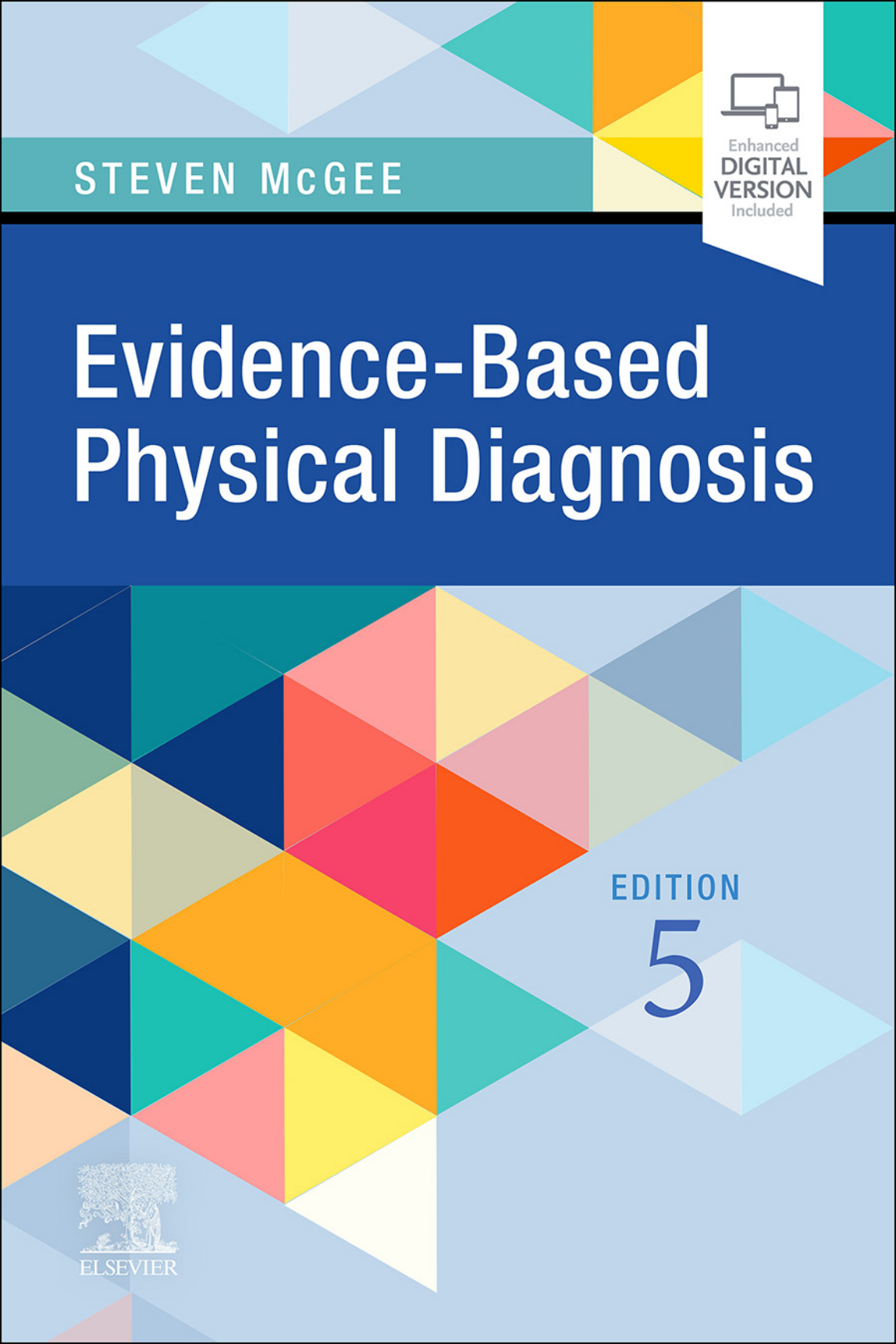

STEVEN McGEE
Enhanced
DIGITAL
VERSION
Included
Evidence-Based
Physical Diagnosis
EDITION
5
ELSEVIER

eBooks for Practicing Clinicians

ELSEVIER

Any screen. Any time. Anywhere.

Activate the eBook version
of this title at no additional charge.

Elsevier eBooks for Practicing Clinicians gives you the power to browse and search content, view enhanced images, highlight and take notes—both online and offline.

Unlock your eBook today.

1. Visit **expertconsult.inkling.com/redeem**
2. Scratch box below to reveal your code
3. Type code into "Enter Code" box
4. Click "Redeem"
5. Log in or Sign up
6. Go to "My Library"

It's that easy!

Place Peel Off Sticker Here

For technical assistance:
email expertconsult.help@elsevier.com
call 1-800-401-9962 (inside the US)
call +1-314-447-8300 (outside the US)

2020_PC

Evidence-Based Physical Diagnosis

This page intentionally left blank

EDITION 5

Evidence-Based Physical Diagnosis

STEVEN McGEE, M.D.
Professor Emeritus, Medicine
University of Washington School of Medicine

ELSEVIER

Elsevier
1600 John F. Kennedy Blvd.
Ste 1800
Philadelphia, PA 19103-2899

EVIDENCE-BASED PHYSICAL DIAGNOSIS,
FIFTH EDITION ISBN: 9780323754835

Library of Congress Control Number: 2021941821

Content Strategist: Charlotta Kryhl
Content Development Specialist: Kathryn DeFrancesco
Publishing Services Manager: Deepthi Unni
Project Manager: Radjan Lourde Selvanadin
Design Direction: Renee Duenow

Printed in the United States of America.
Last digit is the print number: 9 8 7 6 5 4 3 2 1

To Rosalie, Connor, and Matt

PREFACE TO THE FIFTH EDITION

Since the appearance of the fourth edition of this work, hundreds of new studies on physical diagnosis have appeared, and research has further defined how physical findings help identify disease, solve clinical problems, and forecast patient outcomes. In this fifth edition of *Evidence-Based Physical Diagnosis*, 203 of these new studies have been added to the *Evidence-Based Medicine (EBM) Boxes*, thus providing the most up-to-date summary of physical examination and its diagnostic accuracy. New references have replaced outworn ones, although classic articles and original descriptions remain. Every chapter has been updated. Many findings not previously addressed in prior editions appear here, along with their diagnostic accuracy and likelihood ratios. These findings include the Romberg test (spinal stenosis); oximeter paradoxus (cardiac tamponade); platypnea (liver disease); relative afferent pupil defect (unilateral visual loss); pupil size in red eye (acute glaucoma); hum test (hearing loss); jolt accentuation headache (meningitis); nasal flaring, intercostal retractions, and suprasternal retractions (dyspnea); diastolic rumble (mitral stenosis); bendopnea test (congestive heart failure); hand vein estimates of central venous pressure; adenopathy (infectious mononucleosis); abnormal bowel tones (recovery from postoperative ileus); Ipswich touch test (diabetic foot); the Edinburgh algorithm for diplopia; and the bedside diagnosis of orbital fractures. As in the last edition, each chapter begins with a list of *Key Teaching Points*, intended for individuals desiring quick summaries and for teachers constructing concise bedside lessons. Each EBM Box remains linked to the Elsevier *online EBM* calculator to quickly estimate post-test probability using the likelihood ratios in each chapter.

I am grateful to many investigators who supplied me with unpublished information from their original work: Bastiaan Bloem, Jorik Nonnekes, and W.F. Abdo (tandem gait in atypical parkinsonism); Loris Bonetti, Ivana Maria Rosi, Rossella Guastaferro, Alessandra Cerra, Roberto Milos, and Enrico Messina (dehydration in the elderly); Takashi Matono (relative bradycardia in enteric fever); Mark Wright (unilateral visual loss and Marcus Gunn pupil); Alex Butskiy (Rinne test, conductive hearing loss; orientation of tines); William Strawbridge (hearing tests); Vincent Quagliarello (meningeal signs); Johann Steurer (pneumonia); Jason Weatherald (diagnosis of pulmonary hypertension); Gerben ter Riet (red eye); Anikar Chhabra, Keith Jarbo, David E. Hartigan, Kelly Scott, and Karan Patel (tests of anterior cruciate ligament tears); and Kenneth Arinze Ohagwu and Innocent Ijezie Chukwuonye (Siriraj Stroke scale). I greatly appreciate their promptness in responding to my questions and their generosity in sharing data from their research.

Insights into physical signs continue to evolve and progress. This textbook presents the most recent evidence supporting this fundamental clinical skill. By applying this evidence-based approach, clinicians will glean the most from what they hear, see, and feel at the bedside, information that, combined with modern technological testing, will grant clinicians the keys to outstanding patient care.

Steven McGee, M.D.
February 2021

INTRODUCTION TO THE FIRST EDITION

The purpose of this book is to explore the origins, pathophysiology, and diagnostic accuracy of many of the physical signs used today in adult patients. We have a wonderfully rich tradition of physical diagnosis, and my hope is that this book will help square this tradition, now almost two centuries old, with the realities of modern diagnosis, which often rely more on technologic tests such as clinical imaging and laboratory testing. The tension between physical diagnosis and technologic tests has never been greater. Having taught physical diagnosis for 20 years, I frequently observe medical students purchasing textbooks of physical diagnosis during their preclinical years to study and master traditional physical signs, but then neglecting or even discarding this knowledge during their clinical years, after observing that modern diagnosis often takes place at a distance from the bedside. One can hardly fault a student who, caring for a patient with pneumonia, does not talk seriously about crackles and diminished breath sounds when all of his or her teachers are focused on the subtleties of the patient's chest radiograph. Disregard for physical diagnosis also pervades our residency programs, most of which have formal X-ray rounds, pathology rounds, microbiology rounds, and clinical conferences addressing the nuances of laboratory tests. Very few have formal physical diagnosis rounds.

Reconciling traditional physical diagnosis with contemporary diagnostic standards has been a continuous process throughout the history of physical diagnosis. In the 1830s, Professor Pierre Adolphe Piorry, the inventor of topographic percussion taught that there were nine distinct percussion sounds, which he used to outline the patient's liver, heart, lungs, stomach, and even individual heart chambers or lung cavities. Piorry's methods flourished for over a century and once filled 200-page manuals,[1] although today, thanks to the introduction of clinical imaging in the early 1900s, the only vestige of his methods is percussion of the liver span. In his 1819 *A Treatise on Diseases of the Chest*,[2] Laennec wrote that lung auscultation could detect "every possible case" of pneumonia. It was only a matter of 20 years before other careful physical diagnosticians tempered Laennec's enthusiasm and pointed out that the stethoscope had diagnostic limitations.[3] For most of the 20th century, expert clinicians believed that all late systolic murmurs were benign, until Barlow in 1963 showed that they often represented mitral regurgitation, sometimes of significant severity.[4]

There are two contemporary polar opinions regarding physical diagnosis. Holding the less common position are clinicians who believe that all traditional physical signs remain accurate, and these clinicians continue to quiz students about Krönig's isthmus and splenic percussion signs. A more common position is that physical diagnosis has little to offer the modern clinician and that traditional signs, though interesting, cannot compete with the accuracy of our more technologic diagnostic tools. Of course, neither position is completely correct. I hope this book, by examining the best evidence comparing physical signs to current diagnostic standards, will bring clinicians to a more appropriate middle ground: that physical diagnosis is a reliable diagnostic tool that can still help clinicians with many, but not all, clinical problems.

Although some regard evidence-based medicine as "cookbook medicine," this is incorrect, because there are immeasurable subtleties in our interaction with patients that clinical studies cannot address (at least, not as yet) and because the diagnostic power of any physical sign (or any test, for that matter) depends in part on our ideas about disease prevalence, which in turn depend on our own personal interviewing skills and clinical experience.* Instead, evidence-based physical diagnosis simply summarizes the best evidence available on whether a physical sign is accurate

*These subjects are discussed fully in Chapters 2 and 5.

or not. The clinician who understands this evidence can then approach his or her own patients with the confidence and wisdom that would have developed had he or she personally examined and learned from the thousands of patients reviewed in the studies of this book.

Sometimes, comparing physical signs with modern diagnostic standards reveals that the physical sign is outdated and perhaps best discarded (e.g., topographic percussion of diaphragm excursion). Other times the comparison reveals that physical signs are extremely accurate and probably underused (e.g., early diastolic murmur at the left lower sternal area for aortic regurgitation, conjunctival rim pallor for anemia, or a palpable gallbladder for extrahepatic obstruction of the biliary ducts). At other times, the comparison reveals that the physical sign *is* the diagnostic standard, just as most of the physical examination was a century ago (e.g., systolic murmur and click of mitral valve prolapse, hemiparesis for stroke, neovascularization for proliferative diabetic retinopathy). For some diagnoses, conflict remains between physical signs and technologic tests, making it still unclear which should be the diagnostic standard (e.g., the diagnoses of cardiac tamponade and carpal tunnel syndrome). And for still others, the comparison is impossible because clinical studies comparing physical signs to traditional diagnostic standards do not exist.

My hope is that the material in this book will allow clinicians at all levels—students, house officers, and seasoned clinicians alike—to examine patients more confidently and accurately, thus restoring physical diagnosis to its appropriate, and often pivotal, diagnostic role. Once well-versed in evidence-based physical diagnosis, clinicians can then settle the most important clinical questions at the time and place they should be first addressed—the patient's bedside.

Steven McGee, M.D.
July 2000

References

1. Weil A. *Handbuch und Atlas der topographischen Perkussion*. Leipzig: F.C.W. Vogel; 1880.
2. Laennec RTH. *A Treatise on the Diseases of the Chest (Facsimile Edition by Classics of Medicine Library)*. London: T. and G. Underwood; 1821.
3. Addison T. The difficulties and fallacies attending physical diagnosis of diseases of the chest. In: Wilks, Daldy, ed. *A Collection of the Published Writings of the Late Thomas Addison (Facsimile Edition by Classics of Medicine Library)*. London: The New Sydenham society; 1846:242.
4. Barlow JB, Pocock WA, Marchand P, Denny M. The significance of late systolic murmurs. *Am. Heart J.* 1963;66(4):443–452.

CONTENTS

PART 1

Introduction

This page intentionally left blank

CHAPTER 1

What is Evidence-Based Physical Diagnosis?

