Disease Markers 23 (2007) 113–120
IOS Press

Statistical methods for evaluating DNA methylation as a marker for early detection or prognosis

Todd A. Alonzo and Kimberly D. Siegmund*
Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California 90089, USA

Abstract. We summarize standard and novel statistical methods for evaluating the classification accuracy of DNA methylation markers. The choice of method will depend on the type of marker studied (qualitative/quantitative), the number of markers, and the type of outcome (time-invariant/time-varying). A minimum of two error rates are needed for assessing marker accuracy: the true-positive fraction and the false-positive fraction. Measures of association that are computed from the combination of these error rates, such as the odds ratio or relative risk, are not informative about classification accuracy. We provide an example of a DNA methylation marker that is strongly associated with time to death (logrank $p = 0.0003$) that is not a good classifier as evaluated by the true-positive and false-positive fractions. Finally, we would like to emphasize the importance of study design. Markers can behave differently in different groups of individuals. It is important to know what factors may affect the accuracy of a marker and in which subpopulations the marker may be more accurate. Such an understanding is extremely important when comparing marker accuracy in two groups of subjects.

Keywords: Sensitivity, specificity, ROC curves, survival data

1. Introduction

Novel technologies for measuring DNA methylation provide the opportunity to hunt for new markers for early cancer detection and prognosis. Evidence showing DNA methylation occurs early in carcinogenesis has made it a popular target for studies of early detection [17,37]. At the same time, studies showing DNA methylation predicts treatment response have suggested its potential as a marker for prognosis [21,38]. Since DNA methylation patterns change with age [31], controlling for age effects can be important when evaluating a marker's potential.

Various technologies characterize DNA methylation in the human genome differently [7,18,35]. Measurements can be obtained for a single CpG site or for a series of linked CpGs. Depending upon the technology the measurements can be either binary or quantitative. A collection of DNA methylation measurements at a number of different locations across the genome provides a DNA methylation profile or footprint. The statistical method used for marker evaluation will depend on the type of measurement obtained as well as the number of markers studied.

Most studies of DNA methylation for early detection are either preclinical exploratory studies or studies focusing on the development and validation of clinical assays. These have been labeled Phase 1 and Phase 2 studies in the proposed five-phase protocol of biomarker development for early cancer detection [25]. Many of the studies are small, with at most 150 individuals. However, the number of markers studied is growing. Genome scanning techniques can now measure $>$ 5,000 CpG islands simultaneously from a single tissue sample. Therefore, statistical methods that can sift

*Corresponding author: Kimberly D. Siegmund, Department of Preventive Medicine, USC Keck School of Medicine, 1540 Alcazar Street, CHP-220, Los Angeles, CA 90089, USA. Tel.: +1 323 442 1310; Fax: +1 323 442 2349; E-mail: kims@usc.edu.

ISSN 0278-0240/07/$17.00

through a large number of markers in early preclinical exploratory studies are needed.

DNA methylation has been studied as a prognostic factor in cohorts of patients. In general, studies identifying DNA methylation markers for prognosis have based their conclusions on logrank tests of association [5,11,37,38,41]. Several authors have noted that markers that show strong associations with outcome are not necessarily good classifiers [2,10,29]. Pepe et al. [29] illustrate the limitations of the odds ratio in gauging the performance of a marker for screening or prognosis. Similar examples suggest limitations of the logrank test or hazard ratio for evaluating classification accuracy. This suggests predictors of treatment response are potentially poor classifiers and need to be re-evaluated for their classification accuracy. The last five years has seen the development of novel statistical methods for evaluating classification accuracy for treatment response. We will survey these methods along with the classical methods for evaluating biomarkers for early detection.

2. Study design

Studies to assess the accuracy of markers can be performed prospectively or retrospectively. Retrospective studies involve selecting subjects on the basis of their true disease status as determined by the gold standard and assessing the markers on them. These retrospective studies are often called case-control studies where cases are those individuals with disease and controls are those without disease. Prospective studies involve determining marker values for a random sample from the population of interest and determining true disease status for all study subjects. True disease status can be determined concurrently for a cross-sectional cohort study or over a follow-up period for a prospective cohort study.

