2

Critically Assessing the Literature: Understanding Error and Bias

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

Issues

I. What are error and bias?
II. What is random error?
  A. Type I error
  B. Confidence intervals
  C. Type II error
  D. Power analysis
III. What is bias?
IV. What are the inherent biases in screening?
V. Qualitative literature summary

The keystone of the evidence-based imaging (EBI) approach is to critically assess the research data that are provided and to determine if the information is appropriate for use in answering the EBI question. Unfortunately, the published studies 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. Two recent initiatives, the CONSORT (1) and the STARD (2), aim to improve the reporting of clinical trials and studies of diagnostic accuracy, respectively. However, these guidelines are only now being implemented and are not well known to readers of the medical literature.

This chapter summarizes the common sources of error and bias in the imaging literature. Using the EBI approach requires an understanding of these issues.

I. What Are Error and Bias?

Errors in the medical literature can be divided into two main types. *Random error* occurs due to chance variation, causing a sample to be different from the underlying population. Random error is more likely to be problematic when the sample size is small. *Systematic error*, or *bias*, is an incorrect study result due to nonrandom

C.C. Blackmore (✉)
Department of Radiology, Center for Healthcare Solutions, Virginia Mason Medical Center, 1100 Ninth Avenue, Seattle, WA 98111, USA
e-mail: craig.blackmore@vmmc.org

L.S. Medina et al. (eds.), *Evidence-Based Imaging: Improving the Quality of Imaging in Patient Care, Revised Edition*,
DOI 10.1007/978-1-4419-7777-9_2, © Springer Science+Business Media, LLC 2011

distortion of the data. Systematic error is not affected by sample size, but is rather a function of flaws in the study design, data collection, and analysis. A second way to think about random and systematic error is in terms of precision and accuracy (3). Random error affects the precision of a result (Fig. 2.1). The larger the sample size, the more precision in the results and the more likely that two samples from truly different populations will be differentiated from each other. 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 (Fig. 2.1). 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 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).

II. What Is Random Error?

Random error is divided into two main types: Type I, or alpha error, occurs when an investigator concludes that an effect or a difference is present when in fact there is no true difference. Type II, or beta error, occurs when an investigator concludes that there is no effect or no difference when in fact a true difference exists in the underlying population (3).

A. Type I Error

Quantification of the likelihood of alpha error is provided by the familiar p value. A p value 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 effect, the difference observed in a sample is due to chance variation rather than a true underlying difference in the population.

There are limitations to the ubiquitous p values seen in imaging research reports (4). The 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 the use of p values and encouraging research results to be reported by way of confidence intervals (CIs).

B. Confidence Intervals

Confidence Intervals are preferred because they provide much more information than p values. CIs provide information about the precision of an estimate (how wide are the CIs), the size of an estimate (magnitude of the CIs), and the statistical significance of an estimate (whether the intervals include the null) (5).

If you assume that your sample was randomly selected from some population (that follows a normal distribution), you can be 95% certain that the CI includes the population mean. More precisely, if you generate many 95% CIs 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% (4). Therefore, the 95% CI is related to statistical significance at the $p=0.05$ level, which means that the interval itself can be used to determine if an estimated change is statistically significant at the 0.05 level (6). Whereas the p value is often interpreted as being either statistically significant or not, the CI, by providing a range of values, allows the reader to interpret the implications of the results at either end (6, 7). In addition, while p values have no units, CIs are presented in the units of the variable of interest, which helps readers to interpret the results. The CIs shift the interpretation from a qualitative judgment about the role of chance to a quantitative estimation of the biologic measure of effect (4, 6, 7).

CIs can be constructed for any desired level of confidence. There is nothing magical about the 95% that is traditionally used. If greater confidence is needed, then the intervals have to be wider. Consequently, 99% CIs are wider than 95, and 90% CIs are narrower than 95%. Wider CIs are associated with greater confidence but less precision. This is the trade-off (4).

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. However, one study had 50 subjects, while the other one had 500 subjects. At first glance, both studies appear to provide similar information. However, the narrower CIs for the larger study reflect greater precision and indicate the value of the larger sample size. For a smaller sample

$$95\%\text{CI} = 200 \pm 1.96\left(\frac{50}{\sqrt{50}}\right)$$

$$95\%\text{CI} = 200 \pm 14 = 186 - 214$$

For a larger sample

$$95\%\text{CI} = 200 \pm 1.96\left(\frac{50}{\sqrt{500}}\right)$$

$$95\%\text{CI} = 200 \pm 4 = 196 - 204$$

In the smaller series, the 95% CI was 186–214 cm/s, while in the larger series, the 95% CI was 196–204 cm/s. Therefore, the larger series has a narrower 95% CI (4).