When clinicians diagnose disease, their intent is to place the patient's experience into a particular category (or diagnosis), a process implying specific pathogenesis, prognosis, and treatment. This procedure allows clinicians to explain to patients what is happening and to identify the best way to restore the patient's health. A century ago, such categorization of disease rested almost entirely on empiric observation—what clinicians saw, heard, and felt at the patient's bedside. Although some technologic testing was available then (e.g., microscopic examination of sputum and urine), its role in diagnosis was meager, and almost all diagnoses were based on traditional examination (Fig. 1.1). For example, if patients presented a century ago with complaints of fever and cough, the diagnosis of lobar pneumonia rested on the presence of the characteristic findings of pneumonia—fever, tachycardia, tachypnea, grunting respirations, cyanosis, diminished excursion of the affected side, dullness to percussion, increased tactile fremitus, diminished breath sounds (and later bronchial breath sounds), abnormalities of vocal resonance (bronchophony, pectoriloquy, and egophony), and crackles. If these findings were absent, the patient did not have pneumonia. Chest radiography played no role in diagnosis because it was not widely available until the early 1900s.

Modern medicine, of course, relies on technology much more than medicine did a century ago (to our patients' advantage), and for many modern categories of disease the diagnostic standard is a technologic test (Fig. 1.1). For example, if patients present today with fever and cough, the diagnosis of pneumonia is based on the presence of an infiltrate on the chest radiograph. Similarly, the diagnosis of systolic murmurs depends on echocardiography and that of ascites on abdominal ultrasonography. In these disorders, the clinician's principal interest is the result of the technologic test, and decisions about treatment depend much more on that result than on whether the patient has egophony, radiation of the murmur into the neck, or shifting dullness. This reliance on technology creates tension for medical students, who spend hours mastering the traditional examination yet later learn (when first appearing on hospital wards) that the traditional examination pales in importance compared to technology, a realization prompting a fundamental question: What is the diagnostic value of the traditional physical examination? Is it outdated and best discarded? Is it completely accurate and underutilized? Is the truth somewhere between these two extremes?

Examination of Fig. 1.1 indicates that diagnosis today is split into two parts. For some categories of disease, the diagnostic standard still remains empiric observation— what the clinician sees, hears, and feels—just as it was for all diagnosis a century ago. For example, how does a clinician know the patient has cellulitis? The only way is to go to the patient's bedside and observe fever and localized bright erythema, warmth, swelling, and tenderness on the leg. There is no other way to make this diagnosis (technologic or not). Similarly, there is no technologic standard for Parkinson disease (during the patient's life), Bell palsy, or pericarditis. All of these diagnoses—and many

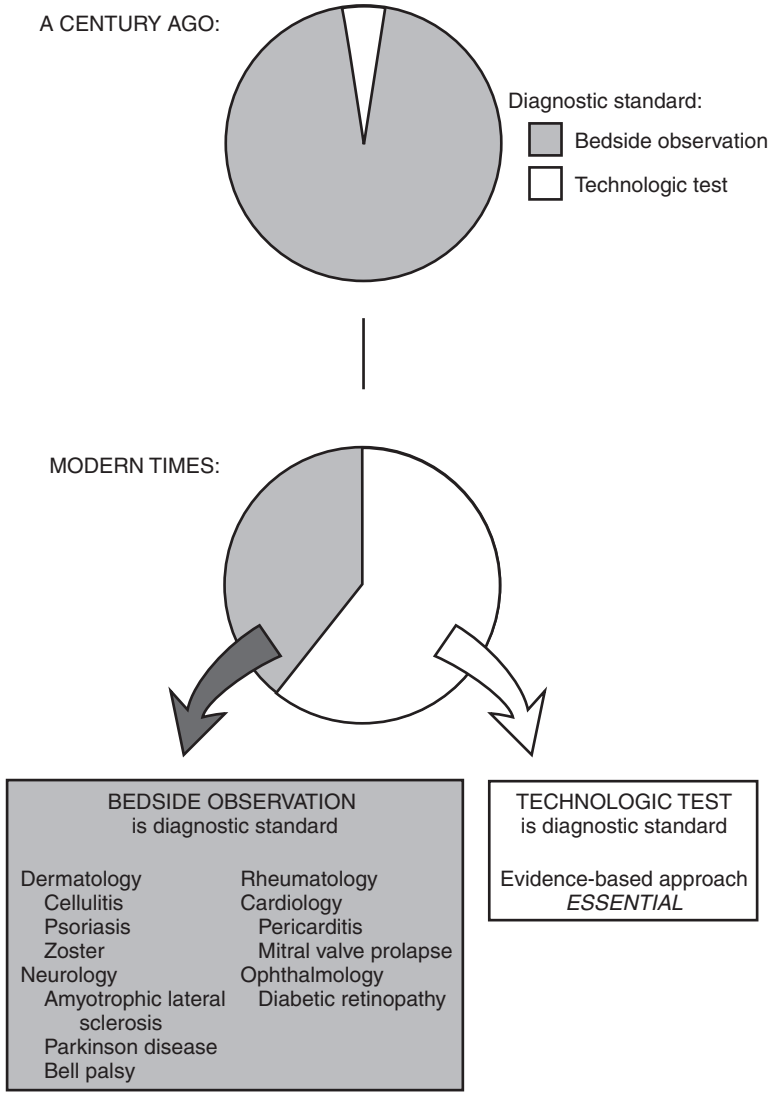


Fig. 1.1 Evolution of Diagnostic Standard. The figure compares the diagnostic process one century ago (*top*, before introduction of clinical imaging and modern laboratory testing) to modern times (*bottom*), illustrating the relative contributions of bedside examination (*grey shade*) and technologic tests (*white shade*) to the diagnostic standard. One century ago, most diagnoses were defined by bedside observation, whereas today technologic standards have a much greater diagnostic role. Nonetheless, there are many examples today of diagnoses based solely on bedside findings (examples appear in *large grey shaded box*). Evidence-based physical diagnosis, on the other hand, principally addresses those diagnoses *defined by technologic standards*, because it identifies those traditional findings that accurately predict the result of the technologic test, as discussed throughout the book.

others in the fields of dermatology, neurology, musculoskeletal medicine, and ophthalmology—are based entirely on empiric observation by experienced clinicians; technology has a subordinate diagnostic role. In fact, the principal reasons medical students still must study and master the traditional examination is the dependence of many diagnoses on bedside findings.

The principal role of evidence-based physical examination, in contrast, is the second category of diseases—that is, those whose categorization today is based on technologic studies. Clinicians want to know the results of the chest radiograph when diagnosing pneumonia, the echocardiogram when diagnosing systolic murmurs, and the ultrasound when diagnosing ascites. For each of these problems, the evidence-based approach compares traditional findings to the technologic standard and then identifies those findings that increase or decrease probability of disease (as defined by the technologic standard), distinguishing them from unhelpful findings that fail to change probability. Using this approach, the clinician will calculate the Heckerling score* to predict the findings on the chest radiograph (Chapter 32), define the topographic distribution of the murmur on the chest wall to predict the findings on the echocardiogram (Chapter 43), and look for a fluid wave or edema to predict the findings on the abdominal ultrasound examination (Chapter 51).

There are thus two distinct ways physical examination is applied at the bedside. For many disorders—those still lacking a technologic standard—the clinician's observations define diagnosis. For other disorders—those based on technologic tests—the clinician's application of an evidence-based approach quickly identifies the relatively few findings that predict the results of technologic standard. Both approaches to bedside examination make physical examination more efficient and accurate, and ultimately more relevant to the care of patients.

*The Heckerling score assigns one point to each of five independent predictors of pneumonia that are present: temperature > 37.8°C; heart rate > 100/min; crackles; diminished breath sounds; and absence of asthma (see Chapter 32).

This page intentionally left blank

PART 2

Understanding the Evidence

This page intentionally left blank

CHAPTER 2

Diagnostic Accuracy of Physical Findings

KEY TEACHING POINTS

- Likelihood ratios (LRs) are nothing more than *diagnostic weights*, numbers that quickly convey to clinicians how much a physical sign argues for or against disease.
- LRs have possible values between 0 to ∞. Values greater than 1 *increase* probability of disease. (The greater the value of the LR, the greater the increase in probability.) LRs less than 1 *decrease* probability of disease. (The closer the number is to zero, the more the probability of disease decreases.) LRs that equal 1 do not change probability of disease at all.
- LRs of 2, 5, and 10 increase probability of disease about 15%, 30%, and 45%, respectively (in absolute terms). LRs of 0.5, 0.2, and 0.1 (i.e., the reciprocals of 2, 5, and 10) decrease probability 15%, 30%, and 45%, respectively.
- EBM Boxes comparing LRs of different physical signs quickly inform clinicians about which findings have the greatest diagnostic value.

I. Introduction

If a physical sign characteristic of a suspected diagnosis is present (i.e., **positive finding**), that diagnosis becomes more likely; if the characteristic finding is absent (i.e., **negative finding**), the suspected diagnosis becomes less likely. How much these positive and negative results modify probability, however, is distinct for each physical sign. Some findings, when positive, shift probability upward greatly, but they change it little when negative. Other signs are more useful if they are absent, because the negative finding practically excludes disease, although the positive one changes probability very little.

Much of this book consists of EBM Boxes that specifically describe how positive or negative findings change the probability of disease, a property called **diagnostic accuracy**. Understanding these tables first requires review of four concepts: pre-test probability, sensitivity, specificity, and LRs.

II. Pre-Test Probability

Pre-test probability is the probability of disease (i.e., prevalence) before application of the results of a physical finding. Pre-test probability is the starting point for all clinical decisions. For example, the clinician may know that a certain physical finding shifts the probability of disease upward 40%, but this information alone is unhelpful unless the clinician also knows the starting point: if the pre-test probability for the particular diagnosis was 50%, the finding is diagnostic (i.e., post-test probability

TABLE 2.1 **Pre-Test Probability**

Setting (Reference)	Diagnosis	Probability (%)
Acute abdominal pain [1–3]	Small bowel obstruction	4–8
Ankle injury[4,5]	Ankle fracture	10–14
Cough and fever[6]	Pneumonia	15–35
Acute calf pain or swelling[7–19]	Proximal deep vein thrombosis	6–43
Pleuritic chest pain, dyspnea, or hemoptysis[20–31]	Pulmonary embolism	9–43
Diabetic foot ulcer[32–34]	Osteomyelitis	52–68

50% + 40% = 90%); if the pre-test probability was only 10%, the finding is less helpful, because the probability of disease is still akin to a coin toss (i.e., post-test probability 10% + 40% = 50%).

Published estimates of disease prevalence, given a particular clinical setting, are summarized in the Appendix for all the clinical problems discussed in this book (these estimates derive from clinical studies reviewed in all the EBM Boxes); Table 2.1 provides a small sample of these pre-test probabilities. Even so, clinicians must adjust these estimates with information from their own practice. For example, large studies based in emergency departments show that 15% to 35% of patients presenting with cough and fever have pneumonia (Table 2.1). The probability of pneumonia, however, is certainly lower in patients presenting with cough and fever to an office-based practice in the community, and it may be higher if cough and fever develop in patients with cancer or chronic lung disease. In fact, because the best estimate of pre-test probability incorporates information from the clinician's own practice—how specific underlying diseases, risks, and exposures make disease more or less likely—the practice of evidence-based medicine is never "cookbook" medicine, but instead consists of decisions based on the unique characteristics of the patients the clinician sees.

III. Sensitivity and Specificity

A. DEFINITIONS

The terms *sensitivity* and *specificity* describe the discriminatory power of physical signs. **Sensitivity** is the proportion of patients *with* the diagnosis who *have* the physical sign (i.e., have the *positive* result). **Specificity** is the proportion of patients *without* the diagnosis who *lack* the physical sign (i.e., have the *negative* result).

Calculation of sensitivity and specificity requires construction of a 2 × 2 table (Fig. 2.1) that has two columns (one for "diagnosis present" and another for "diagnosis absent") and two rows (one for "physical sign present" and another for "physical sign absent"). These rows and columns create four boxes: one for the "true positives" (cell a, sign and diagnosis present), one for "false positives" (cell b, sign present but disease absent), one for the "false negatives" (cell c, sign absent but disease present), and one for the "true negatives" (cell d, sign and disease absent).

Fig. 2.1 presents data from a hypothetical study of 100 patients presenting with pulmonary hypertension. The clinician knows that tricuspid regurgitation is a complication of pulmonary hypertension and wonders how accurately a single physical sign—the presence of a holosystolic murmur at the left lower sternal border—detects this complication*.[35] Forty-two patients have significant tricuspid regurgitation (the sum of column 1), and 58 patients do not (the sum of column 2). The **sensitivity** of the holosystolic murmur is the proportion of patients with disease

*The numbers used in this example are very close to those in reference 35. See also Chapter 46.

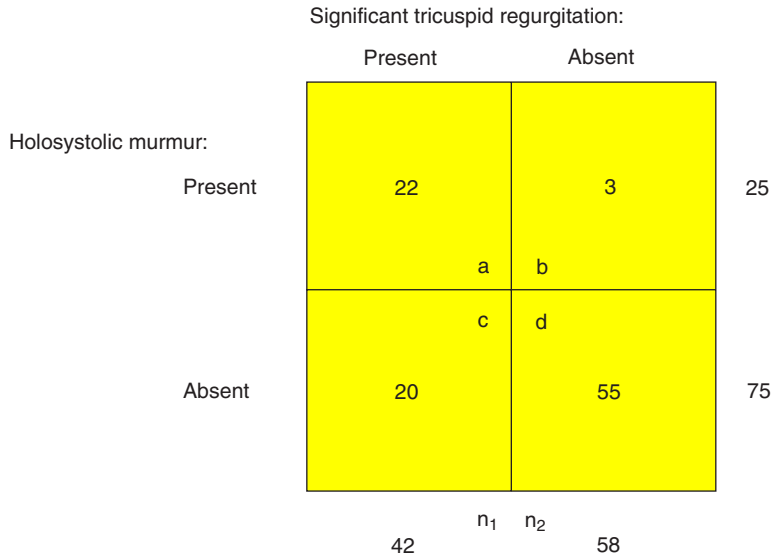


Fig. 2.1 2 × 2 table. The total number of patients with disease (tricuspid regurgitation in this example) is the sum of the first column, or $n_1 = a + c$. The total number of patients without disease is the sum of the second column, or $n_2 = b + d$. The *sensitivity* of a physical finding (holosystolic murmur at the left lower sternal edge, in this example) is the proportion of patients with disease who have the finding [i.e., $a/(a + c)$ or a/n_1]. The *specificity* of a physical finding is the proportion of patients without disease who lack the finding [i.e., $d/(b + d)$ or d/n_2]. The *positive likelihood ratio (LR)* is proportion of patients with disease who have a positive finding (a/n_1) divided by the proportion of patients without disease who have a positive finding (b/n_2), or sensitivity/(1−specificity). The *negative LR* is the proportion of patients with disease who lack the finding (c/n_1) divided by the proportion of patients without disease who lack the finding (d/n_2), or (1−sensitivity)/specificity. In this example, the sensitivity is 0.52 (22/42), the specificity is 0.95 (55/58), the positive LR is 10.1 [(22/42)/(3/58)], and the negative LR is 0.5 [(20/42)/(55/58)].