3. Statistical methods

The statistical methods for evaluating markers of early detection or prognosis are similar. The primary difference is that the tolerance for false-positive results varies in the two settings. In studies of early detection, a very small false-positive rate is necessary to avoid costly follow-up and undue stress to non-diseased individuals receiving a positive test. In studies of prognosis, a higher false-positive rate may be acceptable.

We present statistical methods for evaluating qualitative and quantitative markers as classifiers of disease. Typically, disease status is qualitative, subjects are either diseased or not at the time of study. Such outcomes are called time invariant, meaning that the disease status of the individual does not change over time. Death within a certain follow-up time or disease status at some future time point might also be treated as a time-invariant (qualitative) outcome. However, for these outcomes it might be more natural to allow disease status to change over time. Then the outcome of interest could be time to death or time to disease recurrence. This type of outcome is called time varying and can accommodate different lengths of follow-up for subjects in the study. First we describe statistical methods for the evaluation of single markers depending on the type of marker (qualitative/quantitative) and the type of disease outcome (time invariant/time varying). Later we describe methods for combining markers and methods for comparing different marker panels.

3.1. Evaluating a classifier – Single Marker Models

The accuracy of a qualitative marker is measured by two error rates: the true-positive fraction (TPF) and the false-positive fraction (FPF). The TPF, also known as sensitivity, is the probability of a positive marker in subjects who have disease (i.e., Pr [positive marker | disease]). The FPF, also known as 1-specificity, is the probability of a positive marker in subjects who do not have disease (i.e., Pr [positive marker | no disease]). A perfect marker will have a TPF of 1 and FPF of 0. In practice markers are imperfect and their value depends on the context. Investigators may be willing to withstand a higher FPF in order to obtain a higher TPF for prognostic markers used for classifying diseased individuals to treatment groups. On the other hand an extremely low FPF is desirable for markers used to screen a healthy population because of the follow up costs of a positive result.

The receiver operating characteristic (ROC) curve evaluates the accuracy of quantitative markers. It describes the (TPF, FPF) for every possible cut point of the marker (Fig. 1). ROC curves measure the amount of separation between the distribution of marker values in the diseased population from that in the non-diseased population. When the distributions of marker values for the diseased and non-diseased populations completely overlap, then the ROC curve is the forty-five degree line from (0,0) to (1,1) indicating a non-informative test. The more separated the distributions,

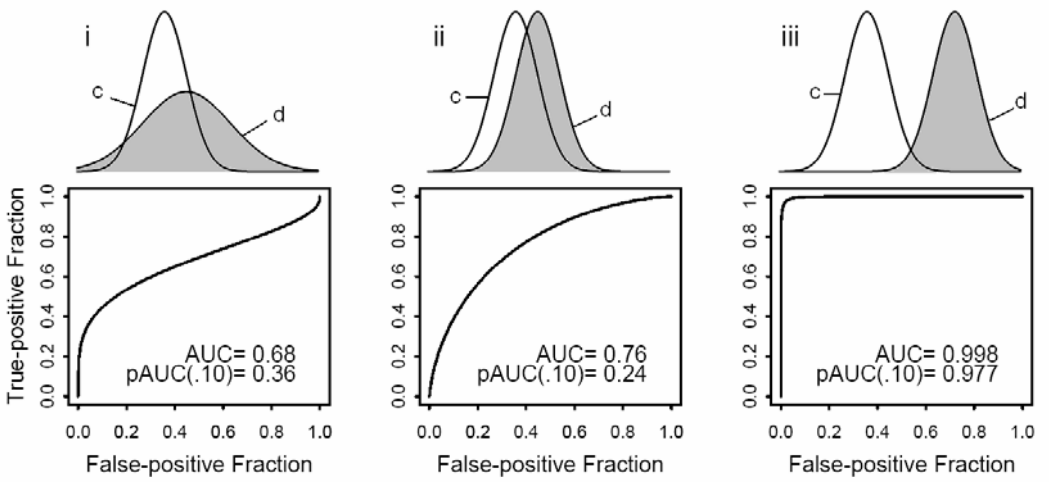

Fig. 1. Hypothetical marker distributions in diseased (d) and non-diseased (c) subjects with corresponding ROC curve below them. pAUC corresponding to the area under the ROC curve from FPF 0 to 0.1 are provided.

the closer the ROC curve is to the upper left-hand corner. The units of measurement have no impact on the ROC curve so that ROC curves can be used to compare the accuracy of different markers even when they are measured either in different units or on different scales.

