Evidence-Based Imaging: Principles

1

L. Santiago Medina, C. Craig Blackmore, and
Kimberly E. Applegate

Contents

———————————————

L.S. Medina (✉)
Division of Neuroradiology-Neuroimaging, Department of Radiology, Miami Children's Hospital, Miami, FL, USA

Herbert Wertheim College of Medicine, Florida International University, Miami, FL, USA

Former Lecturer in Radiology, Harvard Medical School, Boston, MA, USA
e-mail: smedina@post.harvard.edu; Santiago.medina@mch.com

C.C. Blackmore
Department of Radiology, Virginia Mason Medical Center, Seattle, WA, USA
e-mail: craig.blackmore@vmmc.org

K.E. Applegate
Department of Radiology and Imaging Sciences, Emory University School of Medicine, Atlanta, GA, USA
e-mail: Keapple@emory.edu

Medicine is a science of uncertainty and an art of probability.
Sir William Osler

Discussion of Issues

What Is Evidence-Based Imaging?

The standard medical education in Western medicine has emphasized skills and knowledge learned from experts, particularly those encountered in the course of postgraduate medical education, and through national publications and meetings. This reliance on experts, referred to by Dr. Paul Gerber of Dartmouth Medical School as "eminence-based medicine" [1], is based on the construct that the individual practitioner, particularly a specialist devoting extensive time to a given discipline, can arrive at the best approach to a problem through his or her experience. The practitioner builds up an experience base over years and digests information from national experts who have a greater base of experience due to their focus in a particular area. The evidence-based imaging (EBI) paradigm, in contradistinction, is based on the precept that a single practitioner cannot through experience alone arrive at the best course of action. Assessment of appropriate medical care should instead be derived through an evidence-based process. The role of the practitioner, then, is not simply to accept information from an expert but rather to assimilate and critically assess the research evidence that exists in the literature to guide a clinical decision [2–4].

Fundamental to the adoption of the principles of EBI is the understanding that medical care is not optimal. The life expectancy at birth in the United States for males and females in 2005 was 75 and 80 years, respectively (Table 1.1). This is slightly lower than the life expectancies in other industrialized nations such as the United Kingdom and Australia (Table 1.1). In fact, the World Health Organization ranks the USA 50th in life expectancy and 72nd in overall health. The United States spent at least 15.2 % of the gross domestic product (GDP) in order to achieve this life expectancy. This was significantly more than the United Kingdom and Australia, which spent

about half that (Table 1.1). In addition, the US per capita health expenditure was $6,096, which was twice the expenditure in the United Kingdom or Australia. In short, the United States spends significantly more money and resources than other industrialized countries to achieve a similar or slightly worse outcome in life expectancy. This implies that a significant amount of resources is wasted in the US health-care system. In 2007, the United States spent $2.3 trillion in health care or 16 % of its GDP. By 2016, the US health percent of the GDP is expected to grow to 20 % or $4.2 trillion [5]. Recent estimates prepared by the Commonwealth Fund Commission (USA) on a High Performance Health System indicate that $1.5 trillion could be saved over a 10-year period if a combination of options, including evidence-based medicine and universal health insurance, was adopted [6].

Simultaneous with the increase in health-care costs has been an explosion in available medical information. The National Library of Medicine PubMed search engine now lists over 18 million citations. Practitioners cannot maintain familiarity with even a minute subset of this literature without a method of filtering out publications that lack either relevance or appropriate methodological quality. EBI is a promising method of identifying appropriate information to guide practice and to improve the efficiency and effectiveness of imaging.

Evidence-based imaging is defined as medical decision making based on clinical integration of the best medical imaging research evidence with the physician's expertise and with patient's expectations [2–4]. The best medical imaging research evidence often comes from the basic sciences of medicine. In EBI, however, the basic science knowledge has been translated into patient-centered clinical research, which determines the accuracy and role of diagnostic and therapeutic imaging in patient care [3]. New research may make current diagnostic tests obsolete and provide evidence that new tests are more

Table 1.1 Life expectancy and health-care spending in three developed countries

	Life expectancy at birth (2009)		Percentage of GDP in health care (2008) (%)	Per capita health expenditure (2008)
	Male	Female		
United States	75.7	80.6	16.4	$7,720
United Kingdom	78.3	82.5	8.8	$3,281
Australia	79.3	83.9	8.7	$3,445

Source: Organization for Economic Cooperation and Development: http://stats.oecd.org/Index.aspx?DataSetCode=HEALTH
Reprinted, with revisions, with kind permission of Springer Science+Business Media. Medina LS, Blackmore CC, Applegate KE. Principles of evidence-based imaging. In: Medina LS, Applegate KE, Blackmore CC editors. Evidence-based imaging in pediatrics: optimizing imaging in pediatric patient care. New York: Springer Science+Business Media; 2010.
GDP gross domestic product

accurate, less invasive, safer, and less costly [3]. The physician's expertise entails the ability to use the referring physician's clinical skills and past experience to rapidly identify individuals who will benefit from the diagnostic information of an imaging test [4]. Patient's expectations are important because each individual has values and preferences that should be integrated into the clinical decision making [3]. When these three components of medicine come together, clinicians, imagers and patients form a diagnostic team, which will optimize clinical outcomes and quality of life for our patients.

The Evidence-Based Imaging Process

The EBI process involves a series of steps: (a) formulation of the clinical question, (b) identification of the medical literature, (c) assessment of the literature, (d) types of economic analyses in medicine, (e) summary of the evidence, and (f) application of the evidence to derive an appropriate clinical action. This book is designed to bring the EBI process to the clinician and imager in a user-friendly way. This introductory chapter details each of the steps in the EBI process. Chapter 2, "Assessing the Imaging Literature: Understanding Error and Bias" discusses how to critically assess the literature. The rest of the book makes available to practitioners the EBI approach to important neuroimaging issues. Each chapter addresses common disorders encountered by the neuroradiologist evaluating the brain,

spine, and head and neck. Relevant clinical questions are delineated, and then each chapter discusses the results of the critical analysis of the identified literature. Finally, we provide simple recommendations for the various clinical questions, including the strength of the evidence that supports these recommendations.

