Published in final edited form as:
*Acad Radiol*. 2003 December ; 10(12): 1359–1368.

# Statistical Validation Based on Parametric Receiver Operating Characteristic Analysis of Continuous Classification Data[1]

**Kelly H. Zou, PhD**, **Simon K. Warfield, PhD**, **Julia R. Fielding, MD**, **Clare M.C. Tempany, MD**, **William M. Wells III, PhD**, **Michael R. Kaus, PhD**, **Ferenc A. Jolesz, MD**, and **Ron Kikinis, MD**

## Abstract

**Rationale and Objectives—**The accuracy of diagnostic test and imaging segmentation is important in clinical practice because it has a direct impact on therapeutic planning. Statistical validations of classification accuracy was conducted based on parametric receiver operating characteristic analysis, illustrated on three radiologic examples.

**Materials and Methods—**Two parametric models were developed for diagnostic or imaging data. *Example 1:* A semiautomated fractional segmentation algorithm was applied to magnetic resonance imaging of nine cases of brain tumors. The tumor and background pixel data were assumed to have bi-beta distributions. Fractional segmentation was validated against an estimated composite pixel-wise gold standard based on multi-reader manual segmentations. *Example 2:* The predictive value of 100 cases of spiral computed tomography of ureteral stone sizes, distributed as bi-normal after a nonlinear transformation, under two treatment options received. *Example 3:* One hundred eighty cases had prostate-specific antigen levels measured in a prospective clinical trial. Radical prostatectomy was performed in all to provide a binary gold standard of local and advanced cancer stages. Prostate-specific antigen level was transformed and modeled by bi-normal distributions. In all examples, areas under the receiver operating characteristic curves were computed.

**Results—**The areas under the receiver operating characteristic curves were: *Example 1:* Fractional segmentation of magnetic resonance imaging of brain tumors: meningiomas (0.924–0.984); astrocytomas (0.786–0.986); and other low-grade gliomas (0.896–0.983). *Example 3:* Ureteral stone size for treatment planning (0.813). *Example 2:* Prostate-specific antigen for staging prostate cancer (0.768).

**Conclusion—**All clinical examples yielded fair to excellent accuracy. The validation metric area under the receiver operating characteristic curves may be generalized to evaluating the performances of several continuous classifiers related to imaging.

### Keywords

Brain segmentation; magnetic resonance; prostate specific antigen (PSA); genitourinary system; computed tomography; receiver operating characteristic (ROC) analysis

---

The accuracy of diagnostic test and imaging segmentation is important in clinical practice because it has a direct impact on therapeutic planning. Recently, continuous classification tools

---

[1]From the Departments of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA (K.H.Z., S.K.W., C.M.C.T., W.M.W. III, F.A.J., R.K.); the Department of Health Care Policy, Harvard Medical School, 180 Longwood Ave, Boston, MA 02115 (K.H.Z.); the Department of Radiology, University of North Carolina, Chapel Hill, NC (J.R.F.); the Artificial Intelligence Laboratory, Cambridge, MA (W.M.W. III); and Philips Research Laboratories, Sector Technical Systems, Hamburg, Germany (M.R.K.).

**Address correspondence to** K.H.Z..

Supported by National Institutes of Health grant nos. R01LM7861, P41RR13218, P01CA67165, R01RR11747, R01CA86879, R21CA89449-01, U0145356, PO1CA41167, R01AG19513, and R03HS13234-01, and a research grant from the Whitaker Foundation.

are more frequently used in practice, each of which yields continuous rather than ordinal rating data, and methods for evaluation have been developed (1–4). For example, with the availability of three-dimensional (3D) imaging acquisitions and reconstructions, volumetric data are increasingly available. Another example is the use of cancer markers such as CA125, which is important for cancer detection and staging.

In contrast, traditional diagnostic tests were often based on an ordinal rating scale. For example, a five-point scale might be adopted for observer performance evaluations, where 1 = definitely normal, 2 = probably normal, 3 = probably abnormal, 4 = probably abnormal, and 5 = definitely abnormal. A discrete subjective rating method was used in a multi-modal (magnetic resonance [MR], computed tomography [CT], and ultrasound) comparative ovarian cancer technology assessment study (5,6), in one of a series of prospective multicenter Radiologic Diagnostic Oncology Group clinical trials sponsored by the funded by the National Institutes of Health in the 1990s. The advantages of the continuous diagnostic over ordinal scale are that detailed information is preserved, they are more natural with the advancements in measurement tools and computing methods, and enable more objective interpretations. Ordinal rating data will not be the focus of this article. Instead, we will evaluate the performances of continuous classifiers only.

To conduct a validation analysis based on continuous classifiers, the most important component in the validation framework is a binary gold standard, which is the classification truth for each observation in terms of two mutually exclusive classes, eg, tumor versus non-tumor, diseased versus non-diseased. For simplicity, we assume a two-class truth by labeling the control class as $C_0$ and disease class as $C_1$. A popular method for assessing the overall classification accuracy is a receiver operating characteristic (ROC) curve (7,8). It is a function of sensitivity versus (1-specificity) at all possible decision threshold. We have previously developed several methods, including nonparametric, semi-parametric, and parametric, for estimating and comparing ROC curves derived from continuous data (1–4).

In this article, we evaluate and validate the accuracy of several continuous diagnostic classifiers including semiautomated brain MRI segmentation (9), spiral CT of ureteral stone size (10), and prostate-specific antigen (PSA) for cancer staging (11). We conduct secondary analyses to validate their classification accuracy, illustrated on these examples. The connection between these three clinical examples is that the diagnostic and classification systems all generate continuous, rather than categorical, data. In addition, we develop statistical methods to estimate the unknown gold standard and to apply an appropriate transformation of non-normality data that are frequently observed.

