1 Clinical decision making

Philip S. Schoenfeld

What is evidence-based medicine?

David Sackett, the "father" of evidence-based medicine (EBM) stated that EBM is "the conscientious and judicious use of current best evidence from clinical care research in the management of individual patients" [1]. Terms used in this definition can be explained as follows.

• *Conscientious use* implies that physicians review articles about clinical research and apply this information to clinical decision making.

• *Current best evidence from clinical care research* implies that physicians systematically appraise the methods and results of clinical research articles using EBM tools. With these tools, physicians can separate the "wheat from the chaff" when reading medical journals and identify poorly designed studies that will produce biased results and should be discarded before being applied to patient care. This chapter will focus on techniques to identify and interpret the best evidence from properly designed research articles.

• *Judicious use* implies that a physician's experience and patient's preferences are crucial components of decision making and that these judgments must be balanced with the data from best evidence.

Judicious use of best evidence is a particularly important concept to understand [2]. Many critics state that the practice of EBM is "cookbook" medicine that devalues the judgment of a clinician and the values of an individual patient. This interpretation is inaccurate. Physicians must consider a patient's preferences about the potential benefits and side effects and costs of a medication when deciding a specific treatment. Also, a specific patient may not fit the criteria for enrollment of patients into a randomized controlled trial (RCT). For example, an RCT demonstrated that rifaximin, a nonabsorbable antibiotic, improved bloating in Lebanese patients [3]. Will bloating (and other gastrointestinal symptoms) improve if rifaximin is used in patients with irritable bowel syndrome (IBS) in the United States? If we assume that these results are applicable to patients with IBS in the United States, then is it worthwhile to use a treatment that may only produce a temporary relief of symptoms? What if the patient had a past history of *Clostridium difficile* colitis after a course of ciprofloxacin? Would the patient be willing to risk another case of *C. difficile* colitis? What if the patient does not have insurance and would have to pay $200 for this prescription? These questions are qualitative questions that require clinical judgments on the part of the patient and the physician [4]. Although the best evidence from an RCT [3] may identify an effective treatment for bloating, both physician judgment and patient preferences must also be used for effective clinical decision making. Thus, EBM and a reliance on best evidence is not intended to be "cookbook" medicine [2].

Nevertheless, EBM is a helpful tool for the quantitative aspect of clinical decision making, which arises from a systematic examination of study methodology and study results [2]. The medical literature is expanding at an exponential rate [5], and the time available for reading may be hurried and fragmented. Physicians need tools to build a framework for the rapid evaluation of the methodology and results of published studies, and EBM provides these tools (Tables 1.1 and 1.2). With these frameworks, physicians can rapidly identify well-designed studies that produce accurate and unbiased results and should be applied to patient care. Studies using improper methodology and biased results are quickly identified and ignored.

Principles of Clinical Gastroenterology. Edited by Tadataka Yamada, David H. Alpers, Anthony N. Kalloo, Neil Kaplowitz, Chung Owyang, and Don W. Powell. © 2008 Blackwell Publishing. ISBN 978-1-4051-69103

9781405169103_4_001.qxd 3/5/08 5:14 PM Page 2

Table 1.1 Critical approach to an article about a diagnostic test

1. When are diagnostic tests necessary?
   a. Is the pretest probability of disease so high or so low that further diagnostic tests are not needed?
   b. How is pretest probability estimated?
2. Assessing study design
   a. Was there a blinded comparison of the diagnostic test to a gold standard test?
   b. Were negative study tests verified by performing the gold standard test?
   c. Was the diagnostic study tested in patients similar to the population in which the test will be used?
3. Getting from pretest probability to posttest probability
   a. Interpret and apply data about sensitivity and specificity.
   b. Use likelihood ratios to maximize the data from a diagnostic test.
4. Applying the results of clinical research to your patient
   a. Are the results applicable to your patient?
   b. Is the test available with reproducible accuracy?

From Schoenfeld et al. [6].

Table 1.2 Critical approach to an article about a therapy

1. Assessing study design
   a. Did the study use concealed random allocation?
   b. Were patients and physicians blinded about allocation to treatment or placebo groups?
   c. Did groups receive equal treatment (cointerventions) except for the experimental study treatment?
   d. Did the study use an intention-to-treat analysis?
   e. Was follow-up of study patients complete?
2. Clinically significant results and statistically significant results
   a. Estimating treatment-effect size: relative risk reduction, absolute risk reduction, and number needed to treat
   b. Evaluating the sample size in a non-statistically significant study: were enough patients entered into the study?
   c. What is the precision of the treatment effect: how large are the 95% confidence intervals?
3. Applying the results of clinical research to your patient
   a. Are the results applicable to your patient?
   b. Are the potential treatment benefits worth the potential side effects, cost, and inconvenience to your patient?

From Guyatt et al. [19].

This chapter outlines systematic frameworks for the evaluation of methodology and results of studies about diagnostic tests and therapies; it reviews different statistical presentations of study results and discusses the use of physicians' clinical judgments and patients' preferences when applying these results to patient care.

Critical appraisal of an article about a diagnostic test

Case scenario A – part I

A 65-year-old woman presents for colon cancer screening. The patient has never been screened for colon cancer. She is asymptomatic and she denies weight loss, hematochezia, abdominal pain, or family history of colon cancer. You proceed to describe the benefits and risks of colonoscopy. However, the patient has heard radio advertisements for a *virtual colonoscopy* that uses a computed tomography (CT) scanner to look for polyps (this procedure is more appropriately called "CT colonography"). This radio advertisement emphasized that colon perforations can occur during colonoscopy. To provide adequate guidance to this woman, you consider the following questions:

- What is the probability of colon polyps in this woman?
- How accurate is virtual or CT colonography for the diagnosis of colon polyps?

You decide to apply EBM frameworks [6] to appraise a recent study [7] that examines the accuracy of CT colonography and answers these questions.

When are diagnostic tests necessary?

Is the pretest probability of disease so high or so low that further diagnostic tests are not needed?

Pretest probability defines the likelihood that a patient has a specific disorder before any diagnostic test result is available. Diagnostic tests should be ordered when the pretest probability is intermediate. If the pretest probability of a specific disorder is 50%, then accurate diagnostic tests may rule out the disorder or definitively confirm the presence of the disorder. Conversely, if the pretest probability is very high or very low, then ordering additional diagnostic tests may be unnecessary.

Several examples illustrate the concept of pretest probability. What diagnostic tests would you order for a hospitalized patient who suddenly develops diarrhea? Stool tests for ova and parasites are routinely ordered to evaluate diarrhea, but the pretest probability of a parasitic infection in a hospitalized patient with new-onset diarrhea approaches zero [8]. In this situation, the pretest probability is so low that stool tests for ova and parasites should not be obtained. Conversely, patients with peptic ulcers are rarely tested for *Helicobacter pylori* in some countries (e.g., Armenia) because the prevalence of *H. pylori* approaches 100%. These patients are automatically treated for *H. pylori* after identification of peptic ulcers without testing for *H. pylori*. What level of pretest probability is intermediate and suggests the need for diagnostic tests? Would it be 25% likelihood of disease being present or 75% likelihood of disease being present? A physician must use clinical judgment here and consider the cost, accuracy, and side effects of the diagnostic test, the consequences of a "missed" diagnosis (i.e., if a missed diagnosis may have fatal consequences, then clinicians will have a lower threshold to order diagnostic tests), and a patient's preferences. Patient preference is an important concept with respect to diagnostic tests. For example, many patients may be particularly anxious that their symptoms represent the signs of a fatal disease. These patients may need to be reassured by the results of a negative diagnostic test, and physicians may have a lower threshold to order diagnostic tests in this situation.