(a) Formulating the Clinical Question

The first step in the EBI process is formulation of the clinical question. The entire process of EBI arises from a question that is asked in the context of clinical practice. However, often formulating a question for the EBI approach can be more challenging than one would believe intuitively. To be approachable by the EBI format, a question must be specific to a clinical situation, a patient group, and an outcome or action. For example, it would not be appropriate to simply ask which imaging technique is better – computed tomography (CT) or radiography. The question must be refined to include the particular patient population and the action that the imaging will be used to direct. One can refine the question to include a particular population (which imaging technique is better in pediatric victims of high-energy blunt trauma) and to guide a particular action or decision (to exclude the presence of unstable cervical spine fracture). The full EBI question then becomes, in pediatric victims of high-energy blunt trauma, which imaging modality is preferred, CT or radiography, to exclude the presence of unstable cervical

spine fracture? This book addresses questions that commonly arise when employing an EBI approach for conditions encountered by neuroradiologists. These questions and issues are detailed at the start of each chapter. One popular method used to teach how to develop a good clinical question is called the "PICO" (*P*atient, *I*ntervention, *C*omparison, *O*utcome) format. This method provides structure to formulate the necessary elements for a good clinical question that includes information about the patient, the problem to be solved, the intervention (such as a diagnostic test) and its comparison intervention (perhaps a newer diagnostic test), and the outcome of interest (e.g., what the patient wants, or is concerned about).

(b) Identifying the Medical Literature

The process of EBI requires timely access to the relevant medical literature to answer the question. Fortunately, massive on-line bibliographical references such as PubMed, Embase, Cochrane, and the Web of Science databases are available. In general, titles, indexing terms, abstracts, and often the complete text of much of the world's medical literature are available through these on-line sources. Also, medical librarians are a potential resource to aid identification of the relevant imaging literature. A limitation of today's literature data sources is that often too much information is available and too many potential resources are identified in a literature search. There are currently over 50 radiology journals, and imaging research is also frequently published in journals from other medical subspecialties. We are often confronted with more literature and information than we can process. The greater challenge is to sift through the literature that is identified to select that which is appropriate.

(c) Assessing the Literature

To incorporate evidence into practice, the clinician must be able to understand the published literature and to critically evaluate the strength of the evidence. In this introductory chapter on the process of EBI, we focus on discussing types of research studies. Chapter 2, "Assessing the Imaging Literature: Understanding Error and Bias" is a detailed discussion of the issues in determining the validity and reliability of the reported results.

1. What Are the Types of Clinical Studies?

An initial assessment of the literature begins with determination of the type of clinical study: descriptive, analytical, or experimental [7]. *Descriptive* studies are the most rudimentary, as they only summarize disease processes as seen by imaging, or discuss how an imaging modality can be used to create images. Descriptive studies include case reports and case series. Although they may provide important information that leads to further investigation, descriptive studies are not usually the basis for EBI.

Analytic or *observational* studies include cohort, case–control, and cross-sectional studies (Table 1.2). Cohort studies are defined by risk factor status, and case–control studies consist of groups defined by disease status [8]. Both case–control and cohort studies may be used to define the association between an intervention, such as an imaging test, and patient outcome [9]. In a cross-sectional (prevalence) study, the researcher makes all of his measurements on a single occasion. The investigator draws a sample from the population (i.e., headache in 15–45-year-old females) and determines distribution of variables within that sample [7]. The structure of a cross-sectional study is similar to that of a cohort study except that all pertinent measurements (i.e., number of head CT and MRI examinations) are made at once, without a follow-up period. Cross-sectional studies can be used as a major source for health and habits of different populations and countries, providing estimates of such parameters as the prevalence of stroke, brain tumors, and congenital anomalies [7, 10].

In *experimental studies* or *clinical trials*, a specific intervention is performed

Table 1.2 Study design

	Prospective follow-up	Randomization of subjects	Controls
Case report or series	No	No	No
Cross-sectional study	No	No	Yes
Case–control study	No	No	Yes
Cohort study	Yes/no	No	Yes
Randomized controlled trial	Yes	Yes	Yes

Reprinted with the kind permission of Springer Science +Business Media from Medina LS, Blackmore CC. Evidence-based imaging: optimizing imaging in patient care. New York: Springer Science+Business Media; 2006

and the effect of the intervention is measured by using a control group (Table 1.2). The control group may be tested with a different diagnostic test and treated with a placebo or an alternative mode of therapy [7, 11]. Clinical trials are epidemiologic designs that can provide data of high quality that resemble the controlled experiments done by basic science investigators [8]. For example, clinical trials may be used to assess new diagnostic tests (e.g., CT perfusion imaging for stroke diagnosis and management) or new interventional procedures (e.g., catheter embolization for cerebral aneurysms).

Studies are also traditionally divided into retrospective and prospective (Table 1.2) [7, 11]. These terms refer more to the way the data are gathered than to the specific type of study design. In *retrospective studies*, the events of interest have occurred before study onset. Retrospective studies are usually done to assess rare disorders, for pilot studies, and when prospective investigations are not possible. If the disease process is considered rare, retrospective studies facilitate the collection of enough subjects to have meaningful data. For a pilot project, retrospective studies facilitate the collection of preliminary data that can be used to

improve the study design in future prospective studies. The major drawback of a retrospective study is incomplete data acquisition and resultant bias [10]. Case–control studies are usually retrospective because the outcome or disease status needs to have occurred in order to form the comparison groups. For example, in a case–control study, subjects in the case group (patients with hemorrhagic stroke) are compared with subjects in a control group (nonhemorrhagic stroke) to determine factors associated with hemorrhage (e.g., hypertension, duration of symptoms, presence of prior neurologic deficit) [10].

In *prospective studies*, the event of interest transpires after study onset. Prospective studies, therefore, are the preferred mode of study design, as they facilitate better control of the design (accounting for potential bias) and the quality of the data acquired [7]. Prospective studies, even large studies, can be performed efficiently and in a timely fashion if done on common diseases at major institutions, as multicenter trials with adequate study populations [12]. The major drawback of a prospective study is the need to make sure that the institution and personnel comply with strict rules concerning consents, protocols, and data acquisition [11]. Persistence and dogged determination are crucial to completing a prospective study. Cohort studies and clinical trials are usually prospective. For example, a cohort study could be performed in children with sickle-cell disease who are poorly compliant with their transfusion therapy in which the risk factor of positive transcranial Doppler studies is correlated with neurocognitive complications, as the patients are followed prospectively over time [10].

