

# Statistics in Epidemiology Report

Volume 2001. Number 1. Newsletter of the Statistics in Epidemiology Section of the American Statistical Association. Find Section Information on the World Wide Web at <http://www.amstat.org/sections>

---

## Receiver Operating Characteristic (ROC) Analysis

Kelly H. Zou, PhD<sup>1,2</sup> and  
Xiao-Hua (Andrew) Zhou, Ph.D<sup>3</sup>

<sup>1</sup>Department of Health Care Policy, Harvard Medical School; <sup>2</sup>Department of Radiology, Brigham and Women's Hospital, Harvard Medical School  
<sup>3</sup>Division of Biostatistics, Indiana University School of Medicine e-mail: [zou@hcp.med.harvard.edu](mailto:zou@hcp.med.harvard.edu)

### Introduction

Diagnostic testing is frequently used in clinical practice to help clinicians to select different treatment therapies. The receiver operating characteristic (ROC) curve is a useful graphical and evaluation tool for assessing the accuracy of a medical diagnostic test for detecting whether or not a patient has a disease or condition. The ROC analysis was originally developed for analyzing classification accuracy in differentiating a "signal" versus a "noise" in radar signal detection during the Second World War [1, 2].

A diagnostic test can yield one of the three outcome types. It can be a binary classification indicating that the patient is either non-diseased (here called "non-disease" for convenience, and labeled  $D=0$ ) or diseased ( $D=1$ ). It can also be an ordinal-scale; for example, it can classify each patient into one of the five ordinal-rating categories, where 1=definitely non-disease, 2=probably non-disease, 3=equivocal, 4=probably diseased, and 5=definitely diseased. Finally, it can be a continuous-scale, such as tumor volume and cancer antigen assay.

(continued on page 4)

## Causal Methods for Longitudinal Studies

Miguel A. Hernán

Department of Epidemiology  
Harvard School of Public Health  
e-mail: [miguel\\_hernan@post.harvard.edu](mailto:miguel_hernan@post.harvard.edu)

Making causal inferences about the effect of exposures on outcomes is the objective of many epidemiologists. To accomplish this goal, they design studies, collect data, and conduct data analyses. In principle, any association measure (e.g., a regression coefficient) estimated by statistical methods can be interpreted causally. The justification for interpreting causally the results of a statistical method in a particular situation lies beyond statistics. It is a matter of a priori assumptions.

As an example, consider two possible studies to estimate the overall effect of a new antiretroviral therapy on the time to AIDS among HIV-infected patients. First, an ideal experiment (large sample size, full compliance, and no loss to follow-up) in which patients were randomly assigned to either new or standard therapy and then followed for several years. Second, a large observational study in which sicker patients were more likely to receive the new antiretroviral therapy. The data from both studies is analyzed using a correctly specified Cox proportional hazards model with cumulative treatment dose as the only covariate. Most epidemiologists would agree that the hazard ratio from this model could be interpreted as the causal effect of a unit of treatment (on the

(continued on page 2)

## Causal Methods for Longitudinal Studies (Continued from page 1)

hazard ratio scale) in the first study, but not in the second one. The justifications for these causal statements are not based on statistical arguments. We say that the first hazard ratio can be interpreted causally because the treatment was randomized and randomization in large samples ensures that both groups are comparable or exchangeable, i.e., there is no confounding (Greenland and Robins 1986). We say that the second hazard ratio cannot be interpreted causally because of failure to control for joint predictors of treatment and outcome identified by a priori expert knowledge, i.e., there is confounding by, say, CD4 count and viral load (Greenland and Robins 1986).

In both cases the statistical method is the same, but the conclusions regarding causal interpretation of the hazard ratio are different because the a priori non-statistical information employed by the user of the method varies. In general, a parameter of a particular statistical model has a causal interpretation when all confounders have been measured (and appropriately controlled). Because this condition cannot be checked empirically, causal inference from observational data---and from randomized trials with small sample sizes, lack of compliance, or loss to follow-up--- is a risky task. Why then do the so-called causal methods exist? If the causal validity of statistical estimates relies so heavily on the assumption of no unmeasured confounders, then shouldn't epidemiologists concentrate on using their expert knowledge to make this assumption at least approximately true in the study design/data collection phase and then use standard statistical methods in the data analysis phase? (Standard methods are those based on stratifying or conditioning on covariate history, e.g., non parametric stratified analysis, generalized linear models, time-dependent Cox proportional hazards regression, propensity score matching.) Here comes the problem: standard statistical methods may yield invalid estimates of the overall or direct causal effect, even if the assumption of no unmeasured confounders holds true. Specifically, this occurs in longitudinal

studies when the investigator is interested in the causal effect of a time-varying treatment (e.g., antiretroviral therapy) and some of the time-dependent confounders (e.g., CD4 count and viral load) are themselves affected by prior treatment (Robins 1986). The so-called causal methods overcome this problem, and hence ensure that all confounders painstakingly identified by the epidemiologist will be appropriately controlled.