How is pretest probability estimated?

When physicians evaluate a patient's complaint, they intuitively use their experience and clues from the history and physical examination to estimate pretest probabilities for different medical disorders. However, these estimates are often inaccurate [9]. The prevalence of a disorder (the proportion of patients with a specific disorder at a distinct point in time) provides a more accurate estimate of pretest probability.

Valid studies about the prevalence of a particular diagnosis should meet several methodological criteria [10]. First, the technique for confirming the diagnosis should be explicit and credible. For example, colonoscopy would be an explicit and credible test to estimate the prevalence of colon polyps. However, flexible sigmoidoscopy, which does not examine the right side of the colon, would not be a credible test. Second, the technique for confirming the diagnosis should be applied to consecutive patients who present with a specific complex of symptoms, physical examination signs, or laboratory results. For example, an appropriate study about prevalence of colon polyps could be applied to our case scenario if the study enrolled patients who were women 50 years of age or older, who were asymptomatic (e.g., no hematochezia or abdominal pain), and who were undergoing colorectal cancer screening. Finally, clinicians should determine whether the characteristics of their patients are similar to those of the patients examined in the study. For example, if a study estimated the prevalence of colon polyps in women with a family history of colon cancer, then the results from this study might not apply to the patient in our case scenario because the prevalence of colon polyps in patients with a family history of colon cancer is higher than the prevalence of colon polyps in patients with no family history of colon cancer.

Case scenario A – part II

A recent publication in the *New England Journal of Medicine* estimates the pretest probability for colon polyps in asymptomatic, average-risk women referred for colon cancer screening [11]. In this study, asymptomatic women referred for colon cancer screening underwent colonoscopy. Colonoscopy is a credible diagnostic test to define the prevalence of colon polyps. Consecutive women referred for colon cancer screening were offered colonoscopy. Patients were asymptomatic (e.g., denied history of hematochezia, change in bowel habits, or abdominal pain) and were screened with complete blood cell counts and fecal occult blood tests to rule out anemia or occult gastrointestinal (GI) bleeding. In this trial, the prevalence of colon polyps among asymptomatic women was 21%, and the prevalence of advanced colon polyps (i.e., polyps larger than 10 mm, villous adenomas, adenomas with high-grade dysplasia, or colorectal carcinoma) was approximately 8% among women aged 60–69 years. Although the prevalence of advanced colon polyps is not very high, you recognize that a missed diagnosis of advanced colon polyps could lead to a fatal colon cancer. This study also demonstrates that evaluation of the left side of the colon with flexible sigmoidoscopy is a very poor predictor of polyps in the right side of the colon. Therefore, it is clear that it is crucial to evaluate the entire colon despite guideline recommendations which suggest that flexible sigmoidoscopy is still a reasonable alternative for colon cancer screening [12]. With this knowledge, you proceed to review the study about the diagnostic accuracy of CT colonography [7].

Assessing study design

Was there a blinded comparison of a diagnostic test to a gold standard test?

A *gold standard* or reference standard refers to a diagnostic test that definitively establishes the presence or absence of disease. For example, a study examining the diagnostic accuracy of magnetic resonance cholangiography (MRC) for the diagnosis of choledocholithiasis used endoscopic retrograde cholangiopancreatography (ERCP) as the reference test [13]. Reference standards are usually costlier, riskier, or more inconvenient than new diagnostic tests being studied. Otherwise, performing the reference standard would be more sensible. Biopsies, autopsies, surgical pathology, or even prolonged patient follow-up may also be reference standards that determine the presence or absence of disease.

The results from the reference standard and the diagnostic test should be examined by investigators who do not know the patient's history or the results of other tests. This *blinded* comparison is especially important when the interpretation of test results is subjective. For example, in the study assessing the diagnostic accuracy of MRC [13], the radiologist's interpretation of the MRC would be biased if he or she knew that the ERCP demonstrated choledocholithiasis.

Establishing a reference standard test may be an elusive goal. New technologies may be more accurate than the established reference standard test. If a potentially poor diagnostic test is being used as the reference standard, then the diagnostic accuracy of the new test may appear worse than it truly is. For example, one study evaluated the accuracy of ultrasonography to diagnose cholelithiasis but used oral cholecystograms as the gold standard test [14]. In this study, only patients with abnormal results on cholecystography were referred for surgery. Five patients in this study had ultrasounds that showed evidence of cholelithiasis, but normal-appearing cholecystograms. Based on this study's analysis, these positive ultrasound results were false-positive test results (i.e., patients did not truly have disease). Ultimately, several of these patients underwent cholecystectomy because of recurrent symptoms, confirming the presence of cholelithiasis and demonstrating that oral cholecystograms may not be an adequate gold standard test.

Were negative study results verified by performing the gold standard test?

Study results will be distorted if investigators use the results of the new diagnostic test to decide whether or not to perform the gold standard test. This *verification bias* may produce biased study results in more than 50% of diagnostic test studies [15]. For example, in the study about the accuracy of MRC for choledocholithiasis [13], a few patients with normal MRCs actually had choledocholithiasis on ERCP. If ERCPs were withheld from these patients (because investigators assumed choledocholithiasis was absent based on the normal MRCs), then MRC would appear more accurate than it truly is.

Was the diagnostic study tested in patients similar to the population in which the test will be used?

Patients with end-stage disease may have grossly abnormal diagnostic test results, making it easy to differentiate healthy people from ill patients. For example, virtual colonoscopy might easily differentiate patients with normal colons from patients with end-stage, near-obstructing colon cancer. The real value of a diagnostic test is its ability to identify patients with early manifestations of disease (e.g., colon polyps) that could be easily confused with a normal finding (e.g., stool). To assess the accuracy of a diagnostic test properly, the test should be studied in a broad range of patients, similar to the patients seen in clinical practice [16]. The best example of this *spectrum bias* may be carcinoembryonic antigen (CEA).

Measurement of CEA was evaluated as a diagnostic test for colorectal cancer. Initially, the test was studied in patients with advanced colorectal cancer and in healthy controls [17]. The results demonstrated that almost all (98%) of the patients with advanced colorectal cancer had elevated CEA, whereas almost all healthy controls had low levels of CEA. These initial results raised hope that CEA might be a useful screening tool for colorectal cancer. However, when this test was studied in a broad population of patients with early-stage colorectal cancer and patients with other GI disorders, the test was inaccurate and unable to differentiate patients with early cancer from patients with other disorders [18].

Case scenario A – part III

The study assessing the diagnostic accuracy of CT colonography used conventional colonoscopy as a gold standard, which is appropriate. Conventional colonoscopy was performed in all patients who underwent CT colonography, which eliminates verification bias. Patients were average-risk, asymptomatic men and women referred for colorectal cancer screening, so CT colonography was being evaluated in a population of patients that was similar to the population in which it would be used. So, there was no spectrum bias. All study patients proceeded to conventional colonoscopy even if their CT colonography did not show any polyps. In other words, negative study results were verified by performing the gold standard test (i.e., conventional colonoscopy) and verification bias was avoided. Finally, the comparison of the CT colonography results with the conventional colonoscopy results was completely blinded using a segmental unblinding technique. After completion of the CT colonography, the radiological examination was immediately examined by a radiologist who reported the results for the cecum, ascending colon, transverse colon, descending colon–sigmoid colon, and rectum. The patient then immediately went for conventional colonoscopy. During the colonoscopy, the colonoscopist would complete the evaluation of a given segment of colon (e.g., ascending colon). Then, a study coordinator would reveal the results of the CT colonography for that segment of the colon. If a polyp was seen on CT colonography, but not seen on conventional colonoscopy, then the colonoscopist would carefully reexamine that portion of the colon. If no polyps were found, then the result was a false-positive finding on CT colonography. If a polyp was found on reexamination of the colon, then the result would be a false-negative for conventional colonoscopy. This is a particularly elegant technique to compare the new diagnostic test, CT colonography, to the gold standard, conventional colonoscopy, because the gold standard test is not perfect for identification of all polyps and the new diagnostic test, CT colonography, may occasionally identify polyps that are missed by conventional colonoscopy. Overall, you determine that the article has adequate methodology and you proceed to review the results.