The strongest study design is the prospective randomized, blinded clinical trial (Table 1.2) [7]. The randomization process helps to distribute known and

unknown confounding factors, and blinding helps to prevent observer bias from affecting the results [7, 8]. However, there are often circumstances in which it is not ethical or practical to randomize and follow patients prospectively. This is particularly true in rare conditions and in studies to determine causes or predictors of a particular condition [9]. Finally, randomized clinical trials are expensive and may require many years to conduct. Not surprisingly, randomized clinical trials are uncommon in radiology. The evidence that supports much of radiology practice is derived from cohort and other observational studies. More randomized clinical trials are necessary in radiology to provide sound data to use for EBI practice [3]. Also, more "outcomes-based studies" are needed in radiology to generate more relevant EBI data.

2. What Is the Diagnostic Performance of a Test: Sensitivity, Specificity, Positive and Negative Predictive Values, and Receiver Operating Characteristic Curve?

Defining the presence or absence of an outcome (i.e., disease and nondisease) is based on a standard of reference (Table 1.3). While a perfect standard of reference or so-called gold standard can never be obtained, careful attention should be paid to the selection of the standard that should be widely believed to offer the best approximation to the truth [13].

In evaluating diagnostic tests, we rely on the statistical calculations of sensitivity and specificity (see Appendix 1). Sensitivity and specificity of a diagnostic test are based on the two-way (2 × 2) table (Table 1.3). Sensitivity refers to the proportion of subjects with the disease who have a positive test and is referred to as the true positive rate (Fig. 1.1a, b). Sensitivity, therefore, indicates how well a test identifies the subjects with disease [7, 14].

Specificity is defined as the proportion of subjects without the disease who have a negative index test (Fig. 1.1a,b) and is

Table 1.3 Two-way table of diagnostic testing

Test result	Disease (gold standard)	
	Present	Absent
Positive	a (TP)	b (FP)
Negative	c (FN)	d (TN)

Reprinted with the kind permission of Springer Science+Business Media from Medina LS, Blackmore CC. Evidence-based imaging: optimizing imaging in patient care. New York: Springer Science+Business Media; 2006. *FN* false negative, *FP* false positive, *TN* true negative, TP true positive

referred to as the true negative rate. Specificity, therefore, indicates how well a test identifies the subjects with no disease [7, 11]. It is important to note that the sensitivity and specificity are characteristics of the test being evaluated and are therefore usually independent of the prevalence (proportion of individuals in a population who have disease at a specific instant) because the sensitivity only deals with the diseased subjects, whereas the specificity only deals with the nondiseased subjects. However, sensitivity and specificity both depend on a threshold point for considering a test positive and hence may change according to which threshold is selected in the study [11, 14, 15] (Fig. 1.1a). Excellent diagnostic tests have high values (close to 1.0) for both sensitivity and specificity. Given exactly the same diagnostic test, and exactly the same subjects confirmed with the same reference test, the sensitivity with a low threshold is greater than the sensitivity with a high threshold. Conversely, the specificity with a low threshold is less than the specificity with a high threshold (Fig. 1.1b) [14, 15].

The positive predictive value is defined as the probability that a patient will have a disease given that the patient's test is positive. In other words, when a group of patients test positive, we want to know how frequently they will have the disease. The formula for the positive predictive value (PPV) is provided in the table in Appendix 1. Similarly, the negative
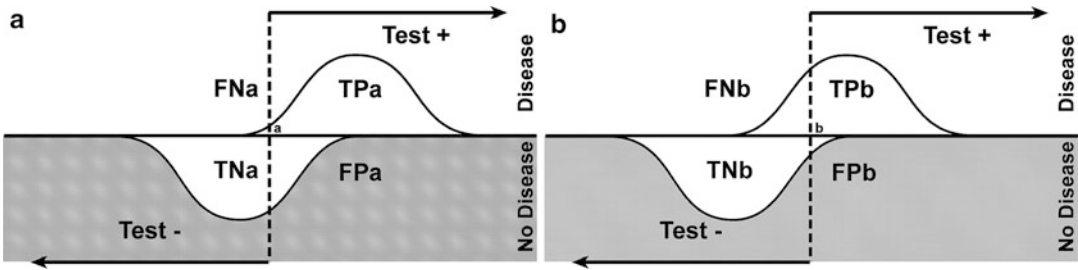
Fig. 1.1 Test with a low (**a**) and high (**b**) threshold. The sensitivity and specificity of a test change according to the threshold selected; hence, these diagnostic performance parameters are threshold dependent. Sensitivity with low threshold (TPa/diseased patients) is greater than sensitivity with a higher threshold (TPb/diseased patients). Specificity with a low threshold (TNa/nondiseased patients) is less than specificity with a high threshold (TNb/nondiseased patients). *FN* false negative, *FP* false positive, *TN* true negative, *TP* true positive (Reprinted with permission of the American Society of Neuroradiology from Medina L. AJNR Am J Neuroradiol 1999; 20:1584–96)

predictive value (NPV) refers to the probability that a group of patients that test negative for a disease or condition will actually not have the disease. It is important to understand that while sensitivity and specificity are relatively independent of disease prevalence, the PPV and NPV are not. Examples 1 and 2 (Appendix 2) provide a demonstration of what happens to the PPV and NPV with a change in disease prevalence. When there is concern about large prevalence effects, the likelihood ratio can be used to estimate the posttest probability of disease. This issue is discussed in the next section.

The effect of threshold on the ability of a test to discriminate between disease and nondisease can be measured by a receiver operating characteristic (ROC) curve [11, 15]. The ROC curve is used to indicate the trade-offs between sensitivity and specificity for a particular diagnostic test and hence describes the discrimination capacity of that test. An ROC graph shows the relationship between sensitivity (y axis) and $1 -$ specificity (x axis) plotted for various cutoff points. If the threshold for sensitivity and specificity is varied, an ROC curve can be generated. The diagnostic performance of a test can be estimated by the area under the ROC curve. The steeper the ROC curve, the greater the area and the better the discrimination of the test (Fig. 1.2a–c). A test with perfect discrimination has an area of 1.0, whereas a test with only random discrimination has an area of 0.5 (Fig. 1.2a–c). The area under the ROC curve usually determines the overall diagnostic performance of the test independent of the threshold selected [11, 15]. The ROC curve is threshold independent because it is generated by using varied thresholds of sensitivity and specificity. Therefore, when evaluating a new imaging test, in addition to the sensitivity and specificity, an ROC curve analysis should be done so that the threshold-dependent and threshold-independent diagnostic performance can be fully determined [10].

