Clinical Epidemiology

Holger J. Schünemann, Gordon H. Guyatt

8.1 Introduction ... 1170

Brief History of Clinical Epidemiology ... 1170
A Definition of Clinical Epidemiology ... 1170
Clinical Epidemiology and Evidence-based Medicine ... 1171

8.2 Case Scenario ... 1173
8.3 Formulating a Clinical Question ... 1174
8.4 Diagnosis ... 1175

Establish the Framework for Bayesian Thinking for Diagnosis ... 1176
Choosing the Right Test ... 1176
Likelihood Ratios ... 1178
How to Obtain Pre-Test Probabilities ... 1182
Sensitivity and Specificity ... 1182

8.5 Therapy/Prevention ... 1184
8.6 Systematic Reviews ... 1193
8.7 Guidelines ... 1196
8.8 Health Related Quality of Life Instruments and Their Application in Clinical Studies ... 1202

Generic Instruments ... 1202
Specific Instruments ... 1203

8.9 Integrating Patient Preferences in the Decision Making Process and Resolution of the Clinical Scenario ... 1205
8.10 Conclusions ... 1214
References ... 1214

8.1 Introduction

This chapter will begin with providing a brief overview of the history of clinical epidemiology and describe its relation with evidence based medicine. Clinical epidemiology differs from classical epidemiology in that clinical epidemiology supports other basic medical sciences such as biochemistry, anatomy and physiology because it facilitates their application in research through formulation of sound clinical research methods and, thus, puts these disciplines into clinical context. Therefore, clinical epidemiology goes beyond clinical trials. We will describe this concept in the following paragraphs (see Sect. 8.1.1 through 8.1.3). The following sections include case scenarios that facilitate the introduction of the key concepts about developing clinical questions, using diagnostic tests, evaluating therapy, appraising systematic reviews, developing guidelines and making clinical decisions.

8.1.1 Brief History of Clinical Epidemiology

Sackett provides an astute historical summary of the development of clinical epidemiology in his recent tribute in memory of Alvan Feinstein (Sackett 2002). Sackett's account gives John Paul credit for introducing the term clinical epidemiology describing it as the "new basic science for preventive medicine" (Paul 1938; Sackett 2002). Over the past 40 years, both Feinstein and Sackett himself made major contributions to the field of clinical epidemiology. Sackett founded, in 1966, the first clinical epidemiology research unit at the University at Buffalo, New York, USA, and in 1967 a department of clinical epidemiology and biostatistics at McMaster University in Hamilton, Canada, which he served as chair. The latter institution has trained numerous clinical epidemiologists, some of whom have taken on chair positions themselves. Today there are departments and units of clinical epidemiology throughout the world, though the development in some jurisdictions has been slower than in others. For example, it was in this millennium that the first department of clinical epidemiology was founded in the German speaking countries of Europe (Basel, Switzerland, H. Bucher, personal communication), although professorships of clinical epidemiology existed in these countries for some time.

8.1.2 A Definition of Clinical Epidemiology

Although seminal, Paul's simple description of clinical epidemiology was perhaps not sufficient in helping investigators and clinicians understand the principles underlying the term clinical epidemiology. Articles and textbooks have provided further definitions. Feinstein portrayed clinical epidemiology as investigating "the occurrence rates and geographic distribution of disease; the pattern of natural and post-therapeutic events that constitute varying clinical courses in the diverse spectrum of disease; and the clinical appraisal of therapy. The contemplation and investigation of these or allied topics constitute a medical domain that can be called clinical epidemiology" (Feinstein 1968; Sackett et al. 1991). Sackett defined clinical epidemiology as "the application, by a physician who provides direct patient care, of epidemiologic and biostatistical methods to the study of diagnostic and therapeutic processes in order to effect an improvement of health" (Sackett 1969, 2002; Sackett and Winkelstein 1967). Fletcher et al. (1996) described clinical epidemiology as the science of making predictions about individual patients by counting clinical events in similar patients, using strong scientific methods for studies of groups of patients to ensure that the predictions are accurate. Weiss (1996) defined clinical epidemiology as the study of variation in the outcome of illness and of the reasons for that variation. Despite the numerous definitions, one might argue that by providing the subheading "a basic science for Clinical Medicine" to their textbook "Clinical Epidemiology" Sackett and colleagues provided a pithy definition that not only turned the wheel back to John Paul, but widened it to all areas of clinical medicine by replacing *preventive medicine* with *clinical medicine* (Sackett et al. 2000).

Definitions are inevitably limited, and in depth understanding requires a more comprehensive discussion. We characterize clinical epidemiology by focusing on its purpose: to ensure that clinicians' practice and decision making is evidence-based. Clinical decision making requires answering questions about diagnosis, therapy, prevention and harm, providing estimates of prognosis and obtaining unbiased and precise estimates of intervention effects. Clinical epidemiology supports other basic medical sciences such as biochemistry, anatomy and physiology because it facilitates their application in research through formulation of sound clinical research methods and, thus, puts these disciplines into clinical context. Thus, clinical epidemiology provides the integrative force of medical science and medical practice.

Clinical Epidemiology and Evidence-based Medicine 8.1.3

When working optimally, clinical epidemiologists communicate results of investigations in ways that clinicians can readily apply in practice. Clinical epidemiology provides the evidence for management decisions resulting in more good than harm. Clinicians should use best evidence for clinical decision making. Thus, clinical epidemiology and evidence-based practice are closely linked. Clinical epidemiology grounds health care research in the mission to deliver optimal care to individual patients. As it turns out, clinicians optimally applying the evidence to their patient care must understand the basic concepts of clinical epidemiology (Guyatt et al. 2000). At the same time, while clinical epidemiology grounds the clinical investigators' viewpoint, evidence-based medicine (EBM) provides the framework for application of research findings in clinical practice. In the next sections of this discussion, we will describe the basics of clinical epidemiology methods and how insights from clinical practice may enlighten clinical epidemiologists.

The History and Philosophy of Evidence-based Medicine

When we first introduced the term evidence-based medicine (EBM) in an informal residency training program document, we described it as "an attitude of enlightened skepticism toward the application of diagnostic, therapeutic, and prognostic technologies in their day-to-day management of patients" (Guyatt 1991, 2002a, b). Through a series of articles published by the evidence-based medicine working group the term as well as the philosophy of EBM became well-known (Evidence-Based-Medicine-Working-Group 1992; Oxman et al. 1993). A Medline search revealed 7 citations including the term "Evidence Based Medicine" in 1993 and 2169 citations in 2002.

EBM evolved out of the efforts of clinicians with methodology training – that is, clinical epidemiologists – to apply their particular insights and approaches to solving clinical problems. In contrast to the traditional paradigm of clinical practice, EBM acknowledges that intuition, unsystematic clinical experience, and pathophysiologic rationale are not sufficient for making the best clinical decisions. Although it acknowledges the importance of clinical experience, EBM postulates that optimal clinical decision-making requires the integration of evidence from clinical research.

EBM places a lower value on authority than the traditional medical paradigm, and explicitly includes patients' and society's values in the clinical decision-making process. Patients or their proxies must always trade the benefits, harm, and costs associated with alternative treatment strategies, and in doing so must consider values and preferences.

To achieve the integration of research results in clinical practice, EBM proposes a formal set of rules to help clinicians interpret and apply evidence. Clinical epidemiologists have, by and large, developed these rules. These rules are characterized by a hierarchy of evidence (Fig. 8.1): confidence in research results is greatest if systematic error (bias) is lowest and increases if bias is more likely to play a role.

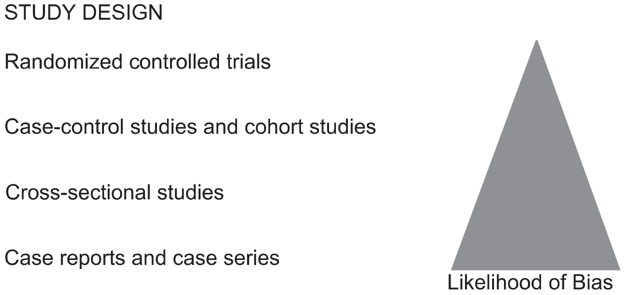
STUDY DESIGN
Randomized controlled trials
Case-control studies and cohort studies
Cross-sectional studies
Case reports and case series
Likelihood of Bias

Figure 8.1. Depicts the hierarchy of quality of evidence. As the research design becomes more rigorous (moving from *bottom* to *top*) the quality of evidence increases and the likelihood of bias decreases

Although randomized and controlled study designs provide the highest quality of evidence, EBM is not a science of randomized controlled trials. Rather, because higher quality evidence is often not available, EBM acknowledges that a large body of highly relevant evidence comes from observational studies. That is, to answer clinical questions clinicians will often depend on observational studies in their evidence-based practice. Therefore, the practice and application of EBM requires an understanding and critical evaluation of all study designs. Clinical epidemiology provides the necessary toolbox for this evaluation. For example, clinical epidemiologists of the Cochrane Collaboration, an international organization dedicated to making up-to-date and accurate information about the effects of healthcare readily available worldwide, have provided important insights into the conduct of systematic reviews and meta-analysis that inform clinicians and patients choices. It produces and disseminates systematic reviews of healthcare interventions and promotes the search for evidence. It can be accessed at www.cochrane.org.

Case Scenario

Example 1. Imagine you are the attending physician on ward rounds with your team. The senior resident presents the case of a 64 year old woman who came to the emergency room early one morning with left sided chest pain lasting for 15 minutes. The pain was severe enough to awaken her. She also had to sit up in her bed because of difficulties with getting her breath. Finally, her symptoms became so severe that she called an ambulance.

You immediately think that this woman has had an acute myocardial infarction. However, other diagnoses such as pulmonary embolism, pneumonia, pericarditis, asthma and a severe case of gastro-esophageal reflux disease also come to mind.

The resident continues with her presentation and tells you that the pain radiated to her left arm and you further hear that she had another episode of similar pain in the ambulance which was relieved within a few minutes by 0.4 mg nitroglycerin given under her tongue.

You feel that this information confirms your early intuition and that it makes a diagnosis of myocardial infarction more likely.

An EKG in the emergency room was unremarkable and cardiac enzymes drawn at arrival to the emergency room were borderline elevated (troponin I, a marker for myocardial injury, was 1.0 g/ml). Her chest X-ray was normal.

You now think that the diagnosis might be one of acute coronary syndrome, perhaps unstable angina or a myocardial infarction without EKG changes, and you continue to entertain pulmonary embolism, pneumonia and pericarditis as alternative – although less likely – diagnoses. The key decision you face is whether to admit the patient to hospital, possibly to a cardiac care unit, or to send her home with provision for subsequent investigation, perhaps an exercise test.

The patient's past medical history includes diabetes mellitus type 2. Her lipid profile is within the limits set by the National Cholesterol Education Program

Guidelines (NCEP 2001) and there is no significant family history of cardiovascular disease. She takes an oral hypoglycemic agent and 325 mg of aspirin daily as recommended by her physician. She has had similar chest pain over the past year when vacuuming her home. However, she did not mention these complaints to her physician because the discomfort always resolved after a few minutes of rest. The patient has no history of cough, wheezing, indigestion, heartburn or changes in her bowel habits. Physical examination shows an anxious patient, but there are no abnormal findings on physical examination.

Your team concludes that the probability that the presentation represents acute coronary syndrome is at least 50%. While you are discussing the patient and her further management, a second set of laboratory results shows that the troponin I is elevated at 4.1 g/l.

At this point you feel that a myocardial infarction without ST segment elevation on the EKG (a NSTEMI) is the most likely diagnosis and together with your team you consider further management. ♦

8.3 Formulating a Clinical Question

Using research evidence to guide clinical practice requires formulating sensible clinical questions (McKibbon et al. 2002; Oxman et al. 1993; Richardson et al. 1995). For most questions the key components are the patients, the intervention or exposure, comparison interventions (or exposure) and the outcomes (Table 8.1).

Table 8.1. Formulating the clinical question

Component	Explanation
Population	Who are the relevant patients?
Interventions or exposures	What are the management strategies clinicians are interested in? For example: Diagnostic test, drugs, toxins, nutrients, surgical procedures, etc.
Comparison (or control) intervention or exposures	What is the comparison, control or alternative intervention clinicians are interested in? For questions about therapy or harm there will always be a comparison or control (including doing nothing, placebo, alternative active treatment or routine care). For questions about diagnosis there may be a comparison diagnostic strategy (for example troponin I compared to creatine kinase MB in the diagnosis of myocardial infarction).
Outcome	What are the patient-important consequences of the exposure clinicians are interested in?

The clinical scenario of an older diabetic women presenting with chest pain potentially generates several clinical questions (about her diagnosis, appropriate therapy, prevention of future events, prognosis). We will use some of these questions to demonstrate how clinical epidemiology helps solve clinical problems.

Diagnosis

Based on the framework for developing a clinical question, we will start with a question about diagnosis:

Population:	In women with chest pain typical for angina pectoris
Intervention/exposure:	What is the test performance of troponin I serum levels
Outcome:	To predict myocardial infarction and associated adverse outcomes (congestive heart failure, death, serious arrhythmia or severe ischemic pain) in the next 72 hours.

The process of diagnosis is a complex cognitive task. There are different approaches to making a diagnosis, but pattern recognition, which is also known as the gestalt method, and logical reasoning play an important role (Glass 1996; Sackett et al. 1991; Sox et al. 1988). Clinicians always look for clues that help them establish a diagnosis, although with increasing clinical experience this process becomes increasingly subconscious. Some clues make a diagnosis more likely, other clues or the absence of certain clues make a diagnosis less likely (Ladenheim et al. 1987). In the scenario described at the beginning of this chapter, the first clue was the presence of left sided chest pain awaking the patient at night. This clue suggested that the patient might suffer from a cardiac problem. The presence of shortness of breath strengthened this suspicion, but brought other possible diagnosis into consideration (asthma, pneumonia and pulmonary embolus).

When clinicians use clues offered by clinical history, symptoms, signs or test results, they routinely, if often subconsciously, apply probabilities associated with these clues. For example, the presence of chest pain makes a heart attack more likely than no chest pain. Thus, a first step in making a diagnosis is to assign probabilities to the contemplated diagnoses. Clinicians then group the findings into coherent clusters, such as left sided (location) chest (heart or lungs) pain (symptom). These clusters inform the differential diagnoses. The differential diagnoses in the case scenario included acute myocardial infarction, pulmonary embolism, pneumonia, asthma or gastro-esophageal reflux disease. In the next step of making a diagnosis, the clinician incorporates new information, which lowers or increases the relative likelihood of the differential diagnoses. The process is therefore sequential. The presence of pain radiating to the left arm increased the probability of coronary heart disease and the absence of cough and gastrointestinal symptoms lowered the likelihood of pneumonia and gastrointestinal disease.

8.4.1 Establish the Framework for Bayesian Thinking for Diagnosis

As described above, the process of diagnosis can take place on a subconscious level where the clinician guesses the associated probabilities and relative likelihoods or it can employ explicit probabilities and relative likelihoods. For some clinical problems, such as diagnosing pulmonary embolism, clinicians intuition is good. The prospective investigation of pulmonary embolism diagnosis (PIOPED) study has shown this (PIOPED-Investigators 1990). Even when intuition is reasonably accurate, use of exact numbers generated by empirical studies can improve clinical decisions (Diamond and Forrester 1979; Diamond et al. 1980, 1981; Dolan et al. 1986). The latter approach of using explicit information is based on epidemiologic and biostatistical concepts, but both the intuitive and the explicit approaches are founded in Bayesian theory, because both approaches depend on probabilites that are altered by subsequent information (Bernardo and Adrian 1994; Berry 1996; Diamond 1999; Ledley and Lusted 1959; cf. Chap. I.1 of this handbook). Using the Bayesian approach in the diagnostic process the clinician starts with a certain probability (often called the pre-test probability) of a disease being present. Then, based on clues from the history, physical exam or test results, the clinician modifies this probability into another probability (often called the posttest probability).

8.4.2 Choosing the Right Test

The best test would be one that excludes or confirms a diagnosis beyond doubt. Using an ideal test, no patient would have the disease if the test is negative and all patients with a positive test would have the disease. For our example, were troponin I perfect, one could assume that a troponin I level $\geq$ 2 g/l proves beyond doubt that the patient has an acute myocardial infarction or will suffer a serious clinical event associated with acute coronary syndrome in the next 72 hours, and a level $<$ 2 g/l establishes that the patient does not have an acute myocardial infarction and will not suffer a serious event. Unfortunately, most information that clinicians obtain in clinical practice comes with uncertainty, and tests that definitively distinguish between disease and no disease are few and far between.

