2

Critically Assessing the Literature for Evidence-Based Imaging: Understanding Error and Bias

C. Craig Blackmore, L. Santiago Medina, James G. Ravenel, Gerard A. Silvestri, and Kimberly E. Applegate

Evidence-based imaging (EBI) requires the critical assessment and application of the best available evidence to patient imaging. Unfortunately, the published studies that comprise the available evidence are often limited by bias, small sample size, and methodological inadequacy. Further, the information provided in published reports may be insufficient to allow estimation of the quality of the research. Initiatives by journal editors to improve the reporting of research studies, including the CONSORT [1], STARD [2], SQUIRE [3], and others, provide useful guides but are incompletely implemented.

The objective of this chapter is to summarize the common sources of error and bias in the imaging literature to guide the critical assessment required for EBI.

What Are Error and Bias?

Errors in the medical literature can be divided into two main types. The first is *random error* that occurs due to chance variation causing a sample to be different from the underlying population. Random error will tend to be more important when sample size is small. *Systematic error, or bias*, is an incorrect study result due to nonrandom distortion of the data. Systematic error is not affected by sample size but rather is a function of flaws in the study design, data collection, or analysis. A second way to think about random and systematic error is in terms of precision and accuracy [4]. Random error affects the precision of a result. Using the bull's eye analogy, precision is how close the measurements are to each other (Fig. 2.1). Higher precision indicates relatively less random error and more likelihood that two samples from truly different populations will be differentiated from each other. Systematic error on the other hand is a distortion in the accuracy of an estimate. Regardless of precision, the underlying estimate is flawed by some aspect of the

C.C. Blackmore (✉)
Department of Radiology, Center for Health Care Improvement Science, Virginia Mason Medical Center, Seattle, WA, USA
e-mail: Craig.blackmore@virginiamason.org

L.S. Medina
Nicklaus Children's Hospital, Miami Children's Health Systems, Miami, FL, USA

Florida International University, Herbert Wertheim College of Medicine, Miami, FL, USA
e-mail: Santiago.medina@mch.com

J.G. Ravenel
Department of Radiology and Radiologic Sciences, Medical University of South Carolina, Charleston, SC, USA
e-mail: ravenejg@musc.edu

G.A. Silvestri
Department of Medicine, Medical University of South Carolina, Charleston, SC, USA
e-mail: silvestri@musc.edu

K.E. Applegate
Department of Radiology, University of Kentucky Children's Hospital, Lexington, KY, USA
e-mail: keapple@uky.edu

A. Kelly et al. (eds.), *Evidence-Based Emergency Imaging*, Evidence-Based Imaging,
https://doi.org/10.1007/978-3-319-67066-9_2