3. What Are Cost-Effectiveness and Cost–Utility Studies?

Cost-effectiveness analysis (CEA) is a scientific technique used to assess alternative health-care strategies on both cost and effectiveness [16–18]. It can be used to develop clinical and imaging practice guidelines and to set health policy [19]. However, it is not designed to be the final answer to the decision-making process; rather, it provides a detailed analysis of the cost and outcome variables and how they are affected by competing medical and diagnostic choices.
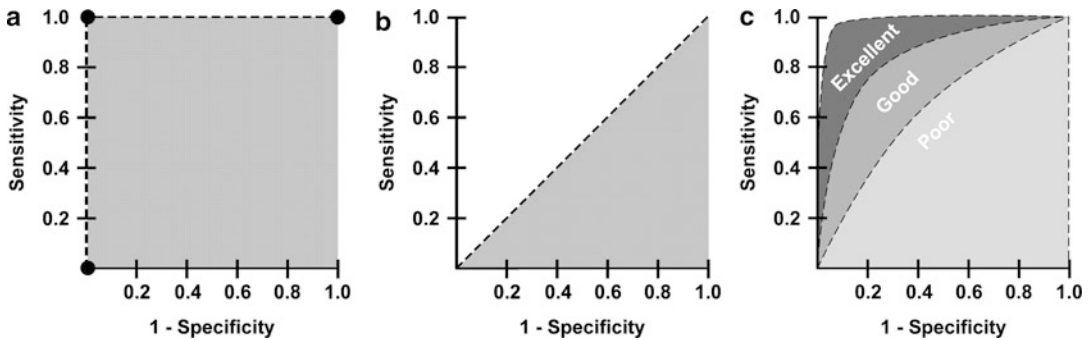
Fig. 1.2 The perfect test (**a**) has an area under the curve (AUC) of 1. The useless test (**b**) has an AUC of 0.5. The typical test (**c**) has an AUC between 0.5 and 1. The greater the AUC (i.e., excellent > good > poor), the better the diagnostic performance (Reprinted with permission of the American Society of Neuroradiology from Medina L. AJNR Am J Neuroradiol. 1999;20: 1584–96)

Health dollars are limited regardless of the country's economic status. Hence, medical decision makers must weigh the benefits of a diagnostic test (or any intervention) in relation to its cost. Health-care resources should be allocated so the maximum health-care benefit for the entire population is achieved [10]. Cost-effectiveness analysis is an important tool to address health cost-outcome issues in a cost-conscious society. Countries such as Australia usually require robust CEA before drugs are approved for national use [10]. Health-care decisions are often made from a "societal perspective," one that looks at a group benefit but which may not result in individual benefit.

Unfortunately, the term *cost-effectiveness* is often misused in the medical literature [20]. To say that a diagnostic test is truly cost-effective, a comprehensive analysis of the entire short- and long-term outcomes and costs needs to be considered. Cost-effectiveness analysis is a technique used to determine which of the available tests or treatments are worth the additional costs [21].

There are established guidelines for conducting robust CEA. The US Public Health Service formed a panel of experts on cost-effectiveness in health and medicine to create detailed standards for cost-effectiveness analysis. The panel's recommendations were published as a book in 1996 [21].

(d) Types of Economic Analyses in Medicine

There are four well-defined types of economic evaluations in medicine: cost-minimization studies, cost–benefit analyses, cost-effectiveness analyses, and cost–utility analyses. They are all commonly lumped under the term *cost-effectiveness analysis*. However, significant differences exist among these different studies.

Cost-minimization analysis is a comparison of the cost of different health-care strategies that are assumed to have identical or similar effectiveness [16]. In medical practice, few diagnostic tests or treatments have identical or similar effectiveness. Therefore, relatively few articles have been published in the literature with this type of study design [22]. For example, a recent study demonstrated that functional magnetic resonance imaging (MRI) and the Wada test have similar effectiveness for language lateralization, but the latter is 3.7 times more costly than the former [23].

Cost–benefit analysis (CBA) uses monetary units such as dollars or euros to compare the costs of a health intervention with its health benefits [16]. It converts all benefits to a cost equivalent and is commonly used in the financial world where the cost and benefits of multiple industries can be changed to only monetary values. One

method of converting health outcomes into dollars is through a contingent valuation or willingness-to-pay approach. Using this technique, subjects are asked how much money they would be willing to spend to obtain, or avoid, a health outcome. For example, a study by Appel et al. [24] found that individuals would be willing to pay $50 for low-osmolar contrast agents to decrease the probability of side effects from intravenous contrast. However, in general, health outcomes and benefits are difficult to transform to monetary units; hence, CBA has had limited acceptance and use in medicine and diagnostic imaging [16, 25].

Cost-effectiveness analysis (CEA) refers to analyses that study both the effectiveness and cost of competing diagnostic or treatment strategies, where effectiveness is an objective measure (e.g., intermediate outcome: number of strokes detected; or long-term outcome: life-years saved). Radiology CEAs often use intermediate outcomes, such as lesion identified, length of stay, and number of avoidable surgeries [16, 18]. However, ideally, long-term outcomes such as life-years saved (LYS) should be used [21]. By using LYS, different health-care fields or interventions can be compared. Given how few exist, there is a need for more "outcome-based studies" in radiology and the imaging sciences.

Cost–utility analysis is similar to CEA except that the effectiveness also accounts for quality of life. Quality of life is measured as utilities that are based on patient preferences [16]. The most commonly used utility measurement is the quality-adjusted life year (QALY). The rationale behind this concept is that the QALY of excellent health is more desirable than the same 1 year with substantial morbidity. The QALY model uses preferences with weight for each health state on a scale from 0 to 1, where 0 is death and 1 is perfect health. The utility score for each health state is multiplied by the length of time the patient spends in that specific health state [16, 26]. For example, assume that a patient

with an untreated Chiari I malformation has a utility of 0.8 and he spends 1 year in this health state. The patient with the Chiari I malformation would have a 0.8 QALY in comparison with his neighbor who has a perfect health and hence a 1 QALY.