The typical cut-off value for troponin I in clinical practice is 2.0 g/l (Meier et al. 2002). However, astute clinicians would not dismiss a diagnosis of myocardial infarction in our scenario after the first troponin I level was $<$ 2.0 g/l, because both EKG and biomarkers may be what is typically defined as normal even when disease is present. Furthermore, they are aware that the troponin I may be normal early, and may rise subsequently. What clinicians expect of a good test is that results change the probability sufficiently to confirm or exclude a diagnosis. If a test result moves the probability below a threshold at which the disease is very unlikely and downsides associated with the treatment outweigh any anticipated benefit, then no further testing and no treatment are indicated. We call this probability the test threshold. If, on the other hand, the test result moves the probability of disease above a threshold at which one would not further test because disease is highly probable and one would start treatment we have found the treatment threshold. This scenario shows how a clinician can estimate the probability of disease and then compare disease probability to these two thresholds (Fig. 8.2).

In a clinical context in which the pre-test probability of a particular diagnosis is above the treatment threshold, further confirmatory testing that raises the probability further would not be helpful. On the other end of the scale, for a disease with a pre-test probability below the test threshold, further exclusionary testing lowering the probability would not be useful. When the probability is between the test and treatment thresholds testing will be diagnostically useful. Test results are of greatest value when they shift the probability across either threshold.

What determines our treatment thresholds? If adverse effects of treatment are frequent and severe, clinicians choose a higher treatment threshold. For example, because a diagnosis of pulmonary embolism involves long-term anticoagulation with appreciable bleeding risk, clinicians are very concerned about falsely labeling patients. The invasiveness of the next test will also impact on the threshold. If results from the next test (such as a ventilation-perfusion scan) are benign, clinicians are ready to choose a high treatment threshold. Clinicians are more reluctant to institute an invasive test associated with risks to the patient, such as a pulmonary angiogram, and this will drive their treatment threshold downward. That is, clinicians are more inclined to accept a risk of a false-positive diagnosis because a higher treatment threshold necessitates putting more patients through the risky test.

Accordingly, the more serious a missed diagnosis, the lower we will set our test threshold. Since a missed diagnosis of a pulmonary embolus could be fatal, clinicians are inclined to set their diagnostic threshold low. At the same time, the risks associated with the next test we are considering have an influence on where

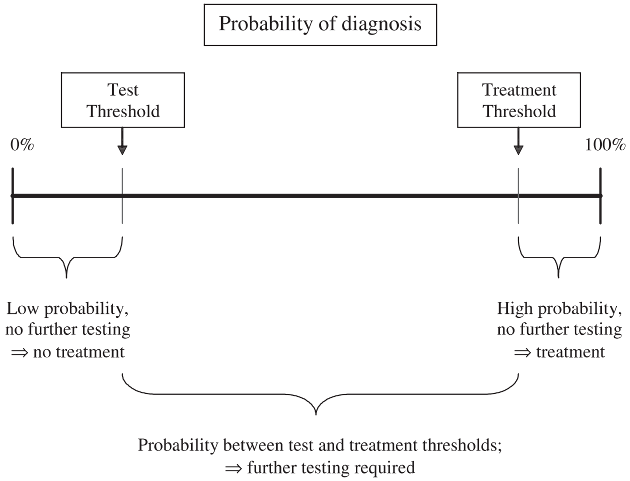
Probability of diagnosis
Test Threshold
Treatment Threshold
0%
100%
Low probability, no further testing ⇒ no treatment
High probability, no further testing ⇒ treatment
Probability between test and treatment thresholds; ⇒ further testing required

Figure 8.2. Test and treatment thresholds

to set the test threshold. If the risks are low, clinicians will be comfortable with a very low diagnostic threshold. The higher the risks, the more the threshold rises.

8.4.3 Likelihood Ratios

The center of any diagnostic process is a patient presenting with a constellation of symptoms and signs. Consider two patients with chest pain and shortness of breath in whom the clinician suspects a myocardial infarction without findings suggestive of pneumonia, airflow obstruction, pulmonary embolism or heart failure or other conditions. One patient is the 64-year-old woman described in the clinical scenario and the other is a 24-year-old man with a history of anxiety disorder. Clinicians would agree that the probability of myocardial infarction for these two patients – that is, their pre-test probabilities – are very different. In the woman described in the scenario, the probability is high; in the young man, it is low. Consequently, even if both patients had borderline elevated troponin I levels of 1.0 g/l at presentation, management is likely to differ between the two. An informed clinician might well treat the elderly woman immediately with aspirin and heparin but order further investigations in the young man.

One can draw two conclusions from these considerations. First, regardless of the results of the troponin I test, they do not definitively establish whether myocardial infarction is in fact the underlying disease, or whether the patient will suffer a serious event associated with an acute coronary syndrome. What they do accomplish is to alter the pre-test probability of the condition, yielding a new post-test probability. The direction and magnitude of this change from pre-test to post-test probability are determined by the test's properties. The test property of greatest value is the likelihood ratio.

Hill and colleagues (2003) investigated the diagnostic properties of troponin I as an early marker of acute myorcardial infarction or acute coronary syndrome with serious sequellae in the next 72 hours in patients who did not have definitively diagnostic EKG changes. The investigators found 20 individuals with a serious cardiac outcome by the reference standard and 332 individuals who did not (Table 8.2). For all patients, troponin I tests were classified into four levels: < 0.5 g/l, 0.5 to < 2.0 g/l, 2.0 to < 10.0 g/l and ≥ 10.0 g/l). Several questions arise.

How likely is a substantially elevated (≥ 10 g/l) troponin I among people who suffered adverse outcomes? Table 8.2 illustrates that 3 of 20 (or approximately 15%) of people with adverse outcomes had troponin I levels ≥ 10 g/l. How often is the same test result, a positive troponin I, found among patients in whom high risk acute coronary syndrome was suspected but ruled out? The answer is 4 out of 332 (or approximately 1.2%). The ratio of these two likelihoods is the likelihood ratio (LR); for a highly elevated troponin I test, it equals 0.15/0.012 (or 12.5). In other words, a highly elevated troponin I is 12.5 times as likely to occur in a patient with – as opposed to without – an ultimate adverse outcome.

In a similar fashion, one can calculate the likelihood ratios for troponin I values of ≤ 0.5 g/l, 0.5 to ≤ 2.0 g/l and 2.0 to ≤ 10.0 g/l. This calculation involves answering

Table 8.2. Test properties of early troponin I testing in myocardial infarction or ischemia-associated adverse outcomes (CI = confidence interval)

	Myocardial infarction or other adverse outcomes				
	Present/proportion		Absent/proportion		Likelihood ratio (95% CI)
Test results					
≥ 10.0 g/l	3	3/20 = 0.15	4	4/332 = 0.012	12.5 (3.0, 51.9)
2.0– < 10.0 g/l	2	2/20 = 0.10	5	5/332 = 0.015	6.6 (1.4, 32.1)
0.5– < 2.0 g/l	3	3/20 = 0.15	20	20/332 = 0.06	2.5 (0.8, 7.7)
< 0.5 g/l	12	12/20 = 0.60	303	303/332 = 0.910	0.7 (0.5, 0.9)
Total	20		332		

two questions: First, how likely is it to obtain a given test result (e.g., a troponin I < 0.5 g/l) among people with the target disorder (myocardial infarction)? Second, how likely is it to obtain the same test result (again, a troponin I < 0.5 g/l) among people without the target disorder? For this troponin I test result, the likelihoods are 12/20 (0.60) and 303/332 (0.91), respectively, and their ratio (the likelihood ratio) is 0.7.

Thus, the likelihood ratios indicate by how much a given diagnostic test result will raise or lower the pre-test probability of the target disorder. A likelihood ratio of 1 indicates that the post-test probability is identical to the pre-test probability. Likelihood ratios above 1.0 increase the probability that the target disorder is present, and the higher the likelihood ratio, the greater is this increase. Likelihood ratios below 1.0 decrease the probability of the target disorder, and the smaller the likelihood ratio, the greater the decrease in probability and the smaller its final value.

Users of likelihood ratios often ask "What are good likelihood ratios for a test?". The answer is that day-to-day clinical practice lets clinicians gain understanding and their own sense of interpretation, but one can consider the following as a guide:

- Likelihood ratios of > 10 or < 0.1 generate large and often conclusive changes from pre- to post-test probability
- Likelihood ratios of 5–10 and 0.1–0.2 generate moderate shifts in pre- to post-test probability
- Likelihood ratios of 2–5 and 0.5–0.2 generate small (but sometimes important) changes in probability
- Likelihood ratios of 1–2 and 0.5–1 alter probability to a small (and rarely important) degree.

How can clinicians use likelihood ratios to move from pre-test to post-test probability? Unfortunately, one cannot combine likelihoods directly, such as one can combine probabilities or percentages. Their formal use requires converting pre-test probability to odds, multiplying the result by the likelihood ratio, and then converting the post-test odds into a post-test probability. Although this calculation is relatively straightforward for an experienced user, it can be time consuming and, fortunately, there is an easier way.

Figure 8.3 shows a nomogram proposed by Fagan that performs the conversions and allows simple transition from pre- to post-test probability (Fagan 1975). The line on the left-hand side represents the pre-test probability, the middle line represents the likelihood ratio, and the line on the right-hand side depicts the resulting post-test probability. One can obtain the post-test probability by anchoring a straight line at the pre-test probability and rotating it until it lines up with the likelihood ratio of the relevant test result.

If we assumed a pre-test probability of 50% (see below) for the elderly woman with multiple risk factors and we applied the LR associated with a troponin I of 1.0 to the nomogram (connecting 0.5 or 50% on the left with a LR of approximately 2.5 on the middle line and extending it through the right line) we would obtain a post-test probability of approximately 70% (or 0.7). If we assumed a pre-test probability of 1 in 1000 or 0.1% for the young man and applied the same LR, the post-test probability remains very low at between 0.2 and 0.3%.

To further explain the application of LRs, let us assume the two patients had troponin levels of 0.3 g/l. Applying the associated LR (0.7) to the pre-test probability of 50% of the elderly women would result in a post-test probability of approximately

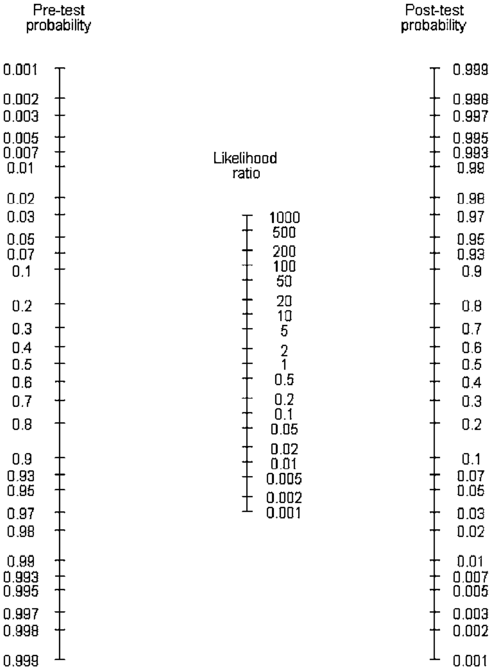

Figure 8.3. Likelihood ratio (Fagan) nomogram (Copyright 1975 Massachusetts Medical Society. All rights reserved. Reproduced with permission from the Massachusetts Medical Society)

45% (or 0.45). Applying this LR (0.7) to the pre-test probability of 0.1% of the young man results in a post-test probability of < 0.1%. It becomes evident from these latter two hypothetical examples that the test result has not altered the post-test probabilities to a large extent and further testing is necessary or that the clinician needs to make a decision on the basis of post-test probabilities that are similar to the pre-test probabilities. These strategies will differ between the two patients. Most clinicians would remain worried about the elderly women, but would safely discharge the young man.

Readers who are interested in the formula for converting pre-test probabilities to post-test probabilities, will note that it is based on Bayes theorem:

$$\text{Post-test odds} = \text{Pre-test odds} \times \text{Likelihood ratio}$$

Mathematically we can write this formula as:

$$O(D|R) = \left(O(D) \times P(R|D)\right)/P(R|\overline{D})\,,$$

where P is the probability of a specific test result R given the status of disease D = disease present, $\overline{D}$ = disease absent, $O(D)$ is the odds of disease to be calculated as $P(D)/[1 - P(D)]$.

Example 1. *(continued)*
Returning to our examples, for our elderly female patient with a test result of 1.0 µg/l and pre-test odds = 0.5/(1 − 0.5) = 1 this formula translates to:

$$\text{post-test odds of myocardial infarction} = 1 \times 0.15/0.06 = 2.5\,.$$

Post-test odds can be converted into post-test probabilities using the following formula:

$$\begin{aligned}\text{post-test probability} &= \text{post-test odds}/(\text{post-test odds} + 1) = 2.5/(2.5 + 1)\\ &= 71.4\%\,.\end{aligned}$$

♦

This estimate is similar to the estimate on the Fagan Nomogram. In fact, had we been able to use the Nomogram as accurately as the calculator we would have obtained identical numbers. This probability moves us into the range of probability where most clinicians would treat patients without further testing because of the morbidity and mortality associated with myocardial infarction.

Example 1. *(continued)*
For the young male with a troponin of 1.0 this formula translates to:

$$\text{post-test odds of myocardial infarction} = \text{pre-test odds} \times 2.5\,.$$

We can derive the pre-test odds from the pre-test probability of 0.1% = 0.001/(1− 0.001) which is about 0.001.

Thus, it follows that the post-test odds can be calculated as

$$\text{post-test odds} = 0.001 \times 2.5 = 0.0025$$

and the post-test probability as

$$\text{post-test probability} = 0.0025/(0.0025 + 1)$$

which is approximately 0.25%. ♦

Again, the estimate of 0.25% corresponds to the estimate using the Fagan Nomogram, but is the exact result of the application of Bayes theorem. Even a troponin I test associated with a LR greater than 1 and commonly considered elevated in the young man does not alter the probability for a myocardial infarction to a great extent.

8.4.4 How to Obtain Pre-Test Probabilities

We guessed at the pre-test probability of the two patients with chest discomfort. How do clinicians obtain valid pre-test probabilities? Intuitively, clinicians use their experience based on previous patients with similar presentations. However, the probabilities clinicians and in particular learners assume are prone to bias and error (Richardson 2002; Richardson et al. 2003).

Richardson has suggested that there are two different forms of clinical research that can guide clinicians estimates of pre-test probabilities (Richardson 2002). The first type are studies that yield disease probability based on representative patient cohorts with a defined clinical problem that carry out careful diagnostic evaluations and apply explicit diagnostic criteria. Examples include studies on causes of syncope (Soteriades et al. 2002) and cancer in involuntary weight loss (Hernandez et al. 2003). Control groups of randomized trials may serve for questions of prognosis. The study by Hill et al. (2003) that provided the estimates for the test properties of troponin I to obtain pre-test probabilities could also serve as an example. The second type of studies are clinical decision rules. These studies assemble cohorts suspected of having the target disorder, apply standard reference tests and report the frequency of diagnoses in subgroups with identifying clinical features (McGinn et al. 2003). High quality studies that provide valid estimates of frequency are applicable to our patients and can provide precise estimates of pre-test probability. For example, Richardson et al. (2003) estimated in a consecutive patient series that for 78% (95% CI: (66%, 96%)) of clinical problems evidence of pre-test probabilities existed in the literature.

8.4.5 Sensitivity and Specificity

Likelihood ratios help users understand diagnostic tests. However, clinicians use two other descriptive terms for diagnostic tests. It is, therefore, helpful for those with interest in clinical epidemiology to understand these two other terms: sensitivity and specificity.

We could have described the test properties of the study by Hill and colleagues using the concepts of sensitivity and specificity defining normal and abnormal test results. We presented the four different interpretations of troponin I levels, each with the associated likelihood ratios. That classification allowed us to omit the terms normal and abnormal or positive and negative. However, this is not the way most investigators present their result. Investigators also often rely on concepts of sensitivity and specificity.

