JCP Online First, published on September 25, 2008 as 10.1136/jcp.2008.061010

Leading Article

Receiver-operating characteristic (ROC) curve analysis in diagnostic, prognostic and predictive biomarker research

Kjetil Søreide MD PhD *

Department of Surgery, Stavanger University Hospital

Stavanger, Norway

* Correspondence to:

Kjetil Søreide MD, PhD

Department of Surgery, Stavanger University Hospital,

POB 8100, Armauer Hansens vei 20, N-4068 Stavanger, Norway

Tel.: +47 5151 8830, fax: +47 5151 9919, email: ksoreide@mac.com.

Key words: biomarker; diagnostic accuracy; ROC analysis; prognosis; cancer

Word count: 3029; abstract 250; refs.: 43; tables: 2; figs.: 2.

ABSTRACT/SUMMARY

From a clinical perspective, biomarkers may have a variety of functions, which correspond to different stages in the disease development, e.g. in the progression of cancer. Biomarkers can assist in the care of patients for screening, diagnosis, prognosis, prediction and surveillance. Fundamental for the use of biomarkers in all situations is biomarker accuracy – the ability to correctly classify one condition and/or outcome from another. Receiver-operating characteristic (ROC) curve analysis is a useful tool in assessment of biomarker accuracy. Its advantages include testing accuracy across the entire range of scores and thereby not requiring a predetermined cut-off point, in addition to easily examined visual and statistical comparisons across tests or scores, and, finally, independence from outcome prevalence. Further, ROC curve analysis is a useful tool for evaluating the accuracy of a statistical model that classifies subjects into one of two categories. Diagnostic models are different from predictive and prognostic models in that the latter incorporate time-to-event analysis, for which censored data may pose a weakness of the model, or the reference standard. However, with the appropriate use of ROC curves, investigators of biomarkers can improve their research and presentation of results. ROC curves help identify the most appropriate classification rules. ROC curves avoid confounding resulting from varying thresholds with subjective ratings. The ROC curve results should always be put in perspective, because a good classifier does not guarantee the eventual clinical outcome, in particular for time-dependant events in screening, prediction, and/or prognosis studies where particular statistical precautions and methods are needed.

INTRODUCTION

From a clinical perspective, biomarkers may have a variety of functions, which correspond to different stages (table 1) in the disease development, such as in the progression in cancer or cardiovascular disease.[1, 2] Biomarkers can assist in the care of patients who are asymptomatic (screening biomarkers), those who are suspected to have the disease (diagnostic biomarkers) and those with overt disease (prognostic biomarkers) for whom therapy may or may not have been initiated. Biomarkers can also be used for treatment response (predictive biomarkers) or surveillance after therapy (monitoring biomarkers). Fundamental for the use of biomarkers in all situations is biomarker accuracy – the ability to correctly classify one condition and/or outcome from another (e.g. healthy vs diseased).

Table 1. Clinical use of biomarkers; rationale and objectives for cancer biomarkers

Type of Biomarker	Objective for use
Risk stratification	Assess the likelihood that cancers will develop (or recur)
Chemoprevention	Identify and target molecular mechanisms of carcinogenesis in (pre-)cancerous tissues
Screening	Detect and treat early-stage (pre-)cancers in the asymptomatic population
Diagnosis	Definitively establish the presence of cancer
Classification	Classify patients by disease subset
Prognosis	Predict the probable outcome of cancer regardless of therapy, to determine the aggressiveness of treatment
Prediction/ treatment stratification	Predict response to particular therapies and choose the drug that is mostly likely to yield a favorable response in a given patient
Risk management	Identify patients with a high probability of adverse effects of a treatment
Monitoring (i.e. chemotherapy)	Determine whether a therapy is having the intended effect on a disease and whether adverse effects arise
Surveillance after treatment/surgery	Early detection and treatment of recurrent disease

For the clinician diagnostic testing plays a fundamental role in clinical practice. For instance, daily surgical decision-making is based on the correct classification by pathology, radiology and/or clinical chemistry reports involving tissue and/or image evaluation and interpretation of disease conditions – many decisions of which the interpretation is based on results in the "grey-area" although requiring "black-and-white" answers for choice of treatment (Figure 1). Further, predictive modelling to estimate expected outcomes such as mortality or adverse events based on patient risk characteristics is common in any type of clinical research. *Receiver-operating characteristic (ROC) curve analysis* is a useful tool in assessment of biomarker accuracy in both situations – acknowledging strengths and weaknesses of the method.

