Skeletal Radiol (2008) 37:393–396
DOI 10.1007/s00256-008-0448-1

PERSPECTIVE

Statistics and methodology

Nancy A. Obuchowski · Michael L. Lieber

Published online: 8 February 2008

Introduction

Study design and analysis are critical parts of a research study. Over the past few decades multiple papers have been written describing common problems in medical research studies, including radiology studies, and methods for correcting these problems [1–3].

Two statisticians (one Ph.D., one M.S.) experienced in study design and analysis of clinical imaging studies reviewed manuscripts published in *Skeletal Radiology* in the first 6 months of 2007 to evaluate the validity of the study designs and analyses. We limited our review to manuscripts of original research with sample sizes of more than ten patients. We recorded the following study design and analysis characteristics for each article: number of patients, number of readers, study design, primary endpoint, whether patients' age and gender were reported, blinding of readers, presence of bias, consensus readings, and validity of statistical analyses. In this paper we report our findings and draw attention to five issues.

Results of 6-month review of *Skeletal Radiology*

Our 6-month review of original research papers in *Skeletal Radiology* included almost 20 original studies of knee, ankle, shoulder, foot, hand, and lumbar imaging. Approximately one-third of studies were prospective designs and the rest were retrospective. Every study clearly indicated the goals of the study and primary endpoints. Every study reported the study subjects' age and gender distributions. These are strengths of the *Skeletal Radiology* papers we reviewed.

The patient sample sizes ranged from 14 to 104 (average 52.5). In the smaller studies, the sample size caused two different kinds of problems when the study results were interpreted: confidence intervals so wide they were useless, and P values>0.05 (i.e., non-significant *statistically*) associated with findings that were *clinically* significant. We discuss these problems in more detail in the section "Limitations of a small sample size".

In studies which included readers' interpretations of imaging findings, all but one clearly stated that the readers had been appropriately blinded to competing test results and reference standard (i.e., gold standard) results. A third of the studies had just one reader; the remaining had two readers. Studies with just one reader are problematic, because the results are limited to that particular reader's cognitive and perceptual abilities. Those studies also provide no information about inter-reader variability. We discuss these limitations in "Limitations of one-reader study designs". In the two-reader studies, more than half used consensus readings. In "Consensus reads" we illustrate the advantages of reporting each reader's results, along with an average of the readers' results, rather than a consensus result.

We found several cases of incorrect data analyses. In "Common data analyses problems" we discuss two common problems: (1) confusion between agreement and correlation, and (2) analysis of clustered data (i.e., multiple observations from the same patient, e.g., multiple vertebral fractures in the same patient).

N. A. Obuchowski (✉) · M. L. Lieber
Department of Quantitative Health Sciences,
Lerner Research Institute and Imaging Institute,
Cleveland Clinic, 9500 Euclid Ave.,
Cleveland, OH 44195, USA
e-mail: obuchon@ccf.org

M. L. Lieber
e-mail: lieberm@ccf.org

Finally, we found several papers with verification bias, the most common bias in radiology studies. We illustrate the problem of verification bias and discuss some solutions in "Verification bias".

Limitations of a small sample size

When a research study is conceived, it is important for one to determine the statistical hypothesis that will be tested. For example, the hypothesis might be that a new test has an accuracy superior to that of the conventional test, or that the new test has an accuracy at least equivalent to that of the conventional test. Once the hypothesis has been determined and the primary endpoint has been chosen (e.g., diagnostic accuracy), one needs to use the appropriate statistical method to determine the sample size required for the study. There is no single method for determining sample size that works for all research studies; rather, one must consider the study design, statistical hypothesis, and primary endpoint, and then find the appropriate statistical method. Someone with training in biostatistics can usually determine the required sample size for a particular study.

Unfortunately, many research studies do not go through this process at the time the study is planned. Rather, the sample size for the study is based on other factors, such as limitations in patient volume, research time, or money. In these circumstances the study's sample size is often too small to address fully the study questions and test the study's hypotheses.

Both *P* values and confidence intervals (CIs) are used extensively in *Skeletal Radiology*, as in all medical literature, to quantify the results of the tested statistical hypotheses and to make conclusions. Both *P* values and CIs are heavily influenced by sample size. In other words, a study can reveal important differences between two imaging tests, but, if the sample size is too small, the *P* value will not reach statistical significance (i.e., it will not be <0.05). When researchers are interpreting their study results, they need to make a distinction between *clinical significance* and *statistical significance*. *Clinical significance* means that the differences observed, for example the differences in accuracy between two imaging tests, are important and would change practice if they are real. *Statistical significance* gives credence that the differences observed are real, not just due to chance. If the sample size is too small, however, clinically significant differences will not reach statistical significance. Thus, the capacity of the study to help us move forward in our understanding of the imaging test's capabilities will be limited. (A study that finds clinically significant results, but fails to achieve statistical significance, may, however, serve to steer future investigators to areas deserving of further study, and may lay the foundation for subsequent research, particularly when dealing with an uncommon medical phenomenon.)