Getting from pretest probability to posttest probability

Interpret and apply data about sensitivity and specificity

Sensitivity and specificity can be calculated from the classic 2×2 table (see Fig. 1.1). The 2×2 table is completed by filling in the true-positive test results (positive test result when disease is present), false-positive test results (positive test result when disease is absent), true-negative test results (negative test result when disease is absent), and false-negative test results (negative test result when disease is present). For example, a recent study assessed the accuracy of ferritin for the diagnosis of iron deficiency anemia, using bone marrow aspirates as a gold standard for the diagnosis of iron deficiency [19]. This trial found that 150 patients had high ferritin levels (more than 45 μg/L) and 85 patients had low ferritin levels (45 μg/L or less). Of the 85 patients with low ferritin levels, 70 had iron deficiency anemia (true-positive test results) and 15 did not (false-positive test results). Of the 150 patients with high ferritin levels, 135 patients did not have iron deficiency anemia (true-negative test results), and 15 did (false-negative test results).

	Iron deficiency anemia		
	Anemia present	Anemia absent	
Test positive (Ferritin ≤ 45 μg/L)	70 (TP)	15 (FP)	PPV: TP/(TP + FP) = 70/(70 + 15) = 82%
Test negative (Ferritin > 45 μg/L)	15 (FN)	135 (TN)	NPV: TN/(FN + TN) = 135/(135 + 15) = 90%
	Sensitivity: TP/(TP + FN) = 70/(70 + 15) = 82%	Specificity: TN/(FP + TN) = 135/(135 + 15) = 90%	

TP: True positive = test positive and disease present
FP: False positive = test positive and disease absent
FN: False negative = test negative and disease present
TN: True negative = test negative and disease absent
PPV: Positive predictive value
NPV: Negative predictive value

Figure 1.1 Sensitivity and specificity of ferritin for the diagnosis of iron deficiency anemia (36% prevalence of iron deficiency anemia) in an elderly population. Data from Guyatt et al. [19].

In the 2 × 2 table (see Fig. 1.1), the formulas for sensitivity (the percentage of patients with the disease in whom the test results are positive) and specificity (the percentage of patients without the disease in whom the test results are negative) are defined. Using these formulas, the sensitivity of ferritin (with a cutoff point of 45 μg/L of ferritin) for iron deficiency anemia is 82%, and the specificity is 90%.

Unfortunately, sensitivity and specificity "work backwards" from clinical practice, evaluating patients with known disease and providing data about the presence or absence of certain diagnostic test results. However, patients present with symptoms and diagnostic test results, and we "work forwards" with these results to determine the likelihood of disease. The positive predictive value (PPV) and negative predictive value (NPV) from the 2 × 2 table (see Fig. 1.1) provide these data. The formulas for PPV (the proportion of patients with positive test results who have the disease) and NPV (the proportion of patients with negative test results who do not have the disease) are also provided in the 2 × 2 table. Using these formulas, the PPV for low ferritin level (45 μg/L or less) in the diagnosis of iron deficiency anemia is 82%. Hence, 82% of patients with ferritin levels of 45 μg/L or less had iron deficiency anemia. The NPV for high ferritin level is 90%, or 90% of patients with ferritin levels of more than 45 μg/L did not have iron deficiency anemia. Before clinicians apply PPV and NPV to their individual patients, the limitations of these statistics must be recognized. Sensitivity and specificity usually remain relatively constant, although they may vary slightly depending on the severity of disease in a specific patient population. However, PPV and NPV vary widely depending on the prevalence of the disease. For example, consider if all internal medicine admissions to a hospital were screened with ferritin for iron deficiency anemia. In this diverse population, the prevalence of iron deficiency anemia might only be 5%, although the prevalence of iron deficiency anemia was 36% among the elderly anemic patients in the ferritin–iron deficiency anemia study [19]. Assuming that the sensitivity and specificity remain constant, a new 2 × 2 table (Fig. 1.2) can be constructed, producing a significantly lower PPV of 32% and a significantly higher NPV of 99%. Hence, when prevalence of a disease decreases, the PPV decreases and the NPV increases.

Use likelihood ratios to maximize the data from a diagnostic test

Likelihood ratios express the likelihood that a particular range of values for a diagnostic test will be found in a patient with a specific disease. They overcome two weaknesses of sensitivity/specificity and PPV/NPV. First, likelihood ratios predict the presence of disease based on a diagnostic test result (similar to PPV/NPV), but likelihood ratios do not change with different disease prevalence (unlike PPV/NPV). Second, studies reporting sensitivity and specificity usually provide data about the accuracy of a diagnostic test around

	Iron deficiency anemia		
	Anemia present	Anemia absent	
Test positive (Ferritin ≤ 45 μg/L)	8 (TP)	17 (FP)	PPV: TP/(TP + FP) = 8/(8 + 17) = 32%
Test negative (Ferritin > 45 μg/L)	2 (FN)	173 (TN)	NPV: TN/(FN + TN) = 173/(173 + 2) = 99%
	Sensitivity: TP/(TP + FN) = 8/(8 + 2) = 80%	Specificity: TN/(FP + TN) = 173/(173 + 17) = 91%	

Figure 1.2 Positive and negative predictive value of ferritin in the diagnosis of iron deficiency anemia in a hypothetical population of 200 hospitalized patients (5% prevalence of iron deficiency anemia).

Table 1.3 Likelihood ratios for ferritin in the diagnosis of iron deficiency anemia

Ferritin	Likelihood ratio
> 100 μg/L	0.13
> 45 to ≤ 100 μg/L	0.46
> 18 to ≤ 45 μg/L	3.12
≤ 18 μg/L	41.47

only one value. For example, the ferritin–iron deficiency anemia study [19] calculated sensitivity and specificity of ferritin around a cutoff point of 45 μg/L (i.e., ferritin level less than or equal to 45 μg/L is consistent with iron deficiency anemia and ferritin level greater than 45 μg/L is not consistent with iron deficiency anemia). Intuitively, a patient with a ferritin level of 5 μg/L is more likely to have iron deficiency anemia than a patient with a ferritin level of 40 μg/L, but the sensitivity and specificity cannot differentiate between these two patients. However, likelihood ratios are usually calculated for multiple ranges of diagnostic test results, thereby maximizing the information from a diagnostic test. Thus, likelihood ratios may facilitate the application of diagnostic test results to patient care.

Mathematically, the likelihood ratio for a positive test result is as follows: sensitivity / (1 – specificity), or true-positive rate/false-positive rate.

The likelihood ratio for a negative test result is as follows: (1 – sensitivity) / specificity, or false-negative rate/true-negative rate.

In the ferritin–iron deficiency anemia study [15], likelihood ratios were calculated for four ranges of ferritin: less than or equal to 18 μg/L, 19–45 μg/L, 46–100 μg/L, and more than 100 μg/L (Table 1.3).