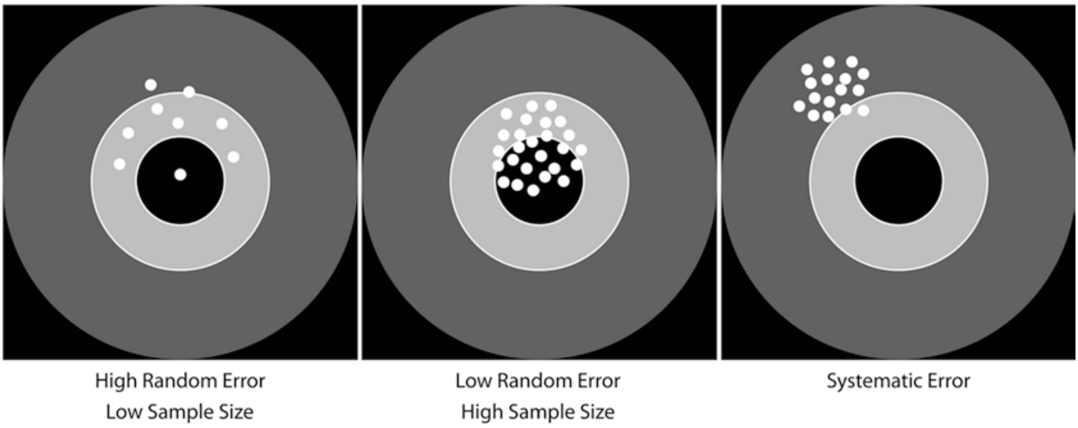

Fig. 2.1 Random and systemic error. Using the bull's-eye analogy, the larger the sample size, the less the random error and the larger the chance of hitting the center of the target. In systemic error, regardless of the sample size, the bias would not allow the researcher to hit the center of the target. *Left*: High random error and low sample size leads to low precision. *Middle*: Low random error and high sample size leads to high precision. *Right*: High precision can be accompanied by low accuracy if systematic error (bias) is present. (Reprinted with kind permission of Springer Science + Business Media from Blackmore CC, Medina LS, Ravenel JG, Silvestri GA. Critically Assessing the Literature: Understanding Error and Bias. In Medina LS, Blackmore DD (eds): *Evidence-Based Imaging: Optimizing Imaging in Patient Care.* New York: Springer Science + Business Media, 2006.)

research procedure. Using the bull's eye analogy, in systematic error regardless of the sample size, the bias would not allow the researcher to hit the center of the target (Fig. 2.1).

What Is Random Error?

Random error is divided into two main types: Type I, or alpha error, is when the investigator concludes that an effect or difference is present when in fact there is no true difference, and Type II or beta error occurs when an investigator concludes that there is no effect or no difference when in the underlying population, a true difference exists [4].

Type I Error

Quantification of the likelihood of alpha error is provided by the familiar *p*-value. A *p*-value of less than 0.05 indicates that there is a less than 5% chance that the observed difference in a sample would be seen if there was in fact no true difference in the population. In fact, the difference observed in a sample is due to chance variation rather than a true underlying difference in the population. It is important to remember that at a *p*-value of 0.05, we will still draw incorrect conclusions (make Type I errors) in 5 of 100 cases.

A second limitation of the ubiquitous *p*-value is that *p*-values are a function of both sample size and magnitude of effect. In other words, there could be a very large difference between two groups under study, but the *p*-value might not be significant if the sample sizes are small. Conversely, there could be a very small, clinically unimportant difference between two groups of subjects or between two imaging tests, but with a large enough sample size, even this clinically unimportant result would be statistically significant. Because of these limitations, many journals are underemphasizing use of *p*-values and encouraging research results to be reported by way of confidence intervals [5].

Confidence Intervals

Confidence intervals are preferred because they provide much more information than *p*-values. Confidence intervals provide information about the precision of an estimate (how wide are the confidence intervals), the size of any effect (magnitude of the confidence intervals), and the statistical significance of an estimate (whether the intervals include the null) [6].

In general, you can be 95% certain that the confidence interval (CI) includes the true population mean. More precisely, if you generate many 95% CI from many data sets, you can expect that the CI will include the true population mean in 95% of the cases and not include the true mean value in the other 5% [5]. Therefore, if the 95% CI interval does not include the null, then the results will be statistically significant at the 0.05 level [7]. Whereas the *p*-value is only interpreted as being either statistically significant or not, the CI has the advantage of providing the range of probable values and allows the reader to understand not just the statistical significance but also the magnitude of any effect [7, 8]. CIs shift the interpretation from a qualitative judgment about the role of chance to a quantitative estimation of the biologic measure of effect [5, 7, 8].

Confidence intervals can be constructed for any desired level of confidence. There is nothing magical about the 95% that is traditionally used, except that it is consistent with the traditional $p < 0.05$ threshold. If greater confidence is needed, then the intervals can be wider (i.e., 99%) or narrower (i.e., 90%) if less confidence is sufficient. The trade-off is that wider CIs are associated with greater confidence but less precision [5].

As an example, two hypothetical transcranial circle of Willis vascular ultrasound studies in patients with sickle-cell disease describe mean peak systolic velocities of 200 cm/s associated with 70% of vascular diameter stenosis and higher risk of stroke. Both articles reported the same standard deviation (SD) of 50 cm/s. At first glance, both articles appear to provide similar information. However, the size of the confidence interval is a function of the sample size, with narrower confidence intervals for the larger study reflecting greater precision. In the smaller series, the 95% CI was 186–214 cm/s, while in the larger series, the 95% CI was narrower, at 196–204 cm/s [5].