Causal methods and the definition of causal effect embedded in them derive from counterfactual theory. Briefly, we say that the binary time-varying treatment  $A(t)$  has a causal effect on subject's  $i$  response  $Y$ , measured at the end of follow-up, if the subject's outcome varies under different hypothetical treatment regimes  $a$  (e.g., treated at all times, never treated). For example, we may compare a subject's outcome when continuously treated ( $Y_{i,a1}$ ) with her outcome when never treated ( $Y_{i,a0}$ ) and say that there exists a causal effect if  $Y_{i,a1} - Y_{i,a0} \neq 0$ . Of course, the subject follows only one treatment regime and therefore one can only possibly observe either  $Y_{i,a1}$  or  $Y_{i,a0}$ . For example, if the subject remained untreated at all times, then  $Y_{i,a0} = Y$  whereas the value of  $Y_{i,a1}$  is missing. Because, in general,  $Y_{i,a1}$  and  $Y_{i,a0}$  represent situations that go against what actually happened (counter to the fact), they are known as counterfactual outcomes. In fact, the subject will generally follow a treatment regime that is neither  $Y_{i,a1}$  nor  $Y_{i,a0}$ , and hence the value of both counterfactual outcomes will be missing. This missing data problem renders it unfeasible to make causal inferences for subject-specific effects. However, under the assumption of no unmeasured confounders, it is possible to make inferences about causal effects averaged over a population of individuals, i.e.,  $E[Y_{a1}] - E[Y_{a0}]$ . For example, if  $Y$  represents five-year survival (1=death, 0=alive),  $E[Y_{a1}] - E[Y_{a0}] \neq 0$  means that the mortality risk had everybody in the population been continuously treated is different from the risk had everybody remained untreated. In other words, treatment has a causal effect on the risk of death in the population under study.

Counterfactual outcomes, also known as potential responses, were introduced in statistics by Neyman (1923), who described them in randomized experiments with time-invariant treatments. Rubin (1974) extended Neyman's theory to observational studies (see also review by Holland (1986)). Robins (1986, 1987) then developed a formal counterfactual theory that generalized the former and that applies to longitudinal studies with time-varying treatments. This is a key development because most exposures of interest to epidemiologists vary over time.

The earliest product of Robins' theory was the g-computation algorithm formula (the "g formula"), a non parametric causal method to compute overall or direct causal effects of time-varying treatment regimes under the assumption of no unmeasured confounders, even in the presence of time-dependent confounders affected by previous treatment. Specifically, the g-formula computes the expected value of the counterfactual outcomes under the treatment regimes of interest (e.g.,  $E[Y_{a1}]$  and  $E[Y_{a0}]$ ). These expected values can then be contrasted, using differences or ratios, to determine if the treatment has a population causal effect. More recently, approaches to causal inference based on causal diagrams (directed acyclic graphs) have led to a method of estimation of causal effects that is mathematically equivalent to the g-formula (Pearl 1995).

Despite its theoretical interest, the g-formula cannot be used in most practical applications because it is a completely nonparametric method, i.e., it makes no modeling assumptions. Hence using the g-formula in longitudinal studies even with a moderate number of repeated measures and covariates would require enormous amounts of data and computing time. On the other hand, fully parametric approaches may lead to bias of the treatment effect estimates. To solve this problem, Robins has developed two classes of semiparametric causal methods that incorporate modeling assumptions: marginal structural models (MSMs; Robins 1998 and 2000) and structural nested models (SNMs; Robins 1997 and 1998). The word 'structural' is commonly utilized as a synonym

for 'causal' in the social sciences. The parameters of MSMs and SNMs---estimated through inverse probability of treatment weighting and g-estimation, respectively--- can always be interpreted as the causal effect of treatment, under the assumptions of no unmeasured confounders and no model misspecification. MSMs are linear, logistic, failure time, etc. models for the counterfactual outcomes under prespecified treatment regimes. For example, a marginal structural Cox model has been used to estimate the mortality hazard ratio for being continuously versus never treated with antiretroviral therapy and prophylaxis for opportunistic infections in HIV-infected patients (Hernán 2001). SNMs are models for the counterfactual outcomes under prespecified or dynamic treatment regimes (i.e., those in which the decision to treat at a certain time depends on the subject's covariate and treatment history). For example, a structural nested failure time model has been used to estimate the causal effect of isolated systolic hypertension on cardiovascular death (Witteaman 1998). SNMs can be used for means and failure time outcomes, but are not generally useful for dichotomous outcomes.

In summary, all statistical methods used for causal inference require that the assumption of no unmeasured confounders is at least approximately true, but only causal methods, such as MSMs and SNMs, permit a valid analysis if the investigator believes that time-dependent confounders are affected by prior treatment. The choice of the analytic approach depends upon the set of assumptions that the investigator is willing to accept.

