CAMBRIDGE

Cambridge University Press
978-0-521-88652-9 - Evidence-Based Diagnosis
Thomas B. Newman and Michael A. Kohn
Excerpt
More information

1

Introduction: understanding diagnosis and diagnostic testing

Two areas of evidence-based medicine: diagnosis and treatment

The term "evidence-based medicine" (EBM) was coined by Gordon Guyatt and colleagues of McMaster University around 1992 (Evidence-Based Medicine Working Group 1992). Oversimplifying greatly, EBM is about using the best available evidence to help in two related areas:

- Diagnosis: How to evaluate a test and then use it to estimate the probability that a patient has a given disease.
- Treatment: How to determine whether a treatment is beneficial in patients with a given disease, and if so, whether the benefits outweigh the costs and risks.

The two areas of evidence-based diagnosis and treatment are closely related. Although a diagnosis can be useful for prognosis, epidemiologic tracking, and scientific study, it may not be worth the costs and risks of testing to diagnose a disease that has no effective treatment. Even if an effective treatment exists, there are probabilities of disease so low that it is not worth testing for the disease. These probabilities depend not only on the cost and accuracy of the test, but also on the effectiveness of the treatment. As suggested by the title, this book focuses more intensively on the diagnosis area of EBM.

The purpose of diagnosis

When we think about diagnosis, most of us think about a sick person going to the health care provider with a collection of signs and symptoms of illness. The provider, perhaps with the help of some tests, identifies the cause of the patient's illness, then tells the patient the name of the disease and what to do to treat it. Making the diagnosis can help the patient by explaining what is happening, predicting the prognosis, and determining treatments. In addition, it can benefit others by establishing the level of infectiousness to prevent the spread of disease, tracking the burden of disease and the success of disease control efforts, discovering etiologies to prevent future cases, and advancing medical science.

Assigning each illness a diagnosis is one way that we attempt to impose order on the chaotic world of signs and symptoms, grouping them into categories that share various characteristics, including etiology, clinical picture, prognosis, mechanism of transmission, and response to treatment. The trouble is that homogeneity with respect to one of these characteristics does not imply homogeneity with respect to the others. So if we are trying to decide how to diagnose a disease, we need to know why we want to make the diagnosis, because different purposes of diagnosis can lead to different disease classification schemes.

For example, entities with different etiologies or different pathologies may have the same treatment. If the goal is to make decisions about treatment, the etiology or pathology may be irrelevant. Consider a child who presents with puffy eyes, excess fluid in the ankles, and a large amount of protein in the urine – a classic presentation of the nephrotic syndrome. When the authors were in medical school, we dutifully learned how to classify nephrotic syndrome in children by the appearance of the kidney biopsy. There were minimal change disease, focal segmental glomerulosclerosis, membranoproliferative glomerulonephritis, and so on. "Nephrotic syndrome" was not considered a *diagnosis*; a kidney biopsy to determine the type of nephrotic syndrome was felt to be necessary.

However, minimal change disease and focal segmental glomerulosclerosis make up the overwhelming majority of nephrotic syndrome cases in children, and both are treated with corticosteroids. So, although a kidney biopsy would provide prognostic information, current recommendations suggest skipping the biopsy initially, starting steroids, and then doing the biopsy later only if the symptoms do not respond or there are frequent relapses. Thus, if the purpose of making the diagnosis is to guide treatment, the pathologic classification that we learned in medical school is usually irrelevant. Instead, nephrotic syndrome is classified as "steroid responsive or non-responsive" and "relapsing or non-relapsing." If, as is usually the case, it is "steroid-responsive and non-relapsing," we will never know whether it was minimal change disease or focal segmental glomerulosclerosis, because it is not worth doing a kidney biopsy to find out.