Type II Error

The familiar *p*-value does not provide information as to the probability of a Type II or beta error. A *p*-value greater than 0.05 does not necessarily mean that there is no difference in the underlying population. The size of the sample studied may be too small to detect an important difference even if such a difference does exist. The ability of a study to detect an important difference, if that difference does in fact exist in the underlying population, is called the power of a study. Power analysis can be performed in advance of a research investigation to avoid Type II error.

Power Analysis

Power analysis plays an important role in determining what an adequate sample size is, so that meaningful results can be obtained [9]. Power analysis is the probability of observing an effect in a sample of patients if the specified effect size, or greater, is found in the population [4]. Mathematically, power is defined as 1 minus β (beta), where β is the probability of having a Type II error. Type II errors are commonly referred to as false negatives in a study population. The other type of error is Type I or α (alpha), also known as false positives in a study population [7]. For example, if β is set at 0.10, then the researchers acknowledge they are willing to accept a 10% chance of missing a correlation between abnormal CT angiographic finding and the diagnosis of carotid artery disease. This represents a power of 1 minus 0.10, or 0.90, which represents a 90% probability of finding a correlation of this magnitude.

Ideally, the power should be 100% by setting β at 0. In addition, ideally, α should also be 0. By accomplishing this, false-negative and false-positive results are eliminated, respectively. In practice, however, powers near 100% are rarely achievable, so, at best, a study should reduce the false negatives β and false positives α to a minimum [4, 10]. Achieving an acceptable reduction of false negatives and false positives requires a large subject sample size. Optimal power, α and β, settings are based on a balance between scientific rigorousness, and the issues of feasibility and cost. For example, assuming an α error of 0.10, your sample size increases from 96 to 118 subjects per study arm (carotid and non-carotid artery disease arms) if you change your desired power from 85% to 90%, respectively [11]. Studies with more complete reporting and better study design will often report the power of the study, for example, by stating that the study has 90% power to detect a difference in sensitivity of 10% between CT angiography and Doppler ultrasound in carotid artery disease. Unfortunately, power calculations are often lacking, and it is left to the reader to determine if a study has sufficient power to interpret if a high *p*-value is actually an indication that a difference does not exist.

What Is Bias?

The risk of an error from bias decreases as the rigorousness of the study design and analysis increases. Randomized controlled trials are considered the best design for minimizing the risk of bias because patients are randomly allocated. This random allocation allows for unbiased distribution of both known and unknown confounding variables between the study groups. However, as described below, even randomized clinical trials are susceptible to some forms of bias. In nonrandomized studies, appropriate study design and statistical analysis can only control for known or measurable bias.

Detection of and correction for bias or systematic error in research is a vexing challenge for both researchers and users of the medical literature alike. Maclure and Schneeweiss have identified 10 different levels at which biases can distort the relationship between published study results and truth [12]. Unfortunately, bias is common in published reports [13], and reports with identifiable biases often overestimate the accuracy of diagnostic tests [14]. It is not uncommon for the initial reports on an imaging test to be enthusiastic in the results, but biased in the methods. Subsequent, more rigorous investigation will often refute, or at least diminish the purported effectiveness of a procedure. Careful surveillance for each type of bias is critical but may be a challenge. Well-reported studies will often include a section on limitations of the work, spelling out the potential sources of bias that the investigator acknowledges from a study as well as the likely direction of the bias and steps that may have been taken to overcome this. However, the final determination of whether a research study is sufficiently distorted by bias to be unusable is left to the discretion of the user of the imaging literature. The imaging practitioner must determine if results of a particular study are true, are relevant to a given clinical question, and are sufficient as a basis to change practice [15].

A common type of bias encountered in imaging research is that of *selection bias* [15]. Because a research study cannot include all individuals in the world who have a particular clinical situation, research is conducted on samples. Selection bias can arise if the sample is not a true representation of the relevant underlying clinical population (Fig. 2.2). Numerous subtypes of selection bias have been identified, and it is a challenge to the researcher to avoid all of these biases when performing a study. One particularly severe form of selection bias occurs if the diagnostic test is applied to subjects with a spectrum of disease that differs from the clinically relevant group. The extreme form of this spectrum bias occurs when the diagnostic test is evaluated on subjects with severe disease who are then compared to normal controls. In an evaluation of the effect of bias on study results, Lijmer found the greatest overestimation of test accuracy with this type of spectrum bias [14]. Selection bias is a particular challenge in nonrandomized studies.

