Published in final edited form as:

Acad Radiol. 2008 November ; 15(11): 1463–1466. doi:10.1016/j.acra.2008.07.010.

Looking back at prospective studies

Carolyn M. Rutter

Carolyn Rutter, PhD Senior Investigator Center for Health Studies Group Health Cooperative 1730 Minor Ave, Ste 1600 Seattle, WA 98101 (206) 287-2190

Gur's perspective article raises important points about analytic methods and the clinical inferences drawn from retrospective statistical analyses of prospective studies. Specifically, he associates three problems with the scientific methods of retrospective analyses: 1) using the parametric receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC) as a performance measure; 2) using Bonferroni adjustments to account for multiple comparisons; and 3) failing to evaluate the variability of results across sites and observers. Gur demonstrates these problems with a case study: a recent paper analyzing the Digital Mammography Imaging Screening Trial (DMIST).¹ The issues he raises are not specific to either retrospective study designs or to secondary exploratory analyses of large studies, but are important issues to consider in many design settings. I address each of these issues below and relate them to the information provided by DMIST papers.

1. Problems with the ROC curve

The ROC curve plots the true-positive rate (TP or sensitivity) on the y-axis against the false-positive rate (FP or 1 - specificity) on the x-axis to demonstrate the inherent tradeoff between TP and FP rates.^{2,3} TP and FP increase together as test criteria are relaxed to allow more individuals to test positive. The AUC describes the average TP over the full, but often unobserved, range of FP (i.e., 0 to 1). Because the AUC is based on the ROC over the entire FP range, a large fraction of the AUC can be based on FP regions that observed data do not support.

Many have criticized the use of the AUC as a summary measure and suggested the partial AUC (pAUC) as an alternative summary,⁴⁻⁹ as does Gur. I agree that there are problems with the AUC, but am not convinced that the pAUC offers an adequate solution for ordinal tests. The pAUC, which (when rescaled) represents the average TP over a selected range of FP values, has problems of its own. Most notably using the pAUC, requires selecting the corresponding FP range before observing the data. In addition, the pAUC does not account for differences in the observed FP range across tests being compared, leaving this criticism of AUC statistics unresolved. Because of these problems, the pAUC is not widely used for ordinal test outcomes. Instead, authors often make comparisons between tests using AUC statistics that capture the overall accuracy and supplement this overall comparison with (TP, FP) pairs that describe accuracy for selected test criteria. This was the approach taken in the DMIST paper¹. Ideally, ROC curves would be presented, along with operating points, though space limitations sometimes preclude inclusion of these graphs. In cases where clinically relevant test criteria can be defined, it may make sense to design studies using TP and FP rates as primary outcomes and AUC statistics as secondary outcomes. The BI-RADS ratings may offer such an example

if ratings are dichotomized into those that indicate the need for additional imaging or surgical work-up versus those that do not.

Gur also points out that parametric ROC curves may have a “hook”: that is, estimated TP values that are less than FP values for some FP values near 1. This is a problem with parametric modeling approaches but not for nonparametric modeling approaches. Nonparametric models for ROC curves can be used either as the primary analytic approach or to explore potential problems with the parametric ROC curves. The DMIST paper reports the results from parametric ROC curve models that account for the paired study design^{11,12} and report using nonparametric analysis^{13,14} to corroborate their findings.

Finally, Gur suggests that the DMIST authors should have focused solely on comparing sensitivities. Problems with focusing solely on sensitivity and not acknowledging the inherent tradeoff between sensitivity and specificity are well-known and well-documented in the diagnostic testing literature. This trade off is the motivation for using ROC curve analysis. When examining screening tests, is especially important to account for the false-positive rates given that the vast majority of people screened will not have the condition of interest. Post-hoc comparisons, carried out only after large differences are observed, can be misleading, because these tests are conditional on observed differences.