(i.e., tricuspid regurgitation, 42 patients) who have the characteristic murmur (i.e., the *positive* result, 22 patients), which is 22/42 = 0.52 or 52%. The **specificity** of the holosystolic murmur is the proportion of patients *without* disease (i.e., no tricuspid regurgitation, 58 patients) who *lack* the murmur (i.e., the *negative* result, 55 patients), which is 55/58 = 0.95 or 95%. To recall how to calculate sensitivity and specificity, Sackett and others have suggested helpful mnemonics: Sensitivity is represented as "PID" for "positivity in disease" (an abbreviation normally associated with "pelvic inflammatory disease"), and specificity is represented as "NIH" for "negativity in health" (an abbreviation normally associated with the "National Institutes of Health").[36,37]

B. USING SENSITIVITY AND SPECIFICITY TO DETERMINE PROBABILITY OF DISEASE

The completed 2 × 2 table can be used to determine the accuracy of the holosystolic murmur, which is how well its presence or absence discriminates between those with tricuspid regurgitation and those without it. In Fig. 2.1, the first row includes all 25 patients with the murmur (i.e. the positive results). Of these 25 patients, 22 have tricuspid regurgitation; therefore, the probability of tricuspid regurgitation, if the murmur is present (*positive* finding), is 22/25 or 88% (i.e., the "post-test probability" if the murmur is present). The second row includes all 75 patients without the murmur. Of these 75 patients, 20 have tricuspid regurgitation; therefore, the post-test probability of tricuspid regurgitation, if the murmur is absent (i.e., *negative* finding), is 20/75 or 27%.

In this example, the pre-test probability of tricuspid regurgitation is 42%. The presence of the murmur (positive result) shifts the probability of disease upward considerably more (i.e., 46%, from 42% to 88%) than the absence of the murmur (negative result) shifts it downward (i.e., 15%, from 42% to 27%). This illustrates an important property of physical signs with a high specificity: when present, physical signs with *high specificity* greatly *increase* the probability of disease. A corollary to this applies to findings with high sensitivity: when *absent*, physical signs with a high *sensitivity* greatly *decrease* the probability of disease. The holosystolic murmur has a high specificity (95%) but only a meager sensitivity (52%), meaning that, at the bedside, a positive result (the presence of a murmur) has greater diagnostic importance than the negative result (the absence of the murmur). The presence of the characteristic murmur argues compellingly for tricuspid regurgitation, but its absence is less helpful, simply because many patients with significant regurgitation lack the characteristic murmur.

Sackett and others have suggested mnemonics for these characteristics as well: "SpPin" (i.e., a *Sp*ecific test, when *P*ositive, rules *in* disease) and "SnNout" (i.e., a *Sen*sitive test, when *N*egative, rules *out* disease).[37]

IV. Likelihood Ratios

LRs, like sensitivity and specificity, describe the discriminatory power of physical signs. Although they have many advantages, the most important is how simply and quickly they can be used to estimate post-test probability.

A. DEFINITION

The LR of a physical sign is the proportion of patients *with* disease who have a particular finding divided by the proportion of patients *without* disease who also have the same finding.

$$\text{LR} = \frac{\text{Probability of a finding in patients } \textit{with} \text{ disease}}{\text{Probability of the same finding in patients } \textit{without} \text{ disease}}$$

The adjectives *positive* or *negative* indicate whether that LR refers to the presence of the physical sign (i.e., positive result) or to the absence of the physical sign (i.e., the negative result).

A **positive LR**, therefore, is the proportion of patients *with* disease who *have* a physical sign divided by the proportion of patients *without* disease who also *have* the same sign. The numerator of this equation—proportion of patients with disease who have the physical sign—is the sign's sensitivity. The denominator—proportion of patients without disease who have the sign—is the complement of specificity, or (1 – specificity). Therefore,

$$\text{Positive LR} = \frac{(\text{sens})}{(1 - \text{spec})}$$

In our hypothetical study (Fig. 2.1), the proportion of patients with tricuspid regurgitation who have the murmur 22/42 or 52.4% (i.e., the finding's sensitivity) and the proportion of patients without tricuspid regurgitation who also have the murmur is 3/58 or 5.2% (i.e., 1 – specificity). The ratio of these proportions [i.e., (sensitivity)/(1−specificity)] is 10.1, which is the positive LR for a holosystolic murmur at the lower sternal border. This number means that patients *with* tricuspid regurgitation are 10.1 times more likely to have the holosystolic murmur than those *without* tricuspid regurgitation.

Similarly, the **negative LR** is the proportion of patients *with* disease *lacking* a physical sign divided by the proportion of patients *without* disease also *lacking* the sign. The numerator of this equation—proportion of patients with disease *lacking* the finding—is the complement of

sensitivity, or (1 − sensitivity). The denominator of the equation—proportion of patients without disease *lacking* the finding—is the specificity. Therefore,

$$\text{Negative LR} = \frac{(1 - \text{sens})}{(\text{spec})}$$

In our hypothetical study, the proportion of patients with tricuspid regurgitation lacking the murmur is 20/42 or 47.6% (i.e., 1- sensitivity) and the proportion of patients without tricuspid regurgitation lacking the murmur is 55/58 or 94.8% (i.e., the specificity). The ratio of these proportions [i.e., (1-sensitivity)/(specificity)] is 0.5, which is the negative LR for the holosystolic murmur. This number means that patients *with* tricuspid regurgitation are 0.5 times less likely to lack the murmur than those *without* tricuspid regurgitation (the inverse statement is less confusing: patients *without* tricuspid regurgitation are 2 times more likely to lack a murmur than those *with* tricuspid regurgitation).

Although these formulae are difficult to recall, the interpretation of LRs is straightforward. Findings with LRs greater than 1 increase the probability of disease; the greater the LR, the more compelling the argument *for* disease. Findings whose LRs lie between between 0 and 1 decrease the probability of disease; the closer the LR is to zero, the more convincing the finding argues *against* disease. Findings whose LRs equal 1 lack diagnostic value because they do not change probability at all. "Positive LR" describes how probability changes when the finding is *present*. "Negative LR" describes how probability changes when the finding is *absent*.

LRs, therefore, are nothing more than diagnostic weights, whose possible values range from 0 (i.e., excluding disease) to infinity (i.e., pathognomonic for disease, Fig. 2.2).

B. USING LRS TO DETERMINE PROBABILITY

The clinician can use the LR of a physical finding to estimate probability of disease in three ways: (1) by using graphs or other easy-to-use nomograms,[38,39] (2) by using bedside approximations, or (3) by using formulae.

1. Using Graphs

a. Parts of the graph

Fig. 2.3 is an easy-to-use graph that illustrates the relationship between pre-test probability (x-axis) and post-test probability (y-axis), given the finding's LR. The straight line bisecting the

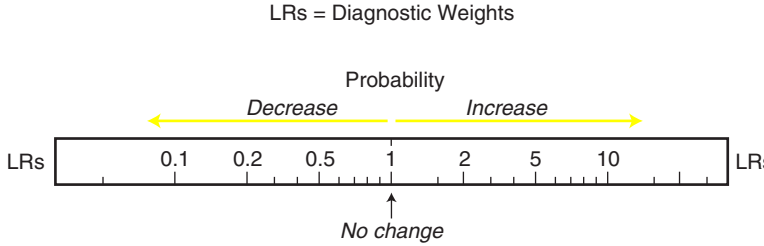


Fig. 2.2 Likelihood ratios as diagnostic weights. The relationship between a specific physical sign and a specific disease is described by a unique number—its likelihood ratio—which is nothing more than a diagnostic weight describing how much that sign argues for or against that specific disease. The possible values of LRs range from zero to infinity (∞). Findings with LRs greater than 1 argue *for* the specific disease (the greater the value of the LR, the more the probability of disease increases). Findings with LRs less than 1 argue *against* the disease (the closer the number is to zero, the more the probability of disease decreases). LRs that equal 1 do not change probability of disease at all.

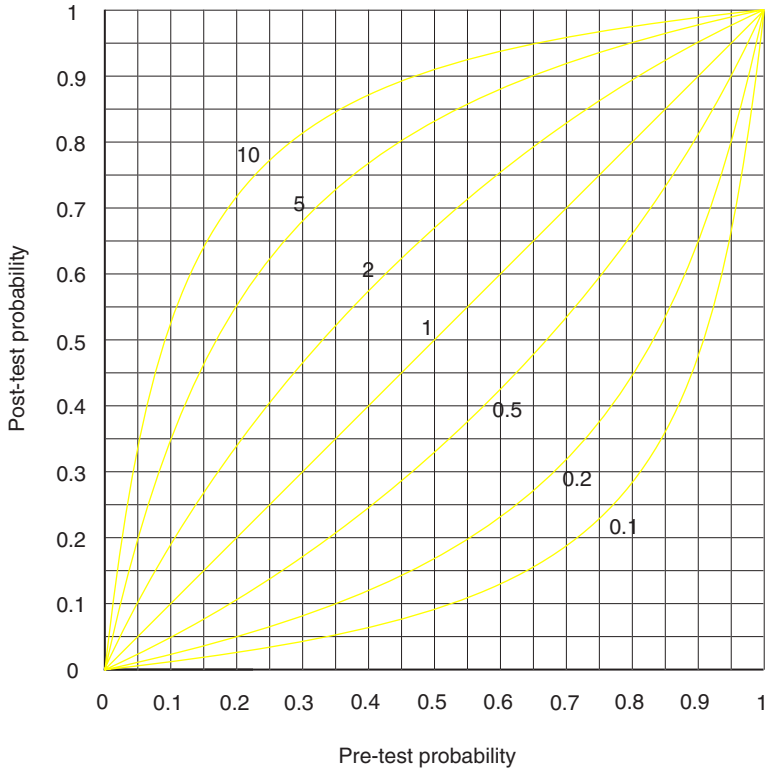


Fig. 2.3 Probability and likelihood ratios. The curves describe how pre-test probability (x-axis) relates to post-test probability (y-axis), given the likelihood ratio (LR) for the physical finding. Only the curves for seven likelihood ratios are depicted (from LR = 0.1 to LR = 10).

graph into an upper left half and lower right half describes the LR of 1, which has no discriminatory value because, for findings with this LR, post-test probability always equals pre-test probability. Physical findings that argue *for* disease (i.e., LRs >1) appear in the upper left half of the graph; the larger the value of the LR, the more the curve approaches the upper left corner. Physical findings that argue *against* disease (i.e., LRs <1) appear in the lower the lower right half of the graph: the closer the LR is to zero, the more the curve approaches the lower right corner.

In Fig. 2.3, the three depicted curves with LRs greater than 1 (i.e., LR = 2, 5, and 10) are mirror images of the three curves with LRs less than 1 (i.e., LR = 0.5, 0.2, and 0.1). (This assumes the "mirror" is the line LR = 1.) This symmetry indicates that findings with an LR of 10 argue as much *for* disease as those with an LR = 0.1 argue *against* disease (although this is true only for the intermediate pre-test probabilities). Similarly, LR = 5 argues as much for disease as LR = 0.2 argues against it, and LR = 2 mirrors LR = 0.5. Keeping these companion curves in mind will help the clinician interpret the LRs throughout this book.[†] If a finding has an LR other than one of these depicted seven curves, its position can be estimated with little loss in accuracy. For example, the curve for LR = 4 lies between LR = 5 and LR = 2, though closer to LR = 5 than to LR = 2.

[†]These companion pairs are easy to recall because they are the inverse of each other: the inverse of 10 is 1/10 = 0.1; the inverse of 5 is 1/5 = 0.2; the inverse of 2 is 1/2 = 0.5.

b. Using the Graph to Determine Probability

To use this graph, the clinician identifies on the x-axis the patient's pre-test probability, derived from published estimates and clinical experience, and extends a line upward from that point to meet the LR curve for the physical finding. The clinician then extends a horizontal line from this point to the y-axis to identify post-test probability.

Figure 2.4 depicts this process for the lower sternal holosystolic murmur and tricuspid regurgitation. The pre-test probability of tricuspid regurgitation is 42%. If the characteristic murmur is present (positive LR = 10), a line is drawn upward from 0.42 on the x-axis to the LR = 10 curve; from this point, a horizontal line is drawn to the y-axis to find the post-test probability (88%). If the murmur is absent (negative LR = 0.5), the post-test probability is the y-value where the vertical line intersects the LR = 0.5 curve (i.e., post-test probability of 27%).

These curves illustrate an additional important point: physical signs are diagnostically most useful when they are applied to patients who have pre-test probabilities in the intermediate range (i.e. 20% to 80%), because in this range the different LR curves diverge the most from the LR = 1 curve (thus, shifting probability up or down a large amount). If instead the pre-test probability is already very low or very high, all the LR curves cluster close to the line LR = 1 curve in either the bottom left or upper right corners, thus with only a relatively small impact on probability.

2. Approximating Probability

The clinician can avoid using graphs and instead approximate post-test probability by remembering the following two points: (1) The companion LR curves in Fig. 2.3 are LR = 2 and LR = 0.5, LR = 5 and LR = 0.2, and LR = 10 and LR = 0.1. (2) The first three multiples of "15" are 15, 30, and 45. Using this rule, the LRs of 2, 5, and 10 *increase* probability about 15%, 30%, and 45%, respectively (see Fig. 2.5). The LRs of 0.5, 0.2, and 0.1 *decrease* probability about 15%, 30%, and 45%, respectively.[40] These estimates are accurate to within 5% to 10% of the actual value, as long as the clinician rounds estimates over 100 to an even 100% and estimates below zero to an even 0%.

Therefore, in our hypothetical patient with pulmonary hypertension, the finding of a holosystolic murmur (LR = 10) increases the probability of tricuspid regurgitation from 42% to 87% (i.e., 42% + 45% = 87%, which is only 1% lower than the actual value). The absence of the murmur (LR = 0.5) decreases the probability of tricuspid regurgitation from 42% to 27% (i.e., 42% − 15% = 27%, which is identical to actual value).