The typical summary measure of an ROC curve is the area under the curve (AUC) which is equal to the Wilcoxon ranksum statistic. The AUC takes on a value between 0 and 1, where AUC is equal to 1 for a perfect marker and is equal to 0.5 for an uninformative marker. AUC is interpreted as the probability that the marker value for a randomly chosen diseased subject is greater than the marker value for a randomly chosen non-diseased subject. AUC can also be interpreted as the average TPF over the whole range of possible FPF.

When markers are compared, it is often by their AUC. Some argue that it makes more sense to compare ROC curves only for the FPF region of interest [23]. For a maximum acceptable FPF ($= f_0$) one might compute two statistics: (1) the TPF when FPF $= f_0$ and (2) the area under the ROC curve from 0 to f_0. This latter statistic is referred to as the partial AUC (pAUC) and corresponds to the average ROC for all FPF $< f_0$. Using simulations, Pepe et al. [27] show that focusing on the area of the FPF region of interest is most important for detecting genes where the distribution of the marker in cases differs from the distribution in controls by more than a simple shift of the mean value. One such example is a larger marker variation in cases compared to controls (Fig. 1i). In Fig. 1, the partial AUC is higher in scenario (i) where the variation of the marker in cases is greater than in controls as compared to scenario (ii) where the variation of the marker is the same in cases and controls and only the mean is slightly higher in the case group. For markers that are highly predictive (Fig. 1iii), the markers are likely to be identified regardless of the method chosen, AUC or pAUC.

For the methods described above, all subjects are classified a priori as diseased or non-diseased. However in prospective cohort studies, the disease status of an individual can change over time. When the outcome status changes over time, e.g. non-diseased individuals become diseased, several approaches to marker evaluation are available. The traditional approach is to define a fixed follow-up time for all subjects, classify them as diseased or non-diseased at the end of that time, and then apply the methods previously described. As an alternative, recent approaches propose to estimate sensitivity and specificity as a function of time [15,16]. This allows the sensitivity and specificity of a marker to change over time. For example, there can be markers associated with an acute event that are both sensitive and specific immediately following data collection but not for long-term predictions. Alternatively there can be markers associated with events having a latency period that appear informative for long-term outcomes but are not predictive in the short term. Two approaches for estimating sensitivity and specificity have been proposed. Sensitivity can be estimated using all new cases at time t (incident sensitivity) or using all cases identified up to time t (cumulative sensitivity). The choice of which sensitivity measure to report depends on whether one is interested in disease

incidence or prevalence, respectively. Specificity can be estimated using all individuals who are not cases at time t (dynamic specificity) or using only those individuals who do not become cases after long-term follow up (static specificity). Most recent work has applied the definition of dynamic specificity. Cumulative sensitivity and dynamic specificity are estimated using the Kaplan-Meier estimator [15] and incident sensitivity and dynamic specificity using the Cox model [16].

To evaluate the accuracy of quantitative markers for time-varying outcome, the time-dependent ROC curve has been proposed [15]. ROC curves can be drawn for different follow-up times and the different curves summarized in a plot of the AUC over time. The ROC curve computed using the cumulative sensitivity measure is called the cumulative ROC and the ROC curve computed using the incident sensitivity, the incident ROC. One nice feature of the incident ROC curve is that it permits a global test over all time points of whether a marker is predictive of disease. This is achieved by averaging the AUC over time in a manner that cannot be applied to the analysis of the cumulative ROC [16].

We apply these new concepts of incident/cumulative sensitivity and dynamic specificity to a study of DNA methylation in a cohort of breast cancer patients. MethyLight was used to measure DNA methylation in primary tumor of 54 breast cancer patients who did not receive tamoxifen therapy. The measurements, percent of methylated reference sample (PMR) [9], are quantitative and the outcome is time until death or end of study. Nineteen of the patients (35%) died during follow up. Eight deaths occurred in the first two years (15%) and 16 in the first five years (30%). The PMR value for one gene, dichotomized at its median, predicted overall survival using a Kaplan-Meier analysis (logrank $p = 0.0003$) (Fig. 2). The hazard ratio estimate from a Cox model was 7.2 (95% confidence interval (CI) = 2.1–24.6). Figure 3 shows the cumulative and incident ROC curves 2 and 5 years after diagnosis. Cumulative ROC curves 2 and 5 years after diagnosis show that the AUC increases slightly with time following diagnosis (Fig. 3i & 3ii). Interestingly, the incident ROC curves at 2 and 5 years show AUCs that are similar at the two time points and higher than the AUCs for the cumulative ROC curves (Fig. 3iv & 3v). Plots of the AUC over time are shown in Fig. 3iii and 3vi for the cumulative and incident ROC curves, respectively. Pointwise 95% confidence intervals are computed using 1,000 bootstrap samples. The confidence intervals cover 0.5 (uninformative marker) up to 6 years following diagnosis for the cumulative ROC curve. For the

