The development of new technology traditionally has been the lifeblood of radiology. Many of the spectacular advances in medicine over the past few decades have centered around radiology. One does not have to go far into the past to predate the development of CT, MR imaging, and sonography, technologies that now are omnipresent, critical components of medical care. Yet for all the advances in the development of imaging technology, radiology research has come under deserved criticism in its efforts to assess the effectiveness and appropriate use of such imaging technology [1–5].

Production of a technologically adequate image is a starting point, but it is only the first step in determining whether such a technology should be used in clinical care. To be useful, an imaging study must also be accurate and provide information that has the potential to change the medical care, and ultimately the health, of the patient [6, 7].

This article is the first of an ongoing series that, taken together, will form a comprehensive teaching primer on basic and advanced concepts in technology assessment and outcomes research as described in the introductory article in this month’s issue of the *American Journal of Roentgenology* (AJR) [8]. This series of articles published in the *AJR* will form one component of the research course cosponsored by the American College of Radiology and Canadian Association of Radiologists on the fundamentals of clinical research for radiologists. Tightly linked with these articles will be Web-based interactive teaching modules. The intent of this integrated series is to be progressive, starting with basic introductory concepts and gradually adding complexity through intermediate and more advanced modules. The objective is to provide a pathway for the novice researcher to learn to critically appraise the literature and to conduct evidence-based radiology, to communicate effectively with methodology experts, and finally, to perform or direct independent, scientifically valid, and clinically useful research.

The concepts introduced in this first article will be by design simplistic. The intent of this first module is to introduce the scope of the material that is to be covered in much greater detail in the sessions to come. Many of the major concepts of rigorous technology assessment will be introduced, with detailed discussions to follow in future modules. This introduction describes the problems of research in radiology and attempts to provide the radiology investigator with an understanding of some of the potential pitfalls to be avoided.

**Evidence-Based Radiology**

Every day in the clinical practice of radiology, we make observations and adjust our practice accordingly. Many of the great advances in science have arisen from just such observations. The fortuitous observation that bacteria colonies did not grow around bread mold led Alexander Fleming to discover penicillin. In radiology, we constantly observe the imaging appearances of diseases and healthy states and subtly adjust our thresholds for interpretation. However, at the same time, this simple anecdote and experience is, by definition, limited to what we personally have seen and is most strongly influenced by what we have seen recently. We have all observed the phenomenon that after a patient...
presents with a rare and difficult-to-diagnose disease, the next group of patients that appear at all similar will be examined for that same malady. Our belief that a disease is rare is shaken by the fact that we have seen it, and have seen it recently. The same is true for the use of diagnostic technology. For example, because we have diagnosed a case of testicular seminoma from CT findings does not mean that CT is the imaging modality of choice for this condition, or that all patients at risk for testicular seminoma should undergo CT.

To supersede this practice based on anecdote, the field of evidence-based medicine has evolved and has become the standard for medical practice [9, 10]. Although less established in radiology than in other areas of medicine, this evidence-based paradigm is no less relevant for radiology [11]. The construct underlying evidence-based medicine is that one individual’s experience is limited. Decisions should be based on the best evidence from the medical literature rather than one’s own limited experience [9, 11]. As a corollary, as physicians we tend to cling to what we were taught in residency or fellowship, often by acknowledged experts in the field. However, the evidence-based paradigm suggests that the experts are also individuals, and we should trust their anecdotal experience only somewhat more than we trust our own. Instead, practice should be guided by rigorous scientific investigation [9, 11, 12].