2. Problems with adjustments for multiple comparisons

Gur points out problems with use of the Bonferroni correction for multiple tests. To place the multiple-comparison problem in context, first consider a single statistical test. The probability of a Type I error—that is, rejecting the null hypothesis (e.g., that there is no difference between digital and film mammography) when it is true—is set to a fixed value, α , and this is referred to as the α -level of the test. Now consider m independent tests, each performed with a Type I error equal to α . In the multiple testing literature, this may be referred to as a “family” of tests. The goal of adjusting for multiple tests is to control the overall probability of rejecting at least one of the tests in the family when all of the null hypotheses are true. For these m independent tests, this family-wise error rate is $1 - (1 - \alpha)^m$. If $\alpha=0.05$ and $m=25$, then the probability of a statistically significant finding when all of the null hypotheses are true rises to 0.72. Put another way, if we conducted 25 independent significance tests at the 0.05 level, where in each case the null hypothesis was true, i.e., no differences exist, then we would expect to reject at least one of the null hypotheses simply by chance. The Bonferroni adjustment is based on the Bonferroni inequality, which tells us that *when all of the null hypotheses are true* the probability of rejecting at least one is less than or equal to $m\alpha$; and the correction is made by dividing the desired family-wise error rate by the number of tests. The Bonferroni approximation works well for families of independent statistical test. The advantage of this approach is its simplicity, and therefore, transparency. The cost is power. The test can be conservative when the tests within the family are not independent, as is the case when the outcomes included in the family of tests are correlated.

A series of articles during the 1990s debated whether to adjust for multiple comparisons.^{7, 15-23} Benjamini and Hochberg (1995) provide an excellent discussion with key references, contrasting control of family-wise error rates with control of the false-discovery rate, defined as the proportion of incorrectly rejected null hypotheses within a family of tests. Genomic research has favored control of the false-discovery rate, focusing on discovery from many simultaneous tests that are often correlated. Part of the appeal of the Bonferroni correction, particularly given debate over whether adjustment is appropriate, is the ability to recover the unadjusted results easily. For exploratory, hypothesis generating studies, it may be reasonable to consider more powerful approaches that control for the false-discover rate. .

The DMIST authors used a very conservative Bonferroni approach, counting both the original 15 comparisons and the second set of 10 comparisons, rather than considering these newly reported analyses a second family of hypothesis tests. Independence of these tests is unlikely. The first series of tests examined primary and secondary hypotheses. These were surely correlated, since they compared overlapping groups: women under 50 years old versus women 50 and older; women with heterogeneously or extremely dense breasts versus those with scattered fibroglandular densities or almost entirely fatty breast; and premenopausal or perimenopausal women versus post menopausal women. The second series of tests takes a closer look, separating women into non-overlapping groups based on age, breast density, and menopausal status. Still, the approach taken allows any reasonably informed reader to determine the unadjusted results and to use these as he or she sees fit. Gur's paper also hints at limitations to using p-values as a method for selecting important findings. P-values certainly have limits, and statisticians have long debated their use.²⁴

3. Reporting results from multisite studies

The DMIST study is a multisite trial, drawing patients from 33 North American institutions. The study enrolled 49,533 women; the number enrolled at each center varied from 197 to 3347, with a mean of about 1500.²⁵ About 14% of these women were ultimately excluded from the study, and breast cancer was diagnosed in about 0.8% of the final 42760 women during the study period. The DMIST analyses provide important information about expected differences in the performance of digital and film mammography in a population of North American women and pre-specified subgroups of women.

Gur requests information about the consistency in the study results across these institutions. I agree that analysis of between site variability can provide important information that can help pinpoint areas for quality improvement, but simple descriptions of site-specific findings are generally not informative. For example, the DMIST study had to include a large number of sites because the prevalence of breast cancer is low. Simple calculations suggest that the number of expected breast cancer cases range from one to 23 across the sites. Therefore, any focus on between-institution variability would ultimately focus on false-positive rates and thus would ignore the tradeoff between sensitivity and specificity. Gur suggests that limiting analyses to institutions or radiologists with enough cases, but this design might induce bias, because interpretive volume may be associated with performance. Evaluation of between site variability also needs to account for differences in the number of observations per site, since results from sites with fewer observations are necessarily more variable than results from sites with many observations. A hierarchical regression model could be used to examine between-institution variability in results while accounting for differences in both the number of women assessed and the characteristics of women across institutions,²⁷ and may be an interesting next analysis of these data.

Conclusions

As researchers, we have a responsibility to provide enough information to let an informed reader evaluate our work completely. Page limits can hinder our ability to disclose all details of a study fully, but space is often sufficient to include the most important points. To avoid biases, and to serve our funders and study participants, researchers are also bound to set a plan for analyses before data collection and then to carry out and report the results of these analyses at the close of the trial. Scientists may disagree about how the DMIST data were analyzed and reported, but the published reports referenced by Gur included sufficient details to allow for their critical evaluation. As readers, we are responsible for carrying out this critical evaluation—and, when moved, to write letters lobbying for a change in approach.