## MATERIALS AND METHODS

### Example 1: Magnetic Resonance Imaging of Brain Tumors

**Imaging protocol—**A total of nine patients were selected from a neurosurgical database of 260 brain tumor patients, of which three had meningiomas (M), three astrocytomas (A), and three other low-grade gliomas (G) (9). The imaging protocol consisted of the following parameters: patient heads were imaged in the sagittal planes with a 1.5T MRI system (Signa, GE Medical Systems, Milwaukee, WI), with a postcontrast 3D sagittal spoiled gradient recalled acquisition with contiguous slices (flip angle, 45°); repetition time, 35 ms; echo time, 7 ms; field of view, 240 mm; slice-thickness,1.5 mm; 256 × 256 × 124 matrix).

**Fractional segmentation—**Instead of applying binary manual segmentation, Warfield et al (12) have proposed an automated segmentation algorithm that yields voxel-wise continuous probabilistic measures indicative of the tumor class. The automated fractional segmentation was applied only to a single, randomly selected two-dimensional MR image containing the

tumor. The relative signal intensity was modeled as a normal mixture of the two classes based on an initial semi-automated binary segmentation.

**Binary gold standard—**Using the same randomly selected slice, an interactive segmentation tool (MRX, GE Medical Systems, Schenectady, NY) was used and executed on an Ultra 10 Workstation (Sun Microsystems, Mountain View, CA). The structures were contoured by three independent imaging readers, blinded to the semiautomated fractional segmentation results. An anatomic object was defined by a closed contour, and the computer program labeled every voxel of the enclosed volume. Of the gold standard, the background and brain tumor pixels defined the $C_0$ and $C_1$ classes, respectively.

### Example 2. Spiral Computed Tomography of Uretral Stones

**Imaging protocol—**A total of 100 unenhanced spiral CT scans were obtained to evaluate flank pain in patients with obstructing ureteral stones documented by means of chart review (10). A standard protocol was used (280 mA; 12 kVp; pitch, 1.0–1.6). The imaging thickness was 5 mm, with images reconstructed at 5-mm increments.

**Ureteral stone size—**Two radiologists initially reviewed the CT scans independently and blindly to derive several imaging features, one of which was the size of the ureteral stone measured in millimeter, using a caliper on CT images. This size variable was treated as the classifier to predict treatment options.

**Binary gold standard—**The actual treatment received by each subject was considered as the gold standard, either spontaneous passage (class $C_0$) or surgical intervention (class $C_1$).

### Example 3. Prostate Specific Antigen for Prostate Cancer Staging

**Prostate cancer biopsy and PSA—**In a subset of patients enrolled in a multicenter prospective Radiologic Diagnostic Oncology Prostate Cancer Staging Clinical Trial (11), magnetic resonance imaging was performed in 213 patients with prostate cancer after excluding the missing data in the baseline biopsy, 180 cases were included here. Results of the PSA were treated as the main diagnostic variable.

**Binary gold standard—**Radical prostatectomy was performed in all patients to provide the gold standard, which was based on histopathology to separate patients further into local (periprostatic invasion of tumor and spread of diseased to the seminal vesicles and lymph nodes) versus advanced (stages A and B) disease. Of the gold standard, the local and advanced stages were regarded as classes $C_0$ and $C_1$, respectively.

### Statistical Methods

**Mixture modeling of distributions—**Depending on the types of data, different types of mixture distributions for the two populations were chosen, we now focus on two parametric models, named the bi-beta and bi-normal models. In Example 1 on brain tumor segmentations, because the semi-automated fractional segmentation yielded pixel-wise probability of the tumor class, the range of the data is restricted to [0, 1]. A convenient model for such probabilistic data is a mixture of two beta distributions, here called the bi-beta model. Characteristics of the beta distribution are found in classical literature on statistical distribution theory (13). Specifically, the distribution of the fractional segmentation in class $C_0$ was assumed to be $F(x) \sim Beta(\alpha_0,\beta_0)$, while the distribution of the fractional segmentation probabilities in class $C_1$ was assumed to be $G(y) \sim Beta(\alpha_1, \beta_1)$. The ROC parameters were the four shape parameters in these beta distributions. For simplicity, we assumed pixel-wise

independence, and more complicated models that incorporate spatial homogeneity will be developed for validation in the future.

In Examples 2 and 3, the diagnostic classifiers were PSA and ureteral stone size (in mm), respectively, thus both outcome variables take on positive values rather than probabilistic values as in Example 1. We applied a nonlinear normality Box-Cox transformation algorithm (see Appendix A.1) (2–4,14). After such a parametric transformation, a mixture of two normal distributions can be assumed, known as the bi-normal model in the literature (1,15,16). We assume here that the measurements in class $C_0$ have a standard normal distribution, while the measurements $C_1$ of the diseased class have a normal distributions. The bi-normal model in the literature is a slight variation in terms of parametrization of these two ROC parameters.

**Composite gold standard estimate from multiple manual segmentations—**In Example 1, for the purpose of validation, it is necessary to derive a composite binary gold standard by combining multiple manual segmentations by three independent image readers. We have applied our recently developed, Simultaneous Truth and Performance Level Estimation (STAPLE) program (outlined in Appendix A.2.) (17–19), an automated expectation-maximization (EM) algorithm (20) for estimating the composite gold standard. For each pixel, a maximum likelihood estimate of the composite gold standard of tumor or background class was optimally determined over all image readers' results. This algorithm has a major advantage over *ad hoc* combination methods such as a pixel-wise or voxel-wise voting scheme across all readers. Instead, it considers a higher weight for readers with estimated better quality in segmentation (17,18).

**Normality transformation—**As mentioned in both Examples 2 and 3, the classification data may not be suitable for the bi-normal model without any transformation. We applied a nonlinear Box-Cox (14) normality transformation to the continuous and positively valued classification data (see Appendix A.1.).

**Receiver operating characteristic analysis—**Statistical validations were carried out using ROC analysis. In the radiology literature, Metz et al (1) have provided a maximum-likelihood solution for estimating ROC parameters, with algorithms implemented with software available from the Kurt Rossmann Laboratories for Radiologic Image Research, Department of Radiology, at the University of Chicago (Chicago, IL). Conventionally, and as in these popular programs, a bi-normal model is often used, with two ROC parameters, a standardized difference in the means, and the ratio of the standard deviations of the distributions of the diseased and control classes. These are simple functions of the two sets of (mean, standard deviation) parameters of these classes. As an extension, we specified the four shape parameters in the bi-beta model as the ROC parameters.