Cost–utility analysis incorporates the patient's subjective value of the risk, discomfort, and pain into the effectiveness measurements of the different diagnostic or therapeutic alternatives. Ideally, all medical decisions should reflect the patient's values and priorities [26]. That is the explanation of why cost–utility analysis is the preferred method for evaluation of economic issues in health [19, 21]. For example, in low-risk newborns with intergluteal dimple suspected of having occult spinal dysraphism, ultrasound was the most effective strategy with an incremental cost-effectiveness ratio of $55,100 per QALY. In intermediate-risk newborns with low anorectal malformation, however, MRI was more effective than ultrasound at an incremental cost-effectiveness of $1,000 per QALY [27].

Assessment of Outcomes: The major challenge to cost–utility analysis is the quantification of health or quality of life. One way to quantify health is descriptive analyses. By assessing what patients can and cannot do, how they feel, their mental state, their functional independence, their freedom from pain, and any number of other facets of health and well-being that are referred to as domains, one can summarize their overall health status. Instruments designed to measure these domains are called health status instruments. A large number of health status instruments exist, both general instruments, such as the SF-36 [28], and instruments that are specific to particular disease states, such as the Roland scale for back pain. These various scales enable the quantification of health benefit. For example, Jarvik et al. [29] found no significant difference in the Roland score between patients randomized to MRI versus radiography for low back pain, suggesting that MRI was not worth the additional cost.

Assessment of Cost: All forms of economic analysis require assessment of cost. However, assessment of cost in medical care can be confusing, as the term *cost* is used to refer to many different things. The use of charges for any sort of cost estimation, however, is inappropriate. Charges are arbitrary and have no meaningful use. Reimbursements, derived from Medicare and other fee schedules, are useful as an estimation of the amounts society pays for particular health-care interventions. For an analysis taken from the societal perspective, such reimbursements may be most appropriate. For analyses from the institutional perspective or in situations where there are no meaningful Medicare reimbursements, assessment of actual direct and overhead costs may be appropriate [30].

Direct cost assessment centers on the determination of the resources that are consumed in the process of performing a given imaging study, including *fixed costs* such as equipment and *variable costs* such as labor and supplies. Cost analysis often utilizes activity-based costing and time motion studies to determine the resources consumed for a single intervention in the context of the complex health-care delivery system. Activity-based accounting is a type of accounting that assigns costs to each resource activity based on resource consumption, decreasing the amount of indirect costs with this method. Time and motion studies are time-intensive observational methods used to understand and improve work efficiency in a process. *Overhead*, or *indirect cost*, assessment includes the costs of buildings, overall administration, taxes, and maintenance that cannot be easily assigned to one particular imaging study. Institutional cost accounting systems may be used to determine both the direct costs of an imaging study and the amount of institutional overhead costs that should be apportioned to that particular test. For example, Medina et al. [31] studied the total direct costs of the Wada test ($1,130.01 $\pm$ $138.40) and of functional MR imaging ($301.82 $\pm$ $10.65) that were significantly different ($P < .001$).

The cost of the Wada test was 3.7 times higher than that of functional MR imaging.

(e) Summarizing the Data

The results of the EBI process are a summary of the literature on the topic, both quantitative and qualitative. *Quantitative analysis* involves, at minimum, a descriptive summary of the data and may include formal *meta-analysis*, where there is sufficient reliably acquired data. *Qualitative analysis* requires an understanding of error, bias, and the subtleties of experimental design that can affect the quality of study results. Qualitative assessment of the literature is covered in detail in Chap. 2, "Assessing the Imaging Literature: Understanding Error and Bias"; this section focuses on meta-analysis and the quantitative summary of data.

The goal of the EBI process is to produce a single summary of all of the data on a particular clinically relevant question. However, the underlying investigations on a particular topic may be too dissimilar in methods or study populations to allow for a simple summary. In such cases, the user of the EBI approach may have to rely on the single study that most closely resembles the clinical subjects upon whom the results are to be applied or may be able only to reliably estimate a range of possible values for the data.

Often, there is abundant information available to answer an EBI question. Multiple studies may be identified that provide methodologically sound data. Therefore, some method must be used to combine the results of these studies in a summary statement. *Meta-analysis* is the method of combining results of multiple studies in a statistically valid manner to determine a summary measure of accuracy or effectiveness [32, 33]. For diagnostic studies, the summary estimate is generally a summary sensitivity and specificity, or a summary ROC curve.

The process of performing meta-analysis parallels that of performing primary research.

However, instead of individual subjects, the meta-analysis is based on individual studies of a particular question. The process of selecting the studies for a meta-analysis is as important as unbiased selection of subjects for a primary investigation. Identification of studies for meta-analysis employs the same type of process as that for EBI described above, employing Medline and other literature search engines. Critical information from each of the selected studies is then abstracted usually by more than one investigator. For a meta-analysis of a diagnostic accuracy study, the numbers of true positives, false positives, true negatives, and false negatives would be determined for each of the eligible research publications. The results of a meta-analysis are derived not just by simply pooling the results of the individual studies but instead by considering each individual study as a data point and determining a summary estimate for accuracy based on each of these individual investigations. There are sophisticated statistical methods of combining such results [34].

Like all research, the value of a meta-analysis is directly dependent on the validity of each of the data points. In other words, the quality of the meta-analysis can only be as good as the quality of the research studies that the meta-analysis summarizes. In general, a meta-analysis cannot compensate for selection and other biases in the primary data. If the studies included in a meta-analysis are different in some way, or are subject to some bias, then the results may be too heterogeneous to combine in a single summary measure. Exploration for such heterogeneity is an important component of a meta-analysis.

The ideal for EBI is that all practice be based on the information from one or more well-performed meta-analyses. However, there is often too little data or too much heterogeneity to support a formal meta-analysis. Understanding the hierarchy of next best available evidence, and how to find it, is then critical for readers of the literature.