ROC curve analysis is said to originally have developed during World War II to analyze classification accuracy in differentiating signal from noise in radar detection, before its principles later were implemented for improving medical desicion-making.[3] Evidently, the methodology has been adapted to several medical areas dependent on accuracy of screening and diagnostic tests, such as laboratory testing,[4, 5] epidemiology,[6, 7] radiology,[8] several clinical disciplines,[9-12] and bioinformatics.[13, 14] Further, ROC analysis has also been applied in histopathology in the attempt to define stages in diseases that show a continuous spectrum of histologic patterns, in which the uncertainty of boundary points and the overlap of features makes such definition difficult.[15] Construction and analysis of ROC curves may help to identify the features with the greatest utility, as, for example, in the grading of mucinous carcinomas of the ovary. ROC curves can also be used to assess diagnostic differences between histopathologists (as an alternative to the standard use of κ-coefficients),[15, 16] whether they are using different criteria or the same criteria but with different weightings, as, for example, in cervical pre-cancer or borderline ovarian tumors.

DIAGNOSTIC ACCURACY

Generally, diagnostic accuracy is referred to as the ability of a (laboratory) test or (bio)-marker to correctly classify subjects into clinically relevant groups (i.e. disease vs no disease; fig. 1). Diagnostic accuracy refers to the quality of the information provided by the classification device (i.e. the chosen cut-off level for a biomarker with a continuous spectrum of results) and should be distinguished from the usefulness, or actual practical value, of the information.[5] Diagnostic tests are usually measured and interpreted in their applicability by a number of features, including:

Sensitivity and specificity

Sensitivity, or the True Positive (TP) rate, which tells how good the test is at picking up people with the condition investigated. A high sensitivity is typically preferred in a screening test to rule out people without the disease.

Specificity, or the True Negative (TN) rate, which tells how good the test is at correctly defining people without the disease. A high specificity is required for diagnostic tests in order to have a low false positive rate.

Sensitivity and specificity are features of the test itself, and "looks backward" in that they show the probability that a person with a disease will have a positive test, rather than "looking forward" and showing the probability that the person (or patient) who tests positive actually has the disease. The latter is better performed by the predictive values:

Positive and negative predictive values

Positive predictive value (PPV; or, the post-test probability of a positive test); is a measure of the probability of having the condition, if a person tests positive.

Negative predictive value (NPV, or the post-test probability of a negative test); will address the situation "if a patient/person tests negative on a test, what is the probability of not having the condition/disease".

Accuracy

Accuracy gives the proportion of all tests that have given the correct result (true positives and true negatives) as proportion of all the results. Assessing the accuracy of any diagnostic procedure remains integral to method evaluation. Evaluation of a diagnostic procedure is assessed by its ability to categorise patients accurately into those with or without a disease state.

Likelihood ratio

Because one test may have higher sensitivity but lower specificity than another, the diagnostic likelihood ratio is sometimes used to combine these measures. Likelihood ratios (LR+; positive LR) is an estimate of the relative predictive value of a test (true positives/false positives), is useful in clinical practice as it indicates how likely a positive result will be found in a person with the disease compared to a person without the disease. LR of a test indicates the increase from pre-test probability (e.g. prevalence of the disease) to post-test probability. Interpretation of LRs can be used by nomograms. As a rule of thumb, LR over 10 is generally regarded as large and a conclusive change in pre- to post-test probability of having the disease. LR of 5 to 10 are considered moderate, and LR<2 are rarely considered important. For tests with a continuous range of test-results (e.g. from 1-100), rather than a dichotomous test-result (positive/negative; yes/no; red/green; present/absent), the sensitivity and specificity (and the LR) heavily relies on the chosen cut-off value for dichotomization into f. ex. healthy vs diseased. In particular for diagnostic test, the corresponding LR for a given test (or biomarker cut-off) result should be presented together with the other diagnostic features.

CONTINGENCY TABLE

The most commonly used analytical model for evaluating a test is the standard "2 × 2" or "contingency table" method in which sensitivity and specificity are calculated. However, there are several limitations to this approach, including the reliance on a single defined criterion or cut-off for determining a true-positive result (such as arbitrarily chosen percentiles for immunohistochemistry markers) or defining "abnormal",[17] use of non-standardized measurement instruments and sensitivity to outcome prevalence. In this setting, the ROC analysis is a more appropriate and useful technique for assessing diagnostic and predictive accuracy.[18] Its advantages include testing accuracy across the entire range of scores and thereby not requiring a predetermined cut-off point, in addition to easily examined visual and statistical comparisons across tests or scores, and, finally, independence from outcome prevalence. Further, ROC curve analysis is a useful tool for evaluating the accuracy of a statistical model (e.g. logistic regression, linear discriminant analysis) that classifies subjects into one of two categories (i.e. sick or healthy). The function as a simple graphical tool for displaying the accuracy of a medical diagnostic test is one of the most well known applications of ROC curve analysis. For example, the expression (positive stain) of a given protein biomarker in precancerous tissues may produce a continuous spectrum of test results (i.e. from zero to 100%). Thus, the diagnostic properties of such a biomarker (as expressed by sensitivity, specificity, predictive values, or likelihood ratios) depend on the chosen cut-off value to differentiate between normal and diseased states.[19, 20]