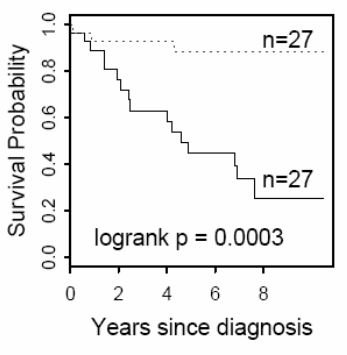

Fig. 2. Kaplan-Meier curves of the breast cancer patients that do not receive tamoxifen therapy (N=54) after dichotomizing the marker at its median value.

incident ROC analysis, the confidence interval does not contain 0.5 suggesting that the marker is somewhat predictive. The average AUC from time of diagnosis until 5 years after follow up is 0.65 (95% CI = 0.52–0.76). The conflicting conclusions from the cumulative and incident ROC analyses suggests that the impressive logrank p-value may not be a good indicator of the markers' accuracy for classification and stresses the importance of evaluating the sensitivity and specificity of a marker when making decisions on accuracy.

3.2. Evaluating a classifier – Combining Markers

When a single marker has inadequate ability to detect the presence or absence of disease (e.g., cancer) it is possible that combining multiple markers could yield a combination that is more accurate than the single marker. Two binary markers can be combined by classifying a subject as diseased if both markers are positive and non-diseased otherwise. This is referred to as the 'believe the negative' (BN) rule [22] or the 'and' rule. The BN rule is more stringent than either marker alone. It decreases FPF and TPF relative to the individual markers, but maintains the TPF of the combination above the sum of the TPFs minus 1. Therefore, this combination strategy is used when both markers have high TPF but also have unacceptably high FPF because the rule decreases FPF while hopefully not reducing TPF very much.

Another approach for combining two binary markers is to consider a subject diseased if either marker is positive. This is referred to as the 'believe the positive' (BP) or the 'or' rule. The BP rule increases TPF relative to the individual markers. It also increases the FPF, but by no more than the sum of the FPFs for the two markers. This combination strategy is used when the

