deyo RA, Berger ZD. Application of "less is more" to low back pain. Arch Intern Med. 2012;172(13):1016-20.

Stamuli E. Health outcomes in economic evaluation: who should value health? Br Med Bull. 2011;97:197-210.

Straus SE, Richardson WS, Glasziou P, Haynes RB. Evidence-Based Medicine: How to Practice and Teach It. 4th ed. New York, NY: Churchill Livingstone Elsevier, 2011.

Taylor KS. Hospital tech assessment teams field requests − and politics. Hosp Health Netw. 1994;68(16):58.

Thiele AT, Leier B. Towards an ethical policy for the prevention of fetal sex selection in Canada. J Obstet Gynaecol Can. 2010 Jan;32(1):54-7.

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

Toh S, Hernández-Díaz S, Logan R, Rossouw JE, Hernán MA. Coronary heart disease in postmenopausal recipients of estrogen plus progestin therapy: does the increased risk ever disappear? A randomized trial. Ann Intern Med. 2010;152(4):211-7. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2936769.

Torrance GW, Feeny D. Utilities and quality-adjusted life years. Int J Technol Assess Health Care. 1989;5(4):559-75.

Tun W, Gange SJ, Vlahov D, Strathdee SA, Celentano DD. Increase in sexual risk behavior associated with immunologic response to highly active antiretroviral therapy among HIV-infected injection drug users. Clin Infect Dis. 2004;38(8):1167-74. http://www.cid.oxfordjournals.org/cgi/pmidlookup?view=long&pmid=15095224.

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 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 Food and Drug Administration. Guidance for Industry and FDA Staff. Pharmacogenetic Tests and Genetic Tests for Heritable Markers. June 19, 2007. Accessed Sept. 1, 2011 at: http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071075.pdf.

Wagner TH, Patrick DL, Bavendam TG, Martin ML, Buesching DP: Quality of life in persons with urinary incontinence: development of a new measure. Urology 1996;47(1):67-72.

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.

Whitehead SJ, Ali S. Health outcomes in economic evaluation: the QALY and utilities. Br Med Bull 2010;96:5-21. http://bmb.oxfordjournals.org/content/96/1/5.long.

Whiting PF, Rutjes AW, Westwood ME, et al.; QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529-36.

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.

Zhou S, Wang F, Hsieh TC, Wu JM, Wu E. Thalidomide-a notorious sedative to a wonder anticancer drug. Curr Med Chem. 2013;20(33):4102-8.

III. PRIMARY DATA METHODS

A. Primary Data Studies: Diverse Attributes

Primary data methods involve collection of original data, ranging from more scientifically rigorous approaches for determining the causal effect of health technologies, such as randomized controlled trials (RCTs), to less rigorous ones, such as case series. These study designs can be described and categorized based on multiple attributes or dimensions, e.g.:

- Comparative vs. non-comparative
- Separate (i.e., external) control group vs. no separate (i.e., internal) control group
- Participants (study populations /groups) defined by a health outcome vs. by having been exposed to, or received or been assigned, an intervention
- Prospective vs. retrospective
- Interventional vs. observational
- Experimental vs. non-experimental
- Random assignment vs. non-random assignment of patients to treatment and control groups

All experimental studies are, by definition, **interventional studies**. Some non-experimental studies can be interventional, e.g., if investigators assign a technology to a patient population but without a control group or with a non-randomized control group, and then assess their outcomes. An interventional cross-sectional design can be used to assess the accuracy of a diagnostic test. Some study designs are better at rigorous demonstration of causality in well-defined circumstances, such as the RCT. Other study designs may be better for reflecting real-world practice, such as pragmatic clinical trials and some **observational studies**, such as cohort, cross-sectional, or case control studies using data from registries, surveillance, electronic health (or medical) records, and payment claims.

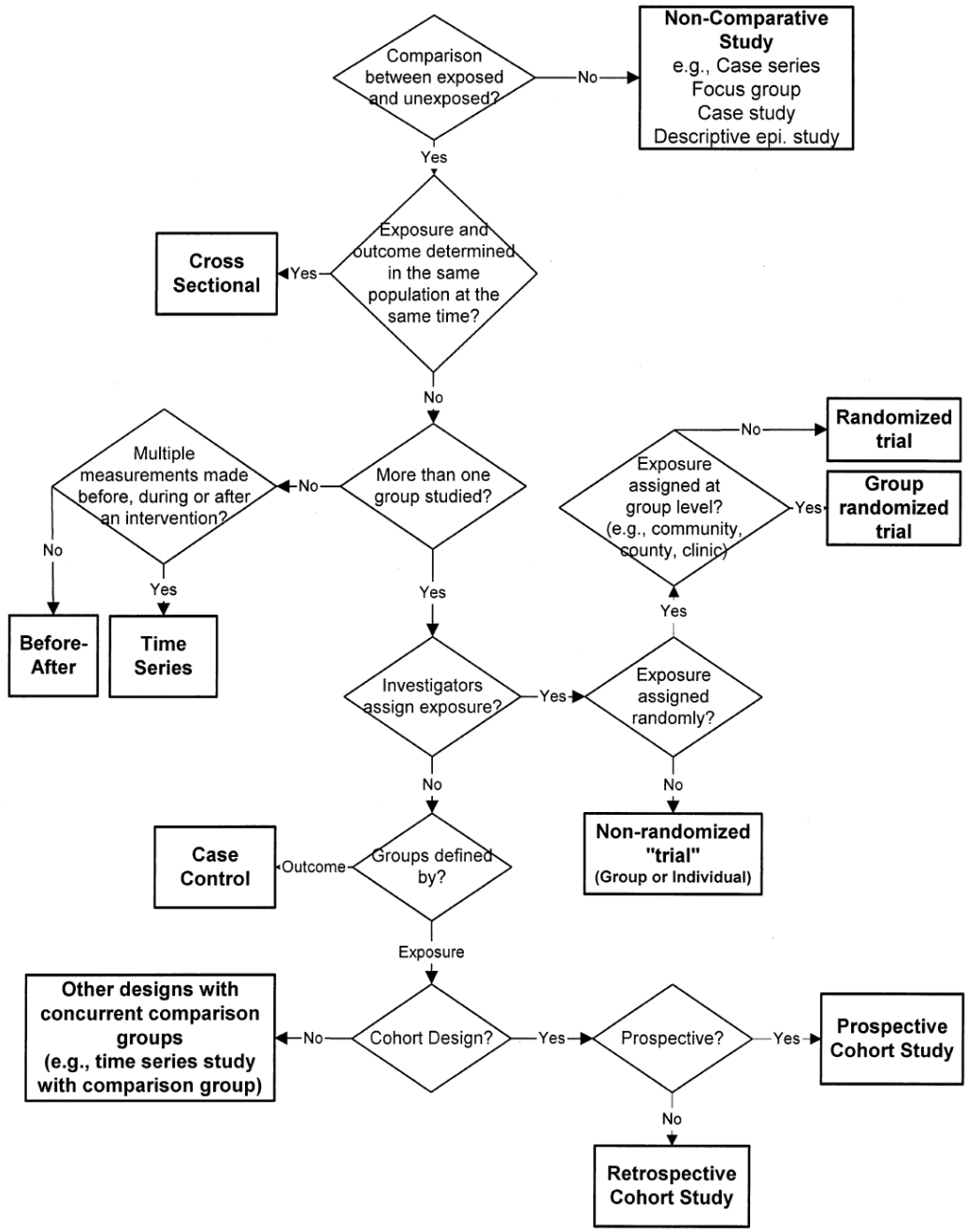
Box III-1. Examples of Experimental and Non-Experimental Study Designs

Experimental Studies	**Non-experimental studies**
Randomized controlled trial	Prospective cohort
Randomized cross-over trial	Retrospective cohort
N-of-1-trial	Case-control
Group randomized trial	Cross-sectional
Non-randomized controlled trial*	Interrupted time series with comparison
Pragmatic trials (randomized or non-randomized)	Non-concurrent cohort
	Interrupted time series without comparison
	Before-and-after
	Time series
	Case Series
	Case study

*A controlled trial in which participants are assigned to treatment and control groups using a method other than randomization, yet intended to form similar groups. Sometimes known as a "quasi-experimental" design.

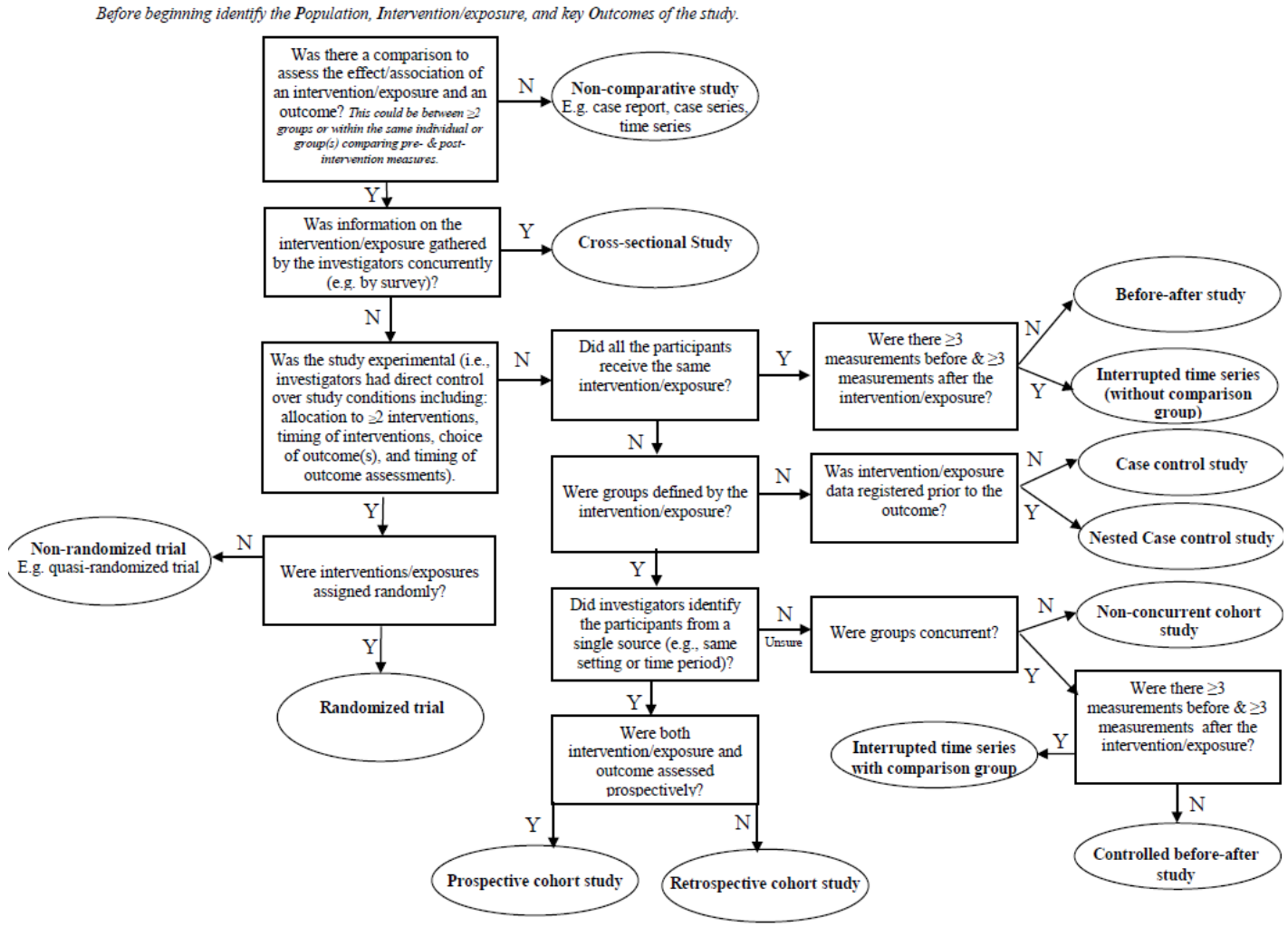
Box III-1 categorizes various types of primary data studies as experimental and non-experimental. Researchers have developed various frameworks, schemes, and other tools for classifying study designs, such as for the purpose of conducting systematic reviews (Hartling 2010). **Box III-2** and **Box III-3** show algorithms for identifying study designs. Some of these study designs have alternative names, and some studies use diverse combinations of design attributes.

Box III-2. Study Design Algorithm, Guide to Community Preventive Services



Source: Briss PA, Zasa S, Pappaioanou M, Fielding J, et al. Developing an evidence-based Guide to Community Preventive Services--Am J Prev Med 2000;18(1S):35-43, Copyright © 2000) with permission from Elsevier.

Box III-3. Design Algorithm for Studies of Health Care Interventions*



*Developed, though no longer advocated by, the Cochrane Non-Randomised Studies Methods Group.

Source: Hartling L, et al. Developing and Testing a Tool for the Classification of Study Designs in Systematic Reviews of Interventions and Exposures. Agency for Healthcare Research and Quality; December 2010. Methods Research Report. AHRQ Publication No. 11-EHC-007.

Although the general type of a study design (e.g., RCT, prospective cohort study, case series) conveys certain attributes about the quality of a study (e.g., control group, random assignment), study design type alone is not a good proxy for study quality. More important are the attributes of study design and conduct that diminish sources of bias and random error, as described below.

New types of observational study designs are emerging in the form of patient-centered online registries and related research platforms. For example, PatientsLikeMe, a patient network, is set up for entry of member patient demographic information, treatment history, symptoms, outcome data, and evaluations of treatments, as well as production of individual longitudinal health profiles and aggregated reports. Such patient-centered registries can supplement clinical trials and provide useful postmarket data across heterogeneous patients and circumstances (Frost 2011, Nakamura 2012).

Most HTA programs rely on integrative methods (especially systematic reviews), particularly to formulate findings based on available evidence from primary data studies that are identified through systematic literature searches. Some HTA programs 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 rigorously designed studies. Indeed, policies and decisions often must be made in the absence, or before completion, of definitive studies. Given their varying assessment purposes, resource constraints, and other factors, HTA programs use evidence from various study designs, although they usually emphasize evidence based on the more rigorous and systematic methods of data collection.

The following sections describe concepts that affect the quality of primary data studies, particularly their ability to yield unbiased and precise estimates of treatment effects and other findings.

B. Assessing the Quality of Primary Data Studies

Our confidence that the estimate of a treatment effect, accuracy of a screening or diagnostic test, or other impact of a health care technology that is generated by a study is correct reflects our understanding of the quality of the study. For various types of interventions, we examine certain attributes of the design and conduct of a study to assess the quality of that study. For example, some of the attributes or criteria that are commonly used to assess the quality of studies for demonstrating the **internal validity** of the impact of therapies on health outcomes are the following:

- Prospective, i.e., following a study population over time as it receives an intervention or exposure and experiences outcomes, rather than retrospective design
- Experimental rather than observational
- Controlled, i.e., with one or more comparison groups, rather than uncontrolled
- Contemporaneous control groups rather than historical ones
- Internal (i.e., managed within the study) control groups rather than external ones
- Allocation concealment of patients to intervention and control groups
- Randomized assignment of patients to intervention and control groups
- Blinding of patients, clinicians, and investigators as to patient assignment to intervention and control groups
- Large enough sample size (number of patients/participants) to detect true treatment effects with statistical significance
- Minimal patient drop-outs or loss to follow-up of patients (or differences in these between intervention and control groups) for duration of study
- Consistency of pre-specified study protocol (patient populations, assignment to intervention and control groups, regimens, etc.) and outcome measures with the reported (post-study) protocol and outcome measures

Similarly, some attributes that are commonly used for assessing the **external validity** of the impact of therapies and other technologies on health outcomes include:

- Flexible entry criteria to identify/enroll patient population that is representative of patient diversity likely to be offered the intervention in practice, including demographic characteristics, risk factors, disease stage/severity, comorbidities
- Large enough patient population to conduct meaningful subgroup analyses (especially for pre-specified subgroups)
- Dosing, regimen, technique, delivery of the intervention consistent with anticipated practice
- Comparator is standard of care or other relevant, clinically acceptable (not-substandard) intervention
- Dosing, regimen, or other forms of delivering the comparator consistent with standard care
- Patient monitoring and efforts to maintain patient adherence comparable to those in practice
- Accompanying/concurrent/ancillary care similar to what will be provided in practice
- Training, expertise, skills of clinicians and other health care providers similar to those available or feasible for providers anticipated to deliver the intervention
- Selection of outcome measures relevant to those experienced by and important to intended patient groups
- Systematic effort to follow-up on all patients to minimize attrition
- Intention-to-treat analysis used to account for all study patients
- Study duration consistent with the course/episode of disease/condition in practice in order to detect outcomes of importance to patients and clinicians
- Multiple study sites representative of type/level of health care settings and patient and clinician experience anticipated in practice

RCTs are designed to maximize internal validity, and are generally regarded as the "gold standard" study design for demonstrating the causal impact of a technology on health care outcomes. However, some attributes that strengthen the internal validity of RCTs tend to diminish RCTs' external validity. Probing the strengths and limitations of RCTs with respect to internal and external validity is also instructive for understanding the utility of other studies. A variety of design aspects intended to improve the external validity of RCTs and related experimental designs are described briefly later in this chapter.

The commonly recognized attributes of study quality noted above that strengthen internal and external validity of primary data studies are derived from and extensive body of methodological concepts and principles, including those summarized below: confounding and the need for controls, prospective vs. retrospective design, sources of bias, random error, and selected other factors.

1. Types of Validity in Methods and Measurement

Whether they are experimental or non-experimental in design, studies vary in their ability to produce valid findings. Validity refers to how well a study or data collection instrument measures what it is intended to measure. Understanding different aspects of validity helps in comparing strengths and weaknesses of alternative study designs and our confidence in the findings generated by those studies. Although these concepts are often addressed in reference to primary data methods, they generally apply as well to integrative methods.

Internal validity refers to the extent to which the results of a study accurately represent the causal relationship between an intervention and an outcome in the particular circumstances of that study. This includes the extent to which the design and conduct of a study minimize the risk of any systematic (non-random) error (i.e., bias) in the study results. Internal validity can be suspect when biases in the design or conduct of a clinical trial or other study could have affected outcomes, thereby causing the study results to deviate from the true magnitude of the treatment effect. True experiments such as RCTs generally have high internal validity.

External validity refers to the extent to which the results of a study conducted under particular circumstances can be generalized (or are applicable) to other circumstances. When the circumstances of a particular study (e.g., patient characteristics, the technique of delivering a treatment, or the setting of care) differ from the circumstances of interest (e.g., patients with different characteristics, variations in the technique of delivering a treatment, or alternative settings of care), the external validity of the results of that study may be limited.

Construct validity refers to how well a measure is correlated with other accepted measures of the construct of interest (e.g., pain, anxiety, mobility, quality of life), and discriminates between groups known to differ according to the construct. **Face validity** is the ability of a measure to represent reasonably (i.e., to be acceptable "on its face") a construct (i.e., a concept, trait, or domain of interest) as judged by someone with knowledge or expertise in the construct.

Content validity refers to the degree to which the set of items of a data collection instrument is known to represent the range or universe of meanings or dimensions of a construct of interest. For example, how well do the domains of a health-related quality of life index for rheumatic arthritis represent the aspects of quality of life or daily functioning that are important to patients with rheumatoid arthritis?

Criterion validity refers to how well a measure, including its various domains or dimensions, is correlated with a known gold standard or definitive measurement, if one exists. The similar concept of **concurrent validity** refers to how well a measure correlates with a previously validated one, and the ability of a measure to accurately differentiate between different groups at the time the measure is applied. **Predictive validity** refers to the ability to use differences in a measure of a construct to predict future events or outcomes. It may be considered a subtype of criterion validity.

Convergent validity refers to the extent to which different measures that are intended to measure the same construct actually yield similar results, such as two measures of quality of life. **Discriminant validity** concerns whether different measures that are intended to measure different constructs actually fail to be positively associated with each other. Convergent validity and discriminant validity contribute to, or can be considered subtypes of, construct validity.

2. Confounding and the Need for Controls

Confounding occurs when any factor that is associated with an intervention has an impact on an outcome that is independent of the impact of the intervention. As such, confounding can "mask" or muddle the true impact of an intervention. In order to diminish any impact of confounding factors, it is necessary to provide a basis of comparing what happens to patients who receive an intervention to those that do not.

The main purpose of control groups is to enable isolating the impact of an intervention of interest on patient outcomes from the impact of any extraneous factors. The composition of the control group is intended to be as close as possible to that of the intervention group, and both groups are managed as similarly as possible, so that the only difference between the groups is that one receives the intervention of interest and the other does not. In controlled clinical trials, the control groups may receive a current standard of care, no intervention, or a placebo.

For a factor to be a confounder in a controlled trial, it must differ for the intervention and control groups and be predictive of the treatment effect, i.e., it must have an impact on the treatment effect that is independent of the intervention of interest. Confounding can arise due to differences between the intervention and control groups, such as differences in baseline risk factors at the start of a trial or different exposures during the trial that could affect outcomes. Investigators may not be aware of all potentially confounding factors in a trial. Examples of potentially confounding factors are age, prevalence of comorbidities at baseline, and different levels of ancillary care. To the extent that potentially confounding factors are present at different rates between comparison groups, a study is subject to **selection bias** (described below).

Most controlled studies use contemporaneous controls alongside (i.e., constituted and followed simultaneously with) intervention groups. Investigators sometimes rely on **historical control** groups. However, a historical control group is subject to known or unknown inherent differences (e.g., risk factors or other prognostic factors) from a current intervention group, and environmental or other contextual differences arising due to the passage of time that may confound outcomes. In some instances, including those noted below, historical controls have sufficed to demonstrate definitive treatment effects. In a **crossover design** study, patients start in one group (intervention or control) and then are switched to the other (sometimes multiple times), thereby acting as their own controls, although such designs are subject to certain forms of bias.

Various approaches are used to ensure that intervention and control groups comprise patients with similar characteristics, diminishing the likelihood that baseline differences between them will confound observed treatment effects. The best of these approaches is **randomization** of patients to intervention and control groups. Random allocation diminishes the impact of any potentially known or unrecognized confounding factors by tending to distribute those factors evenly across the groups to be compared. "Pseudo-randomization" approaches such as alternate assignment or using birthdays or identification numbers to assign patients to intervention and control groups can be vulnerable to confounding.

Placebo Controls

Among the ongoing 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); and (4) when and how to use the placebo effect to patient advantage. As in other health technologies, surgical procedures can be subject to the **placebo effect**. Following previous missteps that raised profound ethical concerns, guidance was developed 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). Even so, the circumstances in which placebo surgery is ethically and scientifically acceptable as well as practically feasible and acceptable to enrolled patients may be very limited (Campbell 2011).

In recent years there has been considerable scientific progress in understanding the physiological and psychological basis of the placebo response, prompting efforts to put it to use in improving outcomes. It remains important to control for the placebo effect in order to minimize its confounding effect on evaluating the treatment effect of an intervention. However, once a new drug or other technology is in clinical use, the patient expectations and learning mechanisms contributing to the placebo effect may be incorporated into medication regimens to improve patient satisfaction and outcomes. Indeed, this approach may be personalized based on patient genomics, medical history, and other individual characteristics (Enck 2013).

3. Prospective vs. Retrospective Design

Prospective studies are planned and implemented by investigators using real-time data collection. These typically involve identification of one or more patient groups according to specified risk factors or exposures, followed by collection of baseline (i.e., initial, prior to intervention) data, delivery of interventions of interest and controls, 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 interventions and outcomes involving one or more patient groups.

Prospective studies are usually subject to fewer types of confounding and bias than retrospective studies. In particular, retrospective studies are more subject to selection bias than prospective studies. In retrospective studies, patients' interventions and outcomes have already transpired and been recorded, raising opportunities for intentional or unintentional selection bias on the part of investigators. In prospective studies, patient enrollment and data collection can be designed to reduce bias (e.g., selection bias and detection bias), which is an advantage over most retrospective studies. Even so, the logistical challenges of maintaining blinding of patients and investigators are considerable and unblinding can introduce performance and detection bias.

Prospective and retrospective studies have certain other relative advantages and disadvantages that render them more or less useful for certain types of research questions. Both are subject to certain types of missing or otherwise limited data. As retrospective studies primarily involve selection and analyses of existing data, they tend to be far less expensive than prospective studies. However, their dependence on existing data makes it difficult to fill data gaps or add data fields to data collection instruments, although they can rely in part on importing and adjusting data from other existing sources. Given the costs of enrolling enough patients and collecting sufficient data to achieve statistical significance, prospective studies tend to be more suited to investigating health problems that are prevalent and yield health outcomes or other events that occur relatively frequently and within short follow-up periods. The typically shorter follow-up periods of prospective studies may subject them to seasonal or other time-dependent biases, whereas retrospective studies can be designed to extract data from longer time spans. Retrospective studies offer the advantage of being able to canvass large volumes of data over extended time periods (e.g., from registries, insurance claims, and electronic health records) to identify patients with specific sets of risk factors or rare or delayed health outcomes, including certain adverse events.

4. Sources of Bias

The quality of a primary data study determines our confidence that the estimated treatment effect in a primary data study is correct. **Bias** refers to any *systematic* (i.e., not due to random error) deviation in an observation from the true nature of an event. In clinical trials, bias may arise from any factor that systematically distorts (increases or decreases) the observed magnitude of an outcome (e.g., treatment effect or harm) relative to the true magnitude of the outcome. As such, bias diminishes the accuracy (though not necessarily the precision; see discussion below) of an observation. Biases may arise from inadequacies in the design, conduct, analysis, or reporting of a study.

Major types of bias in comparative primary data studies are described below, including selection bias, performance bias, detection bias, attrition bias, and reporting bias (Higgins, Altman, Gøtzsche 2011; Higgins, Altman, Sterne 2011; Viswanathan 2014). Also noted are techniques and other study attributes that tend to diminish each type of bias. These attributes for diminishing bias also serve as criteria for assessing the quality of individual studies.

Selection bias refers to systematic differences between baseline characteristics of the groups that are compared, which can arise from, e.g., physician assignment of patients to treatments, patient self-selection of treatments, or association of treatment assignment with patient clinical characteristics or demographic factors. Among the means for diminishing selection bias are random sequence generation (random allocation of patients to treatment and control groups) and **allocation concealment** for RCTs, control groups to diminish confounders in cohort studies, and case matching in case-control studies.

Allocation concealment refers to the process of ensuring that the persons assessing patients for potential entry into a trial, as well as the patients, do not know whether any particular patient will be allocated to an intervention group or control group. This helps to prevent the persons who manage the allocation, or the patients, from influencing (intentionally or not) the assignment of a patient to one group or another. Patient allocation based on personal identification numbers, birthdates, or medical record numbers may not ensure concealment. Better methods include centralized randomization (i.e., managed at one site rather than at each enrollment site); sequentially numbered, opaque, sealed envelopes; and coded medication bottles or containers.

Performance bias refers to systematic differences between comparison groups in the care that is provided, or in exposure to factors other than the interventions of interest. This includes, e.g., deviating from the study protocol or assigned treatment regimens so that patients in control groups receive the intervention of interest, providing additional or co-interventions unevenly to the intervention and control groups, and inadequately blinding providers and patients to assignment to intervention and control groups, thereby potentially affecting whether or how assigned interventions or exposures are delivered. Techniques for diminishing performance bias include blinding of patients and providers (in RCTs and other controlled trials in particular), adhering to the study protocol, and sustaining patients' group assignments.

Detection (or ascertainment) bias refers to systematic differences between groups in how outcomes are assessed. These differences may arise due to, e.g., inadequate blinding of outcome assessors regarding patient treatment assignment, reliance on patient or provider recall of events (also known as **recall bias**), inadequate outcome measurement instruments, or faulty statistical analysis. Whereas allocation concealment is intended to ensure that persons who manage patient allocation, as well as the patients themselves, do not influence patient assignment to one group or another, **blinding** refers to

preventing anyone who could influence assessment of outcomes from knowing which patients have been assigned to one group or another. Knowledge of patient assignment itself can affect outcomes as experienced by patients or assessed by investigators. The techniques for diminishing detection bias include blinding of outcome assessors including patients, clinicians, investigators, and/or data analysts, especially for subjective outcome measures; and validated and reliable outcome measurement instruments and techniques.

Attrition bias refers to systematic differences between comparison groups in withdrawals (drop-outs) from a study, loss to follow-up, or other exclusion of patients/participants and how these losses are analyzed. Ignoring these losses or accounting for them differently between groups can skew study findings, as patients who withdraw or are lost to follow-up may differ systematically from those patients who remain for the duration of the study. Indeed, patients' awareness of whether they have been assigned to a particular treatment or control group may differentially affect their likelihood of dropping out of a trial. Techniques for diminishing attrition bias include blinding of patients as to treatment assignment, completeness of follow-up data for all patients, and intention-to-treat analysis (with imputations for missing data as appropriate).

Reporting bias refers to systematic differences between reported and unreported findings, including, e.g., differential reporting of outcomes between comparison groups and incomplete reporting of study findings (such as reporting statistically significant results only). Also, narrative and systematic reviews that do not report search strategies or disclose potential conflicts of interest raise concerns about reporting bias as well as selection bias (Roundtree 2009). Techniques for diminishing reporting bias include thorough reporting of outcomes consistent with outcome measures specified in the study protocol, including attention to documentation and rationale for any post-hoc (after the completion of data collection) analyses not specified prior to the study, and reporting of literature search protocols and results for review articles. Reporting bias, which concerns differential or incomplete reporting of findings in individual studies, is not the same as **publication bias**, which concerns the extent to which all relevant studies on given topic proceed to publication.

Registration of Clinical Trials and Results

Two related sets of requirements have improved clinical trial reporting for many health technologies. These requirements help to diminish reporting bias and publication bias, thereby improving the quality of the evidence available for HTA. Further, they increase the value of clinical trials more broadly to trial participants, patients, clinicians, and other decision makers, and society (Huser 2013).

In the US, the Food and Drug Administration Amendments Act of 2007 (FDAAA) mandates that certain clinical trials of drugs, biologics, and medical devices that are subject to FDA regulation for any disease or condition be registered with ClinicalTrials.gov. A service of the US National Library of Medicine, ClinicalTrials.gov is a global registry and results database of publicly and privately supported clinical studies. Further, FDAAA requires investigators to register the results of these trials, generally no more than 12 months after trial completion. Applicable trials include those that have one or more sites in the US, are conducted under an FDA investigational new drug application (IND) or investigational device exemption (IDE), or involve a drug, biologic, or device that is manufactured in the US and its territories and is exported for research (ClinicalTrials.gov 2012; Zarin 2011).

The International Committee of Medical Journal Editors (ICMJE) requires clinical trial registration as a condition for publication of research results generated by a clinical trial. Although the ICMJE does not

advocate any particular registry, it is required that a registry meet certain criteria for investigators to meet the condition for publication. (ClinicalTrials.gov meets these criteria.) ICMJE requires registration of trial methodology but not trials results (ICMJE 2013).

As noted above, study attributes that affect bias can be used as criteria for assessing the quality of individual studies. For example, the use of randomization to reduce selection bias and blinding of outcomes assessors to reduce detection bias are among the criteria used for assessing the quality of clinical trials. Even within an individual study, the extent of certain types of bias may vary for different outcomes. For example, in a study of the impact of a technology on mortality and quality of life, the presence of detection bias and reporting bias may vary for those two outcomes.

Box III-4 shows a set of criteria for assessing risk of bias for benefits for several types of study design based on the main types of risk of bias cited above and used by the US Agency for Healthcare Research and Quality (AHRQ) Evidence-based Practice Centers (EPCs).

5. Random Error

In contrast to the systematic effects of various types of bias, **random error** is a source of *non-systematic* deviation of an observed treatment effect or other outcome from a true one. Random error results from chance variation in the sample of data collected in a study (i.e., sampling error). The extent to which an observed outcome is free from random error is **precision**. As such, precision is inversely related to random error.

Random error can be reduced, but it cannot be eliminated. ***P*-values** and **confidence intervals** account for the extent of random error, but they do not account for systematic error (bias). The main approach to reducing random error is to establish large enough sample sizes (i.e., numbers of patients in the intervention and control groups of a study) to detect a true treatment effect (if one exists) at acceptable levels of statistical significance. The smaller the true treatment effect, the more patients may be required to detect it. Therefore, investigators who are planning an RCT or other study consider the estimated magnitude of the treatment effect that they are trying to detect at an acceptable level of statistical significance, and then "**power**" (i.e., determine the necessary sample size of) the study accordingly. Depending on the type of treatment effect or other outcome being assessed, another approach to reducing random error is to reduce variation in an outcome for each patient by increasing the number of observations made for each patient. Random error also may be reduced by improving the precision of the measurement instrument used to take the observations (e.g., a more precise diagnostic test or instrument for assessing patient mobility).

6. Role of Selected Other Factors

Some researchers contend that if individual studies are to be assembled into a body of evidence for a systematic review, precision should be evaluated not at the level of individual studies, but when assessing the quality of the body of evidence. This is intended to avoid double-counting limitations in precision from the same source (Viswanathan 2014).

In addition to evaluating internal validity of studies, some instruments for assessing the quality of individual studies evaluate **external validity**. However, by definition, the external validity of a study depends not only on its inherent attributes, but on the nature of an evidence question for which the study is more or less relevant. An individual study may have high external validity for some evidence questions and low external validity for others. Certainly, some aspects of bias for internal validity noted

above may also be relevant to external validity, such as whether the patient populations compared in a treatment and control group represent same or different populations, and whether the analyses account for attrition in a way that represents the population of interest, including any patient attributes that differ between patients who were followed to study completion and those who were lost to follow-up. Some researchers suggest that if individual studies are to be assembled into a body of evidence for a systematic review, then external validity should be evaluated when assessing the quality of the body of evidence, but not at the level of individual studies (Atkins 2004; Viswanathan 2014).

Box III-4. Design-Specific Criteria to Assess Risk of Bias for Benefits

Risk of bias	Criterion	RCTs	CCTs or cohort	Case-control	Case series	Cross-sectiona
Selection bias	Was the allocation sequence generated adequately (e.g., random number table, computer-generated randomization)?	x				
	Was the allocation of treatment adequately concealed (e.g., pharmacy- controlled randomization or use of sequentially numbered sealed envelopes)?	x				
	Were participants analyzed within the groups they were originally assigned to?	x	x			
	Did the study apply inclusion/exclusion criteria uniformly to all comparison groups?		x			x
	Were cases and controls selected appropriately (e.g., appropriate diagnostic criteria or definitions, equal application of exclusion criteria to case and controls, sampling not influenced by exposure status)			x		
	Did the strategy for recruiting participants into the study differ across study groups?		x			
	Does the design or analysis control account for important confounding and modifying variables through matching, stratification, multivariable analysis, or other approaches?	x	x	x	x	x
Performance bias	Did researchers rule out any impact from a concurrent intervention or an unintended exposure that might bias results?	x	x	x	x	x
	Did the study maintain fidelity to the intervention protocol?	x	x	x	x	
Attrition bias	If attrition (overall or differential nonresponse, dropout, loss to follow-up, or exclusion of participants) was a concern, were missing data handled appropriately (e.g., intention-to-treat analysis and imputation)?	x	x	x	x	x
Detection bias	In prospective studies, was the length of follow-up different between the groups, or in case-control studies, was the time period between the intervention/exposure and outcome the same for cases and controls?	x	x	x		
	Were the outcome assessors blinded to the intervention or exposure status of participants?	x	x	x	x	x
	Were interventions/exposures assessed/defined using valid and reliable measures, implemented consistently across all study participants?	x	x	x	x	x
	Were outcomes assessed/defined using valid and reliable measures, implemented consistently across all study participants?	x	x	x	x	x
	Were confounding variables assessed using valid and reliable measures, implemented consistently across all study participants?		x	x	x	x
Reporting bias	Were the potential outcomes prespecified by the researchers? Are all prespecified outcomes reported?	x	x	x	x	x

*Cases and controls should be similar in all factors known to be associated with the disease of interest, but they should not be so uniform as to be matched for the exposure of interest.

Source: Viswanathan M, Ansari MT, Berkman ND, Chang S, et al. Chapter 9. Assessing the risk of bias of individual studies in systematic reviews of health care interventions. In: Methods Guide for Effectiveness and Comparative Effectiveness Reviews. AHRQ Publication No. 10(14)-EHC063-EF. Rockville, MD: Agency for Healthcare Research and Quality. January 2014.

Some quality assessment tools for individual studies account for funding source (or sponsor) of a study and disclosed conflicts of interest (e.g., on the part of sponsors or investigators) as potential sources of bias. Rather than being direct sources of bias themselves, a funding source or a person with a disclosed conflict of interest may induce bias indirectly, e.g., in the form of certain types of reporting bias or detection bias. Also, whether the funding source of research comes is a government agency, non-profit organization, or health technology company does not necessarily determine whether it induces bias. Of course, all of these potential sources of bias should be systematically documented (Viswanathan 2014).

C. Instruments for Assessing Quality of Individual Studies

A variety of assessment instruments are available to assess the quality of individual studies. Many of these are for assessing internal validity or risk of bias for benefits and harms; others focus on assessing external validity. These include instruments for assessing particular types of studies (e.g., RCTs, observational studies) and certain types of interventions (e.g., screening, diagnosis, and treatment).

A systematic review identified more than 20 scales (and their modifications) for assessing the quality of RCTs (Olivo 2008). Although most of these had not been rigorously developed or tested for validity and reliability, the systematic review found that one of the original scales, the Jadad Scale (Jadad 1996), shown in **Box III-5**, was the strongest.

The Cochrane Risk of Bias Tool for RCTs, shown **in Box III-6**, accounts for the domains of bias noted above (i.e., selection, performance, detection, attrition, and reporting bias), providing criteria for assessing whether there is low, unclear, or high risk of bias for each domain for individual RCTs as well as across a set of RCTs for a particular evidence question (Higgins, Altman, Sterne 2011).

Criteria and ratings for assessing internal validity of RCTs and cohort studies and of diagnostic accuracy studies used by the US Preventive Services Task Force (USPSTF) are shown in **Box III-7** and **Box III-8**, respectively. **Box III-9** shows a framework used by the USPSTF to rate the external validity of individual studies. QUADAS-2 is a quality assessment tool for diagnostic accuracy studies (Whiting 2011).

Among their numerous instruments for assessing the quality of individual studies, the AHRQ EPCs use a PICOS framework to organize characteristics that can affect the external validity of individual studies, which are used as criteria for evaluating study quality for internal validity, as shown in **Box III-10**.

D. Strengths and Limitations of RCTs

For demonstrating the internal validity of a causal relationship between an intervention and one or more outcomes of interest, the well-designed, blinded (where feasible), appropriately powered, well-conducted, and properly reported RCT has dominant advantages over other study designs. Among these, the RCT minimizes selection bias in that any enrolled patient has the same probability, due to randomization, of being assigned to an intervention group or control group. This also minimizes the potential impact of any known or unknown confounding factors (e.g., risk factors present at baseline), because randomization tends to distribute such confounders evenly across the groups to be compared.

When the sample size of an RCT is calculated to achieve sufficient statistical power, it minimizes the probability that the observed treatment effect will be subject to random error. Further, especially with larger groups, randomization enables patient subgroup comparisons between intervention and control groups. The primacy of the RCT remains even in an era of genomic testing and expanding use of

biomarkers to better target selection of patients for adaptive clinical trials of new drugs and biologics, and advances in computer-based modeling that may replicate certain aspects of RCTs (Ioannidis 2013).

Box III-5. 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:

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.

Box III-6. The Cochrane Collaboration's Tool for Assessing Risk of Bias

Domain	Support for Judgment	Review authors' judgement
Selection bias.		
Random sequence generation.	Describe the method used to generate the allocation sequence in sufficient detail to allow an assessment of whether it should produce comparable groups.	Selection bias (biased allocation to interventions) due to inadequate generation of a randomized sequence.
Allocation concealment.	Describe the method used to conceal the allocation sequence in sufficient detail to determine whether intervention allocations could have been foreseen in advance of, or during, enrolment.	Selection bias (biased allocation to interventions) due to inadequate concealment of allocations prior to assignment.
Performance bias.		
Blinding of participants and personnel *Assessments should be made for each main outcome (or class of outcomes).*	Describe all measures used, if any, to blind study participants and personnel from knowledge of which intervention a participant received. Provide any information relating to whether the intended blinding was effective.	Performance bias due to knowledge of the allocated interventions by participants and personnel during the study.
Detection bias.		
Blinding of outcome assessment *Assessments should be made for each main outcome (or class of outcomes).*	Describe all measures used, if any, to blind outcome assessors from knowledge of which intervention a participant received. Provide any information relating to whether the intended blinding was effective.	Detection bias due to knowledge of the allocated interventions by outcome assessors.
Attrition bias.		
Incomplete outcome data *Assessments should be made for each main outcome (or class of outcomes).*	Describe the completeness of outcome data for each main outcome, including attrition and exclusions from the analysis. State whether attrition and exclusions were reported, the numbers in each intervention group (compared with total randomized participants), reasons for attrition/exclusions where reported, and any re-inclusions in analyses performed by the review authors.	Attrition bias due to amount, nature or handling of incomplete outcome data.
Reporting bias.		
Selective reporting.	State how the possibility of selective outcome reporting was examined by the review authors, and what was found.	Reporting bias due to selective outcome reporting.
Other bias.		
Other sources of bias.	State any important concerns about bias not addressed in the other domains in the tool. If particular questions/entries were pre-specified in the review's protocol, responses should be provided for each question/entry.	Bias due to problems not covered elsewhere in the table.

Reprinted with permission: Higgins JPT, Altman DG, Sterne, JAC, eds. Chapter 8: Assessing risk of bias in included studies. In: Higgins JPT, Green S, eds. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011.

Box III-7. Criteria for Assessing Internal Validity of Individual Studies: Randomized Controlled Trials and Cohort Studies, USPSTF

Criteria:

- Initial assembly of comparable groups:
  - For RCTs: adequate randomization, including first concealment and whether potential confounders were distributed equally among groups.
  - For cohort studies: consideration of potential confounders with either restriction or measurement for adjustment in the analysis; consideration of inception cohorts.
- Maintenance of comparable groups (includes attrition, cross-overs, adherence, contamination).
- Important differential loss to follow-up or overall high loss to follow-up.
- Measurements: equal, reliable, and valid (includes masking of outcome assessment).
- Clear definition of interventions.
- All important outcomes considered.
- Analysis: adjustment for potential confounders for cohort studies, or intention to treat analysis for RCTs.

Definitions of ratings based on above criteria:

Good: Meets all criteria: Comparable groups are assembled initially and maintained throughout the study (follow-up at least 80 percent); reliable and valid measurement instruments are used and applied equally to the groups; interventions are spelled out clearly; all important outcomes are considered; and appropriate attention to confounders in analysis. In addition, for RCTs, intention to treat analysis is used.

Fair: Studies will be graded "fair" if any or all of the following problems occur, without the fatal flaws noted in the "poor" category below: Generally comparable groups are assembled initially but some question remains whether some (although not major) differences occurred with follow-up; measurement instruments are acceptable (although not the best) and generally applied equally; some but not all important outcomes are considered; and some but not all potential confounders are accounted for. Intention to treat analysis is done for RCTs.

Poor: Studies will be graded "poor" if any of the following fatal flaws exists: Groups assembled initially are not close to being comparable or maintained throughout the study; unreliable or invalid measurement instruments are used or not applied at all equally among groups (including not masking outcome assessment); and key confounders are given little or no attention. For RCTs, intention to treat analysis is lacking.

Source: US Preventive Services Task Force Procedure Manual. AHRQ Pub. No. 08-05118-EF, July 2008.

Box III-8. Criteria for Assessing Internal Validity of Individual Studies: Diagnostic Accuracy Studies, USPSTF

Criteria:

- Screening test relevant, available for primary care, adequately described.
- Study uses a credible reference standard, performed regardless of test results.
- Reference standard interpreted independently of screening test.
- Handles indeterminate results in a reasonable manner.
- Spectrum of patients included in study.
- Sample size.
- Administration of reliable screening test.

Definitions of ratings based on above criteria:

Good: Evaluates relevant available screening test; uses a credible reference standard; interprets reference standard independently of screening test; reliability of test assessed; has few or handles indeterminate results in a reasonable manner; includes large number (more than 100) broad-spectrum patients with and without disease.

Fair: Evaluates relevant available screening test; uses reasonable although not best standard; interprets reference standard independent of screening test; moderate sample size (50 to 100 subjects) and a "medium" spectrum of patients.

Poor: Has fatal flaw such as: Uses inappropriate reference standard; screening test improperly administered; biased ascertainment of reference standard; very small sample size or very narrow selected spectrum of patients.

Source: US Preventive Services Task Force Procedure Manual. AHRQ Pub. No. 08-05118-EF, July 2008.

Box III-9. Global Rating of External Validity (Generalizability) of Individual Studies, US Preventive Services Task Force

External validity is rated "good" if:

- The study differs minimally from the US primary care population/situation/providers and only in ways that are unlikely to affect the outcome; it is highly probable (>90%) that the clinical experience with the intervention observed in the study will be attained in the US primary care setting.

External validity is rated "fair" if:

- The study differs from the US primary care population/situation/providers in a few ways that have the potential to affect the outcome in a clinically important way; it is only moderately probable (50%-89%) that the clinical experience with the intervention in the study will be attained in the US primary care setting.

External validity is rated "poor" if:

- The study differs from the US primary care population/ situation/ providers in many way that have a high likelihood of affecting the clinical outcomes; the probability is low (<50%) that the clinical experience with the intervention observed in the study will be attained in the US primary care setting.

Source: US Preventive Services Task Force Procedure Manual. AHRQ Pub. No. 08-05118-EF, July 2008.