Once these parameters are estimated by maximizing the likelihood function, at a pre-specified cutoff threshold value, the true positive fraction (or sensitivity) and the true negative fraction (or specificity) = 1 − (false positive fraction) are estimated.

An ROC curve consists of all (1-specificity, sensitivity)-pair by varying all possible threshold values (see notations and assumptions for ROC analysis given in Appendix A.3). In Example 1, stratified ROC analyses were performed in each tumor case and type against the estimated composite voxel-wise gold standard using the STAPLE program (17–19) based on an EM algorithm (20). Over all thresholds $\gamma$ ($\gamma \subseteq [0,1]$), the four bi-beta ROC parameters were estimated via matching moments (see details in Appendix A.4).

In Examples 2 and 3, after estimated Box-Cox transformations, the bi-normal ROC parameters were estimated by maximizing the likelihood function over all possible thresholds $\gamma \subseteq \Re$ (see details in Appendix A.5).

An invariance property of the ROC curve states that with any monotone transformation such as the nonlinear Box-Cox transformation considered here, the underlying ROC curve, as well as the resulting area under the curve (AUC), remain the same (3). However, the parametric assumptions may not be appropriate and the estimated AUC biased if, for example, skewed data are fitted with a bi-normal model, therefore leading to unsatisfactory goodness-of-fit (16).

### Numerical Simulation of the Receiver Operating Characteristic Parameters and the Resulting Area Under the Curve

**Bi-beta model—**We assessed the relationship between possible underlying parameters and the resulting ROC curves via either numerical integration or exact computations. We considered hypothetical scenarios of the mixture distributions having equal variances or unequal variances in the resulting samples in the two classes. As the mean of a two-parameter Beta($\alpha$,$\beta$) distribution is $\alpha/(\alpha + \beta)$ and the variance is $\alpha\beta/\{(\alpha + \beta)^2(\alpha + \beta+1)\}$, we set one of the two shape parameters of the distribution as 1 for simplicity. By varying the value of the other parameter, we achieved different underlying mixtures. The ROC parameters considered were as follows:

$(\alpha_0, \beta_0, \alpha_1, \beta_1)$ = {(1,1,1,1);(1,1.5,1.5,1);(1,2,2,1); (1,2.5,2.5,1);(1,3,3,1); (1,9,9,1)} with equal variances, and = {(1,2,1.5,1); (1,1.5,3,1); (1,9,3,1); (1,2,9,1)} with unequal variances. In this model, the population means had various combinations from 10%–90% in fractions. Similarly, the population variances were flexibly varied.

**Bi-normal model—**In this model, the distribution under class $C_0$ is standard normal, while the population means of class $C_1$ was expressed as $\alpha$. By considering small population standard deviations close to 1, the location-shift in these the distribution of $C_1$ ranged from 0.25–2.5. Similar simulation studies for the bi-normal model have been carried out previously (3). The parameters considered were ($\alpha$,$\beta$) = {(0.25,1); (0.5,1); (0.75,1); (1,1); (1.25,1); (1.5,1); (1.75,1); (2,1); (2.25,1); (2.5,1)} with equal variances, and = {(0.25,1.5); (0.5,1.5); (0.75,1.5); (1,1.5); (1.25,1.5); (1.5,1.5); (1.75,1.5); (2,1.5); (2.25,1.5); (2.5,1.5)} with unequal variances.

## RESULTS

### Example 1: Magnetic Resonance Imaging of Brain Tumors

**Estimated Binary Gold Standard—**In Figure 1, we present the manual segmentation results of a meningioma case according to an index summary of all three imaging readers' performances (Fig 1, left panel). The estimated composite gold standard is also provided (Fig 1, right panel).

**Receiver operating characteristic analysis—**Table 1 presents the estimated bi-beta ROC parameters for all cases, with the corresponding resulting AUCs for these cases. High accuracy values were achieved using the semi-automated segmentation algorithm for all cases. The AUCs for the meningiomas were 0.924, 0.968, and 0.984 for meningiomas, 0.786, 0.926, and 0.986 for astrocytomas, and 0.896, 0.916, and 0.983 for other mixed low-grade gliomas. The ROC curves in Figure 2 showed case-to-case variations of segmentation accuracies, which were the smallest for meningiomas but were the largest for astrocytomas.

### Example 2. Spiral Computed Tomography of Ureteral Stones

**Box-Cox transformation—**The estimated Box-Cox transformation coefficient was $\widehat{\lambda} = 0.11$. The test of normality showed that the *P* values were .05 and <.001 before this transformation, in comparison to .92 and .94 after the transformation, for the distributions of the ureteral stone sizes in the spontaneous passage and the interventional samples, respectively.

**Receiver operating characteristic analysis—**The estimated ROC parameters were $(\widehat{\alpha}, \widehat{\beta}) = (1.17, 0.85)$, with a resulting AUC of 0.813, which was fairly high. See Figure 3 for both the unsmooth nonparametric empirical and the smooth parametric bi-normal ROC curves. The parametric curve followed closely to the empirical curve.

### Example 3. Prostate Specific Antigen for Prostate Cancer Staging

**Box-Cox transformation—**The estimated Box-Cox transformation coefficient $\widehat{\lambda} = 0.33$. The test of normality showed that the *P* values were .003 and <.001 before this transformation, in comparison to .06 and .10 after the transformation, for the PSA values in the local and the advanced prostate cancer stage samples, respectively.

**Receiver operating characteristic analysis—**The estimated ROC parameters were $(\widehat{\alpha}, \widehat{\beta}) = (1.20, 1.30)$, with an AUC of 0.768, suggesting fair to moderate accuracy. See Figure 4 for the ROC curve.