## References

1. Greenland S., Robins J.M. Identifiability, exchangeability, and epidemiological confounding. *International Journal of Epidemiology*, 15:413—419 (1986).
2. Hernán, M.A., Brumback, B., Robins, J.M. Marginal structural models to estimate the joint causal effect of non-randomized treatments. *Journal of the American Statistical Association*, in press (2001).

3. Holland, P.W. Statistics and causal inference (with discussion), *Journal of the American Statistical Association*, 81:945-970 (1986).
4. Neyman, J. On the Application of Probability Theory to Agricultural Experiments: Essay on Principles, Section 9, (1923), translated in *Statistical Science*, 5:465--480 (1990).
5. Pearl J. Causal diagrams for empirical research. *Biometrika*, 82:669--710 (1995).
6. Robins, J.M. A new approach to causal inference in mortality studies with a sustained exposure period Application to the healthy worker survivor effect [published errata appear in *Mathl Modelling*, 14, 917-21 (1987)]. *Mathematical Modelling*, 7:1393-512 (1986).
7. Robins JM. A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *Journal of Chronic Disease*, 40(suppl 2):139s--161s (1987).
8. Robins, J.M. 'Marginal structural models', *1997 Proceedings of the Section on Bayesian Statistical Science*, American Statistical Association, Alexandria, VA, 1-10 (1998).
9. Robins, J.M. Causal inference from complex longitudinal data. In: Berkane M., ed. *Latent Variable Modeling and Applications to Causality*. Lecture Notes in Statistics 120. New York: Springer-Verlag, 69-117 (1997).
10. Robins JM. Structural Nested Failure Time Models. In: Survival Analysis. Andersen PK, Keiding N, section eds. In: Armitage P, Colton C, eds. *The Encyclopedia of Biostatistics*. Chichester, UK: John Wiley and Sons, 4372-89 (1998).
11. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11:550-60 (2000).
12. Rubin, D.B. Bayesian inference for causal effects: The role of randomization, *Annals of Statistics*, 6:34-58 (1978).
13. Witteman, J.C., D'Agostino, R.B., Stijnen, T., Kannel, W.B., Cobb, J.C., de Ridder, M.A., Hofman, A., Robins, J.M. G-estimation of causal effects: isolated systolic hypertension and cardiovascular death in the

Framingham Heart Study. *American Journal of Epidemiology*, 148:390-401 (1998).

### Receiver Operating Characteristic (ROC) Analysis (Continued from page 1)

Suppose the outcome of a medical test results in a continuous-scale measurement  $T$ . Let  $t$  be a threshold (sometimes called a cutoff) value of  $T$ , used to classify patients, such that  $m$  subjects are classified as non-diseased if  $T \leq t$ , and  $n$  subjects are classified as diseased if  $T > t$ . As a result, for any set of  $N=m+n$  test results, once the true disease status (or the gold standard) for each patient is determined, independent of the test result, we have the following two-by-two contingency table at  $t$ :

	Gold	Standard	
Diagnosis	Non-diseased (D=0)	Diseased (D=1)	Total
Negative ( T+ )	True Negative's	False Negative's	$\#\{T > t\}$ (Test "+"s)
Positive ( T- )	False Positive's	True Positive's	$\#\{T \leq t\}$ (Test "-"s)
Total	$m$	$n$	$N=m+n$

The accuracy of such a binary diagnostic test is commonly assessed using the probability that the test correctly classifies a non-diseased subject as negative  $P(T- | D=0)$ , namely the true negative rate  $TNR=1-FPR$ , or specificity, and the probability that the test correctly classifies a diseased subject as positive  $P(T+ | D=1)$ , namely the true positive rate  $TPR$ , or sensitivity. When evaluating a continuous-scale diagnostic test, both  $\text{specificity}=P(T \leq t | D=0)$  and  $\text{sensitivity}=P(T > t | D=1)$  depend on the test threshold  $t$ . As  $t$  varies, so do these two quantities. By considering all possible values of the threshold values for  $t$ , a receiver operating characteristic (ROC) curve can be constructed as a plot of sensitivity (TPR) against (1-specificity) (FPR) pairs. An ROC curve, that is towards the upper left in the  $(0,1) \times (0,1)$ -space and lies above the diagonal line connecting  $(0,0)$  and  $(1,1)$ , represents higher diagnostic accuracy. If we let

$F_d$  be the distribution function of the continuous-scale  $T$  for a patient with  $D=d$ , then the ROC curve of  $T$  can be formally written  $ROC(p)=1-F_1(F_0^{-1}(1-p))$ , where  $p$  is the FPR corresponding to a cutoff point  $t$  in the domain of the distribution function  $F_0$ .

There are several other epidemiological terms frequently used in the ROC literature. The probability of disease is called the prevalence or prior probability. The ratio of TPR and FPR is the likelihood ratio positive, and the ratio of TNR and FNR is the likelihood ratio negative. The probability of disease given a positive test result is the predictive value positive, and the probability of non-disease given a negative test result is the predictive value negative.