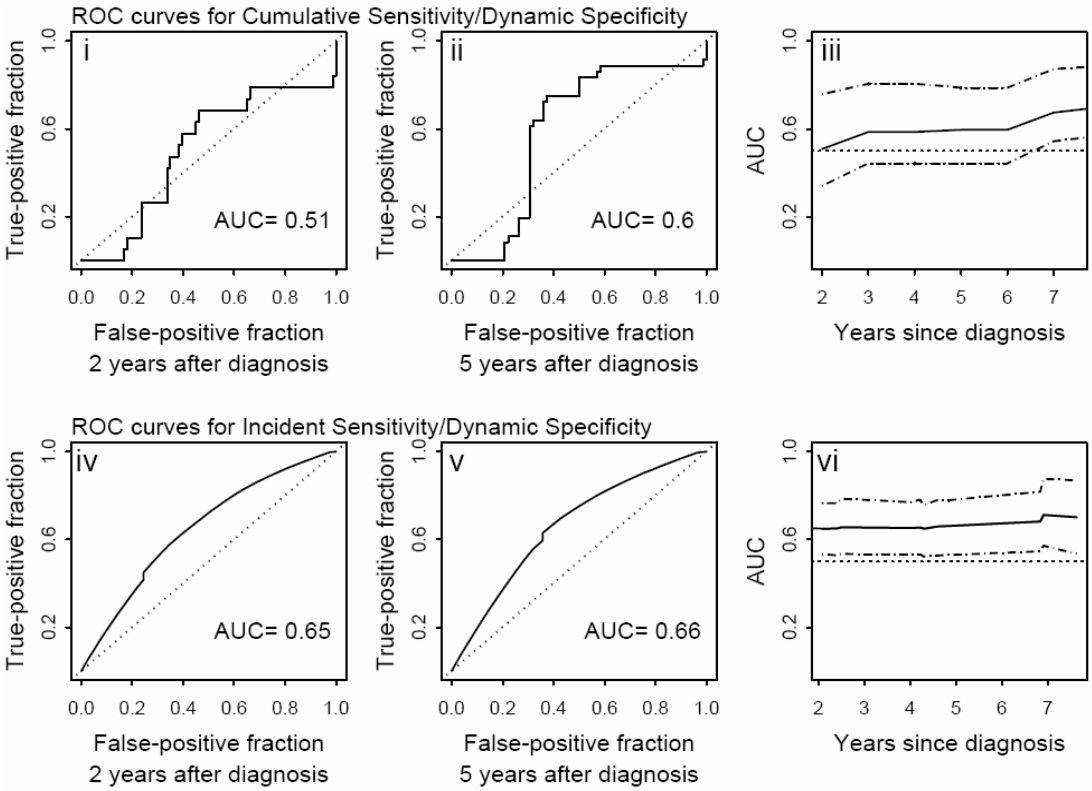

Fig. 3. Cumulative ROC curves (i) 2 years after diagnosis, (ii) 5 years after diagnosis, and (iii) plot of the AUC versus time for cumulative ROC curves (solid line) with 95% pointwise confidence interval (long dashed lines). Incident ROC curves (iv) 2 years after diagnosis, (v) 5 years after diagnosis and (vi) plot of AUC versus time for incident ROC curves (solid line) with 95% pointwise confidence interval (long dashed lines). Confidence intervals are computed using bootstrap resampling.

markers have low FPFs but inadequate TPFs. This second approach was used to evaluate a group of three DNA methylation markers for detecting bladder cancer from urine sediment [13].

Quantitative markers can be combined by creating a risk score to estimate the probability of disease given the marker data. The preferred risk score would optimize the ROC curve, providing the maximum TPF for each FPF and the minimum FPF for each TPF. For a discussion of optimal properties of the risk score see [26]. In practice, many methods can be applied to compute a risk score: logistic regression [4], classification trees [3], logic regression [33], artificial neural networks [32], support vector machines [39] and boosting [12]. These methods fall into the subject area known as supervised learning, where rules are developed to properly classify the diseased and non-diseased individuals. For a description and comparison of the different methods see [14,32].

Supervised learning methods can predict disease status given the marker data; however no single method will be best for all situations. Logistic regression has the desirable properties of being efficient when the model is correct and being robust under model-misspecification [20]. In addition, it can be applied for case-control designs. Still, a recent article found situations in which it was far from estimating the optimal decision rule [30]. In the article, the authors proposed to select the risk score that maximizes the ROC curve directly [30].

Statisticians have been slow to adopt new supervised learning methods for combining markers. This could be due to early comparisons of new methods with standard statistical methods. For example, a paper reporting a review of the literature for applications of artificial neural networks (ANNs) published from 1991 to 1995 found that ANNs were being misused and that they did not perform better than logistic regression for many problems [34]. A recent evaluation of a wide variety of classification methods applied to gene expression data from microarrays found support vector machines to be the most effective classifiers in performing cancer diag-

nosis, outperforming several other supervised learning algorithms [36]. More research is needed to understand the properties of many of these new methods.

3.3. Factors affecting marker accuracy

There are many factors that can affect the performance of a marker. For DNA methylation, age could be such a factor. Other possibilities include other demographic attributes of the subjects tested (e.g., gender, race), characteristics or severity of their disease (e.g., histology and stage of cancer), characteristics of controls (e.g., benign disease or non-diseased), characteristics of testers (e.g., experience, institution), and conditions under which the markers are studied. It is important to identify and understand the influence of these factors because populations and settings where a marker is more or less accurate can be identified, which can be useful in determining how best to use a marker. Assessing the effects of factors on accuracy is also important because study results may not be relevant to populations with different conditions or characteristics. This is referred to as extrapolation bias [26].