### Numerical Simulation of the Receiver Operating Characteristic Parameters and the Resulting Area Under the Curve

**Bi-beta model—**Table 2 presents the corresponding AUC for the specified bi-beta model parameters.

**Bi-normal model—**Table 3 presents the corresponding AUC for the specified bi-beta model parameters.

## DISCUSSION

In this article, we have presented both parametric bi-beta and bi-normal models for validating continuous diagnostic or imaging classification results. In the first example, we focus on the performance of a semi-automated fractional segmentation, which gave probabilistic interpretation of the presence of tumor in all pixels in an MR image. The hidden pixel-wise gold standard was estimated using our recently developed EM-algorithm, STAPLE. In the next two examples, we examined cancer marker PSA and CT of ureteral stone size data derived from individual patients. A nonlinear transformation model was applied to these data, enabling a bi-normal parametric model.

To interpret the AUC values, an area of 1 represents a perfect classifier, and 0.5 represents a classifier that has the same accuracy as flipping an unbiased coin. Subjectively, a rough guide is that fair-to-excellent accuracy is achieved when AUC ≥0.7. In a classical article on AUC by Hanley and McNeil (34), the authors gave the sample size necessary in a table for testing various differences between the two correlated AUCs, when one is minimally 0.7.

In all of these clinical examples, we observed satisfactory accuracy, as evidenced by fairly to high AUC values. In Example 1, the estimated AUCs were between 0.786 and 0.986, although the ROC curves in Figure 2 were variable. Thus, the semi-automated segmentation algorithm was moderately to highly accurate, as compared with the composite gold standard derived from

three manual segmentations using STAPLE. However, the accuracy was rather case-dependent. The AUCs were reasonably high, at 0.813 and 0.768, for ureteral stone size and for PSA, respectively.

There are several advantages of our proposed ROC validation methodology. The unknown gold standard can be estimated via the STAPLE algorithm. The bi-beta and bi-normal parametric modeling are quite flexible for modeling the distributions of fractional data and positive-valued diagnostic data, respectively. The means and variances of the underlying mixture distributions may be characterized by simple functions of the parameters in these models. In particular, under the bi-normal model, a goodness-of-fit test of normality to assess modeling fitting was conducted using the z-test, with *P* values reported. These methods are natural extensions of existing methods, such as the bi-logistic or bi-gamma models, found in the literature (21,22).

As a limitation, in the first clinical example, only three radiologist segmentations were used in the brain tumor example. We have previously conducted a digital phantom experiment consisting of one set covering approximately 11% of a 256 × 256 pixel image (18). We assumed three segmenters, one yielding results equal to the ground truth, and one set equal to the ground truth shifted left 10 columns, and set equal to the ground truth shifted right 10 columns. After 11 iterations, STAPLE discovered that one of the segmenters' result was identical to the ground truth, and the remaining two were slightly incorrect. A comparison between the STAPLE result and other measures, such as those based on a voting rule or using the median of the segmenters, will be investigated in the future.

However, all parametric models should be used with caution. If parametric assumptions are not met, then violations of the corresponding assumptions could lead to poor fitting and inferences. Alternative methods such as nonparametric smoothing methods may be considered, which may be much more computer intensive and less straightforward (23,24).

Statistical validation of classification accuracy may be conducted using several other metrics (16,25), particularly when spatial information is also important in dealing with 2-dimensional imaging pixel or 3D voxel data. For comparing two sets of segmentation results, existing validation metrics other than area under the ROC curve, for example, entropy-based mutual information (26), Jaccard (27) and Dice (28) similarity coefficient, and Hausdorff distance measure (29,30). We have already investigated such metrics in separate articles (19,31).

In summary, we have conducted parametric evaluations of two types of continuous classification data using ROC analysis, with application to three clinical examples. The proposed method may be adapted to several validation tasks in radiologic research, as illustrated in our clinical examples.

**Acknowledgements**

The authors thank the three experts who performed manual segmentations of the brain tumor cases.

## References

1. Metz CE, Herman BA, Shen JH. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously distributed data. Stat Med 1998;17:1033–1053. [PubMed: 9612889]
2. Zou KH, Tempany CM, Fielding JR, Silverman SG. Original smooth receiver operating characteristic curve estimation from continuous data: statistical methods for analyzing the predictive value of spiral CT of ureteral stones. Acad Radiol 1998;5:680–687. [PubMed: 9787838]
3. Zou KH, Hall WJ. Two transformation models for estimating an ROC curve derived from continuous data. J Appl Stat 2000;27:621–631.