### Characteristics and Summary Accuracy Measures

(1) Confidence intervals and bands: A vertical (or horizontal) confidence interval for TPR (or FPR) for a specified FPR (or TPR) may be constructed, often first in an unrestricted space (e.g., a probit or logit space) and then transformed back to ROC space. The reason for the transformation is to improve the performance of large-sample approximation [3, 4]. Schäfer [6] constructed vertical nonparametric confidence bounds based directly on the standard errors obtained by Greenhouse and Mantel [7] rather than using the above transformation. Confidence rectangles (and also ellipses) based on the nonparametric Greenhouse and Mantel formula are proposed and constructed instead [4]. Confidence regions based on distribution-free tolerance intervals were also proposed [8]. Simultaneous inference is desired when sensitivity is examined over a range of specificities (or vice versa). Ma and Hall [9] constructed Working-Hotelling type confidence bands for an ROC curve by mapping out Working-Hotelling bands for a regression line in probit space, using a binormal parametric model. Komogolov-Smirnov type fixed-width nonparametric bands are developed by Campbell [3]. Nonparametric regional confidence bands were constructed [10].

(2) Area and Partial Area under the Curve: The area under the curve (AUC) is a popular

summary measure of diagnostic accuracy. It ranges from 0.5 for accuracy equal to chance to 1.0 for perfect accuracy. The area may even range from 0 to 0.5 for corresponding test accuracy worse than chance [11]. The full area under the curve corresponds to the probability of a pair of independent non-diseased and diseased measurement values being in the correct order, i.e.,  $P(X<Y)$ . For continuous diagnostic data, the nonparametric estimate of AUC is the Mann-Whitney U statistic [12], namely the proportion of all possible pairs of non-diseased and diseased test subjects for which the diseased result is higher than the non-diseased one, plus half the proportion of ties. The area is a simple and convenient overall measure of diagnostic test accuracy. But it gives equal attention to the full range of TPR and FPR, whereas only limited ranges may be of practical interest. Also, areas under two ROC curves that cross provide little discriminating information. Thus, partial area is sometimes preferred. It is the area under the ROC curve between two fixed apriori values for specificities [13].

(3) Point of Intersection: Moses et al. [14] proposed a point of intersection of the ROC curve and the line on which the sum of any FPR and TPR pair is 1, i.e., a diagonal line from (1,0) to (0,1). In other words, this is the ROC point at which the sensitivity and specificity are equal. This common sensitivity value also reflects test accuracy, the higher the more accurate. For example, 1 represents a perfect test (i.e., the gold standard), and 0.5 a test with accuracy no better than flipping a coin.

(4) Optimal Intercept: Phelps and Mushlin [15] obtained an alternative utility-prior-probability based measure named optimal intercept. It is the intercept of a line tangent to the ROC curve with the optimal slope, which gives the Bayes solution to the underlying decision problem reflecting patient utilities and prior probabilities.

### Methods for Estimation

(1) Nonparametric: The crudest method for creating an ROC plot involves plotting pairs of TPR vs. FPR at all possible values for the decision threshold, where TPR and FPR are calculated using the empirical survival

distribution function for the diseased and non-diseased subjects. This method is usually referred to as the empirical or nonparametric method because no parameters are needed to model the behavior of the plot, and the unknown underlying distributions for the two groups are left unstructured [16]. This approach has the advantage of being free of structural assumptions. However, the empirical ROC curve is usually unsmooth and has a jagged form. Since the true ROC curve is a smooth function, the efficiency of the empirical ROC curve is reduced relative to a smoothed ROC curve if the smoothing is done correctly. Smooth nonparametric ROC curves were derived from estimates of density or distribution functions of the two test distributions [4, 17, 18, 19, 53]. The degree of smoothness is determined by the choice of kernel and bandwidth.

(2) Parametric: As an alternative to the nonparametric approach, a parametric model may be assumed, which has the further advantage of allowing incorporation of covariates into ROC curve fitting. Zweig and Campbell [16] discussed various possible parametric forms. For ordinal (rating) data, instead of continuous data, popular methods assume a binormal model for a latent measurement scale. Computer procedures, based on maximum likelihood estimation, were developed by Dorfman and Alf [20]; Metz and colleagues [21], and Zou and Hall [22].

(3) Semiparametric: Semiparametric models are also being considered [19, 22-23]. These models assume that there exists a monotone transformation of the measurement scale that simultaneously makes both non-diseased and diseased distributions normal. These methods are less sensitive to non-normality than the direct parametric method.

(4) Regression: Regression analysis approaches under ordinal regression methodology were developed by Torsteson and Begg [23]. Generalized linear models (GLM) methods, incorporating covariates were proposed by Pepe [24-26].

