Published in final edited form as:

Acad Radiol. 2009 May ; 16(5): 638–640. doi:10.1016/j.acra.2008.11.013.

Imaging Technology and Practice Assessments: What Next?

David Gur, ScD

From the Department of Radiology, University of Pittsburgh, Pittsburgh, PA 15213-3180

A recent study conducted by our research group revealed that there may be a significant “laboratory effect” in retrospective observer performance studies that may in turn limit the relevance of inferences generated as a result of these studies to the clinical environment (1). Even though we found a considerable consistency among the performance levels of the readers regardless of the specific method or rating scale used (i.e., binary, receiver operating characteristic (ROC) or free-response ROC (FROC)) (2,3), our study showed that radiologists may perform significantly different in the clinic than in the laboratory when reading the very same cases. The differences were reflected in terms of both their overall performance levels (e.g. sensitivity and specificity) and, perhaps more important, in terms of variability, or “spread”, among the observers’ performance levels (1). Although the results of our study should definitely be validated experimentally in other studies before general acceptance, there is a reasonably solid rationale for the observed outcome of our own study (1). Indeed radiologists may perform differently in the laboratory in ways that are not a-priori predictable; therefore, differences in their behavior can not always be completely accounted for during retrospective studies. Even attempting to duplicate seemingly simple conditions such as practice guidelines (e.g. aim for 10% recall rate) during the experiment ultimately may not reflect actual behavior in the clinic. Clinical decisions that affect patient management can not be duplicated in laboratory studies. Because differences in behavior are difficult to account for in retrospective observer performance studies we, the investigators, have to ask ourselves what next? How do we proceed with appropriate evaluations of new technologies and practices in a manner that is both practical and, at the same time, clinically relevant?

In the 1980s when a group of investigators, including myself, were working on stroke models and absolute measurements of regional and local cerebral perfusion using non-radioactive xenon computed tomography (XeCT), the results of the Extracranial -Intracranial (EC/IC) Bypass study (4) were published. This randomized trial to assess the efficacy of this seemingly ideal surgical procedure, which at the time was rapidly growing in the number of operations performed around the world, showed that the procedure was actually not as beneficial as originally perceived by most. In effect, it was more harmful than beneficial in the general population it had been applied to at the time. This was a great surprise and a significant disappointment to the field of micro-vascular neurosurgery. As a member of a group of investigators interested in this very question, we strongly believed we had a very “appropriate” way, namely to use XeCT perfusion measurements, to select a sub-set of the population in question who “should clearly benefit” from this surgical procedure, despite the overall negative results of the EC/IC bypass study. Shortly after the results of the EC/IC bypass study were published, we met with the Principal Investigator (PI) of the study and his associates in London,

Correspondence: David Gur, University of Pittsburgh, Department of Radiology Radiology, Imaging Research, 3362 Fifth Avenue, Pittsburgh, PA 15213, Phone: 412-641-2513, Fax: 412-641-2582, Email: E-mail: gurd@upmc.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

(Ontario, Canada) to discuss this very issue. At the end of a long day after much discussion, the point was made in no outstanding terms that “if we as investigators truly believe our hypothesis and there are no conclusive data to support a definitive conclusion, we should design a prospective study that directly tests this hypothesis by randomizing the very select sub-set of the population into “surgery” and “no surgery” arms and there is no way around it!” “You clearly have to have the guts to randomize them to “surgery” and “no surgery” groups, leave the “no surgery group alone””, the PI of the EC/IC bypass study repeated. “Yes, I know that this goes directly against your belief since you “know” you are right and you will try all types of study designs that may circumvent the need to “risk”, at least in your opinion, half the members of the very group you strongly believe are most likely to benefit from the operation by actually not operating on them just to test your own hypothesis.” He paused, took a minute and continued, “however, until you are willing to throw some people in the trash can (these had been his actual words and they stuck in mind since), you will never prove the point the way you should and in the long run society will never benefit from your work as it should or could.” Not being a physician myself it was relatively easy for me to nod my head in agreement, but the neurosurgeons present felt strongly that there must be a way to avoid this “brute force” approach.