Table 2.2 summarizes similar bedside estimates for all LRs between 0.1 and 10.0.

3. Calculating Probability

The post-test probability also can be calculated by first converting pre-test probability (P_{pre}) into pre-test odds (O_{pre}):

$$O_{pre} = \frac{P_{pre}}{\left(1 - P_{pre}\right)}$$

The pre-test odds (O_{pre}) is multiplied times the LR of the physical sign to determine the post-test odds (O_{post}):

$$O_{post} = O_{pre} \times LR$$

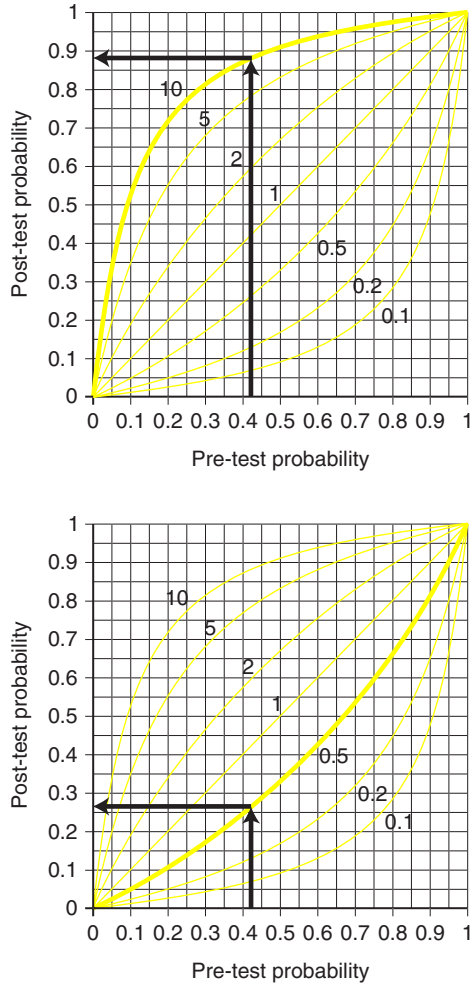


Fig. 2.4 Probability and likelihood ratios: patients with pulmonary hypertension. In our hypothetical clinician's practice, 42% of patients with pulmonary hypertension have the complication of tricuspid regurgitation (i.e., pre-test probability is 42%). To use the curves, the clinician finds 0.42 on the x-axis and extends a line upward. The post-test probability of tricuspid regurgitation is read off the y-axis where the vertical line intersects the curve of the appropriate LR. The probability of tricuspid regurgitation if a holosystolic murmur is present at the left lower sternal edge (LR = 10.1) is 88%; the probability if the finding is absent (LR = 0.5) is 27%.

The post-test odds (O_{post}) converts back to post-test probability (P_{post}), using:

$$P_{post} = \frac{O_{post}}{\left(1 + O_{post}\right)}$$

Therefore, in our hypothetical example of the patients with pulmonary hypertension, the pre-test odds for tricuspid regurgitation would be [(0.42)/(1–0.42)] or 0.72. If the murmur is present (LR = 10), the post-test odds would be [0.72 × 10] or 7.2, which translates to a post-test probability of [(7.2)/(1 + 7.2)] or 0.88 (i.e., 88%). If the murmur wave is absent (LR = 0.5), the post-test odds would be [0.72 × 0.5] or 0.36, which translates to a post-test probability of [(0.36)/(1 + 0.36)] or 0.27 (i.e., 27%).

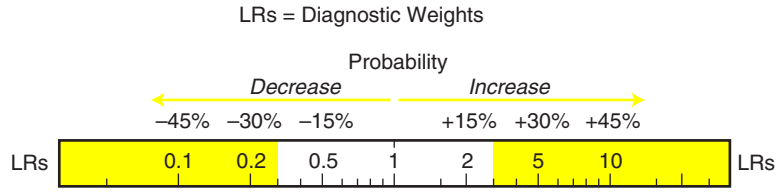


Fig. 2.5 Approximating probability. Clinicians can estimate changes in probability by recalling the LRs 2, 5, and 10 and the first 3 multiples of 15 (i.e., 15, 30, 45). A finding whose LR is 2 increases probability about 15%, one of 5 increases it 30% and one of 10 increases it 45% (these changes are *absolute* increases in probability). LRs whose values are 0.5, 0.2, and 0.1 (i.e., the reciprocals of 2, 5, and 10) decrease probability 15%, 30%, and 45%, respectively. Throughout this book, LRs with values ≥3 or ≤0.3 (represented by the *shaded part* of the diagnostic weight "ruler") are presented in boldface type to indicate those physical findings that change probability sufficiently to be clinically meaningful (i.e., they increase or decrease probability at least 20% to 25%).

TABLE 2.2 **Likelihood Ratios and Bedside Estimates**

Likelihood Ratio	Approximate Change in Probability*
0.1	–45%
0.2	–30%
0.3	–25%
0.4	–20%
0.5	–15%
1	No change
2	+15%
3	+20%
4	+25%
5	+30%
6	+35%
7	
8	+40%
9	
10	+45%

From McGee S. Simplifying likelihood ratios. *J Gen Intern Med*. 2002;17(8):646–649.
*These changes describe *absolute* increases or decreases in probability. For example, a patient with pre-test probability of 20% and physical finding whose LR is 5 would have a post-test probability of 20% + 30% = 50%. The text describes how to easily recall these estimates.

Clinical medicine, however, is rarely as precise as these calculations suggest, and for most decisions at the bedside, the approximations described earlier are more than adequate.

C. ADVANTAGES OF LIKELIHOOD RATIOS

1. Simplicity

In a single number, the LR conveys to clinicians how convincingly a physical sign argues for or against disease. If the LR of a finding is a large number, disease is likely; and if the LR of a finding is close to zero, disease is doubtful. This advantage allows clinicians to quickly compare different diagnostic strategies and thus refine clinical judgment.[40]

2. Accuracy

Using LRs to describe diagnostic accuracy is superior to sensitivity and specificity, because the earlier described mnemonics, SpPin and SnNout, are sometimes misleading. For example, according to the mnemonic SpPin, a finding with a specificity of 95% should argue conclusively for disease, but it does so only if the positive LR for the finding is a high number. If the finding's sensitivity is 60%, the positive LR is 12 and the finding does argue convincingly for disease (i.e., consistent with the SpPin mnemonic); if the finding's sensitivity is only 10%, however, the positive LR is 2 and post-test probability changes only slightly (i.e., inconsistent with SpPin mnemonic). Similarly, a highly sensitive finding argues convincingly against disease (i.e., SnNout) only when its calculated negative LR is a number close to zero.

3. Levels of Findings

Another advantage of LRs is that a physical sign measured on an ordinal scale (e.g., 0, 1+, 2+, 3+) or continuous scale (e.g., blood pressure) can be categorized into different levels to determine the LR for each level, thereby increasing the accuracy of the finding. Other examples include continuous findings such as heart rate, respiratory rate, temperature, and percussed span of the liver, and ordinal findings such as intensity of murmurs and degree of edema.

For example, in patients with chronic obstructive lung disease (i.e., emphysema, chronic bronchitis), breath sounds are typically faint. If the clinician grades the intensity of breath sounds on a scale from 0 (absent) to 24 (very loud), based on the methods discussed in Chapter 30,[41,42] he or she can classify the patient's breath sounds into one of four groups: scores of 9 or less (very faint), 10 to 12, 13 to 15, or greater than 15 (loud). Each category then has its own LR (Table 2.3): scores of 9 or less significantly increase probability of obstructive disease (LR = 10.2), whereas scores greater than 15 significantly decrease it (LR = 0.1). Scores from 10 to 12 argue somewhat for disease (LR = 3.6), and scores from 13 to 15 provide no diagnostic information (LR not significantly different from 1). If the clinician had instead identified breath sounds as simply "faint" or "normal/increased" (i.e., the traditional positive or negative finding), the finding may still discriminate between patients with and without obstructive disease, but it misses the point that the discriminatory power of the sign resides mostly with scores less than 10 and greater than 15.

When findings are categorized into levels, the term *specificity* becomes meaningless. For example, the specificity of a breath sound score of 13 to 15 is 80%, which means that 80% of patients without chronic airflow limitation have values other than 13 to 15, though the "80%" does not convey whether most of these other values are greater than 15 or less than 13. Similarly, when findings are put in more than 2 categories, the LR descriptor *negative* is no longer necessary, because all LRs are *positive* ones for their respective category.

TABLE 2.3 **Breath Sound Intensity and Chronic Airflow Limitation**

Breath sound score	Likelihood ratio
9 or less	10.2
10–12	3.6
13–15	NS
>15	0.1

Data from Bohadana AB, Peslin R, Uffholtz H. Breath sounds in the clinical assessment of airflow obstruction. *Thorax*. 1978;33:345–351; Pardee NE, Martin CJ, Morgan EH. A test of the practical value of estimating breath sound intensity: breath sounds related to measured ventilatory function. *Chest*. 1976;70(3):341–344.
NS, Not significant.

4. Combining Findings

A final advantage of LRs is that clinicians can use them to combine findings, which is particularly important for those physical signs with positive LRs around 2 or negative LRs around 0.5, signs that by themselves change probability little but when combined have significant effects on probability. Individual LRs can be combined, however, only if the findings are "independent."

a. Independence of Findings

Independence means that the LR for second finding does not change once the clinician determines whether the first finding is present or absent. For a few diagnostic problems, investigators have identified which findings are independent of each other. These findings appear as components of "diagnostic scoring schemes" in the EBM Boxes throughout this book (e.g., Wells score for deep venous thrombosis). For most physical findings, however, very little information is available about independence, and the clinician must judge whether combining findings is appropriate.

One important clue is that most independent findings have unique pathophysiology. For example, when considering pneumonia in patients with cough and fever, the clinician could combine the findings of abnormal mental status and diminished breath sounds, using the individual LR of each finding because abnormal mental status and diminished breath sounds probably have separate pathophysiology. Similarly, when considering heart failure in patients with dyspnea, the clinician could combine the findings of elevated neck veins and third heart sound because these findings also have different pathophysiology.

Examples of findings whose individual LRs should not be combined (because the findings share the same pathophysiology) are flank dullness and shifting dullness in the diagnosis of ascites (both depend on intraabdominal contents dampening the vibrations of the abdominal wall during percussion), neck stiffness and Kernig sign in the diagnosis of meningitis (both are caused by meningeal irritation), and edema and elevated neck veins in the diagnosis of heart failure (both depend on elevated right atrial pressure).

Until more information is available, the safest policy for the clinician to follow, when combining LRs of individual findings, is to combine no more than three findings, all of which have distinct pathophysiology.

b. How to Combine Findings

The clinician can use any of the methods previously described to combine findings, simply by making the post-test probability from the first finding the pre-test probability for the second finding. For example, a hypothetical patient with acute fever and cough has two positive findings that we believe have separate pathophysiology and therefore are independent: abnormal mental status (LR = 1.9 for pneumonia) and diminished breath sounds (LR = 2.4 for pneumonia). The pre-test probability of pneumonia, derived from published estimates and clinical experience, is estimated to be 20%. Using the graph, the finding of abnormal mental status increases the probability from 20% to 32%; this post-test probability then becomes the pre-test probability for the second finding, diminished breath sounds, which increases probability from 32% to 53%—the overall probability after application of the two findings. Using the approximating rules, both findings (LRs $\approx$ 2.0) increase the probability about 15%; the post-test probability is thus 20% + 15% + 15% = 50% (an error of only 3%). Using formulas to calculate probability, the LRs of the separate findings are multiplied together and the product is used to convert pre-test into post-test odds. The product of the two LRs is 4.56 (1.9×2.4); the pre-test odds would be $0.2/0.8 = 0.25$; the post-test odds would be $0.25 \times 4.56 = 1.14$, which equals a probability of $1.14/2.14 = 53\%$.

References may be accessed online at *Elsevier eBooks for Practicing Clinicians*.

This page intentionally left blank

References

1. Eskelinen M, Ikonen J, Lipponen P. Contributions of history-taking, physical examination, and computer assistance to diagnosis of acute small-bowel obstruction: a prospective study of 1333 patients with acute abdominal pain. *Scand J Gastroenterol*. 1994;29(8):715–721.
2. Staniland JR, Ditchburn J, De Dombal FT. Clinical presentation of acute abdomen: study of 600 patients. *Br Med J*. 1972;3(5823):393–398.
3. Böhner H, Yang Z, Franke C, Verreet PR, Ohmann C. Simple data from history and physical examination help to exclude bowel obstruction and to avoid radiographic studies in patients with acute abdominal pain. *Eur J Surg*. 1998;164(10):777–784.
4. Stiell IG, Greenberg GH, McKnight RD, Nair RC, McDowell I, Worthington JR. A study to develop clinical decision rules for the use of radiography in acute ankle injuries. *Ann Emerg Med*. 1992;21(4):384–390.
5. Stiell IG, Greenberg GH, McKnight RD, et al. Decision rules for the use of radiography in acute ankle injuries: refinement and prospective validation. *JAMA*. 1993;269(9):1127–1132.
6. Heckerling PS, Tape TG, Wigton RS, et al. Clinical prediction rule for pulmonary infiltrates. *Ann Intern Med*. 1990;113(9):664–670.
7. Wells PS, Anderson DR, Bormanis J, et al. Value of assessment of pretest probability of deep-vein thrombosis in clinical management. *Lancet*. 1997;350(9094):1795–1798.
8. Miron MJ, Perrier A, Bounameaux H. Clinical assessment of suspected deep vein thrombosis: comparison between a score and empirical assessment. *J Intern Med*. 2000;247(2):249–254.
9. Anderson DR, Wells PS, MacLeod B, et al. Thrombosis in the emergency department. *Arch Intern Med*. 1999;159(5):477–482.
10. Funfsinn N, Caliezi C, Baiasiutti FD, et al. Rapid D-dimer testing and pre-test clinical probability in the exclusion of deep venous thrombosis in symptomatic outpatients. *Blood Coagul Fibrinolysis*. 2001;12(3):165–170.
11. Kearon C, Ginsberg JS, Douketis J, et al. Management of suspected deep venous thrombosis in outpatients by using clinical assessment and D-dimer testing. *Ann Intern Med*. 2001;135(2):108–111.
12. Oudega R, Hoes AW, Moons KGM. The Wells rule does not adequately rule out deep venous thrombosis in primary care patients. *Ann Intern Med*. 2005;143(2):100–107.
13. Schutgens REG, Ackermark P, Haas FJLM, et al. Combination of a normal D-dimer concentration and a non-high pretest clinical probability score is a safe strategy to exclude deep venous thrombosis. *Circulation*. 2003;107(4):593–597.
14. Tick LW, Ton E, van Voorthuizen T, et al. Practical diagnostic management of patients with clinically suspected deep vein thrombosis by clinical probability test, compression ultrasonography, and D-dimer test. *Am J Med*. 2002;113(8):630–635.
15. Aschwanden M, Labs KH, Jeanneret C, Gehrig A, Jaeger KA. The value of rapid D-dimer testing combined with structured clinical evaluation for the diagnosis of deep vein thrombosis. *J Vasc Surg*. 1999;30(5):929–935.
16. Penaloza A, Laureys M, Wautrecht JC, Lheureux P, Motte S. Accuracy and safety of pretest probability assessment of deep vein thrombosis by physicians in training using the explicit Wells model. *J Thromb Haemost*. 2006;4(1):278–281.
17. Yamaki T, Nozaki M, Sakurai H, et al. Combined use of pretest clinical probability score and latex agglutination D-dimer testing for excluding acute deep vein thrombosis. *J Vasc Surg*. 2009;50(5):1099–1105.
18. Engelberger RP, Aujesky D, Calanca L, Staeger P, Hugli O, Mazzolai L. Comparison of the diagnostic performance of the original and modified Wells score in inpatients and outpatients with suspected deep vein thrombosis. *Thromb Res*. 2011;127(6):535–539.
19. Dybowska M, Tomkowski WZ, Kuca P, Ubysz R, Jozwik A, Chmielewski D. Analysis of the accuracy of the Wells scale in assessing the probability of lower limb deep vein thrombosis in primary care patients practice. *Thromb J*. 2015;13:18.
20. Chagnon I, Bounameaux H, Aujesky D, et al. Comparison of two clinical prediction rules and implicit assessment among patients with suspected pulmonary embolism. *Am J Med*. 2002;113(4):269–275.
21. Miniati M, Bottai M, Monti S. Comparison of 3 clinical models for predicting the probability of pulmonary embolism. *Medicine*. 2005;84(2):107–114.