4. O'Malley AJ, Zou KH, Fielding JR, Tempany CM. Bayesian regression methodology for estimating a receiver operating characteristic curve with two radiologic applications: prostate biopsy and spiral CT of ureteral stones. Acad Radiol 2001;8:713–725. [PubMed: 11508750]
5. Kurtz AB, Tsimikas JV, Tempany CM, et al. Diagnosis and staging of ovarian cancer: comparative values of Doppler and conventional ultrasound, CT, and MR imaging correlated with surgery and histopathologic analysis–report of the Radiology Diagnostic Oncology Group. Radiology 1999;212:19–27. [PubMed: 10405715]
6. Tempany CM, Zou KH, Silverman SG, Brown DL, Kurtz AB, McNeil BJ. Staging of advanced ovarian cancer: comparison of imaging modalities: report from the Radiological Diagnostic Oncology Group. Radiology 2000;215:761–767. [PubMed: 10831697]
7. Shapiro DE. The interpretation of diagnostic tests. Stat Methods Med Res 1999;8:113–134. [PubMed: 10501649]
8. Zhou XH, McClish DK, Obuchowski NA. Statistical methods in diagnostic medicine. New York, NY: John Wiley & Sons, 2002
9. Kaus MR, Warfield SK, Nabavi A, Black PM, Jolesz FA, Kikinis R. Automated segmentation of MR images of brain tumors. Radiology 2001;218:586–591. [PubMed: 11161183]
10. Fielding JR, Silverman SG, Samuel S, Zou KH, Loughlin KR. Unenhanced helical CT of ureteral stones: a replacement for excretory urography in planning treatment. AJR Am J Roentgenol 1998;171:1051–1053. [PubMed: 9762995]
11. Tempany CM, Zhou X, Zerhouni EA, et al. Staging of prostate cancer: results of Radiology Diagnostic Oncology Group project comparison of three MR imaging techniques. Radiology 1994;192:47–54. [PubMed: 8208963]
12. Warfield SK, Westin CF, Guttmann CRG, Albert M, Jolesz FA, Kikinis R. Fractional segmentation of white matter. In: Proceedings of Second Internationa Conference on Medical Imaging Computing and Computer Assisted Interventions, Cambridge, UK, September 19–22. New York Springer, 1999:62–71.
13. Johnson NL, Kotz SI, Balakrishnan N. Beta distributions. In: Continuous univariate distributions. 2nd Ed. John Wiley and Sons, 1999, 221–235.
14. Box GEP, Cox DR. An analysis of transformations. J Royal Stat Soc (Ser B) 1964;42:71–78.
15. Dorfman DD, Berbaum KS, Metz CE, Lenth RV, Hanley JA, Abu Dagga H. Proper receiver operating characteristic analysis: the bigamma model. Acad Radiol 1997;4:138–149. [PubMed: 9061087]
16. Zou KH, Gastwirth JL, McNeil BJ. A goodness-of-fit test for a receiver operating characteristic curve from continuous diagnostic test data. In: Crossing boundaries: statistics essays in honor of Jack Hall. Lecture Notes in Mathematical Science–Monograph Series, 43. Beechwood, OH: Institute of Mathematical Statistics; 59–68.
17. Warfield SK, Zou KH, Kaus MR, Wells M III. Simultaneous validation of image segmentation and assessment of expert quality. In: Proceeding of IEEE Symposium in Biomedical Imaging, July 7–10 Conference. New York: IEEE, 2002: 1–4.
18. Warfield SK, Zou KH, Wells M III. Validation of image segmentation and expert quality with an expectation-maximization algorithm. In: Proceedings of Fifth International Conference on Medical Imaging Computing and Computer Assisted Interventions, Tokyo, Japan, September 19–22. Berlin: Springer, 2002:298–306.
19. Zou KH, Wells WM, III, Kikinis R, Warfield SK. Three validation metrics for automated probabilistic image segmentation of brain tumors. Stat Med (in press).
20. Dempster AP, Laird NM, Rubin DB. Maximum-likelihood from incomplete data via the EM algorithm. J Royal Stat Soc (Ser B) 1977;39:34–37.
21. Dorfman DD, Berbaum KS, Metz CE. Maximum likelihood estimation of parameters of signal detection theory: a direct solution. Psychometrika 1968;33:117–124. [PubMed: 5239566]
22. Zweig MH, Campbell G. Receiver operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 1993;39:561–577. [PubMed: 8472349]
23. Zou KH, Hall WJ, Shapiro DE. Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. Stat Med 1997;16:2143–2156. [PubMed: 9330425]
24. Zhou XH, Harezlak J. Comparison of bandwidth selection methods for kernel smoothing of ROC curves. Stat Med 2002;21:2045–2055. [PubMed: 12111886]

25. Zijdenbos AP, Dawant BM, Margolin RA, Palmer AC. Morphometric analysis of white matter lesions in MR images: method and validation. IEEE Trans Med Imaging 1994;13:716–724. [PubMed: 18218550]
26. Cover TM, Thomas JA. Elements of information theory. New York: JohnWiley & Sons, 1991.
27. Jaccard P. The distribution of flora in the alpine zone. New Phytologist 1912;11:37–50.
28. Dice LR. Measures of the amount of ecologic association between species. Ecology 1945;26:297–302.
29. Gerig G, Jomier M, Chakos M. Valmet: a new validation tool for assessing and improving 3D object segmentation. In: Proceedings of Fourth International Conference on Medical Imaging Computing and Computer Assisted Interventions, Urecht, The Netherlands, October 14–17. Heidelberg: Springer, 2001: 561–523.
30. Huttenlocher DP, Klauderman GA, Rucklidge WJ. Comparing images using the Hausdorff-distance. Pattern Analysis and Machine Intelligence 1993;15:850–863.
31. Zou KH, Wells M III, Kaus MR, Kikinis R, Jolesz FA, Warfield SK. Statistical validation of automated probabilistic segmentation against composite latent expert ground truth in MR imaging of brain tumors. In: Proceedings of Fifth International Conference on Medical Imaging Computing and Computer Assisted Interventions, Tokyo, Japan, September 25–28. Berlin: Springer, 2002:315–322.
32. Lin CC, Mudholkar GS. A simple test for normality against asymmetric alternatives. Biometrika 1980;67:455–461.
33. Hernandez F, Johnson RA. The large-sample behavior of transformations to normality. J Am Stat Assoc 1980;75:855–861.
34. Hanley JA, McNeil BJ. The meaning and use of the area under an ROC curve. Radiology 1982;143:27–36.
35. McClish DK. Analyzing a portion of the ROC curve. Med Decis Making 1989;9:190–195. [PubMed: 2668680]

## A.1. A Box-Cox Normality Transformation

The Box-Cox transformation (14), from a positive-valued measurement X to a real-valued measurement X′, has the form $X' = (X^{\lambda} - 1).\lambda$ when $\lambda \neq 0$, and = $log(X)$ when $\lambda = 0$. The transformation parameter, λ, is estimated by a maximum likelihood method via nonlinear optimization (2). Subsequently, a test of normality may be performed using the z-test (32) under each binary gold standard class to ensure the bi-normal assumption.

Here we now include an S-Plus (http://www.insightful.com) function to estimate the optimal Box-Cox transformation coefficient (l in the following program codes) after entering the x- and y- sample data under classes $C_0$ and $C_1$, respectively.