A diagnostic classification test typically yields binary, ordinal, or continuous outcomes. The simplest type, binary outcomes, arises from a test indicating whether the patient is healthy or diseased. The test indicates whether the patient is likely to be diseased or not. When more than two categories are used, the test data can be on an ordinal rating scale – such as a 5-point ordinal (0, 1+, 2+, 3+, 4+) scale for disease severity. When a particular cut-off level or threshold is of particular interest, an ordinal scale may be dichotomized (e.g. values ≤ 2+ in one group and, values >2+ in a second group), in which case methods for binary outcomes can be used. Test data such as serum markers (e.g. CEA measurements)[21] or physiological markers also may be acquired on a continuous scale.[2, 9] In particular, ROC plots occupy a central position in the process of assessing and using diagnostic tools.[5, 21, 22]

Reference standard ("Gold Standard")

The presence or absence of the disease state is defined according to some, sometimes arbitrarily selected, "reference standard". Obviously, the nature of the reference standard can itself be a cause for debate. To estimate classification accuracy using standard ROC methods, the disease status for each patient is measured *without error*, that is, the endpoint has to be defined without uncertainty (which is sometimes not the case). The true disease status often is referred to as the *reference standard* (previously named "gold standard"). The reference standard may be available from clinical follow-up, surgical verification, biopsy, and autopsy, or in some cases by a committee of "experts" or in particular situations by the use of results of multiple imperfect tests (referred to as "latent class analysis").[23]

Bias

Obviously, bias in accuracy testing may occur due to such factors as case mix, severity of disease, and selection of control subjects, as well as measurement technique and quality of the reference standard. In selection of the reference standard both verification bias and measurement error can occur. Verification bias results when the accuracy of a test is evaluated only among those with known disease status (excluding the "unknown" or "indeterminate"). Measurement error may result when a true reference standard is absent or an imperfect standard is used for comparison.

THE ROC CURVE

Receiver-operating characteristic (ROC) plots provide a statistical method to assess the diagnostic accuracy of a test (or biomarker) which has a continuous spectrum of testresults.[22] A ROC curve is a graphical display of the tradeoffs of the true-positive rate (sensitivity) and false-positive rate (1 – specificity) corresponding to all possible binary tests that can be formed from this continuous biomarker.[22] Each classification rule, or cut-off level, generates a point on the graph. The closer the curve follows the left-hand border and then the top-border of the ROC space, the more accurate the test. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test. The traditional ROC curve arises when a continuous value is measured in each subject and the classification is positive if the value is above a threshold. As the threshold varies, a new classification rule is created, and the resulting plot is a single curve. The optimal ROC curve is the line connecting the points highest and farthest to the left-upper corner. The rationale for the optimal ROC curve is that it captures the trade-off between sensitivity and specificity over a continuous range. Further, in a ROC curve the slope of the tangent line at a cut-point gives the *likelihood ratio* (LR) for that value of the test.

Area Under the ROC Curve (AUC)

Area under the curve (AUC) is also known as the *c-statistic* or *c index,* and can range from 0.5 (random chance, or no predictive ability; refers to the 45 degree line in the ROC plot; see fig. 2) to 1 (perfect discrimination/accuracy). On rare occasions, the estimated AUC is <0.5, indicating that the test does worse than chance. The AUC is a measure of overall diagnostic accuracy of the test, and the cut-off value providing the highest sensitivity and specificity is calculated (figure 2). Importantly, the results are independent of the prevalence of the disease. Optimal cut-off values are identified by the significant AUC of the ROC analysis and the continuous scores can then be dichotomized accordingly. Depending on the program used for ROC analysis, the AUC may be presented with 95% confidence intervals to indicate if the curve crosses or nears the non-significant 0.5 value at any point.

Other features of the ROC curve may be of interest in particular applications, such as the partial area under the ROC curve (PAUROC) analysis,[24, 25] which could be used, for example, when the specificity for a biomarker for cancer screening must be above a certain threshold to be clinically useful. Cut-offs for 90% sensitivity and 90% specificity, respectively, should thus usually be presented as well.