22. Wells PS, Anderson DR, Rodger M, et al. Excluding pulmonary embolism at the bedside without diagnostic imaging: management of patients with suspected pulmonary embolism presenting to the emergency department by using a simple clinical model and D-dimer. *Ann Intern Med*. 2001;135(2):98–107.
23. Wolf SJ, McCubbin TR, Feldhaus KM, Faragher JP, Adcock DM. Prospective validation of Wells criteria in the evaluation of patients with suspected pulmonary embolism. *Ann Emerg Med*. 2004;44(5):503–510.
24. Anderson DR, Kovacs MJ, Dennie C, et al. Use of spiral computed tomography contrast angiography and ultrasonography to exclude the diagnosis of pulmonary embolism in the emergency department. *J Emerg Med*. 2005;29(4):399–404.
25. Calisir C, Yavas US, Ozkan IR, et al. Performance of the Wells and revised Geneva scores for predicting pulmonary embolism. *Eur J Emerg Med*. 2008;16(1):49–52.
26. Klok FA, Kruisman E, Spaan J, et al. Comparison of the revised Geneva score with the Wells rule for assessing clinical probability of pulmonary embolism. *J Thromb Haemost*. 2008;6(1):40–44.
27. Wong DD, Ramaseshan G, Mendelson RM. Comparison of the Wells and revised Geneva scores for the diagnosis of pulmonary embolism: an Australian experience. *Int Med J*. 2011;41(3):258–262.
28. Penaloza A, Verschuren F, Meyer G, et al. Comparison of the unstructured clinician gestalt, the Wells score, and the revised Geneva score to estimate pretest probability for suspected pulmonary embolism. *Ann Emerg Med*. 2013;62(2):117–124.
29. Penaloza A, Melot C, Motte S. Comparison of the Wells score with the simplified revised Geneva score for assessing pretest probability of pulmonary embolism. *Thromb Res*. 2011;127(2):81–84.
30. Yap KS, Kalff V, Turlakow A, Kelly MJ. A prospective reassessment of the utility of the Wells score in identifying pulmonary embolism. *Med J Aust*. 2007;187(6):333–336.
31. Di Marca S, Cilia C, Campagna A, et al. Comparison of Wells and revised Geneva rule to assess pretest probability of pulmonary embolism in high-risk hospitalized elderly adults. *J Am Geriatr Soc*. 2015;63(6):1091–1097.
32. Newman LG, Waller J, Palestro J, et al. Unsuspected osteomyelitis in diabetic foot ulcers: diagnosis and monitoring by leukocyte scanning with indium In 111 oxyquinoline. *JAMA*. 1991;266(9):1246–1251.
33. Fleischer AE, Didyk AA, Woods JB, Burns SE, Wrobel JS, Armstrong DG. Combined clinical and laboratory testing improves diagnostic accuracy for osteomyelitis in the diabetic foot. *J Foot Ankle Surg*. 2009;48(1):39–46.
34. Ertugrul BM, Savk O, Ozturk B, Cobanoglu M, Oncu S, Sakarya S. The diagnosis of diabetic foot osteomyelitis: examination findings and laboratory values. *Med Sci Monit*. 2009;15(6):CR307–12.
35. Rahko PS. Prevalence of regurgitant murmurs in patients with valvular regurgitation detected by Doppler echocardiography. *Ann Intern Med*. 1989;111(6):466–472.
36. Sackett DL, Richardson WS, Rosenberg W, Haynes RB. *Evidence-Based Medicine: How to Practice and Teach EBM*. Churchill Livingstone; 1997.
37. Sackett DL, Haynes RB, Tugwell P. *Clinical Epidemiology: A Basic Science for Clinical Medicine*. 1st ed. Little, Brown and Co; 1985.
38. Fagan TJ. Nomogram for Bayes' theorem. *N Engl J Med*. 1975;293(5):257.
39. Grimes DA, Schulz KF. Refining clinical diagnosis with likelihood ratios. *Lancet*. 2005;365(9469):1500–1505.
40. McGee S. Simplifying likelihood ratios. *J Gen Intern Med*. 2002;17(8):646–649.
41. Bohadana AB, Peslin R, Uffholtz H. Breath sounds in the clinical assessment of airflow obstruction. *Thorax*. 1978;33(3):345–351.
42. Pardee NE, Martin CJ, Morgan EH. A test of the practical value of estimating breath sound intensity: breath sounds related to measured ventilatory function. *Chest*. 1976;70(3):341–344.

CHAPTER 3

Using the Tables in This Book

KEY TEACHING POINTS

- **Frequency of findings** tables present only the sensitivity of findings (derived from studies of large numbers of patients with a confirmed diagnosis). In these tables, only those findings with *high* sensitivity are clinically useful: if these key findings are *absent* in symptomatic patients, diagnosis of disease is *unlikely*.
- **EBM Boxes**, derived from large numbers of patients presenting with similar symptoms but different final diagnoses, quickly convey to clinicians which physical signs are most accurate for a particular diagnosis. Those findings with likelihood ratios (LRs) having the greatest value *increase* probability of disease the most (i.e., LRs function like diagnostic weights). Those findings with LRs closest to the value of 0 *decrease* probability of disease the most.

I. Introduction

Information about the diagnostic accuracy of physical findings is presented in two types of tables in this book: (1) "frequency of findings" tables, which display only the sensitivity of physical signs, and (2) evidence-based medicine (EBM) boxes, or "diagnostic accuracy" tables, which present the sensitivity, specificity, and LRs of various physical signs.

II. Frequency of Findings Tables

A. DEFINITION

Frequency of findings tables summarize multiple studies of patients with a specific diagnosis and present the sensitivity of physical signs found in that disorder. These tables provide no information about a sign's specificity. An example is Table 3.1, listing the frequency of findings in constrictive pericarditis, a disorder in which a diseased and unyielding pericardium interferes with diastolic filling of the heart.

B. PARTS OF THE TABLE

1. Finding

The first column lists the various physical signs, organized by organ system, with the findings of each organ system listed from most to least frequent.

TABLE 3.1 ■ **Constrictive Pericarditis***

Physical Finding	Frequency (%)†
Neck veins	
Elevated neck veins	94
Prominent *y* descent (Friedreich sign)	57–100
Kussmaul sign	14–50
Arterial pulse	
Irregularly irregular (atrial fibrillation)	36–70
Blood pressure	
Pulsus paradoxus >10 mm Hg	13–64
Auscultation of heart	
Pericardial knock	20–94
Pericardial rub	3–16
Other	
Hepatomegaly	53–100
Edema	58–100
Ascites	37–89

Data from 780 patients from references 1–13.
*Diagnostic standard: for *constrictive pericarditis*, surgical and postmortem findings,[1–5] sometimes in combination with hemodynamic findings.[6–12]
†Results are overall mean frequency or, if statistically heterogeneous, the range of values.

2. Frequency

The second column lists the sensitivity (or frequency) of the physical signs. If the sensitivity from every study is statistically similar, the overall mean frequency is presented (e.g., in Table 3.1, 94% of patients with constrictive pericarditis have elevated neck veins). If the sensitivities from the different studies are statistically diverse ($p < 0.05$ by the chi-squared test), the range of values is instead presented (e.g., in Table 3.1, 20% to 94% have a pericardial knock—a loud heart sound heard near the apex during early diastole).

3. Footnotes

The footnotes to these tables present the source of the information and the diagnostic standards used. For example, the information in Table 3.1 is based on 780 patients from 13 different studies, which based the diagnosis of constrictive pericarditis on surgical, postmortem, or hemodynamic findings.

C. Interpretation

Because the frequency of findings tables provide just information about a sign's sensitivity, they can only be used to support a statement that a physical sign, when *absent*, argues *against* disease. The absence of any finding whose sensitivity (or frequency) is 94% or more is a compelling argument against that diagnosis (i.e., the negative LR is 0.1 or less, even if the specificity of the finding, which is unknown, is as low as 50%). In Table 3.1, elevated venous pressure is such a finding (sensitivity = 94%): if the clinician is considering the diagnosis of constrictive pericarditis, but the patient's bedside estimate of venous pressure is normal, the diagnosis is unlikely.

Similarly, the absence of two or three independent findings having sensitivities greater than 80% is also a compelling argument against disease* (see Chapter 2 for a definition of *independent findings*).

*This statement assumes that the product of the LRs being combined is less than 0.1. Therefore, $LR^n = \left[\frac{(1 - \text{sens})}{(\text{spec})}\right]^n \leq 0.1$ where n is the number of findings being combined. If the specificity of the findings is as low as 50%, each of two findings being combined must have a sensitivity greater than 84%, and each of three findings being combined must have a sensitivity greater than 77%.

III. Diagnostic Accuracy Tables (EBM Boxes)

A. DEFINITION

Diagnostic accuracy tables summarize information from large numbers of patients who present with similar symptoms but different diagnoses. These EBM Boxes present the physical sign's sensitivity, specificity, and positive and negative LRs, which then indicate how well that physical sign discriminates between patients with the particular diagnosis of interest and those without it.

EBM Box 3.1 presents an example summarizing the diagnostic accuracy of physical signs for pneumonia, as applied to large numbers of patients with cough and fever (see Chapter 32 for the complete EBM Box). In these studies, only about 20% of patients had pneumonia; the remainder had other causes of cough and fever such as sinusitis, bronchitis, or rhinitis.

B. PARTS OF THE EBM BOX

1. Finding

The first column presents the physical signs, organized by organ system, and the source of the information. Validated scoring schemes that combine findings appear in the bottom rows of EBM Boxes.

2. Sensitivity and Specificity

The second and third columns present the range of a physical sign's sensitivity and specificity observed in these studies.

3. Likelihood Ratios

The third and fourth columns present the physical sign's positive and negative LR (for clarity, "likelihood ratio if finding *present*" refers to the positive LR, and "likelihood ratio if finding *absent*" refers to the negative LR). In contrast to sensitivity and specificity, which are presented as a range of values, LRs are described by a single number, derived by using a statistical technique called the random effects model (see the section on Summarizing LRs in this chapter).[29] Only statistically significant LRs are presented in the EBM Boxes. If the 95% confidence interval (CI) for an LR, positive or negative, includes the value of 1, that result of the physical finding fails to statistically discriminate between patients with disease and those without it, and the notation "NS" (for "not significant") is recorded in the EBM Box.

4. Footnote

The footnotes to EBM Boxes describe the diagnostic standards used in the studies and, if necessary, definitions of findings. The footnote for EBM Box 3.1, for example, indicates that the diagnostic standard for pneumonia was the chest radiograph; it also describes the components of Heckerling diagnostic scoring scheme presented in the bottom rows of the EBM Box.

C. Interpretation of EBM Box

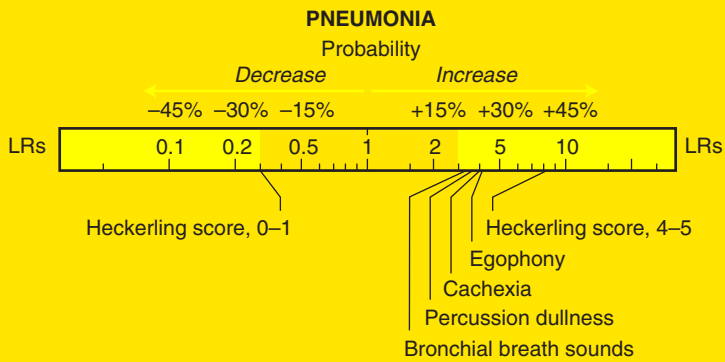
To use these EBM Boxes, the clinician needs to only glance at the LR columns to appreciate the discriminatory power of different findings. LRs with the greatest value increase probability of disease the most; LRs with the value closest to zero decrease probability the most. Boldface type highlights all findings with an LR of 3 or more or of 0.3 or less, thus allowing quick identification of those physical signs that increase probability more that 20% to 25% (LR $\geq$3) and those that decrease it more that 20% to 25% (LR $\leq$0.3; see also Chapter 2).

In patients with cough and fever (EBM Box 3.1), the individual findings increasing probability of pneumonia the most are egophony (LR = 4.1), cachexia (LR = 4), percussion dullness (LR = 3.6), and bronchial breath sounds (LR = 3.3). In contrast, no *individual* finding in this EBM box, whether present or absent, significantly *decreases* probability of pneumonia. (No LR has a value $\leq$0.3.)