(f) Applying the Evidence

The final step in the EBI process is to apply the summary results of the medical literature to the EBI question. Sometimes the answer to an EBI question is a simple yes or no, as for this question: Does a normal clinical exam exclude unstable cervical spine fracture in patients with minor trauma? Commonly, the answers to EBI questions are expressed as some measure of accuracy. For example, how good is MRI for detecting acute ischemic infarction (<6 h)? The answer is that MRI has an approximate sensitivity of 91 % and specificity of 95 % [35]. However, to guide practice, EBI must be able to answer questions that go beyond simple accuracy; for example, should MRI then be used for the early detection of acute infarct? To answer this question, it is useful to divide the types of literature studies into a *hierarchical framework* [36] (Table 1.4). At the foundation in this hierarchy is assessment of *technical efficacy*: studies that are designed to determine if a particular proposed imaging method or application has the underlying ability to produce an image that contains useful information. Information for technical efficacy would include signal-to-noise ratios, image resolution, and freedom from artifacts. The second step in this hierarchy is to determine if the image predicts the truth. This is the *accuracy* of an imaging study and is generally studied by comparing the test results to a reference standard and defining the sensitivity and the specificity of the imaging test. The third step is to incorporate the physician into the evaluation of the imaging intervention by evaluating the effect of the use of the particular imaging intervention on physician certainty of a given diagnosis (physician decision making) and on the actual management of the patient (*therapeutic efficacy*). Finally, to be of value to the patient, an imaging procedure must not only affect management but also improve outcome. *Patient outcome efficacy* is the determination of the effect of a given imaging intervention on the length and quality of life of a patient. A final efficacy level

Table 1.4 Imaging effectiveness hierarchy

Technical efficacy: production of an image or information
Measures: signal-to-noise ratio, resolution, absence of artifacts
Accuracy efficacy: ability of test to differentiate between disease and nondisease
Measures: sensitivity, specificity, receiver operator characteristic curves
Diagnostic-thinking efficacy: impact of test on likelihood of diagnosis in a patient
Measures: pre- and posttest probability, diagnostic certainty
Treatment efficacy: potential of test to change therapy for a patient
Measures: treatment plan, operative or medical treatment frequency
Outcome efficacy: effect of use of test on patient health
Measures: mortality, quality-adjusted life years, health status
Societal efficacy: appropriateness of test from perspective of society
Measures: cost-effectiveness analysis, cost–utility analysis

Adapted with permission of Fryback DG, Thornbury JR. Med Decis Making. 1991;11:88–94
Reprinted with the kind permission of Springer Science+Business Media from Medina LS, Blackmore CC, Applegate KE. Evidence-based imaging: improving the quality of imaging in patient care. Revised Edition. New York: Springer Science+Business Media; 2011

is that of society, which examines the question of not simply the health of a single patient but that of the health of society as a whole, encompassing the effect of a given intervention on all patients and including the concepts of *cost* and *cost-effectiveness* [36].

Some additional research studies in imaging, such as clinical prediction rules, do not fit readily into this hierarchy. *Clinical prediction rules* are used to define a population in whom imaging is appropriate or can safely be avoided. Clinical prediction rules can also be used in combination with CEA as a way of deciding between competing imaging strategies [37].

Bayes' Theorem, Predictive Values, and the Likelihood Ratio

Ideally, information would be available to address the effectiveness of a diagnostic test on all levels of the hierarchy. Commonly in imaging, however, the only reliable information that is available is that of diagnostic accuracy. It is incumbent upon the user of the imaging literature to determine if a test with a given sensitivity and specificity is appropriate for use in a given clinical situation. To address this issue, the concept of Bayes' theorem is critical. Bayes' theorem is based on the concept that the value of the diagnostic tests depends not only on the characteristics of the test (sensitivity and specificity) but also on the prevalence (pretest probability) of the

disease in the test population. As the prevalence of a specific disease decreases, it becomes less likely that someone with a positive test will actually have the disease and more likely that the positive test result is a false positive. The relationship between the sensitivity and specificity of the test and the prevalence (pretest probability) can be expressed through the use of Bayes' theorem (see Appendix 2) [11, 14] and the likelihood ratio. The positive likelihood ratio (PLR) estimates the likelihood that a positive test result will raise or lower the pretest probability, resulting in estimation of the posttest probability [where PLR = sensitivity/(1 − specificity)]. The negative likelihood ratio (NLR) estimates the likelihood that a negative test result will raise or lower the pretest probability, resulting in estimation of the posttest probability [where NLR = (1 − sensitivity)/specificity] [38]. The likelihood ratio (LR) is not a probability but a ratio of probabilities. The positive predictive value (PPV) refers to the probability that a person with a positive test result actually has the disease. The negative predictive value (NPV) is the probability that a person with a negative test result does not have the disease. Since the predictive value is determined once the test results are known (i.e., sensitivity and specificity), it actually represents a posttest probability; hence, the posttest probability is determined by both the prevalence (pretest probability) and the test

information (i.e., sensitivity and specificity). Thus, the predictive values are affected by the prevalence of disease in the study population.

A practical understanding of this concept is shown in examples 1 and 2 in Appendix 2. The example shows an increase in the PPV from 0.67 to 0.98 when the prevalence of carotid artery disease is increased from 0.16 to 0.82. Note that the sensitivity and specificity of 0.83 and 0.92, respectively, remain unchanged. If the test information is kept constant (same sensitivity and specificity), the pretest probability (prevalence) affects the posttest probability (predictive value) results.