References

1. Pisano ED, Hendrick RE, Yaffe MJ, et al. Diagnostic accuracy of digital versus film mammography: exploratory analysis of selected population subgroups in DMIST. *Radiology* Feb;2008 246(2):376–83. [PubMed: 18227537]
2. Obuchowski NA. Receiver operating characteristic curves and their use in radiology. *Radiology* Oct; 2003 229(1):3–8. [PubMed: 14519861]
3. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;143(1):29–36. [PubMed: 7063747]
4. Baker SG. The central role of receiver operating characteristic (ROC) curves in evaluating tests for the early detection of cancer. *J Natl Cancer Inst* Apr 2;2003 95(7):511–5. [PubMed: 12671018]
5. Dodd LE, Pepe MS. Partial AUC estimation and regression. *Biometrics* Sep;2003 59(3):614–23. [PubMed: 14601762]
6. McClish DK. Analyzing a portion of the ROC curve. *Med Decis Making* Jul-Sep;1989 9(3):190–5. [PubMed: 2668680]
7. Thompson ML, Zucchini W. On the statistical analysis of ROC curves. *Stat Med* Oct;1989 8(10):1277–90. [PubMed: 2814075]
8. Jiang Y, Metz CE, Nishikawa RM. A receiver operating characteristic partial area index for highly sensitive diagnostic tests. *Radiology* Dec;1996 201(3):745–50. [PubMed: 8939225]
9. Baker SG, Pinsky P. A proposed design and analysis for comparing digital and analog mammography: Special ROC methods for cancer screening. *J Am Stat Asso* 2001;96:421–8.
11. Metz, C.; Wang, P.; Kronman, H. A new approach for testing the significance of differences between ROC curves measures from correlated data.. In: Deconinck, F., editor. *Information processing in medical imaging*. The Hague; The Netherlands, Nijhoff: 1984.
12. Zhou, XH.; Obuchowski, NA.; McClish, D. *Statistical methods in diagnostic medicine*. Wiley & Sons, Inc; New York: 2002.
13. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;44(3):837–45. [PubMed: 3203132]
14. Toledano AY, Gatsonis C. Generalized estimating equations for ordinal categorical data: arbitrary patterns of missing responses and missingness in a key covariate. *Biometrics* Jun;1999 55(2):488–96. [PubMed: 11318205]
15. Greenland S. Multiple comparisons and association selection in general epidemiology. *Int J Epidemiol* Jun;2008 37(3):430–4. [PubMed: 18453632]
16. Rothman KJ. No adjustments are needed for multiple comparisons. *Epidemiology* 1990;1(1):43–6. [PubMed: 2081237]
17. Greenland S, Robins JM. Empirical-Bayes adjustments for multiple comparisons are sometimes useful. *Epidemiology* Jul;1991 2(4):244–51. [PubMed: 1912039]
18. Poole C. Multiple comparisons? No problem! *Epidemiology* Jul;1991 2(4):241–3. [PubMed: 1912038]
19. Savitz DA, Olshan AF. Multiple comparisons and related issues in the interpretation of epidemiologic data. *Am J Epidemiol* 1995;142(9):904–8. [PubMed: 7572970]
20. Manor O, Peritz E. Re: “Multiple comparisons and related issues in the interpretation of epidemiologic data”. *Am J Epidemiol* Jan 1;1997 145(1):84–5. [PubMed: 8982029]
21. Thompson JR. Invited commentary: Re: “Multiple comparisons and related issues in the interpretation of epidemiologic data”. *Am J Epidemiol* May 1;1998 147(9):801–6. [PubMed: 9583708]
22. Goodman SN. Multiple comparisons, explained. *Am J Epidemiol* May 1;1998 147(9):807–12. [PubMed: 9583709]discussion 15
23. Savitz DA, Olshan AF. Describing data requires no adjustment for multiple comparisons: a reply from Savitz and Olshan. *Am J Epidemiol* May 1;1998 147(9):813–4. [PubMed: 9583710]discussion 5
24. Goodman S. Commentary: The P-value, devalued. *Int J Epidemiol* Oct;2003 32(5):699–702. [PubMed: 14559733]

25. Pisano ED, Gatsonis CA, Yaffe MJ, et al. American College of Radiology Imaging Network digital mammographic imaging screening trial: objectives and methodology. *Radiology* Aug;2005 236(2): 404–12. [PubMed: 15961755]
26. Pisano ED, Gatsonis C, Hendrick E, et al. Diagnostic performance of digital versus film mammography for breast-cancer screening. *N Engl J Med* Oct 27;2005 353(17):1773–83. [PubMed: 16169887]
27. Elmore JG, Miglioretti DL, Reisch LM, et al. Screening mammograms by community radiologists: variability in false-positive rates. *J Natl Cancer Inst* 2002;94(18):1373–80. [PubMed: 12237283]