EBM BOX 3.1 Pneumonia*

Finding (Reference)†	Sensitivity (%)	Specificity (%)	Likelihood Ratio‡ if Finding Is: Present	Absent
General appearance				
Cachexia[14]	10	97	**4.0**	NS
Abnormal mental status[15–17]	12–14	92–95	1.9	NS
Lung findings				
Percussion dullness[14–16,18–21]	4–26	82–99	**3.6**	NS
Diminished breath sounds[15,16,18–20,22–26]	7–60	73–98	2.4	0.8
Bronchial breath sounds[15,20]	14–19	94–96	**3.3**	0.9
Egophony[14–16]	4–16	96–99	**4.1**	NS
Crackles[14–19,22–27]	19–67	36–97	2.8	0.8
Wheezing[15–19,22,24,26,27]	4–36	50–96	0.8	NS
Diagnostic score (Heckerling et al.)[15,28]				
0 or 1 findings	7–29	33–65	**0.3**	…
2 or 3 findings	48–55	…	NS	…
4 or 5 findings	38–41	92–97	**8.2**	…

*Diagnostic standard: for *pneumonia*, infiltrate on chest radiograph.
†Definition of findings: for *Heckerling diagnostic score*, the clinician scores 1 point for each of the following 5 findings that are present: temperature >37.8°C, heart rate >100/min, crackles, diminished breath sounds, and *absence* of asthma.
‡Likelihood ratio (LR) if finding present = positive LR; LR if finding absent = negative LR.
NS, Not significant.



EBM Box 3.1 also shows that 4 or more points using Heckerling's diagnostic scheme significantly *increases* probability of pneumonia (LR = 8.2), whereas the presence of 0 or 1 point significant *decreases* it (LR = 0.3).

IV. Criteria for Selecting Studies Used in Diagnostic Accuracy Tables

All studies of adult patients that meet the following four criteria are included in the EBM Boxes of this book.

A. PATIENTS WERE SYMPTOMATIC

The study must have enrolled patients presenting to clinicians with symptoms or other problems. Therefore, studies using asymptomatic controls, which tend to inflate the specificity of physical signs, are excluded. Clinicians do not need a physical sign to help them distinguish patients with pneumonia from healthy persons (who would not be consulting the doctor); instead, they are interested in those physical signs distinguishing pneumonia from other causes of cough and fever.

B. DEFINITION OF PHYSICAL SIGN

The physical sign in the study must be clearly defined.

C. INDEPENDENT COMPARISON TO A DIAGNOSTIC STANDARD

There must be an independent comparison to an acceptable diagnostic standard. *Independent comparison* means that the physical sign was not used to select patients for testing with the diagnostic standard. Acceptable diagnostic standards include laboratory testing, clinical imaging, surgical findings, or postmortem analysis.

D. 2 × 2 TABLE COULD BE CONSTRUCTED

The studies must provide figures or tables from which numbers could be extracted to construct 2 × 2 tables and calculate sensitivity, specificity, and LRs. If any cell of the 2 × 2 table contained the value of zero, 0.5 was added to all cells, to avoid creating the unlikely LRs of 0 or infinity.

V. Summarizing Likelihood Ratios

The random effects model by Dersimonian and Laird,[29] which considers both within study and between study variance to calculate a pooled LR, was used to summarize the LRs from the various studies. Table 3.2 illustrates how this model works. In the top rows of this table are the individual data from all studies of egophony that appear in EBM Box 3.1, including the finding's sensitivity, specificity, the positive and negative LRs, and the LRs 95% confidence intervals (CIs). The bottom row of Table 3.2 shows how all of this information is summarized throughout the book.

In each of the studies, egophony is specific (96% to 99%) but not sensitive (4% to 16%). The positive LRs are all greater than 1, indicating that the finding of egophony increases probability of pneumonia. For one of the three studies (i.e. Gennis and others[16]), the positive LR lacks statistical significance because its 95% CI includes the value of 1 (i.e., a LR value of 1 has no discriminatory value). For the other two studies, the 95% confidence interval of the positive LR excludes the value of 1, thus making them statistically significant. The summary measure for the positive LR (fourth row of this table) is both clinically significant (4.08, a large positive number) and statistically significant (its 95% CI excludes 1.0). All of this information is summarized, in the notation used in this book (last row), by simply presenting the pooled LR of 4.1. (Interested readers may consult the Appendix for the 95% CIs of all LRs in this book.)

In contrast, the negative LRs from each study have both meager clinical significance (i.e., 0.87 to 0.96, values close to 1) and, for two of the three studies, no statistical significance (i.e., the 95% CI includes 1). The pooled negative LR also lacks clinical and statistical significance. Because it is statistically no different from 1.0 (i.e., the 95% CI of the pooled value, 0.88 to 1.01, includes 1), it is summarized using the notation "NS" for "*not significant*."

TABLE 3.2 **Egophony and Pneumonia - Individual Studies**

Reference	Sensitivity (%)	Specificity (%)	Positive LR (95% CI)	Negative LR (95% CI)
Heckerling[15]	16	97	4.91 (2.88, 8.37)	0.87 (0.81, 0.94)
Gennis[16]	8	96	2.07 (0.79, 5.41)	0.96 (0.90, 1.02)
Diehr[14]	4	99	7.97 (1.77, 35.91)	0.96 (0.91, 1.02)
Pooled result			4.08 (2.14, 7.79)	0.93 (0.88, 1.01)
Notation used in book	4–16	96–99	4.1	NS

CI, Confidence interval; *LR*, likelihood ratio; *NS*, not significant.

Presenting the single pooled result for statistically significant LRs and NS for the statistically insignificant ones simplifies the EBM Boxes and makes it much simpler to grasp the point that the finding of egophony in patients with cough and fever increases probability of pneumonia (LR = 4.1), but the absence of egophony changes probability very little or not at all.

References may be accessed online at *Elsevier eBooks for Practicing Clinicians*.

References

1. Evans W, Jackson F. Constrictive pericarditis. *Br Heart J.* 1952;14(1):53–69.
2. Lange RL, Botticelli JT, Tsagaris TJ, Walker JA, Bani M, Bustamante RA. Diagnostic signs in compressive cardiac disorders: constrictive pericarditis, pericardial effusion, and tamponade. *Circ*. 1966;33(5):763–777.
3. Paul O, Castleman B, White PD. Chronic constrictive pericarditis: a study of 53 cases. *Am J Med Sci*. 1948;216(4):361–377.
4. Mounsey P. The early diastolic sound of constrictive pericarditis. *Br Heart J*. 1955;17(2):143–152.
5. Ling LH, Oh JK, Schaff HV, et al. Constrictive pericarditis in the modern era: evolving clinical spectrum and impact on outcome after pericardiectomy. *Circulation*. 1999;1000(13):1380–1386.
6. Tyberg TI, Goodyer AVN, Langou RA. Genesis of pericardial knock in constrictive pericarditis. *Am J Cardiol*. 1980;46(4):570–575.
7. Schiavone WA. The changing etiology of constrictive pericarditis in a large referral center. *Am J Cardiol*. 1986;58(3):373–375.
8. Wood P. Chronic constrictive pericarditis. *Am J Cardiol*. 1961;7:48–61.
9. El-Sherif A, El-Said G. Jugular, hepatic, and praecordial pulsations in constrictive pericarditis. *Br Heart J*. 1971;33(2):305–312.
10. Talreja DR, Edwards WD, Danielson GK, et al. Constrictive pericarditis in 26 patients with histologically normal pericardial thickness. *Circ*. 2003;108(15):1852–1857.
11. Kumawat M, Lahiri TK, Agarwal D. Constrictive pericarditis: retrospective study of 109 patients. *Asian Cardiovasc Thorac Ann*. 2018;26(5):347–352.
12. Mutyaba AK, Balkaran S, Cloete R, et al. Constrictive pericarditis requiring pericardiectomy at Groote Schuur Hospital, Cape Town, South Africa: causes and perioperative outcomes in the HIV era (1990-2012). *J Thorac Cardiovasc Surg*. 2014;148(6):3058-65.e1.
13. Fernandes F, Melo DTP, Ramires FJA, et al. Importance of clinical and laboratory findings in the diagnosis and surgical prognosis of patients with constrictive pericarditis. *Arq Bras Cardiol*. 2017;109(5):457–465.
14. Diehr P, Wood RW, Bushyhead J, Krueger L, Wolcott B, Tompkins RK. Prediction of pneumonia in outpatients with acute cough—A statistical approach. *J Chron Dis*. 1984;37(3):215–225.
15. Heckerling PS, Tape TG, Wigton RS, et al. Clinical prediction rule for pulmonary infiltrates. *Ann Intern Med*. 1990;113(9):664–670.
16. Gennis P, Gallagher J, Falvo C, Baker S, Than W. Clinical criteria for the detection of pneumonia in adults: guidelines for ordering chest roentgenograms in the emergency department. *J Emerg Med*. 1989;7(3):263–268.
17. Mehr DR, Binder EF, Kruse RL, Zweig SC, Madsen RW, D'Agostino RB. Clinical findings associated with radiographic pneumonia in nursing home residents. *J Fam Pract*. 2001;50(11):931–937.
18. Melbye H, Straume B, Aasebo U, Dale K. Diagnosis of pneumonia in adults in general practice. *Scand J Prim Health Care*. 1992;10(3):226–233.
19. Melbye H, Straume B, Aasebo U, Brox J. The diagnosis of adult pneumonia in general practice. *Scand J Prim Health Care*. 1988;6(2):111–117.
20. Steurer J, Held U, Spaar A, et al. A decision aid to rule out pneumonia and reduce unnecessary prescriptions of antibiotics in primary care patients with cough and fever. *BMC Med*. 2011;9:56.
21. Zusman O, Farbman L, Elbaz M, et al. A decision support model to predict the presence of an acute infiltrate on chest radiograph. *Eur J Clin Microbiol Infect Dis*. 2018;37(2):227–232.
22. Nakanishi M, Yoshida Y, Takeda N, et al. Significance of the progression of respiratory symptoms for predicting community-acquired pneumonia in general practice. *Respirology*. 2010;15(6):969–974.
23. van Vugt SF, Broekhuizen BDL, Lammens C, et al. Use of serum C reactive protein and procalcitonin concentrations in addition to symptoms and signs to predict pneumonia in patients presenting to primary care with acute cough: diagnostic study. *BMJ*. 2013;346(apr30):f2540.
24. Flanders SA, Stein J, Shochat G, et al. Performance of a bedside C-reactive protein test in the diagnosis of community-acquired pneumonia in adults with acute cough. *Am J Med*. 2004;116(8):529–535.
25. Minnaard MC, van de Pol AC, de Groot JA, et al. The added diagnostic value of five different C-reactive protein point-of-care test devices in detecting pneumonia in primary care: a nested case-control study. *Scand J Clin Lab Investig*. 2015;75(4):291–295.

26. Takada T, Yamamoto Y, Terada K, et al. Diagnostic utility of appetite loss in addition to existing prediction models for community-acquired pneumonia in the elderly: a prospective diagnostic study in acute care hospitals in Japan. *BMJ Open*. 2017;7(11):e019155.
27. Singal BM, Hedges JR, Radack KL. Decision rules and clinical prediction of pneumonia: evaluation of low-yield criteria. *Ann Emerg Med*. 1989;18(1):13–20.
28. Emerman CL, Dawson N, Speroff T, et al. Comparison of physician judgment and decision aids for ordering chest radiographs for pneumonia in outpatients. *Ann Emerg Med*. 1991;20(11):1215–1219.
29. DerSimonian R, Laird N. Meta analysis in clinical trials. *Control Clin Trials*. 1986;7(3):177–188.

CHAPTER 4

Using the Online EBM Calculator

I. The Evidence-Based Medicine Calculator

An easy-to-use online calculator is provided on the *Elsevier eBooks for Practicing Clinicians* platform, allowing clinicians to quickly calculate post-test probabilities when applying the likelihood ratios (LRs) in this book.

II. Using the Calculator

A. BLANK CALCULATOR

After opening the evidence-based medicine (EBM) calculator, the **Blank Calculator** appears (see Fig. 4.1). The blank calculator has three horizontal rules: **Pre-test probability, Likelihood ratio (LR)**, and **Post-test probability**, each with its own arrow. The clinician can move the arrows under the first two rules to indicate the appropriate pre-test probability and LR. Then, the third arrow (post-test probability) automatically displays the corresponding post-test probability. For example, dragging the pre-test probability arrow to 32% and LR arrow to 5 reveals the post-test probability to be approximately 70% (Fig. 4.1).

B. CALCULATING PROBABILITY FOR SPECIFIC CONDITIONS

If the clinician taps the arrow to the right of the box titled **Problem** (at the top of the calculator), a drop-down list of over 70 clinical problems will appear. By selecting any problem from this list, 2 additional items of information appear: (1) the pre-test probability for that particular clinical problem derived from the actual studies used in this book, with both the range and median pre-test probabilities displayed automatically on the first rule, and (2) a **View LR Value** button located in the upper right corner of the calculator (Fig. 4.2).