There are many similar examples where, at least at some point in an illness, an exact diagnosis is not necessary to guide treatment. We have sometimes been amused by the number of different Latin names there are for certain similar skin conditions, all of which are treated with topical steroids, which makes distinguishing between them rarely necessary from a treatment standpoint. And, although it is sometimes interesting for an emergency physician to determine which knee ligament is torn, "acute ligamentous knee injury" is a perfectly adequate emergency department diagnosis, because the treatment is immobilization, ice, analgesia, and orthopedic follow-up, regardless of the specific ligament injured.

Disease classification systems sometimes have to expand as treatment improves. Before the days of chemotherapy, a pale child with a large number of blasts (very immature white blood cells) on the peripheral blood smear could be diagnosed simply with leukemia. That was enough to determine the treatment (supportive) and the prognosis (grim) without any additional tests. Now, there are many different types of leukemia based, in part, on cell surface markers, each with a specific prognosis and treatment schedule. The classification based on cell surface markers has no *inherent* value; it is valuable only because careful studies have shown that these markers predict prognosis and response to treatment.

If one's purpose is to keep track of disease burden rather than to guide treatment, different classification systems may be appropriate. Diseases with a single cause (e.g., the diseases caused by premature birth) may have multiple different treatments. A classification system that keeps such diseases separate is important to clinicians, who will, for example, treat necrotizing enterocolitis very differently from hyaline membrane disease. On the other hand, epidemiologists and health services researchers looking at neonatal morbidity and mortality may find it most advantageous to group these diseases with the same cause together, because if we can successfully prevent prematurity, then the burden from all of these diseases should decline. Similarly, an epidemiologist in a country with poor sanitation might only want to group together diarrheal diseases by mode of transmission, rather than keep separate statistics on each of the responsible organisms, in order to track success of sanitation efforts. Meanwhile, a clinician practicing in the same area might only need to sort out whether a case of diarrhea needs antibiotics, an antiparasitic drug, or just rehydration.

In this text, when we are discussing how to quantify the information and usefulness of diagnostic tests, it will be helpful to clarify why a diagnostic test is being considered – what decisions is it supposed to help us with?

Definition of "disease"

In the next several chapters, we will be discussing tests to diagnose "disease." What do we mean by that term? We use the term "disease" for *a health condition that is either already causing illness or is likely to cause illness relatively quickly in the absence of treatment.* If the disease is not currently causing illness, it is presymptomatic. For most of this book, we will assume that the reason for diagnosis is to make treatment decisions, and that diagnosing disease is important for treatment decisions because there are treatments that are beneficial in those who have a disease and not beneficial in those who do not.

For example, acute coronary ischemia is an important cause of chest pain in patients presenting to the emergency department. Its short-term mortality is 25% among patients discharged from the emergency department and only 12.5% among patients who are hospitalized (presumably for monitoring, anticoagulation, and possibly thrombolysis or angioplasty) (Lee and Goldman 2000). Patients presenting to the emergency department with nonischemic chest pain do not have high short-term mortality and do not benefit from hospitalization. Therefore, a primary purpose of diagnosing acute coronary ischemia is to help with the treatment decision about whom to hospitalize.

The distinction between a disease and a risk factor depends on what we mean by "likely to cause illness relatively quickly." Congenital hypothyroidism, if untreated, will lead to severe retardation, even though this effect is not immediate. This is clearly a disease, but because it is not already causing illness, it is considered presymptomatic. On the other hand, mild hypertension is asymptomatic and is important mainly as a risk factor for heart disease and stroke, which are diseases. At some point, however, hypertension can become severe enough that it is likely to cause illness relatively quickly (e.g., a blood pressure of 280/140) and is a disease. Similarly, diseases like hepatitis C, ductal carcinoma in situ of the breast, and prostate cancer illustrate that there is a continuum between risk factors, presymptomatic diseases, and diseases that depends on the likelihood and timing of subsequent illness. For the first five chapters of this book we will be focusing on "prevalent" disease – that is, diseases that patients either do or do not have at the time we test them. In Chapter 6 on screening, we will distinguish between screening for unrecognized symptomatic disease[1], presymptomatic disease, and risk factors. In Chapter 7 on prognostic testing, we will also consider incident diseases and outcomes of illness – that is, diseases and other outcomes (like death) that are not yet present at the time of the test, but have some probability of happening later, that may be predicted by the results of a test for risk factors or prognosis.