C. Type II Error

The familiar p value alone 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. To conclude that no difference exists, the study must be powered sufficiently to detect a clinically important difference and have p value or CI indicating no significant effect.

D. Power Analysis

Power analysis plays an important role in determining what an adequate sample size is, so that meaningful results can be obtained (8). 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 (3). Mathematically, power is defined as 1 minus beta ($1-\beta$), 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. Type I errors, in contrast, are analogous false positives in a study population (7). For example, if β is set at 0.10, then the researchers acknowledge that they are willing to accept a 10% chance of missing a correlation between abnormal computed tomography (CT) angiographic findings in 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% is rarely achievable, so, at best, a study should reduce the false negatives (β) and false positives (α) to a minimum (3, 9). 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 noncarotid artery disease arms) if you change your desired power from 85 to 90% (10). 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.

III. What Is Bias?

The risk of an error from bias decreases as the rigorousness of the study design and analysis increases. Randomized controlled trials (RCTs) 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. In nonrandomized studies,

appropriate study design and statistical analysis can control only 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 (11) have identified ten different levels at which biases can distort the relationship between published study results and truth. Unfortunately, bias is common in published reports (12), and reports with identifiable biases often overestimate the accuracy of diagnostic tests (13). Careful surveillance for each of these individual bias phenomena is critical, but may be a challenge. Different study designs are also susceptible to different types of bias, as will be discussed in this section as well. Well-reported studies 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 it. 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.

A common bias encountered in imaging research is that of *selection bias* (14). 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 and on normal controls. In an evaluation of the effect of bias on study results, Lijmer et al. (13) found the greatest overestimation of test accuracy with this type of spectrum bias.

A second frequently encountered bias in imaging literature is that of *observer bias* (15, 16), also called test-review bias and diagnostic-review bias (17). Imaging tests are largely 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; incentives 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 an intervention is being assessed. Important distortions in research results have been found when observers are not blinded vs. blinded. For example, Schulz et al. (18) showed a 17% greater outcome improvement in studies with unblinded assessment of outcomes versus those with blinded assessment. 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.

Bias can also be introduced by the *reference standard* used to confirm the final diagnosis. 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 (13, 16). Of course, sometimes it is not possible or ethical to perform the same reference standard procedure on all subjects. For example, in a recent meta-analysis of imaging for appendicitis, Terasawa et al. (19) found that all of the identified studies used a different reference standard for subjects with positive imaging (appendectomy and pathologic evaluation) than for those with negative imaging (clinical follow-up). It simply would not be ethical to perform appendectomy on all subjects. Likely the sensitivity and specificity of imaging for appendicitis was overestimated as a result.

IV. What Are the Inherent Biases in Screening?

Investigations of screening tests are susceptible to an additional set of biases. Screening case–control trials are vulnerable to *screening selection bias*. For example, lung cancer case–control studies have been performed in Japan, where long-running tuberculosis control programs have been in place. This allowed for the analysis of those who were screened to be matched with a database of matched unscreened controls to arrive at a relative risk of dying from lung cancer in screened and unscreened populations. Because screening is a choice in these studies, selection bias plays a prominent role. That is, people who present for elective screening tend to have better health habits (20). In assessing the exposure history of cases, the inclusion of 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 (21). 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 on the stage distribution and survival 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 (22) (Fig. 2.3). *Lead-time bias* results 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. *Length-time bias* relates to the virulence of tumors. More indolent tumors are more likely to be detected by screening, whereas aggressive tumors are more likely to be detected by symptoms. This disproportionally assigns more indolent disease to the intervention group 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 it is so indolent that it would have never caused symptoms during life. Thus, survival alone is not an appropriate measure of the effectiveness of screening (23).