As an example, the clinician discovers the physical finding of *clubbing* in a patient with cirrhosis, a finding raising the possibility of hepatopulmonary syndrome (see Chapter 8). To use the calculator, the clinician first selects *Hepatopulmonary syndrome* from the drop-down list (Fig. 4.2), which changes the appearance of the **Pre-test probability** rule to display both the range and median pre-test probabilities (or prevalence) of hepatopulmonary syndrome in patients with cirrhosis derived from the studies in this book (i.e., range, 14% to 37%; median, 26%). In our example, however, the clinician using the calculator believes that the prevalence of hepatopulmonary syndrome in his or her own practice is slightly higher than the median (i.e., he or she believes it is about 30%). Therefore, the clinician sets the **Pre-test probability** arrow to 30%. Next, the clinician clicks on the **View LR Value** button (at the upper right) to reveal the EBM Box for Hepatopulmonary syndrome (from Chapter 8). This EBM Box reveals that the LR for clubbing

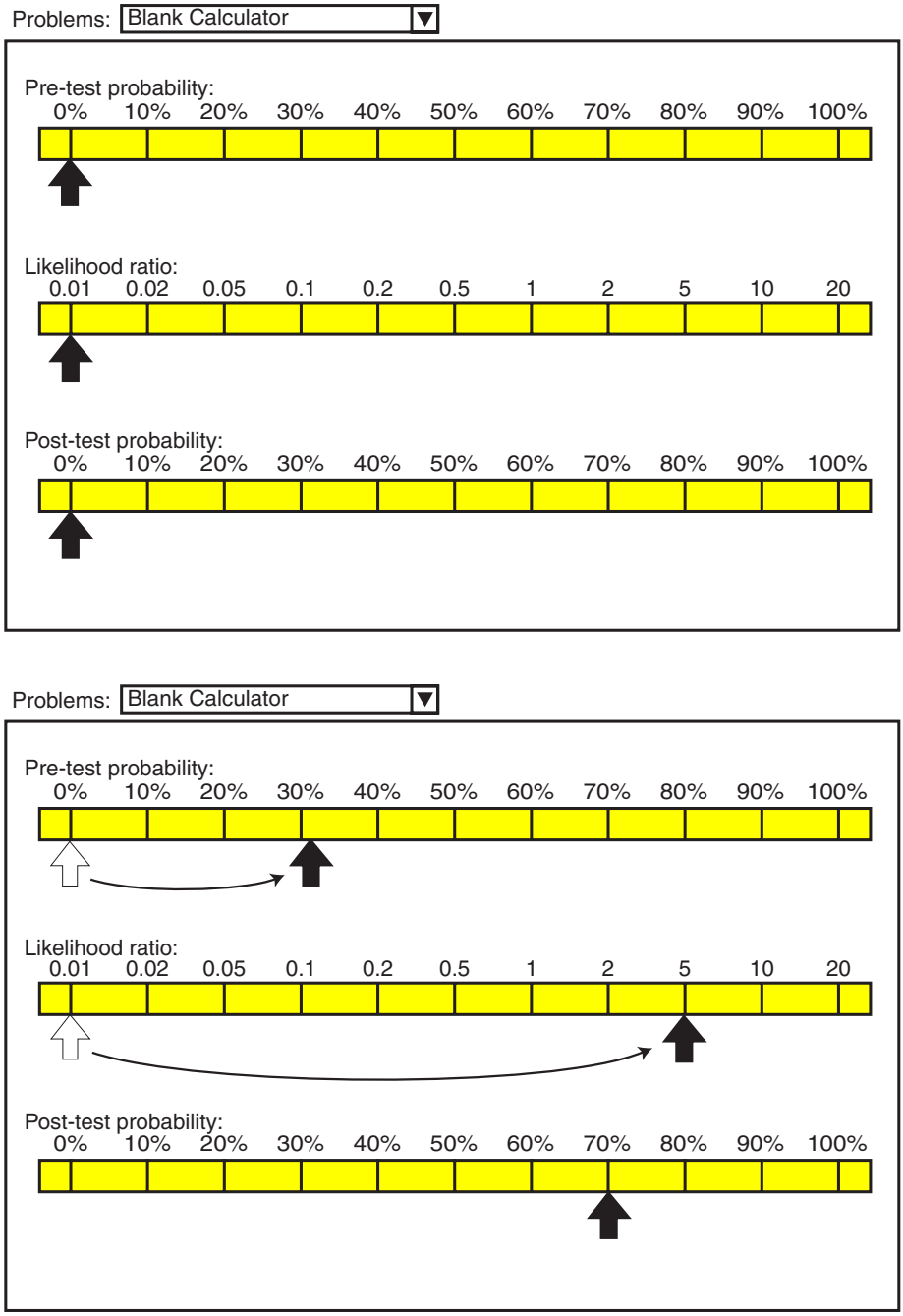


Fig. 4.1 Using the blank calculator. In this example, the clinician knows the pre-test probability is 32% and the finding's LR is 5. Therefore, the clinician drags the arrow under the first rule (*pre-test probability*) to 32% and the arrow under the second rule (*likelihood ratio*) to 5; the arrow under the third rule (*post-test probability*) automatically displays the corresponding post-test probability (70%).

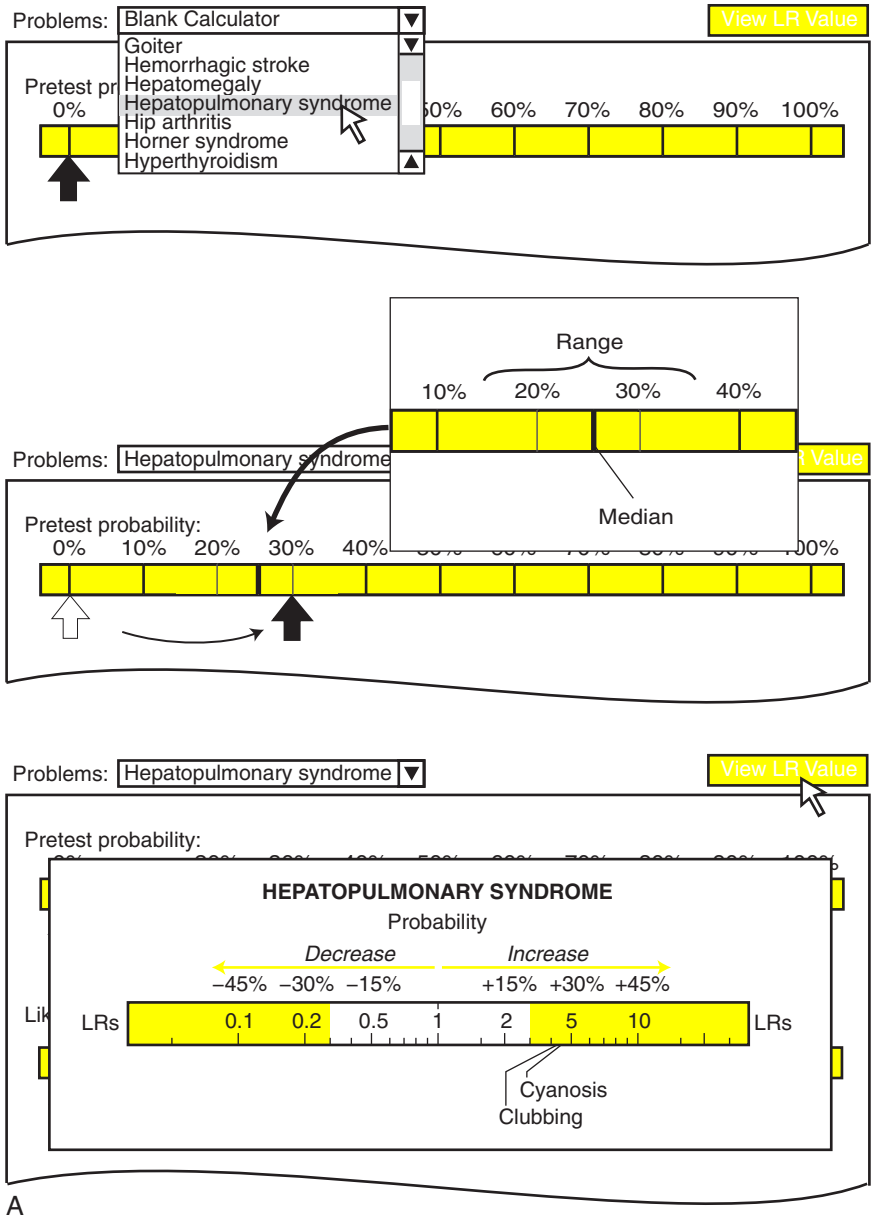
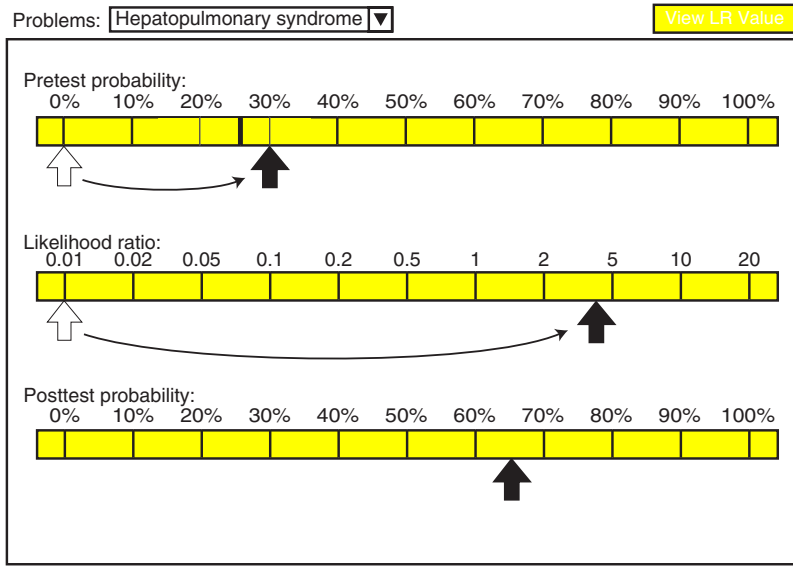


Fig. 4.2 Diagnosing hepatopulmonary syndrome with the EBM Calculator. The clinician is evaluating a patient with cirrhosis and clubbing and wonders about the likelihood of hepatopulmonary syndrome. Selecting *hepatopulmonary syndrome* (*top left*) reveals the pre-test probability in clinical studies ranges from 14% to 37%, with a median probability of 26% (*middle left*). Believing hepatopulmonary syndrome to be more prevalent in his own practice than 26%, the clinician drags the **pre-test probability arrow** to 30% (*middle left*), clicks *view LR value* (*bottom left*) to reveal the LR for clubbing (LR = 4.3). Dragging the **LR arrow** to 4.3 demonstrates the post-test probability of hepatopulmonary to be approximately 65% (*right*).



B

Fig. 4.2 *Continued*

is 4.3. After dragging the **LR arrow** to 4.3, the calculator indicates that the post-test probability of hepatopulmonary syndrome (in this clinician's patient with cirrhosis and clubbing) is 65% (Fig. 4.2).

Following the rules discussed in Chapter 2, the clinician may combine findings using this calculator by simply transferring the post-test probability from the first finding to the pre-test probability rule of the second finding. (See the section on Combining Findings in Chapter 2.)

CHAPTER 5

Reliability of Physical Findings

KEY TEACHING POINTS

- Reliability refers to how often two clinicians examining the same patient agree about the presence or absence of a particular physical finding. Commonly used measurements of reliability are *simple agreement* or the *kappa (κ-) statistic*.
- About 60% of physical findings have κ-statistics of 0.4 or more, indicating that observed agreement is moderately good or better.
- Despite the common belief that technologic tests are more precise than bedside observation, the κ-statistics observed for most diagnostic standards (e.g., chest radiography, computed tomography, angiography, magnetic resonance imaging, endoscopy, and pathology) are similar to those observed for physical signs.
- Some causes of interobserver disagreement can be eliminated, but because clinical medicine is inherently a human enterprise (even when interpreting technologic tests), subjectivity and a certain level of clinical disagreement will always be present.

Reliability refers to how often multiple clinicians, examining the same patients, agree that a particular physical sign is present or absent. As characteristics of a physical sign, reliability and accuracy are distinct qualities, although significant interobserver disagreement tends to undermine the finding's accuracy and prevents clinicians from applying it confidently to their own practice. Disagreement about physical signs also contributes to the growing sense among clinicians, not necessarily justified, that physical examination is less scientific than more technologic tests, such as clinical imaging and laboratory testing, and that physical examination lacks their diagnostic authority.

The most straightforward way to express reliability, or interobserver agreement, is **simple agreement**, which is the proportion of total observations in which clinicians agree about the finding. For example, if two clinicians examining 100 patients with dyspnea agree that a third heart sound is present in 5 patients and is absent in 75 patients, simple agreement would be 80% [i.e., $(5 + 75)/100 = 0.80$]; in the remaining 20 patients, only one of the two clinicians heard a third heart sound. Simple agreement has advantages, including being easy to calculate and understand, but a significant disadvantage is that agreement may be quite high by chance alone. For example, if one of the clinicians in our hypothetical study heard a third heart sound in 10 of the 100 dyspneic patients and the other heard it in 20 of the patients (even though they agreed about the presence of the heart sound in only 5 patients), simple agreement by chance *alone* would be 74%.* With chance agreement this high, the observed 80% agreement no longer seems so impressive.

To address this problem, most clinical studies now express interobserver agreement using the kappa (κ) statistic, which usually has values between 0 and 1 (the Appendix at the end of this

*Agreement by chance approaches 100% as the percentage of positive observations for both clinicians approaches 0% or 100% (i.e., both clinicians agree that a finding is very uncommon or very common). The Appendix at the end of this chapter shows how to calculate chance agreement.

chapter shows how to calculate the κ-statistic). A κ-value of 0 indicates that observed agreement is the same as that expected by chance, and a κ-value of 1 indicates perfect agreement. According to convention, a κ-value of 0 to 0.2 indicates *slight* agreement; 0.2 to 0.4 *fair* agreement; 0.4 to 0.6 *moderate* agreement; 0.6 to 0.8 *substantial* agreement; and 0.8 to 1.0 almost *perfect* agreement.[†]

TABLE 5.1 **Interobserver Agreement and Physical Signs**

Finding (ref)	κ-statistic*
General Appearance	
Mental status examination	
Mini-Mental Status Examination[1]	0.28–0.80
Clock-drawing test (Wolf-Klein Method)[2]	0.73
Confusion Assessment Method for delirium[3–6]	0.70–0.91
Altered mental status[7]	0.71
Stance and gait	
Abnormal gait[8,9]	0.11–0.71
Skin	
Patient appears anemic[10,11]	0.23–0.48
Nailbed pallor[12]	0.19–0.34
Conjunctival pallor (rim method)[13,14]	0.54–0.77
Palmar crease pallor[14]	0.44
Ashen or pale skin[7]	0.34
Cyanosis[10,15]	0.36–0.70
Jaundice[16]	0.65
Loss of hair[17]	0.51
Vascular spiders[16–18]	0.64–0.92
Palmar erythema[16–18]	0.37–1.00
Hydration status	
Patient appears dehydrated[10]	0.44–0.53
Axillary dryness[19]	0.50
Increased moisture on skin[10]	0.31–0.53
Capillary refill > 3 seconds[7]	0.29
Capillary refill > 5 seconds[20]	0.74–0.91
Nutritional assessment	
Abnormal nutritional state[10]	0.27–0.36
Other	
Consciousness impaired[10]	0.65–0.88
Patient appears older than age[10]	0.38–0.42
Patient appears in pain[10]	0.43–0.75
Generally unwell in appearance[10]	0.52–0.64
Vital Signs	
Tachycardia (heart rate >100/min)[21]	0.85
Bradycardia (heart rate <60/min)[21]	0.87
Systolic hypertension (SBP>160 mm Hg)[21]	0.75
Hypotension (SBP <90 mm Hg)[21,22]	0.27–0.90
Osler sign[23–25]	0.26–0.72
Rumpel-Leede ("tourniquet") test[26,27]	0.76–0.88
Elevated body temperature, palpating the skin[10]	0.09–0.23
Tachypnea[7,15,21]	0.25–0.60

[†]No measure of reliability is perfect, especially for findings whose prevalence clinicians agree approaches 0% or 100%. For these findings, simple agreement tends to overestimate reliability, and the κ-statistic tends to underestimate the reliability.