Box III-10. Characteristics of Individual Studies That May Affect Applicability (AHRQ)

Population

- Narrow eligibility criteria and exclusion of those with comorbidities
- Large differences between demographics of study population and community patients
- Narrow or unrepresentative severity, stage of illness, or comorbidities
- Run in period with high-exclusion rate for non-adherence or side effects
- Event rates much higher or lower than observed in population-based studies

Intervention

- Doses or schedules not reflected in current practice
- Intensity and delivery of behavioral interventions that may not be feasible for routine use
- Monitoring practices or visit frequency not used in typical practice
- Older versions of an intervention no longer in common use
- Co-interventions that are likely to modify effectiveness of therapy
- Highly selected intervention team or level of training/proficiency not widely available

Comparator

- Inadequate dose of comparison therapy
- Use of substandard alternative therapy

Outcomes

- Composite outcomes that mix outcomes of different significance
- Short-term or surrogate outcomes

Setting

- Standards of care differ markedly from setting of interest
- Specialty population or level of care differs from that seen in community

Source: Atkins D, et al. Chapter 5. Assessing the Applicability of Studies When Comparing Medical Interventions. In: Methods Guide for Effectiveness and Comparative Effectiveness Reviews. AHRQ Publication No. 10(12)-EHC063-EF. Rockville, MD: Agency for Healthcare Research and Quality. April 2012.

As described below, despite its advantages for demonstrating internal validity of causal relationships, the RCT is not the best study design for all evidence questions. Like all methods, RCTs have limitations. RCTs can have particular limitations regarding external validity. The relevance or impact of these limitations varies according to the purposes and circumstances of study. In order to help inform health care decisions in real-world practice, evidence from RCTs and other experimental study designs should be augmented by evidence from other types of studies. These and related issues are described below.

RCTs can cost in the tens or hundreds of millions of dollars, and exceeding $1 billion in some instances. Costs can be particularly high for phase III trials of drugs and biologics conducted to gain market approval by regulatory agencies. Included are costs of usual care and the additional costs of conducting research. Usual care costs include those for, e.g., physician visits, hospital stays, laboratory tests, radiology procedures, and standard medications, which are typically covered by third-party payers. Research-only costs (which would not otherwise occur for usual care) include patient enrollment and related management; investigational technologies; additional tests and procedures done for research purposes; additional time by clinical investigators; data infrastructure, management, collection, analysis, and reporting; and regulatory compliance and reporting (DiMasi 2003; Morgan 2011; Roy 2012). Costs are higher for trials with large numbers of enrollees, large numbers of primary and secondary endpoints (requiring more data collection and analysis), and longer duration. Costs are generally high for trials that are designed to detect treatment effects that are anticipated to be small (therefore requiring large

sample sizes to achieve statistical significance) or that require extended follow-up to detect differences in, e.g., survival and certain health events.

> *A clinical trial is the best way to assess whether an intervention works, but it is arguably the worst way to assess who will benefit from it (Mant 1999).*

Most RCTs are designed to investigate the effects of a uniformly delivered intervention in a specific type of patient in specific circumstances. This helps to ensure that any observed difference in outcomes between the investigational treatment and comparator is less likely to be confounded by variations in the patient groups compared, the mode of delivering the intervention, other previous and current treatments, health care settings, and other factors. However, while this approach strengthens **internal validity**, it can weaken **external validity**.

Patients who enroll in an RCT are typically subject to inclusion and exclusion criteria pertaining to, e.g., age, comorbidities, other risk factors, and previous and current treatments. These criteria tend to yield homogeneous patient groups that may not represent the diversity of patients that would receive the interventions in real practice. 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. Findings from these RCTs may not be applicable to different practice settings for variations in the technique of delivering the intervention.

When RCTs are conducted to generate sufficient evidence for gaining market approval or clearance, they are sometimes known as "efficacy trials" in that they may establish only short-term **efficacy** (rather than **effectiveness**) and safety in a narrowly selected group of patients. Given the patient composition and the choice of comparator, results from these RCTs can overstate how well a technology works as well as under-represent the diversity of the population that will ultimately use the technology.

Given the high costs of RCTs and sponsors' incentives to generate findings, such as to gain market approval for regulated technologies, these trials may be too small (i.e., have insufficient statistical power) or too short in duration to detect rare or delayed outcomes, including adverse events, and other unintended impacts. On the other hand, even in large, long-term RCTs (as well as other large studies), an observed statistically significant difference in adverse events may arise from random error, or these events may simply happen to co-occur with the intervention rather than being caused by it (Rawlins 2008). As such, the results from RCTs may be misleading or insufficiently informative for clinicians, patients, and payers who make decisions pertaining to more heterogeneous patients and care settings.

Given their resource constraints and use to gain market approval for regulated technologies, RCTs may be designed to focus on a small number of outcomes, especially shorter-term **intermediate endpoints** or **surrogate endpoints** rather than ultimate endpoints such as mortality, morbidity, or quality of life. As such, findings from these RCTs may be of limited use to clinicians and patients. Of course, the use of validated surrogate endpoints is appropriate in many instances, including when the health impact of interventions for some health care conditions will not be realized for years or decades, e.g., screening for certain cancers, prevention of risky health behaviors, and management of hypertension and dyslipidemia to prevent strokes and myocardial infarction in certain patient groups.

RCTs are traditionally designed to test a **null hypothesis**, i.e., the assumption by investigators that there is no difference between intervention and control groups. This assumption often does not pertain for several reasons. Among these, the assumption may be unrealistic when findings of other trials

(including phase II trials for drugs and biologics) of the same technology have detected a treatment effect. Further, it is relevant only when the trial is designed to determine if one intervention is better than another, in contrast to whether they can be considered equivalent or one is inferior to the other (Rawlins 2008). Testing of an "honest" null hypothesis in an RCT is consistent with the principle of **equipoise**, which refers to a presumed state of uncertainty regarding whether any one of alternative health care interventions will confer more favorable outcomes, including balance of benefits and harms (Freedman 1987). However, there is controversy regarding whether this principle is realistic and even whether it is always ethical (Djulbegovic 2009; Fries 2004; Veatch 2007).

RCTs depend on principles of probability theory whose validity may be diminished in health care research, including certain aspects of the use of ***p*-values** and **multiplicity**, which refers to analyses of numerous endpoints in the same data set, stopping rules for RCTs that involve "multiple looks" at data emerging from the trial, and analysis of numerous subgroups. Each of these types of multiplicity involve iterative (repeated) tests of statistical significance based on conventional *p*-value thresholds (e.g., <0.05). Such iterative tests are increasingly likely to result in at least one false-positive finding, whether for an endpoint, a decision to stop a trial, or patient subgroup in which there appears to be a statistically significant treatment effect (Rawlins 2008; Wang 2007).

Using a *p*-value threshold (e.g., $p<0.01$ or $p<0.05$) as the basis for accepting a treatment effect can be misleading. There is still a chance (e.g., 1% or 5%) that the difference is due to random error. Also, a statistically significant difference detected with a large sample size may have no **clinical significance**. On the other hand, a finding of no **statistical significance** (e.g., $p\geq0.01$ or $p\geq0.05$) does not prove the absence of a treatment effect, including because the sample size of the RCT may have been too small to detect a true treatment effect. The reliance of most RCTs on *p*-values, particularly that the probability that a conclusion is in error can be determined from the data in a single trial, ignores evidence from other sources or the plausibility of the underlying cause-and-effect mechanism (Goodman 2008).

As noted below, other study designs are preferred for many types of evidence questions, even in some instances when the purpose is to determine the causal effect of a technology. For investigating technologies for treating rare diseases, the RCT may be impractical for enrolling and randomizing sufficient numbers of patients to achieve the statistical power to detect treatment effects. On the other hand, RCTs may be unnecessary for detecting very large treatment effects, especially where patient prognosis is well established and historical controls suffice.

To conduct an RCT may be judged unethical in some circumstances, such as when patients have a largely fatal condition for which no effective therapy exists. Use of a placebo control alone can be unethical when an effective standard of care exists and withholding it poses great health risk to patients, such as for HIV/AIDS prevention and therapy and certain cancer treatments. RCTs that are underpowered (i.e., with sample sizes too small to detect a true treatment effect or that yield statistically significant effects that are unreliable) can yield overestimated treatment effects and low reproducibility of results, thereby raising ethical concerns about wasted resources and patients' commitments (Button 2013).

E. Different Study Designs for Different Questions

RCTs are not the best study design for answering all evidence questions of potential relevance to an HTA. As noted in **Box III-11**, other study designs may be preferable for different questions. For example, the prognosis for a given disease or condition may be based on a follow-up studies of patient cohorts at uniform points in the clinical course of a disease. Case control studies, which are usually

Box III-11. RCTs Are Not the Best Study Design for All Evidence Questions

Other study designs may include the following (not a comprehensive list):

- Prevalence of a disease/disorder/trait? *Random sample survey* of relevant population
- Identification of risk factors for a disease/disorder/adverse event? *Case control study (for rare outcome) or cohort study (for more common outcome)*
- Prognosis? *Patient cohort studies* with follow-up at uniform points in clinical course of disease/disorder
- Accuracy and reliability of a diagnostic test? *Cross-over study* of index test (new test) vs. reference method ("gold standard") in cohort of patients at risk of having disease/disorder
- Accuracy and reliability of a screening test? *Cross-over study* of index test vs. reference method ("gold standard") in representative cross-section of asymptomatic population at-risk for trait/disorder/preclinical disease
- Efficacy/effectiveness (for health outcomes) of screening or diagnostic tests? *RCT* if time and resources allow; *observational studies and RCTs* rigorously linked for analytic validity, clinical validity, and clinical utility
- Efficacy/effectiveness (for health outcomes) of most therapies and preventive interventions? *RCT*
- Efficacy/effectiveness of interventions for otherwise fatal conditions? *Non-randomized trials, case series*
- Safety, effectiveness of incrementally modified technologies posing no known additional risk? *Registries*
- Safety, effectiveness of interventions in diverse populations in real-world settings? *Registries,* especially to complement findings of available *RCTs, PCTs*
- Rates of recall or procedures precipitated by false positive screening results? *Cohort studies*
- Complication rates from surgery, other procedures? *Registries, case series*
- Identification of a cause of a suspected iatrogenic (caused by a physician or therapy) disorder? *Case-control studies*
- Incidence of common adverse events potentially due to an intervention? *RCTs, nested case-control studies, n-of-1 trial for particular patients, surveillance, registries*
- Incidence of rare or delayed adverse events potentially due to an intervention? *Surveillance; registries; n-of-1 trial for particular patients; large, long-term RCT if feasible*

retrospective, are often used to identify risk factors for diseases, disorders, and adverse events. The accuracy of a new diagnostic test (though not its ultimate effect on health outcomes) may be determined by a cross-over study in which patients suspected of having a disease or disorder receive both the new ("index") test and the "gold standard" test. 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 or delayed adverse events that may be associated with an intervention. For incrementally modified technologies posing no known additional risk, registries may be appropriate for determining safety and effectiveness.

Although experimentation in the form of RCTs is regarded as the gold standard for deriving unbiased estimates of the causal effect of an intervention on health care outcomes, RCTs are not always necessary to reach the same convincing finding. Such instances arise when the size of the treatment effect is very large relative to the expected (well-established and predictable) prognosis of the disease and when this effect occurs quickly relative to the natural course of the disease, as may be discerned using historical controlled trials and certain well-designed case series and non-randomized cohort studies. Some examples include ether for anesthesia, insulin for diabetes, blood transfusion for severe hemorrhagic shock, penicillin for lobar pneumonia, ganciclovir for cytomegalovirus, imiglucerase for Gaucher's disease, phototherapy for skin tuberculosis, and laser therapy for port wine stains (Glasziou 2007; Rawlins 2008).

F. Complementary Methods for Internal and External Validity

> *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 (Eisenberg 1999).*

Given the range of impacts evaluated in HTA and its role in serving decision makers and policymakers with diverse responsibilities, HTA must consider the methodological validity and other attributes of various primary data methods. There is increasing recognition of the need for evidence generated by primary data methods with complementary attributes.

Although primary study investigators and assessors would prefer to have methods that achieve both internal and external validity, they often find that study design attributes that increase one type of validity jeopardize the other. As described above, a well-designed and conducted RCT is widely considered to be the best approach for ensuring internal validity. However, for the reasons that an RCT may have high internal validity, its external validity may be limited.

Findings of some large **observational studies** (e.g., from large cohort 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** and **confounding** that threaten internal validity of any observed relationship between an intervention (or other exposure) and outcomes. These studies are subject, for example, to **selection bias** on the part of patients, who have self-selected or otherwise influenced choice of an intervention, and investigators, who select which populations to study and compare. They are also subject to investigator **detection bias**. Interesting or promising findings from observational studies can generate hypotheses that can be tested using study designs with greater internal validity.

It is often not practical to conduct RCTs in all of the patient populations that might benefit from a particular intervention. Combinations of studies that, as a group, address internal validity and external validity may suffice. For example, RCTs demonstrating the safety and efficacy in a narrowly defined patient population can be complemented with continued follow-up of the original patient groups in those trials and by observational studies following more diverse groups of patients over time. These observational studies might include registries of larger numbers of more diverse patients who receive the intervention in various health care settings, studies of insurance claims data for patients with the relevant disease and intervention codes, studies using medical records, and postmarketing surveillance for adverse events in patients who received the intervention. Further, the RCT and observational data can provide inputs to computer-based simulations of the safety, effectiveness, and costs of using the intervention in various patient populations.

The methodological literature often contends that, due to their inherent lack of rigor, observational studies tend to report larger treatment effects than RCTs. However, certain well-designed observational studies can yield results that are similar to RCTs. An analysis published in 2000 that compared treatment effects reported from RCTs to those reported from observational studies for 19 treatments between 1985 and 1998 found that the estimates of treatment effects were similar for a large majority of the treatments (Benson 2000). Similarly, a comparison of the results of meta-analyses of RCTs and meta-analyses of observational studies (cohort or case control designs) for the same five clinical topics

published between 1991 and 1995 found that the reported treatment effects (including point estimates and 95% confidence intervals) were similar (Concato 2000).

Similar to quality assessment tools for various types of studies, the GRACE (Good ReseArch for Comparative Effectiveness) principles were developed to evaluate the methodological quality of observational research studies of comparative effectiveness. The GRACE principles comprise a series of questions to guide the evaluation, including what belongs in a study plan, key elements for good conduct and reporting, and ways to assess the accuracy of comparative effectiveness inferences for a population of interest. Given the range of types of potentially relevant evidence and the need to weigh applicability for particular circumstances of routine care, GRACE has no scoring system (Dreyer 2010). The accompanying GRACE checklist is used to assess the quality and usefulness for decision making of observational studies of comparative effectiveness (Dreyer 2014).

G. Evidence Hierarchies

> *So should we assess evidence the way Michelin guides assess hotels and restaurants? (Glasziou 2004).*

Researchers often use evidence hierarchies or other frameworks to portray the relative quality of various study designs for the purposes of evaluating single studies as well as a body of evidence comprising multiple studies. An example of a basic evidence hierarchy is:

- Systematic reviews and meta-analyses of RCTs
- Randomized controlled trials (RCTs)
- Non-randomized controlled trials
- Prospective observational studies
- Retrospective observational studies
- Expert opinion

In this instance, as is common in such hierarchies, the top item is a systematic review of RCTs, an integrative method that pools data or results from multiple single studies. (Hierarchies for single primary data studies typically have RCTs at the top.) Also, the bottom item, expert opinion, does not comprise evidence as such, though it may reflect the judgment of one or more people drawing on their perceptions of scientific evidence, personal experience, and other subjective input. There are many versions of such hierarchies, including some with more extensive levels/breakdowns.

> *Hierarchies cannot, moreover, accommodate evidence that relies on combining the results from RCTs and observational studies (Rawlins 2008).*

As noted earlier in this chapter, although the general type or name of a study design (e.g., RCT, prospective cohort study, case series) conveys certain attributes about the quality of a study, the study design name itself is not a good proxy for study quality. One of the weaknesses of these conventional one-dimensional evidence hierarchies is that, while they tend to reflect internal validity, they do not generally reflect external validity of the evidence to more diverse patients and care settings. Depending on the intended use of the findings of a single study or of a body of evidence, an assessment of internal validity may be insufficient. Such hierarchies do not lend themselves to characterizing the quality of a

body of diverse, complementary evidence that may yield fuller understanding about how well an intervention works across a heterogeneous population in different real-world circumstances. **Box III-12** lists these and other limitations of conventional evidence hierarchies.

Box III-12. Limitations of Conventional Evidence Hierarchies

- Originally developed for pharmacological models of therapy
- Poor design and implementation of high-ranking study designs may yield less valid findings than lower-ranking, though better designed and implemented, study types
- Emphasis on experimental control, while enhancing internal validity, can jeopardize external validity
- Cannot accommodate evidence that relies on considering or combining results from multiple study designs
- Though intended to address internal validity of causal effect of an intervention on outcomes, they have been misapplied to questions about diagnostic accuracy, prognosis, or adverse events
- Number and inconsistencies among (60+) existing hierarchies suggest shortcomings, e.g.,
  - ranking of meta-analyses relative to RCTs
  - ranking of different observational studies
  - terminology ("cohort studies," "quasi-experimental," etc.)

Sources: See, e.g.:

Glasziou P, et al. Assessing the quality of research. BMJ. 2004;328:39-41.
Rawlins MD. On the evidence for decisions about the use of therapeutic interventions. The Harveian Oration of 2008. London: Royal College of Physicians, 2008.
Walach H, et al. Circular instead of hierarchical: methodological principles for the evaluation of complex interventions. BMC Med Res Methodol. 2006;24;6:29.

Box III-13 shows an evidence framework from the Oxford Centre for Evidence-Based Medicine that defines five levels of evidence for each of several types of evidence questions pertaining to disease prevalence, screening tests, diagnostic accuracy, therapeutic benefits, and therapeutic harms. The lowest level of evidence for several of these evidence questions, "Mechanism-based reasoning," refers to some plausible scientific basis, e.g., biological, chemical, or mechanical, for the impact of an intervention. Although the framework is still one-dimensional for each type of evidence question, it does allow for moving up or down a level based on study attributes other than the basic study design.

While retaining the importance of weighing the respective methodological strengths and limitations of various study designs, extending beyond rigid one-dimensional evidence hierarchies to more useful evidence appraisal (Glasziou 2004; Howick 2009; Walach 2006) recognizes that:

- Appraising evidence quality must extend beyond categorizing study designs
- Different types of research questions call for different study designs
- It is more important for 'direct' evidence to demonstrate that the effect size is greater than the combined influence of plausible confounders than it is for the study to be experimental.
- Best scientific evidence–for a pragmatic estimate of effectiveness and safety–may derive from a complementary set of methods
  - They can offset respective weaknesses/vulnerabilities
  - "Triangulating" findings achieved with one method by replicating it with other methods may provide a more powerful and comprehensive approach than the prevailing RCT approach
- Systematic reviews are necessary, no matter the research type

Box III-13. Oxford Centre for Evidence-Based Medicine 2011 Levels of Evidence

Question	Step 1 (Level 1*)	Step 2 (Level 2*)	Step 3 (Level 3*)	Step 4 (Level 4*)	Step 5 (Level 5)
How common is the problem?	Local and current random sample surveys (or censuses)	Systematic review of surveys that allow matching to local circumstances**	Local non-random sample**	Case-series**	n/a
Is this diagnostic or monitoring test accurate? (Diagnosis)	Systematic review of cross sectional studies with consistently applied reference standard and blinding	Individual cross sectional studies with consistently applied reference standard and blinding	Non-consecutive studies, or studies without consistently applied reference standards**	Case-control studies, or "poor or non-independent reference standard**	Mechanism-based reasoning
What will happen if we do not add a therapy? (Prognosis)	Systematic review of inception cohort studies	Inception cohort studies	Cohort study or control arm of randomized trial*	Case-series or case-control studies, or poor quality prognostic cohort study**	n/a
Does this intervention help? (Treatment Benefits)	Systematic review of randomized trials or *n*-of-1 trials	Randomized trial or observational study with dramatic effect	Non-randomized controlled cohort/follow-up study**	Case-series, case-control studies, or historically controlled studies**	Mechanism-based reasoning
What are the COMMON harms? (Treatment Harms)	Systematic review of randomized trials, systematic review of nested case-control studies, *n*-of-1 trial with the patient you are raising the question about, or observational study with dramatic effect	Individual randomized trial or (exceptionally) observational study with dramatic effect	Non-randomized controlled cohort/follow-up study (post-marketing surveillance) provided there are sufficient numbers to rule out a common harm. (For long-term harms the duration of follow-up must be sufficient.)**	Case-series, case-control, or historically controlled studies**	Mechanism-based reasoning
What are the RARE harms? (Treatment Harms)	Systematic review of randomized trials or *n*-of-1 trial	Randomized trial or (exceptionally) observational study with dramatic effect			
Is this (early detection) test worthwhile? (Screening)	Systematic review of randomized trials	Randomized trial	Non -randomized controlled cohort/follow-up study**	Case-series, case-control, or historically controlled studies**	Mechanism-based reasoning

*** Level may be graded down on the basis of study quality, imprecision, indirectness (study PICO does not match questions PICO), because of inconsistency between studies, or because the absolute effect size is very small; Level may be graded up if there is a large or very large effect size.**

**** As always, a systematic review is generally better than an individual study.**

Source: OCEBM Levels of Evidence Working Group. The Oxford 2011 Levels of Evidence. Oxford Centre for Evidence-Based Medicine. http://www.cebm.net/index.aspx?o=5653

H. Alternative and Emerging Study Designs Relevant to HTA

Primary data collection methods are evolving in ways that affect the body of evidence used in HTA. Of great significance is the recognition that clinical trials conducted for biomedical research or to gain market approval or clearance by regulatory agencies do not necessarily address the needs of decision makers or policymakers.

Comparative effectiveness research (CER) reflects the demand for real-world evidence to support practical decisions. It emphasizes evidence from direct ("head-to-head") comparisons, effectiveness in real-world health care settings, health care outcomes (as opposed to surrogate or other intermediate endpoints), and ability to identify different treatment effects in patient subgroups. As traditional RCTs typically do not address this set of attributes, CER can draw on a variety of complementary study designs and analytical methods. Other important trends in support of CER are the gradual increase in use of electronic health records and more powerful computing and related health information technology, which enable more rapid and sophisticated analyses, especially of observational data. The demand for evidence on potentially different treatment effects in patient subgroups calls for study designs, whether in clinical trials or observational studies, that can efficiently discern such differences. Another powerful factor influencing primary data collection is the steeply increasing costs of conducting clinical trials, particularly of RCTs for new drugs, biologics, and medical devices; this focuses attention on study designs that require fewer patients, streamline data collection, and are of shorter duration.

Investigators continue to make progress in combining some of the desirable attributes of RCTs and observational studies. Some of the newer or still evolving clinical trial designs include: large simple trials, pragmatic clinical trials, cluster trials, adaptive trials, Bayesian trials, enrichment trials, and clinical registry trials (Lauer 2012), as described below.

Large simple trials (LSTs) retain the methodological strengths of prospective, randomized design, but use large numbers of patients, more flexible patient entry criteria and multiple study sites to generate effectiveness data and improve external validity. Fewer types of data may be collected for each patient in an LST, easing participation by patients and clinicians (Buring 1994; Ellenberg 1992; Peto 1995; Yusuf 1990). Prominent examples of LSTs include the GISSI trials of thrombolytic treatment of acute myocardial infarction (AMI) (Maggioni 1990), the ISIS trials of alternative therapies for suspected AMI (Fourth International Study of Infarct Survival 1991), and the CATIE trial of therapies for schizophrenia (Stroup 2003).

Pragmatic (or practical) clinical trials (PCTs) are a related group of trial designs whose main attributes include: comparison of clinically relevant alternative interventions, a diverse population of study participants, participants recruited from heterogeneous practice settings, and data collection on a broad range of health outcomes. PCTs 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). Some LSTs are also PCTs.

Cluster randomized trials involve randomized assignment of interventions at the level of natural groups or organizations rather than at the level of patients or other individuals. That is, sets of clinics, hospitals, nursing homes, schools, communities, or geographic regions are randomized to receive interventions or comparators. Such designs are used when it is not feasible to randomize individuals or when an intervention is designed to be delivered at a group or social level, such as a workplace-based smoking

cessation campaign or a health care financing mechanism. These are also known as "group," "place," or "community" randomized trials (Eldridge 2008).

Adaptive clinical trials use accumulating data to determine how to modify the design of ongoing trials according to a pre-specified plan. Intended to increase the quality, speed, and efficiency of trials, adaptive trials typically involve interim analyses, changes to sample size, changes in randomization to treatment arms and control groups, and changes in dosage or regimen of a drug or other technology (FDA Adaptive Design 2010; van der Graaf 2012).

A current example of an adaptive clinical trial is the I-SPY 2 (Investigation of Serial Studies to Predict Your Therapeutic Response with Imaging and Molecular Analysis 2), which is investigating multiple drug combinations and accompanying biomarkers for treating locally advanced breast cancer. In this adaptive trial, investigators calculate the probability that each newly enrolled patient will respond to a particular investigational drug combination based on how previous patients in the trial with similar genetic "signatures" (i.e., set of genetic markers) in their tumors have responded. Each new patient is then assigned to the indicated treatment regimen, accordingly, with an 80% chance of receiving standard chemotherapy plus an investigational drug and a 20% chance of receiving standard chemotherapy alone (Barker 2009; Printz 2013).

Bayesian clinical trials are a form of adaptive trials that rely on principles of Bayesian statistics. Rather than waiting until full enrollment and completion of follow-up for all enrolled patients, a Bayesian trial allows for assessment of results during the course of the trial and modifying its design to arrive at results more efficiently. Such midcourse modifications may include, e.g., changing the ratio of randomization to treatment arms (e.g., two patients randomized to the investigational group for every one patient randomized to the control group) to favor what appear to be more effective therapies, adding or eliminating treatment arms, changing enrollee characteristics to focus on patient subgroups that appear to be better responders, changing hypotheses from non-inferiority to superiority or vice-versa, and slowing or stopping patient accrual as certainty increases about treatment effects. These trial modifications also can accumulate and make use of information about relationships between biomarkers and patient outcomes (e.g., for enrichment, as described below). These designs enable more efficient allocation of patients to treatment arms, with the potential for smaller trials and for patients to receive better treatment (Berry 2006). Recent advances in computational algorithms and high-speed computing enable the calculations required for the complex design and simulations involved in planning and conducting Bayesian trials (FDA Guidance for the Use of Bayesian 2010; Lee 2012).

"Enrichment" refers to techniques of identifying patients for enrollment in clinical trials based on prospective use of patient attributes that are intended to increase the likelihood of detecting a treatment effect (if one truly exists) compared to an unselected population. Such techniques can decrease the number of patients needed to enroll in a trial; further, they can decrease patient heterogeneity of response, select for patients more likely to experience a disease-related trial endpoint, or select for patients (based on a known predictive biomarker) more likely to respond to a treatment (intended to result in a larger effect size). In adaptive enrichment of clinical trials, investigators seek to discern predictive markers during the course of a trial and apply these to enrich subsequent patient enrollment in the trial (FDA 2012). While these techniques improve the likelihood of discerning treatment effects in highly-selected patient groups, the findings of such trials may lack external validity to more heterogeneous patients. In one form of enrichment, the **randomized-withdrawal trial**, patients who respond favorably to an investigational intervention are then randomized to continue receiving that intervention or placebo. The study endpoints are return of symptoms or the ability to continue

participation in the trial. The patients receiving the investigational intervention continue to do so only if they respond favorably, while those receiving placebo continue to do only until their symptoms return. This trial design is intended to minimize the time that patients receive placebo (IOM Committee on Strategies for Small-Number-Participant Clinical Research Trials 2001; Temple 1996).

Clinical registry trials are a type of multicenter trial design using existing online registries as an efficient platform to conduct patient assignment to treatment and control groups, maintain case records, and conduct follow-up. Trials of this type that also randomize patient assignment to treatment and control groups are randomized clinical registry trials (Ferguson 2003; Fröbert 2010).

N-of-1 trials are clinical trials in which a single patient is the total population for the trial and in which a sequence of experimental and control interventions are allocated to the patient (i.e., a multiple crossover study conducted in a single patient). A trial in which random allocation is used to determine the sequence of interventions is given to a patient is an N-of-1 RCT. N-of-1 trials are used to determine treatment effects in individuals, and sets of these trials can be used to estimate **heterogeneity of treatment effects** across a population (Gabler 2011).

Patient preference trials are intended to account for patient preferences in the design of RCTs, including their ability to discern the impact of patient preference on health outcomes. Among the challenges to patient enrollment and participation in traditional RCTs are that some patients who have a strong preference for a particular treatment may decline to proceed with the trial or drop out early if they are not assigned to their preferred treatment. Also, these patients may experience or report worse or better outcomes due to their expectations or perceptions of the effects of assignment to their non-preferred or preferred treatment groups. Any of these actions may bias the results of the trial. Patient preference trials enable patients to express their preferred treatment prior to enrolling in an RCT. In some of these trials, the patients with a strong preference, e.g., for a new treatment or usual care, are assigned to a parallel group receiving their preferred intervention. The patients who are indifferent to receiving the new treatment or usual care are randomized into one group or another. Outcomes for the parallel, non-randomized groups (new intervention and usual care) are analyzed apart from the outcomes for the randomized groups.

In addition to enabling patients with strong preferences to receive their preferred treatment and providing for comparison of randomized groups of patients who expressed no strong preference, these trials may provide some insights about the relative impact on outcomes of receiving one's preferred treatment. However, this design is subject to selection bias, as there may be systematic differences in prognostic factors and other attributes between patients with a strong preference for the new treatment and patients with strong preferences for usual care. Selection bias can also affect the indifferent patients who are randomized, as there may be systematic differences in prognostic factors and other attributes between indifferent patients and the general population, thereby diminishing the external validity of the findings. To the extent that patients with preferences are not randomized, the time and cost required to enroll a sufficient number of patients for the RCT to achieve statistical power will be greater. Patient preference trials have alternative designs, e.g., partially randomized preference trials and fully randomized preference trials. In the fully randomized preference design, patient preferences are recorded prior to the RCT, but all patients then randomized regardless of their preference. In that design, subgroup analyses enable determining whether receiving one's preferred treatment has any impact on treatment adherence, drop-outs, and outcomes (Howard 2006; Mills 2011; Preference Collaborative Review Group 2008; Silverman 1996; Torgerson 1998).

I. 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 large and evolving literature on these subjects (Friedman 2010; Piantadosi 2005; Spilker 1991). Also, there is a literature on priority setting and efficient resource allocation for clinical trials, and cost-effective design of clinical trials (Antman 2012; Chilcott 2003; Claxton 1996; Detsky 1990; FDA Adaptive Design 2010).

As noted above, the process of compiling evidence for an assessment may call attention to the need for new primary data. An assessment program may determine that existing evidence is insufficient for informing 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 National Institutes of Health (NIH) such as the National Cancer Institute (NCI); the National Heart, Lung, and Blood Institute (NHLBI); and the National Institute of Allergy and Infectious Diseases (NIAID) sponsor and conduct biomedical research, including clinical trials. The Department of Veterans Affairs (VA) Cooperative Studies Program is responsible for the planning and conduct of large multicenter clinical trials and epidemiological studies within the VA. This program also works with the VA Health Economics Resource Center to perform economic analyses as part of its clinical trials. The Food and Drug Administration (FDA) does not typically conduct primary studies related to the marketing of new drugs and devices; rather, it reviews primary data from studies sponsored or conducted by the companies that make these technologies. The FDA also maintains postmarketing surveillance programs, including the FDA Adverse Event Reporting System on adverse events and medication error reports for drug and therapeutic biologic products, and the MedWatch program, in which physicians and other health professionals and the public voluntarily report serious reactions and other problems with drugs, devices, and other medical products.

In the US, the Patient-Centered Outcomes Research Institute (PCORI) was established as an independent research institute by Congress in the Patient Protection and Affordable Care Act of 2010. PCORI conducts CER and related research that is guided by patients, caregivers and the broader health care community. PCORI's five national research priorities are: assessment of prevention, diagnosis, and treatment options; improving health care systems; enhancing communication and dissemination of evidence; addressing disparities in health and health care; and improving CER methods and data infrastructure. PCORI devotes more than 60% of its research budget to CER, including for pragmatic clinical trials, large simple trials, and large observational studies, with the balance allocated to infrastructure, methods, and communication and dissemination research (Selby 2014).

Third-party payers generally do not sponsor clinical trials. However, they have long supported clinical trials of new technologies indirectly by paying for care associated with trials of those technologies, or by paying unintentionally for non-covered new procedures that were coded as covered procedures. As noted above, payers provide various forms of conditional coverage, such as coverage with evidence development (CED), for certain investigational technologies in selected settings to compile evidence that can be used to make more informed coverage decisions. Two main types of CED are "only in research," in which coverage of a technology is provided only for patients with specified clinical indications in the payer's beneficiary population who are enrolled in a clinical trial of that technology, and "only with research," in which coverage of a technology is provided for all of the patients with specified clinical indications if a subset of those patients is enrolled in a clinical trial of that technology.

An early example of CED was the multicenter RCT of lung-volume reduction surgery, the National Emphysema Treatment Trial (NETT) conducted in the US, funded by the NHLBI and the Centers for Medicare and Medicaid Services (CMS, which administers the US Medicare program) (Fishman 2003; Ramsey 2003). In another form of conditional coverage known as conditional treatment continuation, payment is provided only as long as patients meet short-term treatment goals such as lowered blood cholesterol or cancer tumor response. In performance-linked reimbursement (or "pay-for-performance"), payment for a technology is linked to data demonstrating achievement of pre-specified clinical outcomes in practice; this includes schemes in which a manufacturer must provide rebates, refunds, or price adjustments to payers if their products do not achieve certain patient outcomes (Carlson 2010). Findings about the impact of conditional coverage, performance-linked reimbursement, and related efforts on coverage policies, patient outcomes, and costs are still emerging (de Bruin 2011).

Payers and researchers often analyze data from claims, electronic health records, registries, and surveys to determine comparative effectiveness of interventions, develop coverage policies, or determine provider compliance with coverage policies. These analyses increasingly involve efforts to link claims and other administrative sources to electronic health records and other clinical sources (Croghan 2010; de Souza 2012).

The ability of most assessment programs to undertake new primary data collection, particularly clinical trials, is limited by such factors as programs' remit (which may not include sponsoring primary data collection), financial constraints, time constraints, 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.

References for Chapter III

Antman EM, Harrington RA. Transforming clinical trials in cardiovascular disease: mission critical for health and economic well-being. JAMA. 2012;308(17):1743-4.

Atkins D, Best D, Briss PA, Eccles M, et al., GRADE Working Group. Grading quality of evidence and strength of recommendations. BMJ. 2004;328(7454):1490. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC428525.

Atkins D, Chang S, Gartlehner G, Buckley DI, et al. Chapter 6. Assessing the Applicability of Studies When Comparing Medical Interventions. In: Methods Guide for Effectiveness and Comparative Effectiveness Reviews. AHRQ Publication No. 10(13)-EHC063-EF. Rockville, MD: Agency for Healthcare Research and Quality. September 2013. Accessed Nov. 1, 2013 at: http://effectivehealthcare.ahrq.gov/ehc/products/60/318/CER-methods-guide-130916.pdf.

Barker AD, Sigman CC, Kelloff GJ, Hylton NM, et al. I-SPY 2: an adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. Clin Pharmacol Ther. 2009;86(1):97-100.

Benson K, Hartz AJ. A comparison of observation studies and randomized, controlled trials. N Engl J Med. 2000;342(25):1878-86. http://www.nejm.org/doi/full/10.1056/NEJM200006223422506.

Berry DA. Bayesian clinical trials. Nat Rev Drug Discov. 2006;5(1):27-36.

Boer GJ, Widner H. Clinical neurotransplantation: core assessment protocol rather than sham surgery as control. Brain Res Bull. 2002;58(6):547-53.

Briss PA, Zaza S, Pappaioanou M, Fielding J, et al. Developing an evidence-based Guide to Community Preventive Services--Am J Prev Med 2000;18(1S):35-43.

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;67-91. BP-H-142. Washington, DC: US Government Printing Office; 1994. Accessed Nov. 1, 2013 at: http://ota-cdn.fas.org/reports/9440.pdf.

Button KS, Ioannidis JP, Mokrysz C, Nosek BA, et al. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci. 2013;14(5):365-76.

Carlson JJ, Sullivan SD, Garrison LP, Neumann PJ, Veenstra DL. Linking payment to health outcomes: a taxonomy and examination of performance-based reimbursement schemes between healthcare payers and manufacturers. Health Policy. 2010;96(3):179-90.

Campbell MK, Entwistle VA, Cuthbertson BH, Skea ZC, et al.; KORAL study group. Developing a placebo-controlled trial in surgery: issues of design, acceptability and feasibility. Trials. 2011;12:50. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3052178.

Chilcott J, Brennan A, Booth A, Karnon J, Tappenden P. The role of modelling in prioritising and planning clinical trials. Health Technol Assess. 2003;7(23):iii,1-125. http://www.journalslibrary.nihr.ac.uk/__data/assets/pdf_file/0006/64950/FullReport-hta7230.pdf.

Claxton K, Posnett J. An economic approach to clinical trial design and research priority-setting. Health Econ. 1996;5(6):513-24.

ClinicalTrials.gov. FDAAA 801 Requirements. December 2012. Accessed Aug. 1, 2013 at: http://clinicaltrials.gov/ct2/manage-recs/fdaaa#WhichTrialsMustBeRegistered.

Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med. 2000;342:1887-92. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1557642.

Croghan TW, Esposito D, Daniel G, Wahl P, Stoto MA. Using medical records to supplement a claims-based comparative effectiveness analysis of antidepressants. Pharmacoepidemiol Drug Saf. 2010;19(8):814-8.

de Bruin SR, Baan CA, Struijs JN. Pay-for-performance in disease management: a systematic review of the literature. BMC Health Serv Res. 2011;11:272. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3218039.

de Souza JA, Polite B, Perkins M, Meropol NJ, et al. Unsupported off-label chemotherapy in metastatic colon cancer. BMC Health Serv Res. 2012;12:481. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3544564.

Detsky AS. Using cost-effectiveness analysis to improve the efficiency of allocating funds to clinical trials. Stat Med. 1990;9(1-2):173-84.

DiMasi JA, Hansen RW, Grabowski HG. The price of innovation: new estimates of drug development costs. J Health Econ. 2003;22(2):151–85.

Djulbegovic B. The paradox of equipoise: the principle that drives and limits therapeutic discoveries in clinical research. Cancer Control. 2009;16(4):342-7. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2782889.

Dreyer NA, Schneeweiss S, McNeil BJ, Berger ML, et al. GRACE principles: recognizing high-quality observational studies of comparative effectiveness. Am J Manag Care. 2010;16(6):467-71.

Dreyer NA, Velentgas P, Westrich K, Dubois R. The GRACE Checklist for Rating the Quality of Observational Studies of Comparative Effectiveness: A Tale of Hope and Caution. J Manag Care Pharm. 2014;20(3):301-8.

Eisenberg JM. Ten lessons for evidence-based technology assessment. JAMA. 1999; 282(19):1865-9.

Eldridge S, Ashby D, Bennett C, et al. Internal and external validity of cluster randomised trials: systematic review of recent trials. BMJ. 2008;336(7649):876-80. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2323095.

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

Enck P, Bingel U, Schedlowski M, Rief W. The placebo response in medicine: minimize, maximize or personalize? Nat Rev Drug Discov. 2013;12(3):191-204.

Ferguson TB Jr, Peterson ED, Coombs LP, Eiken MC, 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(1):49-56.

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. http://www.nejm.org/doi/full/10.1056/NEJMoa030287.

Food and Drug Administration. Adaptive Design Clinical Trials for Drugs and Biologics. Draft Guidance. Center for Drug Evaluation and Research, Center for Biologics Evaluation and Research. Rockville, MD, February 2010. Accessed Nov. 1, 2013 at: http://www.fda.gov/downloads/Drugs/.../Guidances/ucm201790.pdf.

Food and Drug Administration. Guidance for Industry. Enrichment Strategies for Clinical Trials to Support Approval of Human Drugs and Biological Products. Draft Guidance. Center for Drug Evaluation and Research, Center for Biologics Evaluation and Research, Center for Devices and Radiological Health. Rockville, MD, December 2012. Accessed Nov. 1, 2013 at: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM332181.pdf.

Food and Drug Administration. Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials. Center for Devices and Radiological Health, Center for Biologics Evaluation and Research. Rockville, MD, February 5, 2010. Accessed Nov. 1, 2013 at: http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071121.pdf/

Fourth International Study of Infarct Survival: protocol for a large simple study of the effects of oral mononitrate, of oral capvtopril, and of intravenous magnesium. ISIS-4 collaborative group. . Am J Cardiol. 1991;68(14):87D-100D.

Freedman B. Equipoise and the ethics of clinical research. N Engl J Med. 1987;317(3):141-5.

Friedman LM. Furberg CD, DeMets DL. Fundamentals of Clinical Trials. (4th edition). New York: Springer, 2010.

Fries JF, Krishnan E. Equipoise, design bias, and randomized controlled trials: the elusive ethics of new drug development. Arthritis Res Ther. 2004;6(3):R250-5. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC416446.

Fröbert O, Lagerqvist B, Gudnason T, Thuesen L, et al. Thrombus Aspiration in ST-Elevation myocardial infarction in Scandinavia (TASTE trial). A multicenter, prospective, randomized, controlled clinical registry trial based on the Swedish angiography and angioplasty registry (SCAAR) platform. Study design and rationale. Am Heart J. 2010;160(6):1042-8.

Frost J, Okun S, Vaughan T, Heywood J, Wicks P. Patient-reported outcomes as a source of evidence in off-label prescribing: analysis of data from PatientsLikeMe. J Med Internet Res. 2011 Jan 21;13(1):e6. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3221356.

Gabler NB, Duan N, Vohra S, Kravitz RL. N-of-1 trials in the medical literature: a systematic review. Med Care. 2011;49(8):761-8.

Glasziou P, Vandenbroucke J, Chalmers I. Assessing the quality of research. BMJ. 2004;328(7430):39-41. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC313908.

Goodman S. A dirty dozen: twelve p-value misconceptions. Semin Hematol. 2008;45(3):135-40.

Hartling L, Bond K, Harvey K, Santaguida PL, et al. Developing and Testing a Tool for the Classification of Study Designs in Systematic Reviews of Interventions and Exposures. Agency for Healthcare Research and Quality; December 2010. Methods Research Report. AHRQ Publication No. 11-EHC-007. http://www.ncbi.nlm.nih.gov/books/NBK52670/pdf/TOC.pdf.

Higgins JP, Altman DG, Gøtzsche PC, Jüni P, et al.; Cochrane Bias Methods Group; Cochrane Statistical Methods Group. The Cochrane Collaboration's tool for assessing risk of bias in randomized trials. BMJ. 2011;343:d5928. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3196245.

Higgins JPT, Altman DG, Sterne, JAC, eds. Chapter 8: Assessing risk of bias in included studies. In: Higgins JPT, Green S, eds. Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011.

Horng S, Miller FG. Ethical framework for the use of sham procedures in clinical trials. Crit Care Med. 2003;31(suppl. 3):S126-30.

Howard L, Thornicroft G. Patient preference randomised controlled trials in mental health research. Br J Psychiatry. 2006;188:303-4. http://bjp.rcpsych.org/content/188/4/303.long.

Howick J, Glasziou P, Aronson JK. The evolution of evidence hierarchies: what can Bradford Hill's 'guidelines for causation' contribute? J R Soc Med. 2009;102(5):186-94. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2677430.

Huser V, Cimino JJ. Evaluating adherence to the International Committee of Medical Journal Editors' policy of mandatory, timely clinical trial registration. J Am Med Inform Assoc. 2013;20(e1):e169-74.

ICMJE (International Committee of Medical Journal Editors). Uniform Requirements for Manuscripts Submitted to Biomedical Journals: Publishing and Editorial Issues Related to Publication in Biomedical Journals: Obligation to Register Clinical Trials. 2013. Accessed Aug. 1, 2013 at: http://www.icmje.org/publishing_10register.html.

Institute of Medicine. Committee on Strategies for Small-Number-Participant Clinical Research Trials. Small Clinical Trials: Issues and Challenges. Washington, DC: National Academies Press; 2001. http://www.nap.edu/openbook.php?record_id=10078&page=1.

Ioannidis JP, Khoury MJ. Are randomized trials obsolete or more important than ever in the genomic era? Genome Med. 2013;5(4):32. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3707036.

Jadad AR, Moore RA, Carrol D, et al. Assessing the quality of reports of randomized clinical trials: Is blinding necessary? Control Clin Trials. 1996;17:1-12.

Lauer MS. Commentary: How the debate about comparative effectiveness research should impact the future of clinical trials. Stat Med. 2012;31(25):3051-3.

Lee JJ, Chu CT. Bayesian clinical trials in action. Stat Med. 2012;31(25):2955-72.

Maggioni AP, Franzosi MG, Fresco C, et al. GISSI trials in acute myocardial infarction. Rationale, design, and results. Chest. 1990;97(4 Suppl):146S-150S.

Mant D. Evidence and primary care: can randomized trials inform clinical decisions about individual patients? Lancet. 1999;353:743–6.

Mills N, Donovan JL, Wade J, Hamdy FC, et al. Exploring treatment preferences facilitated recruitment to randomized controlled trials. J Clin Epidemiol. 2011;64(10):1127-36. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3167372.

Morgan S, Grootendorst P, Lexchin J, Cunningham C, Greyson D. The cost of drug development: a systematic review. Health Policy. 2011;100(1):4-17.

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(2):81-8. http://www.nejm.org/doi/full/10.1056/NEJMoa013259.

Nakamura C, Bromberg M, Bhargava S, Wicks P, Zeng-Treitler Q. Mining online social network data for biomedical research: a comparison of clinicians' and patients' perceptions about amyotrophic lateral sclerosis treatments. J Med Internet Res. 2012;14(3):e90. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3414854.

OCEBM Levels of Evidence Working Group. The Oxford 2011 Levels of Evidence. Oxford Centre for Evidence-Based Medicine. http://www.cebm.net/index.aspx?o=5653.

Olivo SA, Macedo LG, Gadotti IC, Fuentes J, et al. Scales to assess the quality of randomized controlled trials: a systematic review. Phys Ther. 2008;88(2):156-75. http://ptjournal.apta.org/content/88/2/156.long.

Peto R, Collins R, Gray R. Large-scale randomized evidence: large, simple trials and overviews of trials. J Clin Epidemiol. 1995;48(1):23-40.