In the *Skeletal Radiology* papers that we reviewed there were several situations where an investigator concluded that an imaging test did not meet his or her expectations because the *P* value was not significant. The authors, however, had not distinguished clinical significance from statistical significance and, thus, might have come to an incorrect conclusion. In other words, even in the absence of *statistical* significance, clinically significant trends in the data should be reported, along with the caveat that there is insufficient evidence of statistical significance.

Similarly, sample size and the width of confidence intervals are highly negatively correlated. If the sample size is small, the confidence intervals will be wide. In one *Skeletal Radiology* paper we reviewed, the authors reported a confidence interval for the sensitivity of a diagnostic test that was near zero for the lower bound and near 100% for the upper bound. Clearly, this confidence interval tells us nothing, because the sample size was too small.

Limitations of one-reader study designs

Despite the understandable practical appeal of single-reader studies, they have severe potential drawbacks and should thus be avoided in favor of multiple-reader study designs. The diagnostic accuracy of any imaging modality under study is a function of both the machine and the reader using that machine [4]. Since every reader has a different cognitive, visual, and perceptual skill set, and brings to the study a unique combination of background, training, specialization, and level of experience, it is important that any diagnostic accuracy study include more than one reader. In this way, a study can assess the overall performance of the reader–modality combination, as opposed to the accuracy of a specific reader using a specific modality.

Variability among readers has been studied and quantified. In their 1996 paper, Beam et al. [5] had 108 radiologists from 50 US mammography centers look at the same set of 79 screening mammograms. They found that sensitivities ranged from 0.47 to 1.0 and specificities ranged from 0.36 to 0.99. Theirs, and other studies, have documented wide variability in accuracy among the population of radiologists.

Single-reader studies are unable to account for the effects of differences among readers. Instead, a single reader is erroneously assumed to be representative of all readers in the reader population of interest. This is analogous to a study that includes only a single subject, assumes that the lone subject represents the entire patient population, and ignores all the inter-patient variability in the patient

population. Just as we include a sample of patients in a research study to deal with diversity among patients, so too should we include more than one reader in diagnostic imaging studies. In cases where circumstances preclude the use of multiple readers, this limitation should be acknowledged, and the conclusions of the single-reader study should be subject to confirmation by subsequent multi-reader studies.

Consensus reads

In single-reader studies, we collect no data on inter-reader variability. In the case of consensus readings, data from multiple readers is needlessly lost, as a single consensus replaces the assessments of individual radiologists. In both instances, the existence of inter-reader variability is ignored, weakening the entire study.

Using a consensus or majority interpretation is analogous to performing a study with multiple subjects, but, instead of recording and analyzing individual patient responses, patient data are combined into a single pooled value. Similarly, when a single interpretation is recorded for multiple readers, information on inter-reader differences is lost.

A specific risk inherent in consensus readings is that some readers will be inordinately influenced by other readers (e.g., junior readers deferring to senior readers, or more-passive readers going along with more-persuasive readers). In such cases, a consensus opinion represents a biased (i.e., not equally weighted) average of the various readers' opinions.

As with single-reader study designs, consensus readings have a practical appeal. They are less time-consuming overall and easier to conduct than studies in which each reader interprets cases independently. But, as Obuchowski and Zepp [6] have pointed out, "for the results of research studies to have any practical application, they must imitate the normal activities of radiologists and not a peculiar arrangement rarely seen in ordinary practice." Thus, consensus readings not only produce problematic statistical results that ignore inter-reader variability, but such studies also may not be readily generalizable to standard clinical practice among radiologists. Instead, readers should interpret and record their findings independently of other readers in the study. The accuracy of each reader should be assessed and reported separately.

Common data analyses problems

There were two errors that occurred in two or more papers. The first error was confusion between correlation and agreement. Two tests can give results that are perfectly correlated but never agree; yet, if two tests agree, then they are also perfectly correlated. The concept is very simple: If two tests are measuring the size of a lesion, the measurements in one test can be exactly 5 mm greater than the other test's measurement for every lesion measured. The two tests' results are perfectly correlated (i.e., correlation coefficient, r, equals 1.0), yet the measurements never agree with each other.

We usually want to talk about *correlation* when we are comparing two distinct characteristics of a patient, e.g., size of the lesion and severity of symptoms. Correlation is measured by correlation coefficients, such as Pearson's correlation coefficient (for normally distributed variables) and Kendall's tau (a nonparametric measure). On the other hand, we want to talk about *agreement* when we are comparing the same characteristic but measured by two different methods, e.g., size of the lesion as measured by two different tests. Agreement is measured by kappa statistics and intraclass correlation coefficients (ICCs) and analyzed by approaches such as those described by Bland and Altman [7].