Biomarkers for early diagnosis and screening

Biomarkers are increasingly being developed to detect tumors early when disease is in a less progressed state and treatment is likely to be more successful. In contrast to research and clinical implementation of new drugs, the research and incorporation of (new) laboratory tests has not been subject to the same rigid restrictions, trials and surveillance – on the contrary, standards for evaluating new clinical classifiers lag far behind the well established standards that exist for evaluating new clinical treatments.[26, 27] Thus, the disappointing performance of many markers, such as the early reports of carcinoembryonal antigen (CEA) in colorectal cancer,[28] that were initially shown to have a strong association with outcome may be, in part, because of that a marker that is strongly associated with outcome may not be effective for predicting those who are likely and those who are not likely to have the outcome.[29]

Recently, standards for reporting recommendations for tumor marker (REMARK) prognostic studies have been proposed in a 20-step model.[30] Further, five phases of biomarker development for early detection of cancer have been proposed (table 2).[31]

Table 2. Steps for biomarker development

Step/phase	**Study type**	**Objectives**
Phase 1	Preclinical exploratory	Promising directions identified
Phase 2	Clinical assay and validation	Clinical assay detects established disease
Phase 3	Retrospective longitudinal	Biomarker detects disease early before it becomes clinical and a ”screen postive” rule is defined
Phase 4	Prospective screening	Extent and characteristics of disease detected by the test and the false referral rate are defined
Phase 5	Cancer control	Impact of screening on reducing the burden of disease on the population is quantified

Developed from (31).

The REMARK guidelines[30] appear to be most relevant to phase 2 and 3 biomarker studies. In these phases the receiver operating characteristic (ROC) curve can be used to estimate the ability of a continuous biomarker to predict or classify the patients who are likely and those who are not likely to have the outcome. Although the ROC analysis applies only to step 11 of the 20 REMARK steps,[30] it is nonetheless an important and increasingly appreciated tool for the accuracy assessment in biomarker research. Importantly, when investigating biomarkers that are used directly for early detection (i.e. colorectal cancer screening) a low false-positive rate is required.[22] One might also be interested in either biomarkers for possible cancer precursors (e.g., adenomas) or those that would be used in conjunction with a more invasive test (such as colonoscopy).[31] In these other situations, one would often be interested in a different part of the ROC curve (e.g., higher false- and true-positive rates).[22]

ROC and role in predictive/prognostic time-to-event analyses

Diagnostic and prognostic or predictive models serve different purposes. Whereas diagnostic models are usually used for classification, prognostic models incorporate the dimension of time, adding a stochastic element.[32] Although it is useful for classification, evaluation of prognostic models should not rely solely on the ROC curve, but should assess both discrimination and calibration for which criticisms, and solutions, have been proposed.[6, 32, 33] In assessing predictive and prognostic models one should acknowledge that the *time-to-event* is the actual object under analysis – it is thus a surrogate for the actual event (e.g. mortality) and should be interpreted cautiously. In particular this holds true for the situation when many patients are censored to the actual outcome. In fact, ROC analysis methods are not well developed for the analysis of censored time data, and specific statistical considerations may be necessary to adjust for confounding factors in this setting.[6, 34] For example, it is known that PSA tend to rise with age, and thus the predictive/prognostic capacity with any cut-off will be influenced by the time from testing to outcome, if not taken into consideration.[33] As an example used by Pepe et al[31] the ROC curve and cut-offs for PSA sampled at 1, 2, 4, and 8 years prior to the diagnosis of prostate cancer may yield very different results as the time effect influence on biology and thus the obtained test data. Although the ROC analysis does not require longitudinal data, Pepe et al suggest[31] that a series of biomarker values over time from a relatively small number of subjects are preferable to more subjects contributing fewer measurements each. The longitudinal data will allow assessment of within-subject variability and more powerful comparisons of time-specific ROC curves, thereby providing better statistical evaluation of time trends in the ability to discriminate between control subjects and case subjects.[34] Recently, our group demonstrated the value for serial CEA measurements in assessing the diagnostic accuracy on colorectal cancer recurrence detection in a systematic surveillance after surgery. The diagnostic accuracy of CEA was influenced by the chosen cut-off value, and a threefold increase in CEA over time ("slope of increase") indicated recurrent disease just as well, or better, than the absolute value within or outside the normal range.[21]

Tools that ROC

Most statistical programs (i.e SPSS Inc, Chicago, USA) provide some sort of ROC curve analysis, albeit with different degrees of values and analytical tools available. One study evaluated eight programs running under Windows (AccuROC, Analyse-It, CMDT, GraphROC, MedCalc, mROC, ROCKIT, and SPSS) finding strengths and weaknesses in most.[35, 36] Up-dated and improved versions have since been provided for some of the programmes, such as MedCalc™. An anaesthetist website ("The magnificent ROC" at http://www.anaesthetist.com/mnm/stats/roc/Findex.htm) provides a "virtual tour" for hands-on experience by manipulating cut-offs and test distributions and direct visualisation of the effects in the ROC curve.