## Methods for Comparison

An important problem in ROC analysis concerns the comparison of two (or more) diagnostic tests. In a diagnostic accuracy study, if such a test is repeated under several occasions, or different tests are administered onto the same set of subjects, then the test results are typically correlated. A less common design is to enroll two different groups of subjects for these tests, resulting in two independent sets of results. In comparison, the correlated design is much efficient because it controls for subject-to-subject variation. A possible scenario would be, for example, that a test on the same set of subjects is analyzed by two similar types of laboratory instruments or by two slightly different operators, or taken at different points of time such as at the baseline or a few hours later.

Greenhouse and Mantel [7] and Linnet [27] compared the sensitivity values at a common fixed level of specificity. DeLong et al. [28] compared areas based on correlated U-statistics. Beam and Wieand [29] compared the performances of correlated tests, one of which was a discrete test and the rest continuous tests based on sensitivities at a fixed specificity that corresponds to a natural threshold of the discrete test. An approximation procedure was developed by Hanley and McNeil [30] using Pearson correlation coefficients to estimate the correlation of the two full areas. Wieand et al. [31] proposed a family of nonparametric comparisons based on a weighted average of sensitivities, in which both the area under the ROC curve and the sensitivity at any given specificity became special cases. Venkatraman and Begg [32, 33] compared two diagnostic tests using a statistical permutation test. A likelihood ratio test for testing the equivalence of correlated ROC curves by discretizing the continuous measurement scales was developed by Metz et al. [34]. The authors applied standard methods for fitting parametric bivariate binormal ROC curves to ordinal data by discretizing the continuous bivariate test data under many categories. Emil et al. [35, 36] developed nonparametric methods for comparing the sensitivities at a given specificity or the average of sensitivities over a range of specificity-values using repeated diagnostic markers. Furthermore,

generalized estimating equations (GEE) framework for repeated ordinal categorical diagnostic data was developed by Toledano and Gatsonis [37-39] using multiplicative ordinal regression models. Finally, regression analysis approaches under generalized linear models (GLM) methods for binary gold standard data were proposed by Pepe [24-26].

### **Verification Bias and Imperfect Gold Standard Bias**

To provide an unbiased estimator for the test's accuracy, we need to determine the disease status for each patient (present or absent) independent of the patient's test result. The procedure that establishes the patient's disease status is referred to as a gold standard. The gold standard could be based on surgery, autopsy, or clinical assessments. Two major problems relating to the gold standard are: (1) verification bias, and (2) imperfect gold standard bias. Verification bias occurs when only some of the patients with test results received the gold standard and the decision to verify a patient depends on the patient's test results. The bias caused by estimating the test's accuracy for those with only verified disease status is called the verification bias [40]. Imperfect gold standard bias occurs when an imperfect standard is used in the place of the gold standard in estimation of the test's accuracy.

Parametric adjustment methods for verification bias have been proposed, for example, by assuming that the verification process depends only on the diagnostic test results [40-42]. Alternatively, Zhou [43] focused on the positive and negative predictive values. A nonparametric unbiased estimate of the trapezoidal AUC was by Zhou [44]. Toledano and Gatsonis [39] developed an ordinal regression approach for estimating multiple correlated ROC curves using the GEE methodology when a key covariate is missing. Rodenberg and Zhou [45] used an EM algorithm for adjustment of ROC curve when covariates affect the verification process.

### **Additional Resources**

Designing an ROC Study. A review article on sample size calculations in ROC studies was by Obuchowski [46]. Sample size tables can be

found in Obuchowski and McClish [47] and in Obuchowski [48].

Recent Review Articles. There were 91 review articles found by using Medline as a search engine with the key phrase "ROC curve." For example, see Shapiro [49] for an overview of the ROC analysis, Beam [50] for the presence of clustered data, Zhou [51] for correction of verification bias, Hui and Zhou [54] for correction of imperfect gold standard bias, and Rockett et al [52] for ROC in meta-analysis.

Software Programs. For general analytical or sample size calculation purposes, free Fortran software programs are downloadable from the WWW address, <http://www-radiology.uchicago.edu/krl/toppage11.htm>, of Metz and colleagues of the University of Chicago.

### **References**

1. Swets JA, Pickett RM. Evaluation of diagnostic systems. New York, NY: Academic Press, 1982.
2. Swetz JA. Signal Detection Theory and ROC Analysis in Psychology and Diagnostics: Collected Papers. Mahway, NJ: Lawrence Erlbaum Associates, Publishers.
3. Campbell G. General methodology I: Advances in statistical methodology for the evaluation of diagnostic and laboratory tests. *Statistics in Medicine* 1994; 13: 499-508.
4. Zou KH, Hall WJ, Shapiro DE. Smooth nonparametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *Statistics in Medicine* 1997; 16: 2143-2156.
5. Platt RW, Hanley JA, Yang H. Bootstrap confidence intervals for the sensitivity of a quantitative diagnostic test. *Statistics in Medicine* 2000; 19: 313-322.
6. Schäfer H. Efficient confidence bounds for ROC curves. *Statistics in Medicine* 1994; 13: 1551-1561.
7. Greenhouse SW, Mantel N. The evaluation of diagnostic tests. *Biometrics* 1950; 6: 299-412.
8. Hilgers RA. Distribution-free confidence bounds for ROC curves. *Methods of Information in Medicine* 1991; 20: 96-101.