Sensitivity expresses the proportion of people with the target disorder in whom the test result is positive and specificity expresses the proportion of people without the target disorder in whom a test result is negative. Table 8.3 shows the general concept of sensitivity and specificity in a 2 × 2 table. We could transform the 4 × 2 table described above (Table 8.2) into three 2 × 2 tables, depending on what we call positive or negative. Let us assume that only troponin I values $\geq$ 10.0 g/l are positive (or abnormal).

To calculate sensitivity from the data in Table 8.2 for positive troponin I levels, we look at the number of people with proven myocardial infarction ($n = 20$) who were diagnosed as having the target disorder on troponin testing ($n = 3$) showing a sensitivity of 3/20, or approximately 15% ($a/(a + c)$). To calculate specificity, we use the number of people without the target disorder (332) whose troponin test results were classified as normal or < 0.5 g/l (303), yielding a specificity of 303/332, or 91% ($d/(b + d)$).

As indicated above, one can easily calculate LRs for different levels of quantitative test results while sensitivity and specificity require a definition of normal and

Table 8.3. Sensitivity and specificity of diagnostic tests ($a + b + c + d = 100\%$)

	Disease or reference standard (proportion)	
	Present (positive)	Absent (negative)
Test results		
Disease present (positive)	True Positive (a)	False Positive (b)
Disease absent (negative)	False Negative (c)	True Negative (d)
Sensitivity =	True positive/ positive reference standard $a/(a + c)$	
Specificity =		True negative/negative reference standard $d/(b + d)$
Likelihood ratio for positive test (LR+) =	Sensitivity/(1 − Specificity) = $(a/(a + c))/(1 - d/(b + d))$	
Likelihood ratio for negative test (LR-) =	(1 − Sensitivity)/Specificity = $(1 - (a/(a + c))/(d/(b + d)))$	

abnormal that is often arbitrary. In using sensitivity and specificity one has to either discard important information or recalculate sensitivity and specificity for every cut-point. Therefore, the likelihood ratio is much simpler and much more efficient when tests have more than two possible results, which is very often the case (Guyatt et al. 1990, 1992; Guyatt and Rennie 2002).

8.5 Therapy/Prevention

Example 1. *(continued)*
Returning to the clinical scenario and having treated the elderly woman with aspirin and heparin, the team questions what other therapeutic or preventive interventions may be of benefit. The resident points you to a study that aimed at maximizing platelet inhibition in patients with acute coronary syndrome, the Clopidogrel in Unstable Angina to Prevent Recurrent Events Trial (CURE) trial (CURE-Investigators 2001). You vaguely remember this study and that it is the only one you are aware of addressing this question in patients with acute coronary syndrome. You know that neurologists and cardiologists in your institution often use two antiplatelet agents to maximize platelet inhibition, but you wonder about the quality of the evidence supporting this conclusion and the magnitude of benefits and downsides. An electronic database search confirms that there is no additional evidence addressing this specific question in a randomized trial in patients with acute coronary syndrome. The CURE investigators addressed the question "In patients with acute coronary syndromes without ST-segment elevation, does early and long-term use of clopidogrel plus aspirin versus aspirin alone prevent cardiovascular events and is the combination safe?" You find that this question would be highly relevant to your patient in whom you want to prevent further cardiovascular events.

You decide to critically appraise this study together with your team. The resident retrieves the article through your library's online full text journal subscription and you evaluate the article with your team over lunch break. ♦

The concepts of clinical epidemiology help clinicians appraise clinical research. We list the three commonly agreed on steps of critical appraisal for studies on therapy or prevention in Table 8.4 and describe how clinical epidemiology facilitates interpretation and evaluation of clinical research. This form addresses critical appraisal for most clinicians. Other important issues for experienced readers concern the appropriateness of the statistical methods.

The first factor that can influence confidence in research results is systematic error or bias. Bias is directly linked to the design and execution of a study. Therefore, the first step is an appraisal of whether results are valid and to what extend bias is present (Table 8.4 – Are the results valid?). The next step in the evaluation of research is the review of the results. Because clinicians often are unfamiliar with applying the magnitude of effects to patient care and because classical epidemiology applies research results in different contexts using different terminology, clinical epidemiologists can bring quantitative outcome measures closer to the clinician. This helps address the question "How large is the effect of an intervention and how large is the role of random error or chance (What are the results?)?". Finally, clinicians need to know whether the results are applicable to their patients. Therefore, clinicians need to appraise whether the results help with decision-making in individual practice circumstances (How can I apply the results to my patient care?). Clinicians must decide whether their patients are similar to those included in the studies from which they obtain the relevant evidence, but clinical epidemiologists can help them in the decision making process (see Sect. 8.8).

Critical appraisal (Table 8.4) of the CURE trial reveals that patients were randomized using a 24-hour computerized randomization service to conceal randomization. Control patients received placebos and investigators and outcome assessors were blinded to treatment assignment. While blinding refers to not being aware of treatment allocation when treatment has been assigned, concealment refers to avoiding biased allocation of patients because of prior knowledge of forthcoming treatment allocation. Investigators can achieve concealment through measures such as central (telephone) randomization and sealed envelopes. The more stringent the method for concealment the less likely are those allocating patients to tamper with this important aspect of randomized controlled trials. Fulfilling these

Table 8.4. Critical appraisal of studies about therapy

Question	Therapy or prevention
Study design and execution – evaluation of bias	I. Are the results of the study valid? 1. Were patients randomized? 2. Was randomization concealed? 3. Were patients analyzed in the groups to which they were randomized? 4. Were patients in the treatment and control groups similar with respect to known prognostic factors? 5. Were patients aware of group allocation? 6. Were clinicians aware of group allocation? 7. Were outcome assessors aware of group allocation? 8. Was follow up complete?
Results and random error	II. What are the results? 1. How large was the treatment effect? 2. How precise was the estimate of the treatment effect?
Application and uptake	III. How can I apply the results to my patient care? 1. Were the study patients similar to my patients? 2. Were all clinically important outcomes considered? 3. Are the likely treatment benefits worth the potential harms and costs?

validity criteria protects against bias, because systematic reviews suggest that lack of blinding and concealment (see Table 8.4) may lead to systematic overestimation of treatment effects although the effects may differ by clinical specialty only (Balk et al. 2002; Moher et al. 1998).

The CURE investigators achieved a follow-up of greater than 99.9% (only 13 out of 12,562 patients were lost) and analyzed patients in the group they were assigned according to the intention to treat principle. The intention to treat principle refers to analysis of patient outcomes based on which group they were randomized regardless of whether they actually received the planned intervention. This analysis preserves the power of randomization, thus maintaining that important unknown factors that influence outcome are likely equally distributed in each comparison group.

The evaluation of clinical research results requires an understanding of measures of association or effect. As noted above clinical epidemiologists often use terms that are different from those of classical epidemiologists (see relative risk reduction and number needed to treat below in this section). Table 8.5 summarizes the measures we will now describe in more detail (cf. Chapter I.2 of this handbook).

Absolute Risk

The easiest measure of risk to understand in clinical epidemiology is the absolute risk. In the CURE trial, the main outcomes were a composite of death from cardiovascular causes, nonfatal myocardial infarction (MI), or stroke and a composite of death from cardiovascular causes, nonfatal MI, stroke, or refractory ischemia. Safety outcomes included major and minor bleeding. We will focus on the latter composite endpoint. The absolute risk for this combined primary endpoint was 16.5%, and the absolute risk for this outcome in the control group was 18.8% (Table 8.5). As described above, other terminology for the risk of an adverse outcome in the control group are baseline risk, absolute risk, or control event rate.

Absolute Risk Reduction

One can express treatment effects as the difference between the absolute risks in the experimental and control groups, the absolute risk reduction or the risk difference. This effect measure represents the proportion of patients spared from the unfavorable outcome if they receive the experimental therapy (clopidogrel), rather than the control therapy (placebo). In our example, the absolute risk reduction is 18.8% − 16.5% = 2.3 percentage points.

Relative Risk

The relative risk or risk ratio presents the proportion of the baseline risk in the control group that still is present when patients receive the experimental treatment (clopidogrel). The relative risk of the combined outcome after receiving clopidogrel is $1035/(1035+5224)$ divided by $1187/(1187+5116)$ (the risk in the control group), or 0.87. One could also say the risk of experiencing the combined outcome with clopidogrel and aspirin is approximately 87% of that with aspirin alone.

Table 8.5. Measures of association and effect

General 2 × 2 Table

	Outcome		Absolute risk of outcome:
	+	–	
Intervention (Y)	a	b	Intervention = $a/(a+b) = Y$
Control (X)	c	d	Control = $c/(c+d) = X$

2 × 2 Table for the Example from CURE et al. (CURE-Investigators 2001)

	Combined primary outcome		Absolute risk of outcome
	+	–	
Clopidogrel	1035	5224	Clopidogrel = 1035/(1035 + 5224) = 16.5%
Placebo	1187	5116	Placebo = 1187/(1187 + 5116) = 18.8%

Absolute risk reduction (ARR)
Definition: The difference in risk between the control group and the intervention group
ARR = $c/(c+d) - a/(a+b) = X - Y$
Example:
ARR = 1035/(1035 + 5224) – 1187/(1187 + 5116) = 2.3%

Relative risk or risk ratio (RR)
Definition: The ratio of risk in the intervention (Y) to the risk in the control group (X)
RR = Y/X
Example:
RR = (1035/(1035 + 5224))/(1187/(1187 + 5116)) = 0.87

Relative Risk Reduction (RRR)
Definition: The percent reduction in risk in the intervention compared to the control group
RRR = 1 – RR = $(1 - X/Y) \times 100\%$ or
RRR = $[(X - Y)/X] \times 100\%$
Example:
RRR = (1 – 0.87) × 100% = 13%

Number Needed to Treat (NNT)
Definition: Inverse of the ARR
NNT = 1/ARR = $1/(X - Y)$
Example:
NNT = 1/2.3% = 44

Relative Risk Reduction

The most commonly reported measure of dichotomous treatment effects is the complement of this relative risk, the relative risk reduction. One can obtain the relative risk reduction easily from the relative risk because it is the proportion of baseline risk that is removed by the experimental therapy and it is equivalent to 1.0 − relative risk. It can be expressed as a percent: (1 − relative risk) × 100 = (1 − 0.87) × 100 = 13% or 100% − 87% = 13% for this example. Alternatively, one may obtain the relative risk reduction by dividing the absolute risk reduction by the absolute risk in the control group. Therefore, the result is the same if it is calculated from 2.3% (the absolute risk reduction) divided by 18.8% (the risk in the control group) = 0.13 (13%). A relative risk reduction of 13% means that clopidogrel reduced the risk of combined outcome by 13% relative to that occurring among control patients. The greater the relative risk reduction, the more efficacious is the therapy. Investigators may compute the relative risk over a period of time, as in a survival analysis, and call it a hazard ratio, the weighted relative risk over the entire study (see Chap. II.4 of this handbook).

In fact, the CURE investigators calculated the hazard ratio which was slightly more in favor of clopidogrel (0.86). For practical purposes we use the relative risk for the CURE trial in our example. If we had used the hazard ratio for the calculation of the RR the RRR would have been 14%.

Odds Ratio

Instead of evaluating the risk of an event, one can estimate the odds of having an event compared with not having an event. Most individuals are familiar with odds in the context of sporting events, when sport reporters describe the odds of a team or player winning a particular event. When odds are used in the medical sciences it stands for the proportion of patients with the target outcome divided by the proportion without the target outcome. The odds in the control group of the example trial described are 1187 of 6303 divided by 5116 of 6303. Because the denominator is the same in both the numerator and the denominator, it is canceled out, leaving the number of patients with the event (1187) divided by the number of patients without the event (5116). The odds are 1187/5116 or 0.232. To convert from odds to risk, one divides the odds by 1 plus the odds. As the odds of the combined endpoint are 0.232, the risk is 0.232/(1 + 0.232), or 0.188 (18.8%), identical to the baseline risk reported in the CURE trial. Table 8.6 presents the link between risk and odds. The greater the risk, the greater is the divergence between the risk and odds. Odds and risk are about equal if the absolute risk is small.

In the CURE trial, the odds of the combined endpoint in the clopidgrel group are 1035 (those with the outcome) compared with 5224 (those without the outcome), or 1035/5224 = 0.198, and the odds of the combined endpoint in the placebo group are 0.232. Therefore, the ratio of these odds is (1035/5224)/(1187/5116), or 0.854. If one used a terminology parallel to risk (note, that epidemiologists call a ratio of risks in most instances a relative risk), one would call the ratio of odds relative odds. The commonly used term, however, is odds ratio (OR). Until recently the

Table 8.6. Relation between risks and odds

Risk	Risk (proportion)	Odds
80%	0.80	4.0000
60%	0.60	1.5000
50%	0.50	1.0000
40%	0.40	0.6667
33%	0.33	0.5000
25%	0.25	0.3333
20%	0.20	0.2500
10%	0.10	0.1111
5%	0.05	0.0526
1%	0.01	0.0101

odds ratio has been the most popular measure of association. The reason for the use of the odds ratio is that the odds ratio has a statistical advantage because it essentially is independent of the choice between a comparison of the risks of an event (such as death) or the analogous non-event (such as survival), which is not true of the relative risk.

However, clinicians do not easily understand a ratio of odds. Clinicians would like to be able to substitute the relative risk, because it is more intuitively understandable, for the odds ratio. As shown in Table 8.6, as the risk decreases, the odds and risk come closer together. For low event rates, the odds ratio and relative risk are very close. In fact, if the risk is below 25%, odds and risks are approximately equal and many authors calculate relative odds and then report the results as if they calculated relative risks. One can see from the example that the odds ratio of 0.85 is very similar to the relative risk of 0.87. Clinicians should be aware that if events are frequent in either the control group or experimental group, odds ratios can be a very inaccurate estimate of the relative risk. The RR and OR also will be more similar when the treatment effect is small (OR and relative risk are close to 1.0) than when the treatment effect is large. When considering RR, HR and OR, the RR is always closest to unity; the odds ratio is farthest away; and the HR is intermediate (Symons and Moore 2002). These differences can become large when effect sizes increase. When event rates are high and effect sizes are large, there are ways of converting the odds ratio to relative risk. Fortunately, clinicians will need to do this infrequently. One note of caution is that typical case-control studies do not yield relative risks.

The Number Needed to Treat

Having seen that making a distinction between odds ratio and relative risk rarely will be important when evaluating clinical research, because high event rates are rare, one must give much more attention to distinguishing between the odds ratio or relative risk compared with the absolute risk reduction. Let us assume that the absolute risk of experiencing the combined outcome would be twice as high in both groups in the CURE trial. This could have happened if the investigators had conducted a study in patients at greater risk for the endpoint, for example by restricting the study population to older patients. In the clopidogrel group, the absolute risk would become 33% compared with 37.6% in the control group. Therefore, the absolute risk reduction would increase from 2.3% to 4.6% whereas the relative risk (and therefore the relative risk reduction) would remain identical at 33% divided by 37.6% = 0.87 (the relative risk reduction remains 13%). Therefore, the increase in the proportion of those experiencing the endpoint in both groups by a factor of 2 leaves the relative risk (and the relative risk reduction) unchanged, but increases the absolute risk reduction by a factor of 2.

A 13% reduction in the relative risk of the combined endpoint may not sound very impressive, however, its impact on patient groups and practice may be large. This notion is shown using the concept of the number needed to treat, the number of patients who must receive an intervention during a specific period to prevent one additional adverse outcome or produce one positive outcome (Laupacis et al. 1988). When discussing the number needed to treat, it is important to specify the treatment, its duration, and the outcome being prevented. The number needed to treat is the inverse of the absolute risk reduction, calculated as 1/absolute risk reduction. Therefore, in the example above with an absolute risk reduction of 4.6%, the number needed to treat would be 22 (1/4.6%) and it would be 44 (1/2.3%) in the CURE trial. Finally, imagine young patients with no additional risk factors for adverse outcomes. Such patients may carry a baseline risk of 4% for experiencing the endpoint and, therefore, the number needed to treat could increase to 192 [1/(0.13 × 4%)]. Given the duration, potential harms and cost of treatment with clopidogrel and the increased risk for bleeding, it could be reasonable to withhold therapy in that latter patient. Unfortunately, we do not know how best to present risk information to patients. Presenting the relative risk reduction alone is more persuasive for making actual or hypothetical medical decisions, because the actual benefit appears larger (Bucher et al. 1994; Edwards and Elwyn 1999; Edwards et al. 1999). Thus, direct to patient advertising and pharmaceutical industry detailing uses relative risk reduction as measure of effect to persuade clinicians and patients to use drug interventions. However, omitting the presentation of baseline risk or not informing those who receive this information that the baseline risk drives the absolute benefit is misleading.