It can be shown that we may maximize the profile log likelihood, a function of λ, conditioned on other ROC parameters (33):

$$\arg\_\max l(\lambda | x_1, \cdots, x_m, y_1, \cdots, y_n) = \arg\_\max \left[ -m log(s_{x'}) - n log(s_{y'}) + (\lambda - 1) \left\{ \sum_{i=1}^{m} log(x'_i) + \sum_{j=1}^{n} log(y'_i) \right\} + c \right] \quad (1)$$

where c is a constant free of the parameters. Because a nonlinear optimization routine is used, we include the S-Plus codes using a built-in a nonlinear minimization function:

```
bclik<−function(lambda){

m <− length(x); n<− length(y); xlog<− log(x); ylog<− log(y)

if (l = = 0){xnew<-xlog;ynew<-ylog}; else {xnew<− (x^l−1)/l; ynew<− (y^l−1)/l}
```

m×log(stdev(xnew))+n×log(stdev(ynew))-l-1)×(sum(xlog)+sum(ylog))}nlmin(bclik,0)

## A.2. Expectation-Maximization Algorithm for Estimating a Composite Gold Standard

Denote the hidden voxel-wise gold standard by T, and a set of manual segmentations by r = 1, …,R segmenters. For simplicity, assume that their conditionally independent manual segmentations on the same image with a total of l = 1,…,N voxels, yielded binary decisions (eg, tumor versus background; diseased versus non-diseased) $B_{lr}$.

Characterize the r-th segmenter's performance quality by his true specificity ($Q_{0r}$) and and sensitivity ($Q_{1r}$), respectively. The conditional independence assumes that $(B_{lr}|T_l,Q_{0r},Q_{1r})\perp(B_{lr'}|T_l,Q_{0r'},Q_{1r})$ for any pair of segmenters, $r \neq r'$.

To estimate all of the $T_l$'s, the maximum likelihood estimate is $\widehat{T} = \arg\max_{T} P(B|T, Q_0, Q_1)$, for all voxels. However, segmentor-specific quality, $Q_{0r}$ and $Q_{1r}$are unknown. We have developed a software (STAPLE) to iteratively estimate the voxel-wise gold standard using an EM-algorithm (17–20). This algorithm is briefly outlined as follows, with k = 1,…K iterations till convergence:

### The expectation (E) step

In the (k−1)-th iteration, let

$$u^{(k-1)} = \prod_{r:B_{lr}=1} \widehat{Q}_{1r}^{(k-1)} \prod_{r:B_{lr}=0} (1-\widehat{Q}_{1r}^{(k-1)}) \text{ŁŁand } v^{(k-1)} = \prod_{r:B_{lr}=0} \widehat{Q}_{0r}^{(k-1)} \prod_{r:B_{lr}=1} (1-\widehat{Q}_{0r}^{(k-1)}) \quad (2)$$

Define a weight variable, a common notation in an EM-algorithm, for iteration (k−1):

$$w_l^{(k-1)} = f(T_l = 1|B_l, \widehat{Q}_{0r}^{(k-1)}, \widehat{Q}_{1r}^{(k-1)}) = \frac{u^{(k-1)}\overline{\pi}}{u^{(k-1)}\overline{\pi} + v^{(k-1)}\pi}, \quad (3)$$

where $\pi = P(T=0)$ and $\overline{\pi} = 1-\pi = P(T=1)$. At k = 0, the initial segmentation quality parameters may be estimated using the median rating or a voting rule as the initial gold standard over R segmenters.

### The maximization (M) step

At the k-th iterative step, we maximize the log-likelihood, such that

$$(\widehat{Q}_0^{(k)}, \widehat{Q}_1^{(k)}) = arg \text{ŁŁmax}_{(Q_0,Q_1)} \times \text{ŁŁ}E\left[\log\{f(B|T, Q_0, Q_1)f(T)\} \middle| f(T|B, \widehat{Q}_0^{(k-1)}, \widehat{Q}_1^{(k-1)})\right] \quad (4)$$

and that for the r-th segmenter,

$$(\widehat{Q}_{0r}^{(k)}, \widehat{Q}_{1r}^{(k)}) = \arg\max_{(Q_0,Q_1)} \text{ŁŁ} \sum_{l:B_{1r}=0} \left\{ \overline{w}_l^{(k-1)} \log Q_{0r} + \text{ŁŁ} w_l^{(k-1)} \log(1-Q_{1r}) \right\} + \sum_{l:B_{1r}=1} \left\{ \overline{w}_l^{(k-1)} (1-\log Q_{0r}) + \text{ŁŁ} w_l^{(k-1)} \log Q_{1r} \right\} \quad (5)$$

where $w_l^{(k-1)}$ is the weight variable defined earlier for iteration (k−1), and $\overline{w}_l^{(k-1)} = 1 - w_l^{(k-1)}$.

The maximum likelihood estimates of the segmentation quality parameters are:

$$Q_{0r}^{(k)} = \frac{\sum_{l:B_{1r}=0} \overline{w}_l^{(k-1)}}{\sum_{l:B_{1r}=0} \overline{w}_l^{(k-1)} + \sum_{l:B_{1r}=1} \overline{w}_l^{(k-1)}} \text{ŁŁand } Q_{1r}^{(k)} = \frac{\sum_{l:B_{1r}=1} w_l^{(k-1)}}{\sum_{l:B_{1r}=1} w_l^{(k-1)} + \sum_{l:B_{1r}=0} w_l^{(k-1)}}. \quad (6)$$

In our experience, typically only K <20 iterations were required to achieve convergence in several segmentation applications.

## A.3. Mixture Modeling in Receiver Operating Characteristic Analysis

For simplicity, independence is assumed in space or over all individuals, both for classes $C_0$ and $C_1$. We label the measurements of $C_0$ as $X_i$(I = 1,…,m individuals or pixels), and the measurements of $C_1$ as $Y_j$ (j = 1,…,n individuals or pixels).

At each possible threshold γ, the underlying cumulative distributional function (c.d.f.) under class $C_0$ is F(γ), with a survival function, and a probability density function (p.d.f.) f(γ). Similarly, at the same threshold γ, the underlying c.d.f. under class $C_1$ is G(γ), with a survival function, a p.d.f. g(γ).

Each point along the ROC curve, given γ, is the false positive rate $(\text{FPR}) = 1 - \text{specificity} = \overline{F}(\gamma)$ and true positive rate $(\text{TPR}) = \text{sensitivity} = \overline{G}(\gamma)$. When the threshold takes on all possible values (ie, $\gamma \subseteq [0, 1]$ for fractional data and $\gamma \subseteq \Re$ for real-valued data), an ROC curve is formed in the space [0,1] × [0,1].