Piantadosi S. Clinical Trials: A Methodological Perspective (2nd edition). New York: Wiley, 2005.

Preference Collaborative Review Group. Patients' preferences within randomised trials: systematic review and patient level meta-analysis. BMJ. 2008;337:a1864. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2659956.

Printz C. I-SPY 2 may change how clinical trials are conducted: Researchers aim to accelerate approvals of cancer drugs. Cancer. 2013;119(11):1925-7.

Rawlins MD. De Testimonio: On the evidence for decisions about the use of therapeutic interventions. The Harveian Oration of 2008. London: Royal College of Physicians, 2008.

Rothman KJ, Michels KB. The continuing unethical use of placebo controls. N Engl J Med. 1994;331(6):394-7.

Roundtree AK, Kallen MA, Lopez-Olivo MA, Kimmel B, et al. Poor reporting of search strategy and conflict of interest in over 250 narrative and systematic reviews of two biologic agents in arthritis: a systematic review. J Clin Epidemiol. 2009;62(2):128-37.

Roy ASA. Stifling New Cures: The True Cost of Lengthy Clinical Drug Trials. Project FDA Report 5. New York: Manhattan Institute for Policy Research; April 2012. http://www.manhattan-institute.org/html/fda_05.htm.

Selby JV, Lipstein SH. PCORI at 3 years--progress, lessons, and plans. N Engl J Med. 2014;370(7):592-5. http://www.nejm.org/doi/full/10.1056/NEJMp1313061.

Silverman WA, Altman DG. Patients' preferences and randomised trials. Lancet. 1996;347(8995):171-4.

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

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(10):1581-7.

Stroup TS, McEvoy JP, Swartz MS, et al. The National Institute of Mental Health Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) project: schizophrenia trial design and protocol development. Schizophr Bull. 2003;29(1):15-31. http://schizophreniabulletin.oxfordjournals.org/content/29/1/15.long.

Temple R. Problems in interpreting active control equivalence trials. Account Res. 1996;4(3-4):267-75.

Torgerson DJ, Sibbald B. Understanding controlled trials. What is a patient preference trial? BMJ. 1998;316(7128):360. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2665528.

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.

US Preventive Services Task Force Procedure Manual. AHRQ Publication No. 08-05118-EF, July 2008. Accessed Aug. 1, 2013 at: http://www.uspreventiveservicestaskforce.org/uspstf08/methods/procmanual.htm.

van der Graaf R, Roes KC, van Delden JJ. Adaptive trials in clinical research: scientific and ethical issues to consider. JAMA. 2012;307(22):2379-80.

Varmus H, Satcher D. Ethical complexities of conducting research in developing countries. N Engl J Med. 1997;337(14):1003-5.

Veatch RM. The irrelevance of equipoise. J Med Philos. 2007;32(2):167-83.

Viswanathan M, Ansari MT, Berkman ND, Chang S, et al. Chapter 9. Assessing the risk of bias of individual studies in systematic reviews of health care interventions. In: Methods Guide for Effectiveness and Comparative Effectiveness Reviews. AHRQ Publication No. 10(14)-EHC063-EF. Rockville, MD: Agency for Healthcare Research and Quality. January 2014. Accessed Feb. 1, 2014 at: http://www.effectivehealthcare.ahrq.gov/ehc/products/60/318/CER-Methods-Guide-140109.pdf.

Walach H, Falkenberg T, Fønnebø V, Lewith G, Jonas WB. Circular instead of hierarchical: methodological principles for the evaluation of complex interventions. BMC Med Res Methodol. 2006;24;6:29. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1540434.

Wang R, Lagakos WE, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine—reporting of subgroup analyses in clinical trials. N Engl J Med 2007;357(21):2189-94.

Whiting PF, Rutjes AW, Westwood ME, Mallett S, et al.; QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529-36.

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.

Zarin DA, Tse T, Williams RJ, Califf RM, Ide NC. The ClinicalTrials.gov results database--update and key issues. N Engl J Med. 2011;364(9):852-60. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3066456.

IV. INTEGRATIVE METHODS

Integrative methods (or 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.

Having considered the merits of individual studies, an assessment group must then integrate, synthesize, or consolidate the available relevant 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 them may be combined or considered in broader social and economic contexts in order to help inform policies.

Methods used to combine or integrate data from primary sources include the following:

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

Certain 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. Considerable work has been done to improve the validity of decision analysis and meta-analysis in particular over the last 25 years (see, e.g., Eckman 1992; Eddy 1992; Lau 1992). This was augmented by consensus-building approaches of 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 predecessor agency to AHRQ (the Agency for Health Care Policy and Research), and others.

Systematic reviews, meta-analyses, and certain types of modeling consolidate findings of existing relevant research in order to resolve inconsistencies or ambiguities among existing studies and yield findings that may not have been apparent or significant in individual studies. These study designs use predetermined criteria and systematic processes to search for, screen for inclusion and exclusion, and combine the findings of existing studies.

Although systematic reviews, meta-analyses, and modeling can produce new insights from existing evidence, they do not generate new data. Well-formulated inclusion and exclusion criteria can help to diminish various sources of bias that could be introduced by the primary data studies or the selection of these studies for the integrative studies.

The applicability of the findings of integrative methods is constrained by any limitations of their component primary studies with respect to, e.g., patient age groups, comorbidities, health care settings,

and selection of outcomes measures. Regardless of how rigorously a systematic review is conducted, its ability to determine the most effective intervention is limited by the scope and quality of the underlying evidence. Even studies that satisfy rigorous inclusion criteria may, as a group, reflect publication bias. Other factors may limit the external validity of the findings, such as narrowly defined study populations (e.g., with no comorbidities), inappropriate comparison therapies, insufficient duration of follow-up, and restriction to clinical settings with high levels of expertise and ancillary services that may not prevail in community practice. Often, the greatest value of a systematic review or meta-analysis is its ability to identify gaps in evidence that may be helpful in identifying the need for future primary studies.

Four major types of integrative methods, i.e., systematic literature reviews, meta-analysis, decision analysis, and consensus development, are described below.

A. Systematic Literature Reviews

A systematic literature review is a form of structured literature review that addresses one or more evidence questions (or key questions) that are formulated to be answered by analysis of evidence. Broadly, this involves:

- Objective means of searching the literature
- Applying predetermined inclusion and exclusion criteria to this literature
- Critically appraising the relevant literature
- Extraction and synthesis of data from evidence base to formulate answers to key questions

Depending on the purpose of the systematic review and the quality of the included studies, systematic reviews often include meta-analyses. A useful way to define the key questions used in a systematic review and to guide the literature search is the "PICOTS" format (see, e.g., Counsell 1997):

- **P**opulation: e.g., condition, disease severity/stage, comorbidities, risk factors, demographics
- **I**ntervention: e.g., technology type, regimen/dosage/frequency, technique/method of administration
- **C**omparator: e.g., placebo, usual/standard care, active control
- **O**utcomes: e.g., morbidity, mortality, quality of life, adverse events
- **T**iming: e.g., duration/intervals of follow-up
- **S**etting: e.g., primary, inpatient, specialty, home care

Not all evidence questions use all of these elements; some use PICO only.

The main steps of a systematic review include the following (see, e.g., Buckley 2014; Rew 2011; Shea 2007; Sutton 1998):

1. Specify purpose of the systematic review
2. Specify evidence questions. Use appropriate structured format, e.g., PICOTS
3. Specify review protocol that is explicit, unbiased, and reproducible, including:
    - Inclusion and exclusion criteria for studies to be reviewed, including type/status of publication (e.g., peer-reviewed publication vs. grey literature)

- Bibliographic databases (and other sources, if applicable) to be searched
- Search terms/logic for each database
- Methods of review (e.g., number of independent parallel reviewers of each study)
- Intention to conduct meta-analysis (if appropriate and feasible) and specification of methods to combine/pool data
- Register or publish protocol, as appropriate

4. Perform comprehensive literature search
    - Document all search sources and methods
5. Review search results and compare to inclusion/exclusion criteria
    - Account for included and excluded studies (e.g., using a flow diagram)
    - Identify and exclude duplicate studies, as appropriate
    - Compile and provide lists of included studies and excluded studies (with reasons for exclusion)
    - Assess potential sources of publication bias
6. Systematically extract data from each included study
    - Consistent with review protocol
    - Include PICOTS characteristics
    - Present extracted data in tabular form
7. Assess quality of individual studies retrieved/reviewed
    - Document quality for each study
    - Account for potential conflicts of interest
8. Perform meta-analysis (if specified in protocol and if methodologically feasible based on primary data characteristics)
9. Assess quality (or strength) of cumulative body of evidence
    - Assess risks of bias, directness or relevance of evidence (patients, interventions, outcomes, etc.) to the evidence questions, consistency of findings across available evidence, and precision in reporting results
    - Assign grade to cumulative body of evidence
10. Present results/findings
    - Link results/findings explicitly to evidence from included studies
    - Account for quality of the included studies
    - Present clearly to enable critical appraisal and replication of systematic review
11. Conduct sensitivity analysis of review results
    - Examine the impact on review results of inclusion/exclusion criteria, publication bias, and plausible variations in assumptions and estimates of outcomes and other parameters
    - Also conduct analyses (e.g., subgroup analyses and meta-regression) for better understanding of heterogeneity of effects
12. Describe limitations and actual/potential conflicts of interest and biases in the process
    - Account for body of included studies and the systematic review process
    - Describe evidence gaps and future research agenda, as appropriate
13. Disseminate (e.g., publish) results

Assessing the quality of individual studies is described in chapter III. Assessing the quality (or strength) of a cumulative body of evidence is described later in chapter IV. Dissemination of results is described in chapter VIII.

In conducting literature searches for systematic reviews, a more comprehensive and unbiased identification of relevant clinical trials and other studies (consistent with the inclusion criteria) may be realized by expanding the search beyond the major biomedical bibliographic databases such PubMed and Embase. Additional in-scope studies may be identified via specialized databases and clinical trial registries, reference lists, hand-searching of journals, conference abstracts, contacting authors and trials sponsors (e.g., life sciences companies) to find unpublished trials, and Internet search engines. The studies identified in these ways should remain subject to the quality criteria used for the systematic review. The extent to which an expanded search has an impact on the findings of the systematic review varies (Savoie 2003). Systematic reviews of particular types of technologies may use different sets of databases and synthesis approaches for particular types of technologies, such as for imaging procedures and diagnostic tests (Bayliss 2008; Whiting 2008).

One of the tools developed to assess the quality of systematic reviews, the Assessment of Multiple Systematic Reviews (AMSTAR), which was derived using nominal group technique and factor analysis of previous instruments, has the following 11 items (Shea 2007; Shea 2009):

1. Was an 'a priori' design provided?
2. Was there duplicate study selection and data extraction?
3. Was a comprehensive literature search performed?
4. Was the status of publication ([e.g.,] grey literature) used as an inclusion criterion?
5. Was a list of studies (included and excluded) provided?
6. Were the characteristics of the included studies provided?
7. Was the scientific quality of the included studies assessed and documented?
8. Was the scientific quality of the included studies used appropriately in formulating conclusions?
9. Were the methods used to combine the findings of studies appropriate?
10. Was the likelihood of publication bias assessed?
11. Was the conflict of interest stated?

In addition to those for assessing methodological quality of systematic reviews, there are instruments to assess the reporting of systematic reviews and meta-analyses, including the Preferred Reporting Items of Systematic reviews and Meta-Analyses instrument (PRISMA) (Moher 2009), as shown in **Box IV-1**. Various computer software packages are available to manage references and related bibliographic information for conducting systematic reviews; examples are EndNote, Reference Manager, and RefWorks (see, e.g., Hernandez 2008), though no particular recommendation is offered here.

Box IV-1. PRIMSA Checklist of Items to Include When Reporting a Systematic Review

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	

Source: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med. 2009;6(7):e1000097.

Box IV-1. PRIMSA Checklist of Items to Include When Reporting a Systematic Review (cont'd)

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	

Source: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med. 2009;6(7):e1000097.

A resource for minimizing publishing bias in systematic reviews is PROSPERO, an international database of prospectively registered systematic reviews in health and social care. The main objectives of PROSPERO are to reduce unplanned duplication of systematic reviews and provide transparency in the review process with the aim of minimizing reporting/publication bias. By providing a comprehensive listing of systematic reviews and their protocols at their inception, PROSPERO is intended to help counter publication bias by providing a permanent record of the original protocol of each systematic review, whether or not it is published. As such, comparison of this record to any reported findings of the systematic reviews can reveal any differences between the methods and outcomes of the registered protocol with those that are ultimately reported. Also, registration of reviews should diminish instances of duplication of effort. Established in 2011, PROSPERO is managed by the Centre for Reviews and Dissemination and funded by the UK National Institute for Health Research (Booth 2012).

B. Working with Best Evidence

In health care as well as other fields, there are tradeoffs between wanting to rely on the highest quality of evidence and the need to derive useful findings when evidence of the highest quality is limited or unavailable. For example:

> *In law, there is a principle that the same evidence that would be essential in one case might be disregarded in another because in the second case there is better evidence available…. Best-evidence synthesis extends this principle to the practice of research review. For example, if a literature contains several studies high in internal and external validity, then lower quality studies might be largely excluded from the review …. However, if a set of studies high in internal and external validity does not exist, we might cautiously examine the less well designed studies to see if there is adequate unbiased information to come to any conclusion (Slavin 1995).*

A desire to base health care decisions and policies on evidence generated from study designs that are of high quality for establishing internal validity of a causal relationship should not preclude using the best evidence that is available from other study designs. First, as described in detail in chapter III, evidence of internal validity should be complemented by evidence of external validity wherever appropriate and feasible to demonstrate that a technology works in real-world practice. Second, whether for internal validity or external validity, evidence from the highest quality study designs may not be available. For purposes of helping to inform clinical decisions and health care policies, it may be impractical to cease an evidence review because of the absence of high-quality evidence. The "**best evidence**" may be the *best available* evidence, i.e., the best evidence that is currently available and relevant for the evidence questions of interest (Ogilvie 2005).

"Best evidence" is *not* based on a single evidence hierarchy and it is not confined to internal validity. Even where traditional high-quality evidence with internal validity does exist (e.g., based on well-designed and conducted RCTs or meta-analyses of these), complementary evidence from other study designs (e.g., practical clinical trials, observational studies using registry data) may be needed to determine external validity. Where there is little or no high-quality evidence with internal validity, it may be necessary to pursue lower quality evidence for internal validity, such as non-randomized clinical trials, trials using historical controls, case series, or various types of observational studies, while documenting potential forms of bias that might accompany such evidence.

The need to seek lower-quality evidence in the absence of high-quality evidence also depends on the nature of the health problem and evidence question(s) of interest. For example, given a serious health

problem for which one or more existing technologies have been proven safe and effective based on high-quality evidence, the evidence required for a new technology should be based on high-quality evidence, as substitution of an existing proven technology by a new one with poorly established safety and uncertain effectiveness could pose unacceptable risks to patients who are experiencing good outcomes. In the instance of a rare, serious health problem for which no effective treatment exists, it may be difficult to conduct adequately powered RCTs, and lower-quality evidence suggesting a clinically significant health benefit, even with limited data on safety, may be acceptable as the best available evidence. Of course, appraising the evidence and assigning grades to any accompanying recommendations must remain objective and transparent. That is, just because an assessment must rely on the best available evidence does not necessarily mean that this evidence is high-quality (e.g., "Level I") evidence, or that recommendations based on it will be "Strong" or of "Grade A."

Inclusion and exclusion criteria for a systematic review should be informed by the evidence questions to be addressed as well as some knowledge about the types and amounts of evidence available, which can be determined from examining previous reviews and a preliminary literature search. To the extent that there appears to be a body of high-quality evidence with high internal and external validity, it may be unnecessary to pursue evidence of lower quality. However, in the absence of such evidence, it may be necessary to pursue lower-quality evidence (Lyles 2007; Ogilvie 2005).

C. Meta-Analysis

Meta-analysis refers to a group of statistical methods for combining (or "pooling") the data or results of multiple studies to obtain a quantitative estimate of the overall effect of a particular technology (or other 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). A meta-analyses is not the same as a systematic review, although many systematic reviews include meta-analyses, where doing so is methodologically feasible.

The purposes of meta-analysis include:

- 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 HTA 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 statistical

power to detect a true treatment effect if one exists, or at least narrow the confidence interval around the mean treatment effect.

The basic steps in meta-analysis are the following:

1. Specify the problem of interest.
2. Specify the criteria for inclusion and exclusion of studies (e.g., type and quality).
3. Identify and acquire 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.

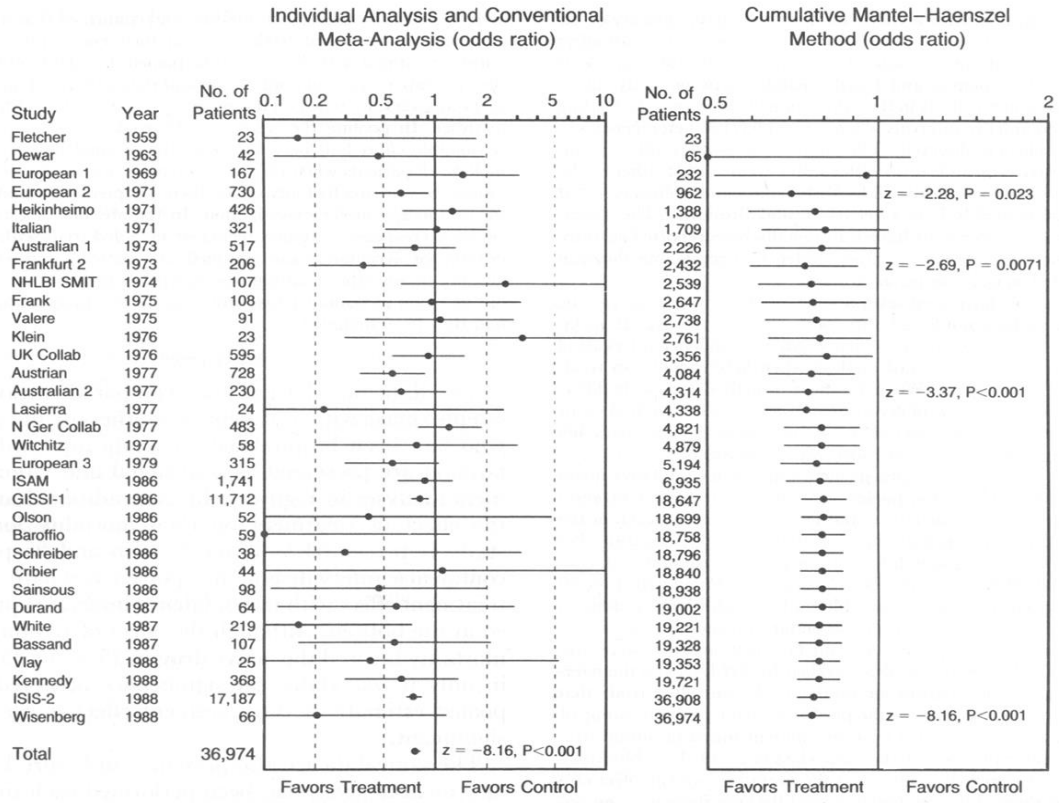
Meta-analysis can be limited by publication bias of the RCTs or other primary studies used, biased selection of available relevant studies, poor quality of the primary studies, unexplainable heterogeneity (or otherwise insufficient comparability) in the primary studies, and biased interpretation of findings (Borenstein 2009; Nordmann 2012). The results of meta-analyses that are based on sets of RCTs with lower methodological quality have been reported 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).

Some of the 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 on the comparability of the circumstances of the individual studies, 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 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).

The shortcomings of meta-analyses, which are shared by—though are generally greater in—unstructured literature reviews and other less rigorous synthesis methods, can be minimized by maintaining a systematic approach. Performing meta-analyses as part of high-quality systematic reviews, i.e., that have objective means of searching the literature and apply 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 IV-2. 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 36,974 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 graph 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.

Box IV-2 shows two types of meta-analysis side-by-side: a conventional meta-analysis and a cumulative meta-analysis of the impact of thrombolytic therapy (to dissolve blood clots) on mortality among patients with myocardial infarction. These meta-analyses are applied to the same set of 33 clinical trials reported over a 30-year period. Most of these trials had tens or hundreds of patients, though two were much larger. The "forest plot" diagram on the left represents a single conventional meta-analysis of those 33 trials. Across the sum of nearly 37,000 patients in the 33 trials, that meta-analysis yielded a statistically significant treatment effect favoring the use of streptokinase. The forest plot on the right depicts a cumulative meta-analysis in which iterative meta-analyses could have been performed each time a report of a new trial appeared. The cumulative meta-analysis suggests that a statistically significant treatment effect of streptokinase on morality could have been discerned many years earlier than the appearance of the last of the 33 trials.

Network meta-analysis (also known as multiple-treatment or mixed-treatment comparisons meta-analysis), is used to compare various alternative interventions of interest when there are limited or no available direct ("head-to-head") trials of those interventions. It enables integration of data from available direct trials and from indirect comparisons, i.e., when the alternative interventions are compared based on trials of how effective they are versus a common comparator intervention (Caldwell 2005; Jansen 2011).

Although meta-analysis has been applied primarily for treatments, meta-analytic techniques also are applied 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).

More advanced meta-analytic techniques can be applied to assessing health technologies, e.g., involving multivariate treatment effects, meta-regression, and Bayesian methods (see, e.g., van Houwelingen 2002). **Meta-regression** refers to techniques for relating the magnitude of an effect to one or more characteristics of the studies used in a meta-analysis, such as patient characteristics, drug dose, duration of study, and year of publication (Thompson 2002).

Various computer software packages are available to help conduct meta-analyses; examples are Comprehensive Meta-analysis (CMA), OpenMeta[Analyst], and RevMan, though no particular recommendation is offered here.

D. Guidelines for Reporting Primary and Secondary Research

The conduct of systematic reviews, meta-analysis, and related integrative studies requires systematic examination of the reports of primary data studies as well as other integrative methods. As integrative methods have taken on more central roles in HTA and other forms of evaluation, methodological standards for conducting and reporting these studies have risen (Egger, Davey Smith, Altman 2001; Moher 1999; Petitti 2001; Stroup 2000). In addition to the PRISMA instrument for systematic reviews and meta-analyses noted above, there are other instruments for assessing the reporting of clinical trials, systematic reviews, meta-analyses of trials, meta-analyses of observational studies, and economic analyses. Some of these are listed in **Box IV-3**. HTA programs that use the inclusion/exclusion rules and other aspects of these instruments are more likely to conduct more thorough and credible assessments. In addition to their primary purpose of improving reporting of research, these guidelines are helpful for planning studies of these types and in reviewing studies as part of systematic reviews and other

integrative methods. See also *Research Reporting Guidelines and Initiatives* compiled by the US NLM at: http://www.nlm.nih.gov/services/research_report_guide.html.

Box IV-3. Guidelines for Reporting Research

- AMSTAR (Assessment of Multiple Systematic Reviews) (Shea 2009)
- CHEERS (Consolidated Health Economic Evaluation Reporting Standards) (Husereau 2013)
- CONSORT (Consolidated Standards of Reporting Trials) (Turner 2012)
- GRACE (Good ReseArch for Comparative Effectiveness) (Dreyer 2014)
- MOOSE (Meta-analysis of Observational Studies in Epidemiology) (Stroup 2000)
- PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) (Moher 2009)
- QUOROM (Quality Of Reporting Of Meta-analyses) (Moher 1999)
- STARD (Standards for Reporting of Diagnostic Accuracy) (Bossuyt 2003)
- STROBE (Strengthening the Reporting of OBservational Studies in Epidemiology) (von Elm 2008)
- TREND (Transparent Reporting of Evaluations with Nonrandomized Designs) (Des Jarlais 2004)

E. Modeling

Quantitative modeling is used to evaluate the clinical and economic effects of health care interventions. Models are often 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 of interest. For example, decision analytic modeling can be used to represent alternative sequences of clinical decisions for a given health problem and their expected health outcomes and cost effectiveness.

The high cost and long duration of large RCTs and other clinical studies also contribute to the interest in developing alternative methods to collect, integrate, and analyze data to answer questions about the impacts of alternative health care interventions. Indeed, some advanced types of modeling are being used to simulate (and substitute in certain ways for) clinical trials.

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 adjusting 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; state-transition modeling, including Markov modeling (described below) and Monte Carlo simulation; survival and hazard functions; and fuzzy logic. 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; Siebert 2012). Infectious disease modeling is used to understand the spread, incidence, and prevalence of disease, including modeling those that model the impact health care interventions such as immunizations (Bauch 2010) and insect control (Luz 2011).

Decision analysis uses available quantitative estimates to represent (model or simulate) 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). As described by Rawlins:

> *Combining evidence derived from a range of study designs is a feature of decision-analytic modelling as well as in the emerging fields of teleoanalysis and patient preference trials. Decision-analytic modelling is at the heart of health economic analysis. It involves synthesising evidence from sources that include RCTs, observational studies, case registries, public health statistics, preference surveys and (at least in the US) insurance claim databases (Rawlins 2008).*

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.

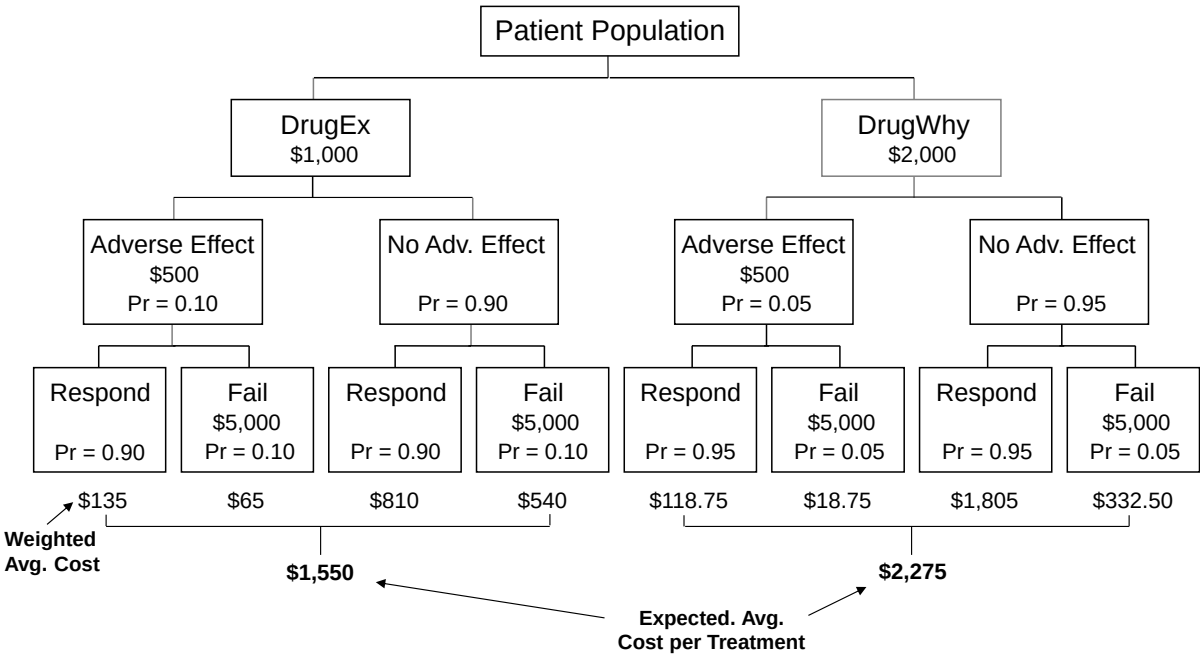
Decision models can be used in different ways. They can be used to predict the distribution of outcomes for patient populations and associated costs of care. They can be used as a tool to support development of clinical practice guidelines for specific health problems. For individual patients, decision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

Although decision analyses can take different forms, the basic steps of a typical approach 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 subject to random variation or based on limited data or simply expert conjecture.)

Box IV-4 shows a decision tree for determining the cost of treatment for alternative drug therapies for a given health problem.

Box IV-4. Decision Analysis Model: Cost per Treatment, DrugEx vs. DrugWhy



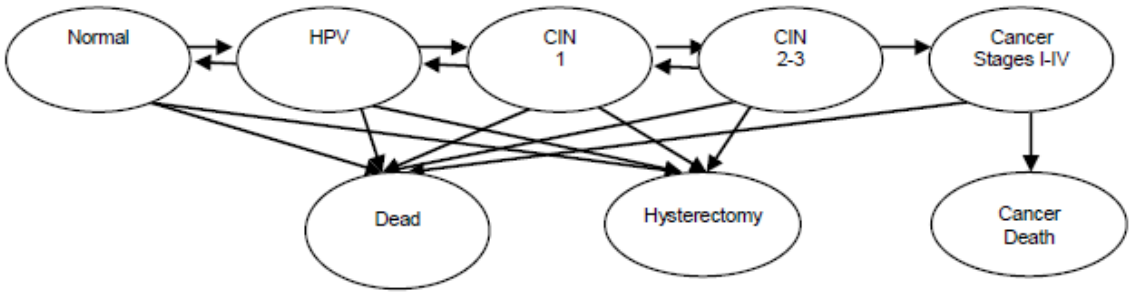
Treatment	Path	Cum. Cost	Cum. Prob.	Weighted Avg. Cost	Expected Avg. Cost/Treatment
DrugEx	1	$1,500	0.09	$135.00	
	2	6,500	0.01	65.00	
	3	1,000	0.81	810.00	
	4	6,000	0.09	540.00	
	Total		1.00		**$1,550**
DrugWhy	1	$2,500	0.0475	$118.75	
	2	7,500	0.0025	18.75	
	3	2,000	0.9025	1,805.00	
	4	7,000	0.0475	332.50	
	Total		1.00		**$2,275**

This decision analysis model compares the average cost per treatment of two drugs for a given patient population. The cost of the new DrugWhy is twice that of DrugEx, the current standard of care. However, the probability (Pr) that using DrugWhy will be associated with an adverse health event, with its own costs, is half of the probability of that adverse event associated with using DrugEx. Also, the response rate of patients (i.e., the percentage of patients for whom the drug is effective) for DrugWhy is slightly higher than that of DrugEx. For patients in whom either drug fails, there is a substantial cost of treatment with other interventions. The model assumes that: the drugs are equally effective when patients respond to them; the cost of an adverse effect associated with either drug is the same; and the cost of treating a failure of either drug is the same. For each drug, there are four potential paths of treatment and associated costs, accounting for whether or not there is an adverse effect and whether or not patients respond to the drug. The model calculates an average cost per treatment of using each drug.

A limitation of modeling with decision trees is representing recurrent health states (i.e., complications or stages of a chronic disease that may come and go, such as in multiple sclerosis). In those instances, a preferable alternative approach is to use state-transition modeling (Siebert 2012), such as in the form of Markov modeling, 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.

A **Markov model** (or chain) is a way to represent and quantify changes from one state of health to another, such as different stages of disease and death. These changes can result from the natural history of a disease or from use of health technologies. These models are especially useful for representing patient or population experience when the health problem of interest involves risks that are continuous over time, when the timing of health states is important, and when some or all these health states may recur. Markov models assume that each patient is always in one of a set of mutually exclusive and exhaustive health states, with a set of allowable (i.e., non-zero) probabilities of moving from one health state to another, including remaining in the same state. These states might include normal, asymptomatic disease, one or more stages of progressive disease, and death. For example, in cardiovascular disease, these might include normal, unstable angina, myocardial infarction, stroke, cardiovascular death, and death from other causes. Patient utilities and costs also can be assigned to each health state or event. In representing recurring health states, time dependence of the probabilities of moving among health states, and patient utility and costs for those health states, Markov models enable modeling the consequences or impacts of health technologies (Sonnenberg 1993). **Box IV-5** shows a Markov chain for transitions among disease states for the natural history of cervical cancer.

Box IV-5. Disease States and Allowed Transitions for the Natural History Component of a Markov Model Used in a Decision Analysis of Cervical Cancer Screening



HPV: human papillomavirus; CIN: cervical intraepithelial neoplasia (grades 1, 2, 3)
Transition probabilities among disease states are not shown here.

Source: Kulasingam SL, Havrilesky L, Ghebre R, Myers ER. Screening for Cervical Cancer: A Decision Analysis for the U.S. Preventive Services Task Force. AHRQ Pub. No. 11-05157-EF-1. Rockville, MD: Agency for Healthcare Research and Quality; May 2011.

High-power computing technology, higher mathematics, and large data systems are being used for simulations of clinical trials and other advanced applications. A prominent example is the Archimedes model, a large-scale simulation system that models human physiology, disease, and health care systems. The Archimedes model uses information about anatomy and physiology; data from clinical trials, observational studies, and retrospective studies; and hundreds of equations. In more than 15 diseases and conditions, it models metabolic pathways, onset and progression of diseases, signs and symptoms of disease, health care tests and treatments, health outcomes, health services utilization, and costs. In diabetes, for example, the Archimedes model has been used to predict the risk of developing diabetes in individuals (Stern 2008), determine the cost-effectiveness of alternative screening strategies to detect new cases of diabetes (Kahn 2010), and simulate clinical trials of treatments for diabetes (Eddy 2003).

One of the challenges of decision analysis is accounting for the varying perspectives of stakeholders in a given decision, including what attributes or criteria (e.g., health benefit, avoidance of adverse events, impact on quality of life, patient copayment) are important to each stakeholder and the relative importance or weight of each attribute. **Multi-criteria decision analysis** (MCDA) has been applied to HTA (Goetghebeur 2012; Thokala 2012). A form of operations research, MCDA is a group of methods for identifying and comparing the attributes of alternatives (e.g., therapeutic options) from the perspectives of multiple stakeholders. It evaluates these alternatives by ranking, rating, or pairwise comparisons, using such stakeholder elicitation techniques as conjoint analysis and analytic hierarchy process.

Models and their results are only aids to decision making; they are 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 (see, e.g., Gold 1996; Philips 2004; Soto 2002; Weinstein 2003).

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

Various computer software packages are available to conduct decision-analytic and other forms of modeling; examples are Decision Analysis, Excel, and TreeAge; no particular recommendation is offered here.

F. Assessing the Quality of a Body of Evidence

Systematic reviews assemble bodies of evidence pertaining to particular evidence questions. Although each body of evidence may comprise studies of one type, e.g., RCTs, they may also comprise studies of multiple designs. Many approaches have been used to assess the quality of a body of evidence since the 1970s. In recent years, there has been some convergence in these approaches, including by such organizations as the Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group (Balshem 2011), the Cochrane Collaboration (Higgins 2011), the US Agency for Healthcare Research and Quality Evidence-based Practice Centers (AHRQ EPCs) (Berkman 2014), the Oxford Centre for Evidence-Based Medicine (OCEBM Levels of Evidence Working Group 2011), and the US Preventive Services Task Force (USPSTF) (US Preventive Services Task Force 2008). According to the GRADE Working Group, more than 70 organizations, including international collaborations, HTA agencies, public health agencies, medical professional societies, and others have endorsed GRADE and are using it or modified versions of it (GRADE Working Group 2013).

Increasingly, organizations such as those noted above consider the following types of factors, dimensions, or domains when assessing the quality of a body of evidence:

- Risk of bias
- Precision
- Consistency
- Directness
- Publication (or reporting) bias

- Magnitude of effect size (or treatment effect)
- Presence of confounders that would diminish an observed effect
- Dose-response effect (or gradient)

Each of these dimensions is described briefly, below.

Risk of bias refers to threats to internal validity, i.e., limitations in the design and implementation of studies that may cause some systematic deviation in an observation from the true nature of an event, such the deviation of an observed treatment effect from the true treatment effect. For a body of evidence, this refers to bias in the overall or cumulative observed treatment effect of the group of relevant studies, for example, as would be derived in a meta-analysis. As described in chapter III regarding the quality of individual studies, the quality of a body of evidence is subject to various types of bias across its individual studies. Among these are selection bias (including lack of allocation concealment), performance bias (including insufficient blinding of patients and investigators), attrition bias, and detection bias. Some quality rating schemes for bodies of evidence compile aggregate ratings of the risk of bias in individual studies.

Precision refers to the extent to which a measurement, such as the mean estimate of a treatment effect, is derived from a set of observations having small variation (i.e., are close in magnitude to each other). Precision is inversely related to random error. Small sample sizes and few observations generally widen the **confidence interval** around an estimate of an effect, decreasing the precision of that estimate and lowering any rating of the quality of the evidence. Due to potential sources of bias that may increase or decrease the observed magnitude of a treatment effect, a precise estimate is not necessarily an accurate one. As noted in chapter III, some researchers contend that if individual studies are to be assembled into a body of evidence for a systematic review, precision should be evaluated not at the level of individual studies, but when assessing the quality of the body of evidence. This is intended to avoid double-counting limitations in precision from the same source (Viswanathan 2014).

Consistency refers to the extent that the results of studies in a body of evidence are in agreement. Consistency can be assessed based on the direction of an effect, i.e., whether they are on the positive or negative side of no effect or the magnitudes of effect sizes across the studies are similar. One indication of consistency across studies in a body of evidence is overlap of their respective confidence intervals around an effect size. Investigators should seek to explain inconsistency (or heterogeneity) of results. For example, inconsistent results may arise from a body of evidence with studies of different populations or different doses or intensity of a treatment. Plausible explanations of these inconsistent results may include that, in similar patient populations, a larger dose achieves a larger treatment effect; or, given the same dose, a sicker population experiences a larger treatment effect than a less sick population. The quality of a body of evidence may be lower when there are no plausible explanations for inconsistent results.

Directness has multiple meanings in assessing the quality of a body of evidence. First, directness refers to the proximity of comparison in studies, that is, whether the available evidence is based on a "head-to-head" (i.e., direct) comparison of the intervention and comparator of interest, or whether it must rely on some other basis of comparison (i.e., directness of comparisons). For example, where there is no direct evidence pertaining to intervention A vs. comparator B, evidence may be available for intervention A vs. comparator C and of comparator B vs. comparator C; this could provide an indirect

basis for the comparison intervention A vs. comparator B. This form of directness can apply for individual studies as well as a body of evidence.

Second, directness refers to how many bodies of evidence are required to link the use of an intervention to the impact on the outcome of interest (i.e., directness of outcomes). For example, in determining whether a screening test has an impact on a health outcome, a single body of evidence (e.g., from a set of similar RCTs) that randomizes patients to the screening test and to no screening and follows both populations through any detection of a condition, treatment decisions, and outcomes would comprise direct evidence. Requiring multiple bodies of evidence to show each of detection of the condition, impact of detection on a treatment decision, impact of treatment on an intermediate outcome, and then impact of the intermediate outcome on the outcome of interest would constitute indirect evidence.

Third, directness can refer to the extent to which the focus or content of an individual study or group of studies diverges from an evidence question of interest. Although evidence questions typically specify most or all of the elements of PICOTS (patient populations, interventions, comparators, outcomes, timing, and setting of care) or similar factors, the potentially relevant available studies may differ in one or more of those respects. As such, directness may be characterized as the extent to which the PICOTS of the studies in a body of evidence align with the PICOTS of the evidence question of interest. This type of directness reflects the external validity of the body of evidence, i.e., how well the available evidence represents, or can be generalized to, the circumstances of interest. Some approaches to quality assessment of a body of evidence address external validity of evidence separately, noting that external validity of a given body of evidence may vary by the user or target audience (Berkman 2014). Some researchers suggest that, if individual studies are to be assembled into a body of evidence for a systematic review, then external validity should be evaluated only once, i.e., when assessing the quality of the body of evidence, not at the level of individual studies (Atkins 2004; Viswanathan 2014).

Publication bias refers to unrepresentative publication of research reports that is not due to the quality of the research but to other characteristics. This includes tendencies of investigators and sponsors to submit, and publishers to accept, reports of studies with "positive" results, such as those that detect beneficial treatment effects of a new intervention, as opposed to those with "negative" results (no treatment effect or high adverse event rates). Studies with positive results also are more likely than those with negative results to be published in English, be cited in other publications, and generate multiple publications (Sterne 2001). When there is reason to believe that the set of published studies is not representative of all relevant studies, there is less confidence that the reported treatment effect for a body of evidence reflects the true treatment effect, thereby diminishing the quality of that body of evidence. Prospective registration of clinical trials (e.g., in ClinicalTrials.gov), adherence to guidelines for reporting research, and efforts to seek out relevant unpublished reports are three approaches used to manage publication bias (Song 2010).

One approach used for detecting possible publication bias in systematic reviews and meta-analyses is to use a **funnel plot** that graphs the distribution of reported treatment effects from individual studies against the sample sizes of the studies. This approach assumes that the reported treatment effects of larger studies will be closer to the average treatment effect (reflecting greater precision), while the reported treatment effects of smaller studies will be distributed more widely on both sides of the average (reflecting less precision). A funnel plot that is asymmetrical suggests that some studies, such as small ones with negative results, have not been published. However, asymmetry in funnel plots is not a definitive sign of publication bias, as asymmetry may arise from other causes, such as over-estimation of treatment effects in small studies of low methodological quality (Song 2010; Sterne 2011).

The use of the terms, publication bias and reporting bias, varies. For example, in the GRADE framework, reporting bias concerns selective, incomplete, or otherwise differential reporting of findings of individual studies (Balshem 2011). Other guidance on assessing the quality of a body of evidence uses reporting bias as the broader concept, including publication bias as described above and differential reporting of results (Berkman 2014). The Cochrane Collaboration uses reporting bias as the broader term to include not only publication bias, but time lag bias, multiple (duplicate) publication bias, location (i.e., in which journals) bias, citation bias, language bias, and outcome reporting bias (Higgins 2011).

Magnitude of effect size can improve confidence in a body of evidence where the relevant studies report treatment effects that are large, consistent, and precise. Overall treatment effects of this type increase confidence that they did not arise from potentially confounding factors only. For example, the GRADE quality rating approach suggests increasing the quality of evidence by one level when methodologically rigorous observational studies show at least a two-fold change in risk ratio and increasing by two levels for at least a five-fold change in relative risk (Guyatt 2011).

Plausible confounding that would diminish observed effect refers to instances in which plausible confounding factors for which the study design or analysis have not accounted would likely have diminished the observed effect size. That is, the plausible confounding would have pushed the observed effect in the opposite direction of the true effect. As such, the true effect size is probably even larger than the observed effect size. This increases the confidence that there is a true effect. This might arise, for example, in a non-randomized controlled trial (or a comparative observational study) comparing a new treatment to standard care. If, in that instance, the group of patients receiving the new treatment has greater disease severity at baseline than the group of patients receiving standard care, yet the group receiving the new treatment has better outcomes, it is likely that the true treatment effect is even greater than its observed treatment effect.

Dose-response effect (or dose-response gradient) refers to an association in an individual study or across a body of evidence, between the dose, adherence, or duration of an intervention and the observed effect size. That is, within an individual study in which patients received variable doses of (or exposure to) an intervention, the patients that received higher doses also experienced a greater treatment effect. Or, across a set of studies of an intervention in which some studies used higher doses than other studies, those study populations that received higher doses also experienced greater treatment effects. A dose-response effect increases the confidence that an observed treatment effect represents a true treatment effect. Dose-response relationships are typically not linear; further, they may exist only within a certain range of doses.

As is so for assessing the quality of individual studies, the quality of a body of evidence should be graded separately for each main treatment comparison for each major outcome for each where feasible. For example, even for a comparison of one intervention to a standard of care, the quality of the bodies of evidence pertaining to each of mortality, morbidity, various adverse events, and quality of life may differ. For example, the GRADE approach calls for rating the estimate of effect for each critical or otherwise important outcome in a body of evidence. GRADE also specifies that an overall rating of multiple estimates of effect pertains only when recommendations are being made (i.e., not just a quality rating of evidence for individual outcomes) (Guyatt 2013).

Among the important ways in which appraisal of evidence quality has evolved from using traditional evidence hierarchies is the accounting for factors other than study design. For example, as shown in the upper portion of **Box IV-6**, the GRADE approach to rating quality of evidence (which has been adopted by the Cochrane Collaboration and others) starts with a simplified categorization of study types, i.e., RCTs and observational studies, accompanied by two main levels of confidence (high or low) in the estimate of a treatment effect. Then, the rating scheme allows for factors that would raise or lower a level of confidence. Factors that would lower confidence in evidence would include, e.g., risk of bias, inconsistency across the RCTs, indirectness, and publication bias; factors that would increase confidence include, e.g., large effect size and an observed dose-response effect. The final levels of confidence rating (high, moderate, low, very low) are shown at the right, and defined in the lower portion of that

Box IV-6. A Summary of the GRADE Approach to Rating Quality of a Body of Evidence

Table: GRADE's approach to rating quality of evidence (aka confidence in effect estimates)

For each outcome based on a systematic review and across outcomes (lowest quality across the outcomes critical for decision making)

1. Establish initial level of confidence		2. Consider lowering or raising level of confidence		3. Final level of confidence rating
Study design	*Initial confidence in an estimate of effect*	*Reasons for considering lowering or raising confidence*		*Confidence in an estimate of effect across those considerations*
		↓ Lower if	↑ Higher if*	
Randomized trials →	High confidence	Risk of Bias Inconsistency Indirectness Imprecision Publication bias	Large effect Dose response All plausible confounding & bias • would reduce a demonstrated effect or • would suggest a spurious effect if no effect was observed	High ⊕⊕⊕⊕
				Moderate ⊕⊕⊕◯
Observational studies →	Low confidence			Low ⊕⊕◯◯
				Very low ⊕◯◯◯

*upgrading criteria are usually applicable to observational studies only.

Significance of the four levels of evidence

Quality level	Current definition
High	We are very confident that the true effect lies close to that of the estimate of the effect
Moderate	We are moderately confident in the effect estimate: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different
Low	Our confidence in the effect estimate is limited: The true effect may be substantially different fro the estimate of the effect
Very low	We have very little confidence in the effect estimate: The true effect is likely to be substantially different from the estimate of effect

Reprinted with permission: GRADE Working Group, 2013. Balshsem H, et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol. 2011(64):401-6.

box. Similarly, the OCEBM 2011 Levels of Evidence (see chapter III, Box III-13) allows for grading down based on study quality, imprecision, indirectness, or small effect size; and allows for grading up for large effect size. **Box IV-7** shows the strength of evidence grades and definitions for the approach used by the AHRQ EPCs, which are based factors that are very similar to those used in GRADE, as noted above.