9. Ma G, Hall WJ. Confidence bands for receiver operating characteristic curves. *Medical Decision Making* 1992; 7: 149-155.
10. Jensen K, Müller H-H, Schäfer. Regional confidence bands for ROC curves. *Statistics in Medicine* 2000; 19: 493-509.
11. Hanley JA and McNeil BJ. The meaning and use of the area under a ROC curve. *Radiology* 1982; 143: 27-36.
12. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating graph. *Jour Math Psychol* 1975; 12: 387-415.
13. McClish DK. Analyzing a portion of the ROC curve. *Medical Decision Making* 1989; 9: 190-195.
14. Moses LE, Shapiro DE, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. *Statistics in Medicine* 1993; 12: 1293-1316.
15. Phelps CE, Mushlin AI. Focusing technology assessment using medical decision theory. *Medical Decision Making* 1988; 8: 279-289.
16. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry* 1993; 39: 561-577.
17. Lloyd CJ. Using smooth receiver operating characteristic curves to summarize and compare diagnostic systems. *Journal of American Statistical Association* 1998; 93: 1356-1364.
18. Lloyd CJ, Yong Z. Kernel estimators of the ROC curves are better than empirical. *Statistics & Probability Letters* 1999; 44: 221-228.
19. Hsieh F, Turnbull BW. Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *Annals of Statistics* 1996; 24: 24-40.
20. Dorfman DD, Alf E. Maximum likelihood estimation of parameters of signal detection theory—a direct solution. *Psychometrika* 1968; 33: 117-124.
21. Metz CE, Herman BA, Shen J. Maximum-likelihood estimation of receiver operating characteristic (ROC) curves from continuous distributed data. *Statistics of Medicine* 1998; 17: 1033-1053.
22. Zou KH, Hall WJ. Two transformation models for estimating an ROC curve derived from continuous data. *Journal Applied Statistics* 2000; 27: 621-631.
23. Torsteson, ANA, Begg CB. A general regression methodology for ROC curve estimation. *Medical Decision Making* 1988; 8: 204-215.
24. Pepe MS. Three approaches to regression analysis of receiver operating characteristic curves for continuous test results. *Biometrics* 1998; 54: 124-135.
25. Pepe MS. A regression modeling framework for ROC curves in medical diagnostic testing. *Biometrika* 1997; 84: 595-608.
26. Pepe MS. An interpretation for the ROC curve and inference using GLM procedures. *Biometrics* 2000; 56: 352-359.
27. Linnett K. Comparison of quantitative diagnostic tests: Type I error, power and sample size. *Statistics in Medicine* 1987; 6: 147-158.
28. 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: 837-845.
29. Beam CA, Wieand HS. A statistical method for the comparison of a discrete diagnostic test with several continuous diagnostic tests. *Biometrics* 1991; 47: 907-919.
30. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* 1983; 148, 839-843.
31. Wieand S, Gail MH, James BR, James KL. A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika* 1989; 55: 1-17.
32. Venkatraman ES, Begg CB. A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment. *Biometrika* 1996; 83: 835-848.
33. Venkatraman ES. A permutation test to compare receiver operating characteristic curves. *Biometrics* 2000; 56: 1134-1138.