Relative to patients without the disease, patients with the disease have a high risk of bad outcomes and will benefit from treatment. We still need to quantify that benefit so we can compare it with the risks and costs of treatment. Because we often cannot be certain that the patient has the disease, there may be some threshold probability of disease at which the projected benefits of treatment will outweigh the risks and costs. In Chapters 3 through 8, we will sometimes assume that such a treatment threshold probability exists. In Chapter 9, when we discuss quantifying the benefits (and harms) of treatments, we will show how they determine the treatment threshold probability of disease.

Dichotomous disease state (D+/D−): a convenient oversimplification

Most discussions of diagnostic testing, including this one, simplify the problem of diagnosis by assuming a dichotomy between those with a particular disease and those without the disease. The patients with disease, that is, with a positive diagnosis, are denoted "D+," and the patients without the disease are denoted "D−." This is an oversimplification for two reasons. First, there is usually a spectrum of disease. Some patients we label D+ have mild or early disease, and other patients have severe or advanced disease; so instead of D+, we should have D+, D++, and D+++. Second, there is usually a spectrum of those who do not have the disease (D−) that includes other diseases as well as varying states of health. Thus for symptomatic patients, instead of D+ and D−, we should have D1, D2, and D3, each potentially at varying levels of severity, and for asymptomatic patients we will have D− as well.

[1] It is possible for patients with a disease (e.g., depression, deafness, anemia) to have symptoms that they do not recognize.

For example, a patient with prostate cancer might have early, localized cancer or widely metastatic cancer. A test for prostate cancer, the prostate-specific antigen, is much more likely to be positive in the case of metastatic cancer. Or consider the patient with acute headache due to subarachnoid hemorrhage. The hemorrhage may be extensive and easily identified by computed tomography scanning, or it might be a small sentinel bleed, unlikely to be identified by computed tomography and identifiable only by lumbar puncture.

Even in patients who do not have the disease in question, a multiplicity of potential diagnoses of interest may exist. Consider a young woman with lower abdominal pain and a positive urine pregnancy test. The primary concern is ectopic pregnancy, but women without ectopic pregnancies may have either normal or abnormal intrauterine pregnancies. One test commonly used in these patients, the β-HCG, is unlikely to be normal in patients with abnormal intrauterine pregnancies, even though they do not have ectopic pregnancies (Kohn et al. 2003). For patients presenting to the emergency department with chest pain, acute coronary ischemia is not the only serious diagnosis to consider; aortic dissection and pulmonary embolism are among the other serious causes of chest pain, and these have very different treatments from acute coronary ischemia.

Thus, dichotomizing disease states can get us into trouble because the composition of the D+ group (which includes patients with differing severity of disease) as well as the D− group (which includes patients with differing distributions of other diseases) can vary from one study and one clinical situation to another. This, of course, will affect results of measurements that we make on these groups (like the distribution of prostate-specific antigen results in men with prostate cancer or of β-HCG results in women who do not have ectopic pregnancies). So, although we will generally assume that we are testing for the presence or absence of a single disease and can therefore use the D+/D− shorthand, we will occasionally point out the limitations of this assumption.

Generic decision problem: examples

We will start out by considering an oversimplified, generic medical decision problem in which the patient either has the disease (D+) or doesn't have the disease (D−). If he has the disease, there is a quantifiable benefit to treatment. If he doesn't have the disease, there is an equally quantifiable cost associated with treating unnecessarily. A single test is under consideration. The test, although not perfect, provides information on whether the patient is D+ or D−. The test has two or more possible results with different probabilities in D+ individuals than in D− individuals. The test itself has an associated cost. The choice is to treat without the test, do the test and determine treatment based on the result, or forego treatment without testing.