Although not perfect, a randomized controlled trial (RCT) is the most natural approach to studying these clinically important questions. In treatment related studies (e.g. medical, surgical, and radiation oncology) RCTs and/or similar approaches have been widely accepted and implemented for many years and in numerous studies. This type of study allows the assessment of outcome in a largely natural progression of the populations being investigated under different intervention/treatment arms. In most, if not all, oncology related studies of this nature it is virtually impossible to perform a logical “OR” type clinical study; hence, the two arms are completely separated. Two large trials, The National Lung Screening Trial (NLST), a large diagnostic imaging based RCT, headed cooperatively by the American College of Radiology Imaging Network (ACRIN), and The Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial, (PLCO) clearly deserve much credit for leading the imaging diagnostic field in this regard (5,6). At the same time, in diagnostic medicine, in general, and in radiology, in particular, there are scenarios in which the nature of the study frequently resulting in a possible follow-up of participants suspected of having the condition of interest enables a somewhat different study design to be implemented without a significant loss of information in regards to the hypotheses being tested. One such example is the Digital Mammographic Imaging Screening Trial (DMIST) (7) that actually preceded the NLST. I have been critical of the DMIST for various reasons (8). At the same time, the investigators of this study deserve much credit for agreeing to increase the risks to participants (e.g. radiation, participant anxiety, increasing number of diagnostic follow-ups that may result in increasing the number of biopsies, etc.) in order to ascertain extremely important information that could ultimately benefit future patients through altering substantially the way radiology practices. The DMIST study design was a logical “OR” type study in which participants undergo both diagnostic imaging arms and are considered “positive” (in this case recommended for future diagnostic workup) when either of the arms result in positive/suspicious findings (warranting a diagnostic workup). This type of a study design is somewhat unique to diagnostic imaging and despite the somewhat, yet often limited, increased risk to participants it should be considered more often in our field.

Clearly, observational studies are important in assessing the actual a-posteriori outcome as a result of introduction of new technologies or practices into the clinic, but they are not perfect and should be viewed with caution (9,10). As related to this paper, observational studies do not provide timely prospective information (data) that is so needed and so hard to come by. Yet, data from observational studies may provide important supporting (or not) information regarding decisions we have already made.

As investigators, we clearly can not afford to perform prospective studies (i.e., RCT or a logical “OR” type study as in DMIST) for every diagnostic or practice related question. In order to obtain clinically valid data that leads to clinically relevant and sustainable conclusions investigators will have to increase “risk” to at least a fraction of the participants in prospective studies. As long as the added risk is relatively small compared with the potential (hypothesized) gain in clinically implementing the “new” approach being tested, it is not only appropriate but also prudent to consider performing prospective studies in support of potentially important practice recommendations.

So, where do we go from here? For cost effectiveness purposes, even if the results of our recent study are experimentally confirmed by other studies, there will still be a need to perform retrospective laboratory studies to assess the potential for changes in performance levels due to proposed changes in practices or the introduction of new technologies. These somewhat artificial laboratory studies may have to remain the primary reason for initial, albeit conditional, approval of new technologies and practices by regulatory bodies. However, until otherwise proven, a condition for the regulatory approval should be that there will be the requirement of some actual, rigorous, a-priori agreed upon, future verification of the changes in performance due to the introduction of said technology or practice into the clinical environment. Accepting that in many situations retrospective laboratory studies may remain pivotal and sufficient for regulatory approval, even if conditional, a variety of study designs can be used for this purpose and there are a number of observers’ rating paradigms (and scales) possible for specific hypothesis testing. However, in general, the relative consistency we found when using the three rating paradigms, namely binary, ROC, and FROC type responses, suggests that, in principle, under many scenarios any of the three rating approaches could be used (3). Nonetheless, the specific clinical question being investigated should drive the study design and not the availability of various analytical tools, some of which may seemingly have statistical power advantages over the others. Each of the three rating approaches we evaluated in our study, namely binary (screening BIRADS) response, ROC, and FROC, tends to address a different underlying question. Therefore, the approach that most directly addresses the actual clinical question should be the one used in these retrospective studies. As important perhaps, to ultimately have practical use in decision making, these studies should be based on testing clinically relevant differences in terms of the magnitude in performance changes.