34. Metz CE, Wang P-L, Kroman HB. A new approach for testing the significance of differences between ROC curves for correlated data. In: Deconick F, ed. *Information processing in medical imaging*. The Hague: Nijhoff, 1984; 432-445.
35. Emil B, Wieand S, Su JQ, Cha S. Analysis of repeated markers used to predict progression of cancer. *Statist Med* 1998; 17: 2563-2578.
36. Emil B, Wieand S, Jung S-H, Ying Z. Comparison of diagnostic markers with repeated measurements: a non-parametric ROC curve approach. 2000; 19: 511-523.
37. Toledano A, Gatsonis C. Regression analysis of correlated receiver operating characteristic data. *Acad Radiol* 1995; 2: S30-S36.
38. Toledano A, Gatsonis C. Ordinal regression methodology for ROC curves derived from correlated data. *Statist Med* 1996; 15: 1807-1826.
39. Toledano AY, Gatsonis C. Generalized estimating equations for ordinal categorical data: Arbitrary patterns of missing responses and missingness in a key covariate. *Biometrics* 1999; 55: 488-496.
40. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics* 1983; 39: 207-215.
41. Gray R, Begg CB, Greenes RA. Construction of receiver operating characteristic curves when disease verification is subject to selection bias. *Medical Decision Making* 1984; 4: 151-164.
42. Greenes RA, Begg CB. Assessment of diagnostic technologies: methodology for unbiased estimation from samples of selectively verified patients. *Investigative Radiology* 1985; 20: 751-756.
43. Zhou XH. Effect of verification bias on positive and negative predictive values. *Statistics in Medicine* 1994; 13: 1737-1745.
44. Zhou XH. A nonparametric maximum likelihood estimator for the receiver operating characteristic curve area in the presence of verification bias. *Biometrics* 1996; 52: 299-305.
45. Rodenberg CA and Zhou XH. ROC curve estimation when covariates affect the verification bias. *Biometrics* 2000; 56: 1256-1262.
46. Obuchowski NA. Sample size calculations in studies of test accuracy. *Statistical Methods in Medical Research* 1998; 7: 371-392.
47. Obuchowski NA, McClish DK. Sample size determination for diagnostic accuracy studies involving binormal ROC curve indices. *Statistics in Medicine* 1997; 16: 1529-1542.
48. Obuchowski NA. Sample size tables for receiver operating characteristic studies. *American Journal of Roentgenology* 2000; 175: 603-608.
49. Shapiro DE. The interpretation of diagnostic tests. *Statistical Methods in Medical Research* 1999; 8: 113-134.
50. Beam CA. Analysis of clustered data in receiver operating characteristic studies. *Statistical Methods in Medical Research* 1998; 7: 324-336.
51. Zhou XH. Correcting for verification bias in studies of a diagnostic test's accuracy. *Statistical Methods in Medical Research* 1998; 7: 337-353.
52. Rockett HE, Gur D, Campbell WL, Thaete FL. Use of meta-analysis in the evaluation of imaging systems. *Academic Radiology* 1994; 1: 63-69.
53. Zhou XH and Harezlak J. Comparison of bandwidth selection methods for kernel smoothing of ROC curves. Submitted.
54. Hui SL and Zhou XH. Evaluation of diagnostic tests without gold standards. *Statistical Methods in Medical Research*, 7, 354-370.

### **ASA 2001 Joint Statistical Meetings**

By Jacqueline M. Hughes-Oliver, Publications  
Officer

Start planning your schedule for the 2001 Joint Statistical Meetings, and make sure to include slots for the Section of Statistics in Epidemiology's four Invited Sessions. SIE's 2001 Program Chair, Xiao-Hua (Andrew) Zhou, has organized an exciting slate of activities.

The Invited Session "Correlated errors, Biased Instruments, and Measurement Error Correction in Nutritional Epidemiology" is organized by Donna Spiegelman (SIE 2001 Section Chair) to highlight new research results when the original error-prone measurement is validated by one or more additional instruments, some of which may be biased, have correlated systematic within-person errors, and have correlated random errors. Ross Prentice, from Fred Hutchinson Cancer Research Center, will talk about biomarkers and self-report data in the context of nutrient consumption and chronic disease associations. Raymond Carroll, from Texas A&M University, will discuss new measurement error models and the power of food frequency questionnaires. Donna Spiegelman, from Harvard School of Public Health, will present competing models for correlated data in dietary validation studies. The discussant for this session is Rudolph Kaaks, from the International Agency on Research on Cancer in Lyons, France.

Organized by Xiao-Hua (Andrew) Zhou, the Invited Session "Methods for Assessing Quality and Costs of Health Care" will highlight new developments in assessing quality and costs of our national health care. Steven Cohen, from the Agency for Healthcare Research and Quality, will discuss design and estimation innovation in the Medical Expenditure Panel Survey. S.L. Normand, from Harvard Medical School, will talk about design and analysis of health care quality studies, particularly for assessing chronic cardiovascular care in the U.S. Wanzhu Tu, from Indiana University, will present bootstrap multiple comparison methods for analysis of cost data. Xiao-Hua Zhou, from

Indiana University, will be the discussant for this session.

Olivia Carter-Pokras, from Centers for Disease Control, has organized a panel titled "Improving Data on Racial/Ethnic Groups," with distinguished panelists Raynard Kingston from Centers for Disease Control, Rose Maria Li from the National Institute on Aging, David Williams from University of Michigan, and Beatrice Rouse from SAMHSA. This invited panel session will examine useful ways of incorporating communities, especially racial and ethnic groups, into the research process to improve data on racial and ethnic groups. Panelists will describe success stories involving the communities studied, including the informed consent process, recruitment of individuals into a study, data collection, analysis, interpretation, and dissemination of findings back to the community.

The Invited Session "Myths, Lies and Statistics" will have a panel of four outstanding speakers to discuss several areas where there is often misunderstanding, including behavioral risk factors and substance use, testing and intelligence, and health insurance coverage versus health care and wealth distribution. This panel is organized by Gladys Reynolds of the Centers for Disease Control and contains panelists Jeff Cronkite from The National Institute of Standards and Technology, Gladys Reynolds, Juarlyn Gaiter from the Centers for Disease Control and Prevention, and Robert Robinson from the Centers for Disease Control and Prevention.