The major source for the evidence on which to base practice is the medical literature. With the rapid proliferation in radiology technology has come a parallel increase in the volume of the radiology literature. There are now more than 40 radiology journals and more than 4000 articles published each year [13]. However, the published literature has its own perils and should be interpreted with a critical eye. First, case reports, even if published, are essentially anecdotes that are codified in print. Although they are often interesting, may be provocative, and can invoke questions for scientific study, they should not form the basis for practice. Second, and more insidious, are published reports that, although well intended, contain biases or flaws in the methodology that attenuate the applicability of the results into practice. A central tenet of evidence-based medicine is that the literature must be analyzed critically, and only those studies that are robust should be used as the basis for practice [11, 14]. A useful framework for evaluating the value of a literature article is promoted by Kent et al. [2], who propose a four-grade scale (Appendix). At the top level (grade A) are methodologically rigorous studies with broad generalizability, including large randomized clinical trials and prospective comparisons of diagnostic test results to an appropriate gold standard. At the bottom level of this hierarchy are grade D studies, which include multiple methodologic flaws, biases in study design, or unsubstantiated opinion [2, 15]. Most of the radiology literature relates to development of new techniques and descriptive work. Actual assessment of these new technologies and determination of any impact on patient outcome is relatively uncommon [4]. Few grade A or B studies exist. New radiology technologies have been rapidly developed and disseminated, often without adequate proof of efficacy [1, 16, 17]. Although radiologists may not have paid great attention to the shortcomings in their research efforts, these limitations may have been more apparent to the remainder of the medical community.

Early studies of MR imaging report an illustrative example of how radiology research has come under external criticism, particularly for methodologic deficiencies. Developed in the 1970s and early 1980s, MR imaging was initially greeted with a variety of investigations and reports in the radiology literature in particular, describing the exciting potential of this new modality. However, most of this early research was merely descriptive. Those studies that attempted to assess even accuracy were limited in size and generally suffered from important design flaws [2, 16, 18, 19]. A 1988 article by Cooper et al. [16] noted that none of the initial 54 research reports on the efficacy of MR imaging met accepted contemporary standards for research design. The article concluded that “health care professionals paying for expensive innovative technology should demand better research on diagnostic efficacy.” In 1994, Kent et al. [2] found that of 142 studies of MR neuroimaging published through 1993, only one provided grade A information, 28 provided grade B or C, and most (113) provided only grade D information. Kent et al. concluded that despite the fact that more than 2000 MR imaging scanners had been installed, the evidence supporting the use of MR imaging in clinical practice was weak.

The credibility of the radiology research community was shaken by these criticisms, with some nonradiologists questioning whether conflicts of interest would influence radiologists and organized radiology [17]. Similar methodologic deficiencies have also been reported for radiology economic analyses [3, 20]. Today, more sophisticated and dependable research methods have been applied to MR imaging and assessment of efficacy with this modality for a number of indications. However, most of the research literature on the use of radiology techniques remains descriptive, with little published work on the influence of radiology on patient treatment or outcome [4]. One of the reasons for these deficiencies is the lack of research training of the individual radiology investigators. Unfortunately, training in research methodology has been underemphasized in radiology residency training in the United States [21]. Many radiologists, although highly skilled clinicians, have only a rudimentary background in research methodology and lack many of the basic tools required to perform a critical review of the medical literature. The objective of this discussion is to introduce some major concepts in research design and in critical literature review. More detailed discussion will be included in subsequent modules.

Anatomy of a Research Project

It is useful to review the anatomy of a research project. This standard framework is the foundation of the scientific literature. In brief, a research question is formulated, methods are derived to answer the question, data are collected and analyzed, and conclusions are drawn. Within this framework are several key concepts that are discussed in the following text, including formulation of the research question, use of efficient study design, avoidance of error and bias, and appropriate data analysis.