Since prospective studies are frequently very costly in terms of both actual expenses and time to a decision point, the decision of whether or not a prospective study is needed, or when a retrospective study is not sufficient for even a conditional approval, will have to be made based on the possible/potential impact of the new technology, or practice, in terms of the magnitude of the operational changes, as well as the potential costs and benefits to the patient (or society as a whole) as compared with the cost of performing a prospective study. Guidelines may need to be developed for making these decisions in collaboration with radiologists, industry, and the regulatory bodies.

It is only natural, and in some cases quite appropriate, to explore other “more efficient” approaches and different analytical tools for the purpose of determining possible effects due to the incorporation of new technologies or practices. These may include but are not limited to data pooling, meta-analyses of a number of smaller studies, the use of simulations, or the use of different analytical approaches within a single study such as Bayesian based analysis schemes. However, the underlying issue discussed here is that we truly do not know well enough how human behavior is affected at the individual or the group level under different conditions and until we do, the results of these “more efficient” approaches that constitute primarily different types of “short cuts” to a definitive conclusion, have to be viewed with the appropriate perspective and some caution.

Since our research group is very cognizant of the possible implications resulting from the findings of our recent study (1–3), we not only continue to assess the data from this study, and others, but we are also investigating the possibility of performing a larger multi-center, yet perhaps less complicated, laboratory study to test again the primary hypothesis posed in our recently reported study. We must admit that the results of our study were a significant surprise to us, not as much in terms of the actual absolute levels of performance changes between the laboratory and the clinic but rather in terms of the substantial and statistically significant change in variability (“spread”) among radiologists’ performance levels between the clinic and the laboratory. The implications of radiologists performing more “similarly” in the clinic than in the laboratory (namely, with a lower inter observer variance component) may be substantial in terms of the relative “ease” and “cost” of testing prospectively the validity and robustness of implementing different technologies and practices in the actual clinical environment.

Acknowledgments

Funding: This work is supported in part by Grant EB003503 (to the University of Pittsburgh) from the National Institute for Biomedical Imaging and Bioengineering (NIBIB), National Institute of Health.

References

1. Gur D, Bandos AI, Cohen CS, et al. The “Laboratory” Effect: Comparing Radiologists’ Performance and Variability during Prospective Clinical and Laboratory Mammography Interpretations. *Radiology* 2008 Oct;249(1):47–53. [PubMed: 18682584]
2. Gur D, Bandos AI, King JL, et al. Binary and Multi-Category Ratings in a Laboratory Observer. Performance Study: A Comparison. *Med Phys* 2008 Oct;35(10):4404–4409. [PubMed: 18975686]
3. Gur D, Bandos AI, Klym AH, et al. Agreement of the Order of Overall Performance Levels Under Different Reading Paradigms. *Acad Radiol* 2008;15(12):1567–1573. [PubMed: 19000873]
4. Failure of extracranial-intracranial arterial bypass to reduce the risk of ischemic stroke. Results of an international randomized trial. The EC/IC Bypass Study Group. *N Engl J Med* 1985;313(19):1191–1200. [PubMed: 2865674]
5. Hillman BJ. Economic, legal, and ethical rationales for the ACRIN national lung screening trial of CT screening for lung cancer. *Acad Radiol* 2003;10:349–350. [PubMed: 12643562]
6. Church TR. National Lung Screening Trial Executive Committee. Chest radiography as the comparison for spiral CT in the National Lung Screening Trial. *Acad Radiol* 2003 Jun;10(6):713–715. [PubMed: 12809426]
7. Pisano ED, Gatsonis CA, Yaffe MJ, et al. American College of Radiology Imaging Network digital mammographic imaging screening trial: objectives and methodology. *Radiology* 2005 Aug;236(2):404–412. [PubMed: 15961755]
8. Gur D. Retrospective Analyses of Pivotal Prospective Studies With Population Segmentation—Statistically Based Inferences and Clinical Relevance. *Acad Radiol* 2008;15(11):1458–1462. [PubMed: 18995196]
9. Gur D, Sumkin JH, Rockette HE, et al. Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system. *J Natl Cancer Inst* 2004 Feb 4;96(3):185–190. [PubMed: 14759985]
10. Fenton JJ, Taplin SH, Carney PA, et al. Influence of computer-aided detection on performance of screening mammography. *N Engl J Med* 2007 Apr 5;356(14):1399–409. [PubMed: 17409321]