The concept of diagnostic performance discussed above can be summarized by incorporating the data from Appendix 2 into a nomogram for interpreting diagnostic test results (Fig. 1.3). For example, two patients present to the emergency department complaining of left-sided weakness. The treating physician wants to determine if they have a stroke from carotid artery disease. The first patient is an 8-year-old boy complaining of chronic left-sided weakness. Because of the patient's young age and chronic history, he was determined clinically to be in a low-risk category for carotid artery disease-induced stroke and hence with a low pretest probability of 0.05 (5 %). Conversely, the second patient is 65 years old and is complaining of acute onset of severe left-sided weakness. Because of the patient's older age and acute history, he was determined clinically to be in a high-risk category for carotid artery disease-induced stroke and hence with a high pretest probability of 0.70 (70 %). The available diagnostic imaging test was unenhanced head CT followed by CT angiography. According to the radiologist's available literature, the sensitivity and specificity of these tests for carotid artery disease and stroke were each 0.90. The positive likelihood ratio (sensitivity/1 − specificity) calculation derived by the radiologist was 0.90/(1 − 0.90) = 9. The posttest probability for the 8-year-old patient is therefore 30 % based on a pretest probability of 0.05 and a likelihood ratio of 9 (Fig. 1.3, dashed line A). Conversely, the posttest probability for the 65-year-old patient is greater than 95 % based on a pretest probability
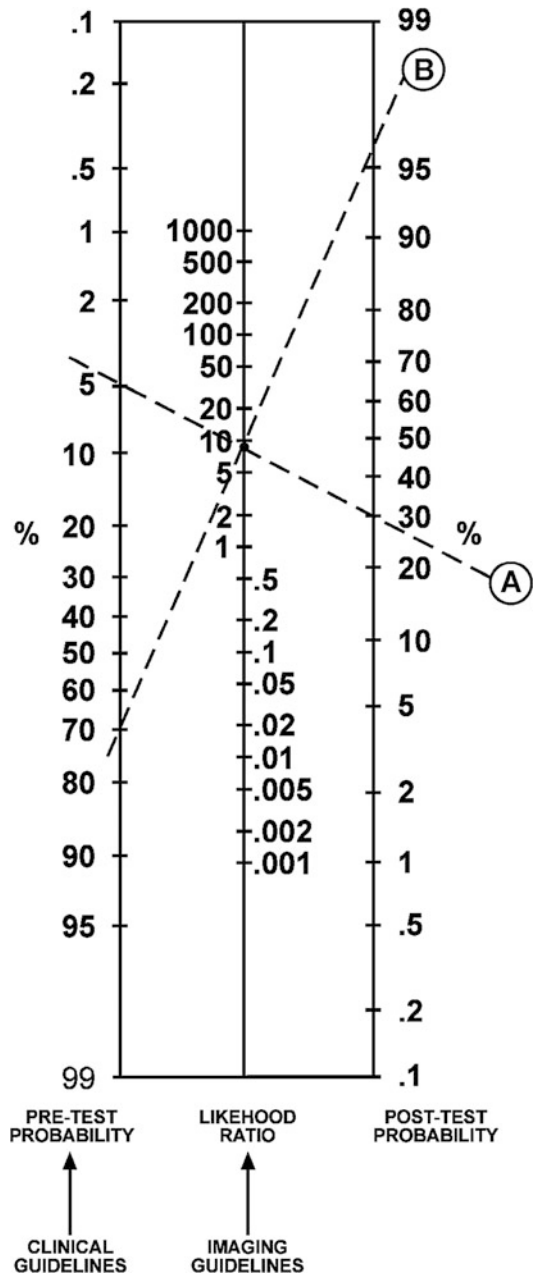

Fig. 1.3 Bayes' theorem nomogram for determining posttest probability of disease using the pretest probability of disease and the likelihood ratio from the imaging test. Clinical and imaging guidelines are aimed at increasing the pretest probability and likelihood ratio, respectively. Worked example is explained in the text (Reprinted with permission from Medina L, Aguirre E, Zurakowski D. Neuroimaging Clin N Am. 2003;13:157–65)

of 0.70 and a positive likelihood ratio of 9 (Fig. 1.3, dashed line B). Clinicians and radiologists can use this scale to understand the probability of disease in different risk groups and for imaging studies with different diagnostic performance.

Jaeschke et al. [38] have proposed a rule of thumb regarding the interpretation of the LR. For PLR, tests with values greater than 10 have a large difference between pretest and posttest probability with conclusive diagnostic impact; values of 5–10 have a moderate difference in test probabilities and moderate diagnostic impact; values of 2–5 have a small difference in test probabilities and sometimes an important diagnostic impact; and values less than 2 have a small difference in test probabilities and seldom have important diagnostic impact. For NLR, tests with values less than 0.1 have a large difference between pretest and posttest probability with conclusive diagnostic impact; values of 0.1 and less than 0.2 have a moderate difference in test probabilities and moderate diagnostic impact; values of 0.2 and less than 0.5 have a small difference in test probabilities and sometimes an important diagnostic impact; and values of 0.5–1 have small difference in test probabilities and seldom have important diagnostic impact.

The role of the clinical guidelines is to increase the pretest probability by adequately distinguishing low risk from high-risk groups. The role of imaging guidelines is to increase the likelihood ratio by recommending the diagnostic test with the highest sensitivity and specificity. Comprehensive use of clinical and imaging guidelines will improve the posttest probability, hence increasing the diagnostic outcome [10].

How to Use This Book

As these examples illustrate, the EBI process can be lengthy [39]. The literature is overwhelming in scope and somewhat frustrating in methodological quality. The process of summarizing data can be challenging to the clinician not skilled in meta-analysis. The time demands on busy practitioners can limit their appropriate use of the EBI approach. This book can mitigate these challenges in the use of EBI and make the EBI accessible to all imagers and users of medical imaging.

This book is organized by major diseases and injuries. In the table of contents within each chapter, you will find a series of EBI issues provided as clinically relevant questions. Readers can quickly find the relevant clinical question and receive guidance as to the appropriate recommendation based on the literature. Where appropriate, these questions are further broken down by age, gender, or other clinically important circumstances. Following the chapter's table of contents is a summary of the key points determined from the critical literature review that forms the basis of EBI. Sections on pathophysiology, epidemiology, and cost are next, followed by the goals of imaging and the search methodology. The chapter is then broken down into the clinical issues. Discussion of each issue begins with a brief summary of the literature, including a quantification of the strength of the evidence, and then continues with detailed examination of the supporting evidence. At the end of the chapter, the reader will find the take-home tables and imaging case studies, which highlight key imaging recommendations and their supporting evidence. Finally, questions are included where further research is necessary to understand the role of imaging for each of the topics discussed.