By using a nomogram [20], likelihood ratios easily convert pretest probabilities to posttest probabilities (Fig. 1.3). Simply place the base of a ruler at the pretest probability and angle the ruler through the likelihood ratio to find the posttest probability. For example, clinicians might assume that a 70-year-old patient with a history of myocardial infarction, daily use of aspirin, a mean corpuscular volume of 78, and a hemoglobin level of 11 g/dL has a 50% pretest probability of iron deficiency anemia (moderately higher than the prevalence of iron deficiency anemia among a general population of elderly anemic patients). If this patient has a ferritin level of 5 μg/L, then the pretest probability of 50% and the likelihood ratio of 41 produces a 98% posttest probability that iron deficiency anemia is present. Conversely, if this patient has a ferritin level of 110 μg/L, then the posttest probability is 10%.

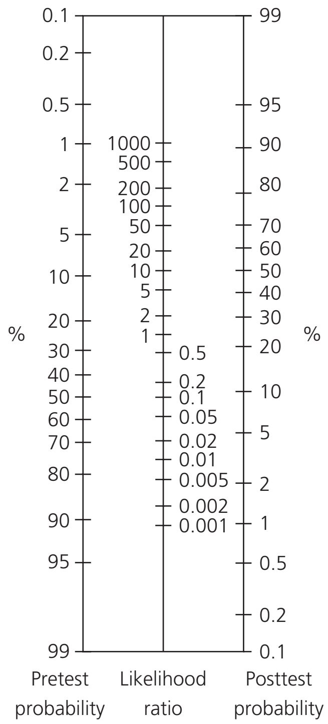

Figure 1.3 Nomogram for interpreting diagnostic test results. Adapted from Fagan [20], with permission from the Massachusetts Medical Society.

Case scenario A – part IV

For patients with large polyps (at least 10 mm), CT colonography produced sensitivity of 93.8% and specificity of 96% with a likelihood ratio for a positive test of 23 and a likelihood ratio for a negative test of 0.06. For patients with polyps larger than 5 mm in size, CT colonography produced sensitivity of 88.7% and specificity of 79.6% with likelihood ratios of 4 for a positive test and 0.1 for a negative test. Based on the data from the screening colonoscopy study in women [11], the prevalence of large (more than 10 mm) polyps in 65-year-old women is 5%, and the prevalence of colon polyps of larger than 5 mm is about 10%. Therefore, using likelihood ratios, positive results on CT colonography for large polyps predicts a 55% posttest probability of large polyps in our 65-year-old woman. Positive results on CT colonography for colon polyps larger than 5 mm predicts a 33% posttest probability of colon polyps in our patient.

The CT colonography is not very accurate for polyps that are 5 mm or less in size. However, this may not be a significant concern because fewer than 1% of these polyps have villous architecture or high-grade dysplasia contained within them [21]. Therefore, these diminutive polyps are very unlikely to develop into colorectal cancer.

Applying the results of clinical research to your patient

Are the results applicable to your patient?

A study may have valid methodology, and the results may indicate that the diagnostic test is accurate. However, if the study patient population is much different from your patient population (e.g., geriatric vs pediatric patients, symptomatic vs asymptomatic patients), then the diagnostic test may perform differently in your patients. If your own patient meets all the inclusion criteria of a study and the prevalence of disease is similar in your setting, the results are probably applicable to your patient.

Physicians need to use their own judgment when there are no definitive research data about the accuracy of the diagnostic test in specific patients. For example, the study being evaluated in the case scenario does not provide specific data about the diagnostic accuracy of CT colonography in 60- to 69-year-old women. However, there is no obvious pathological or physiological reason why CT colonography should be less accurate among women aged 60–69 years compared to other patients in this study. Thus, results from the CT colonoscopy study [7] should be applicable to the case scenario patient.

Is the test available with reproducible accuracy?

Even when a diagnostic test is adequately described to permit replication in your clinical setting, the test may not be available. Some diagnostic tests require special equipment or skilled examiners, which may not be widely available. For example, CT colonography is not widely available outside academic medical centers in 2007. Many diagnostic tests, including CT colonography, require subjective interpretation. A well-designed study of a diagnostic test will have the diagnostic test results reviewed by examiners with different levels of expertise. If experienced and inexperienced practitioners produce similar interpretations of diagnostic test results, then you may be assured that study results may be reproduced in your own clinical setting. For example, previous studies [22] have demonstrated that both experienced and inexperienced radiologists can accurately diagnose choledocholithiasis on MRC. Recent studies [23] demonstrate that properly trained radiologists and gastroenterologists produce reproducible and accurate interpretations of CT colonography when the three-dimensional "fly-through" technique is used. The three-dimensional fly-through technique produces an image that is similar to the endoscopic view of the colon produced during conventional colonoscopy.

Case scenario A – part V

Based on the radio advertisement that she heard, CT colonography is available to the patient and the study that examined the accuracy of CT colonoscopy indicates that it is very accurate for polyps larger than 5 mm in diameter. Although you do not know about the experience of radiologists at this specific CT colonography center, you recognize that properly trained radiologists appear to produce accurate and reproducible evaluations of CT colonography if they use the three-dimensional fly-through technique. Overall, you decide that CT colonography may be a reasonable alternative based on the quantitative data from this particular study [7]. However, you also review several qualitative issues that the patient should consider before scheduling the CT colonography.

CT colonography requires a full preparation to evacuate stool from the colon. In fact, the study from the case scenario also had patients drink a barium-based solution to tag any remaining stool or water in the colon, and then used digital subtraction to remove these wastes from the three-dimensional fly-through image. If the local CT colonography center does not include this in their protocol, then they may not be able to reproduce the accurate polyp detection reported here [7]. Even if the local CT colonography center does add barium-based solutions to their protocol, then patients still need to go through two separate bowel preparations if they are found to have colon polyps (i.e., one bowel preparation for the CT colonography and then a second bowel preparation for the conventional colonoscopy that is subsequently performed to remove polyps identified on CT colonography). Many patients will not want to undergo two separate bowel preparations. Also, CT colonography is not reimbursed by most insurance policies, so the patient will probably have to pay $800–$1000 for the procedure. Finally, you reassure your patient that the reported rate of colon perforation during colonoscopy in asymptomatic individuals undergoing colorectal cancer screening appears to be less than 1 in 10 000 [23]. Given these issues, your patient decides that she would prefer to proceed directly to colonoscopy.

In the future, you recognize that CT colonography may become widespread when Medicare and insurance companies reimburse for this service. Proof-of-concept studies have demonstrated that CT colonography may be possible without evacuation of stool with bowel preparation [24]. (i.e., consumption of barium-based solution alone may be adequate to perform digital subtraction of stool during CT colonography.) If CT colonography is accurate, is reimbursed by insurance, and is performed without evacuation of stool by bowel preparation, then CT colonography could become the preferred tool for colorectal cancer screening. You are glad that recent research [22] demonstrates that gastroenterologists can also be taught to accurately interpret CT colonography images.

Critical appraisal of an article about a therapy

Case scenario B – part I

You are seeing a 75-year-old man with osteoarthritis of the hip and a past history of a nonsteroidal antiinflammatory drug (NSAID)-associated bleeding ulcer. His test results were negative for *H. pylori* infection during the evaluation of his bleeding ulcer. The primary care physician has referred this patient because treatment of osteoarthritis with acetaminophen and physical therapy has been ineffective. The primary care physician wants to restart the patient on an NSAID, but she is still concerned about the risk for recurrent NSAID-associated bleeding ulcers. Specifically, she asks the following questions.

• Among patients using NSAIDs, does the addition of proton pump inhibitors (PPIs) reduce the frequency of GI bleeding?
• If the addition of PPIs does reduce GI bleeding, how large is the reduction?
• Would it be more appropriate to place the patient on a cyclooxygenase-2 (COX-2) -selective NSAID or on a combination of PPI and conventional NSAID.

You recently saw an article [25] that may address the first two questions. This trial randomized patients with a history of *H. pylori* infection and NSAID-associated bleeding peptic ulcers to receive NSAID + PPI vs NSAID + *H. pylori* eradication. You apply EBM frameworks [26] to determine whether the study has appropriate methodology that is likely to produce accurate results (see Table 1.2).