Assessing the effects of factors on accuracy can be accomplished using regression analysis. Specifically, binary regression methods can be used to assess effects on the TPF and FPF for binary markers [19]. There are several approaches for quantitative markers (see Chapter 6 of [26]). Alonzo and Pepe [1] propose a ROC regression model where the outcome variable is a binary variable indicating whether or not marker values for diseased individuals are greater than a specified quantile of the distribution of marker values for non-diseased individuals with the same covariate values. Predictor variables include covariates common to the diseased and non-diseased individuals and covariates that are specific to the diseased state (e.g., disease severity). Generalized linear model methods for binary data can be used to perform parameter estimation. Pepe and Cai [28] and Cai [6] propose alternative approaches to parameter estimation.

3.4. Comparing different classifiers

There are different methods for comparing two ROC curves depending on the approach used to estimate the ROC curves. Typically summary indices are used, such as the AUC. Then the hypothesis test is based on the difference in the AUC for the two classifiers. If the two ROC curves are estimated from the same study subjects, then the correlation between the AUC estimates must be taken into account [8]. If there is particular interest in a restricted portion of the ROC curve, then comparisons of the pAUC can be used.

ROC curves can also be compared using ROC regression models by including a covariate or covariates in the model indicating the different markers [24]. ROC regression models have the nice feature that they can compare the accuracy of markers while simultaneously adjusting for other factors that could affect accuracy (see Section 3.3 for examples of factors that can affect accuracy). This reduces the bias for confounding in observational studies and can increase the precision in which accuracy is estimated in experimental studies. Since DNA methylation patterns change with age [31], it is important to control for age effects when comparing the accuracy of markers.

3.5. Training and test sets, cross-validation

Estimates of sensitivity and specificity from a single study tend to be optimistic and may not generalize to future studies. When the main interest is in estimating sensitivity and specificity in future studies, investigators might consider splitting the data into training and test sets. When sample size is not large enough to split a sample, cross-validation is recommended. To perform cross-validation the data set is divided randomly into K subsets of approximately equal size. The model is fit to (K-1) subsets and evaluated for prediction error on the omitted subset (test set). This is repeated K times, each time omitting a different subset. The average error rate from the K test sets estimates the prediction error for the model. Common choices for K are 5 or 10. Smaller data sets may require a larger K to reduce bias in the error estimate. However this comes at the price of higher variability of the estimate.

For models that require the estimation of a tuning parameter, (e.g. estimating the number of hidden nodes in an artificial neural network) cross-validation is typically used to select the parameter value. In such models, the cross-validation error will underestimate the prediction error [34]. Such methods require splitting the data into three subsets: training, validation, and test set. When the sample size is not large enough to split three ways, nested cross-validation can be applied [34, 36].

3.6. Selecting markers for further study

Selecting markers for further study after they have been ranked on predictive ability is a separate chal-

lenge. One strategy is to annotate the marker list using information on gene function or location. A new approach that outperformed 'annotation' in selecting differentially expressed genes that could be reproduced in an independent test set, was to select genes that are highly connected (co-expressed) with other genes [42]. Zhang and Horvath (2005) [42] found that highly connected genes are more likely to be predictive in a follow-up study than genes that are ranked based on statistical significance only but are not connected. Gene clusters have been observed studying DNA methylation in colorectal cancer [40]. This suggests investigating the application of this approach using DNA methylation profiles.

4. Conclusions

Recent technologic and scientific advances have led to the development of new potential markers for early cancer detection or prognosis. We have reviewed methods for evaluating the performance of markers with respect to early detection and prognosis. The methods use two error rates to evaluate classification accuracy: the true-positive fraction and the false-positive fraction. Statistical methods that combine these two errors into a single measure of association, either an odds ratio or relative risk, can be misleading about the marker's classification abilities. Statistical methods have been recently developed to evaluate markers in prospective cohort studies where the subject's disease status can change over time. These methods will allow the assessment of accuracy where subjects have varying lengths of follow-up. Once a marker is shown to have adequate classification ability, the next step is to determine the practical usefulness of the marker in managing patients. An accurate marker may not be useful if subjects are not willing to undergo further work-up and treatment after screening positive, or if treatment is ineffective for disease detected by the marker. Finally, there is a need to develop novel statistical methods to select and combine markers from high-throughput studies where $>$ 5,000 CpG islands are being measured on sample sizes of only 100 individuals. For this we can probably learn from the recent methods developed for microarray studies of gene expression data.