The second data analysis problem in *Skeletal Radiology* papers occurred in approximately 25% of the studies we reviewed. The investigators were studying more than one finding in each patient, e.g., two feet per patient, multiple ligaments per patient, variable number of vertebral bodies per patient, one or more meniscal tears in the same patient, yet the investigators analyzed the data as if each finding came from a different patient. These data are called *clustered data* and are very common in radiology studies, as in all medical studies. It is a natural phenomenon, but the usual statistical methods do not apply to clustered data. The problem is this: multiple observations from the same patient are almost always correlated, as least to some small degree. Whenever there is correlation between observations, we must account for this in the statistical analyses; otherwise, the *P* values will be wrong (usually too small) and the confidence intervals will be incorrect (usually too narrow). There have been several papers in the literature addressing the issue of clustered data in imaging studies and describing how to analyze these data correctly [8, 9].

There are some incorrect approaches to clustered data that we should warn against. First, it is not appropriate to test for the presence of correlation, and if the result is non-significant, then to ignore the clustered data. Chances are good that the result will be non-significant because the correlation is small (but still real!) and/or the sample size is small. A second approach, although valid, leads to a different kind of problem. In this approach the investigators omit observations until there is only one observation left per patient. The problem here is that the reduced sample size also reduces the statistical power of the study, i.e., the

ability to reach statistical significance when clinically significant findings are present. The best approach to clustered data is to apply appropriate methods that take into account the number of findings, as well as the number of patients that these findings came from.

Verification bias

Verification bias occurs often in studies designed to measure the diagnostic accuracy [i.e., sensitivity, specificity, receiver operating characteristic (ROC) curves] of medical tests. This bias occurs when patients undergoing the test(s) do not always undergo the gold standard (reference test) and the test results influence the decision to perform the gold standard. The consequence of verification bias is incorrect and misleading estimates of test accuracy.

Consider a study comparing the accuracy of MRI and CT to detect meniscal tears. The gold standard is the findings at arthroscopy, but not all patients undergoing imaging to detect tears will undergo arthroscopy. If the MRI or CT results are used to help decide which patients undergo arthroscopy, and if diagnostic accuracy is measured just on those patients who undergo arthroscopy, then verification bias has occurred. Usually, patients with a positive test result (e.g., meniscal tear detected on imaging) are referred to the gold standard (arthroscopy), while patients with a negative result (without a detected meniscal tear) are less likely to undergo the gold standard. That is, patients with positive results (true positives and false positives) are verified by the gold standard and thus included in the study; on the other hand, patients with negative results (true negatives and false negatives) tend not to be verified and thus excluded from the study. When sensitivity is calculated, it is usually over-estimated (too high) because (false) negatives tended not to be in the study sample. Similarly, when specificity is calculated, it is usually under-estimated (too low) because many of the (true) negatives were not in the study sample.

The ideal way to deal with verification bias is to avoid it by designing the study so that the tests being evaluated are not allowed to influence the decision to perform the gold standard. For example, sometimes the decision to perform the gold standard can be based on other tests or clinical findings, and the tests under study are not interpreted until a later date. Another approach is to have more than one gold standard. For example, patients with a negative test result might be followed clinically and/or radiographically for some period of time, assuming that if the condition were present, then it would worsen over time. When these approaches are not feasible, there are statistical models which can be used to estimate test accuracy correctly, as long as the study includes test results from all patients, regardless of whether or not they underwent the gold standard [10].

References

1. Reid MC, Lachs MS, Feinstein AR. Use of methodologic standards in diagnostic test research: getting better but still not good. JAMA 1995; 274: 645–651.
2. Sica GT. Bias in research studies. Radiology 2006; 238: 780–789.
3. Obuchowski NA. Special topics III: bias. Radiology 2003; 229: 617–621.
4. Beam CA. Strategies for improving power in diagnostic radiology research. AJR Am J Roentgenol 1992; 159: 631–637.
5. Beam CA, Layde PM, Sullivan DC. Variability in the interpretation of screening mammograms by US radiologists: findings from a national sample. Arch Intern Med 1996; 156: 209–213.
6. Obuchowski NA, Zepp RC. Simple steps for improving multiple-reader studies in radiology: perspective. AJR Am J Roentgenol 1996; 166: 517–521.
7. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 1: 307–310.
8. Gonen M, Panageas KS, Larson SM. Statistical issues in analysis of diagnostic imaging experiments with multiple observations per patient. Radiology 2001; 221: 763–767.
9. Obuchowski NA. On the comparison of correlated proportions for clustered data. Stat Med 1998; 17: 1495–1507.
10. Zhou XH, Obuchowski NA, McClish DL. Statistical methods in diagnostic medicine. New York: Wiley; 2002.