Here are several examples of the sorts of clinical scenarios that material covered in this book will help you understand better. In each scenario, the decision to be made includes whether to treat without testing, to do the test and treat based on testing results, or to do nothing (neither test nor treat). We will refer to these scenarios throughout the book.

Clinical scenario: febrile infant

A 5-month-old boy has a fever of 39.7°C with no obvious source. You are concerned about a bacterial infection in the blood (bacteremia), which could be treated with an injection of an antibiotic. You have decided that other tests, such as a lumbar puncture or chest x-ray, are not indicated, but you are considering drawing blood to use his white blood cell count to help you decide whether or not to treat.

Disease in question: Bacteremia.
Test being considered: White blood cell count.
Treatment decision: Whether to administer an antibiotic injection.

Clinical scenario: screening mammography

A 45-year-old economics professor from a local university wants to know whether she should get screening mammography. She has not detected any lumps on breast self-examination. A positive screening mammogram would be followed by further testing, possibly including biopsy of the breast.

Disease in question: Breast cancer.
Test being considered: Mammogram.
Treatment decision: Whether to pursue further evaluation for breast cancer.

Clinical scenario: flu prophylaxis

It is the flu season, and your patient is a 14-year-old girl who has had fever, muscle aches, cough, and a sore throat for two days. Her mother has seen a commercial for Tamiflu® (oseltamivir) and asks you about prescribing it for the whole family so they don't catch the flu. You are considering testing the patient with a rapid bedside test for influenza A and B.

Disease in question: Influenza.
Test being considered: Bedside influenza test for 14-year-old patient.
Treatment decision: Whether to prescribe prophylaxis for others in the household. (You assume that prophylaxis is of little use if your 14-year-old patient does not have the flu but rather some other viral illness.)

Clinical scenario: sonographic screening for fetal chromosomal abnormalities

In late first-trimester pregnancies, fetal chromosomal abnormalities can be identified definitively using chorionic villus sampling, but this is an invasive procedure that entails some risk of accidentally terminating the pregnancy. Chromosomally abnormal fetuses tend to have larger nuchal translucencies (a measurement of fluid at the back of the fetal neck) and absence of the nasal bone on 13-week ultrasound, which is a noninvasive test. A government perinatal screening program faces the question of who should receive the screening ultrasound examination and what combination of nuchal translucency and nasal bone examination results should prompt chorionic villus sampling.

Disease in question: Fetal chromosomal abnormalities.
Test being considered: Prenatal ultrasound.
Treatment decision: Whether to do the definitive diagnostic test, chorionic villus sampling.

Preview of coming attractions: criteria for evaluating diagnostic tests

In the next several chapters, we will consider four increasingly stringent criteria for evaluating diagnostic tests:

1. *Reliability*. The results of a test should not vary based on who does the test or where it is done. This consistency must be maintained whether the test is repeated by the same observer or by different observers and in the same location or different locations. To address this question for tests with continuous results, we need to know about things like the within-subject standard deviation, within-subject coefficient of variation, correlation coefficients (i.e., why they can be misleading), and Bland–Altman plots. For tests with categorical results (e.g., normal/abnormal), we should understand kappa and its alternatives. These measures of reliability are discussed in Chapter 2. If you find that repetitions of the test do not agree any better than would be expected by chance alone, you might as well stop: the test cannot be informative or useful. But simply performing better than would be expected by chance, although necessary, is not sufficient to establish that the test is worth doing.
2. *Accuracy*. The test must be able to distinguish between those with disease and those without disease. We will measure this ability using sensitivity and specificity, test-result distributions in D+ and D− individuals, Receiver Operating Characteristic curves, and likelihood ratios. These are covered in Chapters 3 and 4. Again, if a test is not accurate, you can stop; it cannot be useful. It is also important to be able to tell how much new information a test gives. For this we will need to address the concept of "test independence" (Chapter 8).
3. *Usefulness*. In general, this means that the test result should have a reasonable probability of changing decisions (which we have assumed will, on average, benefit the patient).[2] To determine this, we need to know how to combine our test results with other information to estimate the probability of disease, and at what probability of disease different treatments are indicated (Chapters 3, 4, 7, and 8).