Assessing study design

Did the study use concealed random allocation?

A patient's response to treatment may be influenced by many factors other than treatment. Age, severity of illness, and comorbid medical problems will affect a patient's prognosis and limit the effect of treatment. Therefore, these factors should be distributed equally between the treatment and placebo groups (or between a "new" treatment group and a control group) to identify a "true" or accurate estimate about the effectiveness of the treatment. In an RCT, every patient has an equal chance of receiving treatment or placebo when they enter the trial, so that these factors are usually distributed equally between the treatment and placebo groups.

In a nonrandomized trial, physicians determine which patients enter the treatment group and which patients enter the placebo group. For unclear reasons, patients with a good prognosis are disproportionately entered into the treatment group in a nonrandomized trial. Patients with a good prognosis are more likely to have a favorable outcome, regardless of the effectiveness of treatment [27]. Nonrandomized trials illustrate that these studies demonstrate larger treatment effects than randomized trials and are more likely to demonstrate a false-positive result [27].

Concealment of allocation maintains the integrity of randomization. In concealed random allocation, researchers who obtain informed consent and enroll patients into a trial do not know whether the next study patient will receive treatment or placebo. If concealment of allocation was used, then the methods section of a study should indicate this (e.g., sealed, opaque, sequentially numbered envelopes were opened after patients gave informed consent; a central coordinating center was called for treatment assignment after a patient gave informed consent). Researchers may subconsciously wish to show that the therapy being studied is superior to the control therapy. Therefore, without concealed allocation, researchers may subconsciously assess a patient's prognosis and guide patients with good prognosis into the treatment or new therapy group and guide patients with a bad prognosis into the placebo or control therapy group.

Were patients and physicians blinded about allocation to treatment or placebo groups?

Blinding simply means that the patients and physicians do not know if the patient received placebo or treatment. This is particularly important when the outcome is subjective. For example, another study [28] compared rates of peptic ulcer bleeding among patients with rheumatoid arthritis who took a COX-2-selective NSAID or naproxen. As part of this study, the VIGOR trial, the frequency of dyspepsia among study patients was assessed. However, the assessment of dyspepsia is quite subjective and variable. Both the patient and the study physicians may assume that COX-2 inhibitors are less likely to cause dyspepsia than conventional NSAIDs, like naproxen, possibly introducing bias into their subjective assessment of dyspepsia. Blinding both the patients and the health-care personnel (double blinding) is the best method to avoid this bias. Double blinding has been demonstrated to prevent inflated estimates of treatment benefit in randomized trials [29]. Randomization, concealed allocation, and double blinding are the only techniques that have been shown to reduce inflated estimates of treatment benefit in epidemiological studies [27,29].

The importance of blinding is self-evident, but it may be difficult to ensure. For example, a recent double-blind, randomized controlled trial compared lubiprostone with placebo among patients with constipation-predominant IBS [30]. In this trial, the study end point was global assessment of improvement in IBS symptoms, which is clearly a subjective outcome. However, lubiprostone, a calcium channel agonist, stimulates intestinal secretion of water and rapidly increases the frequency of bowel movements in constipated patients. Thus, patients using lubiprostone may have noted the rapid increase in frequency of bowel movements and assumed that they were using lubiprostone. This knowledge may have *unmasked* the blinding process and biased the subjective assessment of improvement in IBS symptoms. One possible resolution would be to ask patients to guess whether they had received lubiprostone or placebo at the end of the trial. If 50% of patients receiving lubiprostone guess correctly and 50% of patients using tegaserod guess incorrectly, then the blinding process still worked.

Even when the study outcome is objective, it is still helpful to maintain double blinding. For example, one study of primary prevention of bleeding esophageal varices with β-blockers examined overall mortality as a study outcome [31]. This is certainly an objective outcome, but double blinding was still maintained in this study because of the risk for cointerventions. Cointerventions are treatments other than the study treatment that may affect the outcome, especially when the cointerventions are unequally distributed between the treatment and placebo groups. For example, isosorbide-5-mononitrate has been shown to reduce variceal bleeding [32]. Without double blinding, the physician or the patient might be tempted to start isosorbide-5-mononitrate in one group more than the other group. Double blinding limits the unequal use of cointerventions.

Did the groups receive equal treatment (cointerventions) except for the experimental study treatment?

Additional treatments or cointerventions are most problematic when they are very effective, such as the additional use of isosorbide-5-mononitrate in a study about the effectiveness of β-blockers to prevent variceal bleeding. Although the methodology of a study may be strengthened if all cointerventions are withheld, it may be unethical to withhold effective treatment from patients enrolled in a study. Research does not occur in a vacuum, and patients receive additional treatments or cointerventions to optimize their health while participating in a study. To balance this conflict, the indications to use cointerventions should be clearly described in the Methods section, their use should be limited, and their use should be recorded for later analysis.

The VIGOR trial [28], which compared a COX-2-selective NSAID with naproxen in patients with isosorbide-5-mononitrate, clearly described the use of cointerventions in the Methods section. Patients were allowed to take other treatments for rheumatoid arthritis, including acetaminophen, methotrexate, and corticosteroids, even though concurrent use of corticosteroids with NSAIDs increases the risk for serious NSAID-associated GI complications [33]. The frequency of serious NSAID-associated GI complications among corticosteroid-using patients who used rofecoxib or naproxen was recorded for subgroup analysis. Concurrent use of other NSAIDs was forbidden because use of multiple NSAIDs increases the risk for NSAID-associated GI complications [33]. Patients on NSAIDs develop dyspepsia, and over-the-counter preparations to treat dyspepsia are readily available. So, recognizing that trials do not occur in a vacuum, researchers allowed patients to use antacids or *limited* doses of H_2-receptor antagonists. In limited doses, these medications are unlikely to affect the occurrence of serious NSAID-associated GI complications, but allowing use of these medications will treat dyspepsia.

Did the study trial use an intention-to-treat analysis?

In almost every study, some patients stop taking the study medication (treatment or placebo). They are noncompliant, or they believe that the study medication is causing side effects. An intention-to-treat analysis includes all randomized patients in the final data analysis, regardless of whether the patients completed the study or were compliant. An adherence-to-protocol analysis excludes patients who did not complete the study owing to noncompliance or side effects. An intention-to-treat analysis preserves the value of randomization because some patients with a poor prognosis may not be able to complete the study, but these patients should be included to understand fully the true effectiveness of a treatment.

An example best illustrates the concept of intention-to-treat analysis. A randomized trial compared desipramine, a tricyclic antidepressant, with placebo in patients with moderate to severe symptoms of functional bowel disease [34]. Some patients randomized to receive desipramine withdrew from the trial because they experienced severe side effects, including constipation, which is a well-known complication of anticholinergic drugs including tricyclic antidepressants. In the intention-to-treat analysis, these patients must be considered *treatment failures*, and the intention-to-treat analysis did not demonstrate that patients using desipramine were significantly more likely to experience global relief of symptoms compared to patients using placebo (60% vs 47%, $P = 0.13$).

An adherence-to-protocol analysis only includes patients who complete therapy, estimating the likelihood of a good outcome for patients who complete therapy. However, an adherence-to-protocol analysis may lose the value of randomization because patients unlikely to have a good outcome are eliminated from analysis. In the adherence-to-protocol analysis of patients with functional bowel disease, desipramine-treated patients were significantly more likely to experience global symptom relief (73% vs 49%, $P = 0.006$). Therefore, if a patient with functional bowel disease can tolerate desipramine without side effects, then they are more likely to experience global symptom relief compared to placebo-treated patients.