The Research Question

The first step in any research endeavor is to frame an appropriate research question. This question must be important (or it is not worth our efforts), but it also must be precise [22, 23]. As an example, we can start with a common and vexing clinical problem that has been the cause of considerable interest in the radiology literature, “Which test is better in patients with possible appendicitis, CT or sonography?” This question is certainly important and clinically relevant, but as framed above it cannot be answered. The question must be defined more precisely with respect to the type of patients in whom the question is being raised, the target population, and what is actually being asked. The imaging accuracy and usefulness of sonography and CT will likely vary on the basis of a number of patient-specific variables. Are the pa-
tients we are interested in adults or children? Are they thin or fat? Are they cooperative or uncooperative? Are they men or women? Disease-specific factors may also affect the imaging. Has the patient been symptomatic for a few hours and we suspect simple unperforated appendicitis, or has the patient been symptomatic for 4 days and appears septic, leading us to suspect an abscess? These factors also might affect the performance of sonography and CT.

Finally, how we are using the findings of an imaging study might affect the determination of optimal imaging modality. Are we using imaging to confirm appendicitis en route to the operating room, or are we using imaging to look for other abnormalities that might mimic appendicitis, such as ureteral calculi, diverticulitis, or even abdominal aortic aneurysm? A better defined research question might be, “In nonpregnant women younger than 40 years with symptoms suggestive of appendicitis but no peritoneal signs, what is the preferred imaging modality to exclude the presence of an abdominal condition that might require surgical intervention?” This reformulated research question is perhaps less “sexy” than “Which test is better?” but it is also much more useful. The reformulated question is no longer an issue of comparing radiology tests. Instead, we are asking a clinical question about a specific group of patients that can potentially affect the health of those patients [22–25]. Some experienced researchers believe that formulating and answering a clinical question about a specific factor may also affect the health of those patients. Some experienced researchers believe that formulating and answering a clinical question about a specific factor may also affect the health of those patients [22–25].

Study Design

Having determined the question to be answered, the next issue is the research methodology itself. To produce evidence that will appropriately drive decision making, experimental design is of critical importance and will be the focus of much of this article series. The goal of study design is to achieve the most with the least (i.e., to achieve efficiency). Fortunately, we have the experience of clinical epidemiologists and biostatisticians with decades of experience from which to draw to determine the most efficient way of designing studies and the most appropriate way to productively critique research. Prospective comparisons of diagnostic test results with a well-defined reference test and randomized double-blinded clinical trials are the study designs that provide the best information to guide clinical practice [2, 26]. However, other study designs, including cohort and case-control investigations and modeling studies can also provide useful information [4, 26]. These study designs will be discussed in detail in future modules.

Error

The research design is intended to arrive at the truth for the question under study. One of the major driving factors of research design is the effort to avoid or control error. Error can be divided into two general categories: random error, and systematic error, also known as bias. Random error, as the name implies, is due to chance events that have the potential to lead to false conclusions. The field of statistics has evolved in large part to deal with the random and therefore unpredictable error that can occur in any study design. Statistics is a methodology for drawing inference about populations from data collected on samples [27]. In medicine, we generally accept events as being true (not related to random chance) if the probability of their random occurrence is less than 5%, expressed as the common statistical p value of 0.05. Of course, unlikely events do occur. Type I (also known as alpha) error occurs when we conclude that a difference exists when in fact two groups are the same. At a significance threshold of p less than 0.05, we will make such type I errors in 5% of comparisons. However, if a study involves multiple comparisons (i.e., comparing six different MR imaging pulse sequences), then the probability of a type I error also increases [28].

The opposite of type I error, known as type II error, is when we conclude that two populations are the same when in fact they are not. Unfortunately, the commonly reported p value gives no information about the potential for this type II, or beta, error. There is a common misconception that a p value greater than 0.05 indicates that two groups are the same. However, this is only true if the study sample has sufficient size to have the power to detect a difference if it is present [27]. Sufficient sample size is determined by the size of difference we are interested in detecting, usually the amount of difference that would be clinically significant, and by the desired power of the study [27, 29]. Power is the chance the study will reveal the clinically significant difference when it exists and equals one minus the type II error probability. As an example, a study might report 90% power to detect a difference of 5%.