   The best proof of a test's usefulness is a randomized trial showing that clinical outcomes are better in those who receive the test than in those who do not. Such trials are mainly practical to evaluate certain screening tests, which we will discuss in Chapter 6. If a test improves outcomes in a randomized trial, we can infer that it meets the three preceding criteria (reliability, accuracy, and usefulness). Similarly, if it clearly does not meet these criteria, there is no point in doing a randomized trial. More often, tests provide some information and their usefulness will vary depending on specific characteristics of the patient that determine the prior probability of disease.

[2] This discussion simplifies the situation by focusing on value provided by affecting management decisions. In fact, there may be some value to providing information even if it does not lead to changes in management (as demonstrated by peoples' willingness to pay for testing). Just attaching a name to the cause of someone's suffering can have a therapeutic effect. However, giving tests credit for these effects, like giving drugs credit for the placebo effect, is problematic.

4. *Value*. Even if a test is potentially useful, it may not be worth doing if it is too expensive, painful, or difficult to do. Although we will not formally cover cost-effectiveness analysis or cost-benefit analysis in this text, we will informally show (in Chapters 3 and 9) how the cost of a test can be weighed against possible benefits from better treatment decisions.

Summary of key points

1. There is no single best way to classify people into those with different diagnoses; the optimal classification scheme depends on the purpose for making the diagnosis.
2. We will initially approach evaluation of diagnostic testing by assuming that the patient either does or does not have a particular disease, and that the purpose of diagnosing the disease is to decide whether or not to treat. We must recognize, in doing this, that dichotomization of disease states may inaccurately represent the spectrum of both disease and nondisease, and that diagnosis of disease may have purposes other than treatment.
3. Increasingly stringent criteria for evaluating tests include reliability, accuracy, usefulness, and value.

References

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

Kohn, M. A., K. Kerr, et al. (2003). “Beta-human chorionic gonadotropin levels and the likelihood of ectopic pregnancy in emergency department patients with abdominal pain or vaginal bleeding.” *Acad Emerg Med* **10**(2): 119–26.

Lee, T. H., and L. Goldman (2000). “Evaluation of the patient with acute chest pain.” *N Engl J Med* **342**(16): 1187–95.

Chapter 1 Problems: understanding diagnosis

1. In children with apparent viral gastroenteritis (vomiting and diarrhea), a test for rotavirus is often done. No specific antiviral therapy for rotavirus is available, but rotavirus is the most common cause of hospital-acquired diarrhea in children and is an important cause of acute gastroenteritis in children attending child care. A rotavirus vaccine is recommended by the CDC’s Advisory Committee on Immunization Practices. Why would we want to know if a child’s gastroenteritis is caused by rotavirus?
2. Randomized trials (Campbell 1989; Forsyth 1989; Lucassen et al. 2000) suggest that about half of formula-fed infants with colic respond to a change from cow’s milk formula to formula in which the protein has been hydrolyzed. Colic in these studies (and in textbooks) is generally defined as crying at least 3 hours per day at least three times a week in an otherwise well infant. You are seeing a distressed mother of a formula-fed 5-week-old who cries inconsolably for about 1 to 2 hours

daily. Your physical examination is normal. Does this child have colic? Would you recommend a trial of hydrolyzed-protein formula?