Take-Home Appendix 1: Equations

Test result	Outcome	
	Present	Absent
Positive	a (TP)	b (FP)
Negative	c (FN)	d (TN)
(a) Sensitivity	a/(a + c)	
(b) Specificity	d/(b + d)	
(c) Prevalence	(a + c)/(a + b + c + d)	
(d) Accuracy	(a + d)/(a + b + c + d)	
(e) Positive predictive value[a]	a/(a + b)	
(f) Negative predictive value[a]	d/(c + d)	

(*continued*)

(g) 95 % confidence interval (CI)	$p \pm 1.96\sqrt{\frac{1-n}{n}}$
	p = proportion
	n = number of subjects
(h) Likelihood ratio	$\dfrac{\text{Senstivity}}{1 - \text{Senstivity}} = \dfrac{a(b+d)}{b(a+c)}$

Reprinted with the kind permission of Springer Science+ Business Media from Medina LS, Blackmore CC, Applegate KE. Evidence-based imaging: improving the quality of imaging in patient care. Revised Edition. New York: Springer Science+Business Media; 2011.
[a]Only correct if the prevalence of the outcome is estimated from a random sample or based on an a priori estimate of prevalence in the general population; otherwise, use of Bayes' theorem must be used to calculate positive predictive value (*PPV*) and negative predictive value (*NPV*). *TP* true positive, *FP* false positive, *FN* false negative, *TN* true negative

Take-Home Appendix 2: Summary of Bayes' Theorem

1. Information before test × Information from test = Information after test
2. Pretest probability (prevalence) sensitivity/ 1 − specificity = posttest probability (predictive value)
3. Information from the test also known as the likelihood ratio, described by the equation: sensitivity/1 − specificity
4. Examples 1 and 2 predictive values: The predictive values (posttest probability) change according to the differences in prevalence (pretest probability), although the diagnostic performance of the test (i.e., sensitivity and specificity) is unchanged. The following examples illustrate how the prevalence (pretest probability) can affect the predictive values (posttest probability) having the same information in two different study groups.

Equations for calculating the results in the previous examples are listed in Appendix 1. As the prevalence of carotid artery disease increases from 0.16 (low) to 0.82 (high), the positive predictive value (PPV) of a positive contrast-enhanced CT increases from 0.67 to 0.98, respectively. The sensitivity and specificity remain unchanged at 0.83 and 0.92, respectively. These examples also illustrate that the diagnostic performance of the test (i.e., sensitivity and specificity) does not depend on the prevalence (pretest probability) of the disease. CTA, CT angiogram.

Example 1: Low prevalence of carotid artery disease

	Disease (carotid artery disease)	No disease (no carotid artery disease)	Total
Test positive (positive CTA)	20	10	30
Test negative (negative CTA)	4	120	124
Total	24	130	154

Example 2: High prevalence of carotid artery disease

	Disease (carotid artery disease)	No disease (no carotid artery disease)	Total
Test positive (positive CTA)	500	10	510
Test negative (negative CTA)	100	120	220
Total	600	130	730

Results: sensitivity = 500/600 = 0.83; specificity = 120/130 = 0.92; prevalence = 600/730 = 0.82; positive predictive value = 0.98; negative predictive value = 0.55
Reprinted with the kind permission of Springer Science+ Business Media from Medina LS, Blackmore CC, Applegate KE. Evidence-based imaging: improving the quality of imaging in patient care. Revised Edition. New York: Springer Science+Business Media, 2011

References

1. Levin A. Ann Intern Med. 1998;128:334–6.
2. Evidence-Based Medicine Working Group. J Am Med Assoc. 1992;268:2420–5.
3. The Evidence-Based Radiology Working Group. Radiology. 2001;220:566–75.
4. Wood BP. Radiology. 1999;213:635–7.
5. Poisal JA, et al. Health Aff. 2007;26:w242–53.
6. Davis K. N Engl J Med. 2008;359:1751–5.
7. Hulley SB, Cummings SR. Designing clinical research. Baltimore: Williams & Wilkins; 1998.
8. Kelsey J, Whittemore A, Evans A, Thompson W. Methods in observational epidemiology. New York: Oxford University Press; 1996.
9. Blackmore C, Cummings P. AJR Am J Roentgenol. 2004;183(5):1203–8.

10. Medina L, Aguirre E, Zurakowski D. Neuroimaging Clin N Am. 2003;13:157–65.
11. Medina L. AJNR Am J Neuroradiol. 1999;20:1584–96.
12. Sunshine JH, McNeil BJ. Radiology. 1997;205:549–57.
13. Black WC. AJR Am J Roentgenol. 1990;154:17–22.
14. Sox HC, Blatt MA, Higgins MC, Marton KI. Medical decision making. Boston: Butterworth; 1988.
15. Metz CE. Semin Nucl Med. 1978;8:283–98.
16. Singer M, Applegate K. Radiology. 2001;219:611–20.
17. Weinstein MC, Fineberg HV. Clinical decision analysis. Philadelphia: WB Saunders; 1980.
18. Carlos R. Acad Radiol. 2004;11:141–8.
19. Detsky AS, Naglie IG. Ann Intern Med. 1990;113:147–54.
20. Doubilet P, Weinstein MC, McNeil BJ. N Engl J Med. 1986;314:253–6.
21. Gold MR, Siegel JE, Russell LB, Weinstein MC. Cost-effectiveness in health and medicine. New York: Oxford University Press; 1996.
22. Hillemann D, Lucas B, Mohiuddin S, Holmberg M. Ann Pharmacother. 1997;31:974–9.
23. Medina L, Aguirre E, Bernal B, Altman N. Radiology. 2004;230:49–54.
24. Appel LJ, Steinberg EP, Powe NR, Anderson GF, Dwyer SA, Faden RR. Med Care. 1990;28:324–37.
25. Evens RG. Cancer. 1991;67:1245–52.
26. Yin D, Forman HP, Langlotz CP. AJR Am J Roentgenol. 1995;165:1323–8.
27. Medina L, Crone K, Kuntz K. Pediatrics. 2001;108:E101.
28. Ware JE, Sherbourne CD. Med Care. 1992;30:473–83.
29. Jarvik J, Hollingworth W, Martin B, et al. J Am Med Assoc. 2003;2810–2818.
30. Blackmore CC, Magid DJ. Radiology. 1997;203:87–91.
31. Medina L, Aguirre E, Byron B, Altman N. Radiology. 2004;230:49–54.
32. Zou K, Fielding J, Ondategui-Parra S. Acad Radiol. 2004;11:127–33.
33. Langlotz C, Sonnad S. Acad Radiol. 1998;5(suppl 2):S269–73.
34. Littenberg B, Moses LE. Med Decis Making. 1993;13:313–21.
35. Fiebach JB, et al. Stroke. 2004;35(2):502–6.
36. Fryback DG, Thornbury JR. Med Decis Making. 1991;11:88–94.
37. Blackmore C. Radiology. 2005;235(2):371–4.
38. Jaeschke R, Guyatt GH, Sackett DL. JAMA. 1994;271:703–7.
39. Malone D. Radiology. 2007;242(1):12–4.