How Clinicians Can Use Confidence Intervals

Until now we presented the results of the CURE trial as if they represented the true effect. The results of any experiment, however, represent only an estimate of the truth. The true effect of treatment actually may be somewhat smaller, or larger, than what researchers found. The confidence interval (CI) tells, within the bounds of plausibility, how much smaller or greater the true effect is likely to be. For each of the measures described, one can use statistical programs to calculate confidence intervals.

The point estimate within the confidence interval and the confidence interval itself help with two questions. First, what is the one value most likely to represent the true difference between treatment and control; and second, given the difference between treatment and control, what is the plausible range of differences within which the true effect might actually lie? The smaller the sample size or the number of events in an experiment, the wider the confidence interval. As the sample size gets very large and the number of events increases, investigators become increasingly certain that the truth is not far from the point estimate and, therefore, the confidence interval is narrower.

One can interpret the 95% confidence interval as "what is the range of values of probabilities within which, 95% of the time, the truth would lie?". If investigators or clinicians would not need to be so certain, one could ask about the range within which the true value would lie 90% of the time. This 90% confidence interval would be somewhat narrower.

How does the confidence interval facilitate interpretation of the results from the CURE trial? As described above, the confidence interval represents the range of values within the truth plausibly lies. Accordingly, one way to use confidence intervals is to look at the boundary of the interval that represents the lowest plausible treatment effect and decide whether the action or recommendation would change compared with when one assumes the point estimate represents the truth. Based on the numbers provided in the CURE trial, one can calculate a confidence interval around the point estimate of the relative risk reduction of 13% ranging from approximately 6 to 21%. Values progressively farther from 13% will be less and less likely. One can conclude that patients receiving clopidogrel are less likely to experience the combined endpoint – but the magnitude of the difference may be either quite small (and not outweigh the increased risk of bleeding) or quite large. This way of understanding the results avoids the yes/no dichotomy of testing a hypothesis. Because the lower limit of the CI is associated with a benefit, certainty about beneficial treatment effects from clopidogrel on the combined endpoint is relatively high. However, toxicity and expense will bear on the final treatment decision.

The chief toxicitiy of clopidogrel the authors of the CURE trial were concerned about was bleeding. They found that the absolute risk increase (ARI – conceptually similar to the absolute risk reduction but indicating an increase in risk from the investigational therapy) for major bleeding was 1 percentage point (an increase from 2.7% in the placebo group to 3.7% in the clopidogrel group). The investigators defined major bleeding as substantially disabling bleeding, intraocular bleeding or the loss of vision, or bleeding necessitating the transfusion of at least 2 units of blood. Similarly to the number needed to treat we can calculate the number of patients who must receive an intervention during a specific period to cause one additional harmful outcome. The number needed to harm (NNH) for major bleeding in the CURE trial is equal to 1/ARI or 1/0.01 = 100. The authors also provided the information for minor bleeding which they defined as hemorrhages that led to interruption of the study medication but did not qualify as major bleeding. The risk for minor bleeding was 5.1% in the clopidogrel group compared to 2.4% in the placebo group. Thus, the NNH for minor bleeding was 1/0.027 = 37. Clinicians should keep in mind that estimates of harm also come with uncertainty and, therefore, authors usually present confidence intervals around these harmful effects.

Use of Composite Endpoints

Investigators often use a composition of endpoints when they compare interventions. For example, the CURE trial investigators used a composite endpoint of death from cardiovascular causes, nonfatal MI, stroke, or refractory ischemia. Safety outcomes included major and minor bleeding with the definitions described in the foregoing paragraph. There are a number of reasons for combining several endpoints. First, investigators may believe that distinguishing between outcomes is unnecessary because the endpoints have the same consequences. For example, many investigators combine the outcome ischemic stroke with hemorrhagic stroke because they believe they may be similarly disabling (Lubsen and Kirwan 2002). Second, investigators may chose composite endpoints to avoid misleading conclusion when an intervention reduced an endpoint (usually less severe) by increasing another endpoint (usually more severe) that precludes patients from suffering the less severe endpoint. For example, surgery for cerebrovascular disease could reduce strokes by directly killing those at highest risk for stroke (those who die at surgery are no longer at risk for strokes) (van Walraven et al. 2002). Thus, the use of stroke alone as the endpoint would be misleading. To avoid an erroneous conclusion, an investigator facing this situation must combine the more and less serious outcomes, creating an endpoint such as "stroke or death". Third, investigators chose to combine endpoints because of the reduced sample size when event rates increase, as one would expect for a combined endpoint. The latter is the probably reason why the CURE investigators used a combined efficacy endpoint.

Using composite endpoints has a number of implications. The most obvious implication is that if these endpoints have different importance to patient (that is, patients have different underlying values and preferences for these outcomes) treating them as equally important is an oversimplification.

Using combined endpoints does not always support reaching conclusive results in clinical trials. Freemantle et al. (2003) reported on the use of composite endpoints in 167 trials published between 1997 and 2001 in 9 leading medical journals. The authors found that in 69 of these trials the difference between treatment arms in the CEP was not statistically significant. Investigators sometimes included component endpoints that reduced trial efficiency by diluting the treatment effect. For example, Freemantle and colleagues described that the CAPRICORN [CArvedilol Post-infaRct survIval COntRol in LV dysfunctioN] trial investigated the effects of carvedilol, a β-blocker, in 1959 patients with left ventricular dysfunction following myocardial infarction (The CAPRICORN Investigators 2001). Originally, the CAPRICORN investigators identified all-cause mortality as primary outcome in the trial protocol. However, while the study was ongoing, the data and safety monitoring board (whose assignment is to protect the patients in a trial) noted that the overall rate of mortality was lower than that predicted for the power analysis and sample size calculation. The board informed the CAPRICORN steering committee of the trial's insufficient power to identify the primary end point as significant (a preset level of significance α of 0.05). Taking the uncommon measure of altering the primary outcome, the steering committee defined a new composite outcome (all-cause mortality or cardiovascular hospital admissions). The steering committee assigned a critical level of significance of $\alpha = 0.045$ to this new composite outcome that they introduced while the trial was ongoing and reduced the significance level of the original primary outcome to achieve statistical significance to $\alpha = 0.005$. They reduced the level of α of the original primary outcome to penalize their retrospective action, but the board decided that if the p-value for either primary outcome achieved statistical significance at the new critical level, the study would be deemed positive.

At the end of the study, the original primary end point (all-cause mortality) achieved a p-value of 0.03 (ie, substantially larger than the new 0.005 allocated after consultation from the data safety and monitoring board, but smaller than the original critical level of significance), but the alternative primary outcome achieved a p-value of 0.30. Thus, the original primary outcome did not reach statistical significance at the new and more stringent level and neither did the new composite outcome. Although 12% of patients died in the carvedilol group, compared with 15% in the placebo group, 23% of patients in the carvedilol group and 22% of patients in the placebo group qualified for the composite outcome on the basis of hospitalizations alone, a result that undermined the relatively small reduction in mortality in the carvedilol group. Thus, CAPRICORN provides a neutral result, although the study would have been modestly statistically significant had the original primary outcome of all-cause mortality been maintained.

In summary, evaluating trials that use composite outcome requires scrutiny in regards to the underlying reasons for combining endpoints and its implications and has impact on medical decision making (see below in Sect. 8.8).

Systematic Reviews

8.6

The patient in our scenario took aspirin at a dose of 325 mg on admission. Aspirin use is associated with gastrointestinal bleeding and the risk for bleeding increases with the aspirin dose used (García Rodríguez et al. 2001; Weil et al. 1995). On the other hand, the beneficial effects of lower doses of aspirin on cardiovascular events likely does not differ from those of higher doses (Antithrombotic Trialists' Collaboration 2002). Taken together, studies comparing higher and lower doses of aspirin show similar effects. However, a clinician would ask the question "Which of the available studies should I trust and consider for decision making?". Even for clinicians trained in critical appraisal, evaluating all available studies would be a time-intensive solution.

Traditionally, clinicians – when they did not invest the time and resources to review individual studies – have relied on review articles by authorities in the field. However, experts may be unsystematic in their approach to summarizing the evidence. Unsystematic approaches to identification and collection of evidence risks biased ascertainment. That is, treatment effects may be underestimated or, more commonly, overestimated, and side effects may be exaggerated or ignored. Even if the evidence has been identified and collected in a systematic fashion, if reviewers are then unsystematic in the way they summarize the collected evidence, they run similar risks of bias. In one study, self-rated expertise was inversely related to the methodologic rigor of the review (Oxman and Guyatt 1993). One result of unsystematic approaches may be recommendations advocating harmful treatment; in other cases, there may be a failure to encourage effective therapy. For example, experts supported routine use of lidocaine for patients with acute myocardial infarction when available data suggested the intervention was ineffective and possibly even harmful, and they failed to recommend thrombolytic agents for the treatment of acute myocardial infarction when data showed patient benefit (Antman et al. 1992).

Systematic reviews deal with this problem by explicitly stating inclusion and exclusion criteria for evidence to be considered, conducting a comprehensive search for the evidence, and summarizing the results according to explicit rules that include examining how effects may vary in different patient subgroups. When a systematic review pools data across studies to provide a quantitative estimate of overall treatment effect, we call this summary a meta-analysis (cf. Chap. II.7 of this handbook). Systematic reviews provide strong evidence when the quality of the primary study design is high and sample sizes are large; they provide weaker evidence when study designs are poor and sample sizes are small. Because judgment is involved in many steps in a systematic review (including specifying inclusion and exclusion criteria, applying these criteria to potentially eligible studies, evaluating the methodologic quality of the primary studies, and selecting an approach to data analysis), systematic reviews are not immune to bias, for example publication bias.

Nevertheless, in their rigorous approach to identifying and summarizing data, systematic reviews reduce the likelihood of bias in estimating the causal links between management options and patient outcomes.

Over the past 10 to 15 years, the literature describing the methods used in systematic reviews, including studies that provide an empiric basis for guiding decisions about the methods used in summarizing evidence has rapidly expanded. Clinical epidemiologists have contributed significantly to this development (Egger et al. 2000).

Table 8.7 demonstrates the process of conducting systematic reviews.

As we described above for answering questions in clinical practice (see also Table 8.1), investigators who conduct a systematic review should begin by formulating a clinical question. This question formulation constitutes the essential specific selection criteria for deciding which studies to include in a review. These criteria define the population, the exposures or interventions, the comparison intervention, and the outcomes of interest. A systematic review will also restrict the included studies to those that meet minimal methodologic standards. For example, systematic reviews that address a question of therapy will often include only randomized controlled trials.

Table 8.7. The process of conducting systematic reviews

Define the question

- Specify inclusion and exclusion criteria
  - Population
  - Intervention or exposure (and comparison)
  - Outcome
  - Methodology

- Establish a priori hypotheses to explain heterogeneity

Conduct literature search

- Decide on information resources: databases, experts, funding agencies, pharmaceutical companies, hand-searching, personal files, registries, citation lists or retrieved articles
- Determine restrictions: time frame, unpublished data, language
- Identify titles and abstracts

Apply inclusion and exclusion criteria

- Apply inclusion and exclusion criteria to titles and abstracts
- Obtain full articles for eligible titles and abstracts
- Apply inclusion and exclusion criteria to full articles
- Select final eligible articles
- Assess agreement on study selection

Create data abstraction

- Data abstraction: participants, interventions, comparison interventions, study design
- Results
- Methodologic quality
- Assess agreement on validity assessment between data abstractors

Conduct analysis

- Determine method for pooling results
- Pool results (if appropriate)
- Decide on handling of missing data

Having evaluated the potential eligibility of titles and abstracts, and obtained the full text of potentially eligible studies, reviewers apply the selection criteria to the complete reports. Having completed the data collection process, they assess the methodologic quality of the eligible articles and abstract data from each study. Finally, they summarize the data, including, if appropriate, a quantitative synthesis or meta-analysis. The analysis includes an examination of differences among the included studies, an attempt to explain differences in results (exploring heterogeneity), a summary of the overall results, and an assessment of their precision and validity. Guidelines for assessing the validity of reviews and using the results correspond to this process and are available (Oxman et al. 2002).

Returning to our example, the question you want to address could be formulated as: "In patients with coronary artery disease does low dose aspirin (75 mg daily or less) compared with high dose aspirin (325 mg daily or more) confer similar mortality benefits". Keep in mind that gastrointestinal side effects are more likely with higher doses of aspirin.

A prudent clinician will look for a systematic review to answer this question. A recent systematic review by the Antithrombotic Trialists' Collaboration provides useful information to answer this question (Antithrombotic Trialists' Collaboration 2002). The investigators carefully evaluated 448 trials for inclusion and finally performed a meta-analysis of 195 trials of antiplatelet effects in cardiovascular disease. Overall they observed that aspirin markedly reduced the risk of recurrent events in patients with acute myocardial infarction (odds ratio 0.7) and patients with previous myocardial infarction (odds ratio 0.75). They identified three trials in patients at high risk for cardiovascular events that compared doses of aspirin of $\leq$ 75 mg daily to higher doses. Higher doses of aspirin conferred no additional benefit in preventing cardiovascular events when compared with lower doses. In fact the point estimate indicated a benefit of lower doses of aspirin (relative odds reduction 8%). However, these results were not statistically significant (95% CI ranging from approximately -12% indicating a slight benefit with higher doses to 28% indicating a benefit with lower doses). When the investigators compared the effects of low dose aspirin versus placebo and higher dose aspirin versus placebo the effects were similar. Although the evidence from direct comparisons is limited and the investigators included additional patient populations, such as patients with stroke, the systematic review and meta-analysis of all available trials provides additional indirect evidence for comparable effects of higher and low dose aspirin. Most clinicians would judge the biology for the role of aspirin in coronary artery disease and other cardiovascular disease to be sufficiently similar to have confidence in this extrapolation. Combining the available evidence in a properly conducted systematic review helps the clinician to obtain answers that are valid and based on methodological evaluation of the literature. Chapter II.7 of this handbook addresses methodological issues related to meta-analysis.

Having explained some of the benefits of systematic reviews we will explain how systematic reviews can be used to guide patient care. Available guidance on therapeutic or prevention goes beyond systematic reviews, because factors in addition to the quality of the evidence and treatment effects are important to make recommendations about treatment (Freemantle et al. 1999). However, systematic reviews should be conducted in the process of generating evidence-based guidelines that provide treatment recommendations.

8.7 Guidelines

Guidelines are systematically developed syntheses to support practitioners and patients in decision making about specific clinical circumstances. The key elements of each synthesis include the scope of the guidelines, the interventions and practices considered, the major recommendations and the and strength of the evidence and recommendations, and the underlying values and preferences (Guyatt et al. 2002a, 2004b; Schünemann et al. 2004). A list of guidelines maintained by the US Agency for Healthcare Research and Quality is available at the National Guidelines Clearinghouse (National-Guideline-Clearing-House 2004).

Organizations such as professional societies or governing boards, government agencies, academic or private institutions, typically develop guidelines by convening expert panels. Usually, guideline developers will define the topic of a guideline before evaluating the evidence for clinical sensible questions. Dialogue among clinicians, patients, and the prospective users of the guideline contribute to its refinement. Although it is possible to develop guidelines that are broad in scope, it requires considerable time and resources.

One method of defining and focusing the clinical questions of interest and also identifying the processes for which evidence needs to be collected and assessed is the construction of models or causal pathways (Woolf 1994). The causal pathway is a diagram showing the relation of the population, intervention(s) of interest and the intermediate, surrogate or definitive health outcomes (Shekelle et al. 1999). When designing the pathway, guideline developers should describe explicitly which outcomes (benefits and harms) they consider important and the associated values. This process reveals specific questions that the evidence must address and where high quality evidence is lacking, identifying areas for additional research.

While investigators have not tested alternative approaches to guideline development, one suggestion is to include all groups whose activities would be covered by the guidelines and any others with legitimate reasons for having input (Shekelle et al. 1999). A group size of six to 15 individuals with clearly identified roles, including group leader and members, specialist resource, technical support, and administrative support may be advisable (Shekelle et al. 1999). The group should comprise members proficient in the following areas: literature searching and retrieval, epidemiology, biostatistics, health services research, clinical area of interest (generalists and specialists), group process, writing, and editing.