Box IV-7. Strength of Evidence Grades and Definitions

Grade	Definition
High	**We are very confident that the estimate of the effect lies close to the true effect for this outcome.** The body of evidence has few or no deficiencies. We believe that the findings are stable, i.e., another study would not change the conclusions.
Moderate	**We are moderately confident that the estimate of effect lies close to the true effect for this outcome.** The body of evidence has some deficiencies. We believe that the findings are likely to be stable, but some doubt remains.
Low	**We have limited confidence that the estimate of effect lies close to the true effect for this outcome.** The body of evidence has major or numerous deficiencies (or both). We believe that additional evidence is needed before concluding either that the findings are stable or that the estimate of effect is close to the true effect.
Insufficient	**We have no evidence, we are unable to estimate an effect, or we have no confidence in the estimate of effect for this outcome.** No evidence is available or the body of evidence has unacceptable deficiencies, precluding reaching a conclusion.

Source: Berkman ND, et al. Chapter 15. Grading the Strength of a Body of Evidence When Assessing Health Care Interventions for the Effective Health Care Program of the Agency for Healthcare Research and Quality: An Update. In: Methods Guide for Effectiveness and Comparative Effectiveness Reviews. AHRQ Publication No. 10(14)-EHC063-EF. Rockville, MD: Agency for Healthcare Research and Quality. January 2014.

G. Consensus Development

In various forms, group judgment or **consensus development** is used to set standards, make regulatory recommendations and decisions, make payment recommendations and policies, make technology acquisition decisions, formulate practice guidelines, define the state-of-the-art, and other purposes. The term "consensus development" can refer to particular group processes or techniques that generally are intended to derive best estimates of parameters or general (or unanimous) agreement on a set of findings or recommendations. It also can refer to particular methodological paradigms or approaches, e.g., the consensus development conferences that were conducted by the US 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; video and web conferencing and related telecommunications* approaches also are used.

In HTA, consensus development is not used as the sole approach to deriving findings or recommendations, but rather as supported by systematic reviews and other analyses and data. Virtually all HTA efforts involve some form of consensus development at some juncture, including one or more of three main steps of HTA: interpret evidence, integrate evidence, and formulate findings and

recommendations. Consensus development also can be used for ranking, such as to set assessment priorities, and for rating, such as drawing on available evidence and expert opinion to develop practice guidelines.

The opinion of an expert committee concerning, e.g., the effectiveness of a particular intervention, does not in itself constitute strong evidence. The experience of experts in the forms of, e.g., individual cases or series of cases could comprise poor evidence, as it is subject to multiple forms of bias (selection bias, recall bias, reporting bias, etc.). Where they exist, the results of pertinent, rigorous scientific studies 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. HTA must be explicit regarding where the evidence stops and where the expert group opinion begins.

Many consensus development programs in the US and around the world were 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 modified and experimented with its process over the years. Especially in later years, these conferences usually involved a systematic review (such as prepared by an AHRQ Evidence-based Practice Center), in addition to invited expert speaker testimony and public (audience) testimony. The NIH program was discontinued in 2013, after having conducted nearly 130 consensus development conferences and nearly 40 state-of-the-science conferences that used a similar format. Australia, Canada, Denmark, France, Israel, Japan, The Netherlands, Spain, Sweden and the UK are among the countries that used various forms of consensus development programs to evaluate health technologies, some of which were later adapted or incorporated into HTA programs (McGlynn 1990).

Various evaluations and other reports have defined attributes or made recommendations concerning how to strengthen consensus development programs (Goodman 1990; Institute of Medicine 1990; Olsen 1995; Portnoy 2007). Much of this material has contributed to HTA and related fields that use forms of group process.

References for Chapter IV

Atkins D, Best D, Briss PA, Eccles M, et al., GRADE Working Group. Grading quality of evidence and strength of recommendations. BMJ. 2004;328(7454):1490.

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(22):2973-82.

Balshem H, Helfand M, Schünemann HJ, Oxman AD, et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol. 2011;64(4):401-6.

Bauch CT, Li M, Chapman G, Galvani AP. Adherence to cervical screening in the era of human papillomavirus vaccination: how low is too low? Lancet Infect Dis. 2010;10(2):133-7.

Bayliss SE, Davenport C. Locating systematic reviews of test accuracy studies: how five specialist review databases measure up. Int J Technol Assess Health Care. 2008;24(4):403-11.

Berkman ND, Lohr KN, Ansari M, McDonagh M, et al. Chapter 15. Grading the Strength of a Body of Evidence When Assessing Health Care Interventions for the Effective Health Care Program of the Agency for Healthcare Research and Quality: An Update. In: Methods Guide for Effectiveness and Comparative Effectiveness Reviews. AHRQ Publication No. 10(14)-EHC063-EF. Rockville, MD: Agency for Healthcare Research and Quality. January 2014. Accessed Feb. 1, 2014 at: http://effectivehealthcare.ahrq.gov/ehc/products/60/318/CER-Methods-Guide-140109.pdf.

Buckley DI, Ansari M, Butler M, Williams C, Chang C. Chapter 4. The Refinement of Topics for Systematic Reviews: Lessons and Recommendations from the Effective Health Care Program. In: Methods Guide for Effectiveness and Comparative Effectiveness Reviews. AHRQ Publication No. 10(14)-EHC063-EF. Rockville, MD: Agency for Healthcare Research and Quality. January 2014. Accessed Feb. 1, 2014 at: http://effectivehealthcare.ahrq.gov/ehc/products/60/318/CER-Methods-Guide-140109.pdf.

Booth A, Clarke M, Dooley G, Ghersi D, et al. PROSPERO at one year: an evaluation of its utility. Syst Rev. 2013;2:4.

Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Introduction to Meta-Analysis. Chapter 43: Criticisms of meta-analysis. New York, NY: John Wiley & Sons; 2009.

Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, et al.; Standards for Reporting of Diagnostic Accuracy. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. Ann Intern Med. 2003;138(1):40-4.

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(4):282-91.

Caldwell DM, Ades AE, Higgins JP. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. BMJ. 2005;331(7521):897–900.

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

Deeks JJ. Systematic reviews in health care: systematic reviews of evaluations of diagnostic and screening tests. BMJ. 2001;323(7305):157-62.

Des Jarlais DC, Lyles C, Crepaz N; TREND Group. Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: the TREND statement. Am J Public Health. 2004;94(3):361-6. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1448256.

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.

Dreyer NA, Velentgas P, Westrich K, Dubois R. The GRACE Checklist for Rating the Quality of Observational Studies of Comparative Effectiveness: A Tale of Hope and Caution. J Manag Care Pharm. 2014;20(3):301-8.

Eckman MH, Levine HJ, Pauker SG. Decision analytic and cost-effectiveness issues concerning anticoagulant prophylaxis in heart disease. Chest. 1992;102(suppl. 4):538-549S.

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

Eddy DM, Schlessinger L. Validation of the Archimedes diabetes model. Diabetes Care. 2003;26(11):3102-10.

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

Egger M, Smith GD, Sterne JA. Uses and abuses of meta-analysis. Clin Med. 2001;1(6):478-84.

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

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.

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(4):544-55.

Goetghebeur MM, Wagner M, Khoury H, et al. Bridging health technology assessment (HTA) and efficient health care decision making with multicriteria decision analysis (MCDA): applying the EVIDEM framework to medicines appraisal. Med Decis Making. 2012;32(2):376-88.

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

Goodman C, Baratz SR, eds. Improving Consensus Development for Health Technology Assessment: An International Perspective. Washington, DC: National Academy Press; 1990. Accessed Nov. 1, 2013 at: http://www.nap.edu/openbook.php?record_id=1628&page=1.

GRADE Working Group. Organizations that have endorsed or that are using GRADE. 2013. Accessed Sept. 1, 2013 at: http://www.gradeworkinggroup.org/society/index.htm.

Guyatt G, Oxman AD, Sultan S, Brozek J, et al. GRADE guidelines: 11. Making an overall rating of confidence in effect estimates for a single outcome and for all outcomes. J Clin Epidemiol. 2013;66(2):151-7.

Guyatt GH, Oxman AD, Sultan S, Glasziou P, et al., GRADE Working Group. GRADE guidelines: 9. Rating up the quality of evidence. J Clin Epidemiol. 2011;64(12):1311-6.

Hasselblad V, Hedges LV. Meta-analysis of screening and diagnostic tests. Psychol Bull. 1995;117(1): 167-78.

Hernandez DA, El-Masri MM, Hernandez CA. Choosing and using citation and bibliographic database software (BDS). Diabetes Educ. 2008;34(3):457-74.

Higgins JPT, Green S, eds. Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Accessed Sept. 1, 2013 at: http://handbook.cochrane.org.

Husereau D, Drummond M, Petrou S, Carswell C, et al.; CHEERS Task Force. Consolidated Health Economic Evaluation Reporting Standards (CHEERS) statement. Int J Technol Assess Health Care. 2013;29(2):117-22.

Institute of Medicine. Consensus Development at the NIH: Improving the Program. Washington, DC: National Academy Press; 1990. Accessed Nov. 1, 2013 at: http://www.nap.edu/openbook.php?record_id=1563&page=1.

Irwig L, Tosteson AN, Gatsonis C, Lau J, et al. Guidelines for meta-analyses evaluating diagnostic tests. Ann Intern Med. 1994;120(8):667-76.

Jairath N, Weinstein J. The Delphi methodology (part two): a useful administrative approach. Can J Nurs Admin. 1994;7(4):7-20.

Jansen JP, Fleurence R, Devine B, et al. Interpreting indirect treatment comparisons and network meta-analysis for health-care decision making: report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices: part 1. Value Health. 2011;14(4):417-28.

Kahn R, Alperin P, Eddy D, et al. Age at initiation and frequency of screening to detect type 2 diabetes: a cost-effectiveness analysis. Lancet. 2010;375(9723):1365-74.

Kulasingam SL, Havrilesky L, Ghebre R, Myers ER. Screening for Cervical Cancer: A Decision Analysis for the U.S. Preventive Services Task Force. AHRQ Pub. No. 11-05157-EF-1. Rockville, MD: Agency for Healthcare Research and Quality; May 2011.Laird NM, Mosteller F. Some statistical methods for combining experimental results. Int J Technol Assess Health Care. 1990;6(1):5-30. Accessed Nov. 1, 2013 at: http://www.ncbi.nlm.nih.gov/books/NBK92546.

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.

Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS Med. 2009;6(7):e1000100.

Littenberg B, Moses LE. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. Med Dec Making. 1993;13(4):313-21.

Luz PM, Vanni T, Medlock J, Paltiel AD, Galvani AP. Dengue vector control strategies in an urban setting: an economic modelling assessment. Lancet. 2011;377(9778):1673-80.

Lyles CM, Kay LS, Crepaz N, Herbst JH, et al. Best-evidence interventions: findings from a systematic review of HIV behavioral interventions for US populations at high risk, 2000-2004. Am J Public Health. 2007;97(1):133-43.

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.

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(9193):1896-900.

Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med. 2009;6(7): e1000097.

Moher D, Pham B, Jones A, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet. 1998;352(9128):609-13.

Nordmann AJ, Kasenda B, Briel M. Meta-analyses: what they can and cannot do. Swiss Med Wkly. 2012;142:w13518.

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

OCEBM Levels of Evidence Working Group. The Oxford 2011 Levels of Evidence. Oxford Centre for Evidence-Based Medicine. Accessed Sept. 1, 2013 at: http://www.cebm.net/index.aspx?o=5653.

Ogilvie D, Egan M, Hamilton V, Petticrew M. Systematic reviews of health effects of social interventions: 2. Best available evidence: how low should you go? J Epidemiol Community Health. 2005;59(10):886-92.

Olson CM. Consensus statements: applying structure. JAMA. 1995;273(1):72-3.

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.

Philips Z, Ginnelly L, Sculpher M, Claxton K, et al. Review of guidelines for good practice in decision-analytic modelling in health technology assessment. Health Technol Assess. 2004;8(36):iii-iv, ix-xi, 1-158.

Portnoy B, Miller J, Brown-Huamani K, DeVoto E. Impact of the National Institutes of Health Consensus Development Program on stimulating National Institutes of Health-funded research, 1998 to 2001. Int J Technol Assess Health Care, 2007;23(3):343-8.

Rawlins MD. On the evidence for decisions about the use of therapeutic interventions. The Harveian Oration of 2008. London: Royal College of Physicians, 2008.

Rew L. The systematic review of literature: synthesizing evidence for practice. J Spec Pediatr Nurs. 2011;16(1):64-9.

Sassi, F. Setting priorities for the evaluation of health interventions: when theory does not meet practice. Health Policy. 2003;63(2):141-54.

Savoie I, Helmer D, Green CJ, Kazanjian A. Beyond Medline: reducing bias through extended systematic review search. Int J Technol Assess Health Care. 2003;19(1):168-78.

Shea BJ, Grimshaw JM, Wells GA, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. BMC Med Res Methodol. 2007;7:10.

Shea BJ, Hamel C, Wells GA, Bouter LM, et al. AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. J Clin Epidemiol. 2009;62(10):1013-20.

Siebert U, Alagoz O, Bayoumi AM, Jahn B, et al.; ISPOR-SMDM Modeling Good Research Practices Task Force. State-transition modeling: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force--3. Value Health. 2012;15(6):812-20. Accessed Feb. 1, 2014 at: http://mdm.sagepub.com/content/32/5/690.full.pdf+html.

Slavin RE. Best evidence synthesis: an intelligent alternative to meta-analysis. J Clin Epidemiol. 1995;48(1):9-18.

Song F, Parekh S, Hooper L, Loke YK, et al. Dissemination and publication of research findings: an updated review of related biases. Health Technol Assess. 2010;14(8):iii, ix-xi, 1-193.

Sonnenberg FA, Beck JR. Markov models in medical decision making: a practical guide. Med Decis Making. 1993;13(4):322-38.

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(1):94-111.

Stern M, Williams K, Eddy D, Kahn R. Validation of prediction of diabetes by the Archimedes model and comparison with other predicting models. Diabetes Care. 2008;31(8):1670-1.

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(7304):101-5.

Sterne JA, Sutton AJ, Ioannidis JP, Terrin N, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomized controlled trials. BMJ. 2011;343:d4002.

Stroup DF, Berlin JA, Morton SC, et al. Meta-analysis of observational studies in epidemiology. A proposal for reporting.Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. JAMA. 2000;283:2008-12.

Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. Systematic reviews of trials and other studies. Health Technol Assess. 1998;2(19):1-276.

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

Thokala P, Duenas A. Multiple criteria decision analysis for health technology assessment. Value Health. 2012;15(8):1172-81.

Thompson SG, Higgins JP. How should meta-regression analyses be undertaken and interpreted? Stat Med. 2002;21(11):1559-73.

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

Turner L, Shamseer L, Altman DG, Weeks L, et al. Consolidated standards of reporting trials (CONSORT) and the completeness of reporting of randomised controlled trials (RCTs) published in medical journals. Cochrane Database Syst Rev. 2012 Nov 14;11:MR000030.

US Preventive Services Task Force. Procedure Manual. AHRQ Publication No. 08-05118-EF, July 2008. Accessed Aug. 1, 2013 at: http://www.uspreventiveservicestaskforce.org/uspstf08/methods/procmanual.htm.

van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. Stat Med. 2002;21(4):589-624.

Viswanathan M, Ansari MT, Berkman ND, Chang S, et al. Chapter 9. Assessing the risk of bias of individual studies in systematic reviews of health care interventions. In: Methods Guide for Effectiveness and Comparative Effectiveness Reviews. AHRQ Publication No. 10(14)-EHC063-EF. Rockville, MD: Agency for Healthcare Research and Quality. January 2014. Accessed Feb. 1, 2014 at: http://www.effectivehealthcare.ahrq.gov/ehc/products/60/318/CER-Methods-Guide-140109.pdf.

von Elm E, Altman DG, Egger M, Pocock SJ, et al.; STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol. 2008;61(4):344-9.

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.

Whiting P, Westwood M, Burke M, Sterne J, Glanville J. Systematic reviews of test accuracy should search a range of databases to identify primary studies. J Clin Epidemiol. 2008;61(4):357-64.

V. ECONOMIC 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, for example, as part of RCTs and other clinical studies as well as administrative ("claims") 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 analyses 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 great increase in the number of reports of cost analyses in the literature and further refinement of these methods.

A. Main Types of Economic Analysis in HTA

There is a variety of approaches to economic analysis, the suitability of any of which depends on 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 outcomes (or outputs or benefits), and the units used to quantify these may differ.

Main types of economic analysis used in HTA 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 diabetes, 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.
- *Budget-impact analysis (BIA):* determines the impact of implementing or adopting a particular technology or technology-related policy on a designated budget, e.g., of a drug formulary or health plan.

The differences in valuation of costs and outcomes among these alternative are shown in **Box V-1**.

Box V-1. Types of Economic Analysis Used in HTA

Analysis Type	Valuation of costs[1]		Valuation of outcomes
Cost of Illness	$	vs.	None
Cost Minimization	$	vs.	Assume same
Cost Effectiveness	$	÷	Natural units
▪ Cost Consequence	$	vs.	Natural units
▪ Cost Utility	$	÷	Utiles (e.g., QALYs)
Cost Benefit	$	÷ or[2] -	$
Budget Impact	$	vs.	None[3] or maximize various[4]

[1]Any currency
[2]Cost-benefit ratio (÷) or net of costs and benefits (-)
[3]That is, determine impact of an intervention/program on a designated non-fixed budget
[4]That is, maximize some outcome within a designated fixed ("capped") budget

Cost-minimization analysis, CEA and CUA necessarily involve comparisons of alternative interventions. A technology cannot be simply cost effective on its own, though it may be cost effective compared to something else.

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 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 CBA are the 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 ratio approach and a net benefits 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 can yield different preferences among alternative technologies. **Box V-2** shows basic formulas for determining CEA, CUA, and CBA.

Box V-2. Basic Formulas for CEA, CUA, and CBA

A: Technology A
C: Technology C (a comparator)

Cost-Effectiveness Ratio:

$$\text{CE Ratio} = \frac{\$\text{Cost}_A - \$\text{Cost}_C}{\text{Effect}_A - \text{Effect}_C}$$

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

Cost-Utility Ratio:

$$\text{CU Ratio} = \frac{\$\text{Cost}_A - \$\text{Cost}_C}{\text{Utile}_A - \text{Utile}_C}$$

Utiles, units of utility or preference, are often measured in QALYs. So, for example: "$150,000 per QALY gained" or "$12,000 per QALY gained"

Cost-Benefit, Ratio Approach:

$$\text{CB Ratio} = \frac{\$\text{Cost}_A - \$\text{Cost}_C}{\$\text{Benefit}_A - \$\text{Benefit}_C}$$

For example: "Cost-benefit ratio of 1.2"

Cost-Benefit, Net Benefit Approach:

$$\text{CB Net} = (\$\text{Cost}_A - \$\text{Cost}_C) - (\$\text{Benefit}_A - \$\text{Benefit}_C)$$

For example: "Net cost of $10,000"

B. Key Attributes of Cost Analyses

The approaches to accounting for costs and outcomes in cost analyses can vary in a number of important respects, such as choice of comparator, perspective of economic analysis, and time horizon of analysis. A basic "checklist" of these attributes is shown in **Box V-3**, and some are described briefly below. These attributes should be carefully considered by investigators who design and conduct cost analyses, assessors who review or appraise reports of these analyses, and policymakers who intend to make use of findings based on these analyses. Given the different ways in which costs and outcomes may be determined, all studies should make clear their methodology in these respects.

Box V-3. Attributes to Consider When Designing and Reviewing Cost Analyses

- ❑ Comparator
- ❑ Perspective
- ❑ Outcomes/endpoints selected
- ❑ Efficacy vs. effectiveness
- ❑ Data capture method
- ❑ Direct costs (health care and non-health care)
- ❑ Indirect costs (e.g., loss of productivity)
- ❑ Actual costs vs. charges/prices
- ❑ Marginal costs vs. average costs
- ❑ Time horizon of analysis
- ❑ Discounting
- ❑ Correction for inflation
- ❑ Modeling use
- ❑ Sensitivity analysis
- ❑ Reporting results
- ❑ Funding source

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 are realized. For instance, the perspective 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 societal perspective that seeks to identify all costs and all outcomes accordingly. However, "society" as such may not be the decision maker, and what is cost effective from that perspective may not be what is cost effective from the standpoint of a ministry of health, insurance company, hospital manager, patient, or other decision maker. It is possible that this perspective may resemble that of a national or regional 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. 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., taken

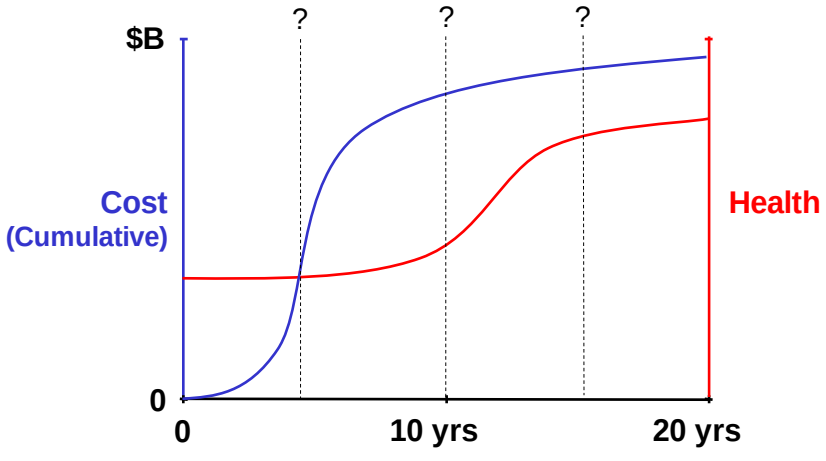
from price lists) rather than *true costs,* whose determination may require special analyses of resource consumption. Charges (as well as actual payments) tend to reflect provider cost-shifting and other factors that decrease the validity of using charges to represent the true costs of providing care.

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 (sometimes known as "presenteeism"), 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 associated with a particular intervention usually do not accrue in steady streams over time. As suggested in **Box V-4,** where most of the health benefits resulting from investing in a health program are realized years after most of the spending, any determination of the cost effectiveness of a particular health program can vary widely depending on the time horizon (e.g., 5, 10, 15, or 20 years since inception of the program) of the analysis. The meaningful time horizons for assessing the cost effectiveness of, e.g., emergency appendectomies, cholesterol-lowering in high-risk adults, and smoking prevention 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 reached a break-even point in about three years, after which transplantation provided a net financial gain compared to dialysis, according to the analysis (Rettig 1991).

Box V-4. Time Horizon of Economic Analysis

Determination of cost effectiveness can depend on relative accruals of costs and health benefits at selected time horizon of analysis. Here, improvements in population health lag behind (are delayed relative to) costs. What is an appropriate time to conduct a cost-effectiveness analysis?



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 the standard of care or another 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 V-5**, the average cost per desired outcome of a protocol of iterative screening tests 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.

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 3% or 5% 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 the same 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 being evaluated. **Box V-6** 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 changes 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 cost or cost-effectiveness for one year is 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.

Box V-5. Average Cost Analysis vs. Marginal Cost Analysis

The importance of determining marginal costs is apparent in this analysis 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.

While cost analysis methods are generally becoming more harmonized, considerable variation in their use remains. Although some variation is unavoidable, some differences in economic perspective, accounting for direct and indirect costs, time horizon, discounting and other aspects are arbitrary, result from lack of expertise, and may reflect biases on the part of investigators or study sponsors (Elixhauser 1998; Hjelmgren 2001; Nixon 2000). Guidelines and related standards for conducting costs analyses in HTA have evolved, especially since the 1990s, helping to improve their quality, reporting, and interpretation of their findings (Byford 1998; Drummond 2005; Eccles 2001; Gold 1996; Husereau 2013; Mauskopf 2007; McGhan 2009; Weinstein 1977). Also, HTA agencies and collaborations are increasingly transparent about their methods involving economic analyses (IQWiG 2009; Wonderling 2011).

Box V-6. 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

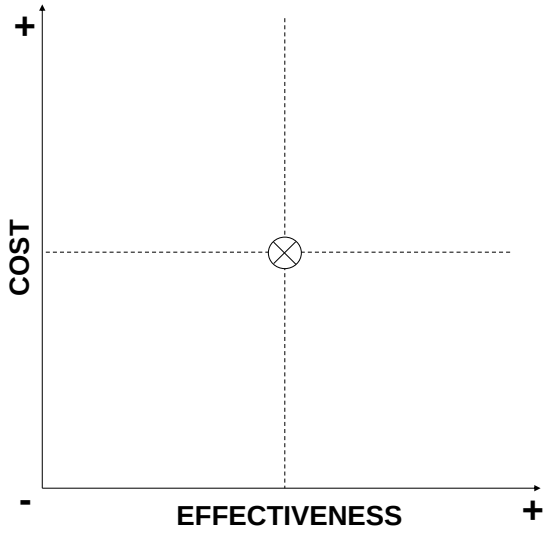
For example, the present value of a cost (or benefit) of $1,000 occurring:

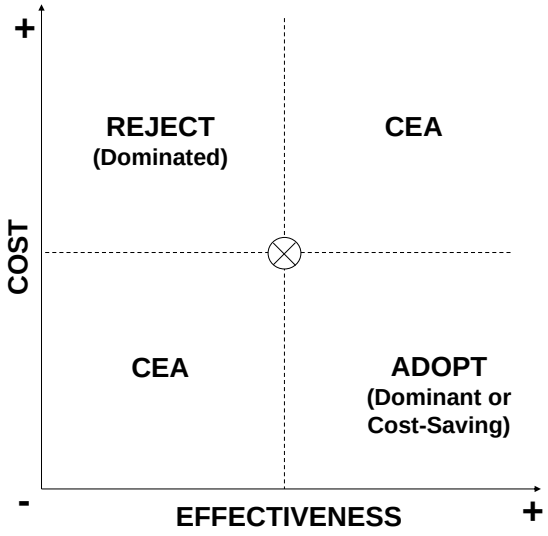
- 1 year in the future, using 10% discount rate, is $910
- 5 years in the future, using 3% discount rate, is $860
- 50 years in the future, using 5% discount rate, is $90

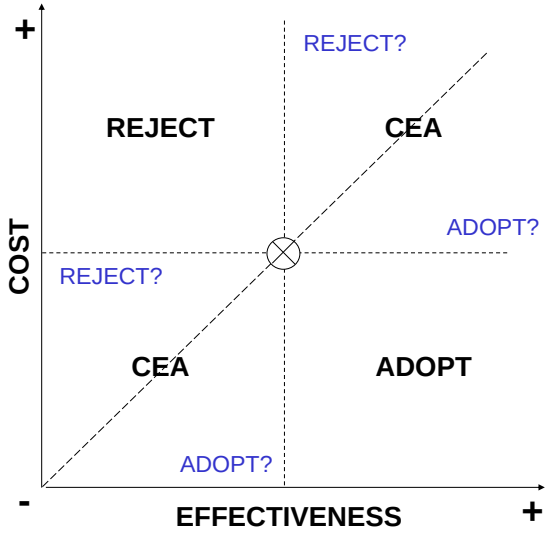
C. Cost-Effectiveness Plane

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 the new intervention in four quadrants of a cost-effectiveness plane as shown in **Box V-7.** Starting with the upper graph, the level of costs and the level of effectiveness for the current standard of care are indicated by the "X" at the center. 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. As shown in the middle graph, if 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 is "dominated" and should be rejected), or it has lower costs and higher effectiveness (indicating that it is "dominant" and 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.

Box V-7. Quadrants of Cost Effectiveness







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 graph of **Box V-7**, 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?).

D. Cost-Utility Analysis Using Cost per QALY

As noted above, QALYs are often used in cost-utility analysis for the purposes of optimizing allocation of health care spending to maximize QALYs gained, and thereby maximize social welfare. Cost per QALY gained, i.e., the marginal (additional or incremental) cost required to gain 1.0 QALY by using a technology, is one means to quantify the value to society of using that technology instead of the alternative(s). Because the QALY incorporates length of life and quality of life but is not specific to any particular disease state or condition, it enables cost-utility comparisons across virtually the entire spectrum of health care interventions.

As shown in **Box V-8,** a cost per QALY analysis can account for large differences in technology costs, survival, and quality of life. Here, cost utilities are compared for three alternative therapies for a particular disease, end-stage heart disease.

An early, controversial example of using cost-utility analysis to assess the relative societal benefits of a diverse range of technologies is shown in **Box V-9**. In this type of list (sometimes known as a "league table"), allocating health care spending for the technologies higher on the list is more efficient (i.e., purchases the next QALY at lower cost and therefore enables maximizing the QALYs purchased) than allocating those resources to technologies further down the list. That is, the technologies on that list are arranged in order of greatest to least cost utility.

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

Outcomes and Costs by Therapy*

Therapy	Life years gained (yr)	Mean utility	QALY gained (yr)	Aggregate cost ($)
A. Conventional medical treatment	0.50	0.06	0.03	28,500
B. Heart transplantation	11.30	0.75	8.45	298,200
C. Total artificial heart (TAH)	4.42	0.65	2.88	327,600

*Costs and outcomes were discounted at 3% 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.

Cost-Utility Ratios for Therapy Comparisons

Comparison	Incremental QALY (yr)	Incremental Cost ($)	Marginal Cost per QALY ($/yr)
Heart transplantation vs. Conventional medical (B – A)	8.42	269,700	32,031
Total artificial heart vs. Conventional medical (C – A)	2.85	299,100	104,947
Total artificial heart vs. Heart transplantation (C – B)	-5.57	29,400	Dominated

This cost-utility comparison indicates that, for patients with end-stage heart disease, both heart transplantation and the total artificial heart yield more quality-adjusted life years at higher costs compared to conventional medical therapy. However, the cost-utility ratio of heart transplantation vs. conventional medical treatment is preferred to (i.e., lower than) the cost-utility ratio of total artificial heart vs. conventional medical therapy. Also, compared to heart transplantation, the total artificial heart costs more and results in fewer quality-adjusted life years, and therefore the total artificial heart is "dominated" by heart transplantation.

Adapted from estimates provided in: Hogness JR, Van Antwerp M. The Artificial Heart: Prototypes, Policies, and Patients. Washington, DC: National Academy Press; 1991.

Box V-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 the allocation rule is intended to optimize societal gain in net health benefits (e.g., as measured using QALYs).

In some instances, the impact of a technology on survival may be sufficiently great as to diminish its relative impact on HRQL, such that there is little need to adjust survival for HRQL. In other instances, the impact of an intervention on HRQL is great, and adjusting survival for it to determine the QALYs gained will affect the relative cost-utility of alternative interventions (Chapman 2004; Greenberg 2011).

An example of a league table of costs per DALY gained for several interventions in low-to-middle-income countries is shown in **Box V-10**.

Box V-10. Cost per DALY Gained for Selected Interventions in Low- and Middle-Income Countries

	Cost per DALY[1] (US$)
Basic childhood vaccines	7
Tuberculosis treatment[2]	102
Improved emergency obstetric care[3]	127
Polypill to prevent heart disease	409
Drug and psychosocial treatment of depression	1,699
Coronary artery bypass graft	37,000

[1]Cost per DALY represents an average for low- and middle-income countries, except where noted.

[2]Directly observed treatment short course (DOTS) for epidemic infectious tuberculosis

[3]Refers to South Asia only; includes measures to address life-threatening complications

Source: Disease Control Priorities Project, Using Cost-Effectiveness Analysis for Setting Health Priorities. March 2008. Calculations based on Chapters 2, 16, 26, and 33 of: Jamison DT, Breman G, Measham AR, et al., eds., Disease Control Priorities in Developing Countries. 2nd ed. New York: Oxford University Press; 2006.

Certain methodological aspects and the proposed use of QALYs or similar units in setting health care priorities remain controversial (Arnesen 2000; Gerard 1993; Nord 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 an approach of maximizing QALYs (or other HALY or similar measure) 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 have for the same health care expenditure, making resource allocation decisions based on cost-utility is viewed by some as being biased against the elderly and disabled. A review of such concerns in the context of analyses of cancer care by the UK NICE cited three potential limitations: 1) insufficient sensitivity of the EQ-5D HRQL instrument to changes in health status of cancer patients, 2) diminished validity of certain assumptions of the time-tradeoff method for estimating the values of various health states for patients at the end of life, and 3) relying on using members of the general population rather than actual cancer patients to estimate the values of various health states (Garau 2011).

Certain cost-per-QALY-gained levels have been cited as informal decision thresholds for acceptance of new interventions (e.g., the equivalent of $50,000 or $100,000 per QALY in the wealthy nations); however, analyses of societal preferences suggest much higher levels of acceptance (Braithwaite 2008). Further, without recognition of any limits to providing all potentially beneficial health technologies to all people, such a threshold for the acceptable cost of a QALY has little relevance (Weinstein 2008). Comparisons of the cost per QALY gained from various health care interventions in widespread use can be revealing about how efficient health care systems are in allocating their resources. A continuously updated, detailed set of standardized cost-utility analyses, including tables of cost-utility ratios for many

types of health care interventions, can be found at the *Cost-Effectiveness Analysis Registry,* maintained by the Tufts Medical Center [www.cearegistry.org].

QALYs and other HALYs can enable comparisons among health technologies that have different types of health effects and help to inform resource allocation. Given their relative strengths and weaknesses, these measures are preferable to other measures of health improvement when it is important to make comparisons across diverse interventions. Aside from methodological considerations, their use has been limited by various political and social concerns (Neumann 2010). The relevance of cost-utility analysis for resource allocation depends, at least in part, on how health care is organized and financed.

E. Role of Budget Impact Analysis

Budget-impact analysis determines the impact of implementing or adopting a particular technology or program on a designated budget. The designated budget is generally the responsibility of a particular health care program or authority, e.g., a drug plan or formulary, a hospital, an employer-sponsored health plan, or a regional or national health authority (Mauskopf 2007). It does not necessarily account for the broader economic impact (e.g., societal impact) of implementing or adopting the technology. BIAs can take different forms. For example, a BIA can be conducted simply to determine how much a technology would increase or decrease a particular budget. Or, it could be conducted to determine whether, or to what extent, one or more technologies could be implemented within a fixed (or "capped") budget. BIAs have appeared with increasing frequency during the past decade, along with greater attention to improvement and standardization of their methods, transparency, and reporting, including with respect to the uses of BIA given decision makers' various economic perspectives (Orlewska 2009).

A BIA might incorporate a CEA to determine the most cost-effective combination of technologies that can be implemented subject to a budget constraint. However, a CEA is not a substitute for a BIA; indeed, a CEA may yield an inappropriate finding for making budgetary decisions. For example, a CEA of alternative population screening tests might indicate that technology A is the most cost-effective (e.g., in dollars per case of cancer detected). However, allocating resources efficiently (e.g., maximizing cost-effectiveness) may not be consistent with affordability, i.e., remaining within a fixed budget. As such, a fixed budget amount may be too small to implement technology A for the full designated population, forcing the use of an alternative (less cost-effective) technology, no technology, or a policy to limit the use of technology A to fewer people in order to remain within the fixed budget. **Box V-11** presents a hypothetical example of a BIA in which a budget constraint is contrary to selecting the most cost-effective alternative. The need for a health program (such as a drug formulary) to operate within a particular budget constraint may be contrary to selecting a technology that is cost-effective or even cost-saving for the broader health care system. This is a form of the "silo budgeting" problem, in that each budget is managed independently of other budgets and of the overall health system. The inability to transfer funds across these silos can undermine system-wide efficiency.

Box V-11. Cost Effectiveness of New Intervention 'A' in the Context of a Fixed Budget: A Hypothetical Budget Impact Analysis

Patient subgroup age (years)	Δ Cost per life-year gained (£/life-yr)	Net cost of intervention 'A' over existing treatment (£/patient)	Number of patients per year	Potential budget impact (£/yr)
<45	200,000	500	250	125,000
45-60	75,000	500	1,000	500,000
61-75	25,000	500	1,750	875,000
>75	15,000	500	2,000	1,000,000

How should a fixed annual budget of £500,000 be allocated?

It would be most cost-effective to provide Intervention 'A' to patients age >75 years. However, there is insufficient budget to provide the intervention to all patients in that subgroup. Although there is sufficient budget to provide the intervention to all patients age 45-60 years, this is not the most cost-effective approach.

Source: With kind permission from Springer Science+Business Media: Pharmacoeconomics, Developing guidance for budget impact analysis, 19(6), 2001, 609-21, Trueman P, Drummond M, Hutton J, Figure 1, and any original (first) copyright notice displayed with material.

F. Collecting Cost Data in Clinical Studies

The validity of a cost-related study depends on the sources of the data for costs and outcomes. Increased attention is being given to collection of cost data in more rigorous, prospective studies. The closer integration of economic and clinical studies raises important methodological issues. In order to promote more informed resource allocation for 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 for seeking regulatory approval for marketing. Although an RCT would be expected to yield the most reliable data concerning efficacy of an intervention, 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 2005; Graves 2002; Poe 1995). The prevalence of multinational clinical trials of drugs and other technologies can complicate estimating country-specific treatment effects and cost-effectiveness, given differences in epidemiological factors, health care delivery models, resource use, and other factors (Willke 1998).

References for Chapter V

Arnesen T, Nord E. The value of DALY life: problems with ethics and validity of disability adjusted life years. BMJ. 2000;320(7246):1398. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1117148.

Braithwaite RS, Meltzer DO, King JT Jr, et al. What does the value of modern medicine say about the $50,000 per quality-adjusted life-year decision rule? Med Care. 2008;46(4):349-56.

Briggs A, Clark T, Wolstenholme J, Clarke P. Missing… presumed at random: cost-analysis of incomplete data. Health Econ. 2003;12(5):377-92.

Byford S, Palmer S. Common errors and controversies in pharmacoeconomic analyses. Pharmacoeconomics. 1998;13(6):659-66.

Chapman RH, Berger M, Weinstein MC, Weeks JC, Goldie S, Neumann PJ. When does quality-adjusting life-years matter in cost-effectiveness analysis? Health Econ. 2004;13(5):429-36.

Disease Control Priorities Project, Using Cost-Effectiveness Analysis for Setting Health Priorities. March 2008. Accessed Oct 1, 2013 at: http://www.dcp2.org/file/150/DCPP-CostEffectiveness.pdf.

Drummond MF, Sculpher MJ, Torrance GW, O'Brien BJ, Stoddart GL. Methods for the Economic Evaluation of Health Care Programmes. Third Edition. Oxford: Oxford University Press; 2005.

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.

Eccles M, Mason J. How to develop cost-conscious guidelines. Health Technol Assess. 2001;5(16):1-69.

Garau M, Shah KK, Mason AR, Wang Q, Towse A, Drummond MF. Using QALYs in cancer: a review of the methodological limitations. Pharmacoeconomics. 2011;29(8):673-85.

Gerard K, Mooney G. QALY league tables: handle with care. Health Economics. 1993;2(1):59-64.

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

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(8):735-9.

Greenberg D, Neumann PJ. Does adjusting for health-related quality of life matter in economic evaluations of cancer-related interventions? Expert Rev Pharmacoecon Outcomes Res. 2011;11(1):113-9.

Hjelmgren J, Berggren F, Andersson F. Health economic guidelines--similarities, differences and some implications. Value Health. 2001;4(3):225-50.

Hogness JR,Van Antwerp M. The Artificial Heart: Prototypes, Policies, and Patients. Washington, DC: National Academy Press; 1991. Used with permission from the National Academy of Sciences, courtesy of the National Academies Press, Washington, DC. http://www.nap.edu/openbook.php?record_id=1820.

Husereau D, Drummond M, Petrou S, Carswell C, et al.; CHEERS Task Force. Consolidated Health Economic Evaluation Reporting Standards (CHEERS) statement. Int J Technol Assess Health Care. 2013;29(2):117-22.

IQWiG (Institute for Quality and Efficiency in Health Care). General Methods for the Assessment of the Relation of Benefits to Costs. Version 1.0 – 19/11/2009. Accesed Aug. 1, 2013 at: https://www.iqwig.de/download/General_Methods_for_the_Assessment_of_the_Relation_of_Benefits_to_Costs.pdf.

Mauskopf JA, Sullivan SD, Annemans L, Caro J, et al. Principles of good practice for budget impact analysis: report of the ISPOR Task Force on good research practices--budget impact analysis. Value Health. 2007;10(5):336-47.

Maynard A. Developing the health care market. Econ J. 1991;101(408):1277-86.

McGhan WF, Al M, Doshi JA, Kamae I, Marx SE, Rindress D. The ISPOR Good Practices for Quality Improvement of Cost-Effectiveness Research Task Force Report. Value Health. 2009;12(8):1086-99.

Neuhauser D, Lewicki AM. What do we gain from the sixth stool guaiac? N Engl J Med. 1975;293(5):226-8..

Neumann PJ, Weinstein MC. Legislating against use of cost-effectiveness information. N Engl J Med 2010;363(16):1495-7. http://www.nejm.org/doi/full/10.1056/NEJMp1007168.

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; 16(3): 731-42.

Nord E. The QALY − a measure of social value rather than individual utility? Health Econ. 1994;3(2):89-93.

Orlewska E, Gulácsi L. Budget-impact analyses: a critical review of published studies. Pharmacoeconomics. 2009;27(10):807-27.

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. http://ota-cdn.fas.org/reports/9440.pdf.

Rettig RA, Levinsky NG, eds. Kidney Failure and the Federal Government. Washington, DC: National Academy Press; 1991. http://www.nap.edu/catalog.php?record_id=1818.

Ubel PA, Nord E, Gold M, et al. Improving value measurement in cost-effectiveness analysis. Med Care. 2000;38(9):892-901.

Weinstein MC. How much are Americans willing to pay for a quality-adjusted life year? Med Care. 2008;46(4):343-5.

Weinstein MC, Stason WB. Foundations of cost-effectiveness analysis for health and medical practices. N Engl J Med. 1977;296(13):716-21.

Willke RJ, Glick HA, Polsky D, Schulman K. Estimating country-specific cost-effectiveness from multinational clinical trials. Health Econ. 1998;7(6):481-93.

Wonderling D, Sawyer L, Fenu E, Lovibond K, Laramée P. National Clinical Guideline Centre cost-effectiveness assessment for the National Institute for Health and Clinical Excellence. Ann Intern Med. 2011;154(11):758-65.

VI. DETERMINE TOPICS

Organizations that conduct or sponsor HTAs are subject to limited resources. With the great supply of potential assessment topics, HTA organizations need practical and accountable means of determining what to assess. This chapter considers how assessment programs identify candidate assessment topics and set priorities among these.

A. Identify Candidate Topics

To a large extent, assessment topics are determined, or bounded, by the mission or purpose of an organization. For example, national and regional health plans and other 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), some 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 (people who are eligible for Medicare), physicians, health product companies, health professional associations, or government entities. CMS may request assistance in the form of "evidence reports" or HTAs from a sister agency, AHRQ, which typically commissions these from one of its Evidence-based Practice Centers (part of the AHRQ Effective Healthcare Program).

Apart from requests from CMS, the AHRQ Effective Healthcare Program solicits topic nominations from the public. Its online topic nomination form requests information about: the health care intervention of interest; any specific comparator(s); patient groups and subgroups affected; health benefits or outcomes; risks/harms/side effects; which (if any) of 14 priority health conditions/diseases are involved; which (if any) of six priority populations is involved; which (if any) federal health program (e.g., Medicare, Medicaid) is involved; why the topic is important; whether the question represents uncertainty for clinicians or policymakers; stakeholders in this topic; how the findings will be used; technical experts relevant to the topic; and any supporting documentation.

For the UK National Institute for Health and Care Excellence (NICE), topics are not determined internally, but are referred from the UK Department of Health. Topics are selected based on such factors as burden of disease, impact on resources, and whether there is inappropriate variation in practice across the UK (NICE 2013).

For the Cochrane Collaboration, potential topics generally arise from members of the more than 50 review groups, who are encouraged to investigate topics of interest to them, subject to the agreement of their review groups. However, there is as yet no standard or common priority-setting process used across the Cochrane Collaboration (Nasser 2013).

Horizon Scanning

The demand for early information about new, emerging, and existing health care interventions and related trends has prompted the development and evolution of "horizon scanning" functions (Carlsson 1998; Douw 2003; Harper 1998; Packer 2012). Horizon scanning is intended to serve multiple purposes, including, e.g., the following:

- Identify potential topics for HTA and information for setting priorities among them

- Identify areas of technological change
- Anticipate and identify new indications or uses of technologies
- Identify variations in use of technologies
- Identify inappropriate use of technologies, including over-use, under-use, and improper use
- Forecast the health and economic impacts of technologies
- Identify levels in improvement in effectiveness in relation to additional costs that would demonstrate the cost-effectiveness of a new technology
- Anticipate potential social, ethical, or legal implications of technologies
- Plan data collection to monitor adoption, diffusion, use, and impacts of technologies
- Enable health care providers, payers, and patients to plan for, adapt to, and manage technological change, including "rising"/emerging technologies and "setting" (becoming obsolescent) technologies (for potential disinvestment)

Most horizon scanning programs generate rapidly completed, brief descriptions of new or emerging technologies and their potential impacts. Certainly, there are tradeoffs inherent in using early information that may be incomplete or unreliable as opposed to waiting long enough for more definitive information that opportunities to benefit from it may have passed. HTA programs have made use of horizon scanning in important ways. While the major thrust of horizon scanning has been to identify "rising" (new and emerging) technologies that eventually may merit assessment, horizon scanning can identify "setting" technologies that may be outmoded, superseded by newer ones, and candidates for disinvestment (Henshall 2012). In either case, horizon scanning provides an important input into setting assessment priorities.