Continued

TABLE 5.1 ■ **Interobserver Agreement and Physical Signs—Cont'd**

Finding (ref)	κ-statistic*
Head and Neck	
Pupils	
Swinging flashlight test (relative afferent pupil defect)[28]	0.63
Diabetic retinopathy	
Microaneurysms[29,30]	0.58–0.66
Intraretinal hemorrhages[29,30]	0.89
Hard exudates[29,30]	0.66–0.74
Cotton wool spots[29,30]	0.56–0.67
Intraretinal microvascular abnormalities ("IRMA")[29,30]	0.46
Neovascularization near disc[29,30]	0.21–0.48
Macular edema[29,30]	0.21–0.67
Overall grade[29,30]	0.65
Hearing	
Whispered voice test[31,32]	0.16–1.0
Finger rub test[33]	0.83
Thyroid	
Thyroid gland diffuse, multinodular or solitary nodule[34]	0.25–0.70
Goiter[35,36]	0.38–0.77
Meninges	
Nuchal rigidity, present or absent[37–39]	0.24–0.76
Lungs	
Inspection	
Clubbing (general impression)[15,40]	0.33–0.45
Clubbing (interphalangeal depth ratio)[41]	0.98
Clubbing (Schamroth sign)[41]	0.64
Breathing difficulties[10]	0.54–0.69
Gasping respirations[7]	0.63
Reduced chest movement[15,42,43]	0.14–0.38
Kussmaul respirations[44]	0.70
Pursed lip breathing[43]	0.45
Asymmetric chest expansion[45]	0.85
Scalene or sternocleidomastoid muscle contraction[7,43,46]	0.52–0.57
Kyphosis[40]	0.37
Barrel chest[43]	0.62
Thoracic ratio ≥0.9[43]	0.32
Displaced trachea[15]	0.01
Palpation	
Tracheal descent during inspiration[46]	0.62
Laryngeal height ≤5.5 cm[43]	0.59
Impalpable apex beat[15,40]	0.33–0.44
Decreased tactile fremitus[15,45]	0.24–0.86
Increased tactile fremitus[15]	0.01
Subxiphoid point of maximal cardiac impulse[47]	0.30
Paradoxical costal margin movement[46,48]	0.56–0.82
Percussion	
Hyperresonant percussion note[15,42,47]	0.26–0.50
Dull percussion note[15,42,45,49]	0.16–0.84
Diaphragm excursion more or less than 2 cm by percussion[47]	–0.04
Diminished cardiac dullness[47]	0.49
Auscultatory percussion abnormal[45,50]	0.18–0.76

Continued

TABLE 5.1 **Interobserver Agreement and Physical Signs—Cont'd**

Finding (ref)	κ-statistic*
Auscultation	
Reduced breath sound intensity[15,42,43,45,47,49,51,52]	0.16–0.89
Bronchial breathing[15,42]	0.19–0.32
Whispering pectoriloquy[15]	0.11
Reduced vocal resonance[45]	0.78
Crackles[15,49,51,53–56]	0.21–0.65
Wheezes[15,47,49,51,52]	0.43–0.93
Rhonchi[42,52]	0.38–0.55
Pleural rub[15,45]	−0.02–0.51
Special tests	
Snider test <10 cm[47]	0.39
Forced expiratory time[43,47,57,58]	0.27–0.70
Hoover sign[52]	0.74
Wells simplified rule for pulmonary embolism[59]	0.54–0.62
Heart	
Neck veins	
Neck veins, elevated or normal[53–55,60,61]	0.08–0.71
Abdominojugular test[60]	0.92
Palpation	
Palpable apical impulse present[62–64]	0.68–0.82
Palpable apical impulse measureable[65]	0.56
Palpable apical impulse displaced lateral to midclavicular line[53,62,63,66]	0.43–0.86
Apical beat normal, sustained, double, or absent[66]	0.88
Palpable right ventricular heave[61]	0.18–0.23
Percussion	
Cardiac dullness >10.5 cm from midsternal line[67,68]	0.57
Auscultation	
S2 diminished or absent, vs. normal[69]	0.54
Third heart sound[53–55,60,70–72]	−0.17–0.84
Fourth heart sound[71,73]	0.15–0.71
Systolic murmur, present or absent[69]	0.19
Systolic murmur radiates to right carotid[69]	0.33
Systolic murmur, long systolic or early systolic[74]	0.78
Murmur intensity (Levine grade)[75]	0.43–0.60
Systolic murmur grade >2/6[76]	0.59
Carotid pulsation	
Delayed carotid upstroke[69]	0.26
Reduced carotid volume[69]	0.24
Abdomen	
Inspection	
Abdominal distention[77,78]	0.35–0.42
Abdominal wall collateral veins, present vs. absent[16]	0.47
Palpation and percussion	
Ascites[16,18,55]	0.47–0.75
Abdominal tenderness[77–79]	0.31–0.68
Surgical abdomen[78]	0.27
Abdominal wall tenderness test[80,81]	0.52–0.81
Rebound tenderness[77]	0.25
Guarding[77,78]	0.36–0.49
Rigidity[77]	0.14
Abdominal mass palpated[78]	0.82

TABLE 5.1 ■ **Interobserver Agreement and Physical Signs—Cont'd**

Finding (ref)	κ-statistic*
Palpable spleen[16,18,82]	0.33–0.75
Palpable liver edge[83]	0.44–0.53
Liver consistency, normal or abnormal[16]	0.4
Liver firm to palpation[84]	0.72
Liver, nodular or not[16]	0.29
Liver, tender or not[18]	0.49
Liver, span >9 cm by percussion[53]	0.11
Spleen palpable or not[85]	0.56–0.70
Spleen percussion sign (Traube), positive or not[86]	0.19–0.41
Spleen percussion sign (Castell), positive or not[82]	0.45
Abdominal aortic aneurysm, present vs. absent[87]	0.53
Auscultation	
Normal bowel sounds[78]	0.36
Extremities	
Peripheral vascular disease	
Peripheral pulse, present vs. absent[88–91]	0.52–0.92
Peripheral pulse, normal or diminished[88]	0.01–0.15
Cool extremities[55]	0.46
Severity of skin mottling over leg[92,93]	0.87
Diabetic foot	
Monofilament sensation, normal or abnormal[94–96]	0.48–0.83
Probe-to-bone test[97–99]	0.59–0.84
Edema and deep venous thrombosis	
Dependent edema[53–55]	0.39–0.73
Wells pretest probability for deep vein thrombosis[100,101]	0.74–0.75
Musculoskeletal system-shoulder	
Shoulder tenderness[102]	0.32
Painful arc[102–105]	0.45–0.64
Neer impingement sign[106]	0.64
Hawkins impingement sign[106]	0.54
External rotation of shoulder <45 degrees[102]	0.68
Supraspinatus test (empty can)[102,105,107]	0.44–0.94
Infraspinatus test (resisted external rotation)[102,103]	0.49–0.67
Impingement sign (Hawkins-Kennedy)[102,103,105,107]	0.29–1.0
Drop arm test[102,105]	0.28–0.35
Musculoskeletal system-hip	
Patrick test[108]	0.47
Passive internal rotation ≤25 degrees[108]	0.51
Musculoskeletal system-knee	
Ottawa knee rules[109,110]	0.51–0.77
Knee effusion visible[109,111–113]	0.28–0.78
Knee flexion <90 degrees[109]	0.74
Patellar tenderness [109,111]	0.69–0.76
Head of fibula tenderness[109]	0.64
Inability to bear weight immediately and emergency room after knee injury[109,111]	0.75–0.81
Bony swelling of knee[113,114]	0.55–0.66
Joint line tenderness[112,114–116]	0.11–0.43
Patellofemoral crepitus[114]	0.24
Mediolateral instability of knee[114]	0.23
McMurray sign[112,116,117]	0.16–0.35
Lachman test[118]	0.72

Continued

TABLE 5.1 **Interobserver Agreement and Physical Signs—Cont'd**

Finding (ref)	κ-statistic*
Musculoskeletal system-ankle	
Inability to walk 4 steps immediately and in emergency room after ankle injury[119,120]	0.71–0.97
Medial malleolar tenderness[120]	0.82
Lateral malleolar tenderness[120]	0.80
Navicular tenderness[120]	0.91
Base of 5th metatarsal tenderness[120]	0.94
Ottawa ankle rule[121,122]	0.41–0.45
Ottawa midfoot rule[121]	0.77
Neurologic Examination	
Visual fields	
Visual fields by confrontation[123]	0.63–0.81
Cranial nerves	
Pharyngeal sensation, present or absent[124]	1.0
Facial palsy, present or absent[125,126]	0.57
Dysarthria, present or absent[127,128]	0.41–0.77
Water swallow test (50 mL)[129]	0.60
Oxygen desaturation test (for aspiration risk)[129]	0.60
Abnormal tongue strength[127]	0.55–0.63
Motor examination	
Muscle strength, MRC scale[130–133]	0.69–0.93
Foot tapping test[134,135]	0.73–0.83
Muscle atrophy[136,137]	0.32–0.82
Spasticity, 6 point scale[138]	0.21–0.61
Rigidity, 4 point scale[139]	0.64
Asterixis[16]	0.42
Tremor[137]	0.74
Pronator drift[140]	0.39
Forearm rolling test[140]	0.73
Sensory examination	
Light touch sensation, normal, diminished, or increased[136,137]	0.22–0.63
Pain sensation, normal, diminished, or increased[131,136,137]	0.41–0.57
Vibratory sensation, normal or diminished[136,137]	0.28–0.54
Romberg test[137]	0.64
Reflex examination	
Reflex amplitude, NINDS scale[141]	0.51–0.61
Ankle jerk, present or absent[131,142,143]	0.34–0.94
Asymmetric knee jerk[131]	0.42
Babinski response[125,126,134,135,137,144,145]	0.17–0.60
Finger flexion reflex[146]	0.65
Palmomental reflex[147]	0.53
Primitive reflexes, amplitude and persistence[148]	0.46–1.0
Coordination	
Finger-nose test[125,126,137,140]	0.14–0.65
Heel-shin test[137]	0.58
Peripheral nerve	
Spurling test[149]	0.60
Katz hand diagram[150]	0.86
Flick sign[151]	0.90
Hypalgesia index finger[151]	0.50
Tinel sign[151]	0.47
Phalen sign[151]	0.79

TABLE 5.1 ■ **Interobserver Agreement and Physical Signs—Cont'd**

Finding (ref)	κ-statistic*
Straight leg raising test[131,152–156]	0.21–0.80
Crossed leg raising test[131]	0.49
Other	
Head impulse test[157]	0.86
Knee lift test (for nonorganic weakness)[158]	0.91

*Interpretation of the κ-statistic: 0 to 0.2 slight agreement, 0.2 to 0.4 fair agreement, 0.4 to 0.6 moderate agreement, 0.6 to 0.8 substantial agreement, 0.8 to 1.0 almost perfect agreement.
MRC, Medical Research Council; *NINDS*, National Institute of Neurological Disorders and Stroke.

Rarely, physical signs have κ-values less than 0 (theoretically as low as –1), indicating the observed agreement was worse than chance agreement.

Table 5.1 presents the κ-statistic for most of the physical signs discussed in this book, demonstrating that, with rare exceptions, observed agreement is better than chance agreement (i.e., κ-statistic exceeds 0). About 60% of findings have a κ-statistic of 0.4 or more, indicating that observed agreement is moderate or better.

Clinical disagreement occurs for many reasons—some causes clinicians can control, but others are inextricably linked to the very nature of clinical medicine and human observation in general. The most prominent reasons include the following: (1) The physical sign's definition can be vague or ambiguous. For example, experts recommend about a dozen different ways to perform auscultatory percussion of the liver, thus making the sign so nebulous that significant interobserver disagreement is guaranteed. Ambiguity also results if signs are defined with terms that are not easily measurable. For example, clinicians assessing whether a peripheral pulse is present or absent demonstrate moderate-to-almost perfect agreement ($\kappa = 0.52–0.92$, Table 5.1), but when the same clinicians are asked to record whether the palpable pulse is normal or diminished, they have great difficulty agreeing about the sign ($\kappa = 0.01–0.15$) simply because they have no idea what the next clinician means by "diminished." (2) The clinician's technique is flawed. For example, common mistakes are using the diaphragm instead of the bell of the stethoscope to detect the third heart sound, or stating a muscle stretch reflex is absent without first trying to elicit it using a reinforcing maneuver (e.g., Jendrassik maneuver). (3) Biologic variation of the physical sign. The pericardial friction rub, pulsus alternans, cannon A waves, Cheyne-Stokes respirations, and many other signs are notoriously evanescent, tending to come and go over time. (4) The clinician could be careless or inattentive. The bustle of an active practice may lead clinicians to listen to the lungs while conducting the patient interview, or to search for a subtle murmur in a noisy emergency room. Reliable observations require undistracted attention and an alert mind. (5) The clinician's biases can influence the observation. When findings are equivocal, expectations influence perceptions. For example, in a patient who just started blood pressure medications, borderline hypertension may become normal blood pressure; in a patient with increasing bilateral edema, borderline distended neck veins may become clearly elevated venous pressure; or in a patient with new weakness, the equivocal Babinski sign may become clearly positive. Sometimes, biases actually create the finding: if the clinician holds a flashlight too long over an eye with suspected optic nerve disease, he or she may temporarily bleach the retina of that eye and produce the Marcus Gunn pupil, thus confirming the original suspicion.

The lack of perfect reliability with physical diagnosis is sometimes regarded as a significant weakness, leading to the charge that physical diagnosis is less reliable and scientific than clinical imaging and laboratory testing. Nonetheless, Table 5.2 shows that, for most of our **diagnostic standards**—chest radiography, computed tomography, screening mammography, angiography, magnetic resonance imaging, ultrasonography, endoscopy, and pathology—interobserver