## A.4. A Bi-Beta Receiver Operating Characteristic Curve

Assume $F(x) \sim \text{Beta}(\alpha_0,\beta_0)$ and $G(y) \sim \text{Beta}(\alpha_1,\beta_1)$. The estimates of the four beta shape parameters are obtained by matching the mean and variances of the beta distributions with what are found in each sample, separately from the nondiseased and from the diseased sample data (I do not understand how exactly this is done). This moment approach is much simpler than solving iteratively for the maximum likelihood estimates, and works well for a large number of pixel data typically encountered in image processing, which we will present with simulations in a separate article.

From the sample data in class $C_0$, let the mean of be and the standard deviation be $s_x$; similarly, from the sample data in class $C_1$, let the mean and standard deviations be &*y*macr; and the standard deviation be $s_y$, respectively, then the estimates of the parameters in the bi-beta model

are: $\widehat{\alpha_0} = \bar{x}\left\{\bar{x}(1-\bar{x})s_x^2 - 1\right\}, \widehat{\beta_0} = (1-\bar{x})\left\{\bar{x}(1-\bar{x})s_x^2 - 1\right\}$; similarly, $\widehat{\alpha_1} = \bar{y}\left\{\bar{y}(1-\bar{y})s_y^2 - 1\right\}$ and $\widehat{\beta_1} = (1-\bar{y})\left\{\bar{y}(1-\bar{y})s_y^2 - 1\right\}$.

The definition of the area under the ROC curve (AUC) (34) is $AUC = P(X < Y) = \int_\gamma \bar{G}(\gamma) d\bar{F}(\gamma)$, which is approximated by numerical integration using S-Plus or other software packages under the bi-beta model.

## A.5. A Bi-Normal Receiver Operating Characteristic Curve

Based on a bi-normal model for Examples 2 and 3, after the Box-Cox transformation, assume $F(x) \sim N(0,1)$ and $G(y) \sim N(\alpha,\beta)$, the maximum likelihood estimates of the parameters may also be derived from the sample means and standard deviations, $\bar{x}, s_x, \bar{y}$, and $s_y$. The estimated parameters are: $\widehat{\alpha} = (\bar{y}-\bar{x})/s_x$ and $\widehat{\beta} = s_y/s_x$.

In the bi-normal model, the AUC is an explicit function of the ROC parameters $AUC = \Phi\left(\frac{\alpha}{\sqrt{1+\beta^2}}\right)$, where Φ is the cumulative probability function of a standard normal distribution (35). The standard error for making inferences based on AUC may be found in the literature (2).

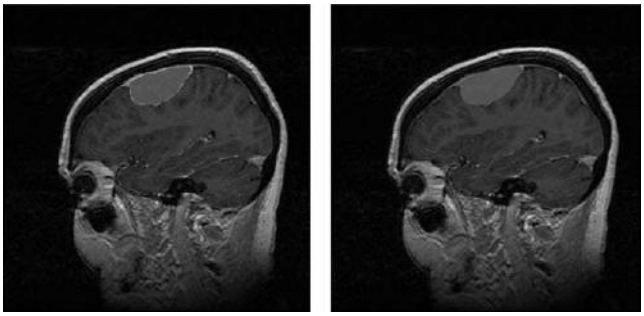

**Figure 1.**
Three segmenters' manual segmentation results (left) and the estimated composite binary pixel-wise ground truth (right) for Example 1 on MRI brain segmentations of meningiomas, astrocytomas, and gliomas.

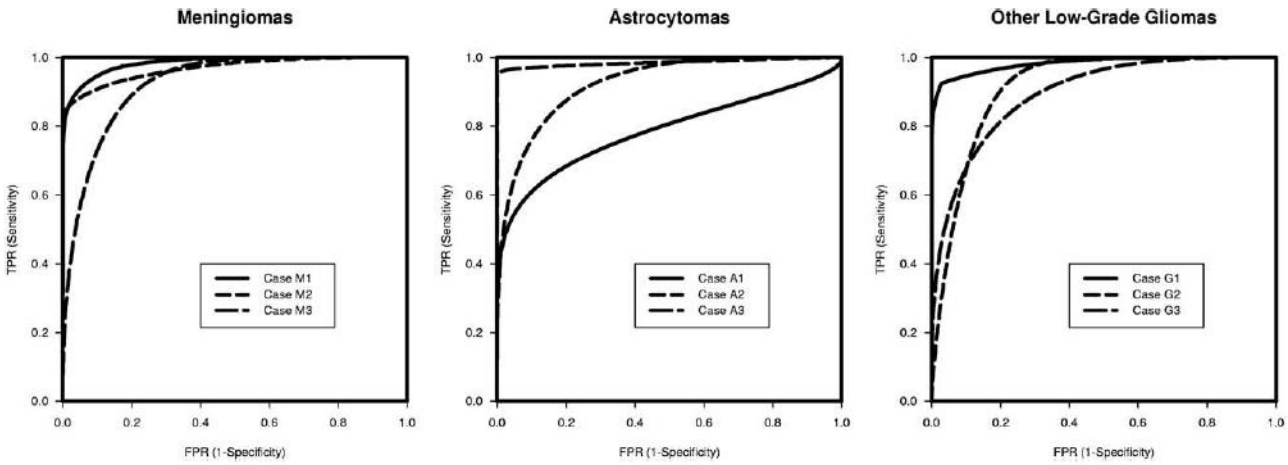


**Figure 2.**
Bi-beta ROC curves for Example 1 on MRI brain segmentations of meningiomas (left), astrocytomas (center), and gliomas (right).