Examples of national and international horizon scanning programs are:

- AHRQ Healthcare Horizon Scanning System (US)
- Canadian Network for Environmental Scanning in Health (CNESH, secretariat at Canadian Agency for Drugs and Technologies in Health [CADTH])
- Centers for Medicare and Medicaid Services Council for Technology and Innovation (US)
- ECRI Institute Health Technology Forecast
- EuroScan International Network (International Information Network on New and Emerging Technologies) (secretariat at NHSC, UK)
- Health Policy Advisory Committee on Technology (HealthPACT, Australia and New Zealand)
- National Evidence-based Healthcare Collaborating Agency Horizon Scanning Center for Innovative Global Health Technology (NECA H-SIGHT, South Korea)
- National Horizon Scanning Centre (NHSC, UK)

For example, the purposes of EuroScan, a collaborative network involving about 20 HTA agencies, are to collect and share information on innovative health care technologies to support decision making and adoption and use of effective, useful, and safe technologies, as well as to provide a forum for sharing

and developing methods for early identification and assessment of new and emerging technologies and predicting their potential impacts.

The Canadian Network for Environmental Scanning in Health (CNESH) identifies information on new, emerging, or new applications of health technologies and shares this information across Canada. It also develops and promotes methods for identifying, filtering, and setting priorities among new or emerging health technologies. CNESH produces a "top 10" list of new and emerging health technologies in Canada.

The Health Policy Advisory Committee on Technology (HealthPACT) provides evidence-based advice about potentially significant new and emerging technologies to health departments in Australia and New Zealand. This supports information exchange and evaluation of the potential impact of these technologies on those national health systems, including informing financing decisions and the managed introduction of new technologies. HealthPACT produces *New and Emerging Health Technology Reports* and *Technology Briefs*.

The AHRQ Healthcare Horizon Scanning System provides AHRQ with a systematic process to identify and monitor target technologies and create an inventory of those that have the highest potential for impact on clinical care, the health care system, patient outcomes, and costs. This system is also intended to serve as a tool for the public to identify and find information on new health care technologies (ECRI Institute 2013).

EUnetHTA developed a web-based Planned and Ongoing Projects (POP) database to enable HTA agencies to share information about planned and ongoing projects at each agency, with the aim of avoiding duplication and encouraging collaborative efforts (EUnetHTA 2013).

A 2013 systematic review of international health technology horizon scanning activity identified 23 formal programs, most of which are members of EuroScan, along with a variety of other less structured horizon scanning functions of government and private sector organizations. Although the formal programs had somewhat varying emphases on target technologies, time horizons of interest, and methods of scanning and assessment, they generally shared the main functions of identification and monitoring of technologies of interest and evaluation of potential impacts of technologies (Sun 2013).

As shown in **Box VI-1**, a considerable variety of electronic bibliographic databases, newsletters, regulatory documents, and other sources provide streams of information pertaining to new and emerging health care interventions. The *AHRQ Horizon Scanning Protocol and Operations Manual* provides a detailed list of databases, news sources, and other information sources for horizon scanning, as well as search filters for horizon scanning of PubMed and Embase (ECRI Institute 2013).

Box VI-1. Information Sources for New and Emerging Health Care Interventions

- Large bibliographic databases (e.g., *PubMed, Embase, SciSearch)*
- Specialized bibliographic databases (e.g., *CINAHL, PEDro, PsycINFO)*
- Databases of ongoing research and results (e.g., *ClinicalTrials.gov, HSRProj)*
- Priority lists and forthcoming assessments from HTA agencies and vendors
- Cochrane Collaboration protocols (plans for forthcoming/ongoing systematic reviews)
- Trade publications (e.g., *The Pink Sheet, The Gray Sheet, In Vivo, Medtech Insight, Pharmaceutical Approvals Monthly, Medical Device Daily, GenomeWeb Daily News, Telemedicine and e-Health)*
- General news *(e.g., PR Newswire, New York Times, Wall Street Journal)*
- General health care/medical journals and specialty health care/medical journals
- Health professions and industry news (e.g., *American Health Line, Medscape, Reuters Health Industry Briefing, Reuters Health Medical News)*
- Conference abstracts and proceedings of health professions organizations, health industry groups
- Technology company web sites
- Industry association (e.g., AdvaMed, BIO, PhRMA) sites (e.g., *AdvaMed SmartBrief, PhRMA New Medicines Database)*
- Market research reports (e.g., Frost & Sullivan; GlobalData; IHS Global Insight; Thomson Reuters)
- Regulatory agency announcements of market approvals, other developments for new pharmaceuticals, biological, and devices (e.g., FDA Advisory Committee Alerts, FDA Approval Alerts, FDA Drug Daily Bulletin, FDA Device Daily Bulletin)
- Adverse event/alert announcements (e.g., from FDA MedWatch, NIH Clinical Alerts and Advisories, United States Pharmacopeia)
- Payer policies, notifications (e.g., *CMS Updates to Coverage Pages, Aetna Clinical Policy Bulletins)*
- Reports and other sources of information on significant variations in practice, utilization, or payment policies (e.g., *The Dartmouth Atlas)*
- Special reports on health care trends and futures (e.g., from Institute for the Future Health Horizons Program; Institute for Healthcare Improvement)

B. Setting Assessment Priorities

Some assessment programs have explicit procedures for setting priorities; others set priorities only in an informal or *ad hoc* manner. 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.

Most assessment programs have criteria for topic selection, although these criteria are not always explicit. For example, 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 shown in **Box VI-2**.

Box VI-2. Examples of HTA Selection Criteria Used in Setting Assessment Priorities

- High individual burden of morbidity, mortality, or disability
- High population burden of morbidity, mortality, or disability
- High unit/individual cost of a technology or health problem
- High aggregate/population cost of a technology or health problem
- Substantial variations in practice
- Unexpected adverse event reports
- Potential for HTA findings to have impact on practice
- Potential for HTA findings to have impact on patient outcomes or costs
- Available findings not well disseminated or adopted by practitioners
- Need to make regulatory decision
- Need to make payment decision (e.g., provide coverage or include in health benefits)
- Need to make a health program acquisition or implementation decision
- Recent or anticipated "breakthrough" scientific findings
- Sufficient research findings available upon which to base HTA
- Feasibility given resource constraints (funding, time, etc.) of the assessment program
- Public or political demand
- Scientific controversy or great interest among health professionals

The timing for undertaking an assessment may depend on the availability of evidence. For example, the results of a recently completed major RCT or meta-analysis may challenge current 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.

As noted in section II. Fundamental Concepts, the demand for HTA by health care decision makers has increasingly involved requests for faster responses to help inform emergent regulatory, payment, or acquisition decisions. The urgency of such a request may raise the priority of an assessment topic and prompt an HTA organization to designate it for a more focused, less-comprehensive "rapid HTA." See discussion of **rapid HTA** in chapter X.

Systematic priority-setting processes typically include such steps as the following (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 VI-3** shows a quantitative model for priority setting. The Cochrane Collaboration has used a more decentralized approach in which review groups use a range of different priority-setting systems (Clarke 2003; Nasser 2013). Starting with topics suggested by their members, many Cochrane Collaboration review groups have 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 have been offered to potential reviewers who might be interested in preparing and maintaining relevant reviews in these areas.

Box VI-3. 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.

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 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.

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, do not address priority setting in the context of a research portfolio, and 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.

C. 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. This understanding might not be established at the outset of the assessment; it may take more probing, discussion and clarification.

The intended users or target audiences of an assessment should affect the content, presentation, and dissemination of results of the HTA. Clinicians, patients, politicians, researchers, hospital managers, company executives, and others have different interests and levels of expertise. They tend to have varying 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.

1. Problem Elements

There is no single correct way to state an assessment problem. The elements typically include specifying most or all of the following:

- Health problem of interest
- Patient population (including subgroups as appropriate)
- Technology of interest
- Comparator(s)
- Setting of care
- Provider/clinician delivering the intervention(s)
- Properties, impacts, or outcomes

- Timeframe, duration, or follow-up period
- Timeframe, duration, or follow-up period
- Study design or type of evidence/data to be included in the HTA
- Target audiences for the HTA findings

One commonly used framework is known as PICOTS (sometimes only PICO or PICOT): **P**opulation, **I**ntervention(s), **C**omparator(s), **O**utcome(s), **T**iming, and **S**tudy design (Counsell 1997). This framework can be used for describing individual studies or HTAs that might examine evidence from multiple studies. For example, a basic specification of one assessment problem would be the following. (This example uses some characteristics of a particular RCT [Stewart 2005].)

- **Population:** males and females age 55-75 years with mild hypertension, i.e., diastolic blood pressure 85-99 mm Hg, systolic blood pressure 130-159 mm Hg; no other serious health problems
- **Intervention:** standardized, moderate exercise program (aerobic and resistance training)
- **Comparator:** usual physical routine and diet
- **Outcomes:** changes in: general and abdominal obesity, systolic blood pressure, diastolic blood pressure, aerobic fitness, aortic stiffness (measured as aortofemoral pulse-wave velocity)
- **Timing:** 6-24 months
- **Study design:** randomized controlled trials

2. Analytic Frameworks for Presenting HTA Problems

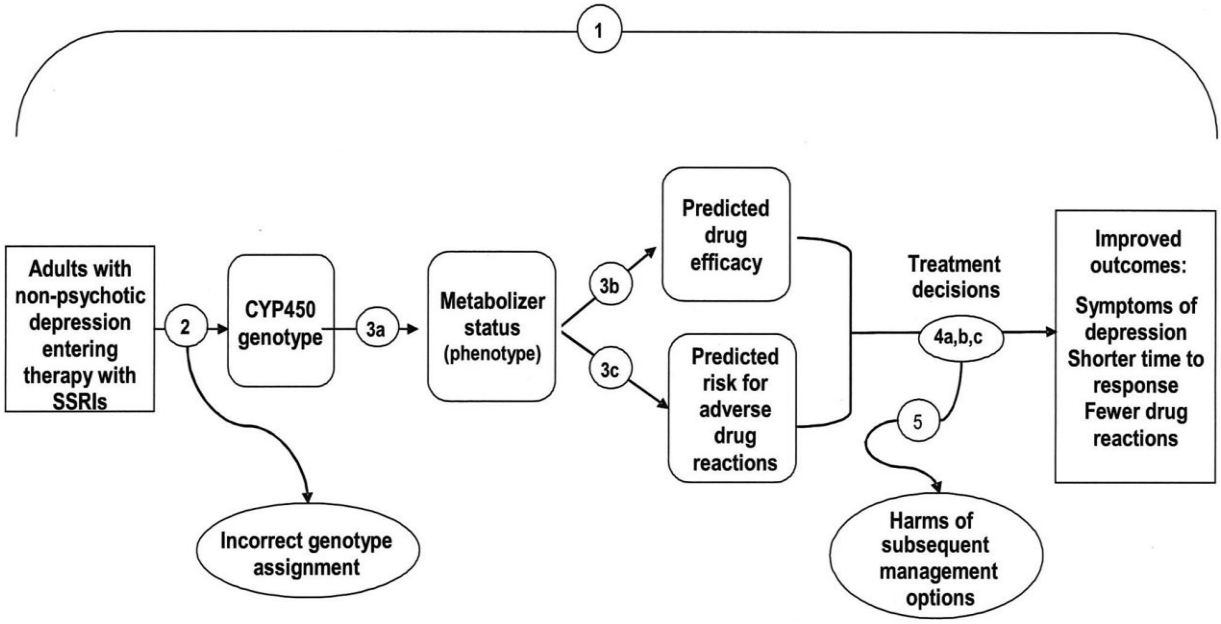
A useful graphical means of presenting an assessment problem is an "analytic framework," sometimes known as a "causal pathway." Analytic frameworks depict direct and indirect relationships between interventions and outcomes. Although often used to present clinical interventions for health problems, they can be used as well for other types of interventions in health care.

Analytic frameworks provide clarity and explicitness in defining the key questions to be addressed in an HTA, and draw attention to important relationships for which evidence may be lacking. They can be useful tools to formulate or narrow the focus of an assessment problem. For a clinical problem, an analytic framework 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 involves a single intervention for narrowly defined indications and outcomes, these frameworks 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 subgroups), analytic frameworks can be detailed.

An example of an analytic framework of the impact of a diagnostic test on health outcomes is shown in **Box VI-4**. In particular, this framework presents a series of key questions intended to determine whether testing for a particular genotype in adults with depression entering treatment with selective serotonin reuptake inhibitors (SSRIs) will have an impact on health outcomes. The framework includes an overarching key question about the impact of the test on outcomes, as well as a series of linked key questions about the accuracy of the test; its ability to predict metabolism of SSRIs, efficacy of SSRIs, and

risk of adverse drug reactions; the test's impact on treatment decisions; and the ultimate impact on health outcomes.

Box VI-4. Analytic Framework: CYP450 Genotype Testing for Selective Serotonin Reuptake Inhibitors



The numbers above correspond to the following key questions:

1. Overarching question: Does testing for cytochrome P450 (*CYP450*) polymorphisms in adults entering selective serotonin reuptake inhibitor (SSRI) treatment for nonpsychotic depression lead to improvement in outcomes, or are testing results useful in medical, personal, or public health decision-making?

2. What is the *analytic validity* of tests that identify key *CYP450* polymorphisms?

3. *Clinical validity*: **a:** How well do particular *CYP450* genotypes predict metabolism of particular SSRIs? **b:** How well does *CYP450* testing predict drug efficacy? **c:** Do factors such as race/ethnicity, diet, or other medications, affect these associations?

4. *Clinical utility*: **a:** Does *CYP450* testing influence depression management decisions by patients and providers in ways that could improve or worsen outcomes? **b:** Does the identification of the *CYP450* genotypes in adults entering SSRI treatment for nonpsychotic depression lead to improved clinical outcomes compared to not testing? **c:** Are the testing results useful in medical, personal, or public health decision-making?

5. What are the harms associated with testing for *CYP450* polymorphisms and subsequent management options?

Source: Teutsch SM, Bradley LA, Palomaki GE, et al. The Evaluation of Genomic Applications in Practice and Prevention (EGAPP) initiative: methods of the EGAPP Working Group. Genet Med. 2009;11(1):3-14.

D. 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 HTA findings and their 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 (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, primary vs. secondary prevention, etc.)
- Protocols or care pathways in which the technology is used 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., quality of life, types of costs)
- 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)
- Cost (or price) of a technology or its comparators or of the associated episode or course of care
- Adoption or use of guidelines, payment policies, or other decisions 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 in the late 1970s of percutaneous transluminal coronary angioplasty (PTCA, approved by the US FDA in 1980), its clinical role in relation to 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, bare metal and drug-eluting 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 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 (Richardson 2002; Zhou 2013).

While HTA programs cannot avoid the moving target problem, they can manage and be responsive to it. **Box VI-5** lists approaches for managing the moving target problem.

Box VI-5. Managing the Moving Target Problem

- Recognize that HTA must have the capacity to revisit topics as needed, whether periodically (e.g., every two or five years) or as prompted by important changes since preparation of the original HTA report.
- Document in HTA reports the information sources, assumptions, and processes used. This "baseline" information 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 type of drug, patient adherence rates, 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, or other developments that might prompt the need for reassessment.
- Have or subscribe to a horizon scanning or monitoring function to help detect significant changes in technologies, how they are used, or 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 to, 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 (Olson 2001). More recently, an assessment by the US Preventive Services Task Force of the same issue prompted re-examination of available evidence, the process used by this task force to arrive at its findings, how the findings were transmitted to the public, and how the findings were interpreted by patients and clinicians (Thrall 2010; US Preventive Services Task Force 2009).

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 commentaries, may signal trends that 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 useful to update it, but to conduct an entirely new assessment.

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.

A method for detecting signals for the need to update systematic reviews was validated for a set of reports produced by the AHRQ Comparative Effectiveness Review program. This method for determining whether a report needed to be updated involved applying the literature search strategy for the original systematic review to five leading general interest medical journals plus four-to-six specialty journals most relevant to the topic. The method also involved providing a questionnaire to experts in the field that requested them to indicate whether the conclusions in the original review were still valid and, if not, to identify any relevant new evidence and citations. This information was used to identify reports to be updated. After the new (i.e., updated) reports were completed, the researchers conducted a systematic comparison of the conclusions of the original and new reports, and found that the determination of priority for updating the original reports was a good predictor of actual changes to conclusions in the updated reports (Shekelle 2014).

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 contend 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).

References for Chapter VI

Carlsson P, Jørgensen T. Scanning the horizon for emerging health technologies. Int J Technol Assess Health Care. 1998;14(4):695-704.

Counsell C. Formulating questions and locating primary studies for inclusion in systematic reviews. Ann Intern Med. 1997;127(5):380-7.

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

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.

ECRI Institute. AHRQ Healthcare Horizon Scanning System Protocol and Operations Manual: January 2013 Revision. (Prepared by ECRI Institute under Contract No. 290-2010-00006-C.) Rockville, MD: Agency for Healthcare Research and Quality. August 2013. Accessed November 1, 2013 at: http://effectivehealthcare.ahrq.gov/ehc/products/393/886/AHRQ-Healthcare-Horizon-Scan-Protocol-Operations-Manual_130826.pdf.

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

EUnetHTA (European network for Health Technology Assessment). EUnetHTA POP Database. Accessed Sept. 1, 2013 at: http://eunethta.dimdi.de/PopDB.

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.

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(4):652-62.

Henshall C, Schuller T, Mardhani-Bayne L. Using health technology assessment to support optimal use of technologies in current practice: the challenge of "disinvestment". Int J Technol Assess Health Care. 2012;28(3):203-10.

Lara ME, Goodman C, eds. National Priorities for the Assessment of Clinical Conditions and Medical Technologies. Washington, DC: National Academy Press; 1990. http://www.nap.edu/catalog.php?record_id=1529.

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):i-vi, 1-149.

Nasser M, Welch V, Tugwell P, Ueffing E, et al. Ensuring relevance for Cochrane reviews: evaluating processes and methods for prioritizing topics for Cochrane reviews. J Clin Epidemiol. 2013;66(5):474-82.

National Institute for Health and Care Excellence. Medical Technologies Evaluation Programme. Accessed Dec. 1, 2013 at: http://www.nice.org.uk/aboutnice/whatwedo/aboutmedicaltechnologies/medicaltechnologiesprogramme.jsp.

Olsen O, Gøtzsche PC. Cochrane review on screening for breast cancer with mammography. Lancet. 2001;358(9290):1340-2.

Packer C, Gutierrez-Ibarluzea I, Simpson S. The evolution of early awareness and alert methods and systems. Int J Technol Assess Health Care. 2012;28(3):199-200.

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

Richardson P, Hideshima T, Anderson K. Thalidomide: emerging role in cancer medicine. Annu Rev Med. 2002;53;629-57.

Sassi, F. Setting priorities for the evaluation of health interventions: when theory does not meet practice. Health Policy. 2003;63(2):141-54.

Shekelle P, Eccles MP, Grimshaw JM, Woolf SH. When should clinical guidelines be updated? BMJ. 2001;323(7305):155-7. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1120790.

Shekelle PG, Motala A, Johnsen B, Newberry SJ. Assessment of a method to detect signals for updating systematic reviews. Syst Rev. 2014;3:13. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3937021/pdf/2046-4053-3-13.pdf.

Shekelle PG, Ortiz E, Rhodes S, Morton SC, et al. Validity of the Agency for Healthcare Research and Quality clinical practice guidelines: how quickly do guidelines become outdated? JAMA. 2001;286(12):1461-7.

Stewart KJ, Bacher AC, Turner KL, Fleg JL, Hees PS, et al. Effect of exercise on blood pressure in older persons: a randomized controlled trial. Arch Intern Med. 2005;165(7):756-62.

Sun F, Schoelles K. A systematic review of methods for health care technology horizon scanning. (Prepared by ECRI Institute under Contract No. 290-2010-00006-C.) AHRQ Publication No. 13-EHC104-EF. Rockville, MD: Agency for Healthcare Research and Quality; August 2013. http://effectivehealthcare.ahrq.gov/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productID=1679.

Teutsch SM, Bradley LA, Palomaki GE, et al. The Evaluation of Genomic Applications in Practice and Prevention (EGAPP) initiative: methods of the EGAPP Working Group. Genet Med. 2009;11(1):3-14. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2743609.

Thrall JH. US Preventive Services Task Force recommendations for screening mammography: evidence-based medicine or the death of science? J Am Coll Radiol. 2010;7(1):2-4.

US Preventive Services Task Force. Screening for breast cancer: U.S. Preventive Services Task Force recommendation statement. Ann Intern Med. 2009;151(10):716-26, W-236.

Zhou S, Wang F, Hsieh TC, Wu JM, Wu E. Thalidomide-a notorious sedative to a wonder anticancer drug. Curr Med Chem. 2013;20(33):4102-8.

VII. RETRIEVE EVIDENCE

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, and to do so efficiently and in a timely manner. For a new or emerging topic, this information may be sparse and difficult to find. For many topics, the evidence is readily available, yet profuse 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.

A. 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 biomedical, social, and other sciences more broadly. Multiple sources should be searched to increase the likelihood of retrieving relevant reports. The types of sources that may be useful for HTA include:

- Journal and other published literature databases
- Clinical and administrative databases
- Practice guideline databases
- 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

Some of these are traditional peer-reviewed, published sources; others are recognized as "grey" literature, as described below.

Across the hundreds of publicly available electronic databases of potential relevance to HTA are various general types. Among the major categories, **bibliographic databases** have indexed citations (or "records") for journal articles and other publications. **Factual databases** provide substantive content in the form of, e.g., guidelines for diagnosis and treatment; compendia of information about pharmacologic properties, indications, contraindications, and other information about drugs; 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. Various other useful listings of HTA information resources and related tools have been prepared, e.g., *Health Technology Assessment on the Net International* [http://www.ihe.ca/documents/HTA%20on%20the%20Net%202013.pdf] (Institute of Health Economics 2013), *Grey Matters: A Practical Search Tool for Evidence-Based Medicine* [http://www.cadth.ca/en/resources/finding-evidence-is/grey-matters] (CADTH 2014), and the HTAi Vortal [http://vortal.htai.org].

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 [http://www.nlm.nih.gov/pubs/factsheets/dif_med_pub.html]. In addition, there are certain specialized or more focused databases in such areas as ongoing clinical trials and their results (e.g., ClinicalTrials.gov of NLM) (Zarin 2011), ongoing health services research (e.g., HSRProj), and practice guidelines (e.g., the National Guideline Clearinghouse of AHRQ and the International Guideline Library of the Guidelines International Network).

The Cochrane Collaboration is an international organization that prepares, maintains and disseminates systematic reviews of RCTs (and other evidence when appropriate) of interventions for many diseases and conditions. Nearly 7,000 systematic reviews have been produced by more than 50 Cochrane review groups in such areas as acute respiratory infections; breast cancer; dementia and cognitive improvement; hypertension; infectious diseases; pregnancy and childbirth; and pain, palliative, and supportive care. These reviews are made available via the Cochrane Library, which also includes certain databases and registers produced by the Cochrane Collaboration as well as some produced by other organizations. The Cochrane Central Register of Controlled Trials includes more than 660,000 records. The Database of Abstracts of Reviews and Dissemination (DARE) and the NHS Economic Evaluation Database (NHS EED) are produced by the Centre for Reviews and Dissemination (CRD) at the University of York. The HTA Database is produced by CRD with information supplied by members of the International Network of Agencies for Health Technology Assessment (INAHTA) and other contributing HTA programs.

The selection of sources for literature searches should depend on the evidence questions and related matters of content of the HTA as well as pertinent time and resource constraints. Most searches are likely to involve one or more of the large bibliographic databases, which have their relative merits and can be complementary in some instances (Wilkins 2005). These sources can be supplemented by specialized bibliographic databases relevant to the topic, such as those devoted to mental health, allied health, or health economics. Further, there are strategies and search filters for identifying particular types of studies in such bibliographic databases, such for records on economic evaluations (Glanville 2010). The selection of other databases is likely to differ by purpose of the inquiry, e.g., horizon scanning, ascertaining regulatory or payment status of technologies, comprehensive systematic review, or identifying literature in particular clinical areas. The yield from searching the more specialized sources in addition to the large, generic ones varies (Royle 2003).

Some of the most commonly used bibliographic and factual databases in HTA are listed in **Box VII-1**. Many additional bibliographic and factual databases that can be used in HTA, including additional generic ones and others that are more specialized, are listed in **Box VII-2**.

Box VII-1. Some Core Sources: Bibliographic and Factual Databases for HTA

Among the most commonly used are the following:

- *PubMed* (including *MEDLINE):* citations for life sciences journal articles and online books (NLM)
- *Embase:* citations for biomedical journal articles (Elsevier)
- *Cochrane Database of Systematic Reviews:* systematic reviews of controlled trials on hundreds of clinical topics
- *Cochrane Central Register of Controlled Trials:* 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, appraised by NHS Centre for Reviews and Dissemination (Cochrane Library)
- *Health Technology Assessment Database:* records of completed and ongoing projects of members of INAHTA and other organizations (Cochrane Library)
- *NHS Economic Evaluation Database (NHS EED):* abstracts and other information about published economic evaluations of health care interventions (Cochrane Library)

Box VII-2. Additional Sources: Bibliographic and Factual Databases for HTA

The variety of additional generic and more specific sources includes the following:

- *Other NLM sources:*
  - *Bookshelf:* free full-text access to online biomedical books
  - *ClinicalTrials.gov:* registry of ongoing, and summary results of, clinical trials and observational studies funded by U.S. government and private sources conducted in the U.S. and around the world
  - *HSRProj:* ongoing health services research and public health projects
  - *HSRR (Health Services and Sciences Research Resources):* information about research datasets, instruments/indices, software
  - *HSTAT:* full text of US technology assessments, evidence reports, CER reports, clinical practice guidelines, consensus development reports, methods, reports, etc.
  - *PubMed Health:* systematic reviews of clinical trials, with related information for clinicians and patients
- *ACP Journal Club:* selected studies and systematic reviews on clinical and economic topics for attention of clinicians, with "value added" abstracts and commentary
- *AMED (Allied and Complementary Medicine):* citations to journals in physiotherapy, occupational therapy, rehabilitation, speech therapy, podiatry, complementary medicine, palliative care (British Library)
- *Bandolier:* journal of summaries of evidence-based health care (University of Oxford, UK)
- *BIOSIS Previews:* citations from life sciences journals and reports, reviews, meetings (Thomson Reuters)
- *Campbell Collaboration Library of Systematic Reviews:* systematic reviews of the effects of social interventions, including education, crime and justice, and social welfare (Campbell Collaboration)
- *CEA Registry:* database of standardized cost-utility analyses (Tufts Medical Center)
- *CINAHL*: citations for nursing and allied health literature (Cinahl Information Systems)
- *CDC Wonder*: gateway to epidemiological and public health reports, other information, and data (CDC, US)
- *Cochrane Methodology Register:* bibliography of articles and books on research synthesis (Cochrane Library)
- *ERIC (Educational Resources Information Center):* citations of education information, including in health, for journals, books, research syntheses, technical reports, conferences, policy papers, etc. (US Dept. Education)
- *Evidence Base:* evidence briefing papers, reviews on state of existing evidence, research gaps; research recommendations (NICE, UK)
- *Evidence for Policy and Practice Information and Coordinating Centre (EPPI-Centre):* databases of health promotion: *Trials Register of Promoting Health Interventions (TRoPHI), Bibliomap, Database of Promoting Health Effectiveness Reviews (DoPHER)* (Social Science Research Unit, UK)
- *EVIPNet Virtual Health Library:* for evidence-informed policymaking, particularly in low- and middle-income countries (WHO)
- *Google Scholar:* scholarly literature, with citation information
- *Health Economic Evaluations Database (HEED):* records and bibliographies of economic evaluations of health care interventions, reviewed by health economists (Wiley InterScience)
- *Health Systems Evidence:* syntheses of research evidence about governance, financial and delivery arrangements within health systems (McMaster University, Canada)
- *International Guideline Library:* guidelines, evidence reports developed or endorsed by G-I-N member organizations (Guidelines International Network)
- *National Guideline Clearinghouse (NGC):* evidence-based clinical practice guidelines (AHRQ, US)
- *NHS Evidence:* portal to organizations and other sources providing health and social care evidence and best practices; most but not limited to UK (NICE, UK)
- *OTSeeker:* abstracts of systematic reviews and RCTs (appraised and rated) on occupational therapy
- *PAIS International:* citations of policy information, including for health and social conditions, demography, from journals, books, government documents, research reports, grey literature, conferences, etc. (ProQuest)
- *PEDro (Physiotherapy Evidence Database):* citations for RCTs, systematic reviews, and clinical practice guidelines in physiotherapy (University of Sydney, Australia)
- *PsycINFO:* citations of psychological literature (American Psychological Association)
- *PROSPERO:* prospective registry of systematic reviews in health and social care (NHS Centre for Reviews and Dissemination)
- *REHABDATA:* citations for reports, articles, other documents on disability and rehabilitation (US National Rehabilitation Information Center)
- *SciSearch:* citations for journals in science, technology, biomedicine, and related disciplines, including all records of *Science Citation Index* (Dialog)
- *Scopus:* citations for peer-reviewed journals, conference proceedings, books, patents in science, medicine, social sciences, other areas; also patent records (Elsevier)
- *SUMSearch 2:* integrated search of original studies, systematic review, and practice guidelines from PubMED, DARE, and NGC (University of Kansas, US)
- *Systematic Review Data Repository (SRDR):* archive of systematic reviews and their data, and data extraction tool (AHRQ, US)
- *TRIP Database:* integrated search tool of multiple sites to identify clinical evidence for practice (TRIP Database, Ltd.)
- *UpToDate®:* evidence-based, physician-authored clinical decision support resource (Wolters Kluwer)
- *Web of Science:* citations for journals and conference proceedings in life sciences, social sciences, other fields (Thomson Reuters)

B. Grey Literature

Beyond the traditional sources of published literature, much valuable information is available from the "grey" literature. (Some or all of this is also known as "fugitive" or "ephemeral" literature.) The concept of grey literature is evolving; some define it as any literature except peer-reviewed journals and books cited in the main bibliographic databases such as MEDLINE or Embase. A contemporary definition (not limited to HTA) is as follows:

> *Grey literature stands for manifold document types produced on all levels of government, academics, business and industry in print and electronic formats that are protected by intellectual property rights, of sufficient quality to be collected and preserved by library holdings or institutional repositories, but not controlled by commercial publishers, i.e., where publishing is not the primary activity of the producing body (Schöpfel 2010).*

For HTA, grey literature is found in such sources as:

- Health and economic statistical reports
- Regulatory documents
- Coverage policies of health authorities and insurance plans
- Drug and biologics compendia
- Health professional association reports and guidelines
- Government monographs
- Industry (e.g., life sciences company) reports
- Market research reports for particular health care sectors/industries
- Policy and research institute studies
- Reports of special expert panels and commissions
- Conference proceedings
- Technical specifications and standards
- Working papers
- Memoranda

Much of the non-proprietary grey literature and some proprietary (for a fee) grey literature are accessible via the Internet.

The main concern with using grey literature in HTA is that it is usually not subject to **peer review**, and therefore may be subject to certain biases that the peer review process is designed to diminish. The benefits of including grey literature in any particular literature search are highly variable (Dundar 2006). In some instances, the available peer-reviewed sources may suffice, with little to be gained by expanding a search to include grey sources. In other instances, such as when the available peer-reviewed sources are few, outdated, or have gaps or suspected biases (e.g., publication bias), certain grey sources may strengthen the body of evidence to be considered in HTA. As such, plans for conducting literature searches should weigh the pros and cons of including grey literature, and the inclusion and exclusion criteria for searches should be designed and implemented accordingly.

C. Publication Bias

As described in chapter III, various forms of **bias** can affect the validity of evidence used in HTA. One reason for careful planning and conduct of search strategies for HTA is to recognize and minimize the effects of **publication bias**. Studies of the composition of the biomedical research literature have found imbalances in the publication patterns of methodologically sound studies (Chalmers 1990), including in HTA (Song 2010). For instance, positive studies (i.e., that find statistically significant effects of a technology) are more likely than negative studies (i.e., that find no effects of a technology) to be published in peer-reviewed journals (Dickersin 1993, 1997). A study sponsored by a company or other organization with a financial or other interest in the results may be less likely to be submitted for publication if the findings are not favorable to the interests of that organization. RCTs and other clinical trials conducted for market approval (e.g., by the US FDA) often are not published for proprietary reasons (MacLean 2003). Some research indicates that, among studies of health technologies that are published, smaller studies tend to report positive results more frequently than larger ones (Agema 2002). Publication bias is not the same as **reporting bias**, which refers to differential (e.g., between treatment and control groups) or incomplete reporting by investigators of findings in individual studies. (See further discussion in chapter III.)

Multiple analyses have found that positive studies are more likely to be published in English-language journals, reported in multiple publications, reported for late-phase drug trials, and cited in other articles (Easterbrook 1991, Gøtzsche 1989; Hall 2007). 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). Certain forms of bias may be linked to the clinical area under study. A 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).

Bias in selection of studies used in HTA may arise when a literature searching strategy (or protocol) excludes reports by language (i.e., **language bias**), when it draws only on MEDLINE or other major bibliographic databases, when it includes only reports in the peer-reviewed literature, or when it excludes studies of lesser (or poorly reported) methodological quality (Jüni 2002). The recent global trend toward publication in English-language journals may be reducing the impact of language bias (Galandi 2006; Higgins 2011).

Prospective, systematic literature searching protocols for HTA should seek to identify relevant evidence beyond traditional published sources where possible. This should include circumstances in which potential conflicts of interest might affect submission or publication of clinical trial reports or other forms of evidence.

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. This can have implications for the timing of conducting HTA and may be of particular importance when the number of relevant studies is small (Hopewell, Clarke 2007; Ioannidis 1998).

While the validity of an HTA is likely affected by 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). Even so, exclusion of non-English-language studies may result in higher risk of bias in some areas of research, e.g., complementary and alternative medicine (Song 2010). Lowering the standard of methodological quality for inclusion of published studies in an HTA may bias the findings if these included studies tend to report positive findings more often that higher-quality studies.

Recognition of publication bias in peer-reviewed literature may increase the acceptance of grey literature, to the extent that the grey literature can contribute to a more representative capture of the available relevant evidence pertaining to a given topic. An analysis of the impact of grey literature in meta-analyses of RCTs found that published trials tended to be larger and report greater treatment effects than RCTs reported in the grey literature (i.e., abstracts and unpublished data). The analysis also reported that there was limited evidence to show whether grey RCTs were of poorer methodological quality than published trials (Hopewell, McDonald 2007).

Publication bias can be diminished by prospective registration of clinical trials, as in ClinicalTrials.gov; adherence to guidelines for reporting clinical trials and other types of studies in journals; efforts to locate relevant unpublished studies and non-English language studies for literature searches; and certain statistical techniques (e.g., "funnel plots") to detect potential publication bias (Song 2010; Sterne 2011).

The emergence, validation, and ongoing improvement of guidelines and checklists for reporting research is helping to strengthen the quality of the biomedical and health services research literature, including regarding aspects of publication bias, for HTA reports. Over the last decade in particular, such guidelines have been developed for published reports of diverse study types relevant to HTA. As detailed in chapter IV, some of these include CONSORT (Consolidated Standards of Reporting Trials), PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses), and STROBE (Strengthening the Reporting of OBservational Studies in Epidemiology). In addition to guiding the reporting of completed research, researchers' knowledge of such guidelines and related standards is improving the design and conduct of research. An extensive list of these is maintained in *Research Reporting Guidelines and Initiatives* by NLM. See: http://www.nlm.nih.gov/services/research_report_guide.html.

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.

D. Help for Searchers

Given the great number of bibliographic, factual, and other databases and the variety in their content, means of access, controlled vocabularies and search commands, it is advisable to involve librarians and health information specialists in planning and conducting literature searches. These experts can be especially helpful when planning which databases to search, designing search strategies to address evidence questions, and reviewing search strategies (Institute of Medicine 2011). Many such experts communicate via international networks. One such network of HTA information specialists who work with HTA agencies and related types of organizations around the world comprise the Health Technology Assessment International (HTAi) Interest Sub-Group on Information Resources, which has helped to extend the capacity, processes, methods, and collaboration in HTA. The Cochrane Library Users' Group

provides a forum for exchanging information about using the Cochrane Library, including its usefulness, conducting searches, and format.

More powerful search engines, improved indexing, text word searching, user-friendly interfaces, reference management software, and other advances in health informatics are helping non-expert searchers to retrieve valuable information more effectively and efficiently. As such, they are able to gain a more immediate, hands-on understanding of the scope and quality of literature of interest.

There is an expanding set of resources for supporting search strategies for HTA (Goodman 1993; Institute of Health Economics 2013; Niederstadt 2010; Ormstad 2010; Straus 2011). A recent analysis identified eight methods for determining the optimal retrieval of studies in HTA searches: the capture-recapture technique, obtaining feedback from the commissioner of the HTA report, seeking the disconfirming case, undertaking comparison against a known gold standard, evaluating retrieval of known items, recognizing the law of diminishing returns, specifying a priori stopping rules, and identifying a point of theoretical saturation (Booth 2010). Particularly instructive for clinicians is the series of articles originally published in the *Journal of the American Medical Association: Users' Guides to the Medical Literature,* which has been compiled and updated by the Evidence-Based Medicine Working Group (Guyatt 2008).

The NLM continues to improve its **MeSH** (*Me*dical *S*ubject *H*eadings) **controlled vocabulary** (used to index and search literature in PubMed and other NLM sources) in HTA and such related fields as health services research, evidence-based medicine, and comparative effectiveness research. Continued work has improved the indexing and searching of RCTs in PubMed and other bibliographic databases (Dickersin 1994; Glanville 2006; Royle 2007; Zhang 2006). 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.

The search for pertinent existing evidence is normally one of the first major tasks of an HTA, 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 sets, and other documentation. Although access to PubMed and other government-source databases is generally free or 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 Institute, the BlueCross BlueShield Association Technology Evaluation Center (BCBSA TEC), and Hayes Inc. sell their reports on a subscription basis. (BCBSA TEC reports are also accessible on its web site to any user for private research or study only.) Some market research monographs and other reports oriented for health product companies, investors and other business interests are priced in the thousands of dollars.

References for Chapter VII

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.

Booth A. How much searching is enough? Comprehensive versus optimal retrieval for technology assessments. Int J Technol Assess Health Care. 2010;26(4):431-5.

Canadian Agency for Drugs and Technologies in Health (CADTH), Information Services. Grey Matters: A Practical Search Tool for Evidence-Based Medicine. February 2014. Accessed April 10, 2014 at: http://www.cadth.ca/en/resources/finding-evidence-is/grey-matters.

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

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. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2541778.

Dundar Y, Dodd S, Dickson R, Walley T, et al. Comparison of conference abstracts and presentations with full-text articles in the health technology assessments of rapidly evolving technologies. Health Technol Assess. 2006;10(5):iii-iv, ix-145. http://www.journalslibrary.nihr.ac.uk/hta/volume-10/issue-5.

Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. Lancet. 1991;337(8746):867-72.

Egger M, Smith GD, Sterne JA. Uses and abuses of meta-analysis. Clin Med. 2001;1(16):478-84.

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. http://www.journalslibrary.nihr.ac.uk/hta/volume-7/issue-1.

Galandi D, Schwarzer G, Antes G. The demise of the randomised controlled trial: bibliometric study of the German-language health care literature, 1948 to 2004. BMC Med Res Methodol. 2006;6:30. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1533847.

Glanville J, Paisley S. Identifying economic evaluations for health technology assessment. Int J Technol Assess Health Care. 2010;26(4):436-40.

Gøtzsche PC. Multiple publication of reports of drug trials. Eur J Clin Pharmacol. 1989;36(5):429-32.

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

Guyatt G, Rennie D, Meade MO, Cook DJ. Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice, 2nd ed. New York: McGraw-Hill Education; 2008.

Hall R, de Antueno C, Webber A; Canadian Research Ethics Board. Publication bias in the medical literature: a review by a Canadian Research Ethics Board. Can J Anaesth. 2007;54(5):380-8.

Higgins JPT, Green S, eds. Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Accessed Sept. 1, 2013 at: http://handbook.cochrane.org.

Hopewell S, Clarke M, Stewart L, Tierney J. Time to publication for results of clinical trials. Cochrane Database Syst Rev. 2007;(2):MR000011.

Hopewell S, McDonald S, Clarke M, Egger M. Grey literature in meta-analyses of randomized trials of health care interventions. Cochrane Database Syst Rev. 2007;(2):MR000010.

Institute of Health Economics, Osteba, AUnETS. Health Technology Assessment on the Net International: 2013. Edmonton AB: Institute of Health Economics. 2013. Accessed February 1, 2014 at: http://www.ihe.ca/documents/HTA%20on%20the%20Net%202013.pdf.

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. http://books.nap.edu/openbook.php?record_id=10027.

Ioannidis JP. Effect of statistical significance of results on the time to completion and publication of randomized efficacy trials. JAMA. 1998;279(4):281-6.

Jüni P, Holenstein F, Sterne J, Bartlett C, Egger M. Direction and impact of language bias in meta-analyses of controlled trials: empirical study. Int J Epidemiol. 2002;31(1):115-23. http://ije.oxfordjournals.org/content/31/1/115.long.

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;56(1):44-51.

Ormstad SS, Isojärvi J. Information retrieval for health technology assessment: standardization of search methods. Int J Technol Assess Health Care. 2010;26(4):359-61.

Roundtree AK, Kallen MA, Lopez-Olivo MA, Kimmel B, et al. Poor reporting of search strategy and conflict of interest in over 250 narrative and systematic reviews of two biologic agents in arthritis: a systematic review. Clin Epidemiol. 2009;62(2):128-37.

Royle P, Waugh N. Literature searching for clinical and cost-effectiveness studies used in health technology assessment reports carried out for the National Institute for Clinical Excellence appraisal system. Health Technol Assess. 2003;7(34):iii, ix-x,1-51. http://www.journalslibrary.nihr.ac.uk/hta/volume-7/issue-34.

Royle PL, Waugh NR. Making literature searches easier: a rapid and sensitive search filter for retrieving randomized controlled trials from PubMed. Diabet Med 2007;24(3):308-11.

Schöpfel J. Towards a Prague Definition of Grey Literature. GL 12. Twelfth International Conference on Grey Literature. Prague, Czech Republic. December 2010. Accessed November 27, 2013 at: http://www.textrelease.com/images/GL12_Abstract_S1N1.pdf.

Song F, Parekh S, Hooper L, Loke YK, et al. Dissemination and publication of research findings: an updated review of related biases. Health Technol Assess. 2010;14(8):iii, ix-xi, 1-193.

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(7304):101-5. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1120714.

Sterne JA, Sutton AJ, Ioannidis JP, Terrin N, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomized controlled trials. BMJ. 2011;343:d4002.

Straus SE, Richardson WS, Glasziou P, Haynes RB. Evidence-Based Medicine: How to Practice and Teach It. 4th ed. New York, NY: Churchill Livingstone Elsevier, 2011.

Wilkins T, Gillies RA, Davies K. EMBASE versus MEDLINE for family medicine searches: can MEDLINE searches find the forest or a tree? Can Fam Physician. 2005;51:848-9. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1479531.

Zarin DA, Tse T, Williams RJ, Califf RM, Ide NC. The ClinicalTrials.gov results database--update and key issues. N Engl J Med. 2011;364(9):852-60. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3066456.

Zhang L, Ajiferuke I, Sampson M. search strategies to identify randomized controlled trials in MEDLINE. BMC Med Res Methodol. 2006;6:23. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1488863.

VIII. DISSEMINATE FINDINGS AND RECOMMENDATIONS

For those who sponsor, conduct, or otherwise participate in an HTA, the importance of its findings and recommendations may be self-evident. However, there is increasing recognition that dissemination of HTA findings and recommendations, whether for internal use by the sponsoring organization or into the national or international health information mainstream, must be carefully planned and implemented in order to enable any HTA to achieve its purpose.

Dissemination of HTA findings and recommendations must inform decisions and policies for improving population risk factors, patient care, health outcomes, and resource allocation, as appropriate. Worthy HTA messages get lost because of misidentified or misunderstood target audiences, poor presentation or format, ineffective or poorly selected communication media, poor timing, and other factors. There is some convergence in the field toward good practices on the format and organization of information to be included in an HTA report, e.g., the INAHTA checklist for HTA reports (INAHTA 2007) and the EUnetHTA core model (EUnetHTA 2008). However, as described in chapter X regarding barriers to HTA, the implementation of HTA findings and recommendations can be hindered by, e.g., lack of access to HTA reports, complex technical formats of HTA reports, lack of confidence in the evidence base, and apparent absence of real-world applications. There is continued need for research regarding how to optimize the dissemination of HTA findings and recommendations.

A. Competing for Attention

The uptake of HTA findings depends in part on the responsibilities, incentives, behavior, other attributes of the target audiences, and various contextual factors. It also depends on the existing knowledge and perceptions of these target audiences. HTA **dissemination** efforts must compete with the burgeoning flow of health-related information being transmitted via diverse media using increasingly sophisticated means. Advanced communications technologies provide alternative means to transmit more information where and when it can influence clinicians and other decision makers. Health technology companies are highly adept at using advanced and continually evolving marketing techniques that are being adapted throughout the health care sector.

The matter of the accuracy and consistency of information presented in health technology promotional information to clinicians and consumers remains a great challenge in many regions around the world (Mali 2010) and will continue to affect the ability of HTA dissemination efforts to inform evidence-based behavior change in practice. Three particularly relevant areas are direct-to-consumer advertising (DTCA), promotion of off-label use of drugs and other health technologies, and making economic claims about health technologies, as described briefly, below.

In many countries, the interest in DTCA includes whether it should be permitted at all and, if so, what regulatory requirements should pertain to the content and format of its messages and whether such requirements are being sufficiently enforced (Donohue 2007; Mintzes 2009). Steep increases in DTCA spending since the late 1990s, especially in the US, have contributed to increased prescription drug sales and changed consumer behavior. Although DTCA can avert underuse of appropriate indications of prescription drugs, it may contribute to overuse, though the evidence on inappropriate prescribing is limited (Atherly 2009; Donohue 2007; Lyles 2002). The increase in the volume of drug purchasing may be class-wide rather than product specific (Atherly 2009). DTCA appears to increase public awareness of prescription drugs but may be less effective at educating patients about appropriate use of medications for their condition. With DTCA exposure, patients are reported to be more likely to ask about advertised

conditions, request prescriptions, and seek further information about medications, with some evidence that patients have expressed some caution about DTCA due to publicity about drug recalls, e.g., pertaining to certain pain medications (Liu 2008). All of these factors affect the receptiveness of target audiences to HTA findings and recommendations.