A second frequently encountered bias in imaging literature is that of *observer bias* [16, 17], also called test-review bias and diagnostic-review bias [18]. Imaging tests are often subjective. The radiologist interpreting an imaging study forms an impression based on the appearance of the image, not based on an objective number or measurement. This subjective impression can be biased by numerous factors including the radiologist's experience; the context of the interpretation (clinical vs. research setting); the information about the patient's history that is known by the radiologist; the incentives

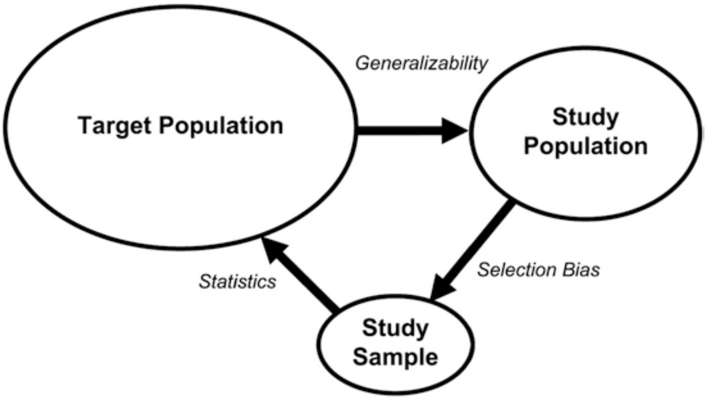

Fig. 2.2 Population and sample. The target population represents the universe of subjects who are at risk for a particular disease or condition. In this example, all subjects with abdominal pain are at risk for appendicitis. The sample population is the group of eligible subjects available to the investigators. These may be at a single center or group of centers. The sample is the group of subjects who are actually studied. Selection bias occurs when the sample is not truly representative of the study population. How closely the study population reflects the target population determines the generalizability of the research. Finally, statistics are used to determine what inference about the target population can be drawn from the sample data. (Reprinted with kind permission of Springer Science + Business Media from Blackmore CC, Medina LS, Ravenel JG, Silvestri GA. Critically Assessing the Literature: Understanding Error and Bias. In Medina LS, Blackmore DD (eds): *Evidence-Based Imaging: Optimizing Imaging in Patient Care.* New York: Springer Science + Business Media, 2006.)

that the radiologist may have, both monetary and otherwise, to produce a particular report; and the memory of a recent experience. But because of all these factors, it is critical that the interpreting physician be blinded to the outcome or gold standard when a diagnostic test or intervention is being assessed. Important distortions in research results have been shown and observers are blinded vs. not blinded. For example, Schulz showed a 17% greater risk reduction in studies with unblinded assessment of outcomes versus those with blinded assessment [19]. In order to obtain objective scientific assessment of an imaging test, all readers should be blinded to other diagnostic tests and final diagnosis, and all patient-identifying marks on the test should be masked. Basically, the research setting should replicate clinical practice as closely as possible. Since the diagnosis is not known when an imaging test is interpreted in clinical practice, it should not be known in the research setting. Observer bias is important for both randomized and nonrandomized studies.

Bias can also be introduced by the *reference standard* used to confirm the final diagnosis, called verification bias. First, the interpretation of the reference standard must be made without knowledge of the test results. Reference standards, like the diagnostic tests themselves, may have a subjective component and therefore may be affected by knowledge of the results of the diagnostic test. In addition, it is critical that all subjects undergo the same reference standard. The use of different reference standards (called differential reference standard bias) for subjects with different diagnostic test results may falsely elevate both sensitivity and specificity [14, 17]. Of course, sometimes it is not possible or ethical to perform the same reference standard procedure on all subjects. For example, in a meta-analysis of imaging for appendicitis, Terasawa found that all of the identified studies used a different reference standard for subjects with positive imaging (appendectomy and pathological evaluation) than for those with negative imaging (clinical follow-up). It simply wouldn't be ethical to perform appendectomy on all subjects. Likely, the sensitivity and specificity of imaging for appendicitis was overestimated as a result [20]. Verification bias and differential reference standard bias are important in both randomized and nonrandomized studies.