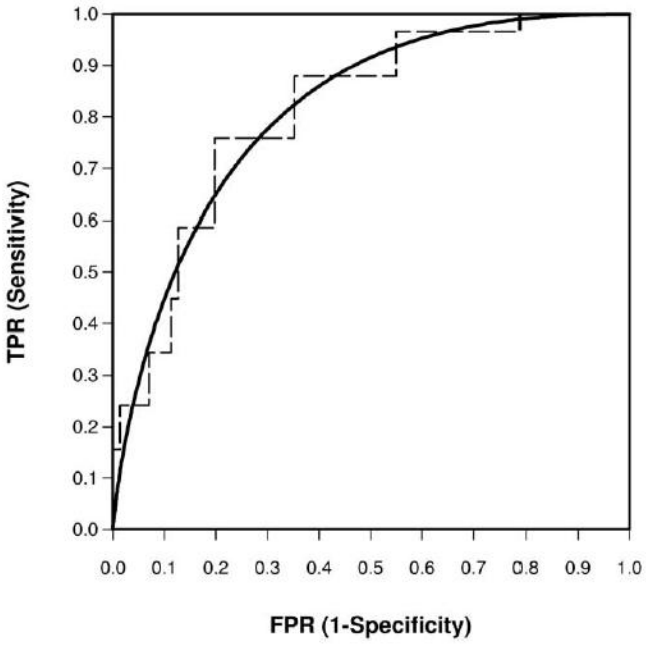


**Figure 3.**
Empirical and bi-normal ROC curve for Example 2 on CT of ureteral stone sizes for predicting treatment outcomes.

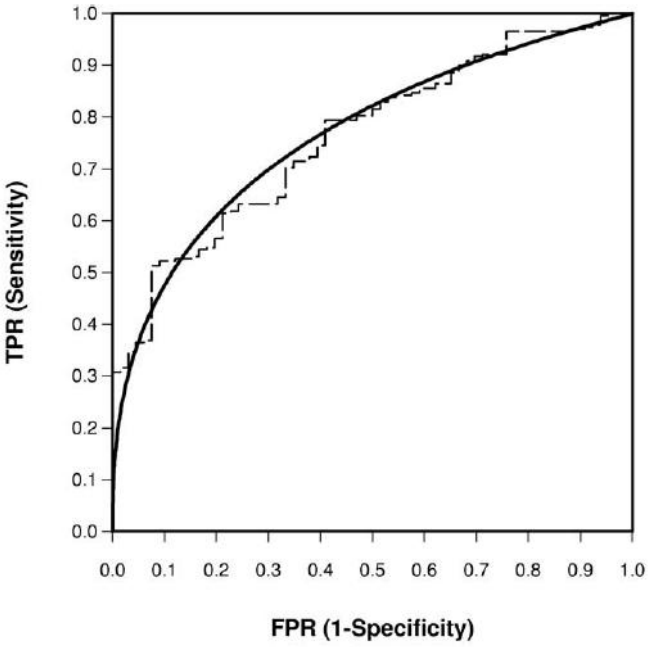


**Figure 4.**
Empirical and bi-normal ROC curve for Example 3 on PSA for prostate cancer staging.

**Table 1**

The Bi-Beta ROC Parameters and AUCs for Example 1

| Tumor Type | $\widehat{\alpha_0}$ | $\widehat{\beta_0}$ | $\widehat{\alpha_1}$ | $\widehat{\beta_1}$ | AUC |
|---|---|---|---|---|---|
| Meningioma | 0.029 | 0.885 | 0.269 | 0.041 | 0.984 |
| | 0.032 | 1.523 | 0.130 | 0.023 | 0.968 |
| | 0.172 | 0.783 | 1.184 | 0.339 | 0.924 |
| Astrocytoma | 3.208 | 5.504 | 1.379 | 0.794 | 0.786 |
| | 0.250 | 1.130 | 1.010 | 0.304 | 0.926 |
| | 0.177 | 2.679 | 0.217 | 0.009 | 0.986 |
| Glioma | 0.004 | 0.339 | 0.109 | 0.016 | 0.983 |
| | 0.106 | 0.573 | 1.169 | 0.411 | 0.916 |
| | 0.351 | 1.190 | 1.131 | 0.404 | 0.896 |

**Table 2**

Bi-Beta ROC Parameters and the Resulting AUCs in a Simulated Study

| Variances | $\widehat{\alpha_0}$ | $\widehat{\beta_0}$ | $\widehat{\alpha_1}$ | $\widehat{\beta_1}$ | AUC |
|---|---|---|---|---|---|
| Equal | 1 | 1 | 1 | 1 | 0.500 |
| | 1 | 1.5 | 1.5 | 1 | 0.706 |
| | 1 | 2 | 2 | 1 | 0.834 |
| | 1 | 2.5 | 2.5 | 1 | 0.908 |
| | 1 | 3 | 3 | 1 | 0.950 |
| | 1 | 9 | 9 | 1 | 1.000 |
| Unequal | 1 | 3 | 1.5 | 1 | 0.847 |
| | 1 | 1.5 | 3 | 1 | 0.847 |
| | 1 | 9 | 3 | 1 | 0.995 |
| | 1 | 3 | 9 | 1 | 0.995 |

**Table 3**

Bi-Normal ROC Parameters and the Resulting AUCs in a Simulated Study

| Variances | $\widehat{\alpha}$ | $\widehat{\beta}$ | AUC |
|---|---|---|---|
| Equal | 0.25 | 1 | 0.570 |
| | 0.50 | 1 | 0.638 |
| | 0.75 | 1 | 0.702 |
| | 1.00 | 1 | 0.760 |
| | 1.25 | 1 | 0.811 |
| | 1.50 | 1 | 0.856 |
| | 1.75 | 1 | 0.892 |
| | 2.00 | 1 | 0.921 |
| | 2.25 | 1 | 0.944 |
| | 2.50 | 1 | 0.961 |
| Unequal | 0.25 | 1.5 | 0.555 |
| | 0.50 | 1.5 | 0.609 |
| | 0.75 | 1.5 | 0.661 |
| | 1.00 | 1.5 | 0.710 |
| | 1.25 | 1.5 | 0.756 |
| | 1.50 | 1.5 | 0.797 |
| | 1.75 | 1.5 | 0.834 |
| | 2.00 | 1.5 | 0.866 |
| | 2.25 | 1.5 | 0.894 |
| | 2.50 | 1.5 | 0.917 |