Acknowledgments

This work was supported by NIH grants CA097346 (K.S.) and R01 GM54438 (T.A.).

References

[1] T.A. Alonzo and M.S. Pepe, Distribution-free ROC analysis using binary regression techniques, *Biostatistics* **3** (2002), 421–432.
[2] S.G. Baker, B.S. Kramer and S. Srivastava, Markers for early detection of cancer: statistical guidelines for nested case-control studies, *BMC Med Res Methodol* **2** (2002), 4.
[3] L. Breiman, J. Friedman, R. Olshen and C. Stone, *Classification and Regression Trees*, Wadsworth, Belmont, CA, 1984.
[4] N.E. Breslow and N.E. Day, Statistical methods in cancer research. Vol. 1 – The analysis of case-control studies, IARC Scientific Publications No. 32, Lyon, France, 1980.
[5] M.V. Brock, M. Gou, Y. Akiyama, A. Muller, T.T. Wu, E. Montgomery, M. Deasel, P. Germonpre, L. Rubinson, R.F. Heitmiller, S.C. Yang, A.A. Forastiere, S.B. Baylin and J.G. Herman, Prognostic importance of promoter hypermethylation of multiple genes in esophageal adenocarcinoma, *Clin Cancer Res* **9** (2003), 2912–2919.
[6] T. Cai, Semi-parametric ROC regression analysis with placement values, *Biostatistics* **5** (2004), 45–60.
[7] S.E. Cottrell, Molecular diagnostic applications of DNA methylation technology, *Clin Biochem* **37** (2004), 595–604.
[8] E.R. DeLong, D.M. DeLong and D.L. Clarke-Pearson, Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach, *Biometrics* **44** (1988), 837–845.
[9] C.A. Eads, K.D. Danenberg, K. Kawakami, L.B. Saltz, C. Blake, D. Shibata, P.V. Danenberg and P.W. Laird, MethyLight: a high-throughput assay to measure DNA methylation, *Nucleic Acids Res* **28** (2000), E32.
[10] B. Emir, S. Wieand, J.Q. Su and S. Cha, Analysis of repeated markers used to predict progression of cancer, *Stat Med* **17** (1998), 2563–2578.
[11] M. Esteller, G. Gaidano, S.N. Goodman, V. Zagonel, D. Capello, B. Botto, D. Rossi, A. Gloghini, U. Vitolo, A. Carbone, S.B. Baylin and J.G. Herman, Hypermethylation of the DNA repair gene O(6)-methylguanine DNA methyltransferase and survival of patients with diffuse large B-cell lymphoma, *J Natl Cancer Inst* **94** (2002), 26–32.
[12] Y. Freund and R. Shapire, A decision-theoretic generalization of online learning and an application to boosting, *Journal of Computer and System Sciences* **55** (1997), 119–139.
[13] M.G. Friedrich, D.J. Weisenberger, J.C. Cheng, S. Chandrasoma, K.D. Siegmund, M.L. Gonzalgo, M.I. Toma, H. Huland, C. Yoo, Y.C. Tsai, P.W. Nichols, B.H. Bochner, P.A. Jones and G. Liang, Detection of methylated apoptosis-associated genes in urine sediments of bladder cancer patients, *Clin Cancer Res* **10** (2004), 7457–7465.
[14] T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning*, Springer-Verlag, New York, 2001.
[15] P.J. Heagerty, T. Lumley and M.S. Pepe, Time-dependent ROC curves for censored survival data and a diagnostic marker, *Biometrics* **56** (2000), 337–344.
[16] P.J. Heagerty and Y. Zheng, Survival model predictive accuracy and ROC curves, *Biometrics* **61** (2005), 92–105.
[17] P.W. Laird, Oncogenic mechanisms mediated by DNA methylation, *Mol Med Today* **3** (1997), 223–229.
[18] P.W. Laird, The power and the promise of DNA methylation markers, *Nature reviews* **3** (2003), 253–266.
[19] W. Leisenring, M.S. Pepe and G. Longton, A marginal regression modelling framework for evaluating medical diagnostic tests, *Stat Med* **16** (1997), 1263–1281.