**Statistics in Epidemiology Section Sponsored Activities in Atlanta:**

Title	Type	Date and Time
Multiple Imputation for Missing Data	Continuing Education	Sunday, August 5 <sup>th</sup> 8:00 AM to 4:00 PM
Myths, Lies and Statistics	Invited Panel	Sunday, August 5 <sup>th</sup> 2:00 PM to 3:50 PM
Survival Analysis in Epidemiological Studies	Contributed Papers	Sunday, August 5 <sup>th</sup> 2:00 PM to 3:50 PM
Methods for Met-Analysis and Disease Mapping	Contributed Papers	Monday, August 6 <sup>th</sup> 8:30 AM to 10:20 AM
Correlated Errors, Biased Instruments and Measurement Error Correction in Nutritional Epidemiology	Invited Papers	Monday, August 6 <sup>th</sup> 10:30 AM to 12:20 PM
Bayesian Methods in Medical Studies	Contributed Papers	08/06/2001 2:00 PM to 3:50 PM
Improving Data on Racial/Ethnic Groups	Invited Panel	Tuesday, August 7 <sup>th</sup> 8:30 AM to 10:20 AM
Statistical Issues in Medical Device Clinical Studies	Topic Contributed Papers	Tuesday, August 7 <sup>th</sup> 8:30 AM to 10:20 AM
Estimating the Accuracy of Screening Tests and Prevalence Rates	Contributed Papers	Tuesday, August 7 <sup>th</sup> 8:30 AM to 10:20 AM
Methods for Assessing Quality and Costs of Health Care	Invited Papers	Wednesday, August 8 <sup>th</sup> 8:30 AM to 10:20 AM
Analysis of Nhanes and Other Large Epidemiologic Studies	Topic Contributed Papers	Wednesday, August 8 <sup>th</sup> 8:30 AM to 10:20 PM
Estimating Treatment Effects in Observational Studies	Contributed Papers	Wednesday, August 8 <sup>th</sup> 2:00 PM to 3:50 PM
<b>Section on Statistics in Epidemiology Members Meeting</b>	<b>Committee / Business</b>	<b>Wednesday, August 8<sup>th</sup> 5:30 PM to 7:00 PM</b>
Methods for Missing Data and Logistic Regression	Contributed Papers	Thursday, August 9 <sup>th</sup> 8:30 AM to 10:20 AM

## **2001 Section Officers**

### **Section Chair:**

**Donna Spiegelman**

[\(stdls@channing.harvard.edu\)](mailto:stdls@channing.harvard.edu)

Harvard School of Public Health

### **Section Chair Elect:**

**Raymond Hoffmann**

[\(hoffmann@mcw.edu\)](mailto:hoffmann@mcw.edu)

Medical College of Wisconsin

### **Past Section Chair:**

**Deborah Dawson**

[\(dvd2@po.cwru.edu\)](mailto:dvd2@po.cwru.edu)

Case Western Reserve University

### **Program Chair:**

**Xiao Hua (Andrew) Zhou**

[\(azhou@iupui.edu\)](mailto:azhou@iupui.edu)

Indiana University School of  
Medicine

### **Program Chair Elect:**

**Kung-Jong Lui**

[\(kjl@rohan.sdsu.edu\)](mailto:kjl@rohan.sdsu.edu)

San Diego State University

### **Secretary/Treasurer:**

**Maya R. Sternberg**

[\(mrs7@cdc.gov\)](mailto:mrs7@cdc.gov)

Center for Disease Control

### **Publications Officer:**

**Jacqueline M. Hughes-Oliver**

[\(hughesol@stat.ncsu.edu\)](mailto:hughesol@stat.ncsu.edu)

North Carolina State University

### **Section Representatives:**

**Ralph d'Agostino, Jr.**

[\(rdagosti@wfubmc.edu\)](mailto:rdagosti@wfubmc.edu)

Wake Forrest University

### **Alula Hadgu**

[\(axh1@cdc.gov\)](mailto:axh1@cdc.gov)

Center for Disease Control

### **Web Page Editor:**

**Ed Frome**

[\(FromeEL@ornl.gov\)](mailto:FromeEL@ornl.gov)

Oak Ridge National Laboratory

### **Newsletter Editors:**

**Raymond Hoffmann**

[\(hoffmann@mcw.edu\)](mailto:hoffmann@mcw.edu)

Medical College of Wisconsin

### **Deborah Dawson**

[\(dvd2@po.cwru.edu\)](mailto:dvd2@po.cwru.edu)

Case Western Reserve University

### **Assistant Editor:**

**Paul Hoffmann**

Marquette University

**Statistics in Epidemiology Newsletter**  
**Raymond Hoffmann, Editor**  
**c/o American Statistical Association**  
**1429 Duke Street**  
**Alexandria, VA 22314-3415**

**Non-Profit Org.**  
**U.S. Postage**  
**Paid**  
**Alexandria, Virginia**  
**Permit No. 351**