While this article deals with the principle of evaluating the accuracy of a single diagnostic test for a binary outcome summarized by the AUC, more sophisticated tools are developed for ROC assessment for complex situations including microarray data or mass spectrometry involving multiple markers,[13, 37-39] and sample size estimations (cases to controls).[40] The use of ROC curve analysis, as any other method, may be subject to bias,[41-43] but will more likely help in reducing the chance for overfitting in choosing appropriate thresholds when investigating new biomarkes.[22]

Conclusion

With the appropriate use of ROC curves, investigators of biomarkers can improve their research and presentation of results. ROC curves help identify the most appropriate classification rules. ROC curves avoid confounding resulting from varying thresholds with subjective ratings. The ROC curve results should always be put in perspective, because a good classifier does not guarantee the eventual clinical outcome, in particular for time-dependant events in screening, prediction, and/or prognosis studies where particular statistical precautions and methods are needed.[22] In the end, such information may be most efficiently and reliably derived from randomized controlled trials.

Competing interests: None

References

1. Søreide K, Nedrebø BS, Knapp JC, Glomsaker TB, Søreide JA, Kørner H. Evolving molecular classification by genomic and proteomic biomarkers in colorectal cancer: Potential implications for the surgical oncologist. *Surg Oncol* 2008 in press.
2. Gerszten RE, Wang TJ. The search for new cardiovascular biomarkers. *Nature* 2008;**451**(7181): 949-952.
3. Lusted LB. Decision-making studies in patient management. *N Engl J Med* 1971;**284**(8): 416-424.
4. Obuchowski NA, Lieber ML, Wians FH, Jr. ROC curves in clinical chemistry: uses, misuses, and possible solutions. *Clin Chem* 2004;**50**(7): 1118-1125.
5. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin Chem* 1993;**39**(4): 561-577.
6. Pepe MS, Janes HE. Gauging the performance of SNPs, biomarkers, and clinical factors for predicting risk of breast cancer. *J Natl Cancer Inst* 2008;**100**(14): 978-979.
7. Walter SD, Sinuff T. Studies reporting ROC curves of diagnostic and prediction data can be incorporated into meta-analyses using corresponding odds ratios. *J Clin Epidemiol* 2007;**60**(5): 530-534.
8. Obuchowski NA. ROC analysis. *AJR Am J Roentgenol* 2005;**184**(2): 364-372.
9. Zou KH, O'Malley AJ, Mauri L. Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation* 2007;**115**(5): 654-657.
10. Musial J, Swadzba J, Motyl A, Iwaniec T. Clinical significance of antiphospholipid protein antibodies. Receiver operating characteristics plot analysis. *J Rheumatol* 2003;**30**(4): 723-730.
11. Cheung R, Altschuler MD, D'Amico AV, Malkowicz SB, Wein AJ, Whittington R. Using the receiver operating characteristic curve to select pretreatment and pathologic predictors for early and late postprostatectomy PSA failure. *Urology* 2001;**58**(3): 400-405.
12. Kørner H, Søreide JA, Sondenaa K. Diagnostic accuracy of inflammatory markers in patients operated on for suspected acute appendicitis: a receiver operating characteristic curve analysis. *Eur J Surg* 1999;**165**(7): 679-685.
13. Li J, Fine JP. ROC analysis with multiple classes and multiple tests: methodology and its application in microarray studies. *Biostatistics* 2008;**9**(3): 566-576.
14. Zheng Y, Cai T, Feng Z. Application of the time-dependent ROC curves for prognostic accuracy with multiple biomarkers. *Biometrics* 2006;**62**(1): 279-287.
15. Langley FA, Buckley CH, Tasker M. The use of ROC curves in histopathologic decision making. *Anal Quant Cytol Histol* 1985;**7**(3): 167-173.
16. Taube A. To judge the judges--kappa, ROC or what? *Arch Anat Cytol Pathol* 1995;**43**(4): 227-233.
17. Smellie WS. When is "abnormal" abnormal? Dealing with the slightly out of range laboratory result. *J Clin Pathol* 2006;**59**(10): 1005-1007.