[20] K.-C. Li and N. Duan, Regression analysis under link violation, *Annals of Statistics* **17** (1989), 1009–1052.
[21] S. Maier, C. Dahlstroem, C. Haefliger, A. Plum and C. Piepenbrock, Identifying DNA methylation biomarkers of cancer drug response, *Am J Pharmacogenomics* **5** (2005), 223–232.
[22] R.J. Marshall, The predictive value of simple rules for combining two diagnostic tests, *Biometrics* **45** (1989), 1213–1222.
[23] D.K. McClish, Analyzing a portion of the ROC curve, *Med Decis Making* **9** (1989), 190–195.
[24] M.S. Pepe, A regression modelling framework for receiver operating characteristic curves in medical diagnostic testing, *Biometrika* **84** (1997), 595–608.
[25] M.S. Pepe, R. Etzioni, Z. Feng, J.D. Potter, M.L. Thompson, M. Thornquist, M. Winget and Y. Yasui, Phases of biomarker development for early detection of cancer, *J Natl Cancer Inst* **93** (2001), 1054–1061.
[26] M.S. Pepe, *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford University Press, Oxford, United Kingdom, 2003.
[27] M.S. Pepe, G. Longton, G.L. Anderson and M. Schummer, Selecting differentially expressed genes from microarray experiments, *Biometrics* **59** (2003), 133–142.
[28] M.S. Pepe and T. Cai, The analysis of placement values for evaluating discriminatory measures, *Biometrics* **60** (2004), 528–535.
[29] M.S. Pepe, H. Janes, G. Longton, W. Leisenring and P. Newcomb, Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker, *Am J Epidemiol* **159** (2004), 882–890.
[30] M.S. Pepe, T. Cai and G. Longton, Combining predictors for classification using the area under the receiver operating characteristic curve, *Biometrics* (2005).
[31] B. Richardson, Impact of aging on DNA methylation, *Ageing Res Rev* **2** (2003), 245–261.
[32] B.D. Ripley, *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge, United Kingdom, 1996.
[33] I. Ruczinski, C. Kooperberg and M. LeBlanc, Logic Regression, *Journal of Computational and Graphical Statistics* **12** (2003), 475–511.
[34] G. Schwarzer, W. Vach and M. Schumacher, On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology, *Stat Med* **19** (2000), 541–561.
[35] K.D. Siegmund and P.W. Laird, Analysis of complex methylation data, *Methods* **27** (2002), 170–178.
[36] A. Statnikov, C.F. Aliferis, I. Tsamardinos, D. Hardin and S. Levy, A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis, *Bioinformatics* **21** (2005), 631–643.
[37] Y. Tada, M. Wada, K. Taguchi, Y. Mochida, N. Kinugawa, M. Tsuneyoshi, S. Naito and M. Kuwano, The association of death-associated protein kinase hypermethylation with early recurrence in superficial bladder cancers, *Cancer Res* **62** (2002), 4048–4053.
[38] M. Van Rijnsoever, H. Elsaleh, D. Joseph, K. McCaul and B. Iacopetta, CpG island methylator phenotype is an independent predictor of survival benefit from 5-fluorouracil in stage III colorectal cancer, *Clin Cancer Res* **9** (2003), 2898–2903.
[39] V. Vapnik, *The Nature of Statistical Leearning Theory*, Springer-Verlag, New York, 1996.
[40] D.J. Weisenberger, K. Siegmund, M. Campan, J. Young, T.I. Long, M.A. Faasse, G.H. Kang, M. Widschwendter, D. Frieswyk, A.J. Levine, J. Jass, R. Haile and P.W. Laird, CpG Island Methylator Phenotype in Human Colorectal Cancer is Tightly Associated with BRAF Mutation and Underlies Sporadic Mismatch Repair Deficiency, *Nature Genetics* **38**(7) (2006), 787–793.
[41] M. Widschwendter, K.D. Siegmund, H.M. Muller, H. Fiegl, C. Marth, E. Muller-Holzner, P.A. Jones and P.W. Laird, Association of breast cancer DNA methylation profiles with hormone receptor status and response to tamoxifen, *Cancer Res* **64** (2004), 3807–3813.
[42] B. Zhang and S. Horvath, A general framework for weighted gene co-expression network analysis, *Statistical Applications in Genetics and Molecular Biology* **4** (2005), 1–43.