3. An 86-year-old man presents with weight loss for 2 months and worsening shortness of breath for 2 weeks. An x-ray shows a left pleural effusion; thoracocentesis shows undifferentiated carcinoma. History, physical examination, and routine laboratory tests do not disclose the primary cancer. Could "metastatic undifferentiated carcinoma" be a sufficient diagnosis, or are additional studies needed? Does your answer change if he is demented?
4. Your patient, an otherwise healthy 20-month-old girl, was diagnosed with a urinary tract infection at an urgent care clinic where she presented with fever last week. At this writing, the American Academy of Pediatrics recommends a voiding cystourethrogram (VCUG, an x-ray preceded by putting a catheter through the urethra into the bladder to instill contrast) to evaluate the possibility that she has vesicoureteral reflux (reflux of urine from the bladder up the ureters to the kidneys) (AAP 1999). A diagnosis of vesicoureteral reflux typically leads to two alterations in management: 1) prophylactic antibiotics and 2) additional VCUGs. However, there is no evidence that prophylactic antibiotics are effective at decreasing infections or scarring in this setting (Hodson et al. 2007), and they may even be harmful (Garin et al. 2006, Newman 2006). How does this information affect the decision to do a VCUG in order to diagnose reflux?

References for problem set

AAP (1999). "Practice parameter: the diagnosis, treatment, and evaluation of the initial urinary tract infection in febrile infants and young children. American Academy of Pediatrics. Committee on Quality Improvement. Subcommittee on Urinary Tract Infection." *Pediatrics* **103**(4 Pt 1): 843–52.

Campbell, J. P. (1989). "Dietary treatment of infant colic: a double-blind study." *J R Coll Gen Pract* **39**(318): 11–4.

Forsyth, B. W. (1989). "Colic and the effect of changing formulas: a double-blind, multiple-crossover study." *J Pediatr* **115**(4): 521–6.

Garin, E. H., F. Olavarria, et al. (2006). "Clinical significance of primary vesicoureteral reflux and urinary antibiotic prophylaxis after acute pyelonephritis: a multicenter, randomized, controlled study." *Pediatrics* **117**(3): 626–32.

Hodson, E. M., D. M. Wheeler, et al. (2007). "Interventions for primary vesicoureteric reflux." *Cochrane Database Syst Rev 3*: CD001532.

Lucassen, P. L., W. J. Assendelft, et al. (2000). "Infantile colic: crying time reduction with a whey hydrolysate: a double-blind, randomized, placebo-controlled trial." *Pediatrics* **106**(6): 1349–54.

Newman, T. B. (2006). "Much pain, little gain from voiding cystourethrograms after urinary tract infection." *Pediatrics* **118**(5): 2251.

2

Reliability and measurement error

Introduction

A test cannot be useful unless it gives the same or similar results when administered repeatedly to the same individual within a time period too short for real biological changes to take place. Consistency must be maintained whether the test is repeated by the same measurer or by different measurers. This desirable characteristic of a test is generally called "reliability," although some authors prefer "reproducibility." In this chapter, we will look at several different ways to quantify reliability of a test. Measures of reliability depend on whether the test is being administered repeatedly by the same observer, or by different people or different methods, as well as what type of variable is being measured. Intra-rater reliability compares results when the test is administered repeatedly by the same observer, and inter-rater reliability compares the results when measurements are made by different observers. Standard deviation and coefficient of variation are used to determine reliability between multiple measurements of a continuous variable in the same individual. These differences can be random or systematic. We usually assume that differences between repeated measurements by the same observer and method are purely random, whereas differences between measurements by different observers or by different methods can be both random and systematic. The term "bias" refers to systematic differences, distinguishing them from "random error." The Bland–Altman plot describes reliability in method comparison, in which one measurement method (often established, but invasive, harmful, or expensive) is compared with another method (often newer, easier, or cheaper).

As we discuss these measures of reliability, we are generally comparing measurements with each other, not with the "truth" as determined by a reference "gold standard." Comparing a test result with a gold standard assesses its accuracy or validity, something we will cover in Chapters 3 and 4. However, sometimes one of the continuous measurements being compared in a Bland–Altman plot is considered the gold standard. In this case, the method comparison is called "calibration."