18. Lang TA, Secic M. Chapter 10. Determining the Presence or Absence of Disease: reporting the performance Characteritics of Diagnostic tests. In: *How to report Statistics in Medicine Annotated Guidelines for Authors, Editors, and Reviewers*, Lang TA, Secic M (eds). American College of Physicians: Philadelphia, 2006; 125-149.
19. Zlobec I, Steele R, Terracciano L, Jass JR, Lugli A. Selecting immunohistochemical cut-off scores for novel biomarkers of progression and survival in colorectal cancer. *J Clin Pathol* 2007;**60**(10): 1112-1116.
20. Søreide K, Gudlaugsson E, Skaland I, Janssen EA, Van Diermen B, Kørner H, Baak JP. Metachronous cancer development in patients with sporadic colorectal adenomas-multivariate risk model with independent and combined value of hTERT and survivin. *Int J Colorectal Dis* 2008;**23**(4): 389-400.
21. Kørner H, Søreide K, Stokkeland PJ, Søreide JA. Diagnostic accuracy of serum-carcinoembryonic antigen in recurrent colorectal cancer: a receiver operating characteristic curve analysis. *Ann Surg Oncol* 2007;**14**(2): 417-423.
22. Baker SG. The central role of receiver operating characteristic (ROC) curves in evaluating tests for the early detection of cancer. *J Natl Cancer Inst* 2003;**95**(7): 511-515.
23. Pepe MS, Janes H. Insights into latent class analysis of diagnostic test performance. *Biostatistics* 2007;**8**(2): 474-484.
24. Li CR, Liao CT, Liu JP. A non-inferiority test for diagnostic accuracy based on the paired partial areas under ROC curves. *Stat Med* 2008;**27**(10): 1762-1776.
25. Dodd LE, Pepe MS. Partial AUC estimation and regression. *Biometrics* 2003;**59**(3): 614-623.
26. Walley T. Evaluating laboratory diagnostic tests. *BMJ* 2008;**336**(7644): 569-570.
27. Pepe MS. Evaluating technologies for classification and prediction in medicine. *Stat Med* 2005;**24**(24): 3687-3696.
28. Ransohoff DF. Rules of evidence for cancer molecular-marker discovery and validation. *Nat Rev Cancer* 2004;**4**(4): 309-314.
29. Ware JH. The limitations of risk factors as prognostic tools. *N Engl J Med* 2006;**355**(25): 2615-2617.
30. McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM. Reporting recommendations for tumor marker prognostic studies. *J Clin Oncol* 2005;**23**(36): 9067-9072.
31. Pepe MS, Etzioni R, Feng Z, Potter JD, Thompson ML, Thornquist M, Winget M, Yasui Y. Phases of biomarker development for early detection of cancer. *J Natl Cancer Inst* 2001;**93**(14): 1054-1061.
32. Cook NR. Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. *Clin Chem* 2008;**54**(1): 17-23.
33. Janes H, Pepe MS. Adjusting for covariates in studies of diagnostic, screening, or prognostic markers: an old concept in a new setting. *Am J Epidemiol* 2008;**168**(1): 89-97.
34. Pepe MS, Zheng Y, Jin Y, Huang Y, Parikh CR, Levy WC. Evaluating the ROC performance of markers for future events. *Lifetime Data Anal* 2008;**14**(1): 86-113.
35. Stephan C, Wesseling S, Schink T, Jung K. Comparison of eight computer programs for receiver-operating characteristic analysis. *Clin Chem* 2003;**49**(3): 433-439.