A high-quality clinical practice guideline produced by such groups should consider the following steps (Schünemann et al. 2004):

1. Define explicit criteria to search for evidence. Similar to the clinical sensible questions identified above for searching the evidence and conducting systematic reviews, this step should include a clear definition of the population, the intervention and comparison intervention and the outcome of interest. The development of a clinical pathway helps in identifying the components of the clinical question and in identifying gaps.
2. Define explicit eligibility criteria for the identified evidence. Guideline development groups should define eligibility criteria for the evidence that they wish to include. Examples include restricting evidence to randomized controlled trials or studies that have used validated instruments for functional outcomes or assessed mortality.

3. Conduct or use comprehensive searches for evidence. A guideline development group should ensure that they conduct a complete evaluation of the evidence. The group may either use a high quality systematic review developed by others or conduct their own systematic review for each recommendation they make in their guideline.
4. Perform a standard consideration of study quality. If a systematic review is identified that answers the group's clinical question and may suffice for developing the guideline, the group should still consider an evaluation of the individual study quality. Should the group conduct a systematic review or update an existing review, the evaluation of study quality becomes an essential step in conducting a systematic review. Quality evaluation includes looking at the basic study design, the detailed study design and execution, and the directness of evidence. The directness of the evidence refers to reporting of surrogate outcomes, for example deep venous thrombosis by ultrasonography as risk factor for fatal events, versus outcomes such as mortality.
5. Summarize the evidence. Guideline development groups who use high quality systematic reviews often will find meta-analysis of studies included in the systematic review. The summaries help in obtaining estimates of intervention effects. It is important to note that summary estimates should be available for all important outcomes, both beneficial and harmful. Ideally the summary estimates also would be available for cost.
6. Acknowledge values and preferences underlying the group's recommendations. Many guideline reports take for granted that guideline developers adequately represent patients' interests. The latter is not necessarily correct and there is a risk that, for example, a specialty society may recommend procedures where the benefits may not outweigh the risks or costs (Woolf et al. 1999). For example, the American Urologic Society and the American Cancer Society recommend prostate cancer screening with prostate specific antigen (PSA) for men older then 50 while the American College of Preventive Medicine and the US Preventive Task Force Services (USPSTF) do not make this recommendation (American Urological Association 2000; Ferrini and Woolf 1998; Smith et al. 2003; USPSTF 2002). Thus, there should be clear statements about which principles, such as patient autonomy, nonmaleficence, or distributive justice, were given priority in guiding decisions about the value of alternative interventions to inform users of the guidelines (Shekelle et al. 1999). Guidelines should report whether it is intended to optimize values for individual patients, reimbursement agencies, or society as a whole. Groups that ensure representation by experts in research methodology, practicing generalists and specialists, and public representatives are more likely to have considered diverse views in their discussions than groups limited to content area experts.
7. Grade the strength of recommendations. Because clinicians are interested in the strength of a recommendation and the balance of benefits and risks, the next section is devoted to recommendations and the grading of the strength of recommendations.

Recommendations

Treatment decisions involve balancing likely benefits against harms and costs. Evidence-based guidelines and treatment recommendations are systematic syntheses of the best available evidence that provide clinicians with guidance for treating average patients in clinical practice. To integrate recommendations with their own clinical judgment, clinicians need to understand the basis for the clinical recommendations that experts offer them. A common systematic approach to grading the strength of treatment recommendations can minimize bias and aid interpretation.

As part of the first American College of Chest Physician (ACCP) Consensus Conference on Antithrombotic Treatment in 1986, Sackett suggested a formal rating scheme, derived from the Canadian Task Force on the Periodic Health Examination, for assessing levels of evidence (Canadian 1979; Sackett 1986). During the past 15 years, clinical epidemiologists from McMaster University have lead the evolution of these "rules of evidence" (Cook et al. 1992; Guyatt et al. 1995, 2001, 2002a, b), which experts have applied to generate grades of recommendations.

The strength of any recommendation depends on two factors: the trade-off between benefits and downsides and the quality of the methodology that leads to estimates of the treatment effect. The ACCP approach to grading of recommendations captures the magnitude of random error in the decision about the confidence in the tradeoff between benefits, harms and cost. The uncertainty associated with this tradeoff will determine the strength of recommendations. The grades that experts generate using the ACCP approach are 1A, 1C+, 1B, 1C, 2A, 2C+, 2B and 2C (Table 8.8). If experts are very certain that benefits do, or do not, outweigh harms and cost, they will make a strong recommendation – in the ACCP formulation, Grade 1. If they are less certain of the magnitude of the benefits and harms, and thus their relative impact, they must make a weaker Grade 2 recommendation. Grade 2 recommendations are those in which variation in patient values or individual physician values often will mandate different treatment choices, even among average or typical patients. The ACCP approach expresses the primacy of the benefit versus downxside judgment by placing it first in the grade of recommendation.

However, today a number of organizations other than the ACCP, including the US Preventive Services Task Force (Harris et al. 2001), the US Task Force on Community Preventive Services (Briss et al. 2000), Scottish Intercollegiate Guidelines Network (SIGN) (Harbour and Miller 2001), the National Institute for Clinical Excellence (NICE), and more than 100 other groups use various systems of codes to communicate grades of evidence and recommendations. All these organizations have definitions of varying length and detail for each letter or number code and a few use single words, such as "Strong" or "Weak", in addition to or in place of a code.

Health care practitioners, in particular learners, are often puzzled by the message the grade of these systems convey. For example, the administration of oral anticoagulation in patients with atrial fibrillation and rheumatic mitral valve dis-

Table 8.8. ACCP approach to grades of recommendations

Grade of recommendation	Clarity of risk/benefit	Methodologic strength of supporting evidence	Implications
1 A	Risk/benefit clear	RCTs without important limitations	Strong recommendation, can apply to most patients in most circumstances without reservation
1 C+	Risk/benefit clear	No RCTs but strong RCT results can be unequivocally extrapolated, or overwhelming evidence from observational studies	Strong recommendation, can apply to most patients in most circumstances
1 B	Risk/benefit clear	RCTs with important limitations (inconsistent results, methodological flaws)*	Strong recommendations, likely to apply to most patients
1 C	Risk/benefit clear	Observational studies	Intermediate strength recommendation; may change when stronger evidence available
2 A	Risk/benefit unclear	RCTs without important limitations	Intermediate strength recommendation, best action may differ depending on circumstances or patients' or societal values
2 C+	Risk/benefit unclear	No RCTs but strong RCT results can be unequivocally extrapolated, or overwhelming evidence from observational studies	Weak recommendation, best action may differ depending on circumstances or patients' or societal values
2 B	Risk/benefit unclear	RCTs with important limitations (inconsistent results, methodological flaws)*	Weak recommendation, alternative approaches likely to be better for some patients under some circumstances
2 C	Risk/benefit unclear	Observational studies	Very weak recommendations; other alternatives may be equally reasonable

* These situations include RCTs (randomized clinical trials) with both lack of blinding and subjective outcomes, where the risk of bias in measurement of outcomes is high, or RCTs with large loss to follow up.

Note: Since studies in categories B and C are flawed, it is likely that most recommendations in these classes will be level 2.

The following considerations will bear on whether the recommendation is Grade 1 or 2: the magnitude and precision of the treatment effect, patients' risk of the target event being prevented, the nature of the benefit, and the magnitude of the risk associated with treatment, variability in patient preferences, variability in regional resource availability and health care delivery practices, and cost considerations (see Table 8.2). Inevitably, weighing these considerations involves subjective judgment (reproduced with permission from Guyatt et al. (2004a).)

ease receives various grades of recommendation from different organizations. Oral anticoagulation in these patients is recommended as Class I based on level B evidence by the American Heart Association (ACC 2001), as a grade C recommendation based on level IV evidence by SIGN (http://www.guidelines.gov/) and as grade 1C+, where the 1 indicates the balance between benefit and downsides and C+ the methodological quality of the underlying evidence, by the American College of Chest Physicians (Albers et al. 2001). It is therefore possible that the different grading systems do not fulfill their intended function: to quickly and concisely communicate a clear message. In particular, if the same code, used by different systems, represents different meanings, bewilderment and incomprehension may result.

A group of guideline developers and clinical epidemiologists formed the Grades or Recommendation Assessment, Development and Evaluation (GRADE) Working Group with the hope of reaching agreement on a common, sensible approach to grading the quality of evidence and strength of recommendations (Atkins et al. 2004; Schünemann et al. 2003). The group has defined grade of evidence as indicating the extent to which one can be confident that an estimate of effect is correct and grades of recommendations as indicating the extent to which one can be confident that adherence to a recommendation will do more good than harm. The Grade group suggests that those developing recommendations should make sequential judgments about the quality of evidence for each important outcome, the overall quality of evidence across outcomes, and the recommendations. Judgments about the quality of evidence require consideration of study design, study quality, consistency and directness of the evidence. Additional considerations include reporting bias, sparse data and strength of associations. Judgments about recommendations require consideration of the balance between benefits and harms, the quality of the evidence, translation of the evidence into specific circumstances, and the certainty of the baseline risk. Recommendations should consider costs (resource utilization) as well as benefits and harms. The GRADE group further concludes that inconsistencies among systems for grading evidence and formulating recommendations reduce their potential to facilitate critical appraisal and improve communication of these judgments, and suggests a system that is new (though bears many similarities to the ACCP approach) that they hope will receive wide adoption.

Returning to our clinical scenario and applying the ACCP approach to grading of recommendation (ideally one will apply the GRADE approach when available), the use of clopidogrel in addition to aspirin would generate a 2A recommendation where the 2 indicates that the individual preferences and values influence the treatment decision and the A denomination indicates that evidence stems from one or more high-quality randomized controlled trials.

In the next section we will describe the importance of health related quality of life (HRQL) outcomes followed by a section on integrating patient preferences in decision making and how clinical epidemiology is key in obtaining patients' preferences and values.

8.8 Health Related Quality of Life Instruments and Their Application in Clinical Studies

Clinical journals have published trials in which HRQL instruments are the primary outcome measures. With the expanding importance of HRQL in evaluating new therapeutic interventions, investigators (and readers) are faced with a large array of instruments. Researchers have proposed different ways of categorizing these instruments, according to the purpose of their use, into instruments designed for screening, providing health profiles, measuring preference, and making clinical decisions (Osoba et al. 1991), or into discriminative and evaluative instruments.

We have also suggested a taxonomy based on the domains of HRQL which an instrument attempts to cover (Guyatt et al. 1989). According to this taxonomy, a HRQL instrument may be categorized, in a broad sense, as generic or specific. *Generic instruments* cover (or at least aim to cover) the complete spectrum of function, disability, and distress of the patient, and are applicable to a variety of populations. Within the framework of generic instruments, health profiles and utility measures provide two distinct approaches to measurement of global quality-of-life. *Specific instruments* are focused on disease or treatment issues specifically relevant to the question at hand.

8.8.1 Generic Instruments

Health Profiles

Health profiles are single instruments that measure multiple different aspects of quality-of-life. They usually provide a scoring system that allows aggregation of the results into a small number of scores and sometimes into a single score (in which case, it may be referred to as an index). As generic measures, their design allows their use in a wide variety of conditions. For example, one health profile, the Sickness Impact Profile (SIP) contains 12 "categories" which can be aggregated into two dimensions and five independent categories, and also into a single overall score (Bergner et al. 1981). The SIP has been used in studies of cardiac rehabilitation (Ott et al. 1983), total hip joint arthroplasty (Liang et al. 1985), and treatment of back pain (Deyo et al. 1986). In addition to the SIP, there are a number of other health profiles available: the Nottingham Health Profile (Hunt et al. 1980), the Duke-UNC Health Profile (Parkerson et al. 1981), and the McMaster Health Index Questionnaire (Sackett et al. 1977). Increasingly, a collection of related instruments from the Medical Outcomes Study (Tarlov et al. 1989), has become the most popular and widely-used generic instruments. Particularly popular is one version that includes 36 items, the SF-36 (Brook et al. 1979; Ware et al. 1995; Ware and Sherbourne 1992). The SF-36 is available in over 40 languages and normal values for the general population in many countries are available.

While each health profile attempts to measure all important aspects of HRQL, they may slice the HRQL pie quite differently. For example, the McMaster Health

Index Questionnaire follows the World Health Organization approach and identifies three dimensions: physical, emotional, and social. The Sickness Impact Profile includes a physical dimension (with categories of ambulation, mobility, body care, and movement), a psychosocial dimension (with categories including social interaction and emotional behavior), and five independent categories including eating, work, home management, sleep and rest, and recreations and pastimes.

General health profiles offer a number of advantages clinical investigators. Their reproducibility and validity have been established, often in a variety of populations. When using them for discriminative purposes, one can examine and establish areas of dysfunction affecting a particular population. Identification of these areas of dysfunction may guide investigators who are constructing disease-specific instruments to target areas of potentially greatest impact on the quality-of-life. Health Profiles, used as evaluative instruments, allow determination of the effects of an intervention on different aspects of quality-of-life, without necessitating the use of multiple instruments (and thus saving both the investigator's and the patient's time). Because health profiles are designed for a wide variety of conditions, one can potentially compare the effects on HRQL of different interventions in different diseases. Profiles that provide a single score can be used in a cost-effectiveness analysis, in which the cost of an intervention in dollars is related to its outcome in natural units.

The main limitation of health profiles is that they may not focus adequately on the aspects of quality-of-life specifically influenced by a particular intervention. This may result in an inability of the instrument to detect a real effect in the area of importance (i.e., lack of responsiveness). In fact, disease specific instrument offer greater responsiveness compared with generic instruments (Guyatt et al. 1999; Wiebe et al. 2003). We will return to this issue when we discuss the alternative approach, specific instruments.

Specific Instruments 8.8.2

An alternative approach to HRQL measurement is to focus on aspects of health status that are specific to the area of primary interest. The rationale for this approach lies in the increased responsiveness that may result from including only those aspects of HRQL that are relevant and important in a particular disease process or even in a particular patient situation. One could also focus an instrument only on the areas that are likely to be affected by a particular drug.

In other situations, the instrument may be specific to the disease (instruments for chronic lung disease, for rheumatoid arthritis, for cardiovascular diseases, for endocrine problems, etc.); specific to a population of patients (instruments designed to measure the HRQL of the frail elderly, who are afflicted with a wide variety of different diseases); specific to a certain function (questionnaires which examine emotional or sexual function); or specific to a given condition or problem (such as pain) which can be caused by a variety of underlying pathologies. Within a single condition, the instrument may differ depending on the intervention. For example, while success of a disease-modifying agent in rheumatoid arthritis should result in improved HRQL by enabling a patient to increase performance of physically stressful activities of daily living, occupational therapy may achieve improved HRQL by encouraging family members to take over activities formerly accomplished with difficulty by the patient. Appropriate disease-specific HRQL outcome measures should reflect this difference.

Specific instruments can be constructed to reflect the "single state" (how tired have you been: very tired, somewhat tired, full of energy) or a "transition" (how has your tiredness been: better, the same, worse) (MacKenzie and Charlson 1986). Theoretically, the same could be said of generic instruments, although none of the available generic instruments has used the transition approach. Specific measures can integrate aspects of morbidity, including events such as recurrent myocardial infarction (Olsson et al. 1986).

The disease-specific instruments may be used for discriminative purposes. They may aid, for example, in evaluating the extent to which a primary symptom (for example dyspnea) is related to the magnitude of physiological abnormality (for example exercise capacity) (Mahler et al. 1987). Disease-specific instruments can be applied for evaluative purposes to establish the impact of an intervention on a specific area of dysfunction, and hence aid in elucidating the mechanisms of drug action (Jaeschke et al. 1991). Guidelines provide structured approaches for constructing specific measures (Guyatt et al. 1986). Whatever approaches one takes to the construction of disease-specific measures, a number of head-to-head comparisons between generic and specific instruments suggests that the latter approach will fulfill its promise of enhancing an instrument's responsiveness, the ability to detect change in HRQL (Chang et al. 1991; Goldstein et al. 1994; Laupacis et al. 1991; Smith et al. 1993; Tandon et al. 1989; Tugwell et al. 1990).