For this reason, a RCT with disease-specific mortality as an end point is the preferred methodology. Randomization should even out the selection process in both arms, eliminating the bias of case–control studies and allowing direct comparison of groups that underwent the intervention and those that did 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 biases that can occur in RCTs and are important to understand: *sticky diagnosis* and *slippery linkage* (24). 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, such as 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 (24). Ultimately, to show the effectiveness of screening, not only more early-stage cancers need to be found in the screened group, but also there must be fewer late-stage cancers (stage shift) (22).

V. 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 book is derived from that of Kent et al. (25) 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.

Conclusion

Critical analysis of a research publication can be a challenging task. The reader must consider the potential for type I and type II random errors, 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 lead time, length bias, and overdiagnosis. These challenges may seem daunting, yet without an understanding of them, a medical practitioner can not learn efficiently from the literature and in so doing, help their patients with the best evidence we have to offer them.

Take Home Tables and Figures

Table 2.1 and Figs. 2.1–2.3 serve to highlight key recommendations and supporting evidence.

Table 2.1. Evidence classification for evaluation of a study

Level I: 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 or when assessing therapeutic impact or patient outcomes. Well-designed meta-analysis based on level I or II studies**
Level II: Moderate evidence **Prospective or retrospective studies with narrower spectrum of generalizability, with only a few flaws that are well described so that 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**
Level III: 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 IV: 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.

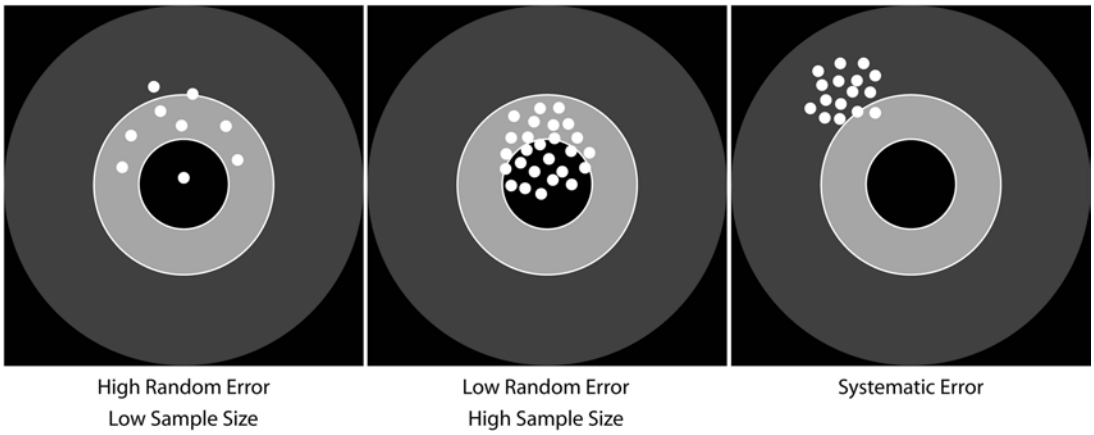


Figure 2.1. Random and systematic errors. Using the bull's-eye analogy, the larger the sample size, the less the random error and the larger the chance of hitting the center of the target. In systematic error, regardless of the sample size, the bias would not allow the researcher to hit the center of the target. (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.)

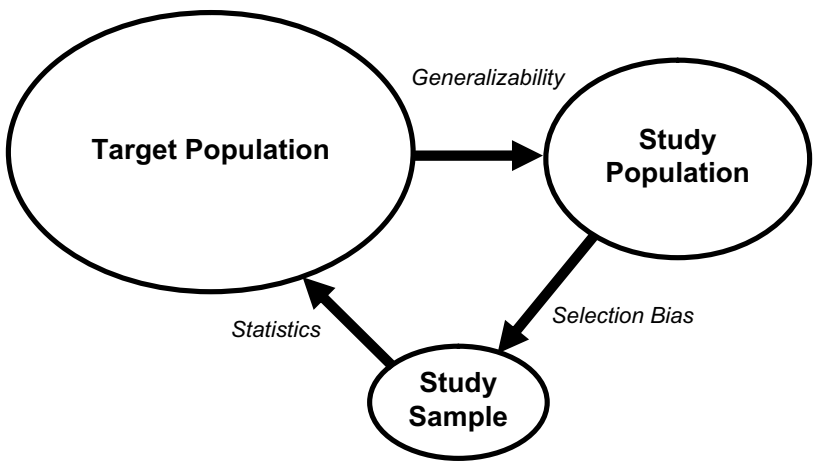


Figure 2.2. Population and sample. The target population represents the universe of subjects who are at risk for a particular disease or condition. In this example, all subjects with abdominal pain are at risk for appendicitis. The sample population is the group of eligible subjects available to the 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.)