A related area of interest is the ability of health product companies to distribute published and unpublished reports of clinical trials of their products for indications that have not been cleared for marketing by the appropriate regulatory authority, known as "off-label" use. In the US, the FDA does not regulate the practice of medicine, and physicians may prescribe drugs and other regulated technologies for indications other than those for which they were approved by the FDA, unless doing so violates certain safety regulations or ethical standards. Off-label use of drugs is widespread. According to a comprehensive study of outpatient drug prescribing patterns in the US during 2001, 21% of prescriptions were off-label, led by cardiac medications and anticonvulsants; with 31% off-label prescriptions of psychiatric drugs. A majority (73%) of the off-label prescriptions were reported to have had little or no scientific support (Radley 2006). A study of primary physician prescribing patterns in Quebec, Canada, during 2005-2009 found that the prevalence of off-label use was 11%, 79% of which lacked strong scientific support. Off-label use was highest for central nervous system drugs (26%), particularly anticonvulsants, antipsychotics, and antidepressants (Eguale 2012). Among the legal concerns is whether providing "truthful information" about "widely accepted" off-label use of a drug, whether or not there is supportive scientific evidence, should be permissible (Gilhooley 2011). In the US, current FDA guidance addresses the types of literature that may be distributed to health care providers (e.g., it should be peer-reviewed, not in special journal supplements, and not significantly influenced by the manufacturer) and how that literature can be disseminated (unabridged, accompanied by approved product labeling, etc.), including that it must state that the uses described in the information have not been approved or cleared by the FDA, and other requirements (US FDA 2009).

Another aspect of information dissemination concerns the conditions under which heath technology companies can make economic claims (regarding prices, market share, cost-effectiveness, etc.) in their marketing information, what the rigor of supporting evidence should be, and which agencies should have regulatory oversight for such economic claims (Neumann 2000). Although patterns of economic claims by pharmaceutical and medical device companies about their products have changed in recent years, the use of supporting evidence in pharmaceutical formulary claims and medical device advertisements in particular remains low (Ackerly 2010; Palmer 2008).

"Academic detailing" has arisen in part to balance or offset the industry-based information dissemination efforts noted above. Academic detailing is a form of evidence-based educational outreach to physicians and other drug and device prescribers by health care professionals with no ties to industry. It is patterned after the highly effective practice in the US and some other countries of physician "detailing" (in-person visits) by pharmaceutical and device company representatives to promote prescribing of their products by practicing physicians (Solomon 2001; Soumerai 1990). To the extent that is provides clinicians with unbiased evidence-based syntheses of the best available evidence, academic detailing is intended to overcome challenges of interpreting evidence, provide trusted sources of information, and translate research findings into clinical practice (Fischer 2012).

Findings of a systematic review suggest that printed educational materials may improve process outcomes (e.g., certain physician and patient behaviors) but do not lead to significant improvement in patient outcomes. Further, there is insufficient knowledge about how to improve the effectiveness of

educational materials and the effectiveness of educational materials compared to other interventions (Farmer 2008).

B. Basic Dissemination Framework

While HTA report findings can be delivered in various forms and communications media, the mere provision of information often results in little behavior change. More is being learned about techniques to increase the likelihood of changing behavior, such as face-to-face interaction, active involvement of the learner, message repetition, making recommendations explicit and relevant to clinical practice, use of opinion leaders (or "champions"), and influence of peers. As the health care information and communications environment evolves, further insights are needed about clinicians' motivations and decision making, as well as those of patients and consumers. In the last decade, research findings on the effectiveness of various approaches for disseminating or sharing evidence-based information among clinicians has been assembled in systematic reviews, with identification of areas for further research, e.g., on continuing education (Forsetlund 2009), educational outreach visits (O'Brien 2007), opinion leaders (Flodgren 2011), audit and feedback (Jamtvedt 2006), and ways to promote interprofessional education and collaboration (Reeves 2013; Zwarenstein 2009).

HTA report dissemination can be described using three basic dimensions: target groups, media, and implementation techniques or strategies, as shown in **Box VIII-1**. HTA reports and documentation derived from those reports can be packaged for dissemination in different formats, e.g., for patients, clinicians, health care facility managers, payers, policymakers, researchers, the public, and news organizations. Reaching and influencing any target audience may require multiple messages and media (Kahan 1988).

Box VIII-1. Options for Dissemination of HTA Findings and Recommendations

Target groups

- Clinicians (individuals, professional/specialty associations)
- Patients/consumers/general public (individuals, organizations)
- Disease/disorder associations (voluntary/non-profit)
- Provider institutions (hospitals, clinics, laboratories, etc.)
- Third party payers/health plans/drug formularies (government, private sector)
- Government policymakers (international, national, state, local)
- Judicial systems (e.g., for legal challenges to payer coverage policies)
- Journal editors (general/popular, scientific/health/specialty clinical)
- News services/professionals (general/popular, scientific/health/specialty clinical)
- Researchers (biomedical, health services, public health, etc.)
- Life sciences companies (pharmaceuticals, biologics, medical devices, etc.)
- Academic institutions (schools, continuing professional education programs)
- Quality assurance and utilization review organizations

Media

- Printed: scientific/professional journals and newsletters, direct mail, newspapers and popular journals, posters, pocket cards
- Electronic: Internet (including cloud), television, radio, digital video disks, flash drives, webinars, podcasts, other computer databases (online, disk, etc.)
- Social i.e., Internet-based and mobile interactive media, including: collaborative sites (e.g., wikis); blogs, vlogs, and microblogs (e.g., Twitter); social networking (e.g., Facebook, LinkedIn); content communities (e.g., YouTube), crowdsourcing
- Word of mouth: in-person/face-to-face interaction, e.g., informal consultation, formal lectures and presentations, focus groups

Implementation techniques or strategies

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

C. Dissemination Plan

Dissemination should be designed to inform and influence behavior of the relevant decision makers, as appropriate. Although each HTA program should have certain standard dissemination protocols and success measures, this does not mean that dissemination plans should be the same for all HTAs. HTA programs should develop, continue to improve, and adapt their dissemination activities for particular HTAs. Dissemination planning should start at or before the initiation of each HTA. The costs, time, and other resources allocated for dissemination should be budgeted accordingly, yet should allow flexibility. As they emerge during the course of the HTA, the findings and recommendations themselves may affect target group selection, the types of messages to be delivered, and the media for delivering these messages.

There is growing recognition of the benefits of early involvement of HTA target groups in improving the design, conduct, and use of HTA reports. This includes gaining insights about overcoming scientific, professional, social, cultural, organizational, economic, and other barriers to acting on HTA findings and recommendations (Hivon 2005). For patients and consumers, this includes efforts to:

- Identify patient and consumer organizations that are stakeholders on particular HTA topics
- Acknowledge patients and consumers as target groups for HTA findings, where appropriate
- Determine ways in which such organizations could help to disseminate HTA findings
- Involve representatives of these groups during the HTA process where appropriate (e.g., to gain input on key questions, interventions, comparators, outcomes/endpoints of importance, concerns about access, review of draft reports)
- Engage these groups in HTA report dissemination strategies and implementation (Fattal 2008)

Chapter X has further discussion of the patient and consumer role in HTA, including matters related to dissemination. Although there is wider recognition of the need to improve HTA dissemination, much work is needed to augment the body of knowledge and methods for doing so. Perhaps because HTA originally focused on generating reports for policymakers, current efforts are addressing more active and adaptive dissemination to other stakeholders who use or are affected by HTA findings, such as hospital and other health care facility managers; clinicians; patient, consumer, and care-giver organizations; the general public; the health technology industry; researchers; and news organizations and other "vectors" of health information (Battista 2009; Lehoux 2005). This includes packaging HTA reports into more diverse products, conveying HTA findings in the contexts or decision-making environments of target groups; translations into target users' languages, and adjusting clinical and technical terminology as needed (Lehoux 2004; Sobrido Prieto 2007).

D. Managing Access

There are many approaches to managing or enhancing access to HTA reports. As noted above, some HTA reports are public documents available at no cost via the Internet or other public sources, while others are available only to paid subscribers or membership organizations. Others are for internal use only and not made public. Access to HTA reports and related literature is also subject to the capacity, policies, and priorities of bibliographic organizations (e.g., the US National Library of Medicine and commercial database vendors) regarding the scope of their online databases and other information sources (e.g., inclusion of peer-reviewed journals, grey literature, or other sources) and capacity to index the literature. Also, the word choice of HTA authors for report titles and abstracts can influence the indexing (e.g., designation of MeSH terms and other controlled vocabularies) and retrieval of HTA

reports. Over time, these controlled vocabularies evolve with the terminology used in the field of HTA, toward improvement of indexing and retrieval.

References for Chapter VIII

Ackerly DC, Glickman SW, Schulman KA. Economic content in medical journal advertisements for medical devices and prescription drugs. Pharmacoeconomics. 2010;28(5):429-38.

Atherly A, Rubin PH. The cost-effectiveness of direct-to-consumer advertising for prescription drugs. Med Care Res Rev. 2009;66(6):639-57.

Battista RN, Hodge MJ. The ”natural history” of health technology assessment. Int J Technol Assess Health Care. 2009;25 Suppl 1:281-4.

Donohue JM, Cevasco M, Rosenthal MB. A decade of direct-to-consumer advertising of prescription drugs. N Engl J Med. 2007;357(7):673-81. http://www.nejm.org/doi/full/10.1056/NEJMsa070502.

Eguale T, Buckeridge DL, Winslade NE, Benedetti A, et al. Drug, patient, and physician characteristics associated with off-label prescribing in primary care. Arch Intern Med. 2012;172(10):781-8.

EUnetHTA (European Network for Health Technology Assessment). HTA Core Model for Medical and Surgical Interventions Version 1.0 Work Package 4. The HTA Core Model. December 2008. http://www.eunethta.eu/sites/5026.fedimbo.belgium.be/files/HTA%20Core%20Model%20for%20Medical%20and%20Surgical%20Interventions%201.0r.pdf.

Farmer AP, Légaré F, Turcot L, Grimshaw J, et al. Printed educational materials: effects on professional practice and health care outcomes. Cochrane Database Syst Rev. 2008;(3):CD004398.

Fattal J, Lehoux P. Health technology assessment use and dissemination by patient and consumer groups: why and how? Int J Technol Assess Health Care. 2008;24(4):473-80.

Fischer MA, Avorn J. Academic detailing can play a key role in assessing and implementing comparative effectiveness research findings. Health Aff (Millwood). 2012;31(10):2206-12.

Flodgren G, Parmelli E, Doumit G, Gattellari M, et al. Local opinion leaders: effects on professional practice and health care outcomes. Cochrane Database Syst Rev. 2011;(8):CD000125.

Forsetlund L, Bjørndal A, Rashidian A, Jamtvedt G, et al. Continuing education meetings and workshops: effects on professional practice and health care outcomes. Cochrane Database Syst Rev. 2009;(2):CD003030.

Gilhooley M. Commercial speech and off-label drug uses: what role for wide acceptance, general recognition and research incentives? Am J Law Med. 2011;37(2-3):258-77.

Hivon M, Lehoux P, Denis JL, Tailliez S. Use of health technology assessment in decision making: coresponsibility of users and producers? Int J Technol Assess Health Care. 2005;21(2):268-75.

INAHTA. International Network of Agencies in Health Technology Assessment Secretariat. A Checklist for Health Technology Assessment Reports. Version 3.2. August 2007. Accessed November 27, 2013 at: http://www.inahta.org/HTA/Checklist.

Jamtvedt G, Young JM, Kristoffersen DT, O'Brien MA, Oxman AD. Audit and feedback: effects on professional practice and health care outcomes. Cochrane Database Syst Rev. 2006;(2):CD000259.

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.

Lehoux P, Denis JL, Tailliez S, Hivon M. Dissemination of health technology assessments: identifying the visions guiding an evolving policy innovation in Canada. J Health Polit Policy Law. 2005;30(4):603-41.

Lehoux P, Tailliez S, Denis JL, Hivon M. Redefining health technology assessment in Canada: diversification of products and contextualization of findings. Int J Technol Assess Health Care. 2004;20(3):325-36.

Liu Y, Doucette WR. Does direct-to-consumer advertising affect patients' choice of pain medications? Curr Pain Headache Rep. 2008;12(2):89-93.

Lyles A. Direct marketing of pharmaceuticals to consumers. Annu Rev Public Health. 2002;23:73-91.

Mali SN, Dudhgaonkar S, Bachewar NP. Evaluation of rationality of promotional drug literature using World Health Organization guidelines. Indian J Pharmacol. 2010;42(5):267-72. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2959206.

Mintzes B, Morgan S, Wright JM. Twelve years' experience with direct-to-consumer advertising of prescription drugs in Canada: a cautionary tale. PLoS One. 2009;4(5):e5699. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2682647.

Neumann PJ, Claxton K, Weinstein MC. The FDA's regulation of health economic information. Health Aff (Millwood). 2000;19(5):129-37. http://content.healthaffairs.org/content/19/5/129.long.

O'Brien MA, Rogers S, Jamtvedt G, Oxman AD, et al. Educational outreach visits: effects on professional practice and health care outcomes. Cochrane Database Syst Rev. 2007;(4):CD000409.

Palmer JA, Timm AR, Neumann PJ. Drug company advertising in medical journals about the health-economic advantages of their products for 2000-2006 versus 1990-1999. J Manag Care Pharm. 2008;14(8):749-55.

Radley DC, Finkelstein SN, Stafford RS. Off-label prescribing among office-based physicians. Arch Intern Med. 2006;166(9):1021-6.

Reeves S, Perrier L, Goldman J, Freeth D, Zwarenstein M. Interprofessional education: effects on professional practice and healthcare outcomes (update). Cochrane Database Syst Rev. 2013 Mar 28;3:CD002213.

Sobrido Prieto M, Gonzalez Guitian C, Cerda Mota T. Strategies for the diffusion and dissemination of health technology assessment (HTA) products. Santiago de Compostela: Galician Agency for Health Technology Assessment (AVALIA-T). avalia-t No 2007-072009; 2007. http://www.sergas.es/docs/Avalia-t/diffusion-Strategies.pdf.

Solomon DH, Van Houten L, Glynn RJ, Baden L. Academic detailing to improve use of broad-spectrum antibiotics at an academic medical center. Arch Intern Med. 2001;161(15):1897-902.

Soumerai SB, Avorn J. Principles of educational outreach ('academic detailing') to improve clinical decision making. JAMA. 1990;263(4):549-56.

US Food and Drug Administration. Guidance for Industry - Good Reprint Practices for the Distribution of Medical Journal Articles and Medical or Scientific Reference Publications on Unapproved New Uses of Approved Drugs and Approved or Cleared Medical Devices. January 2009. Accessed November 27, 2013, at: http://www.fda.gov/RegulatoryInformation/Guidances/ucm125126.htm.

Zwarenstein M, Goldman J, Reeves S. Interprofessional collaboration: effects of practice-based interventions on professional practice and healthcare outcomes. Cochrane Database Syst Rev. 2009 Jul 8;(3):CD000072.

IX. MONITOR IMPACT OF HTA

The impact of HTA is variable and inconsistently understood. Among the most important factors influencing the impact of HTA reports is the directness of the relationship between an HTA program and policymaking bodies and health care decisions. Whereas some HTA reports are translated directly into policies with clear and quantifiable impacts, the findings of some authoritative, well-documented assessment reports, even those based on "definitive" RCTs and other rigorous studies, often go unheeded or are not readily adopted into general practice (Banta 1993; Ferguson, Dubinsky 1993; Henshall 2002; Institute of Medicine 1985). Indeed, even when the reporting of HTA findings is followed by changes in policies, use of a technology, or other potential indicators of impact, it may be difficult to demonstrate the causal effect of the HTA on those changes.

HTA reports can make an impact by changing one or more of:

- Regulatory policy (e.g., market access of a technology)
- Third-party payment policy (e.g., coverage, pricing, reimbursement of a technology)
- Rate of use of a technology
- Clinical practice guidelines
- Clinician awareness and behavior
- Patient awareness and behavior
- Acquisition, adoption, or diffusion of a technology
- Organization or delivery of care
- R&D priorities and associated spending levels
- Data collection (e.g., to fill evidence gaps identified by HTA reports)
- Marketing of a technology
- Allocation of local, regional, national, or global health care resources
- Investment decisions (e.g., by industry, investors)
- Incentives to innovate

The impacts of HTA can occur in an interrelated series (although not necessarily in strict sequence), such as that described by EUnetHTA (Garrido 2008):

Awareness → Acceptance → Policy process → Policy decision → Practice → Outcome

Historically, systematic attempts to document the dissemination processes and impacts of HTA programs have been infrequent (Banta 1993; Goodman 1988; Institute of Medicine 1985; Jacob 1997), though a few, notably the NIH Consensus Development Program (Ferguson 1993), have been studied in detail. More recently, there is growing recognition that monitoring the impact of individual HTAs and HTA programs is a "good practice" or "key principle" of HTA (see, e.g., Drummond 2012; Drummond 2008; Goodman 2012; Velasco 2002). A small but steadily growing literature has reported a range of impacts of HTA in specific countries and other jurisdictions on technology adoption, disinvestment, reimbursement, and other policies and practices (Hailey 2000; Hanney 2007; Zechmeister 2012).

Although precise estimates of the impact of individual HTAs and HTA programs will seldom be possible, ongoing efforts to systematically document the changes that are known to result from HTA, or that are associated with HTA, are feasible (Hanney 2007; Jacob 1997).

An area of increasing interest is the impact of HTA on disinvestment in technologies that do not offer value for money or have been superseded by others that are more safe, effective, and/or cost-effective. The use of such technologies often persists due to financial incentives, professional interests, and resistance to change among clinicians, patients, and health care delivery and payment systems (Garner 2011). For example, in the US, despite the findings from two rigorous, double-blind RCTs demonstrating that percutaneous vertebroplasty for painful vertebral fractures provided no better pain relief than a sham procedure, third-party payers continued to cover the procedure more than two years after publication of the trial results (Wulff 2011). The ability of HTA to inform evidence-based disinvestment is of particular importance in health care systems with fixed budgets, where spending on low-value technologies limits expenditures on more cost-effective ones (Kennedy 2009).

A. Attributing Impact to HTA Reports

The impact of a HTA depends on diverse factors. Among these are target audiences' legal, contractual, or administrative obligations, if any, to comply with the HTA findings or recommendations (Anderson 1993; Ferguson, Dubinsky 1993; Gold 1993). Regulatory agency (e.g., the FDA in the US) approvals or clearances for marketing new drugs and devices are translated directly into binding policy. In the US, HTAs conducted by AHRQ at the request of CMS are used to inform technology coverage policies for the Medicare program, although CMS is not obligated to comply with findings of the AHRQ HTA. The impacts of NIH consensus development conference statements, which were not statements of government policy, were inconsistent and difficult to measure. Their impact appeared to depend on a variety of factors intrinsic to particular topics, the consensus development process itself, and a multitude of contextual factors (Ferguson 1993; Ferguson 2001).

The task of measuring the impact of HTA can range from elementary to infeasible. As noted above, 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 leads to recommendations to increase 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, third-party payment policy is already shifting in favor of the technology, a strong marketing effort is being made by industry, or results of a definitive RCT are being made public.

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 on the conditions under which the findings were made known and the methodological approach used to determine the impact. Evaluations of the impact of an HTA often are unavoidably observational in nature; however, under some circumstances, quasi-experimental or experimental evaluations have been used (Goldberg 1994). To the extent that impact evaluations are prospective, involve pre- and post-report 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 any true causal connection between an HTA report and change in policy or behavior. 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.

B. Factors Influencing Impact

Many factors can affect the impact of HTA reports. Beyond the particular dissemination techniques used, characteristics of the target groups, the environment and the HTAs themselves can influence their impact (Goldberg 1994; Mittman and Siu 1992; Mittman and Tonesk 1992). Examples are shown in **Box IX-1**. Knowledge about these factors can be used prospectively to improve the impact of HTA.

As described in another chapter of this document, in seeking to maximize the impact of their reports, HTA programs can involve target audiences early, such as in priority setting of assessment topics and determination of assessment questions. Further, they can consider how to properly present their reports and plan their dissemination strategies to reach and influence those various target audiences.

The impact of HTA findings may be increased to the extent that the HTA process is local, i.e., conducted by or involving people in the target decision-making organization, such as a hospital network or major payer agency. Such "local" HTA can increase the utility of HTA findings due to the relevance of the HTA topic (e.g., by having input on topic selection and use of local data), timeliness, and formulating policy reflecting the local values and context (Bodeau-Livinec 2006; McGregor 2005). Findings from HTA that is conducted with rigorous, well-documented methodology on topics that are priorities or otherwise of interest to sponsors with policymaking authority ("policy customers") are more likely to be adopted and have an impact (Hanney 2007; Raftery 2009).

In summary, the following are ways in which HTA programs can increase the likelihood of their reports having the intended impacts (see, e.g., Hailey 2000; McGregor 2005; Sorensen 2008):

- Conduct a transparent, credible, unbiased, rigorous, and well-documented HTA process
- Gain prior commitment, where feasible, from decision makers to use HTA findings
- Ensure that assessments are designed to address decision makers' questions
- Seek to establish formal links between producers and users of HTA
- Involve key stakeholders throughout the HTA process (e.g., in priority setting, determination of assessment questions) in a transparent, well-managed manner
- Gain input of representatives of anticipated target audiences and communication experts in planning knowledge transfer strategies, including different formats, languages, media, and related messaging of HTA findings to different target audiences, as appropriate
- Anticipate the resource requirements, incentives, delivery system characteristics, and other diverse factors that will influence the feasibility of implementing HTA findings
- Ensure that HTA findings are delivered on a timely basis to inform decision making
- Promote collaboration and transfer of knowledge and skills across jurisdictions (e.g., across nations, regions, localities)

Box IX-1. Examples of Factors That Can Affect Impact of HTA Reports

Target clinician characteristics

- Type of clinician: physician, mid-level practitioner, nurse, dentist, etc.
- Specialty; training
- Professional activities/affiliations
- Institutional affiliations (e.g., community hospital, university hospital)
- Financial, professional. quality incentives to implement findings/recommendations
- Awareness of performance relative to peers
- Access to and familiarity with current evidence, practice guidelines
- Malpractice concerns/exposure

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, market competition
- Physicians' offices: group practice vs. solo practice, hospital affiliation, teaching affiliation, board certification, distribution of payment sources, market competition
- Financial, organizational, or quality incentives to implement findings/recommendations

Target patient characteristics

- Insurance (type) and cost sharing status (deductible, copayment, etc.)
- Access to regular primary care provider, other care
- Health status
- Health awareness, use of health information media, health literacy
- Socioeconomic/demographic/cultural factors
- Home, workplace, other environmental factors
- Social interaction (family, friends, peers, etc.)

Environmental characteristics

- Urban, suburban, rural
- Competition
- Economic status
- Third-party payment (e.g., market distribution of fee-for-service vs. bundled payment)
- State and local laws, regulations
- Activities of pressure groups/lobbyists, other interest groups
- Malpractice potential/activity
- Political factors

Characteristics of HTA findings/recommendations

- Type/extent of engagement of target audiences/stakeholders in process
- Timeliness/responsiveness relative to needs of target audiences
- Reputation/credibility of HTA organization, analysts, expert panels
- Transparency/rigor of assessment process
- Quality and strength of evidence base
- Application of findings: evidence review only; policy implications/recommendations; input to practice guidelines, coverage/reimbursement, technology acquisition, quality standards, etc.
- Perceived appropriateness of rigidity or flexibility of findings/recommendations
- Dissemination media, format, content/frequency
- Proximity to decision makers or policymakers and extent of their obligation (e.g., legal mandate or optional) to implement findings/recommendations
- Resources required to implement findings/recommendations

References for Chapter IX

Anderson GF, Hall MA, Steinberg EP. Medical technology assessment and practice guidelines: their day in court. Am J Public Health. 1993;83(3):1635-9. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1694869.

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

Bodeau-Livinec F, Simon E, Montagnier-Petrissans C, Joël ME, Féry-Lemonnier E. Impact of CEDIT recommendations: An example of health technology assessment in a hospital network. Int J Technol Assess Health Care. 2006;22(2):161-8.

Drummond M, Neumann P, Jönsson B, Luce B, et al. Can we reliably benchmark health technology assessment organizations? Int J Technol Assess Health Care. 2012 Apr;28(2):159-65.

Drummond MF, Schwartz JS, Jönsson B, Luce BR, et al. Key principles for the improved conduct of health technology assessments for resource allocation decisions. Int J Technol Assess Health Care. 2008;24(3):244-58.

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.

Garner S, LIttlejohns P. Disinvestment from low value clinical interventions: NICEly done? BMJ 2011;343:d4519.

Garrido MV, Kristensen FB, Nielsen CP, Busse R. Health Technology Assessment and Health Policy-Making in Europe: Current Status, Challenges, and Potential. European Observatory for Health Systems and Policies. Copenhagen: WHO Regional Office for Europe, 2008.

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.

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. Toward international good practices in health technology assessment. Int J Technol Assess Health Care. 2012;28(2):169-70.

Goodman C, ed. Medical Technology Assessment Directory: A Pilot Reference to Organizations, Assessments, and Information Resources. Washington, DC: Institute of Medicine; 1988. http://www.nap.edu/openbook.php?record_id=1090&page=R1.

Hailey D, Corabian P, Harstall C, Schneider W. The use and impact of rapid health technology assessments. Int J Technol Assess Health Care. 2000;16(2):651-6.

Hanney S, Buxton M, Green C, Coulson D, Raftery J. An assessment of the impact of the NHS Health Technology Assessment Programme. Health Technol Assess. 2007;11(53):iii-iv, ix-xi, 1-180. http://www.journalslibrary.nihr.ac.uk/hta/volume-11/issue-53/

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.

Institute of Medicine. Assessing Medical Technologies. Washington, DC: National Academy Press; 1985. http://www.nap.edu/openbook.php?record_id=607.

Jacob R, McGregor M. Assessing the impact of health technology assessment. Int J Technol Assess Health Care. 1997;13(1):68-80.

Kennedy I. Appraising the Value of Innovation and Other Benefits. A Short Study for NICE. July 2009. Accessed November 29, 2013 at: http://www.nice.org.uk/media/98F/5C/KennedyStudyFinalReport.pdf.

McGregor M, Brophy JM. End-user involvement in health technology assessment (HTA) development: a way to increase impact. Int J Technol Assess Health Care. 2005;21(2):263-7.

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 the1990s. 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 behavior change. QRB Qual Rev Bull. 1992;18(12):413-22.

Raftery J, Hanney S, Green C, Buxton M. Assessing the impact of England's National Health Service R&D Health Technology Assessment program using the "payback" approach. Int J Technol Assess Health Care. 2009;25(1):1-5.

Sorensen C, Drummond M, Kristensen FB, Busse R. How can the impact of health technology assessments be enhanced? European Observatory for Health Systems and Policies. Copenhagen: WHO Regional Office for Europe, 2008. http://www.euro.who.int/__data/assets/pdf_file/0019/73225/E93420.pdf.

Velasco M, Perleth M, Drummond M, Gürtner F, et al. Best practice in undertaking and reporting health technology assessments. Working group 4 report. Int J Technol Assess Health Care. 2002;18(2):361-422.

Wulff KC, Miller FG, Pearson SD. Can coverage be rescinded when negative trial results threaten a popular procedure? The ongoing saga of vertebroplasty. Health Aff (Millwood). 2011;30(12):2269-76.

Zechmeister I, Schumacher I. The impact of health technology assessment reports on decision making in Austria. Int J Technol Assess Health Care. 2012;28(1):77-84.

X. SELECTED ISSUES IN HTA

A. Barriers to HTA

Although the general trend in health care is toward an increased role of HTA, improved HTA processes, greater transparency of HTA, and greater involvement of stakeholders in HTA, various countervailing forces to HTA remain. Some stakeholders may oppose HTA due to their inherent interests, some of which may conflict with evidence-based processes. Innovators seek to maintain incentives to develop new technologies and lower hurdles to market access. Investors in health care products and services seek high, short-term returns on their investments. Health technology companies seek to maximize profit through increased sales of their products at high prices. Physicians and other clinicians seek to maintain their decision-making autonomy regarding ordering tests, performing procedures, prescribing therapies, and making referrals to other clinicians. Managers of hospitals and other health care facilities seek to improve their competitive status in their markets. Patient interest groups seek to retain unlimited choice of treatment options (Fineberg 2009). Among the barriers to HTA are the following:

Technological imperative. Particularly in the US and other wealthy countries, there is 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 to conduct HTA (Hoffman 2002).

Limited resources for HTA. As is so for other efforts, resources for HTA are limited. Although some HTA programs and certain HTA findings are nationally or internationally recognized, the resources allocated for HTA, even in the wealthy nations, are vanishingly small compared to national health care spending.

Insufficient primary data. Lack of primary studies and other data sources limits the evidence base for HTA. This also includes lack of local or regional data to conduct HTA pertaining to a particular nation, region, or health care facility and lack of data pertaining to particular population subgroups for which a technology may be applicable.

Timing misalignment. The timing of HTA may be misaligned with decision-making needs and other events. This may arise in delays in requesting or allocating funding for an HTA, ongoing adoption and use of technologies while HTA is being conducted, delays between release of HTA findings and their adoption in policy and practice, and the "moving target problem," in which the relevance of HTA findings is diminished by changes in technologies, their comparators, or how they are used.

Prestigious proponents of technology. The opinions of highly regarded or powerful proponents or "champions" of adopting a technology may prevail, even in the absence of credible supporting evidence.

Marketing. Increasingly effective and targeted marketing and promotion of health technologies, including short courses sponsored by health care product companies to train physicians in using these products and direct-to-consumer advertising (where this is permitted) can weigh against HTA findings.

Financial incentives. Health care systems that reimburse hospitals, physicians, and other providers every time a technology is provided, i.e., "fee-for-service" health care, tend to increase the volume of technology use, even when supporting evidence is lacking. Hospitals and physician groups that have invested in major capital equipment and supporting infrastructure such as for diagnostic radiology,

radiation oncology, and robotic surgery, have incentives to use these technologies (Garrison 2011; Jacobs 2013). Also, patients with little or no exposure to costs tend to seek more health care. These financial incentives can contribute to the inertia of existing payment systems that reward use of technologies despite lack of supporting evidence and related policies and clinical practice guidelines.

Political actions. HTA may be circumvented by political actions, often prompted by "lobbying" or "pressure groups." This occurs, for example, when laws are passed to mandate (or eliminate) coverage by government or private sector payers for certain technologies, in contrast to findings based on available evidence, or in the absence of rigorous evidence.

Implementation barriers. There are various barriers to implementing some HTA findings and recommendations 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 findings and recommendations may be difficult to implement given clinicians' and other providers' reluctance to change long-standing practice routines, inertia of existing payment policies, and rapidly outdated education and training in some instances. Insufficient opportunity or encouragement for scientific inquiry and skepticism in clinical education contributes to this inertia. Implementation of HTA findings also may be limited due to practical external constraints, such as when adopting a new technology requires a particular environment (e.g., special shielded rooms, instrumentation, and related facilities for diagnostic and therapeutic procedures using ionizing radiation), professional training, or other resources that are unavailable in a particular facility.

As noted above, political forces can circumvent evidence-based processes (Fletcher 1997). One of the main applications of HTA is informing coverage policies. While many payers use HTA findings to inform coverage policies, they are also subject to laws in their respective countries, states, provinces, and other jurisdictions. Legislative bodies at these levels can mandate that health programs provide certain services. In the US, historical examples of technologies that have been mandated by Congress for the Medicare program or by state legislatures for public or private payers have included 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, were not based on the types of evidence-based methods preferred in HTA, including some instances where the evidence was sufficient for some populations with a given disease or risk factors but not for others. These mandates were influenced by pressure groups representing certain patients, physicians, 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). Mandating coverage of a technology, rather than subjecting it to HTA, can mask more complex clinical consequences. As noted above, in the 1990s, many health plans in the US reluctantly agreed to cover ABMT-HDC 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, along with revelations of falsified data by a researcher who reported early findings in support of the procedure, demonstrated that the procedure

provided no benefit compared to standard-dose treatment for breast cancer, and caused unnecessary suffering in some women (Berger 1999; ECRI 1995; Mello 2001; Sharf 2001).

B. Quality of Care and HTA

HTA is an important source of information for quality of care. Although a thorough discussion of this relationship is not within the scope of this document, 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 prevailing standards of health care. HTA and quality assurance are distinct yet interdependent processes that contribute to quality of care.

HTA generates findings that add knowledge about the relationship between health care interventions and outcomes. This knowledge can be used to develop and update a range of standards and guidelines for improving health care quality, including clinical 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. Participation of clinicians, and particularly opinion leaders, in HTA and in developing evidence-based clinical practice guidelines can improve the acceptance and adoption of those guidelines and, thereby, quality of 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 systematic approaches to implementing quality assurance that have been 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 2003; Pilnick 2001). For example, CQI has been evaluated in a multicenter RCT as a means improve the adoption of two process-of-care measures for coronary artery bypass graft surgery (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 (e.g., 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 summary, HTA is among the sources of 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). 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, Committee on Quality of Health Care in America 2001).

C. Comparative Effectiveness Research and HTA

Comparative effectiveness research (CER) is the generation and synthesis of evidence comparing the benefits and harms of alternative technologies to prevent, diagnose, treat, and monitor diseases and other health care conditions in "real-world" settings in order to improve the delivery of health care (Federal Coordinating Council on Comparative Effectiveness Research 2009; Institute of Medicine 2009). The purpose of CER is to strengthen the evidence base that is used for assisting patients, other consumers, clinicians, health care managers, policymakers and others to make more informed health care decisions for individuals and populations.

Various attributes of what is known today as CER have been incorporated into research on the impact of health care technologies over the last several decades or more. Indeed, CER draws selected attributes from such inquiries and evaluations as RCTs, HTA, outcomes research, effectiveness research, and evidence based medicine. The emergence of CER as an explicit, coherent field of research in the early 2000s arose from a growing recognition of such factors as:

- Evidence of inappropriate use of health care technologies, including over-use, under-use, and improper use
- Evidence of large variations (geographic and other) in practice
- Insufficiency of evidence developed for market approval/clearance by regulatory agencies (e.g., FDA) to also support clinical and policy decisions; this is typically because such evidence:
    - emphasizes efficacy rather than effectiveness
    - is often not derived from controlled clinical trials (especially for many medical devices)
    - often lacks active comparators (e.g., where placebos are the only controls)
    - tends to exclude certain patient populations (those who are elderly, have multiple co-morbidities, etc.)
    - often is not derived from studies that enable subgroup analyses
- Insufficiently rigorous or absent evidence for the many technologies that are not subject to oversight by regulatory agencies (e.g., surgical procedures)
- Lack of evidence from "head-to-head" comparisons of alternative interventions for particular health problems
- Lack of evidence in "real-world" practice (efficacy vs. effectiveness)
- Continued steep increases in health care costs prompting interest in more efficient care delivery

The main attributes of CER are:

- Direct ("head-to-head") comparisons of alternative interventions (rather than comparison with placebo or indirect comparisons)

- Applies to all types of technologies
- Measures effectiveness in real-world populations and health care settings
- Emphasizes health care outcomes (e.g., morbidity, mortality, symptoms, quality of life, adverse events) rather than surrogate or other intermediate endpoints
- Draws on variety of complementary research methods, data sources, and analytical tools
- Enables subgroup analyses to yield findings about different responses across patient types
- Includes emphasis on priority diseases and priority populations

In the US, there is no consensus on the role of economics, such as determining value for money, in the conduct of CER itself. Indeed, there are legislative constraints on the use of CER and cost-effectiveness analysis in making coverage decisions for the Medicare program and other health programs administered by the Department of Health and Human Services (Neumann 2012; Pearson 2010). Even so, the findings of CER and other evidence that may involve health economic analyses, such as cost-effectiveness analysis, is available for use by various analysts and decision makers. While the term "CER" is most often used in the US, other countries and regions use related, though not necessarily synonymous terms. For example, in Europe, there has been growing interest in "relative efficacy" and "relative effectiveness" trials for new drugs. This interest derives in part from differing evidence requirements that sometimes arise between regulatory agencies, particularly the European Medicines Agency (EMA), and payment authorities. Approaches under consideration involve identifying circumstances in which pre-licensing (pre-marketing) efficacy trials should use active comparators, and ways to close gaps between efficacy and effectiveness, including whether to conduct pre-licensing or post-licensing practical clinical trials (with more heterogeneous patient populations in real-world settings), or to better understand how such extrinsic factors as physician prescribing or patient adherence affect variability in drug response (Eichler 2009; Glaeske 2012). These proposals may affect the relative roles and relationships of regulators, payers, and HTA agencies.

CER is generating more of certain types of evidence that have been of increasing relevance to HTA. Among the broad set of attributes or impacts assessed in HTA, those that CER emphasizes are effectiveness and safety in real-world patients and health care settings, patient outcomes, and direct comparisons to standards of care. Further, CER is contributing to tools and methods that will broaden the scope and strengthen the quality of available evidence, including the development of better health outcome and quality of life measures, observational data sources such as registries and insurance claims, and alternative designs for clinical trials. As such, when HTA develops evidence questions, conducts systematic literature searches, conducts meta-analyses and other evidence syntheses, and develops findings and recommendations, it will draw on an evidence base enriched by CER findings.

D. Patient-Centered Outcomes Research and HTA

HTA is directly and indirectly related to certain patient-oriented concepts, including **patient-centered care**, **patient-centered outcomes**, **patient-reported outcomes**, and **patient-centered outcomes research**.

The term "patient-centered medicine" was introduced more than 40 years ago by Balint and colleagues (Balint 1970) in contrast to what they called "illness-centered medicine." Whereas "traditional diagnosis" was based on disease-centered thinking, an "overall diagnosis" involved an understanding of

the patient's complaints based on patient-centered thinking (Stewart 2013). A contemporary definition of **patient-centered care** is:

> *The experience (to the extent the informed, individual patient desires it) of transparency, individualization, recognition, respect, dignity, and choice in all matters, without exception, related to one's person, circumstances, and relationships in health care (Berwick 2009).*

The main dimensions of patient-centered care include:

- Disease and illness experience (patient-as-person)
- Whole person (biopsychosocial perspective)
- Common ground (sharing power and responsibility)
- Patient-doctor relationship (therapeutic alliance) (Hudon 2011; Mead 2000; Stewart 2013)

As a particular form or variant of health care, patient-centered care can have a favorable impact on safety, effectiveness, and related health care outcomes, although the evidence to date of its impact is mixed (Street 2012; Zandbelt 2007). As such, patient-centered care can be a topic of HTA.

Patient-centered care can be evaluated using such measures as:

- Patient Perception of Patient-Centeredness (PPPC): patient perceptions of patient-centered care during the last visit with a family physician; uses 14 items
- Consultation Care Measure (CCM): patients' perceptions of patient-centered care during the last visit with a family physician; 5 subscales: communication and partnership, personal relationship, health promotion, positive and clear approach to the problem, interest in effect on life
- CAHPS (Consumer Assessment of Healthcare Providers and Systems): brief general measure comparing overall quality of interpersonal care across health care settings; includes some patient-centered care domains: access (getting care quickly, getting needed care), provider communication (Epstein 2011; Hudon 2011)

In addition, patient-centeredness may be considered a dimension of health care quality (Berwick 2009), i.e., a standard attribute of practice.

Patient-centered outcomes (or patient-oriented outcomes) comprise a subset of the broad array of health outcomes noted above (mortality, morbidity, adverse events, quality of life, etc.). Patient-centered outcomes refer to outcomes that patients experience across the variety of real-world settings, including: survival, functional status, quality of life, quality of death, symptoms, pain, nausea, psychosocial well-being, health utility (patient-perceived value of particular states of health), and patient satisfaction. Excluded are outcomes or other endpoints that patients do not experience directly, e.g., blood pressure, lipid levels, bone density, viral load, or cardiac output. Patient-centered outcomes can be assessed at a generic level or a disease/condition-specific level. Examples of generic instruments for assessing patient-centered outcomes include:

- CAHPS
- EuroQol (EQ-5D)
- Functional Status Questionnaire (FSQ; 34 items)

- Health Utilities Index
- Nottingham Health Profile
- Quality of Well-Being Scale
- Short Form (12) Health Survey (SF-12)
- Short Form (36) Health Survey (SF-36)
- Sickness Impact Profile (SIP; 136 items)

There are patient-centered outcomes instruments for such diseases and conditions as angina, asthma, epilepsy, kidney disease, migraine, multiple sclerosis (MS), and vision. For example, standard outcomes for assessing MS activity and severity include annualized relapse rate and the Expanded Disability Status Scale. Patient-centered outcomes instruments that may better reflect patient experience with MS include MS-specific instruments and hybrid instruments (i.e., generic instruments with additional MS items). Examples of MS-specific patient centered outcome instruments are: the Multiple Sclerosis Quality of Life Inventory (MSQLI), Multiple Sclerosis Quality of Life-54 scale (MSQoL-54), Functional Assessment of Multiple Sclerosis (FAMS), Multiple Sclerosis Impact Scale (MSIS-29), and Leeds Multiple Sclerosis Quality of Life scale (LMSQoL) (Miller 2010).

Patient-reported outcomes (PROs) are those patient-centered outcomes that are self-reported by patients (or reported on their behalf by their caregivers or surrogates) or obtained from patients by an interviewer without interpretation or modification of the patient's response by other people, including clinicians (Rothrock 2010; US FDA 2009). Patient ratings of health status, adverse events, functional status, quality of life, and other outcomes may differ from measures recorded by clinicians (see, e.g., Ebell 2004; Nakamura 2012), and patients may weigh benefits and risks differently among themselves as well as differently than clinicians when considering treatment alternatives (Wong 2012). Non-patient reported outcomes are those that matter to patients but are not reported directly by them.

Patient-centered outcomes research (PCOR) generates evidence comparing the impact of health care on patient-centered outcomes. PCOR can draw on a wide variety of methods, including, e.g., practical or pragmatic RCTs, cluster randomized trials, and other trial designs; registries, insurance claims data, systematic reviews, and others (Methodology Committee 2012). As described by the recently established Patient-Centered Outcomes Research Institute (PCORI) in the US, PCOR answers these patient-focused questions:

- Given my personal characteristics, conditions and preferences, what should I expect will happen to me?
- What are my options and what are the benefits and harms of those options?
- What can I do to improve the outcomes that are most important to me?
- How can clinicians and the care delivery systems they work in help me make the best decisions about my health and healthcare? (Patient-Centered Outcomes Research Institute 2013)

In summary, HTA can assess patient-centered care as a type of intervention. HTA can assess the impact of health care interventions on patient-centered outcomes, including patient-reported outcomes, of the full range of health technologies. Using a variety of primary and secondary data methods, patient-centered outcomes research generates data and evidence that can be used in HTA.

E. Personalized Health Care and HTA

Clinical trials and other studies that report only average treatment effects may be misleading. Therapies that yield a statistically significant treatment effect across a study population may not necessarily work for all treated patients; they may be ineffective for some patients and harmful for others. Also, therapies that do not yield a statistically significant treatment effect across a study population—and that may be dismissed as ineffective—may work for certain subsets of the population.

Personalized health care (a broader concept that personalized medicine) refers to the tailoring of health care to the particular traits (or circumstances or other characteristics) of a patient that influence response to a heath care intervention. These may include genetic, sociodemographic, clinical, behavioral, environmental, and other personal traits, as well as personal preferences. Personalized health care does not mean the creation of interventions that are unique to a patient; rather, it recognizes differences in how patient subgroups respond to particular interventions, and uses that information to treat individual patients.

Some examples of technologies used in personalized health care are:

- *CYP2C9* and *VKORC1* genetic testing for warfarin anticoagulation response for patients with atrial fibrillation, mechanical heart valves, deep vein thrombosis, etc.
- *HER-2/neu* receptor testing for trastuzumab for breast cancer
- *BRCA 1,2* testing for pharmaceutical and surgical prevention options for and surveillance for breast cancer
- *KRAS* testing for use of EGFR inhibitors (e.g., cetuximab, panitumumab) for colon cancer
- Socioculturally-tailored therapy to treat certain ethnic minority patients with diabetes and depression (Ell 2011)
- Alternative procedure techniques (gastric banding, gastric bypass, etc.) for bariatric (morbid obesity) surgery
- Alternative types of coronary artery revascularization (e.g., coronary artery bypass graft surgery, percutaneous coronary interventions) for symptomatic ischemic coronary artery disease
- Alternative regimens to treat infertility

In order to support personalized health care, information is needed about how alternative technologies affect not just the average patient with a given health problem, but how those technologies affect various subgroups of patients with that health problem. For example, more specific information about how response differs by age, sex, comorbidities (other health problems), or genetic traits can be used by patients with a certain health problem who share those characteristics. Heterogeneity of treatment effects (HTEs) refers to the variation in patient responses observed across levels or types of patient characteristics such as these (Kravitz 2004).

HTA is increasingly oriented to identifying, assembling, and evaluating evidence on HTEs. This applies to all types of technologies, including screening, diagnostic, therapeutic, palliative, and others. Deriving findings about effects on subgroups depends in large part on the availability of data from studies that have been designed to detect such subgroup differences. This depends not only on how well the study population represents various subgroups, but whether the study designs and methods of subgroup

analysis are capable of detecting HTEs for the patient subgroups of interest. For example, prospective subgroup analyses (i.e., identification of subgroups prior to rather than after data collection) tend to be more reliable than retrospective ones, and sample sizes for the subgroups under study must be large enough to detect true subgroup differences where they exist (Oxman 1992; Wang 2007.) Meta-analyses and other integrative methods may be used to pool subgroup data from different studies. HTA can also help to strengthen the evidence base for personalized health care by encouraging the development, validation, and use of patient-centered (including patient-reported) outcome measures; involvement of patients in planning clinical trials; and use of alternative data sources, such as health services utilization data (including from insurance claims) and patient registries, to help identify potential subgroup effects.