In addition to the improved responsiveness, specific measures have the advantage of relating closely to areas routinely explored by the physician. For example, a disease-specific measure of quality-of-life in chronic lung disease focuses on dyspnea during day-to-day activities, fatigue, and areas of emotional dysfunction, including frustration and impatience (Guyatt et al. 1987). Specific measures may therefore appear clinically sensible to the clinician.

The disadvantages of specific measures are that they are (deliberately) not comprehensive, and cannot be used to compare across conditions or, at times, even across programs. This suggests that there is no one group of instruments that will achieve all the potential goals of HRQL measurement. Thus, investigators may choose to use multiple instruments. Some of these instruments are preferences or value instruments that can also be used for clinical decision making described in the next section.

Integrating Patient Preferences in the Decision Making Process and Resolution of the Clinical Scenario

In reviewing the data from the CURE trial, you conclude that if the patient before you remains untreated, the best estimate of the risk for recurrent myocardial infarction or any of the other endpoints summarized in the composite endpoint in the trial during the next year is 16.5%, and, further, that clopidogrel is likely to decrease this risk by approximately 13%, corresponding to absolute risk reductions (ARR) of 2.3% over a one-year period. As described above, this translates into a number needed to treat (NNT) for 1 year to prevent one of the endpoints summarized in the composite endpoint of approximately 44 for treatment with clopidogrel (Table 8.5).

Examining the likelihood of major bleeding, the CURE trial suggests an absolute risk increase of 1.0%. This estimate translates into a NNH of 100. In light of your knowledge that the patient before you is intelligent, conscientious, and very concerned about his health, you anticipate a high rate of adherence; in addition, you anticipate that the bleeding risk rate of 1% (or the NNT of 100) represents a good estimate for risk of the patient in front of you.

Considering these numbers, you are aware that the treatment decision may depend on the relative value the patient places on avoiding a recurrent myocardial infarction or any of the other endpoints summarized in the composite endpoint in the CURE trial and avoiding a major bleeding including those leading to vision loss. We have pointed out that since there are always advantages and disadvantages to an intervention, evidence alone cannot determine the best course of action. Patients, their proxies, or if a parental approach to decision making is desirable, the clinician as decision-maker, must always trade the benefits, harm, and costs associated with alternative treatment strategies, and values and preferences always bear on those trade-offs. Findings that patients vary greatly in the value they place on different outcomes will come as no surprise. Given this variability in patient's values, clinicians should proceed with great care; it is easy to assume that the patient's values are similar to one's own, yet this may be incorrect. For example, facing a decision concerning anticoagulation in atrial fibrillation, clinicians are more concerned about bleeding risk, and place less weight on the associated stroke reduction, than patients (Devereaux et al. 2001). Thus, a fundamental principle of EBM is the explicit inclusion of patients and society's values and clinical circumstances in the clinical decision-making process (Fig. 8.4) (Haynes et al. 2002). Clinical epidemiology can help identifying and applying the key issues in the decision making process.

Considering the model by Haynes et al. (2002) you are now faced with the problem of how to best incorporate the patient's values into the decision. Before resolving the scenario we will describe different ways to optimize decision making.

For many – perhaps most – of our clinical decisions, the tradeoff is sufficiently clear that clinicians need not concern themselves with variability in patient values. Previously healthy patients will all want antibiotics to treat their pneumonia or their urinary tract infection, anticoagulation to treat their pulmonary embolus, or aspirin to treat their myocardial infarction. Under such circumstances, a brief explanation of the rationale for treatment and the expected benefits and side effects will suffice.

When benefits and risks are balanced more delicately and the best choice may differ across patients, clinicians must attend to the variability in patients values (such as in a Grade 2A recommendation in the McMaster approach to grading recommendations). One fundamental strategy for integrating evidence with preferences involves communicating the benefits and risks to patients, thus permitting them to incorporate their own values and preferences in the decision. One advantage of this approach is that it avoids the vexing problem of measuring patients' values. Unfortunately, the problem of communicating the evidence to patients in a way that allows patients to clearly and unequivocally understand their choices is almost as vexing as the direct measurement of patient values.

A second basic strategy is to ascertain the relative value patients place on the key outcomes associated with the management options. One can then consider the likely outcomes of alternative courses of action and use the patient's values as the basis of trading off benefits and risks. When done in a fully quantitative way, this approach becomes a decision analysis using individual patient preferences (Guyatt et al. 2002a,b). A number of texts provide information on decision analysis and decision analyses are available for a number of topics, such as the prevention of

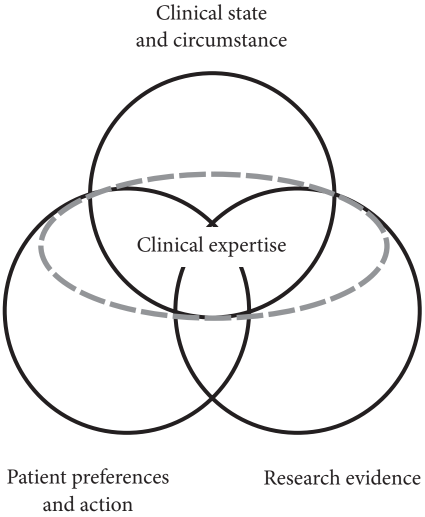

Figure 8.4. Model for Evidence-Based Decision Making. Reproduced with permission from Haynes et al. (2002)

ischemic stroke with warfarin in atrial fibrillation (Guyatt et al. 2002a,b; Petitti 1994; Thomson et al. 2000).

In addition, patients often have preferences not only about the outcomes, but about the decision-making process itself. These preferences can vary, and the patient's desired level of involvement should determine which approach the clinician takes (Degner et al. 1997; Stiggelbout and Kiebert 1997; Strull et al. 1984). Ethicists have characterized the alternative strategies (Emanuel and Emanuel 1992). At one end of the spectrum, the physician acts as a technician, providing the patient with information and taking no active part in the decision-making process. This corresponds to the first strategy for incorporating patient values, presenting patients with the likely benefits, risks, inconvenience and cost and then letting patients decide. At the opposite extreme, corresponding to the second strategy, ascertaining the patient's values and then making a recommendation in light of the likely advantages and disadvantages of alternative management approaches, the clinician takes a "paternalistic" approach and decides what is best for the patient in light of that patient's preferences.

However, intermediate approaches of shared decision making are generally more popular than those at either extreme. Shared decision making uses both of the two fundamental approaches to decision making presented above: The clinician typically shares the evidence, in some form, with the patient, while simultaneously attempting to understand the patient's values. Evidence that more active patient involvement in the process of health care delivery can improve outcomes and reported quality of life – and, possibly, reduce health care expenditures – provides empirical evidence in support of secular trends toward patient autonomy and away from paternalistic approaches (Greenfield et al. 1988; Stewart 1995; Szabo et al. 1997).

Clinicians should temper their enthusiasm for active patient involvement in decision making with an awareness that many patients prefer paternalistic approaches. For example, the results of a survey of 2472 patients suffering from chronic disease (hypertension, diabetes, heart failure, myocardial infarction, or depression) completed between 1986 and 1990 supported this approach (Arora and McHorney 2000). In response to the statement: "I prefer to leave decisions about my medical care up to my doctor", 17.1% strongly agreed, 45.5% agreed, 11.1% were uncertain, 22.5% disagreed, and only 4.8% strongly disagreed. In a more recent study of node-negative breast cancer patients considering adjuvant chemotherapy, 84% of 171 women preferred an independent or shared role in decision-making (Whelan et al. 2003). Increasing general levels of education, the advent of the Internet and the resulting access to medical information, and an increasingly litigious and consumerist environment have all contributed to patients wishing to play a more active role in decision-making and may explain a shift in patient preferences for decision making. Shared decision-making and patient-centeredness have become attractive approaches to resolving the profusion of challenging choices facing patients and clinicians (Charles et al. 1999a, b; Edwards and Elwyn 2001; Guyatt et al. 2004a).

Regardless of the decision-making approach chosen by the patient and clinician, integrating values and preferences and communicating options injects challenges into the process by insisting that clinicians consider quantitative estimates of benefits and risks, rather than just whether a treatment works or whether toxicity occurs. If clinicians leave the decisions to patients, they must effectively communicate the probabilities associated with the alternative outcomes to them. If they opt for taking responsibility for combining patient values with the evidence, they must quantify those values. A vague sense of the patient's preferences cannot fully satisfy the rigor of the optimal decision making approach.

We will now describe some of the specific strategies associated with two decision-making models: one in which the clinician presents the patient with the likely consequences of alternative management strategies and leaves the choice to the patient, and the other in which the clinician ascertains the patient's values and provides a recommendation.

Patient as Decision-Maker: Decision Aids

If the patient wishes to play the primary role in decision making, clinicians may use intuitive approaches to communicating concepts of risk and risk reduction that they have developed through clinical experience. They will answer the patient's questions and ultimately act on the patient's decision. Alternatively, if available for a particular decision, clinicians can use a decision aid that presents descriptive and probabilistic information about the disease, treatment options, and potential outcomes in a patient-friendly manner (Barry 2002; Holmes-Rovner et al. 2001; Levine et al. 1992; O'Connor 2001).

A well-constructed decision aid has two advantages. One is that someone has reviewed the literature and produced a rigorous summary of the probabilities. Clinicians who doubt that the summary of probabilities is rigorous can go back to the original literature on which those probabilities are based and determine their accuracy. A second advantage of a well-constructed decision aid is that it will offer a pre-tested and effective way of communicating the information to patients who may have little background in quantitative decision making. Most commonly, decision aids use visual props to present the outcome data in terms of the percentage of people with a certain condition who do well without intervention, compared to the percentage who do well with intervention. Decision aids will summarize the data regarding all outcomes of importance to patients.

Theoretically, decision aids present an attractive strategy for ensuring that patient values guide clinical decision making. What impact do decision aids actually have on clinical practice? O'Connor and colleagues conducted a systematic review, finding 17 randomized trials that used 11 different decision aids, for example the decision for or against hormone replacement therapy in women after menopause or decisions related to breast surgery in breast cancer (O'Connor et al. 2003, 2004). Of these 17 trials, decision aid impact on knowledge was evaluated in four. All four found greater knowledge in the decision aid group, with a pooled difference of 19 on a 100-point scale (95% CI: $(14, 25)$). Decision aids reduced decisional conflict using a validated decisional conflict scale in three of four trials in which investigators addressed this issue (mean effect: 0.3; 95% CI: $(0.1; 0.4)$ on the 5-point scale decisional conflict scale). Three studies failed to show a difference in satisfaction with the decision made, although one of these three showed increased satisfaction with the decision-making process.

In summary, decision aids markedly increase patient knowledge and decrease discomfort with decision making as reflected in decisional conflict scores. The importance of the reduction in decisional conflict remains uncertain. Simple decision aids that clinicians can integrate into regular patient care could increase the extent to which patient values truly determine health care decisions.

Patient as Provider of Values

The second set of approaches all begin with, at minimum, establishing the relative value the patient places on the target outcomes. Doing so requires that the patient understand the nature of those outcomes. How, for instance, would a patient with atrial fibrillation facing a decision about using oral anticoagulation to prevent strokes imagine living with a stroke, or the experience of having a gastrointestinal bleeding episode as a side effect of the oral anticoagulation? Patients may find a written description of the health states (Table 8.9) useful in the process of describing their preferences (Devereaux et al. 2001).

Having made their best effort to ensure that patients understand the outcomes, clinicians can choose from among a number of ways of obtaining their values for those outcomes. They can gain a qualitative sense of their patients' preferences from a discussion without a formal structure. Alternatively, a direct comparison between outcomes may prove useful. For instance, with only two outcomes, the patient can make a direct comparative rating. The question may be: "How much worse would it be to have a stroke versus a gastrointestinal bleeding

Table 8.9. Sample descriptions of major stroke and gastrointestinal bleed

Major Stroke	Bleeding
You suddenly are dizzy and blackout You are unable to move one arm and one leg You cannot swallow or control bladder and bowel You are unable to understand what is being said You are unable to talk You feel no physical pain You are admitted to hospital You cannot dress The nurse feeds you You cannot walk After 1 month with physiotherapy, you are able to wiggle your toes and lift your arm off the bed You remain this way for the rest of your life Another illness will likely cause your death	You feel unwell for two days then suddenly you vomit blood You are admitted to hospital You stop taking warfarin A doctor puts a tube down your throat to see where you are bleeding from You receive sedation to ease the discomfort of the test You do not need an operation You receive blood transfusions to replace the blood you lost You stay in hospital one week You feel well at the end of your hospital stay You need to take pills for the next six months to prevent further bleeding You do not take warfarin any more After that you are back to normal

episode? Would it be equally bad? Twice as bad to have a stroke? Three times as bad?"

Using a somewhat more complex strategy, the clinician can ask the patient to place a mark on a visual analogue scale or "feeling thermometer", in which the extremes are anchored at dead and full health, to represent how the patient feels about the health states in question (Fig. 8.5).

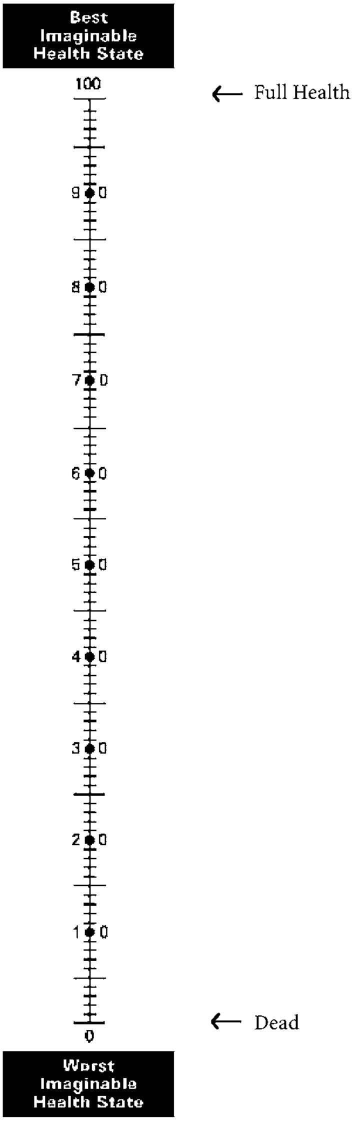

Figure 8.5. The feeling thermometer

When, as in the case of a gastrointestinal bleeding and a stroke, some health states are temporary and others are permanent, the clinician must ensure that patients incorporate the duration of the health state in their rating.

More sophisticated approaches include the time tradeoff and the standard gamble (Torrance 1986). In completing the time tradeoff, patients choose between a longer period in a state of impaired health (such as recovery from severe stroke) and a shorter period in a state of full health. With the standard gamble, by contrast, patients are asked to choose between living in a state of impaired health versus taking a gamble in which they may return to full health or die immediately. These latter approaches may come much closer to meeting assumptions that health economists argue are necessary for accurate ratings of the relative value of health states in the context of choice with uncertain outcomes.

Regardless of the strategy clinicians use to obtain patient values, they must somehow integrate these values with the likely outcomes of the alternative management strategies. Formal decision analysis provides the most rigorous method for making this integration. Practical software for plugging in the patients' values and conducting a patient-specific decision analysis for common clinical problems is being developed, although not yet available for routine use in daily clinical practice. Investigators have shown that, when patients' values are used in individualized decision analyses, their decisions about anticoagulation in atrial fibrillation differ from those suggested by existing guidelines (Protheroe et al. 2000). Whether the decisions would have differed had the patients been provided with the probabilities and asked to choose their preferred management strategy – as with a decision aid – remains unknown.

Even if the tools for individual decision analysis were widely available, application of the approach would depend on the availability of clinicians who could devote time to eliciting patient values. Such a process may be resource intensive, and issues of how much gain there is from the investment, or the intervention's cost-effectiveness, may become very important. Exactly the same considerations apply to the use of decision aids, in which the improvement of knowledge is clear but the impact on anxiety, or on the choices patients actually make, is not as obvious.

Another method of expressing information to patients that incorporates their values is the likelihood of being helped versus harmed (Sackett et al. 2000). Clinicians can apply the likelihood of being helped versus harmed to any clinical decision, and preliminary evidence suggests the approach may be useful on busy clinical services. The clinician begins by calculating the NNT and NNH for the average patients in a study or studies from which the data about treatment effectiveness and harm come. The clinician then adjusts the average NNT and NNH for the individual patient according to that patient's likelihood of suffering the target event that treatment is intended to prevent, and the risks it may precipitate, relative to the average patient. Having established the relative likelihood of help versus harm, the clinician explores the patient's values about the severity of adverse events that might be caused by the treatment relative to the severity of the target event that treatment helps prevent. The final adjustment of the likelihood of being helped versus harmed incorporates the patient's values without providing formal help by a decision aid.