What Are the Inherent Biases in Screening?

Investigations of screening tests are susceptible to an additional set of biases. Screening trials are vulnerable to *healthy volunteer bias*. For example, in the Prostate, Lung, Colorectal, and Ovarian Screening Trial, the individuals who volunteered to undergo screening were generally healthier and had lower mortality than the general population, even before the screening began. Hence, comparing only those who actually undergo screening to those randomized not to be invited to be screened will cause falsely elevated estimates of screening effectiveness. This bias can be avoided by including all of those invited to be screened, not just those who actually undergo screening [21]. Case-control studies are particularly problematic for screening, as screening is a choice in these studies, and people who present for elective screening tend to have better health habits [22]. In assessing the exposure history of cases, including the test on which the diagnosis is made, regardless of whether it is truly screen or symptom detected, can lead to an odds ratio greater than 1 even in the absence of benefit [23]. Similarly, excluding the test on which the diagnosis is made may underestimate screening effectiveness. The magnitude of bias is further reflected in the disease preclinical phase; the longer the preclinical phase, the greater the magnitude of the bias.

Prospective nonrandomized screening trials perform an intervention on subjects, such as screening for lung cancer, and follow them for many years. These studies can give information of the stage distribution and survival from diagnosis of a screened population; however, these measures do not allow an accurate comparison to an unscreened group due to lead time, length time, and overdiagnosis bias [24] (Fig. 2.3). *Lead time bias* results

