God, Like the Devil, is in the Details[1]

Kevin S. Berbaum, PhD

In this issue, Cagnon and colleagues (1) present their quality control program for the American College of Radiology Imaging Network (ACRIN) part of the National Lung Screen Trial (NLST). For the ACRIN-NLST study, more than 18,000 participants were screened for lung cancer with radiography and computed tomography. There were many different imaging centers, models of multi-detector helical CT scanners, and models of digital, computed and film radiography equipment. The role of the program described goes well beyond what we would ordinarily think of as "quality control." Scientifically, it is the heart and soul of the trial, particularly because there seems to have been little consensus prior to the trial on an appropriate low-dose CT lung screening protocol. Without a clear definition of the imaging that could apply across the entire study and ways of checking whether that definition was applied, it is unlikely that a trial with so many patients, radiologists, sites, and scanners would succeed in answering the question of whether CT screening for lung cancer is better than radiography. The judgments that the physicists made to formulate this definition (e.g., how to "strike a balance between image quality and ionizing radiation") will determine the usefulness of the conclusions from the ACRIN trial. Although the paper may leave some readers hungry for more detail about protocol compliance, it does suggest that, despite extensive measures, half of all sites required corrective intervention but that the problems were quickly and effectively settled.

The large, multi-center, randomized, controlled trials provided by ACRIN are impressive in their ability to marshal large amounts of data to address important questions. That said, it must also be acknowledged that most of our scientific knowledge about medical imaging does not come from this source. Many smaller experiments from individual investigators at single institutions address questions in diagnostic radiology. Some of these experiments are not supported by government grants at all, and others are supported by standard R01 level budgets. The cost of a large trial—perhaps 25 million dollars or more—and of a smaller experiment—typically less than a million—means that only a few questions can be addressed with a clinical trial; many important questions must be answered with more modest means. Whatever their limitations, the many small experiments in imaging provide some scientific evidence where otherwise there would be none.

Some things are more troublesome and costly to find out about than others. Large trials are necessary if research questions can only be handled effectively in a large prospective trial. In particular, screening typically involves low prevalence of disease. Because sensitivity can be traded for specificity, we need sufficient samples of both diseased and non-diseased patients. The best but most costly way to do so is to study many more normal patients than needed to measure specificity in order to sample as many diseased patients as needed to measure sensitivity. The alternative is to gather a stratified sample with a greater proportion of diseased patients than would be found in clinical practice. Real sampling has advantages over stratified sampling. A stratified sample cannot usually be collected prospectively, and may not always be representative of typical image reading with its larger proportion of diseased patients. Often stratified sampling is what we can afford.

An example of a small study may be helpful. Franken, *et al.* (2) reported a study of 100 neonatal intensive care radiographs read by each of four pediatric radiologists on a film viewer or monitors that presented digitized versions of the same radiographs. The study was not supported by a grant; it was just part of a general program of research aimed at finding out whether early adoption of digital display was feasible. It used neonatal examinations as a

Acad Radiol 2006; 13:1311–1316

[1] From the University of Iowa, 200 Hawkins Drive, Iowa City, Iowa 52242-1077. Received Sept 22, 2006; accepted Sept 22, 2006. **Address correspondence to:** K.S.B. e-mail: kevin-berbaum@uiowa.edu

doi:10.1016/j.acra.2006.09.053

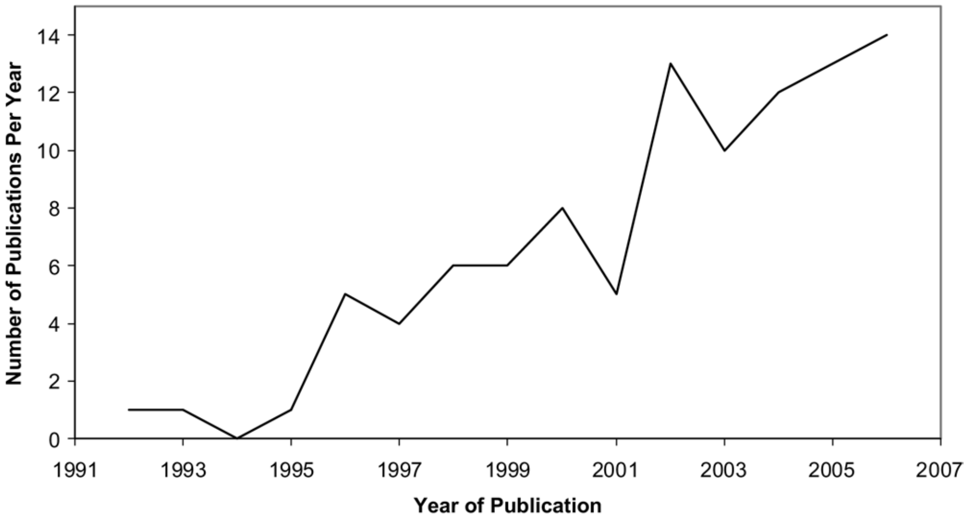

Figure 1. Published experiments per year using DBM MRMC ROC methodology. (Only nine months of 2006 are included.)

bellwether for sufficient resolution. Dorfman, Berbaum, and Metz (3) (DBM) developed a multireader, multicase (MRMC) ROC methodology for analyzing such data and showed how the data of Franken *et al.* (2) could be treated by their method and alternatively the methods of Metz, Wang and Kronman. (4) The results were that digital viewing and film viewing did not differ in diagnostic accuracy.

DBM has since been modified (5–10) and extended. Some of the extensions involved better ROC modeling and curve fitting; (11–13) some extended multireader techniques to multiple responses per case (14–16). Some extensions could also be called alternative methodologies. There were precursors to the DBM methodology (17–22) and alternative methods later become available (15,16, 23–37). Neither has the pace of MRMC ROC development decreased (38–42). Figure 1 gives some idea of the extent of MRMC studies as it shows the studies that used DBM MRMC ROC methodology to analyze their data (2,43–140). Although utilization of the DBM MRMC methodology far outstrips that of any other MRMC method in diagnostic radiology research, there are numerous additional experiments that use alternative methods and the use of the alternative approaches is increasing. There are also many experiments analyzed with single reader methods (4). So Figure 1 only tells a part of the story.

The Franken, *et al.* (2) experiment is representative of much of radiology research. Because it examined abnormalities in intensive care where most patients are sick, it was not difficult to develop a representative sample. Because a pediatric radiologist was needed to assemble and prove the sample and the Radiology Department at the University of Iowa had only three other pediatric radiology faculty and a fellow, only four expert readers were available for the study. The sample of cases was limited in size by the time needed to prove the cases and to read them. More cases could have been included, but at that time it was unclear how many would be needed. In fact, more cases were offered once Franken and colleagues discovered how long it would take Dorfman and colleagues to develop their MRMC methodology—two years from when data collection had been completed.

Since the time of that study, newer methods have become available. With the use of a proper ROC model (13) and more modern statistical approaches (7–10) (DBM MRMC 2.1 software, available from http://perception.radiology.uiowa.edu and from http://xray.bsd.uchicago.edu/krl/roc_soft.htm), the results of that study changed. Area under the ROC curve was still 0.87 for digital display and 0.85 for film, but the difference has become significant ($F(1,3) = 18.34$, $p = 0.023$). The conclusion of the study would not change. (Those who have embraced digital can now breathe a sigh of relief.) If we were to plan a new study using similar readers and patients, we could use the data of Franken *et al.* (2) as a pilot experiment to support our calculations of minimum reader and case sample sizes. We could look up the figures we need at a website that provides such projections using various published multi-reader multi-case data (see