Robust studies will present both an intention-to-treat analysis and an adherence-to-protocol analysis, allowing readers to assess fully the results and to make up their own minds about the benefits of treatment. Because an intention-to-treat analysis includes compliant patients and patients who discontinue therapy as a result of side effects, it estimates the likelihood of achieving a desired outcome when a patient first starts a treatment, consistent with "real-world" medical practice. The adherence-to-protocol analysis estimates the likelihood of achieving a desired outcome when a patient can tolerate therapy without severe side effects. Notably, the study from our example [34] recognized that patients with constipation-predominant IBS were more likely to have side effects, while patients with diarrhea-predominant IBS would benefit from the constipation induced by desipramine. Therefore, the investigators performed a subgroup analysis of diarrhea-predominant IBS patients and demonstrated that desipramine was superior to placebo using an intention-to-treat analysis. Therefore, it appears that desipramine is most appropriate for patients with functional bowel disease who have diarrhea as their predominant bowel habit.

Was follow-up of study patients complete?

When patients drop out of studies, several explanations are possible. The patients may have disappeared because they experienced a side effect or even died from the study treatment, or they may stop follow-up because their symptoms have resolved. How can you determine whether the loss to follow-up biased the study's results? In a treatment study with a positive result, the study results could be recalculated, assuming that all treatment group patients lost to follow-up had a poor outcome and assuming that all control group patients lost to follow-up had a good outcome. If the recalculated results still demonstrate a treatment benefit, then the loss to follow-up did not cause a falsely positive study result. To avoid recalculations, one short cut is available [26]: only rare studies will still demonstrate a positive treatment effect upon recalculation of study results if more than 15%–20% of patients are lost to follow-up.

Case scenario B – part II

The study comparing NSAID + PPI with NSAID + *H. pylori* eradication in patients with a history of *H. pylori* infection and NSAID-associated peptic ulcer bleeding met all the criteria for a well-designed study: randomization, concealed allocation, equal and minimal use of cointerventions, use of double blinding, use of an intention-to-treat analysis, and minimal number of patients lost to follow-up. Based on this analysis, you decide that the study is likely to produce accurate and unbiased results, and you proceed to review the statistical representations of the results.

Clinically significant results and statistically significant results

Estimating treatment-effect size: relative risk reduction, absolute risk reduction, and number needed to treat

The relative risk reduction (RRR) expresses the decreased risk for an adverse outcome in the treatment group compared with the risk for an adverse outcome in the placebo or control group. For example, in a randomized, double-blind, intention-to-treat trial, patients with endoscopically treated bleeding ulcers received either intravenous PPI or placebo. The study end point was recurrent peptic ulcer bleeding: 22.5% of patients receiving placebo had recurrent bleeding, whereas 7% of patients receiving intravenous PPI had recurrent bleeding [35]. The RRR may be calculated as: (% placebo patients with bleeding – % treatment patients with bleeding) / % placebo patients with bleeding, i.e., (22.5% – 7%) / 22.5%.

This RRR is 69%. Hence, a patient with an endoscopically treated bleeding ulcer is 69% less likely to develop recurrent bleeding from the ulcer when receiving an intravenous PPI compared with a similar patient not receiving an intravenous PPI.

Absolute risk reduction (ARR) is the reduction in adverse outcomes between the placebo group and the treatment group. Although the RRR compares the risk for adverse outcomes between treated and placebo patients, the ARR identifies the actual reduction in adverse outcomes for treated patients. In the study of patients with endoscopically treated bleeding ulcer treated with intravenous PPI, the ARR may be calculated as: % placebo patients with recurrent bleeding ulcer – % patients receiving intravenous PPI with recurrent ulcer bleeding, i.e., 25% – 7% = 15.5%.

Therefore, patients with an endoscopically treated bleeding ulcer who receive intravenous PPI may decrease their individual risk for recurrent ulcer bleeding by 15.5%.

The RRR can be misleadingly large if adverse outcomes are infrequent in patients receiving placebo or no treatment. This concept is best illustrated by comparing the number needed to treat (NNT) and ARR with the RRR. Consider the results from the VIGOR trial [28]. The results from this trial were reported as frequency of serious NSAID-associated GI complications (severe upper GI bleeding, perforation, or obstruction) per 100 patient-years of NSAID use. There were 0.6 complications per 100 patient-years among patients using a COX-2-selective NSAID, rofecoxib, and 1.4 complications per 100 patient-years among patients using naproxen. Based on these data, the RRR is as follows: (% naproxen patients with complications – % rofecoxib patients with complications) / % naproxen patients with complications, i.e., (1.4% – 0.6%) / 1.4% = 60%.

In other words, a patient who uses the COX-2-selective NSAID, rofecoxib, is 60% less likely to have a serious NSAID-associated GI complication compared with a similar patient who uses naproxen. This sounds impressive until you consider the ARR. The ARR is as follows: % naproxen patients with complications – % rofecoxib patients with complications, i.e., 1.4% – 0.6% = 0.8%.

In other words, an average NSAID-using patient who uses a COX-2-selective NSAID, like rofecoxib, reduces his or her individual risk for a serious NSAID-associated GI complication by only about 0.8%. The NNT allows interpretation of study results in terms of patient care, especially when the RRR is large and the incidence of adverse outcomes is small. Specifically, the NNT is the inverse of the ARR, or 1 / ARR, estimating the number of patients who need to be treated to prevent one additional adverse outcome. In the VIGOR trial, the NNT is 1 / ARR, i.e., 1 / 0.8% = 1 / 0.008 = 125.

In other words, for every 125 average patients treated with COX-2-selective NSAIDs instead of naproxen for 1 year, one additional serious GI complication will be prevented. Patients and physicians may be less likely to choose a potentially better, but more expensive, treatment if the study results are presented as the ARR or NNT instead of the RRR [36]. Considering all three statistics (ARR, RRR, NNT) helps both patients and physicians assess the potential benefits of therapy.

Although the ARR and NNT may appear to be more useful than the RRR, the RRR is valuable because of its versatility.

It provides the best estimate of treatment benefit among patients with varying risks of adverse outcomes [37]. Patients using NSAIDs receiving naproxen in the VIGOR trial only had a 1.4% risk for serious NSAID-associated complications per 100 patient-years of use. However, how beneficial would rofecoxib be in a 78-year-old man with a past history of upper GI bleeding? Based on previously published data [33], this patient would have a significantly higher baseline risk for serious NSAID-associated GI complications, approaching 10 per 100 patient-years of use (i.e., 10% risk per year). Given this baseline risk, applying the RRR decreases the risk for serious NSAID-associated GI complications from 10% to 4%. This produces an ARR of about 6% (10% – 4% = 6%) and an NNT of about 17 (1 / 6% = 1 / 0.06 = 17). For this high-risk patient, the added expense of a COX-2 inhibitor is outweighed by the significantly improved safety profile. The value of the RRR is that it can be applied to patient populations with different inherent risks for adverse outcomes [37].

Evaluating the sample size in a non–statistically significant study: were enough patients entered into the study?

Studies that do not demonstrate statistical significance may be interpreted as negative studies (i.e., the treatment is no more likely than placebo to reduce adverse outcomes). However, an adequate number of patients have to enter a trial to demonstrate a statistically significant RRR, regardless of the effectiveness of the treatment. Many trials that do not yield statistically significant results have not entered enough patients into the trial to demonstrate reliably an RRR of 25% or even 50% [38]. When assessing study results, it should be clear if a non-statistically significant result represents a truly ineffective treatment or an inadequate enrollment of patients into a study.