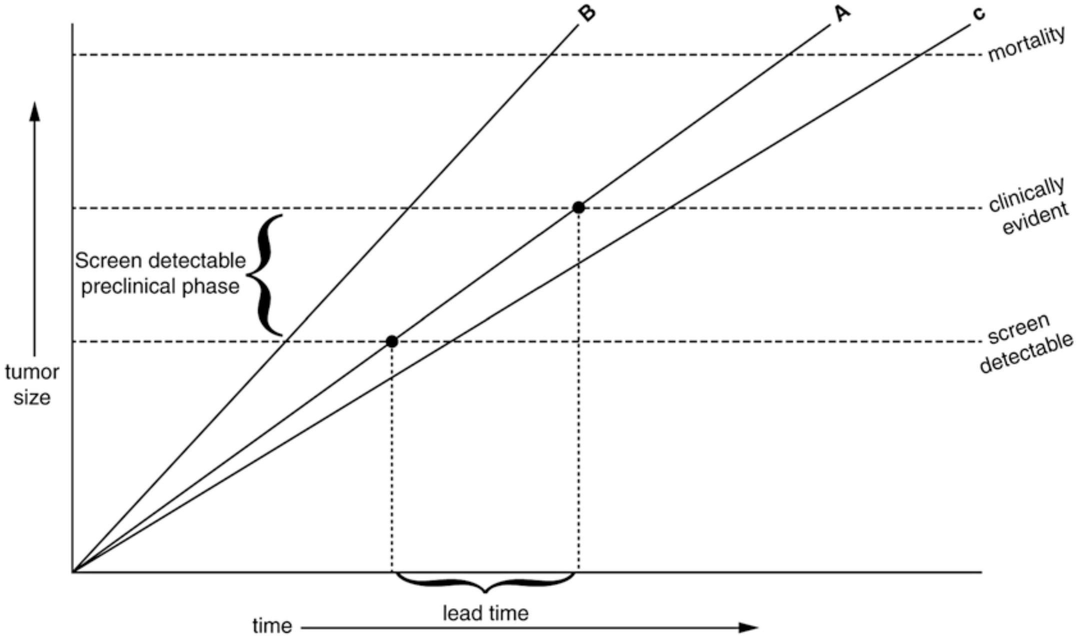

Fig. 2.3 Screening biases. For this figure, cancers are assumed to grow at a continuous rate until they reach a size at which death of the subject occurs. At a small size, the cancers may be evident on screening but not yet evident clinically. This is the preclinical screen detectable phase. Screening is potentially helpful if it detects cancer in this phase. After further growth, the cancer will be clinically evident. Even if the growth and outcome of the cancer is unaffected by screening, merely detecting the cancer earlier will increase apparent survival. This is the screening lead time. In addition, slower growing cancers (such as C) will exist in the preclinical screen detectable phase for longer than faster growing cancers (such as B). Therefore, screening is more likely to detect more indolent cancers, a phenomenon known as length bias. (Reprinted with kind permission of Springer Science + Business Media from Blackmore CC, Medina LS, Ravenel JG, Silvestri GA. Critically Assessing the Literature: Understanding Error and Bias. In Medina LS, Blackmore DD (eds): *Evidence-Based Imaging: Optimizing Imaging in Patient Care.* New York: Springer Science + Business Media, 2006.)

from the earlier detection of the disease which leads to longer time from diagnosis and an apparent survival advantage but does not truly impact the date of death. In effect, individuals live longer with the disease as diagnosis is made earlier, but still die at the same age. *Length time bias* relates to the virulence of tumors. More indolent or slowly growing tumors will persist longer at a size that can be detected by screening but is not yet clinically evident (referred to as the preclinical screen detectable phase) longer than faster-growing tumors that are more likely to be detected by symptoms. Thus, screen-detected tumors will tend to be less aggressive even at the same size, when compared to clinically detected tumors. This disproportionally assigns more indolent disease to the intervention group in screening trials and results in the appearance of a benefit. *Overdiagnosis* is the most extreme form of length time bias in which a disease is detected and "cured" but is so indolent it would never have caused symptoms during life and therefore, in the absence of screening, would never have been diagnosed. Thus, survival from diagnosis alone is not an appropriate measure of the effectiveness of screening [25].

For this reason, a randomized control trial (RCT) with disease-specific mortality as an endpoint is the preferred methodology. Randomization should even out the selection process in both arms, eliminating the bias of case-control studies and allow direct comparison of groups who were invited to undergo the intervention and those who were not, to see if the intervention lowers deaths due to the target disease. The disadvantage of the RCT is that it takes many years and is expensive to perform. There are two additional biases that can occur in RCTs and are important to understand: *sticky diagnosis* and *slippery linkage* [26]. Because the target disease is more likely to be detected in a screened population, it is more likely to be listed as a cause of death, even if not the true cause. As such, the diagnosis "sticks" and tends to underestimate the true value of the test. On the other hand, screening may set into motion a series of events in order to diagnose and treat the illness. If these procedures remotely lead to mortality, say a myocardial infarction during surgery with death several months later, the linkage of the cause of death to the screening may no longer be obvious (slippery linkage). Because the death is not appropriately assigned to the target disease, the value of screening may be overestimated. For this reason, in addition to disease-specific mortality, all-cause mortality should also be evaluated in the context of screening trials [26].

Because of these biases in screening trials, it important not to focus on irrelevant metrics, including survival, test sensitivity, disease prevalence, and detection of early stage disease. All of these are susceptible to bias that may make an ineffective screening test appear effective. Only disease-specific and all-cause mortality reduction (from invitation to screen or intention to treat analysis) are valid as measures of the effectiveness of screening trials [27].

Qualitative Literature Summary

The potential for error and bias makes the process of critically assessing a journal article complex and challenging, and no investigation is perfect. Producing an overall summation of the quality of a research report is difficult. However, there are grading schemes that provide a useful estimation of the value of a research report for guiding clinical practice. The method used in this textbook is derived from that of Kent [28] and is shown in Table 2.1. Use of such a grading scheme is by nature an oversimplification. However, such simple guidelines can provide a useful quick overview of the quality of a research report.