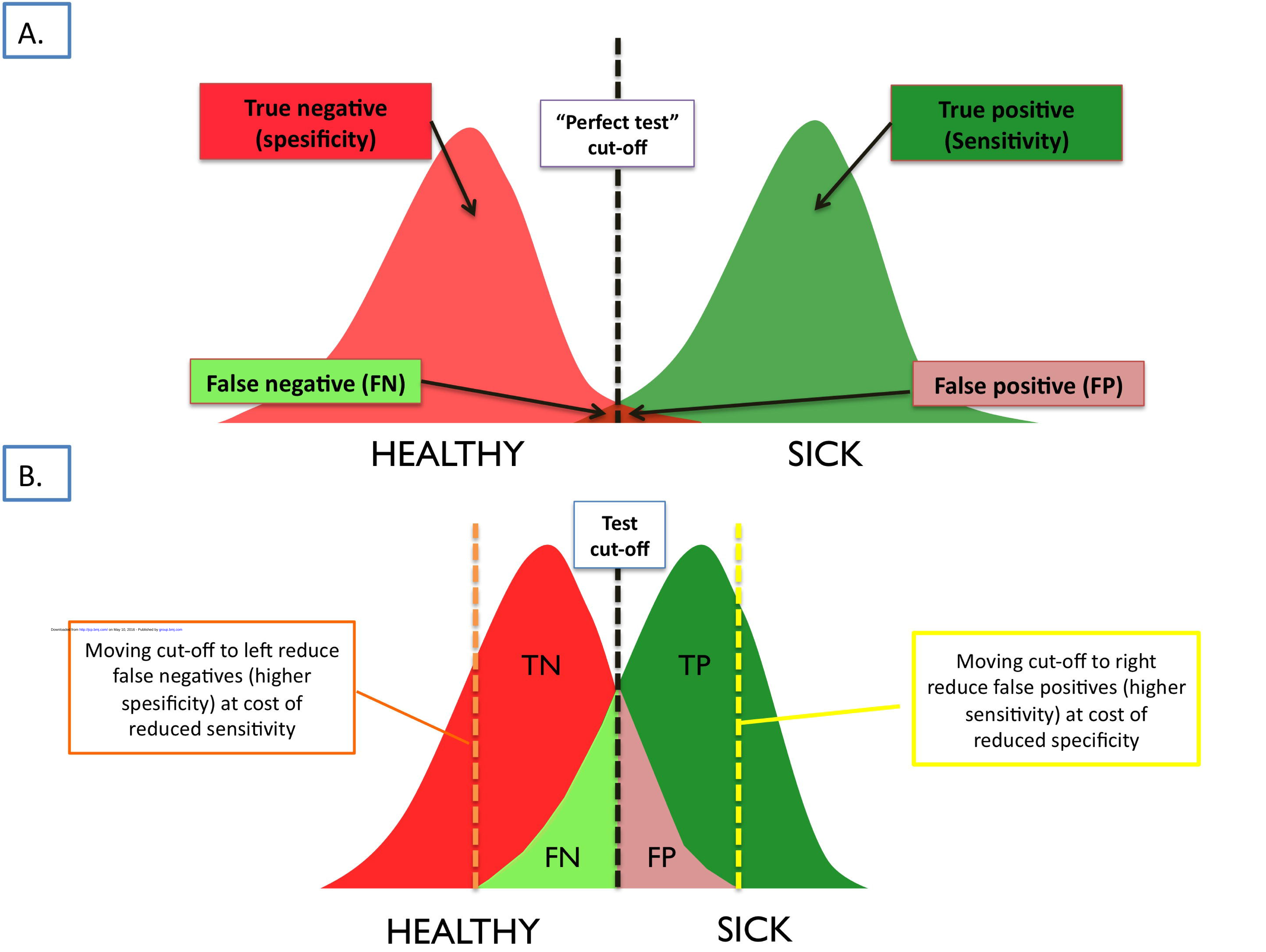
36. Ison JC, Blades MJ. ROCPLOT: a generic software tool for ROC analysis and the validation of predictive methods. *Appl Bioinformatics* 2005;**4**(2): 131-135.
37. Parodi S, Muselli M, Fontana V, Bonassi S. ROC curves are a suitable and flexible tool for the analysis of gene expression profiles. *Cytogenet Genome Res* 2003;**101**(1): 90-91.
38. Ye J, Liu H, Kirmiz C, Lebrilla CB, Rocke DM. On the analysis of glycomics mass spectrometry data via the regularized area under the ROC curve. *BMC Bioinformatics* 2007;**8**: 477.
39. Pepe MS, Cai T, Longton G. Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics* 2006;**62**(1): 221-229.
40. Janes H, Pepe M. The optimal ratio of cases to controls for estimating the classification accuracy of a biomarker. *Biostatistics* 2006;**7**(3): 456-468.
41. Hanley JA. Receiver operating characteristic (ROC) methodology: the state of the art. *Crit Rev Diagn Imaging* 1989;**29**(3): 307-335.
42. Clark RD, Webster-Clark DJ. Managing bias in ROC curves. *J Comput Aided Mol Des* 2008;**22**(3-4): 141-146.
43. Ransohoff DF. Bias as a threat to the validity of cancer molecular-marker research. *Nat Rev Cancer* 2005;**5**(2): 142-149.

Figure legends

Figure 1 A and B. In **A)** is depicted a “good” discriminatory test with fairly “perfect” discriminatory ability between the diseased and the healthy population. The two populations show little overlap in test spectrum, which also contributes to good discriminatory ability of the test. In **B)** is depicted a “real life” situation where the populations show considerable overlap in test spectrum causing a reduced discriminatory ability of the given marker. The sensitivity may be improved at the cost of specificity by changing the cut-off for the test, and vice versa.