Bias

The opposite of random error is systematic error that is introduced through inadequacy in the study design, subject selection, or analysis. Statistics are for the most part unable to compensate for systematic error. Avoidance of such systematic error, or bias, is one of the major challenges of research design. Unfortunately, many of the apparently simple research designs that are common in the radiology literature succumb to bias. As an example, one could imagine a study designed to compare CT and MR imaging for detection of liver metastases in patients with known adenocarcinoma of another organ. To identify patients for such a study, one might review all the patients who underwent both tests, and using some external gold standard, make a comparison. However, would this study design be free of bias? Likely, there would be significant bias in the selection of the subjects. For example, if at a given center CT is generally used as the initial imaging modality for the evaluation of possible liver metastases, then the patients who undergo both imaging studies would be the ones in whom the initial CT was equivocal. The comparison would not be CT versus MR imaging, but rather, CT versus MR imaging in patients in whom the CT was equivocal. Of course, the results of such a study would underestimate the accuracy of CT, because only those cases that are difficult to diagnose with CT were included. This is a simple but unfortunately common example of selection bias in recruiting patients for a study. Selection bias occurs when the subjects studied are not representative of the target population. In the previous example, the target population is all patients with known adenocarcinoma of another organ. However, the study group is only those patients with known adenocarcinoma who underwent both CT and MR imaging. To avoid this bias, subject selection should be based on clinical criteria (i.e., all subjects with a new diagnosis of adenocarcinoma) rather than availability of imaging studies [14, 22].

When using a test to screen a population, selection bias can be more subtle but equally problematic. Intuitively, one would expect that if a cohort of subjects is randomly selected to undergo a radiologic screening test, we could compare the subjects who actually undergo screening with those who elect not to undergo screening and make reasonable conclusions. However, convincing evidence from previous screening studies indicates that differences exist between subjects who elect screening and those who refuse. Subjects who elect to undergo screening may be more health conscious, or more optimistic, or there may be some other factor that is not understood [4, 30]. Thus, in a research study designed to investigate patient outcome for a new screening
study, comparison of those who undergo screening with those who elect against screening could show improved outcomes in the screened group even if the test has no benefit, or is even harmful. Therefore, to investigate the effectiveness of a screening study, it is essential to compare patients who are randomized to be invited for screening to those who are randomized not to be invited. In the analysis, all subjects are included, regardless of whether they actually undergo the screening study. This is known as an intention-to-treat analysis and avoids the subtle bias I have described [4].

Other bias can develop from the way in which data are collected. All humans have preconceived notions, both conscious and unconscious. These preconceptions alter the way in which we observe our surroundings and can unintentionally affect data that we collect, which is referred to as review bias. To remove any review bias, it is necessary to ensure that the individual who collects the data is unaware of the outcome under study. For example, the individual who determines if a test is positive should not know whether the subject truly has the disease in question. Also, when comparing two tests, the results of the first test should not be known before interpretation of the second. A recent analysis of research on diagnostic tests performed by Reid et al. [1] included some radiology studies that reported that 62% of research studies did not document that appropriate steps had been taken to avoid such review bias.

Similarly, if different gold standards are used for patients with disease than for those without, then results of accuracy studies may be overestimated. Lijmer et al. [31] found that the reported accuracy of diagnostic studies was significantly greater if different verification standards were applied to patients with and without disease than if the same gold standard was applied to all. The term “verification bias” has been applied to this problem [31, 32].

Additional potential biases in diagnostic test evaluation include spectrum bias, in which only patients with overt disease are used for patients with disease than for those without, then results of accuracy studies may be overestimated. Lijmer et al. [31] found that the reported accuracy of diagnostic studies was significantly greater if different verification standards were applied to patients with and without disease than if the same gold standard was applied to all. The term “verification bias” has been applied to this problem [31, 32].

Prospective data collection is generally less subject to bias than retrospective collection and is therefore preferred when designing a study. However, retrospective data collection may be preferred in a few circumstances, such as when prospective data collection would remove the ability to blind the observers and would therefore potentially introduce greater bias.