Likelihood of Help Versus Harm

For sake of simplicity, we will assume that the patient in our scenario places a mean value on the composite endpoints cardiovascular disease (defined by the CURE investigators as death from cardiovascular causes, nonfatal MI, stroke, or refractory ischemia), major bleeding (substantially disabling bleeding, intraocular bleeding or the loss of vision, or bleeding necessitating the transfusion of at least 2 units of blood) and minor bleeding (any other bleeding leading to interruption of study medication), respectively. We will ignore other factors bearing on the decision, such as taking an additional pill daily.

During your discussion with the patient about the consequences of further cardiovascular disease and major bleeding you asked her to use the "feeling thermometer" (see Fig. 8.5) to estimate how she feels about each of the two combined outcomes. We will ignore minor bleeding episodes in this example, because your patient is not concerned at all about the risk and consequences of minor bleeding. However, she places a mean value of suffering additional consequences of cardiovascular disease at 0.2 and of living with a major bleed at 0.7. You use these on your handheld personal digital assistant to calculate her likelihood of being helped or harmed (LHH) from clopidogrel therapy versus placebo therapy.

Using the NNTs calculated in the scenario, the LHH for clopidogrel versus placebo becomes (Table 8.5):

$$\mathrm{LHH} = (1/\mathrm{NNT}) : (1/NNH) = (1/44) : (1/100) = 100/44 = 2.3\,.$$

Note: we could also use (1/absolute risk reduction) : (1/absolute risk increase) but this uses decimal fractions and may increase the likelihood of arithmetic errors.

Therefore, you can tell the patient that clopidogrel is approximately twice as likely to help her as to harm her, when compared with placebo.

Incorporating her values that you elicited, the LHH becomes:

$$\begin{aligned}\mathrm{LHH} &= (1/\mathrm{NNT}) \times (1 - \mathrm{Uevent}) : (1/\mathrm{NNH}) \times (1 - \mathrm{Utoxicity})\\ &= (1/44) \times (1 - 0.2) : (1/100) \times (1 - 0.7) = 6.1\,,\end{aligned}$$

where Uevent is the value of the outcome prevented (composite endpoint of death from cardiovascular causes, nonfatal MI, stroke, or refractory ischemia) and Utoxicity (Major bleeding) is the value of the side effect.

You can now inform the patient that clopidogrel is approximately six times as valuable to help her as to harm her. Including additional outcomes would increase the number of terms in the numerator (benefits) or denominator (adverse consequences).

Alternatively, a quicker way of incorporating the patient's values is to ask the patient to rate one event against another. For example, is the adverse effect about as severe as the event the treatment prevents – or 10 times as bad or only half as severe? This rating ("s") can then be used to adjust the LHH as:

$$\mathrm{LHH} = (1/\mathrm{NNT}) \times \mathrm{s} : (1/\mathrm{NNH})\,.$$

Having ascertained the likely outcomes of the alternate courses of action, the clinician must either present patients with the options and outcomes and leave it for them to choose, try to discover the patient's values and having done so suggest a course of action to the patient (the paternalistic approach), or choose the middle course of shared decision making. The patient's preferred decision-making style will guide the clinician in this regard. However, communicating the nature of the outcomes and their probabilities in a way the patient will understand, or accurately ascertaining the patient's values regarding the outcomes, remains problematic.

The challenges of optimal clinical decision making should not obscure the realization that clinicians face these challenges in helping patients with every management decision. For each choice, clinicians guide patients with their best estimate of the likely outcomes. They then help patients balance these outcomes in making their ultimate decision. Finding better strategies to carry out these tasks remains a frontier for clinical epidemiology.

Semistructured Conversation and Resolution

Example 2. You discuss the option of clopidogrel therapy with your patient who is feeling better now and appears to have a good understanding of the information you are providing. You explain that – based on your assessment and the patients' values and preferences – benefit and harm of clopidogrel are finely balanced: for every 44 patients treated for one year in the CURE trial there was one less occurrence of the combined endpoint. However, for every 100 patients treated with clopidogrel for one year one additional patient suffered a major bleeding episode and for every 37 patients treated for one year there was one additional minor bleeding episode. You also explain that, because these are only estimates, the true effect might be somewhat smaller or larger for both the benefit and harms. Your patient states she would like to use clopidogrel.

Because the decision regarding taking clopidogrel depends on the patient's values and preferences regarding preventing the combined endpoint versus incurring additional risk of bleeding and you have the results of the calculation of the likelihood of being helped or harmed. You explain that the results of your calculation using the software on your personal digital assistant support her preference for taking clopidogrel. In terms of expense, you are uncertain about the cost of clopidogrel. Because you feel that this question will come up with additional patients and that you had wanted to address it for some time, you call the hospital pharmacist. She informs you that the cost for clopidogrel is approximately $90 per month and that at least one analysis has suggested the drug is not cost-effective (Gaspoz et al. 2002). The patient tells you that she has minimal co-pay for most medications and remains interested in taking the medication. Together, you decide that beginning clopidogrel treatment ultimately is in her best interest and you start the patient on a 300 mg loading dose of clopidogrel and continue with 75 mg daily. You also suggest reducing the dose of aspirin as lower doses of aspirin confer similar benefits and doses of 75 mg to 325 mg were given in the CURE trial together with Clopidogrel. ♦

8.10 Conclusions

We have presented several concepts of clinical epidemiology. Working through the clinical problems has implicitly highlighted some of clinical epidemiology's research challenges. The following makes explicit some of these challenges that clinical epidemiologists and other investigators need to tackle in future research. A number of important areas have been identified and we will list them here briefly. It is not clear what are the best ways to educate clinicians and students in the methodology of clinical epidemiology and educational researchers will have to focus on this aspect. The Cochrane Collaboration, other organizations and researchers will further elaborate the methodology of systematic reviews (e.g., of diagnostic studies, observational studies and health related quality of life outcomes). Obtaining further information about the most valid and informative ways of presenting statistical information and education of patients and clinicians about these issues is an important task for clinical epidemiologists. Research should also focus on improving the development of clinical practice guidelines and integrating cost information in guidelines and recommendations (Schünemann et al. 2004). Furthermore, research on implementing guidelines into clinical practice is an area of intensive research. It is clear that studies of guideline implementation should follow the same methodological rigor as other studies, but they are presented with different challenges, such as the need for large cluster randomized clinical trials. We described the integration of preferences and values in medical decision making as well as bedside decision making in particular above. Tools that facilitated this difficult task are in development. Health decision aids, in particular electronic decision aids promise to advance this science. Along with health decision aids, the integration of HRQL information into clinical practice and guidelines presents challenges that investigators need to resolve (Frost et al. 2003). Finally, conducting additional research of integrating electronic health (eHealth) (including multimedia decision aids) into clinical practice presents a fascinating but challenging outlook.

References

ACC (2001) ACC/AHA/ESC guidelines for the management of patients with atrial fibrillation. A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines and the European Society of Cardiology Committee for Practice Guidelines and Policy Conferences. J Am Coll Cardiol 38:1231–1266

Albers G, Dalen JE, Laupacis A, Manning WJ, Petersen P, Singer DE (2001) Antithrombotic therapy in atrial fibrillation. Chest 119:194S–206S

American Urological Association (2000) Prostate-specific antigen (PSA) best practice policy. Oncology:267–286

Antithrombotic Trialists' Collaboration (2002) Collaborative metaanalysis of randomised trials of antiplatelet therapy for prevention of death, myocardial infarction, and stroke in high risk patients. BMJ 324:71–86

Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC (1992) A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. Treatments for myocardial infarction. JAMA 268:240–248

Arora N, McHorney C (2000) Patient preferences for medical decision making: who really wants to participate? Med Care 38:335–341

Atkins D, Best D, Briss P, Eccles M, Falck-Ytter Y, Flottorp S, Guyatt GH, Harbour R, Haugh M, Henry D, Hill S, Jaeschke R, Leng G, Liberati A, Magrini N, Mason J, Middleton P, Mrukowicz J, O'Connell D, Oxman A, Phillips B, Schünemann H, Tan-Torres Edejer T, Varonen H, Vist G, Williams J, Zaza S (2004) Grading evidence and formulating recommendations. BMJ (in press)

Balk EM, Bonis PA, Moskowitz H, Schmid CH, Ioannidis JP, Wang C, Lau J (2002) Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials.[comment]. JAMA 287(22):2973–2982

Barry MJ (2002) Health decision aids to facilitate shared decision making in office practice. Ann Intern Med. Online 136(2):127–135

Bergner M, Bobbitt R, Carter W, Gilson B (1981) The Sickness Impact Profile: Development and final revision of a health status measure. Med Care 19:787–805

Bernardo JM, Adrian FM (1994) Bayesian theory. Smith Wiley, Chichester

Berry DA (1996) Statistics: A Bayesian perspective. Duxbury Press, Wadworth

Briss P, Zaza S, Pappaioanou M, Feilding J, Wright-de Aguero L, Truman BI, Hopkins DP, Mullen PD, Thompson RS, Woolf SH, Carande-Kulis VG, Anderson L, Hinman AR, MdQueen DV, Teutsch SM, Harris JR, The Task Force on Community Preventive Services (2000) Developing an evidence-based guide to community preventive services – methods. Am J Prev Med 18:35–43

Brook RH, Ware J, Davies-Avery A, Stewart A, Donald C, Williams KN, Johnston SA (1979) Overview of adult health status measures fielded in Rand's health insurance study. Med Care 17(Suppl 7):1–131

Bucher HC, Weinbacher M, Gyr K (1994) Influence of method of reporting study results on decision of physicians to prescribe drugs to lower cholesterol concentration. BMJ 309(6957):761–764

Canadian (1979) Task Force on the Periodic Health Examination. The periodic health examination. CMAJ 121:1193–1254

Chang S, Fine R, Siegel D, Chesney M, Black D, Hulley S (1991) The impact of diuretic therapy on reported sexual function. Arch Intern Med 151:2402–2408

Charles C, Gafni A, Whelan T (1999a) Decision-making in the physician-patient encounter: revisiting the shared treatment decision-making model. Soc Sci Med 49(5):651–661

Charles C, Whelan T, Gafni A (1999b) What do we mean by partnership in making decisions about treatment? BMJ 319(7212):780–782

Cook DJ GG, Laupacis A, Sackett DL (1992) Rules of evidence and clinical recommendations on the use of antithrombotic agents. Chest (102):305S–11S

CURE-Investigators (2001) Effects of Clopidogrel in addition to Aspirin in Patients with Acute Coronary Syndromes without ST-Segmen Elevation. N Engl J Med 345:494–502

Degner L, Kristjanson L, Bowman D, Sloan J, Carriere K, O'Neil J, Bilodeau B, Watson P, Mueller B (1997) Information needs and decisional preferences in women with breast cancer. JAMA 277:1485–1492

Devereaux PJ, Anderson DR, Gardner MJ, Putnam W, Flowerdew GJ, Brownell BF, Nagpal S, Cox JL (2001) Differences between perspectives of physicians and patients on anticoagulation in patients with atrial fibrillation: observational study. BMJ 323(7323):1218–1222

Deyo R, Diehl A, Rosenthal M (1986) How many days of bed rest for acute low back pain? a randomized clinical trial. N Engl J Med 315:1064–1070

Diamond GA (1999) The wizard of odds: Bayes theorem and diagnostic testing. Mayo Clin Proc 74(11):1179–1182

Diamond GA, Forrester JS (1979) Analysis of probability as an aid in the clinical diagnosis of coronary-artery disease. N Engl J Med 300(24):1350–1358

Diamond GA, Forrester JS, Hirsch M, Staniloff HM, Vas R, Berman DS, Swan HJ (1980) Application of conditional probability analysis to the clinical diagnosis of coronary artery disease. J Clin Invest 65(5):1210–1221

Diamond GA, Hirsch M, Forrester JS, Staniloff HM, Vas R, Halpern SW, Swan HJ (1981) Application of information theory to clinical diagnostic testing: the electrocardiographic stress test. Circulation 63:915–921

Dolan JG, Bordley DR, Mushlim AI (1986) An evaluation of clinicians' subjective prior probability estimates. Med Decis Making 6:216–223

Edwards A, Elwyn G (1999) How should effectiveness of risk communication to aid patients' decisions be judged? A review of the literature.[comment]. Med Dec Making 19(4):428–434

Edwards A, Elwyn G (2001) Evidence-based patient choice. Inevitable or impossible? Oxford University Press, New York

Edwards A, Elwyn G, Stott N (1999) Communicating risk reductions. Researchers should present results with both relative and absolute risks. BMJ 318(7183):603; author reply 603–604

Egger M, Davey Smith G, Altman D (2000) Meta-analysis in context. In: Egger M, Davey Smith G, Altman DG (eds) Systematic reviews in health care. BMJ Books, London

Emanuel E, Emanuel L (1992) Four models of the physician-patient relationship. JAMA 267:2221–2226

Evidence-Based-Medicine-Working-Group (1992) Evidence-based medicine. A new approach to teaching the practice of medicine. JAMA 268:2420–2425

Fagan T (1975) Nomogram for Bayes theorem. N Engl J Med 293:257

Feinstein A (1968) Clinical epidemiology I. The population experiments of nature and of man in human illness. Ann Intern Med 69:807–820

Ferrini R, Woolf SH (1998) American College of Preventive Medicine Practice Policy: screening for prostate cancer in American men. Am J Prev Med 15:81–84

Fletcher RH, Fletcher SW, Wagner EH (1996) Clinical epidemiology: The essentials, 3rd edn. Williams and Wilkins, Baltimore

Freemantle N, Mason J, Eccles M (1999) Deriving Treatment Recommendations from Evidence within Randomized Trials: The Role and Limitation of Meta-Analysis. Int J Technol Assess Health Care 15(2):304–315

Freemantle N, Calvert M, Wood J, Eastaugh J, Griffin C (2003) Composite outcomes in randomized trials: greater precision but with greater uncertainty? JAMA 289:2554–2559

Frost M, Bonomi A, Capelleri J, Schünemann H, Moynihan T, Aaronson N (2003) Applying quality of life data formally and systematically into clinical practice. Clin Ther 25:D10

García Rodríguez L, Hernández-Díaz S, deAbajo FJ (2001) Association between aspirin and upper gastrointestinal complications: Systematic review of epidemiologic studies. Br J Clin Pharmacol 52:563–571

Gaspoz JM, Coxson PG, Goldman PA, Williams LW, Kuntz KM, Hunink MG, Goldman L (2002) Cost effectiveness of aspirin, clopidogrel, or both for secondary prevention of coronary heart disease. N Engl J Med 346(23):1800–1806

Glass RD (1996) Diagnosis: A brief introduction. Oxford University Press, Melbourne

Goldstein R, Gort E, Guyatt GH, Stubbing D, Avendano M (1994) Prospective randomized controlled trial of respiratory rehabilitation. Lancet 344:1394–1397

Greenfield S, Kaplan SH, Ware JE, Jr., Yano EM, Frank HJ (1988) Patients' participation in medical care: effects on blood sugar control and quality of life in diabetes. J Gen Intern Med 3(5):448–457

Guyatt GH (1991) Evidence-based medicine. ACP J Club 114:A-16

Guyatt GH (2002a) Introduction. In: Guyatt GH, Rennie D (eds) Users' guide to the medical literature: A manual for evidence-based clinical practice. AMA Press, Chicago

Guyatt GH (2002b) Preface. In: Guyatt GH, Rennie D (eds) Users' guide to the medical literature: A manual for evidence-based clinical practice. AMA Press, Chicago

Guyatt GH, Rennie D (2002) Users' guide to the medical literature: A manual for evidence-based clinical practice. AMA Press, Chicago

Guyatt GH, Bombardier C, Tugwell P (1986) Measuring disease-specific quality-of-life in clinical trials. CMAJ 134:889–895

Guyatt GH, Walter S, Norman G (1987) Measuring change over time: assessing the usefulness of evaluative instruments. J Chronic Dis 40:171–178