When investigators plan a study, multiple outcomes may be analyzed, but the sample size is calculated based on only one outcome. For example, in a study comparing ligation plus octreotide with ligation alone in the prevention of recurrent bleeding from esophageal varices [39], the investigators estimated that ligation plus octreotide would reduce recurrent bleeding from varices by 70% compared with ligation alone. The study entered enough patients to demonstrate a statistically significant RRR of 70% for recurrent bleeding, and a statistically significant RRR of 76% was measured in the study. Investigators also evaluated 30-day mortality and found that combined treatment reduced 30-day mortality by 52% compared with ligation alone, but this RRR was not statistically significant ($P = 0.09$). However, only enough patients to demonstrate a statistically significant RRR of 70% for recurrent bleeding were enrolled in this study. Because the study demonstrated a strong trend for reduced 30-day mortality with combined treatment, this therapy may truly be efficacious in reducing 30-day mortality, and this non-statistically significant result is likely the result of inadequate sample size.

What is the precision of the treatment effect: how large are 95% confidence intervals?

Traditionally, a P value of 0.05 or less indicates statistical significance. Studies with P values of 0.05 or less indicate that there is a 5% or less likelihood that the difference between treatment and placebo occurred due to chance. However, P values do not provide data about the accuracy or precision of study results. Confidence intervals do provide information about the precision of study results.

The 95% confidence interval (95% CI) estimates the range within which the true RRR or ARR resides 95% of the time (i.e., if you repeated the same trial 100 times, then the RRR would fall within the 95% CI in 95 of 100 trials). When the lower limit of the confidence interval for RRR is greater than zero, then the treatment is significantly better than placebo. If the upper limit of the confidence interval around the RRR is less than zero, then the treatment is actually harmful or worse than placebo. For example, the RRR in the VIGOR trial for fewer serious NSAID-associated GI complications with rofecoxib is 60% with 95% CI of 20%–80% [28]. Thus, the RRR for serious GI complications with rofecoxib has a 95% likelihood of being between 20% and 80% with the best estimate being 60%.

The magnitude of confidence intervals is determined by the sample size [40]. Studies with large sample sizes have narrower 95% CIs and a more precise estimate of the true RRR. When the upper limit of a confidence interval around an RRR is greater than zero, but the lower limit of the confidence interval is less than zero, it is possible that the treatment could be better, worse, or no different than placebo. If the magnitude of benefit is moderate and the confidence interval is wide and barely crosses zero, then the treatment is probably beneficial and the trial simply did not enter enough patients. For example, one study [41] examined patients with bleeding esophageal varices and compared band ligation with sclerotherapy for reducing rebleeding. The RRR for recurrent variceal bleeding with ligation was 48%, with 95% CI of – 15% to 68%; however, only 77 patients were entered into this study, which was not adequate to demonstrate a statistically significant RRR of 48%. A metaanalysis [42] pooled the results from several studies, allowing the analysis of 547 patients. With this larger sample size, the RRR for recurrent variceal bleeding with ligation was 42% (almost the same RRR as the original randomized trial), but with a statistically significant 95% CI of 16%–60%.

Case scenario B – part III

In the NSAID + PPI vs NSAID + *H. pylori* eradication study [25], patients treated with NSAID + PPI had a 4.4% rate of recurrent bleeding ulcer over 6 months compared with an

18.8% rate of recurrent bleeding ulcer among patients treated with NSAID + *H. pylori* eradication. You assume that the NSAID + *H. pylori* eradication group represents a placebo-type group for your patient. Given these results, you conclude that NSAID + PPI is much better than NSAID + placebo with an RRR = 77%, an ARR = 14.4, and an NNT = 7.

Applying the results of clinical research to your patient

Are the results applicable to your patient?

If your patient meets the inclusion and exclusion criteria of a study, then the results should be applicable to that patient. Even if your patient is a year too old to be included in the study or has a history of a comorbid disease not allowed in the trial, these issues may not prevent the application of study results to your patient. Ultimately, physicians must use their clinical judgment to decide whether a patient differs significantly from study patients, preventing application of study results to that patient.

A study may not identify a statistically significant difference between placebo and treatment for all patients entered in a study, but it may find a significant difference for a subgroup of patients. Often, this subgroup of patients has more severe disease with a higher frequency of adverse outcomes. However, physicians should be cautious before applying subgroup analyses to their patients, even if their patients fit into the subgroups. If investigators evaluate multiple outcomes in many subgroups, eventually one subgroup analysis will be statistically significant simply owing to chance.

The results of subgroup analyses are more likely to be valid if:

- the treatment effect is large
- only a few subgroup analyses were performed
- the subgroup analyses were hypothesized a priori (i.e., before performance of the study)
- it is consistent with current understanding of pathophysiology
- other, independent studies have produced similar findings.

For example, the VIGOR trial demonstrated that rofecoxib decreased clinical upper GI events (i.e., symptomatic ulcers, significant upper GI bleeding, perforation, or obstruction) in the subgroup of patients with a past history of clinical upper GI events. This subgroup analysis was one of only a few subgroup analyses; it was hypothesized a priori. Pathophysiology suggests that selective inhibition of COX-2 isoenzymes reduces the inflammatory response without inhibiting COX-1 isoenzymes in GI mucosal cells and platelets, which should prevent clinical upper GI events.

Are the potential treatment benefits worth the potential side effects, costs, and inconvenience to your patient?

When deciding whether to start a new treatment, both the patient and physician should consider the inconvenience, side effects, and costs associated with the treatment. When balancing the potential benefit of a treatment vs the potential consequences, the NNT is a helpful tool. If the treatment is cheap and convenient and the consequences of the adverse event are potentially severe, then a large NNT may be acceptable. For example, hundreds of health-care providers are vaccinated with hepatitis B vaccine to prevent one case of hepatitis B. If the treatment is expensive and inconvenient and has potentially significant side effects, then the treatment may still be acceptable if the NNT is small and the consequences of an adverse outcome are life-threatening.

Case scenario B – part IV

The NSAID + PPI vs NSAID + *H. pylori* eradication study did not have a true placebo group, although it is likely that some of the patients in the NSAID + *H. pylori* eradication group actually had recurrent ulcers prevented by the *H. pylori* eradication treatment. Therefore, the 18.8% recurrent bleeding peptic ulcer rate is probably an underestimation. Our patient does not have *H. pylori* infection, so we can use the NSAID + *H. pylori* eradication group as a rough estimate of recurrent bleeding peptic ulcer rate in an NSAID + placebo group.

This study was performed in Hong Kong in patients of Asian descent. You are uncertain if you can apply these data to your United States-born patient. There would have to be genetic or environmentally created differences in ulcer pathophysiology or PPI pharmacology between United States-born and Asian patients to prevent you from applying the results of this study to your patient.

Finally, you wonder if you should consider recommending a COX-2-selective NSAID for this patient. COX-2-selective NSAIDs have been associated with increased risks of cardiovascular side effects, which could be detrimental for your patient. Fortunately, your literature search identified a recent study that compared a COX-2-selective NSAID, celecoxib, vs a conventional NSAID, diclofenac, + PPI. This study did not find a statistically significant difference in recurrent ulcer bleeding rates between the two groups: 4.9% vs 6.4% [43]. You also note that the average price of generic PPI + generic NSAID is much less than the cost of branded COX-2-selective NSAIDs. Therefore, you conclude that NSAID + PPI is the most appropriate choice for your patient.

Conclusions

Ultimately, decisions about the use of diagnostic tests and treatments are balanced by possible benefits, harms, costs, patient preferences, and availability of these interventions. In this chapter, principles for the systematic appraisal and application of clinical research have been reviewed. This information will provide a basis to interpret the diagnostic and therapeutic applications of GI technology, which will be discussed in the ensuing chapters.