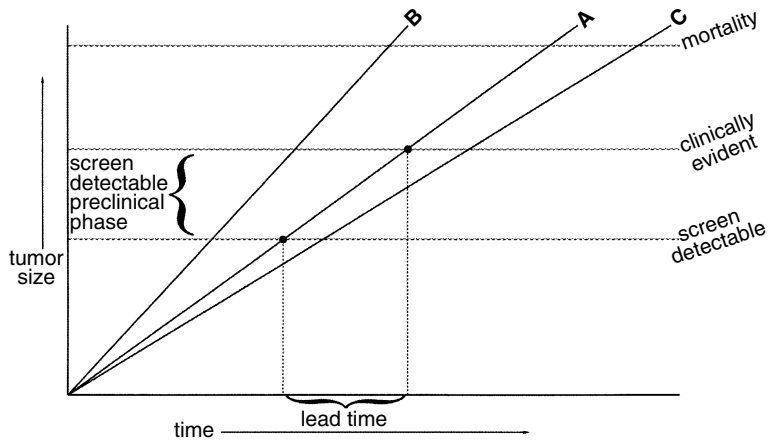


Figure 2.3. Screening biases. For this figure, cancers are assumed to grow at a continuous rate until they reach a size at which death of the subject occurs. At a small size, the cancers may be evident on screening, but not yet evident clinically. This is the preclinical screen-detectable phase. Screening 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 do 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.)

References

1. Moher D, Schulz K, Altman D. JAMA 2001;285: 1987–1991.
2. Bossuyt PM, Reitsma J, Bruns D, et al. Acad Radiol 2003;10:664–669.
3. Hulley SB, Cummings SR. Designing Clinical Research. Baltimore: Williams and Wilkins, 1998.
4. Medina L, Zurakowski D. Radiology 2003;226: 297–301.
5. Gallagher E. Acad Emerg Med 1999;6: 1084–1087.
6. Lang T, Secic M. How to Report Statistics in Medicine. Philadelphia: American College of Physicians, 1997.
7. Gardener M, Altman D. Br Med J 1986;292: 746–750.
8. Medina L, Aguirre E, Zurakowski D. Neuroimag Clin North Am 2003;13:157–165.
9. Medina L. AJNR Am J Neuroradiol 1999;20: 1584–1596.
10. Donner A. Stat Med 1984;3:199–214.
11. Maclure M, Schneeweiss S. Epidemiology 2001;12:114–122.
12. Reid MC, Lachs MS, Feinstein AR. JAMA 1995;274:645–651.
13. Lijmer JG, Mol BW, Heisterkamp S, et al. JAMA 1999;282:1061–1066.
14. Blackmore C. Acad Radiol 2004;11:134–140.
15. Ransohoff DF, Feinstein AR. N Engl J Med 1978;299:926–930.
16. Black WC. AJR Am J Roentgenol 1990;154:17–22.
17. Begg CB, McNeil BJ. Radiology 1988;167: 565–569.
18. Schulz K, Chalmers I, Hayes R, Altman D. JAMA 1995;273:408–412.
19. Terasawa T, Blackmore C, Bent S, Kohlwes R. Ann Intern Med 2004;141:537–546.
20. Marcus P. Lung Cancer 2003;41:37–39.
21. Hosek R, Flanders W, Sasco A. Am J Epidemiol 1996;143:193–201.
22. Patz E, Goodman P, Bepler G. N Engl J Med 2000;343:1627–1633.
23. Black WC, Welch HG. AJR Am J Roentgenol 1997;168:3–11.
24. Black W, Haggstrom D, Welch H. J Natl Cancer Inst 2002;94:167–173.
25. Kent DL, Haynor DR, Longstreth WT Jr, Larson EB. Ann Intern Med 1994;120:856–871.

http://www.springer.com/978-1-4419-7776-2

Evidence-Based Imaging
Improving the Quality of Imaging in Patient Care
Medina, L.S.; Blackmore, C.C.; Applegate, K. (Eds.)
2011, XX, 680 p. 264 illus., 20 illus. in color., Softcover
ISBN: 978-1-4419-7776-2