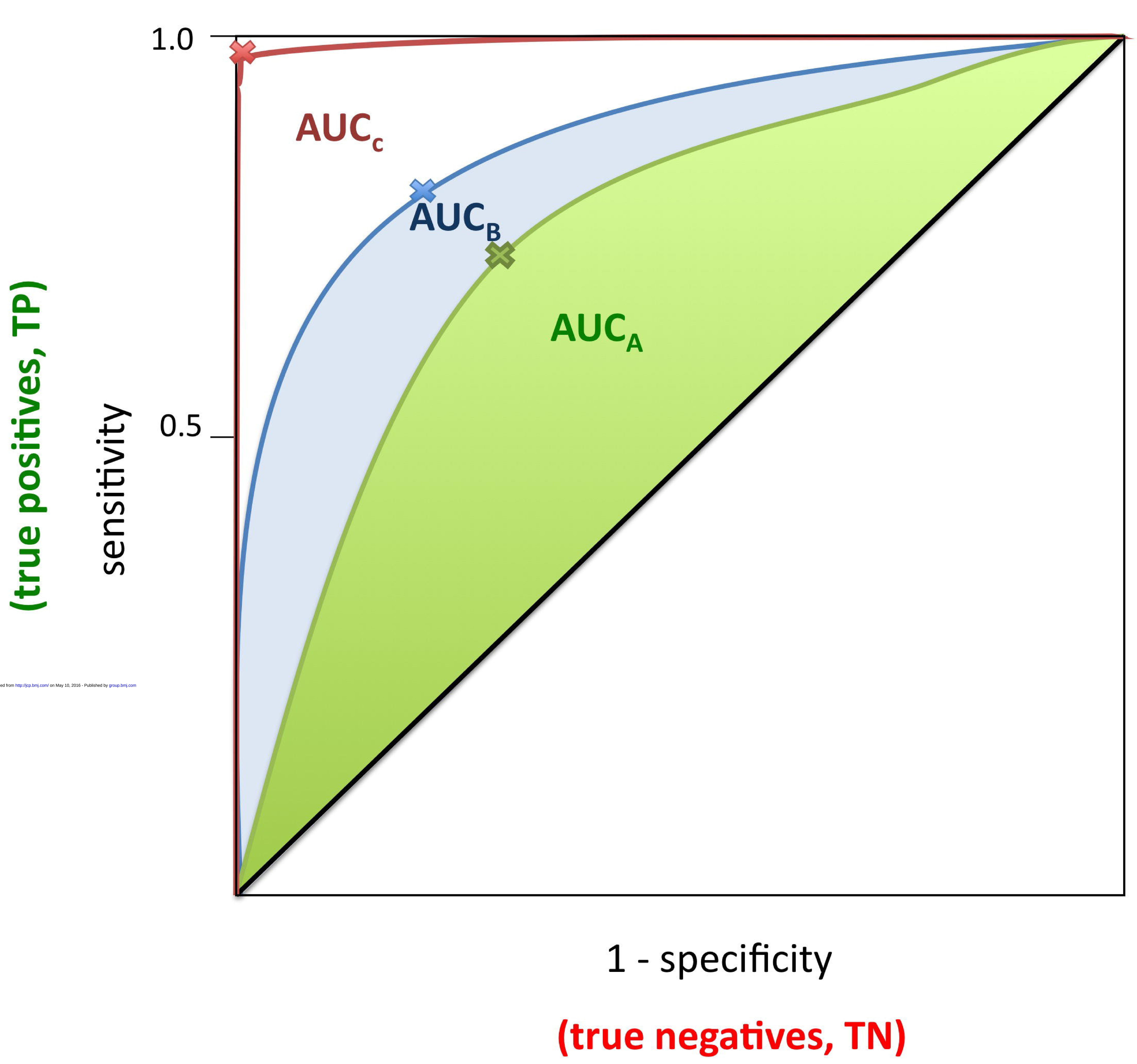
TP denotes true positives, TN true negatives; FN false negatives; FP false positives.

Figure 2. Illustration of ROC curves. Three plots and their respective “area under the curve” (AUC) are given. The diagnostic accuracy of marker C (red) is better than that of B and A, as the AUC of C>B>A. Cross (“x”) marks point of best cut-off for the biomarker.



Moving cut-off to left reduce false negatives (higher spesificity) at cost of reduced sensitivity

Moving cut-off to right reduce false positives (higher sensitivity) at cost of reduced specificity


1.0
0.5
AUC$_C$
AUC$_B$
AUC$_A$
sensitivity
(true positives, TP)
1 - specificity
(true negatives, TN)

JCP

Receiver-operating characteristic (ROC) curve analysis in diagnostic, prognostic and predictive biomarker research

Kjetil Søreide

J Clin Pathol published online September 25, 2008