References

1. Sackett DL, Rosenberg WC, Muir Gray JA, et al. Evidence based medicine: what it is and what it isn't. BMJ 1996;312:71.
2. Straus S, Haynes B, Glasziou P, et al. Misunderstandings, misperceptions, and mistakes. ACP J Club 2007;146:A8.
3. Sharara AI, Aoun E, Abdul-Baki H, et al. A randomized double-blind placebo-controlled trial of rifaximin in patients with abdominal bloating and flatulence. Am J Gastroenterol 2006;101:326.
4. Weiner S. From research evidence to context: the challenge of individualizing care. ACP J Club 2004;141:A11.
5. Haynes RB, Sackett DL, Muir Gray JA, et al. Transferring evidence from research into practice. 1. The role of clinical care research evidence in clinical decisions. ACP J Club 1996;Nov/Dec:A14.
6. Schoenfeld P, Guyatt G, Hamilton F, et al. An evidence-based approach to gastroenterology diagnosis. Gastroenterology 1999;116:1230.
7. Pickhardt PJ, Choi RJ, Hwang I, et al. Computed tomographic virtual colonoscopy to screen for colorectal neoplasia in asymptomatic adults. N Engl J Med 2003;349:2191.
8. Siegel DL, Edelstein PH, Nachamkia I. Inappropriate testing for diarrheal diseases in the hospital. JAMA 1990;263:979.
9. Dolan JG, Bordley DR, Mushlin AI. An evaluation of clinicians' subjective prior probability estimates. Med Decis Making 1986;6:216.
10. Richardson WS, Wilson MW, Guyatt G, Nishikawa J. Users' guides to the medical literature: how to use an article about disease probability for differential diagnosis. JAMA 1999;281:1214.
11. Schoenfeld P, Cash B, Flood A, et al. Colonoscopic screening of average-risk women for colorectal neoplasia. N Engl J Med 2005;352:2061.
12. Winawer S, Fletcher R, Rex D, et al. Colorectal cancer screening and surveillance: clinical guidelines and rationale-update based on new evidence. Gastroenterology 2003;124:544.
13. Chan YL, Chan AC, Lam WW, et al. Choledocholithiasis: comparison of MR cholangiography and ERCP. Radiology 1996;200:85.
14. Barton RJ, Crow HC, Fook SR. Ultrasonographic and radiographic cholecystography. N Engl J Med 1977;296:538.
15. Reid MC, Lachs MS, Feinstein A. Use of methodological standards in diagnostic test research. JAMA 1995;274:6445.
16. Ransohoff D, Feinstein A. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med 1978;299:926.
17. Thomson DM, Krupey J, Freedman SO, Gold P. The radioimmunoassay of circulating carcinoembryonic antigen of the human digestive system. Proc Natl Acad Sci U S A 1969;64:161.
18. Bates SE. Clinical applications of serum tumor markers. Ann Intern Med 1991;115:623.
19. Guyatt G, Patterson C, Ali M, et al. Diagnosis of iron deficiency anemia in the elderly. Am J Med 1990;88:205.
20. Fagan TJ. Nomogram for Bayes theorem. N Engl J Med 1975;293:257.
21. Schoenfeld P. Small colonic polyps – do they matter? Clin Gastroenterol Hepatol 2006;4:293.
21. Becker C, Grossholz M, Becker M, et al. Choledocholithiasis and bile duct stenosis: diagnostic accuracy of MRC. Radiology 1997;205:523.
22. Ray Q, Kim C, Scott T, et al. Gastroenterologist interpretation of CTC: Pilot study demonstrating feasibility and similar accuracy compared to radiologists. [abstract] Gastroenterology 2007;132:A6389.
23. Ko CW, Riffle S, Morris C, et al. Complications after screening and surveillance colonoscopy [abstract]. Gastroenterology 2007;132:A994.
24. Zalis ME, Perumpillichira J, Del Frate C, Hahn PF. CT colonography: digital subtraction bowel cleansing with mucosal reconstruction initial observations. Radiology 2003;226:911.
25. Chan FK, Chung SC, Suen BY, et al. Preventing recurrent upper gastrointestinal bleeding in patients with *Helicobacter pylori* infection who are taking low-dose aspirin or naproxen. N Engl J Med 2001;344:967.
26. Schoenfeld P, Cook D, Hamilton F, et al. An evidence-based approach to gastroenterology therapy. Gastroenterology 1998;114:1318.
27. Chalmers TC, Celano P, Sacks HS, Smith H. Bias in treatment assignment in controlled clinical trials. N Engl J Med 1983;309:1358.
28. Bombardier C, Laine L, Reicin A, et al. Comparison of upper gastrointestinal toxicity of rofecoxib and naproxen in patients with rheumatoid arthritis. N Engl J Med 2000;343:1520.
29. Schulz K, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: dimensions of methodological quality with estimates of treatment effects in controlled trials. JAMA 1995;273:408.
30. Drossman DA, Chey W, Panas R, et al. Lubiprostone significantly improves symptom relief rates in adults with irritable bowel syndrome and constipation (IBS-C): data from two, 12-week, randomized, placebo-controlled, double-blind trials [abstract]. Gastroenterology 2007;132:639f.
31. Pascal JP, Cales P. and a Multicenter Study Group. Propranolol in the prevention of first upper gastrointestinal hemorrhage in patients with cirrhosis of the liver and esophageal varices. N Engl J Med 1987;317:856.
32. Angelico M, Carli C, Piat C, et al. Isosorbide-5-mononitratre versus propranolol in the prevention of first bleeding in cirrhosis. Gastroenterology 1993;104:1460.
33. Schoenfeld P, Kimmey M, Scheiman J, et al. Non-steroidal anti-inflammatory drug associated gastrointestinal complications: guidelines for prevention and treatment. Aliment Pharmacol Ther 1999;13:1273.
34. Drossman DA, Toner BB, Whitehead W, et al. Cognitive-behavioral therapy versus education and desipramine versus placebo for moderate to severe functional bowel disorders. Gastroenterology 2003;125:19.
35. Lau JY, Sung JY, Lee KK, et al. Effect of intravenous omeprazole on recurrent bleeding after endoscopic therapy of bleeding peptic ulcers. N Engl J Med 2000;343:310.
36. Forrow L, Taylor WC, Arnold RM. Absolutely relative: how research results are summarized can affect treatment decisions. Ann Intern Med 1992;92:121.
37. Oxman AD, Guyatt G. A consumer's guide to sub-group analysis. Ann Intern Med 1992;116:78.
38. Frieman JA, Chalmers TC, Smith H, et al. The importance of beta, the type II error, and sample size in the design and interpretation of the randomized control trial. Survey of 71 negative trials. N Engl J Med 1978;299:690.
39. Sung JJ, Chung SC, Yung MY, et al. Prospective randomized study of effect of octreotide on re-bleeding from oesophageal varices after endoscopic ligation. Lancet 1995;346:1666.
40. Altman DG. Confidence intervals. In: Sackett DL, Richardson WS, Rosenberg W, Haynes RB (eds). Evidence-based Medicine: How to Practice and Teach EBM. London: Churchill Livingstone, 1997:228.
41. Laine L, El-Newihi HM, Migikovsky B, et al. Endoscopic ligation compared with sclerotherapy for the treatment of bleeding esophageal varices. Ann Intern Med 1993;119:1.
42. Laine L, Cook D. Endoscopic ligation compared with sclerotherapy for treatment of esophageal variceal bleeding. A meta-analysis. Ann Intern Med 1995;123:280.
43. Chan FK, Hung LC, Suen BY, et al. Celecoxib versus diclofenac and omeprazole in reducing the risk of recurrent ulcer bleeding in patients with arthritis. N Engl J Med 2002;347;2104.