F. Patient and Consumer Involvement in HTA

The field of HTA increasingly recognizes the importance of involving patients and other "non-professional" or "non-expert" individuals. Indeed, these individuals often have developed expertise about their disease, condition, and care. Their involvement can provide perspectives and experiences that are not available from others, but are integral to HTA. Their participation in HTA also helps to fulfill HTA good practices for openness, transparency, and accountability. The value placed on patient participation and input should be commensurate with that of other participants in HTA.

In this context, the terms "patients" and "patient representatives" may also be used to mean patient advocates, consumer representatives, family members, non-professional caregivers, citizens, and the public. These individuals and groups may have different interests and bring different perspectives. While recognizing these differences, this section generally uses the broadly inclusive term "patients" to refer to one or more of these groups. Also, the terms "involvement," "engagement," and "participation" of patients are sometimes used interchangeably, although some users consider the term "involvement" to be more inclusive or integral to the HTA process than the other terms. This section uses all of those terms, and "involvement" is used broadly to encompass various levels of inclusiveness in the HTA process.

In most countries, patient involvement in HTA is still emerging, with considerable variation in how patients are recruited and supported for involvement, the HTA processes in which they are involved, the role they have in determining HTA findings, and how potential conflicts of interest are addressed (Kreis 2013). Across HTA organizations, patients are participating in more aspects of HTA, including various combinations of the following (see, e.g., Facey 2010; HTAi Patient and Citizen Involvement Interest Sub-Group, Good Practice Examples 2012):

- Servings as members of HTA boards, committees, and workgroups
- Identifying potential topics for HTA
- Early identification of potential target groups for HTA reports
- Setting priorities among HTA topics
- Identifying health outcomes and other impacts (economic, social, etc.) to be assessed
- Reviewing proposals or bids by outside organizations/contractors to conduct HTAs
- Submitting evidence for HTAs
- Reviewing draft HTA reports and recommendations
- Helping to design and prepare patient-friendly HTA report summaries

- Dissemination of HTA findings to policymakers, patient groups, and other target groups
- Evaluating the use of HTA recommendations

Patients are also increasingly involved in helping to set research agendas and design clinical trials and other primary data collection (e.g., selecting health outcomes and other endpoints for study that are of particular relevance to patients).

A systematic review of 142 studies published through 2011 of patient engagement in health care research more broadly found that most patient engagement occurred in agenda setting and protocol development, with less occurring during the conduct and translation of research. The review reported that patient engagement improved clinical trial enrollment and retention, design of study protocols, selection of relevant outcomes, and securing research funding. This review found no studies comparing methods of patient engagement. The main challenges identified included the time and funding to implement patient engagement and concerns by some stakeholders that patient engagement is "tokenistic," i.e., not genuine (Domecq 2014).

According to a 2013 report of an international web-based survey, a majority of HTA organizations have formal public involvement or engagement activities, most of which involve lay (i.e., non-professional) representatives of organized groups or, to a lesser extent, individual patients, consumers, and community members. Most public involvement or engagement activities have entailed communication or consultation with those members of the public, while some activities have entailed direct participation in the form of citizens' juries, consensus councils, and similar groups (Whittey 2013).

Through direct involvement in the HTA process, patients can share experiences, perspectives, and evidence that complement or even contradict evidence from such traditional sources as clinical trials and clinicians' observations. Certainly, direct patient involvement can improve the understanding and accuracy of patients' needs and preferences. Further, patient experience, perspectives, and evidence might address the burden of a disease or condition on patients, family members, and caregivers, including dimensions of health, pain, quality of life, social and economic burdens. Such patient input might also address the impacts on patients and others of health technologies, such as benefits and harms compared to alternative therapies, ability to work, difficulties in access or adherence, impact on their life, and economic impact, including any out-of-pocket costs, associated travel costs, income loss, and caregiver costs. Formal patient evidence can be obtained via qualitative or quantitative means through primary data collection or secondary analyses (Bastian 2011; Danner 2011; Facey 2010; Health Equality Europe 2008).

An analysis published in 2009 of a random selection of HTA reports (published in English or Scandinavian languages) from INAHTA-member agencies found that about one-third addressed patient-related issues (i.e., concerning patient views on the use or impact of technologies). These mainly addressed psychological issues such as patient worry and discomfort, ethical issues such as patient acceptance of the technology, social issues such as impact on patient daily life, effect on quality of life, and patient information. There was great variation in the comprehensiveness and methodological transparency of assessment of patient-related matters; only a minority of reports had explicit questions about patient issues (Lee 2009).

One of the more comprehensive efforts to gain patient input is that used for technology appraisals of the UK National Institute for Health and Care Excellence (NICE). Patients and caregivers provide input

on: the personal impact of their condition; outcomes patients want the technology to improve; the impact of a technology on health outcomes, symptoms, physical and social functioning, quality of life; impact on family, friends and employers; ease of use, side effects and costs of the technology; patient preferences; subgroups who might benefit more or less from the technology; challenges to professional or researcher views; and areas needing further research (HTAi Patient and Citizen Involvement Interest Sub-Group, Good Practice Examples 2012; NICE Guide to the Methods 2013). Among other forms of patient and citizen input, NICE has a Citizens Council of 30 members that is demographically representative of the UK. The Citizens Council provides NICE with public perspectives on moral and ethical issues of which NICE takes account when producing its guidances (NICE Citizens Council 2014).

The Consumer Network of the Cochrane Collaboration involves health care consumers in the production of Cochrane systematic reviews. The most common role is consumer referee, which involves, e.g., commenting on protocols (plans) for new Cochrane systematic reviews, commenting on Cochrane reviews prior to publication, and helping to prepare or provide consumer perspectives on the plain language summaries of Cochrane reviews. Consumers may also help to search journals, disseminate information about particular Cochrane reviews, co-author reviews, and translate reviews and plain language summaries (Cochrane Consumer Network 2014).

Based on recent survey findings from HTA organizations, some of the main challenges to patient engagement in HTA include: lack of organizational capacity to involve patients, extensive time requirements, lack of agreed-upon methods for involving patients in HTA, lack of clarity about when in the HTA process to involve patients, lack of funding, and concerns regarding the scientific credibility of patient evidence (European Patients' Forum; undated).

Greater experience with patient involvement in HTA and information sharing about these are leading to good practices for involvement of patients as members of boards, committees, and working groups as well as involvement and communications with patient groups and the broader public. These emerging good practices address such areas as those shown in **Box X-1**. While it may be impractical for most HTA programs to adopt all or even most of these immediately, those programs seeking to benefit from greater patient involvement can adopt more of them as time and resources permit. Also, greater involvement of patients and other target populations is gradually extending to development of evidence-based clinical practice guidelines, including to incorporate information about PROs, values, and preferences (see, e.g., Basch 2011; Brouwers 2010).

The HTAi Interest Sub-Group on Patient and Citizen Involvement in HTA has developed a set of resource materials, including examples of good practices for patient and citizen involvement from various HTA agencies, video and slide presentations introducing HTA and patient roles in it, a glossary, answers to frequently-asked questions, and others (see: http://www.htai.org/index.php?id=730#c2785).

Box X-1. Emerging Good Practices for Patient Involvement in HTA

- Ensure that all HTA processes are appropriately accessible, transparent, fairly applied, documented, and accountable
- Gain commitment of HTA organization to patient involvement
- Establish, review, and update policies for patient involvement in HTA
- Establish a dedicated unit or staff function for patient involvement, with explicit channels of communication and personal points of contact in the HTA organization
- Plan and provide for adequate budgets and other resources for patient involvement
- Conduct broadly accessible, transparent, and equitable (including relative to other types of participants) recruitment of patients to HTA boards, committees, and working groups
- Provide training and education for patient involvement in HTA, including appropriate materials, handbooks, reviewer/evaluation forms, etc.
- Provide training for HTA staff and other participants on working with patients and groups that represent them
- Ensure appropriately equitable rights (e.g., in voting) and responsibilities for patient members of HTA boards, committees, and working groups
- Provide adequate notice of meetings, due dates for input, and related scheduling information to patient participants to enable their sufficient preparation and timely participation
- Inform the variety of patient groups and other potentially interested groups about planned and ongoing HTA reports and ways to provide input to these reports
- Communicate in plain, patient-appropriate language, including in background materials and other documentation to ensure that patients are adequately informed for participation in HTA
- Provide patient participants with adequate financial support, including fees commensurate with those for other participants and reimbursement for travel and other out-of-pocket expenses (e.g., for child and other dependent care)
- Ensure that the context for patient involvement is welcoming and is not intimidating or coercive
- Make reasonable accommodations to ensure accessibility to HTA processes for patients and other participants who may have activity limitations (or are otherwise differently-abled)
- Provide HTA reports (and/or summaries of these) and other outputs in plain, patient-appropriate language and format, including for users with sensory (e.g., visual or hearing) limitations
- Provide for active, targeted dissemination of HTA reports and related products to patient groups
- Provide links from HTA organization web sites to those of relevant patient groups
- Organize special/targeted events, communications, or other activities for engaging individual patients and patient groups
- Seek to involve patient groups that are traditionally difficult to reach (e.g., ethnic minorities, underserved populations, stigmatized populations)
- Document/report patient involvement in HTA, including, e.g., levels of participation, impact on HTA deliberations, findings, and recommendations
- Provide recognition and feedback to patients regarding their participation

Sources: European Patients' Forum (undated); Health Equality Europe 2008; HTAi Patient and Citizen Involvement Interest Sub-Group PIE Good Practice Principles 2012, Good Practice Examples 2013.

G. Rapid HTA

In recent years, the demand for HTA by health care decision makers has increasingly involved requests for faster responses to help inform emergent decisions. This has led to development of "rapid HTAs" (or "rapid assessments" or "rapid reviews"). Among the reported purposes of rapid HTAs are to inform coverage decisions, capital funding, formulary decisions, treatment referrals, guideline formulation, and indications for further research (Hailey 2009). For these, HTA organizations generate more focused assessments in the course of, e.g., four-to-eight weeks. Rapid HTAs offer a tradeoff: provide less-than-comprehensive and less certain information in time to act on a decision versus comprehensive and more certain information when the opportunity to make an effective decision may have passed. In addition to the shorter completion time, rapid HTAs can differ from full HTAs in such ways as: limiting scope to fewer types of impact or evidence questions, focusing searches on fewer bibliographic databases, relying on fewer types of studies (e.g., only systematic reviews or only RCTs), use of shorter and more qualitative syntheses with categorization of results without meta-analyses, and more limited or conditional interpretation of findings or recommendations (Khangura 2012).

Despite the increased use of rapid HTAs on diverse topics and a growing literature on their responsiveness and utility for decision makers, there is no consensus regarding methodology or guidance for rapid HTAs (Harker 2012). EUnetHTA has a version of rapid HTA that focuses on four of the nine total domains of its HTA core model, i.e., health problem and current use of technology, description and technical characteristics, safety, and clinical effectiveness (EUnetHTA Joint Action WP5 2013). The Ottawa Hospital Research Institute has developed an eight-step approach to rapid reviews, including: needs assessment, question development and refinement, proposal development and approval, systematic literature search, screening and selection of studies, narrative synthesis of included studies (including assignment of evidence levels), report production, and ongoing follow-up and dialog with knowledge users (Khangura 2012). Given the need to tailor rapid HTAs to the particular needs and time constraints of decision-makers, providing transparency of various rapid HTA methodologies is likely to be more important that achieving consensus on any standard approach (Watt 2008).

H. Decentralization of HTA

Although technology assessment originated as a mostly centralized function conducted by government agencies and other national- or regional-level organizations, HTA evolved into a more decentralized function, conducted by a wide variety of organizations in the public and private sectors (Goodman 1998; Rettig 1997). As noted above, an HTA done from a particular perspective may not serve the policymaking needs of other perspectives. Even for the same technology or clinical problem, there can be widely different assessment needs of regulatory agencies, health technology companies, hospitals, payers, physicians, policymakers, and others.

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, for example, there remain multiple government centers with ongoing HTA or other technology evaluation responsibilities to fulfill particular purposes, e.g., FDA regulation of drugs, biologics, and medical devices; CMS coverage policies for the Medicare program; and the Effective Health Care Program, Technology Assessment Program, and US Preventive Services Task Force 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. Large health care provider institutions and major health care product companies have established groups or units devoted to

"technology assessment," "**pharmacoeconomics**," "clinical effectiveness," and "health economics and outcomes research," and related areas. More health plans (including various managed care organizations and insurance companies) have established formal programs to assess new pharmaceuticals, procedures, and other technologies in support of payment decisions. The number and magnitude of private firms and university centers involved in HTA and related technology evaluation functions continue to increase. HTA and evidence-based clinical practice guideline 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 and formularies), equipment and other technologies, and other technology-related needs throughout their systems (Kaden 2002).

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 Medicare Evidence Development & Coverage Advisory Committee (MEDCAC) of independent experts that provides findings and recommendations to CMS pertaining to the quality of available evidence to help inform national coverage policies for health technologies, based on the clinical literature, and HTAs and "evidence reports" prepared by selected AHRQ Evidence-based Practice Centers.

AHRQ's Evidence-based Practice Centers (EPC) program has contracts with 11 EPCs, most of which are affiliated with academic health centers, that generate HTAs and evidence reports in support of clinical practice guidelines, coverage policies, and other practices and policies. Some EPC reports are conducted at the request, via AHRQ, from 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 EPCs, on behalf of government and non-government organizations. While the AHRQ EPC program is a decentralized model, it also benefits from collaborative efforts to develop and use standardized approaches, such as for evidence appraisal and other aspects of conducting systematic reviews (see, e.g., Methods Guide for Effectiveness 2014). As noted above, AHRQ also administers the US Preventive Services Task Force, 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, another highly decentralized model for evaluating health care, involves 53 workgroups of volunteer experts from more than 100 countries coordinated through 14 centers with additional branches based in about 30 countries that conduct systematic reviews of a diverse variety of health care interventions. As do the AHRQ EPCs, the Cochrane Collaboration develops standardized approaches to conducting systematic reviews and related methods development (Higgins 2011).

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 diminish concerns that assessments reflect narrow or self-serving interests of a particular agency or organization.

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 US Department of Veterans Affairs (VA) continually seek to build economies of scale and buying leverage

for health care products, ranging from drugs to 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 health technologies and related services.

As health care providers and payers understand the resource requirements for conducting evaluations of health care technologies, they consider the respective benefits of conducting their own HTAs and subscribing to HTA reports from outside assessment groups. Assessment requirements vary widely depending on the type of technology involved. Acquisition of commodity products such as most types of syringes, surgical gloves, and generic drugs is based largely on price, whereas acquisition of the latest deep-brain stimulator for movement disorders such as Parkinson's disease requires a more considered evaluation of safety, effectiveness, cost, patient preferences, 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 products. The large GPOs 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 depends on factors other than price alone. In turn, many GPOs also subscribe to HTA report services (Lewin Group 2002; Hu 2012).

I. Locus of Assessment: Make or Buy?

Health care decision makers can "make or buy" HTAs. The nature of an assessment problem will affect the determination of the most appropriate organization to conduct it. 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 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 and therapeutics (P&T), strategic planning, and capital planning committees (Kaden 2002).

Other organizations rely on HTA reports acquired from organizations that have devoted functions or otherwise specialize in HTA. As noted above, AHRQ commissions evidence reports, HTAs, and other reports from its EPC, most of which are affiliated with academic medical centers. In the UK, NICE commissions one internal center and four external centers (the National Clinical Guideline Centre, National Collaborating Centre for Cancer, National Collaborating Centre for Women's and Children's Health, and National Collaborating Centre for Mental Health) to produce its clinical guidelines. NICE also commissions designated academic centers to develop assessments/evidence reports in support of its various types of technology appraisals and guidance.

Other US-based vendors of HTA include, e.g., Blue Cross and Blue Shield Association Technology Evaluation Center, Cochrane Collaboration, ECRI Institute, Hayes Inc., Institute for Clinical and Economic Review (ICER), Center for Medical Technology Policy (CMTP), Institute for Clinical Systems Improvement (ICSI), and the Oregon Drug Effectiveness Review Project (DERP). Depending upon the producing HTA organization, these reports may be available at no cost, for members only, on a subscription basis, or for a specific price per report.

Determining the responsibility for sponsoring or conducting an assessment depends on the nature of the problem, financial resources available, expertise of available personnel, time constraints, and other factors. For any assessment, an organization must determine to what extent it will conduct the HTA itself or commission it from other sources. Some organizations commission selected components of an HTA, such as evidence retrieval and synthesis, and perform the other steps in-house.

One of the advantages of requesting or commissioning an outside group to conduct HTAs is to gain an independent, outside view in instances 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 regarding its willingness to pay for the use of 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 involved, does that consensus need to incorporate and reflect the opinions of an organization's own clinicians? Will local clinicians accept the results and report recommendations if they do not participate in the assessment?

J. Underused Technologies and HTA

When used properly, HTA can result in the reduction or elimination of the use of technologies that are not safe and effective, or whose costs are too high relative to their benefits. Less attention is given to the ability of HTA to identify technologies that are underused and to determine why they are underused (Asch 2000; McNeil 2001). Underuse is prevalent in preventive, acute, and chronic care; applies to care for children, youth, and adults; and contributes to tens of thousands of deaths and billions of dollars of losses to the economy and unnecessary health care costs (GAO 2002; Mangione-Smith 2007; McGlynn 2003). According to one estimate, replacing more costly, non-superior health care with proven cost-effective care would save approximately one-fourth of US health care spending (Fuchs 2011).

Underuse of some proven therapies affects large proportions of at-risk populations. For example, there is high-quality evidence that smoking cessation interventions, including nicotine replacement therapy, the antidepressant bupropion, and counseling, are safe, effective, and cost effective (Anderson 2002; Foulds 2002; Stead 2008; Woolacott 2002). However, in Europe, North America, and elsewhere, these interventions are used far less than is indicated. Two further examples in the US are management of hypertension and childhood vaccinations, as follows.

Based on data from the US National Health and Nutrition Examination Survey (NHANES) 2003-2010, the Centers for Disease Control and Prevention (CDC) estimates that the prevalence of hypertension among US

adults was 30.4%, or about 66.9 million people. (Hypertension was defined as an average systolic blood pressure ≥140 mmHg or an average diastolic blood pressure ≥90 mmHg, or currently using blood pressure-lowering medication.) Among those with hypertension, 53.5% did not have their hypertension controlled. Among these people, 39.4% were not aware of their hypertension, 15.8% were aware of their hypertension but were not receiving medication for it, and 44.8% were aware of their hypertension and were being treated with medication (CDC 2012). Hypertension accounted for an estimated $131 billion in direct US health care costs in 2010, including its contribution to costs of complications associated with other cardiovascular conditions (Heidenreich 2011).

Vaccination rates in the US remain highly variable by vaccine type and age group, although these rates have generally increased in recent years. For example, according to 2012 National Immunization Survey data of the CDC, vaccination rates (with minimum appropriate number of doses) for children 19-35 months old were at or near the US objective of 90% for poliovirus (92.8%); measles, mumps, and rubella (90.8%); varicella (90.2%), and hepatitis B (89.7%). Rates were lower for other vaccines, including diphtheria, tetanus and pertussis (82.5%); pneumococcal conjugate vaccine (81.9%); Haemophilus influenzae b (80.9%); and hepatitis B birth dose (71.6%). Rates were low for hepatitis A (53.0%, vs. target of 85%), and rotavirus (68.6%, vs. target of 80%). Rates of vaccination coverage were generally lower for children living below poverty level, including 6.0% to 9.5% lower for several of these vaccination types (CDC, National, state, and local area vaccination coverage among children, 2013).

Underuse of particular technologies is attributed to various reasons. For smoking cessation, these reasons include: 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) or 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). Available evidence suggests, for example, that greater insurance coverage of smoking cessation interventions increases the likelihood that smokers will stop smoking (Reda 2012).

Rates of underuse can be highly variable in a country or region. In the instance of pediatric vaccinations in the US, underuse is affected by such factors as differences in socioeconomic status, payment mechanisms, and shortages of certain vaccines (CDC, National, state, and local area vaccination coverage among children, 2013; CDC, National and state vaccination coverage among adolescents, 2013; Hinman 2006). In developing countries, the major cause of vaccine shortages is lack of resources to purchase them (Hinman 2006).

Box X-2. Some 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
- Childhood vaccinations
- 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
- Hemoglobin A1c testing every 6 months in diabetic patients
- Hepatitis B virus vaccination of infants
- HIV testing and treatment with antiretroviral therapy
- Hypertension management
- Implantable cardioverter-defibrillators for survivors of cardiac arrest
- Incontinence diagnosis and treatment
- Influenza immunization
- Inhaled corticosteroids in adults with asthma
- Intraocular pressure screening for glaucoma
- Mammography (especially women age 50+)
- Oral rehydration therapy for dehydrated children
- Organ transplantation
- Pain management
- Pap smears
- Pneumococcal vaccine for high-risk patients
- Prenatal care
- Smoking cessation interventions
- Thrombolytic therapy for acute myocardial infarction
- Thrombolytic therapy for ischemic stroke

Box X-2 shows examples of health technologies for which good evidence exists of safety, effectiveness, and 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 wealthy 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 or ineffective information dissemination
- Limited coverage policies (e.g., for preventive services)
- Limited access to primary care
- Poor socioeconomic status
- Fee-for-service payment systems that reward use of high-paying services without regard for cost-effectiveness
- 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, patients, or the public (e.g., false information suggesting that vaccines increase the risk of other disorders)
- Clinical inertia, i.e., unwillingness to change practice habits
- Insufficient supply (e.g., organs for transplantation)
- Health technology companies' marketing efforts to emphasize brand name products rather than less-expensive technologies that are at least as effective and safe
- Disproportionate concerns about adverse effects (e.g., warfarin to reduce risk of stroke)
- Concerns about **adherence** to therapies by some patient subgroups
- Fear of stigma (e.g., treatment of mental health disorders)
- Professional conflicts and "turf battles" on the part of physician specialists, provider institutions, industry, and others

Merely generating better evidence is not enough to meet the decision-making needs of consumers, patients, health care providers, and purchasers. To maximize its impact on the quality and value of health care, these parties must cite and use evidence when making clinical and policy decisions. Disseminating evidence into clinical practice must be accompanied by ongoing evaluation and feedback to decision makers, the key characteristic of a true learning health care system. However, this is not happening consistently. In a review of adherence to 439 indicators of health care quality for 30 acute and chronic conditions as well as preventive care, McGlynn and colleagues concluded that American adults received only 55% of recommended care (McGlynn 2003). The same study found that children and youth received only 46.5% of recommended care (Mangione-Smith et al., 2007).

These gaps between evidence and execution underscore the need to identify more effective tools to help patients, providers, and policy makers to use the available evidence. Those in HTA must consider how implementation of HTA findings and recommendations, including the extent to which technologies may be underused, overused, or inappropriately used, may be affected by such factors as those listed above. Data suggesting underuse of health technologies known to be safe, effective, and cost-effective may be used to develop priorities for HTA and evidence questions. Explicit attention to underuse of proven technologies conveys that HTA is concerned with optimal use of proven health technologies, not

just reducing use of inappropriate ones or the mistaken view that it seeks to diminish innovation or diffusion of health technology.

K. Managed Entry and HTA

When HTA is used to inform reimbursement decisions about a new technology (or new application of an existing technology), it may encounter promising, yet non-definitive evidence regarding effectiveness, safety, economic impacts, or other attributes that are important to payers as well as other stakeholders. In these instances, rather than delaying any coverage until stronger evidence is available, payers may seek an arrangement for "**managed entry**" of the technology. These arrangements offer several forms of potential societal benefit. They can enable access for certain types of patients for whom existing evidence suggests net health benefit, provide some financial compensation for generating better evidence sooner than in the absence of reimbursement, enable refinement of clinical technique and services delivery, and build expertise and experience among physicians and other providers.

Some government and private sector payers have provided certain forms of coverage for selected "investigational" or "experimental" technologies since the 1990s in order to compile evidence for making more informed coverage policies (Beebe 1997; Brenner 2002; McGivney 1992; Medical Technology Leadership Forum 1999; Sheingold 1998; Wood 2001). More recently, such forms of "conditional coverage" have been subsumed under the broader term, "managed entry."

Intentionally or otherwise, payers have long been providing reimbursement for investigational technologies, including prior to the 1990s in some instances. In the US, such instances have included reimbursement for many off-label uses of drugs approved for other indications, drugs with "treatment investigational new drug" (IND) designations by the FDA, certain devices being evaluated under investigational device exemptions (IDE) designations by the FDA, certain medical and surgical procedures being provided in qualifying clinical trials, and certain technologies whose coverage has been legislatively mandated (Steinberg, Tunis 1995).

M**anaged entry** refers to a range of innovative payment approaches that provide patient access under certain conditions. Three main purposes are to manage: uncertainty about safety, effectiveness, or cost effectiveness; budget impact; and technology utilization for optimizing performance (Klemp 2011). In one general framework of managed entry (Carlson 2010), two main types that are based, at least in part, on health outcomes are conditional coverage and performance-linked reimbursement. Conditional coverage includes **coverage with evidence development (CED)** and conditional treatment continuation.

CED, a broad category of managed entry, refers to limited coverage for a new technology in parallel with specified data collection to provide stronger evidence about the effectiveness, safety, or other impacts of the technology of interest to payers. This enables some patient access to the technology while reducing uncertainty through real-world experience and generating evidence to inform revised coverage policies as well as clinical practice decisions (Trueman 2010). CED includes coverage "only in research" (i.e., coverage for a technology only when used in members of the payer's patient population who are participating in clinical trials of the technology) and coverage "only with research" (i.e., coverage for a technology only when also being used contemporaneously in a sample of the payer's patient population participating in clinical trials of the technology). Two types of performance-linked reimbursement are outcomes guarantee, where payment (or rebates) depend on achievement of predetermined health outcomes targets, and pattern or process of care, such as demonstrated impact on clinical decisions or patient adherence to prescribed regimens (Carlson 2010).

Certainly, various forms of managed entry can help to optimize use of a technology. CED and other forms of managed entry have had various levels of success to date, and face certain practical challenges (Berger 2001; Cooper 2001; Hutton 2007; Miller 2008). These arrangements can be difficult to negotiate and manage, and further evaluation is needed to assess their ability to achieve their purposes (Klemp 2011).

L. Innovation and HTA

Innovation generally refers to creation or development of something that is new or different. It is something that addresses the demand of, is accepted or adopted by, or is otherwise beneficial to users or markets. Innovations usually are in some way claimed to be more effective or otherwise better than existing technologies, thereby offering a meaningful desirable change. In health care and other fields, meaningful change is true (i.e., statistically significant) and of practical importance (i.e., clinically significant). To be successful in a market, innovation must be replicable at an acceptable cost.

Innovation in health care shares attributes of other fields. Examples of such attributes include: better (more effective), safer, faster, cheaper, easier to use, smaller, portable, more standardized, more reliable, and more environmentally friendly ("greener"). In recent years, such attributes as targeted, personalized, less-invasive, and secure (or confidential) have been increasingly sought in health care. HTA can account for all of these attributes of innovation. In particular, though, as a form of inquiry intended to support decision making and policymaking, HTA is especially oriented toward discerning improvements in health care outcomes, i.e., "better" and "safer." Also, where decision makers and policymakers seek such information, HTA is oriented toward meaningful improvements in health care outcomes per incremental health expenditure. (In some decision-making contexts, the inverse is of interest, i.e., whether the cost per incremental improvement in health outcomes is acceptable.) In health care, this is **innovation of value** (Porter 2010).

Certainly, HTA is subject to ongoing controversy pertaining to the relative value of investing in socially desirable goals when resources are scarce. For example, a goal of allocating resources efficiently for the broad population may compete with the goal of allocating resources for those who are most severely affected by ill health. Such tradeoffs or conflicting goals can affect what is considered to be innovation of value. The value of innovation may be considered to extend beyond health care impacts to sustaining the progress of science and technology and to employment, balance of trade, and other economic attributes. These issues pertaining to the value of innovation have arisen in various national HTA programs, such as for NICE in the UK (Green 2010, Kennedy 2009).

HTA could be viewed as yet another hurdle for innovation, beyond proof-of-concept, regulatory approval, and coverage (based merely on whether a technology qualifies as a covered benefit). Indeed, "new or different" may suffice for those hurdles without HTA; in most major regulatory schemes, a new technology can be approved or cleared for market as long as it functions as labeled, or is efficacious at an acceptable level of safety for a particular patient population, without regard to "value." When HTA inquires about tradeoffs of health outcomes and costs, it supports alignment of payment with innovation of value. Further, it sends signals that may help to improve the efficiency of innovation by focusing efforts on developing and validating innovations that will be valued by purchasers on behalf of clinicians, patients, and provider institutions. This emphasis is shifting the direction of innovation, including diminishing or eliminating development with little prospect of value as well as creating new opportunities. Some technologies that would have been "innovative" may no longer be developed because they will not provide sufficient value; their reimbursement outlook will not justify further investment. At the same time, the ability to demonstrate comparative effectiveness at a lower cost or

superiority at an acceptably higher cost can confer market advantages. As the "blockbuster" model of drugs and other technologies for broad populations (e.g., statins for lowering blood cholesterol and certain antidepressants) fades, there are opportunities to develop technologies targeted to patient subgroups (e.g., identified using genomics) that may merit high prices.

Some innovations are "disruptive." Disruptive innovations alter and even displace existing systems, networks, or markets, and may create new business models and lead to emergence of new markets (Bower 1995). The disruption may not be inherent in a new technology itself, but in how it is marketed. Examples in health care are: diagnostic imaging replacing exploratory surgery, ultrasound replacing certain x-ray imaging, percutaneous coronary interventions substituting for some coronary artery bypass graft surgeries, and gene therapy substituting for some pharmaceutical or biological therapies (Wilson 2012). Disruptive innovations can present challenges to existing systems of regulation, payment, health care delivery, and professional training. Consistent with the original scope of technology assessment to consider the broad array of potential impacts of technology, HTA should consider not just whether a new technology is safe, effective, or cost-effective, but how its adoption might affect systems of care; where, how, and by whom care is delivered; and what direct and indirect effects it may have on patients, at-risk populations, use of other technologies, access to care, and broader societal impacts.

Various forms of process innovation in clinical trials, such as adaptive and Bayesian trial designs, and in manufacturing technologies also contribute to value by increasing standardization, improving product quality, and lowering production costs. As health technology companies and other innovators better perceive the need to demonstrate the value of their innovations for regulators, payers, clinicians, and other gatekeepers, some are reorganizing and better integrating their technology R&D, validation, and commercialization functions accordingly.

M. Managing Individual Bias and Conflict of Interest

HTA should seek to ensure that the credibility of its reports is not compromised by any significant biases or conflicts of interest. Bias and conflict of interest are distinct yet related concepts.

As described in chapter III, **bias** generally refers to any systematic deviation in an observation from the true nature of an event (e.g., a treatment effect in a clinical trial). Further, individual bias can refer to factors that might affect one's interpretation of evidence or formulation of findings and recommendations. This form of bias has been defined as "views stated or positions taken that are largely intellectually motivated or arise from close identification or association of an individual with a particular point of view or the positions or perspectives of a particular group" (National Academies 2003). This may include positions taken in public statements, publications, or other media; institutional or professional affiliations; recognition for personal achievement; intellectual passion; political or ideological beliefs; or personal relationships (Knickrehm 2009). As long as such positions have some recognized scientific or policy-related merit, they need not disqualify a person from participating in research or participating in HTA. Indeed, to provide for a competent expert review panel or set of advisors for an assessment, it may be useful to represent a balance of potentially biased perspectives.

Although such stated views or positions are a potential source of bias, they do not necessarily pose a conflict of interest. However, when an individual also has a significant, directly related interest or role, such as leading a professional society, industry association, or advocacy organization that has taken the same fixed position, this may pose a conflict of interest (National Academies 2003). Conflict of interest guidelines often address matters of individual bias as well.

Conflict of interest (or competing interest) refers to "any financial or other interest which conflicts with the service of the individual [person] because it (1) could significantly impair the individual's objectivity or (2) could create an unfair competitive advantage for any person or organization" (National Academies 2003). Conflict of interest policies typically apply to current, rather than past or expired interests, or possible interests that may arise in the future. In HTA, a conflict of interest could cause an individual to be biased in interpreting evidence or formulating findings and recommendations. In most instances, the existence of a significant conflict of interest pertaining to an HTA topic should disqualify an individual from participating in that HTA as a staff person, expert panel member, or consultant. However, persons with conflicts of interest may provide information to an HTA process, including relevant evidence, background information, other perspectives, or comments on draft reports.

Biases and conflicts of interest are conditions, not behaviors (Smith 2006; Thompson 1993). That is, an individual does not have to act on a bias or conflict of interest for it to exist. The existence of a bias or conflict of interest is reason for an HTA process to address them, e.g., whether to seek a balance of reasonable biases on an expert panel or to disqualify individuals with significant conflicts of interest from participating in an HTA, as appropriate.

HTA should consider the potential for conflict of interest on at least three levels:

- Sponsors of clinical trials and other studies that are part of the body of evidence under review
- Investigators who conducted and reported on the clinical trials and other studies that are part of the body of evidence under review
- Health technology assessors, including staff members , expert panel members, or other experts involved in reviewing the evidence and making findings and recommendations

1. Sponsors

Health technology companies and other sponsors of primary research typically determine or influence what research is conducted as well as such aspects as designation of the intervention and control treatments, endpoints, and follow-up periods, and whether research results are submitted for publication.

Financial conflicts of interest are common in clinical trials and other biomedical research. Industry sponsorship of research has been found to be associated with restrictions on publication and data sharing (Bekelman 2003). Clinical trials and cost-effectiveness analyses that are sponsored by industry yield positive (favorable) results more often than studies that are funded or conducted by others (Barbieri 2001; Chopra 2003; Friedberg 1999; Jang 2010). Among the potential reasons for this discrepancy are that industry is more likely to withhold reports of studies with negative results (e.g., that do not demonstrate a treatment effect). Another is that industry is more likely to sponsor studies (including RCTs) designed to increase the likelihood of positive results, i.e., where there is an expectation that one intervention (e.g., a new drug or diagnostic test) is superior to the alternative intervention (Polyzos 2011). In the case of RCTs, this latter tendency could undermine the principle of **equipoise** for enrolling patients in an RCT, although some contend that this principle can be counterproductive to progress in clinical research (Djulbegovic 2009; Fries 2004; Veatch 2007).

An analysis of clinical trials listed in ClinicalTrials.gov database found that health technology companies sponsor trials that are largely focused on their own products, while head-to-head comparisons with different active interventions from other companies are rare. This diminishes the evidence base for

assessing the relative benefits and harms of technologies for the same diseases (Lathyris 2010) and is one of the main reasons for the increased interest in comparative effectiveness research.

ClinicalTrials.gov helps protect against publication bias. Effective July 2005, the International Committee of Medical Journal Editors established a requirement that, as a condition of consideration for publication, all clinical trials be entered in a public registry (not necessarily ClinicalTrials.gov) that meets specified criteria before the onset of patient enrollment. As such, a sponsor cannot wait to see the final results of a trial before deciding to submit a manuscript about it to participating journals (International Committee of Medical Journal Editors 2013).

2. Investigators

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. Investigator conflict of interest is reported to be prevalent among clinical trials in various fields of pharmaceutical therapy and to be associated with a greater likelihood of reporting a drug to be superior to placebo (Perlis 2005). A systematic review of research on financial conflicts of interest among biomedical researchers found that approximately one-fourth of investigators had industry affiliations, and two-thirds of academic institutions held equity in start-up companies that sponsored research performed at the same institutions (Bekelman 2003).

Investigators with conflicts are more likely to report positive findings. This may arise from such factors as preferential funding of research that is likely to report positive findings, biased study designs, investigators' biased interpretation of results, or suppression of negative results (Okike 2008). As this research often appears in influential, "high-impact" journals, editors have adopted more systematic requirements for disclosure by investigators of their financial interests and the funding sources of studies, and applying greater scrutiny when potential conflicts arise (International Committee of Medical Journal Writers 1993; Kassirer 1993; Lo 2000; Jagsi 2009). Such requirements also have been applied to economic analyses (Kassirer 1994), although accompanied by controversy regarding whether certain sponsors (e.g., for-profit vs. not-for-profit) or methods (e.g., pharmacoeconomic modeling) are more acceptable than others (Schulman 1995; Steinberg 1995).

3. Health Technology Assessors

When interpreting the available evidence, health technology assessors should consider the existence of potential conflicts of interest that may have affected the conduct of a study or presentation of results. In addition, those participating in HTA should be subject to provisions that protect against their own potential conflicts of interest.

When interpreting evidence, HTA programs should consider information about sponsorship of a study, investigators, or other factors that suggest the potential for conflict of interest. Studies that are subject to potential conflicts of interest may have to be given less weight or dropped from the body of evidence under consideration.

For purposes of those conducting or otherwise involved in HTA, INAHTA defines conflict of interest as:

> *A situation in which the private interests of someone involved in the assessment or evaluation process (e.g. interviewer, rater, scorer, evaluator) have an impact (either positive or negative)*

on the quality of the evaluation activities, the accuracy of the data, or the results of the evaluation (INAHTA 2006).

Financial conflicts may include holding stock in, serving as a consultant to, or receiving honoraria from health technology companies or other organizations (e.g., medical professional groups) with financial interests in particular medical procedures or other technologies. Conflicts may be personal, i.e., apply to individuals associated with the HTA program and their immediate family members. Conflicts may also be non-personal, e.g., financial benefits to one's organization (e.g., university) or an industry-endowed fellowship held by an individual. Conflicts may be specific to a given technology or non-specific, such as a different technology made by the same company that makes the one being assessed.

HTA programs should take active measures, including adoption and implementation of formal guidelines or requirements, to protect against potential conflicts of interest among their managers, analysts, and expert panel members (Fye 2003; Phillips 1994). Similar measures should apply, as appropriate, to HTA program consultants, contractors, and outside reviewers of draft HTA reports. For example, as part of its extensive conflict of interest policy, ECRI Institute, a US-based independent nonprofit organization that conducts HTA, examines each employee's federal income tax return forms after they are filed to ensure that its employees do not own stock shares in medical device or pharmaceutical firms (ECRI Institute 2014). In addition to minimizing potential conflicts of interest, HTA programs should take active measures to minimize or balance bias among assessment teams and panel members.

HTA programs may have guidelines regarding when certain types of conflict affecting an individual require withdrawal (recusal) from the assessment process and when disclosure of the conflict is sufficient and participation is still permitted. This can involve various aspects or stages of HTA, including priority setting of HTA topics, selecting literature and data sources (including confidential versus open access data) for assessment, and preparing the assessment report. The INAHTA Checklist for HTA Reports includes a question regarding whether an HTA report provides a statement regarding conflict of interest on the part of those who prepared an HTA report or if funding for the HTA was provided by sources other than those responsible for the HTA agency's usual budget (INAHTA 2007).

References for Chapter X

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.

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(18):2325-33.

Balint M, et al. Treatment or Diagnosis: A Study of Repeat Prescriptions in General Practice. Philadelphia, PA: JB Lippincott; 1970.

Barbieri M, Drummond MF. Conflict of interest in industry-sponsored economic evaluations: real or imagined? Curr Oncol Rep. 2001;3(5):410-3.

Basch E, Prestrud AA, Hesketh PJ, Kris MG, et al.; American Society of Clinical Oncology. Antiemetics: American Society of Clinical Oncology clinical practice guideline update. J Clin Oncol. 2011;29(31):4189-98. http://jco.ascopubs.org/content/29/31/4189.long.

Bastian H, Scheibler F, Knelangen M, Zschorlich B, et al. Choosing health technology assessment and systematic review topics: the development of priority-setting criteria for patients' and consumers' interests. Int J Technol Assess Health Care. 2011;27(4):348-56.

Beebe DB, 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(4):454-65.

Berger A. High dose chemotherapy offers little benefit in breast cancer. BMJ. 1999 May 29;318(7196):1440. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1115834.

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;72(2):649-57.

Berwick DM. What 'patient-centered' should mean: confessions of an extremist. Health Aff (Millwood). 2009;28(4):w555-65. http://content.healthaffairs.org/content/28/4/w555.long.

Bower JL, Christensen CM. Disruptive technologies: catching the wave. Harv Bus Rev. 1995;73:43-53.

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;50(2):95-100.

Brouwers MC, Kho ME, Browman GP, Burgers JS, et al.; AGREE Next Steps Consortium. AGREE II: advancing guideline development, reporting and evaluation in health care. J Clin Epidemiol. 2010;63(12):1308-11.

Carlson JJ, Sullivan SD, Garrison LP, Neumann PJ, Veenstra DL. Linking payment to health outcomes: a taxonomy and examination of performance-based reimbursement schemes between healthcare payers and manufacturers. Health Policy. 2010;96(3):179-90.

Centers for Disease Control and Prevention (CDC). National and state vaccination coverage among adolescents aged 13 through 17 years — United States, 2012. MMWR Morb Mortal Wkly Rep. 2013;62(34):685-93. http://www.cdc.gov/mmwr/preview/mmwrhtml/mm6234a1.htm.

Centers for Disease Control and Prevention (CDC). National, state, and local area vaccination coverage among children aged 19-35 months — United States, 2012. MMWR Morb Mortal Wkly Rep. 2013 Sep 13;62(36):733-40. http://www.cdc.gov/mmwr/preview/mmwrhtml/mm6236a1.htm.

Centers for Disease Control and Prevention (CDC). Vital signs: awareness and treatment of uncontrolled hypertension among adults – United States, 2003-2010. MMWR Morb Mortal Wkly Rep. 2012;61(35):703-9.

http://www.cdc.gov/mmwr/preview/mmwrhtml/mm6135a3.htm/ Based on data from the US National Health and Nutrition Examination Survey (NHANES) 2003-2010. http://www.cdc.gov/nchs/nhanes.htm.

Chopra SS. Industry funding of clinical trials: Benefit or bias? JAMA. 2003;290(1):113.

Cochrane Consumer Network. About the Cochrane Consumer Network (CCNet). Accessed Jan. 2, 2014 at: http://consumers.cochrane.org/healthcare-users-cochrane.

Cooper JD. Paying the piper: the NETT strikes a sour note. National Emphysema Treatment Trial. Ann Thorac Surg. 2001;Aug;72(2):330-3.

Danner M, Hummel JM, Volz F, van Manen JG, et al. Integrating patients' views into health technology assessment: Analytic hierarchy process (AHP) as a method to elicit patient preferences. Int J Technol Assess Health Care. 2011;27(4):369-75.

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.

Domecq JP, Prutsky G, Elraiyah T, Wang Z, et al. Patient engagement in research: a systematic review. BMC Health Serv Res. 2014;14:89. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3938901.

Djulbegovic B. The paradox of equipoise: the principle that drives and limits therapeutic discoveries in clinical research. Cancer Control. 2009;16(4):342-7. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2782889.

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

Ebell MH, Siwek J, Weiss BD, Woolf SH, et al. Strength of recommendation taxonomy (SORT): a patient-centered approach to grading evidence in the medical literature. J Am Board Fam Pract. 2004;17(1):59-67. http://www.jabfm.org/content/17/1/59.long.

ECRI Health Technology Assessment Information Service. High-dose chemotherapy with autologous bone marrow transplantation and/or blood cell transplantation for the treatment of metastatic breast cancer. Plymouth Meeting, PA : ECRI, 1995.

ECRI Institute. Policies and Mission Statement. Accessed February 1, 2014 at: https://www.ecri.org/About/Pages/institutepolicies.aspx.

Eichler HG, Bloechl-Daum B, Abadie E, Barnett D, et al. Relative efficacy of drugs: an emerging issue between regulatory agencies and third-party payers. Nat Rev Drug Discov. 2010;9(4):277-91.

Ell K, Katon W, Xie B, Lee PJ, et al. One-year postcollaborative depression care trial outcomes among predominantly Hispanic diabetes safety net patients. Gen Hosp Psychiatry. 2011;33(5):436-42. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3175272.

Epstein RM, Street RL Jr. The values and value of patient-centered care. Ann Fam Med 2011;9(2):100-3. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3056855.

EUnetHTA Joint Action WP5 – Relative Effectiveness Assessment (REA) of Pharmaceuticals – Model for Rapid Relative Effectiveness Assessment of Pharmaceuticals, 1 March 2013 – V3.0. Accessed December 11, 2013 at: http://www.eunethta.eu/sites/5026.fedimbo.belgium.be/files/Model%20for%20Rapid%20REA%20of%20pharmaceuticals_final_20130311_reduced.pdf.

European Patients' Forum. Patient Involvement in Health Technology Assessment. (Undated) Accessed Dec. 1, 2013 at: http://www.eu-patient.eu/Documents/Projects/HTA/EPF-report_HTA-survey_HTA-agencies.pdf.

Facey K, Boivin A, Gracia J, Hansen HP, et al. Patients' perspectives in health technology assessment: a route to robust evidence and fair deliberation. Int J Technol Assess Health Care. 2010;26(3):334-40.

Federal Coordinating Council for Comparative Effectiveness Research. Report to the President and the Congress. Washington, DC: US Department of Health and Human Services, June 2009.

http://www.tuftsctsi.org/~/media/Files/CTSI/Library%20Files/FCC%20for%20CER%20Rpt%20to%20Pres%20and%20Congress_063009.ashx.

Ferguson TB Jr, Peterson ED, Coombs LP, Eiken MC, 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(1):49-56.

Fineberg HV. Keynote Address. Health Technology Assessment International 2009 Annual Meeting, Singapore; June 26, 2009.

Fletcher SW. Whither scientific deliberation in health policy recommendations? N Engl J Med. 1997;336(16):1180-3.

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(15):1453-7.

Fries JF, Krishnan E. Equipoise, design bias, and randomized controlled trials: the elusive ethics of new drug development. Arthritis Res Ther. 2004;6(3):R250-5. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC416446.

Fuchs VR. The doctor's dilemma--what is "appropriate" care? N Engl J Med. 2011;365(7):585-7.

Fye WB. The power of clinical trials and guidelines, and the challenge of conflicts of interest. J Am Coll Cardiol. 2003;41(8):1237-42.

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.