The effect of these various biases has been documented. In general, studies with bias tend to report more encouraging results than those without bias [31]. In addition, preliminary studies of a diagnostic technology, performed with small sample size and vulnerable to bias, often will be highly optimistic about the capabilities of that technology. Subsequent reports may present a more realistic appraisal [32].

Data Analysis

Research is conducted on samples. We measure outcome or accuracy on a relatively small number of subjects. Yet the intent of research is (eventually) to influence clinical care. To achieve this, the research results must be valid on subjects other than those included in the study. Statistics is the science that allows us to make inferences about populations from measurements made on samples. A vast array of tools is available to the biostatistician to enable such inference. These tools must be familiar to the research radiologist and will be discussed in future modules. In this discussion I will limit myself to introduction of the concepts of validity and reliability.

Validity can be divided into internal validity and external validity, which is also known as generalizability. Internal validity refers to the extent to which the results and conclusions of a study actually relate to true events in the sample under study. Some of the biases and study design considerations described previously relate to validity. For example, an observer who is aware of the results of the reference test might unintentionally overestimate the accuracy of the diagnostic test under study. Thus, the recorded results might not be an internally valid representation of the actual sample. The method of data analysis and the statistical tests used are also critical to the internal validity of the study, because use of inappropriate analysis can lead to false conclusions.

Similarly, the external validity of a study is dependent on both the research design and the analytic methods. The extent to which the sample selected truly reflects the target population is a strong determinant of the generalizability of a study [22]. Also, the use of appropriate statistics allows determination of what inferences can be drawn about the target population on the basis of the sample data.

A final consideration is study reliability. Reliability refers to the extent to which the study is reproducible [1, 24]. The opposite of reliability is variability. Interpretation of some diagnostic tests can be quite subjective. If different observers cannot agree on the test result on the same subject, then interobserver variability is high. Similarly, if the same observer determines the results of the same test to be different at different times, then intraobserver variability is high. If a test has low reliability, then the test cannot achieve high accuracy in general practice [1].

Conclusion

Performing methodologically rigorous scientific research is not a trivial task. The optimal research study will be directed at an important, precisely defined clinical question, with a specified target population matched by the subject selection. The most efficient study design will be used and the sample size will be sufficient to limit type II error to an acceptable level. Further, bias will be avoided, and the results will be reliable, internally valid, and generalizable to the target population and possibly beyond. Success at such demanding research endeavors is certainly within the reach of radiologists and radiology researchers. However, training—the goal of this series of articles—is necessary.

In this article, I have attempted to introduce the problem—the need for improved research methodology in radiology research. I have also begun to outline the solution through briefly introducing the concept of evidence-based radiology and discussing the basics of research methodology: posing the research question, and study design, error, bias, and data analysis. I am certain that this discussion has been too basic for some and too sophisticated for others. However, in the modules that follow, increasing depth, clarity, and detail will be added to the rough outline that has been described in this article. By the conclusion of this project, the radiology investigator will have a comprehensive resource to aid the transition from relative novice to skilled researcher.

References

1. Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research: getting better but still not good. JAMA 1995;274:645–651
### Grade A: Studies with broad generalizability
- No significant flaws
- Prospective comparison of a diagnostic test with a well-defined diagnosis
- Large randomized, blinded clinical trial assessing therapeutic efficacy or patient outcome

### Grade B: Studies with narrower spectrum of generalizability
- Few well-described flaws with definable impact on the results
- Prospective study of diagnostic tests
- Randomized trial of therapeutic effects and patient outcomes

### Grade C: Studies with limited generalizability
- Multiple flaws in research methods, small sample size, incomplete reporting
- Retrospective studies of diagnostic accuracy

### Grade D: Studies with multiple flaws in research methods
- Obvious selection bias
- Opinions without substantiating data

---

(Modified from Kent et al. [2, 15])