Table 2.1 Evidence classification for evaluation of a study

Level 1—Strong evidence: Studies with broad generalizability to most patients suspected of having the disease of concern: a prospective, blinded comparison of a diagnostic test result with a well-defined final diagnosis in an unbiased sample when assessing diagnostic accuracy or blinded randomized control trials when assessing therapeutic impact or patient outcomes. Well-designed meta-analysis based on level 1 or 2 studies.
Level 2—Moderate evidence: Prospective or retrospective studies with narrower spectrum of generalizability, with only a few flaws that are well described so their impact can be assessed but still requiring a blinded study of diagnostic accuracy on an unbiased sample. This includes well-designed cohort or case-control studies and randomized trials for therapeutic effects or patient outcomes.

(continued)

Table 2.1 (continued)

Level 3—Limited evidence: Diagnostic accuracy studies with several flaws in research methods, small sample sizes, or incomplete reporting OR nonrandomized comparisons for therapeutic impact or patient outcomes.
Level 4—Insufficient evidence: Studies with multiple flaws in research methods, case series, descriptive studies, or expert opinions without substantiating data.

(Reprinted with kind permission of Springer Science + Business Media from Blackmore CC, Medina LS, Ravenel JG, Silvestri GA. Critically Assessing the Literature: Understanding Error and Bias. In Medina LS, Blackmore DD (eds): *Evidence-Based Imaging: Optimizing Imaging in Patient Care.* New York: Springer Science + Business Media, 2006.)

Conclusion

In summary, critical analysis of a research publication can be a challenging task. The reader must consider the potential for Type I and Type II random error as well as systematic error introduced by biases including selection bias, observer bias, and reference standard bias. Screening includes an additional set of challenges related to the healthy volunteer effect, lead time, length bias, and overdiagnosis.

References

1. Moher D, Schulz K, Altman D. JAMA. 2001;285:1987–91.
2. Bossuyt PM, Reitsma J, Bruns D, et al. Acad Radiol. 2003;10:664–9.
3. Davidoff F, Batalden P, Stevens D, et al. Qual Saf Health Care. 2008;17S1:i3–9.
4. Hulley SB, Cummings SR. Designing clinical research. Baltimore: Williams and Wilkins; 1998.
5. Medina L, Zurakowski D. Radiology. 2003;226: 297–301.
6. Gallagher E. Acad Emerg Med. 1999;6:1084–7.
7. Lang T, Secic M. How to report statistics in medicine. Philadelphia: American College of Physicians; 1997.
8. Gardener M, Altman D. Br Med J. 1986;292:746–50.
9. Medina L, Aguirre E, Zurakowski D. Neuroimaging Clin Am. 2003;13:157–65.
10. Medina L. AJNR. 1999;20:1584–96.
11. Donner A. Stat Med. 1984;3:199–214.
12. Maclure M, Schneeweiss S. Epidemiology. 2001;12: 114–22.
13. Reid MC, Lachs MS, Feinstein AR. JAMA. 1995;274:645–51.
14. Lijmer JG, Mol BW, Heisterkamp S, et al. JAMA. 1999;282:1061–6.
15. Blackmore C. Acad Radiol. 2004;11:134–40.
16. Ransohoff DF, Feinstein AR. NEJM. 1978;299: 926–30.
17. Black WC. AJR. 1990;154:17–22.
18. Begg CB, McNeil BJ. Radiology. 1988;167:565–9.
19. Schulz K, Chalmers I, Hayes R, et al. JAMA. 1995;273:408–12.
20. Terasawa T, Blackmore CC, Bent S, et al. Ann Intern Med. 2004;141:537–46.
21. Pinsky PF, Miller A, Kramer BS, et al. Am J Epidemiol. 2007;165:874–81.
22. Marcus P. Lung Cancer. 2003;41:37–9.
23. Hosek R, Flanders W, Sasco A. Am J Epidemiol. 1996;143:193–201.
24. Patz E, Goodman P, Bepler G. N Engl J Med. 2000;343:1627–33.
25. Black WC, Welch HG. AJR. 1997;168:3–11.
26. Black W, Haggstrom D, Welch H. J Natl Cancer Inst. 2002;94:167–73.
27. Wegwarth O, Schwartz LM, Woloshin S, et al. Ann Intern Med. 2012;156:340–9.
28. Kent DL, Haynor DR, Longstreth WT, et al. Ann Intern Med. 1994;120:856–71.