GAO (General Accounting Office). Medicare: Beneficiary use of clinical preventive services. Report to the Chairman, Subcommittee on Oversight and Investigations, Committee on Energy and Commerce, House of Representatives. GAO-02-422. Washington, DC; 2002. http://www.gao.gov/assets/240/234232.html.

Garrison LP Jr, Bresnahan BW, Higashi MK, et al. Innovation in diagnostic imaging services: assessing the potential for value-based reimbursement. Acad Radiol. 2011;18(9):1109-14.

Glaeske G. The dilemma between efficacy as defined by regulatory bodies and effectiveness in clinical practice. Dtsch Arztebl Int. 2012;109(7):115-6. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3301971.

Goodman CS. Healthcare technology assessment: methods, framework, and role in policy making. Am J Manag Care. 1998;4:SP200-14. http://www.ajmc.com/publications/issue/1998/1998-09-vol4-n2SP/Sep98-1101pSP200-SP21.

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.

Green C. Considering the value associated with innovation in health technology appraisal decisions (deliberations): a NICE thing to do? Appl Health Econ Health Policy. 2010;8(1):1-5.

Hailey D. A preliminary survey on the influence of rapid health technology assessments. Int J Technol Assess Health Care. 2009;25(3):415-8.

Harker J, Kleijnen J. What is a rapid review? A methodological exploration of rapid reviews in Health Technology Assessments. Int J Evid Based Healthc. 2012;10(4):397-410.

Health Equality Europe. Understanding Health Technology Assessment (HTA). July 2008. Accessed Jan. 2, 2014 at: http://www.htai.org/fileadmin/HTAi_Files/ISG/PatientInvolvement/EffectiveInvolvement/HEEGuideToHTAforPatientsEnglish.pdf.

Heidenreich PA, Trogdon JG, Khavjou OA, et al. Forecasting the future of cardiovascular disease in the United States: a policy statement from the American Heart Association. Circulation. 2011;123(8):933–44. http://circ.ahajournals.org/content/123/8/933.long.

Higgins JPT, Green S, eds. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Accessed Sept. 1, 2013 at: http://handbook.cochrane.org.

Hinman AR, Orenstein WA, Santoli JM, Rodewald LE, Cochi SL. Vaccine shortages: history, impact, and prospects for the future. Annu Rev Public Health. 2006;27:235-59.

Hoffman B. Is there a technological imperative in health care? Int J Technol Assess Health Care. 2002;18(3):675-89.

HTAi Patient and Citizen Involvement Interest Sub-Group. Good Practice Examples of PPI. 2012. Accessed Jan. 2, 2014 at: http://www.htai.org/fileadmin/HTAi_Files/ISG/PatientInvolvement/Materials/Good_Practice_Examples.doc.

HTAi Patient and Citizen Involvement Interest Sub-Group. PIE Good Practice Principles for Patient Involvement in Health Technology Assessment—Draft. August 2012. Accessed Jan. 2, 2014 at: http://www.htai.org/fileadmin/HTAi_Files/ISG/PatientInvolvement/Materials/PIE_principles_2012_august.pdf.

HTAi Patient and Citizen Involvement Interest Sub-Group. Good Practice Examples of Patient and Public Involvement in Health Technology Assessment. Sept. 2013. Accessed Jan. 2, 2014 at: http://www.htai.org/fileadmin/HTAi_Files/ISG/PatientInvolvement/GeneralSIGdocuments/Good_Practice_Examples_September_2013.pdf.

Hu Q, Schwarz LB, Uhan NA. The impact of group purchasing organizations on healthcare-product supply chains. MSOM. 2012;14(1):7-23.

Hudon C, Fortin M, Haggerty JL, et al. Measuring patients' perceptions of patient-centered care: a systematic review of tools for family medicine. Ann Fam Med. 2011;9(2):155-64. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3056864.

Hutton J, Trueman P, Henshall C. Coverage with evidence development: an examination of conceptual and policy issues. Int J Technol Assess Health Care. 2007;23(4):425-32.

International Committee of Medical Journal Editors. Uniform Requirements for Manuscripts Submitted to Biomedical Journals: Publishing and Editorial Issues Related to Publication in Biomedical Journals: Obligation to Register Clinical Trials. 2013. Accessed December 11, 2013 at: http://www.icmje.org/publishing_10register.html.

INAHTA. International Network of Agencies in Health Technology Assessment Secretariat. A Checklist for Health Technology Assessment Reports. Version 3.2. August 2007. Accessed December 11, 2013 at: http://www.inahta.org/HTA/Checklist.

INAHTA. International Network of Agencies for Health Technology Assessment. Health Technology Assessment (HTA) Glossary. First Edition. INAHTA Secretariat, c/o SBU, Stockholm, July 5, 2006. Accessed December 11, 2013 at: http://www.inahta.org/upload/HTA_resources/Edu_INAHTA_glossary_July_2006_final.pdf.

Institute of Medicine. Committee on Comparative Effectiveness Prioritization. Initial National Priorities for Comparative Effectiveness Research. Washington, DC: National Academies Press; 2009. http://books.nap.edu/openbook.php?record_id=12648.

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. http://books.nap.edu/openbook.php?record_id=10027

International Committee of Medical Journal Writers. Conflict of interest. Lancet. 1993;341(8847):742-3.

Jacobs BL, Zhang Y, Schroeck FR, et al. Use of advanced treatment technologies among men at low risk of dying from prostate cancer. JAMA. 2013;309(24):2587-95.

Jagsi R, Sheets N, Jankovic A, Motomura AR, Amarnath S, Ubel PA. Frequency, nature, effects, and correlates of conflicts of interest in published clinical cancer research. Cancer. 2009;115(12):2783-91. http://onlinelibrary.wiley.com/doi/10.1002/cncr.24315/pdf.

Jang S, Chae YK, Haddad T, Majhail NS. Conflict of interest in economic analyses of aromatase inhibitors in breast cancer: a systematic review. Breast Cancer Res Treat. 2010;121(2):273-9.

Kaden RJ, Vaul JH, Palazola PA. Negotiating payment for new technology purchases. Healthc Financ Manage. 2002;56(12):44-8.

Kassirer JP, Angell M. Financial conflicts of interest in biomedical research. N Engl J Med. 1993;329(8):570-1. http://www.nejm.org/doi/full/10.1056/NEJM199308193290810.

Kassirer JP, Angell M. The journal's policy on cost-effectiveness analyses. N Engl J Med. 1994;331(10):669-70. http://www.nejm.org/doi/full/10.1056/NEJM199409083311009.

Kennedy I. Appraising the Value of Innovation and Other Benefits. A Short Study for NICE. July 2009. Accessed December 11, 2013 at: http://www.nice.org.uk/media/98F/5C/KennedyStudyFinalReport.pdf.

Khangura S, Konnyu K, Cushman R, et al. Evidence summaries: the evolution of a rapid review approach. Syst Rev. 2012;1:10. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3351736.

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.

Klemp M, Frønsdal KB, Facey K; HTAi Policy Forum. What principles should govern the use of managed entry agreements? Int J Technol Assess Health Care 2011;27(1):77-83.

Knickrehm S. Non-Financial Conflicts of Interest. Slide Presentation from the AHRQ 2009 Annual Conference. December 2009. Agency for Healthcare Research and Quality, Rockville, MD. Accessed December 11, 2013 at: http://www.ahrq.gov/about/annualconf09/knickrehm.htm.

Kravitz RL, Duan N, Braslow J. Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. Milbank Q 2004;82(4):661-87. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2690188.

Kreis J, Schmidt H. Public engagement in health technology assessment and coverage decisions: a study of experiences in France, Germany, and the United Kingdom. J Health Polit Policy Law. 2013;38(1):89-122.

Kwan J, Sandercock P. In-hospital care pathways for stroke: a Cochrane systematic review. Stroke. 2003;34(2):587-8. http://stroke.ahajournals.org/content/34/2/587.long.

Lathyris DN, Patsopoulos NA, Salanti G, Ioannidis JP. Industry sponsorship and selection of comparators in randomized clinical trials. Eur J Clin Invest. 2010;40(2):172-82.

Lee A, Skött LS, Hansen HP. Organizational and patient-related assessments in HTAs: state of the art. Int J Technol Assess Health Care. 2009;25(4):530-6.

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. http://www.lewin.com/~/media/Lewin/Site_Sections/Publications/1602.pdf.

The Lewin Group. Outlook for Medical Technology Innovation. Report 2: The Medicare Payment Process and Patient Access to Technology. Washington, DC: AdvaMed; 2000.

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. http://www.nap.edu/openbook.php?record_id=1547&page=1.

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.

Mangione-Smith R, DeCristofaro AH, Setodji CM, Keesey J, et al. The quality of ambulatory care delivered to children in the United States. N Engl J Med. 2007;357(15):1515-23.

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. J Natl Cancer Inst . 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;348(26):2635-45.

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

Mead N, Bower P. Patient-centredness: a conceptual framework and review of the empirical literature. Soc Sci Med 2000;51(7):1087-110.

Medical Technology Leadership Forum. MTLF Summit: Conditional Coverage of Investigational Technologies. Prepared by The Lewin Group. Washington, DC; October 1999.

Mello MM, Brennan TA. The controversy over high-dose chemotherapy with autologous bone marrow transplant for breast cancer. Health Aff (Millwood). 2001;20(5):101-17. http://content.healthaffairs.org/content/20/5/101.long.

Methodology Committee of the Patient-Centered Outcomes Research Institute. Methodological standards and patient-centeredness in comparative effectiveness research. The PCORI perspective. JAMA. 2012;307(15):1636-40.

Methods Guide for Effectiveness and Comparative Effectiveness Reviews. AHRQ Publication No. 10(14)-EHC063-EF. Rockville, MD: Agency for Healthcare Research and Quality. January 2014. Accessed Feb. 1, 2014 at: http://effectivehealthcare.ahrq.gov/ehc/products/60/318/CER-Methods-Guide-140109.pdf.

Miller D, Rudick RA, Hutchinson M. Patient-centered outcomes: translating clinical efficacy into benefits on health-related quality of life. Neurology. 2010 Apr 27;74 Suppl 3:S24-35.

Miller FG, Pearson SD. Coverage with evidence development: ethical issues and policy implications. Med Care 2008;46(7):746-51.

Nakamura C, Bromberg M, Bhargava S, Wicks P, Zeng-Treitler Q. Mining online social network data for biomedical research: a comparison of clinicians' and patients' perceptions about amyotrophic lateral sclerosis treatments. J Med Internet Res. 2012;14(3):e90. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3414854.

National Academies. Policy on Committee Composition and Balance and Conflicts of Interest for Committees Used in the Development of Reports. May 12, 2003. Accessed December 13, 2013 at: http://www.nationalacademies.org/coi/bi-coi_form-0.pdf.

National Institute for Health and Care Excellence (NICE). Citizens Council. Accessed Jan. 2, 2014 at: http://www.nice.org.uk/aboutnice/howwework/citizenscouncil/citizens_council.jsp.

National Institute for Health and Care Excellence (NICE). Guide to the Methods of Technology Appraisal. Accessed Jan. 2, 2014 at: http://publications.nice.org.uk/guide-to-the-methods-of-technology-appraisal-2013-pmg9/involvement-and-participation#patient-and-carer-groups.

Neumann PJ. What we talk about when we talk about health care costs. N Engl J Med. 2012;366(7):585-6. http://www.nejm.org/doi/full/10.1056/NEJMp1200390.

Okike K, Kocher MS, Mehlman CT, Bhandari M. Industry-sponsored research. Injury. 2008;39(6):666-80.

Oxman AD, Guyatt G. A consumer's guide to subgroup analyses. Ann Intern Med 1992;116(1):76-84.

Patient-Centered Outcomes Research Institute. Patient-centered outcomes research. 2013. Accessed December 13, 2013 at: http://pcori.org/research-we-support/pcor.

Pearson SD, Bach PB. How Medicare could use comparative effectiveness research in deciding on new coverage and reimbursement. Health Aff (Millwood). 2010;29(10):1796-804. http://content.healthaffairs.org/content/29/10/1796.long.

Perlis RH, Perlis CS, Wu Y, et al. Industry sponsorship and financial conflict of interest in the reporting of clinical trials in psychiatry. Am J Psychiatry. 2005;162(10):1957-60.

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. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2566501.

Polyzos NP, Valachis A, Mauri D, Ioannidis JP. Industry involvement and baseline assumptions of cost-effectiveness analyses: diagnostic accuracy of the Papanicolaou test. CMAJ. 2011;183(6):E337-43. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3071415.

Porter ME. What is value in health care? N Engl J Med. 2010;363(26):2477-81. http://www.nejm.org/doi/full/10.1056/NEJMp1011024.

Reda AA, Kotz D, Evers SM, van Schayck CP. Healthcare financing systems for increasing the use of tobacco dependence treatment. Cochrane Database Syst Rev. 2012 Jun 13;(6):CD004305.

Rettig RA. Health Care in Transition: Technology Assessment in the Private Sector. Santa Monica, Ca: RAND; 1997. http://www.rand.org/content/dam/rand/pubs/monograph_reports/2007/MR754.pdf.

Rothrock NE, Hays RD, Spritzer K, Yount SE, et al. Relative to the general US population, chronic diseases are associated with poorer health-related quality of life as measured by the Patient-Reported Outcomes Measurement Information System (PROMIS). J Clin Epidemiol. 2010;63(11):1195-204. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2943571.

Schauffler HH, Barker DC, Orleans CT. Medicaid coverage for tobacco-dependence treatments. Health Aff (Millwood). 2001;20(1):298-303. http://content.healthaffairs.org/content/20/1/298.long.

Schulman K. Cost-effectiveness analyses. N Engl J Med. 1995;332(2):124.

Sharf BF. Out of the closet and into the legislature: breast cancer stories. Health Aff (Millwood). 2001;20(1):213-8. http://content.healthaffairs.org/content/20/1/213.long.

Sheingold, SH. Technology assessment, coverage decisions, and conflict: the role of guidelines. Am J Manag Care. 1998;4:SP117-25. http://www.ajmc.com/publications/issue/1998/1998-09-vol4-n2SP/Sep98-1094pSP117-SP12.

Smith R. Conflicts of interest: how money clouds objectivity. J R Soc Med. 2006;99(6):292–7. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1472724.

Stead LF, Perera R, Bullen C, Mant D, et al. Nicotine replacement therapy for smoking cessation. Cochrane Database Syst Rev. 2012 Nov 14;11:CD000146.

Steinberg EP. Cost-effectiveness analyses. N Engl J Med. 1995;332(2):123. http://www.nejm.org/doi/full/10.1056/NEJM199501123320212.

Steinberg EP, Tunis S, Shapiro D. Insurance coverage for experimental technologies. Health Aff (Millwood). 1995;14(4):143-58. http://content.healthaffairs.org/content/14/4/143.long.

Stewart M, et al. Patient-Centered Medicine: Transforming the Clinical Method. 3rd ed. United Kingdom: Radcliffe Health; 2013.

Street RL Jr, Elwyn G, Epstein RM. Patient preferences and healthcare outcomes: an ecological perspective. Expert Rev Pharmacoecon Outcomes Res. 2012;12(2):167-80.

Thompson DF. Understanding financial conflicts of interest. N Engl J Med. 1993;329(8):573-6.

Trueman P, Grainger DL, Downs KE. Coverage with evidence development: applications and issues. Int J Technol Assess Health Care 2010;26(1):79-85.

UK National Institute for Health and Care Excellence (NICE). Citizens Council. Accessed Jan. 2, 2014 at: http://www.nice.org.uk/aboutnice/howwework/citizenscouncil/citizens_council.jsp.

Veatch RM. The irrelevance of equipoise. J Med Philos. 2007;32(2):167-83.

von Below GC, Boer A, Conde-Olasagasti JL, Dillon A,, et al. Health technology assessment in policy and practice. Working Group 6 Report. Int J Technol Assess Health Care. 2002;18(2):447-55.

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.

Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine--reporting of subgroup analyses in clinical trials. N Engl J Med. 2007;357(21):2189-94. http://www.nejm.org/doi/full/10.1056/NEJMsr077003.

Watt A, Cameron A, Sturm L, Lathlean T, et al. Rapid reviews versus full systematic reviews: an inventory of current methods and practice in health technology assessment. Int J Technol Assess Health Care. 2008;24(2):133-9.

Whitty JA. An international survey of the public engagement practices of health technology assessment organizations. Value Health. 2013;16(1):155-63.

Wilson JM. It's time for gene therapy to get disruptive! Hum Gene Ther. 2012;23(1):1-3.

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. http://www.journalslibrary.nihr.ac.uk/hta/volume-6/issue-16.

Wong MK, Mohamed AF, Hauber AB, Yang JC, et al. Patients rank toxicity against progression free survival in second-line treatment of advanced renal cell carcinoma. J Med Econ. 2012;15(6):1139-48.

Zandbelt LC, Smets EM, Oort FJ, et al. Medical specialists' patient-centered communication and patient-reported outcomes. Med Care. 2007;45(4):330-9.

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 on an unbiased method that provides observations having great variation or random error (i.e., not close in magnitude to each other). (Contrast with **precision.**)

Adaptive licensing: (or progressive licensing) refers to proposals for prospectively planned, stepwise, and adaptive approaches to market approval for drugs, biologics, and other regulated technologies. Recognizing that the evidence available at the time of conventional market approval for many technologies is often insufficient for well-informed clinical decisions, adaptive licensing is intended to improve tradeoffs of timely patient access with accumulating evidence on safety and effectiveness. These approaches involve iterative steps of data collection, regulatory review, and modification of licensing (or labeling). For example, this would enable earlier approval (at phase II or even phase I) of a drug for narrowly defined indications while RCTs or other trials continue to generate confirmatory data, data for broader indications, or data in real-world settings that could be used to modify licensing.

Adherence: (or compliance or concordance) a measure of the extent to which patients undergo, continue to follow, or persist with a treatment or regimen as prescribed, e.g., taking drugs, undergoing a medical or surgical procedure, doing an exercise regimen, or abstaining from smoking.

Allocation concealment: refers to the process of ensuring that the persons assessing patients for potential entry into a trial, as well as the patients themselves, do not know whether any particular patient will be allocated to an intervention group or control group. This diminishes **selection bias** by preventing the persons who are managing patient allocation, or the patients, from influencing (intentionally or not) patient assignment to one group or another. Centralized randomization (i.e., managed at one site rather than at each enrollment site) using certain techniques is a preferred method ensuring allocation concealment. (This is distinct from **blinding** of patients, providers, and others.)

Adaptive clinical trial: a form of trial that uses accumulating data to determine how to modify the design of ongoing trials according to a pre-specified plan. Intended to increase the quality, speed, and efficiency of trials, adaptive trials typically involve interim analyses, changes to sample size, changes in randomization to treatment arms and control groups, and changes in dosage or regimen of a drug or other technology.

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.

Attrition bias: refers to systematic differences between comparison groups in withdrawals (drop-outs) from a study, loss to follow-up, or other exclusions of patients and how these losses are analyzed.

Ignoring these losses or accounting for them differently between groups can skew study findings, as patients who withdraw or are lost to follow-up may differ systematically from those patients who remain for the duration of the study. Patients' awareness of whether they have been assigned to a particular treatment or control group may differentially affect their likelihood of dropping out of a trial. Techniques for diminishing attrition bias include blinding of patients as to treatment assignment, completeness of follow-up data for all patients, and intention-to-treat analysis (with imputations for missing data as appropriate).

Bayesian clinical trial: a type of **adaptive clinical trial** that allows for assessment of results during the course of the trial and modifying the trial design to arrive at results more efficiently. Such modifications during trials may include, e.g., changing the ratio of randomization to treatment arms to favor what appear to be more effective therapies, adding or eliminating treatment arms, changing enrollee characteristics to focus on patient subgroups that appear to be better responders, and slowing or stopping enrollment as certainty increases about treatment effects. Bayesian clinical trials are based on **Bayesian statistics**.

Bayesian statistics: a branch of statistics that involves learning from evidence as it accumulates. It is based on Bayes' Rule (or Bayes' Theorem), a mathematical equation that expresses how the prior (initial) probability of an event (or the probability that a hypothesis is true or the distribution for an unknown quantity) changes to a posterior (updated) probability given relevant new evidence. For example, in the diagnosis of a condition in an individual patient, a *prior probability* of a diagnosis may be based on the known prevalence of that condition in a relevant population. This can be updated to a *posterior probability* based on whether the result of a diagnostic test for that condition in that patient is positive or negative.

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.

Best evidence: refers, in general, to evidence that best attains certain quality criteria for internal and external validity. This also refers to a principle that a desire to base health care decisions and policies only on evidence generated from the best study designs for establishing internal and external validity should not preclude using the best evidence that is available from other study designs. That is, the "best evidence" may be the *best available* evidence that is relevant for the evidence questions of interest. This does not necessarily mean that the best available evidence should be designated as being of high quality.

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, a systematic (i.e., not due to random error) deviation in an observation from the true nature of an event. In clinical trials, bias may arise from any factor other than the intervention of interest that systematically distorts the magnitude of an observed treatment effect from the true effect. Bias diminishes the accuracy (though not necessarily the precision) of an observation. Biases may arise from inadequacies in the design, conduct, analysis, or reporting of a study. Among the main forms of bias are **selection bias, performance bias, detection bias, attrition bias, reporting bias**, and **publication bias**. **Confounding** of treatment effects can arise from various sources of bias.

Bias also refers to factors that may affect an individual's interpretation of evidence or formulation of findings and recommendations, i.e., "views stated or positions taken that are largely intellectually motivated or arise from close identification or association of an individual with a particular point of view or the positions or perspectives of a particular group" (National Academies 2003). This may include positions taken in public statements, publications, or other media; institutional or professional affiliations; recognition for personal achievement; intellectual passion; political or ideological beliefs; or personal relationships.

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.*

Biomarker: (or biological marker) an objectively measured variable or trait that is used as an indicator of a normal biological process, a disease state, or effect of a treatment. It may be a physiological measurement (height, weight, blood pressure, etc.), blood component or other biochemical assay (red blood cell count, viral load, HbA1c level, etc.), genetic data (presence of a specific genetic mutation), or measurement from an image (coronary artery stenosis, cancer metastases, etc.). (See also **intermediate endpoint** and **surrogate endpoint**.)

Blinding: the process of preventing one or more of patients, clinicians, investigators, and data analysts from knowing whether individual patients are receiving the investigational intervention(s) or the control (or standard) intervention(s) in a clinical trial. (Also known as masking.) Blinding is intended to eliminate the possibility that knowledge of which intervention is being received will affect patient outcomes, investigator behaviors that may affect outcomes, or assessment of 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. The terms "single-blinded," "double-blinded," and "triple-blinded" refer to which parties are blinded, e.g., one or more of patients, investigators, and data analysts; however, these terms are used inconsistently and the specific parties who are blinded in a trial should be identified.

Budget impact analysis (BIA): determines the how implementing or adopting a particular technology or technology-related policy will affect a designated budget, e.g., of a drug formulary or health plan. A BIA typically does not account for the broad economic impact (e.g., societal impact) of implementing or adopting the technology. A BIA can be conducted simply to determine the impact of alternative technologies or programs on a budget, or it could be conducted to determine whether, or how much of, a technology or program (or combination of these) could be implemented subject to resource constraints, such as a fixed (or "capped") budget.

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 endpoint: an event or other outcome that can be measured objectively to determine whether an intervention achieved its desired impact on patients. Usual clinical endpoints are mortality (death), morbidity (disease progression), symptom relief, quality of life, and adverse events. These are often categorized as primary (of most importance) endpoints and secondary (additional though not of greatest interest) endpoints.

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 of patient care 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 registry trials are a type of multicenter trial design using existing online registries as an efficient platform to conduct patient assignment to treatment and control groups, maintain case records, and conduct follow-up. Such trials that randomize patient assignment to treatment and control groups are randomized clinical registry trials (see Fröbert 2010).

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.)

Cluster randomized trials: trials that randomize assignment of interventions at the level of natural groups or organizations rather than at the level of patients or other individuals. The clusters may be a set of clinics, hospitals, nursing homes, schools, communities, or geographic regions that are randomized to receive one or more interventions and comparators. Such designs are used when it is not feasible to randomize individuals or when an intervention is designed to be delivered at a group or

social level, such as a workplace-based smoking cessation campaign. These are also known as group, place, or community randomized trials

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.

Cluster randomized trials: involve randomization of interventions at the level of natural groups or organizations rather than at the level of patients or other individuals. That is, sets of clinics, hospitals, nursing homes, schools, communities, or geographic regions are randomized to receive interventions or comparators. Such designs are used when it is not feasible to randomize individuals or when an intervention is designed to be delivered at a group or social level, such as a workplace-based smoking cessation campaign or a health care financing mechanism. These are also known as group, place, or community randomized trials.

Comparative effectiveness research (CER): generation and synthesis of evidence comparing the benefits and harms of technologies, with the attributes of: direct ("head-to-head") comparisons, effectiveness in real-world health care settings, health care outcomes (as opposed to surrogate or other intermediate endpoints), and ability to identify different treatment effects in patient subgroups. CER can draw on a variety of complementary study designs and analytical methods.

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.

Concurrent validity: refers to how well a measure correlates with a previously validated one, and the ability of a measure to accurately differentiate between different groups at the time the measure is applied.

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.

Conflict of interest: (or competing interest) refers to "any financial or other interest which conflicts with the service of the Individual because it (1) could significantly impair the individual's objectivity or (2) could create an unfair competitive advantage for any person or organization" (National Academies 2003). Conflict of interest means something more than individual bias; it involves an interest, usually financial, that could directly affect the conduct of HTA.

Confounding: occurs when any factor that is associated with an intervention has an impact on an outcome that is independent of the impact of the intervention. In controlled clinical trials, confounding refers to circumstances in which the observed treatment effect of an intervention is biased due to a difference between the intervention and control groups, such as a difference in baseline risk factors at the start of a trial or different exposures during the trial that could affect outcomes.

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 processes may be informal or formal, involving such techniques as the nominal group and Delphi techniques.

Construct validity refers to how well a measure is correlated with other accepted measures of the construct (i.e., the concept or trait) of interest (e.g., pain, anxiety, mobility, quality of life), and discriminates between groups known to differ according to the variable.

Content validity refers to the degree to which the set of items in a data collection instrument is known to represent the range or universe of meanings or dimensions of a construct of interest, e.g., how well the domains of a health-related quality of life index for arthritis represent the aspects of quality of life or daily functioning that are important to patients with arthritis.

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 the MEDLINE /PubMED database and other bibliographic databases of the US National Library of Medicine.

Convergent validity, opposite discriminant validity, refers to the extent to which two different measures that are intended to measure the same construct do indeed yield similar results. Convergent validity contributes to, or can be considered a subtype of, construct validity.

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 a disease or health condition, including treatment costs; this form of study does not address benefits/outcomes.

Coverage with evidence development: refers to limited coverage (e.g., for particular clinical indications or practice settings) for a new technology in parallel with specified data collection to provide stronger evidence about the effectiveness, safety, or other impacts of the technology, or additional evidence pertaining to broader uses of the technology, of interest to payers. This enables some patient access to the technology while reducing uncertainty through new evidence generation and real-world experience to inform revised coverage policies as well as clinical practice decisions. CED includes coverage "only in research" (i.e., coverage for a technology only when used in members of the payer's patient population who are participating in clinical trials of the technology) and coverage "only with research" (i.e., coverage for a technology only when also being used contemporaneously in a sample of the payer's patient population participating in clinical trials of the technology). CED is a form of **managed entry**.

Criterion validity: how well a measure, including its various domains or dimensions, is correlated with a known gold standard or definitive measurement, if one exists.

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.

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).

Detection bias: (or ascertainment or observer bias) refers to bias arising from differential assessment of outcomes, whether by patients or investigators, influenced by their knowledge of the assignment of patients to intervention or control groups. **Blinding** of patients and investigators to treatment assignment is a technique used to manage detection bias. Prospective studies help to manage certain forms of detection bias that arise in retrospective studies.

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 and 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.

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. Discriminant validity contributes to, or can be considered a subtype of, construct validity.

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.

Disinvestment: refers to completely or partially withdrawing resources from currently used health technologies that are potentially harmful, ineffective, or cost-ineffective. It is a means of optimizing the use of health care resources. Disinvestment does not imply replacement by alternatives, though it may be accompanied by, or provide "head-room" or a niche for, a new technology or other replacement. Active disinvestment refers to purposely withdrawing resources from or otherwise discontinuing use of a technology. Implicit disinvestment refers to instances in which a technology falls from use or is superseded by another in the absence of an explicit decision to discontinue its use.

Disruptive innovation: an innovation that alters and may even displace existing systems, networks, or markets, and that may create new business models and lead to emergence of new markets. In health care, disruptive innovations challenge and may alter existing systems of regulation, payment, health care delivery, or professional training.

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

Drug compendium: a comprehensive listing or index of summary information about drugs and biologicals (or a subset of these, e.g., anticancer treatments), including their dosing, adverse effects, interactions, contraindications, and recommended indications, including those that are approved by regulatory agencies ("on-label") and those that are beyond regulatory agency approval yet may be "medically accepted" ("off-label") and other pharmacologic and pharmacokinetic information.

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.

Enrichment of trials: techniques of identifying patients for enrollment in clinical trials based on prospective use of patient attributes that are intended to increase the likelihood of detecting a treatment effect (if one truly exists) compared to an unselected population. Such techniques may be designed, e.g., to decrease patient heterogeneity of response, select for patients more likely to experience a disease-related trial endpoint, or select for patients (based on a known predictive

biomarker) more likely to respond to a treatment (intended to result in a larger effect size). In adaptive enrichment of clinical trials, investigators seek to discern predictive markers/attributes during the course of a trial and apply these to enrich subsequent patient enrollment in the trial.

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 an RCT only if there is genuine uncertainty (an expectation for equal likelihood) about which intervention will benefit and which will harm the patient most; and, across a large number of RCTs, the number of RCTs that reject and that fail to reject the null hypothesis will be approximately equal. The assumption of equipoise is the basis for testing the **null hypothesis** in RCTs.

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.

Exclusions after randomization bias: refers to bias arising from inappropriate accounting for patient dropouts, withdrawals, and deviations from trial protocols. Patients who leave a trial or whose data are not otherwise adequately collected as per the trial protocol may differ systematically from the remaining patients, introducing potential biases in observed treatment effects. **Intention-to-treat analysis** and worst-case scenario analysis are two techniques for managing bias due to exclusions after randomization.

External validity: the extent to which the results of a study conducted under particular circumstances can be generalized to other patients, populations, or other circumstances. To the extent that the circumstances of a particular study (e.g., patient characteristics or the manner of delivering a treatment) differ from the circumstances of interest, the external validity of the results of that study may be questioned. Also known as applicability.

Face validity is the ability of a measure to represent reasonably (that is, to be acceptable "on its face") a construct (i.e., a concept, trait, or domain of interest) as judged by someone with expertise in the construct.

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.

Funnel plot: in systematic reviews and meta-analyses, a graph (scatter plot) of the distribution of reported treatment effects of individual studies (along the horizontal axis) against the sample sizes of the studies (along the vertical axis). Because studies with larger sample sizes should generate more precise estimates of treatment effect, they are likely to be grouped more narrowly around an average along the horizontal axis; while the studies with smaller sample sizes are likely to be scattered more widely on both sides of the average along the horizontal axis. As such, in the absence of bias (e.g., publication bias), the scatter plot will be narrower at the top (large sample sizes, small variation) and wider at the bottom (small sample sizes, large variation), resembling an inverted funnel.

Genomics: the branch of molecular genetics that studies the genome, i.e., the complete set of DNA in the chromosomes of an organism. This may involve application of DNA sequencing, recombinant DNA, and related bioinformatics to sequence, assemble, and analyze the structure, function, and evolution of genomes. Whereas genetics is the study of the function and composition of individual genes, genomics addresses all genes and their interrelationships in order to understand their combined influence on the organism. (See also **pharmacogenetics** and **pharmacogenomics**.)

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.

Heterogeneity of treatment effects (HTEs): refers to variation in effectiveness, safety (adverse events), or other patient responses observed across a patient population with a particular health problem or

condition. This variation may be associated with such patient characteristics as genetic, sociodemographic, clinical, behavioral, environmental, and other personal traits, or personal preferences.

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.

Horizon scanning: refers to the ongoing tracking of multiple, diverse information sources (bibliographic databases, clinical trial registries, regulatory approvals, market research reports, etc.) to identify potential topics for HTA and provide input for setting priorities. While horizon scanning is most often used to identify new technologies that eventually may merit assessment, it can also involve identifying technologies that may be outmoded or superseded by newer ones. It can also be used to, e.g., identify areas of technological change; anticipate new indications of technologies; identify variations in, and potential inappropriate use of, technologies; and plan data collection to monitor adoption, diffusion, use, and impacts of technologies.

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. (Contrast with **prevalence**.)

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.

Integrative methods: (or secondary or synthesis methods) involve combining data or information from multiple existing primary studies such as clinical trials. These include a range of more or less systematic quantitative and qualitative methods, including systematic literature reviews, meta-analysis, decision analysis, consensus development, and unstructured literature reviews. (Contrast with **primary data methods**.)

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.

Intermediate endpoint: a non-ultimate endpoint (e.g., not mortality or morbidity) that may be associated with disease status or progression toward an ultimate endpoint such as mortality or morbidity. They may be certain biomarkers (e.g., HbA1c in prediabetes or diabetes, bone density in osteoporosis, tumor progression in cancer) or disease symptoms (e.g., angina frequency in heart disease, measures of lung function in chronic obstructive pulmonary disease). (See also **biomarker** and **surrogate endpoint**.)

Internal validity: the extent to which the results of a study accurately represent the causal relationship between an intervention and an outcome in the particular circumstances of that study. This includes the extent to which the design and conduct of a study minimize the risk of any systematic (non-random) error (i.e., **bias**) in the study results. True experiments such as RCTs generally have high internal validity.

Interventional study: a prospective study in which investigators assign or manage an intervention or other exposure of interest to patients (including RCTs, other experiments, and certain other study designs) and interpret the outcomes. In an interventional study, investigators manage assignment of patients to interventions (e.g., treatment and control groups), timing of interventions, selection of outcomes, and timing of data collection. (Contrast with **observational study**.)

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.)

Managed entry: refers to a range of innovative payment approaches that provide patient access under certain conditions. Three main purposes are to manage: uncertainty about safety, effectiveness, or cost effectiveness; budget impact; or technology use for optimizing performance. Two main types of managed entry are conditional coverage (including **coverage with evidence development**) and performance-linked reimbursement.

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.)

Meta-regression: in meta-analysis, techniques for relating the magnitude of an effect (e.g., change in a health outcome) to one or more characteristics of the primary studies used, such as patient characteristics, drug dose, duration of study, and year of publication.

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).

Multi-criteria decision analysis (MCDA): a transparent and objective method for decomposing a decision problem into a set of attributes or other criteria, including those that may be conflicting. It identifies and compares the attributes of alternatives (e.g., therapeutic options) from the perspectives of multiple stakeholders, and evaluates these alternatives by ranking, rating, or pairwise comparisons, using such stakeholder elicitation techniques as conjoint analysis and analytic hierarchy process.

Multiplicity: (or multiple comparisons) refers to errors in data interpretation that may arise from conducting multiple statistical analyses of the same data set. Such iterative analyses increase the probability of **false positive** (Type I) **error**, i.e., concluding incorrectly that an intervention is effective when the finding of a statistically significant treatment effect is due to random error. Types of multiplicity include analyses of numerous endpoints, stopping rules for RCTs that involve "multiple looks" at the data emerging from the same trial, and analyses of numerous subgroups.

N-of-1 trial: a clinical trial in which a single patient is the total population for the trial and in which a sequence of investigational and control interventions are allocated to the patient (i.e., a multiple crossover study conducted in a single patient). A trial in which random allocation is used to determine the sequence of interventions is given to a patient is an N-of-1 RCT. N-of-1 trials are used to determine

treatment effects in individuals, and sets of these trials can be used to estimate heterogeneity of treatment effects across a population.

Negative predictive value: an operating characteristic of a diagnostic test; negative predictive value 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 **positive predictive value.**)

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 considered to be 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 considered to be statistically significant, and the null hypothesis is rejected. If a confidence interval (e.g., of 95% or 99%) includes a net zero treatment effect (or a risk ratio of 1.0), then the null hypothesis cannot be rejected. The assumption of **equipoise** is the basis for testing the null hypothesis in RCTs.

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 ÷ (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 intervene, but simply observe the course of events over time. That is, investigators do not manipulate the use of, or deliver, an intervention or exposure (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) receive the intervention or exposure, and interpret the outcomes. These studies are more subject to selection bias than experimental studies such as randomized controlled trials. (Contrast with **interventional study**.)

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., $[P_t \div (1 - P_t)] \div [P_C \div (1 - P_C)]$. 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 ÷ (1.0 - 0.10)] ÷ [(0.25 ÷ (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 that act as a control for the other.)

Patient-centered outcomes (or patient-oriented outcomes): refers to health outcomes that patients experience across the variety of real-world settings, including: survival, functional status, quality of life, quality of death, symptoms, pain, nausea, psychosocial well-being, health utility (patient-perceived value of particular states of health), and patient satisfaction. (Excluded are outcomes that patients do not directly experience, such as blood pressure, lipid levels, bone density, viral load, or cardiac output.) Patient-centered outcomes can be assessed at a generic level or a disease/condition-specific level.

Patient-centered outcomes research (PCOR): generates evidence comparing the impact of health care on **patient-centered outcomes**. PCOR can draw on a wide variety of primary and secondary methods, including, e.g., practical or pragmatic RCTs, cluster randomized trials, and other trial designs, registries, insurance claims data, systematic reviews, and others.

Patient preference trials: trials designed to account for patient preferences, including evaluating the impact of preference on health outcomes. These trials have various designs. In some, the patients with a strong preference, e.g., for a new treatment or usual care, are assigned to a parallel group receiving their preferred intervention. The patients who are indifferent to receiving the new treatment or usual care are randomized into one group or another. Outcomes for the parallel, non-randomized groups are analyzed apart from the outcomes for the randomized groups. In other designs, patient preferences are recorded prior to the RCT, but all patients are randomized regardless of their stated preference, and subgroup analyses are conducted to determine the impact of preferences on outcomes.

Patient-reported outcomes: are those patient-centered outcomes that are self-reported by patients or obtained from patients (or reported on their behalf by their caregivers or surrogates) by an interviewer without interpretation or modification of the patient's response by other people, including clinicians.

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.

Performance bias refers to systematic differences between comparison groups in the care that is provided, or in exposure to factors other than the interventions of interest. This includes, e.g., deviating from the study protocol or assigned treatment regimens so that patients in control groups receive the intervention of interest, providing additional or co-interventions unevenly to the intervention and control groups, and inadequately blinding providers and patients to assignment to intervention and control groups, thereby potentially affecting whether or how assigned interventions or exposures are delivered. Techniques for diminishing performance bias include blinding of patients and providers (in RCTs and other controlled trials in particular), adhering to the study protocol, and sustaining patients' group assignments.

Personalized medicine: the tailoring of health care (including prevention, diagnosis, therapy) to the particular traits (or circumstances or other characteristics) of a patient that influence response to a heath care intervention. These may include genomic, epigenomic, microbiomic, sociodemographic, clinical, behavioral, environmental, and other personal traits, as well as personal preferences. Personalized medicine generally care does not refer to the creation of interventions that are unique to a patient, but the ability to classify patients into subpopulations that differ in their responses to particular interventions. (Also known as personalized health care.) The closely related term, precision medicine, is used synonymously, though it tends to emphasize the use of patient molecular traits to tailor therapy.

Pharmacogenetics: is the study of single gene interactions with drugs, including on metabolic variations that influence efficacy and toxicity. (See also **genomics** and **pharmacogenomics**.)

Pharmacogenomics: is the application of **pharmacogenetics** across the entire genome. (See also genomics.)

PICOTS: formulation of an evidence question based on: **P**opulation (e.g., condition, disease severity/stage, comorbidities, risk factors, demographics), **I**ntervention (e.g., technology type, regimen/dosage/frequency, technique/method of administration), **C**omparator (e.g., placebo, usual/standard care, active control), **O**utcomes (e.g., morbidity, mortality, quality of life, adverse events), **T**iming (e.g., duration/intervals of follow-up), and **S**etting (e.g., primary, inpatient, specialty, home care).

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.

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 - β).

Pragmatic (or practical) clinical trials (PCTs): are trials whose main attributes include: comparison of clinically relevant alternative interventions, a diverse population of study participants, participants recruited from heterogeneous practice settings, and data collection on a broad range of health outcomes. Some large simple trials are also PCTs.

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); also, the extent to which the mean estimate of a treatment effect is free from **random error**. 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.**)

Precision medicine: the tailoring of health care (particularly diagnosis and treatment using drugs and biologics) to the particular traits of a patient that influence response to a heath care intervention. Though it is sometimes used synonymously with **personalized medicine**, precision medicine tends to emphasize the use of patient molecular traits to tailor therapy.

Predictive validity refers to the ability to use differences in a measure of a construct to predict future events or outcomes. It may be considered a subtype of criterion validity.
Predictive value negative: see **negative predictive value**.

Predictive value positive: see **positive predictive value**.

Positive predictive value: an operating characteristic of a diagnostic test; positive predictive value 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 negative **predictive value.**)

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. (Contrast with **incidence**.)

Primary data methods involve collection of original data, including from randomized controlled trials, observational studies, case series, etc. (Contrast with **integrative methods**.)

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 and sponsors to submit, and publishers to accept, "positive" research reports, e.g., ones that detect beneficial treatment effects of a new intervention. Prospective registration of clinical trials and efforts to ensure publication of "negative" trials are two methods used to manage publication bias. Contrast with **reporting bias**.

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 error: (or random variation) the tendency for the estimated magnitude of a parameter (e.g., based on the average of a sample of observations of a treatment effect) to deviate randomly from the true magnitude of that parameter. Random error is due to chance alone; it is independent of the effects of systematic biases. In general, the larger the sample size is, the lower the random error is of the estimate of a parameter. As random error 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. (See **selection bias**.)

Randomized controlled trial (RCT): an experiment (and therefore a prospective study) 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**.)

Randomized-withdrawal trial: a form of "enriched" clinical trial design in which patients who respond favorably to an investigational intervention are then randomized to continue receiving that intervention or placebo. The study endpoints are return of symptoms or the ability to continue participation in the trial. The patients receiving the investigational intervention continue to do so only if they respond favorably, while those receiving placebo continue to do only until their symptoms return. This trial design is intended to minimize the time that patients receive placebo.

Rapid HTA: a more focused and limited version of HTA that is typically performed in approximately 4-8 weeks. Rapid HTAs are done in response to requests from decision makers who seek information support for near-term decisions. They offer a tradeoff between providing less-than-comprehensive and more uncertain information in time to act on a decision versus comprehensive and more certain information when the opportunity to make an effective decision may have passed. Rapid HTAs may involve some or all of: fewer types of impacts assessed or evidence questions, searching fewer bibliographic databases, relying on fewer types of studies (e.g., only systematic reviews or only RCTs), use of shorter and more qualitative syntheses with categorization of results without meta-analyses, and more limited or conditional interpretation of findings or recommendations.

Recall bias: refers to under-reporting, over-reporting, or other misreporting of events or other outcomes by patients or investigators who are asked to report these after their occurrence.

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).

Registries: any of a wide variety of repositories (usually electronic) of observations and related information about a group of patients (e.g., adult males living in a particular region), a disease (e.g., hypertension), an intervention (e.g., device implant), biological samples (e.g., tumor tissue), or other events or characteristics. Depending on criteria for inclusion in the database, the observations may have controls. As sources of observational data, registries can be useful for understanding real-world patient experience, including to complement safety and efficacy evidence from RCTs and other clinical trials. Registries can be used to determine the incidence of adverse events and to identify and follow-up with registered people at risk for adverse events. For determining causal relationships between interventions and outcomes, registries are limited by certain confounding factors (e.g., no randomization and possible selection bias in the process by which patients or events are recorded).

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**.)

Reporting bias: refers to systematic differences between reported and unreported findings, including, e.g., differential reporting of outcomes between comparison groups and incomplete reporting of study findings. Techniques for diminishing reporting bias include thorough reporting of outcomes consistent with outcome measures specified in the study protocol, attention to documentation and rationale for any post-hoc analyses not specified prior to the study, and reporting of literature search protocols and results for review articles. Differs from **publication bias**, which concerns the extent to which all relevant studies on given topic proceed to publication.

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 error on the estimate of a treatment effect.

Selection bias: refers to systematic distortions in assigning patients to intervention and control groups. This bias can result in baseline differences between the groups that could affect their prognoses and bias their treatment outcomes. In clinical trials, **allocation concealment** and **randomization** of treatment assignment are techniques for managing selection bias.

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 demonstrating **clinical significance**.)

Statistical test: a mathematical formula (or function) that is used to determine if the difference in outcomes between 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: a measure that is used as a substitute for a clinical endpoint of interest such as morbidity and mortality. They are used in clinical trials when it is impractical to measure the primary endpoint during the course of the trial, such as when observation of the clinical endpoint would require long follow-up. A surrogate endpoint is assumed, based on scientific evidence, to be a valid and reliable predictor of a clinical endpoint of interest. Examples are decrease in blood pressure as a predictor of decrease in strokes and heart attacks in hypertensive patients, increase in CD4+ cell counts as an indicator of improved survival of HIV/AIDS patients, and a negative culture as a predictor of cure of a bacterial infection. (See also **biomarker** and **intermediate endpoint**.)

Systematic review: a form of structured 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.

Teleoanalysis: an analysis that combines data from different types of study. In biomedical and health care research, specifically, it is "the synthesis of different categories of evidence to obtain a quantitative general summary of (a) the relation between a cause of a disease and the risk of the disease and (b) the extent to which the disease can be prevented. Teleoanalysis is different from meta-analysis because it relies on combining data from different classes of evidence rather than one type of study" (Wald 2003).

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 based on the difference between the observed outcomes of a treatment group and a control group. Commonly expressed as a difference in means (for continuous outcome variables); risk ratio (relative risk), odds ratio or risk difference (for binary outcomes such as mortality or health events); or number needed to treat to benefit the outcome of one person. (Also known as **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 or variable accurately reflects the concept that it is intended to measure. See **internal validity** and **external validity**.