Guyatt GH, Zanten SVV, Feeny D, Patrick D (1989) Measuring quality of life in clinical trials: a taxonomy and review. CMAJ 140:1441–1447

Guyatt GH, Patterson C, Ali M, Singer J, Levine M, Turpie I, Meyer R (1990) Diagnosis of iron-deficiency anemia in the elderly. Am J Med 88(3):205–209

Guyatt GH, Oxman AD, Ali M, Willan A, McIlroy W, Patterson C (1992) Laboratory diagnosis of iron-deficiency anemia: an overview. [erratum appears in J Gen Intern Med 1992 Jul-Aug;7(4):423]. J Gen Intern Med 7(2):145–153

Guyatt GH, Sackett D, Sinclair JC, Hayward R, Cook DJ, Cook RJ (1995) Users' guides to the medical literature IX. A method for grading health care recommendations. Evidence-based medicine working group. JAMA 274:1800–1804

Guyatt GH, King DR, Feeny DH, Stubbing D, Goldstein RS (1999) Generic and specific measurement of health-related quality of life in a clinical trial of respiratory rehabilitation. J Clin Epidemiol 52:187–192

Guyatt GH, Meade MO, Jaeschke RZ, Cook DJ, Haynes RB (2000) Practitioners of evidence based care. Not all clinicians need to appraise evidence from scratch but all need some skills. BMJ 320(7240):954–955

Guyatt GH, Schünemann H, Cook D, Pauker S, Sinclair J, Bucher H, Jaeschke R (2001) Grades of Recommendation for Antithrombotic Agents. Chest 119:3S-7S

Guyatt GH, Sinclair J, Cook D, Jaeschke R, Schünemann H (2002a) Moving from evidence to action. Grading recommendations – a qualitative approach. In: Guyatt GH, Rennie D (eds) Users' guides to the medical literature: A manual for evidence-based clinical practice. AMA Press, Chicago

Guyatt GH, Hayward RS, Richardson WS, Green I, Wilson MC, Sinclair J, Cook D, Glasziou P, Detsky A, Bass E (2002b) Moving from evidence to action. In: Guyatt GH, Rennie D (eds) Users' guide to the medical literature: A manual for evidence-based clinical practice. AMA Press, Chicago

Guyatt GH, Devereaux P, Montori V, Schünemann H, Bhandari M (2004a) Putting the patient first: In our practice, and in our use of language. Evidence Based Medicine 9:6–7

Guyatt GH, Schünemann H, Cook D, Jaeschke R, Pauker S (2004b) Applying the grades of recommendation for antithrombotic and thrombolytic agents. Chest (in press)

Harbour R, Miller J (2001) A new system for grading recommendations in evidence based guidelines. BMJ 323:334–336

Harris R, Helfand M, Woolf SH, Lohr KN, Mulrow CD, Teutsch SM, Atkins D, for the Methods Work Group, Third U.S. Preventive Services Task Force (2001) Current methods of the U.S. Preventive Services Task Force: a review of the process. Am J Prev Med 20(3S):21–35

Haynes RB, Devereaux PJ, Guyatt GH (2002) Clinical expertise in the era of evidence-based medicine and patient choice. ACP J Club 136: A11

Hernandez JL, Riancho JA, Matorras P, Gonzalez-Macias J (2003) Clinical evaluation for cancer in patients with involuntary weight loss without specific symptoms. Am J Med 114(8):631–637

Hill SA, Devereaux PJ, Griffith L, Opie J, McQueen MJ, Panju A, Stanton E, Guyatt GH (2003) Can Troponin I measurement predict short term serious cardiac outcomes in patients presenting to the emergency department with possible acute coronary syndrome. Can J Emerg Med (in press)

Holmes-Rovner M, Llewellyn-Thomas H, Entwistle V, Coulter A, O'Connor A, Rovner DR (2001) Patient choice modules for summaries of clinical effectiveness: a proposal. BMJ 322(7287):664–667

http://www.guidelines.gov/ Accessed March 14, 2004

Hunt S, McKenna S, McEwen J, Backett E, Williams J, Papp E (1980) A quantitative approach to perceived health status: a validation study. J Epidemiol Community Health 34:281–286

Jaeschke R, Singer J, Guyatt GH (1991) Using quality-of-life measures to elucidate mechanism of action. CMAJ 144:35–39

Ladenheim ML, Kotler TS, Pollock BH, Berman DS, Diamond GA (1987) Incremental prognostic power of clinical history, exercise electrocardiography and myocardial perfusion scintigraphy in suspected coronary artery disease. Am J Cardiol 59(4):270–277

Laupacis A, Sackett D, Roberts RS (1988) An assessment of clinically useful measures of the consequences of treatment. N Engl J Med 318:1728–1733

Laupacis A, Wong C, Churchill D (1991) The use of generic and specific quality-of-life measures in hemodialysis patients treated with erythropoietin. Control Clin Trials 12:168S-179S

Ledley RS, Lusted LB (1959) Reasoning foundations of medical diagnosis: symbolic logic, probability, and value theory aid our understanding of how physicians reason. Science 130:9–21

Levine MN, Gafni A, Markham B, MacFarlane D (1992) A bedside decision instrument to elicit a patient's preference concerning adjuvant chemotherapy for breast cancer.[comment]. Ann Intern Med 117(1):53–58

Liang M, Larson M, Cullen K, Schwartz J (1985) Comparative measurement efficiency and sensitivity of five health status instruments for arthritis research. Arthritis Rheum 28:542–547

Lubsen J, Kirwan B (2002) Combined endpoints: can we use them? Stat Med 21:2959–2970

MacKenzie M, Charlson M (1986) Standards for the use of ordinal scales in clinical trials. BMJ 292:40–43

Mahler D, Rosiello R, Harver A, Lentine T, McGovern J, Daubenspeck J (1987) Comparison of clinical dyspnea ratings and psychophysical measurements of respiratory sensation in obstructive airway disease. Am Rev Respir Dis 135:1229–1233

McGinn TG, Guyatt GH, Wyer PC, Naylor CD, Stiell IG (2003) Clinical prediction rules. In: Guyatt GH, Rennie D (eds) Users' guides to the medical literature: A manual for evidence-based clinical practice. AMA Press, Chicago

McKibbon A, Hunt D, Richardson SW, Hayward R, Wilson M, Jaeschke R, Haynes RB, Wyer P, Craig J, Guyatt GH (2002) Finding the evidence. In: Guyatt GH, Rennie D (eds) Users' guides to the medical literature: A manual for evidence-based clinical practice. AMA Press, Chicago

Meier MA, Al-Badr WH, Cooper JV, Kline-Rogers EM, Smith DE, Eagle KA, Mehta RH (2002) The new definition of myocardial infarction: diagnostic and prognostic implications in patients with acute coronary syndromes. Arch Intern Med 162(14):1585–1589

Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, Tugwell P, Klassen TP (1998) Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet 352(9128):609–613

National-Guideline-Clearing-House (2004) (http://www.guideline.gov/) Accessed March 14, 2004

CNEP (2001) Executive summary of the third report of the National Cholesterol Education Program (NCEP) Expert Panel on detection, evaluation, and treatment of high blood cholesterol in adults (Adult Treatment Panel III). JAMA 285(19):2486–2497

O'Connor A (2001) Using patient decision aids to promote evidence-based decision making. ACP J Club 135(1):A11–A12

O'Connor AM, Legare F, Stacey D (2003) Risk communication in practice: the contribution of decision aids. BMJ 327:736–740

O'Connor AM, Stacey D, Entwistle V, Llewellyn-Thomas H, Rovner D, Holmes-Rovner M, Tait V, Tetroe J, Fiset V, Barry M, Jones J (2004) Decision aids for people facing health treatment or screening decisions. Cochrane Database of Systematic Reviews 1

Olsson G, Lubsen J, VanEs G, Rehnqvist N (1986) Quality-of-life after myocardial infarction: effect of long term metoprolol on mortality and morbidity. BMJ 292:1491–1493

Osoba D, Aaronson N, Till J (1991) A practical guide for selecting quality-of-life measures in clinical trials and practice. In: Osoba D (ed) Effect of cancer on quality-of-life. CRC Press, Boston

Ott C, Sivarajan E, Newton K, Almes MJ, Bruce RA, Bergner M, Gilson BS (1983) A controlled randomized study of early cardiac rehabilitation: the sickness impact profile as an assessment tool. Heart Lung 12:162–170

Oxman A, Guyatt GH (1993) The science of reviewing research. Ann NY Acad Sci 703:125–133:discussion 133–134

Oxman A, Guyatt GH, Cook D, Montori V (2002) Summarizing the evidence. In: Guyatt GH, Rennie D (eds) Users' guide to the medical literature: A manual for evidence-based clinical practice. AMA Press, Chicago

Oxman AD, Sackett DL, Guyatt GH (1993) Users' guides to the medical literature. I. How to get started. The Evidence-Based Medicine Working Group. JAMA 270(17):2093–2095

Parkerson G, Gehlback S, Wagner E, Sherman A, Clapp N, Muhlbaier L (1981) The Duke-UNC Health Profile: an adult health status instrument for primary care. Med Care 19:806–828

Paul J (1938) Clinical epidemiology. J Clin Invest 17:539–541

Petitti D (1994) Meta-analysis, decision analysis and cost-effectiveness analysis. Oxford University Press, Oxford

PIOPED-Investigators (1990) Value of ventilation/perfusion scan in acute pulmonary embolism. Results of the Prospective Investigation of Pulmonary Embolism (PIOPED). JAMA 263:2753–2759

Protheroe J, Fahey T, Montgomery AA, Peters TJ (2000) The impact of patients' preferences on the treatment of atrial fibrillation: observational study of patient based decision analysis. BMJ 320(7246):1380–1384

Richardson WS (2002) Five uneasy pieces about pre-test probability. J Gen Intern Med 17(11):882–883

Richardson WS, Wilson MC, Nishikawa J, Hayward RS (1995) The well-built clinical question: a key to evidence-based decisions. ACP J Club 123(3):A12-A13

Richardson WS, Wilson MC, Williams JW, Jr., Moyer VA, Naylor CD (2003) Clinical Manifestation of Disease. In: Guyatt GH, Rennie D (eds) Users' guide to the medical literature: A manual for evidence-based clinical practice. AMA Press, Chicago

Sackett D (1969) Clinical epidemiology. Am J Epidemiol 89:125–128

Sackett D (1986) Rules of evidence and clinical recommendations on the use of antithrombotic agents. Arch Int Med 146:464–465

Sackett D (2002) Clinical Epidemiology: what, who, and whither. J Clin Epidemiol 55:1161–1166

Sackett D, Winkelstein W (1967) The relationship between cigarette usage and aortic atherosclerosis. Am J Epidemiol 86:264–270

Sackett D, Chambers L, MacPherson A, Goldsmith C, McAuley R (1977) The development and application of indices of health: general methods and a summary of results. Am J Public Health 67:423–428

Sackett D, Haynes RB, Guyatt GH, Tugwell P (1991) Clinical epidemiology: a basic science for clinical medicine. Little, Brown, Boston, pp 5–17

Sackett D, Straus S, Richardson W, Rosenberg W, Haynes RB (2000) Evidence-based medicine. Churchill Livingston, Toronto, p 166

Schünemann HJ, Best D, Vist G, Oxman AD, for The GRADE Working Group (2003) Letters, numbers, symbols and words: how to communicate grades of evidence and recommendations. CMAJ 169(7):677–680

Schünemann H, Munger H, Brower S, O'Donnell M, Crowther M, Cook D, Guyatt GH (2004) Developing evidence-based guidelines for the ACCP Conference on Antithrombotic and Thrombolytic Therapy. Chest in press

Shekelle PG, Woolf SH, Eccles M, Grimshaw J (1999) Clinical guidelines: developing guidelines. BMJ 318(7183):593–596

Smith D, Baker G, Davies G, Dewey M, Chadwick. D (1993) Outcomes of add-on treatment with Lamotrigine in partial epilepsy. Epilepsia 34:312–322

Smith R, Cokkinides V, Eyre H (2003) American Cancer Society guidelines for the early detection of cancer, 2003. CA Cancer J Clin 53:27–43

Soteriades ES, Evans JC, Larson MG, Chen MH, Chen L, Benjamin EJ, Levy D (2002) Incidence and Prognosis of Syncope. N Engl J Med 347(12):878–885

Sox HC, Blatt MA, Higgins MC, Marton KL (1988) Medical decision making. Butterworths, Boston

Stewart M (1995) Effective physician-patient communication and health outcomes: a review. CMAJ 152:1423–1433

Stiggelbout A, Kiebert G (1997) A role for the sick role. Patient preferences regarding information and participation in clinical decision-making. CMAJ 157:383–389

Strull W, Lo B, Charles G (1984) Do patients want to participate in medical decision making? JAMA 252:2990–2994

Symons MJ, Moore DT (2002) Hazard rate ratio and prospective epidemiological studies. J Clin Epidemiol 55(9):893–899

Szabo E, Moody H, Hamilton T, Ang C, Kovithavongs C, Kjellstrand C (1997) Choice of treatment improves quality of life. A study on patients undergoing dialysis. Arch Intern Med 157:1352–1356

Tandon P, Stander H, Jr. RS (1989) Analysis of quality of life data from a randomized, placebo controlled heart-failure trial. J Clin Epidemiol 42:955–962

Tarlov A, Ware J, Greenfield S, Nelson E, Perrin E, Zubkoff M (1989) The medical outcomes study. JAMA 262:925–930

The CAPRICORN Investigators (2001) Effects of carvedilol on outcome after myocardial infarction in patients with left ventricular dysfunction: the CAPRICORN randomised trial. Lancet 357:1385–1390

Thomson R, Parkin D, Eccles M, Sudlow M, Robinson A (2000) Decision analysis and guidelines for anticoagulant therapy to prevent stroke in patients with atrial fibrillation. Lancet 355:956–962

Torrance G (1986) Measurement of health state utilities for economic appraisal. J Health Econ:1–30

Tugwell P, Bombardier C, Buchanan W, Goldsmith C, Grace E, Bennett K, Williams J, Egger M, Alarcon GS, Guttadauria M (1990) Methotrexate in Rheumatoid Arthritis. Impact on quality of life assessed by traditional standard-item and individualized patient preference health status questionnaires. Arch Intern Med 150:59–62

USPSTF (2002) Screening for prostate cancer: recommendations and rationale. Ann Intern Med 137:915–916

van Walraven C, Hart R, Singer D (2002) Oral anticoagulants vs aspirin in nonvalvular atrial fibrillation: an individual patient meta-analysis. JAMA 288:2441–2448

Ware J, Kozinski M, Bayliss M, McHorney CA, Rogers WH, Raczak A (1995) Comparison of methods for the scoring and statistical analysis of SF-36 health profile and summary measures: Summary of results of the Medical Outcomes Study. Med Care 33:AS264-AS279

Ware J, Sherbourne C (1992) The MOS 36-item short-form health survey (SF-36). Med Care 30:473–483

Weil J, Colin-Jones D, Langman M, Lawson D. Logan R, Murphy M, Rawlins M, VesseyM, Wainright P (1995) rophylactic aspirin and risk of peptic ulcer bleeding. BMJ 310:827–830

Whelan T, Sawka C, Levine M, Gafni A, Reyno L, Willan A, Julian J, Dent S, Abu-Zahra H, Chouinard E, Tozer R, Pritchard K, Bodendorfer I (2003) Helping patients make informed choices: a randomized trial of a decision aid for adjuvant chemotherapy in lymph node-negative breast cancer.[comment]. J Nat Cancer Inst 95(8):581–587

Wiebe S, Guyatt GH, Weaver B, Matijevic S, Sidwell C (2003) Comparative responsiveness of generic and specific quality-of-life instruments. J Clin Epidemiol 56(1):52–60

Weiss N (1996) Clinical epidemiology: the study of the outcome of illness, 2nd edn. Oxford University Press, Oxford

Woolf SH (1994) An organized analytic framework for practice guideline development: using the analytic logic as a guide for reviewing evidence, developing recommendations, and explaining the rationale. In: McCormick KA, Moore S, Siegel RA (eds) Methodology perspectives. US Department of Health and Human Services, Agency for Health Care Policy and Research, Washington, DC

Woolf SH, Grol R, Hutchinson A, Eccles M, Grimshaw J (1999) Clinical guidelines: potential benefits, limitations, and harms of clinical guidelines. BMJ 318(7182):527–530