TIME-DEPENDENT ROC CURVES IN GENETICALLY DETERMINED SUBGROUPS*

Alberto Morabito[1], Emanuela Morenghi[1], Monica Ferraroni[1], Giovanni Radaelli[1] and Fabio Macciardi[2]

[1] *Dipartimento di Medicina Chirurgia e Odontoiatria, Università di Milano, Via Di Rudinì, 8, 20144 Milano.*

[2] *Dipartimento di Biologia e Genetica per le Scienze Mediche, Università di Milano*

Abstract

The ultimate goal of genetic association studies is to identify and map the gene(s) responsible for a given disease. This paper discusses a new and simple statistical method for detecting a genetic association, based on time-dependent Receiver Operating Characteristic (ROC) curves. This method resorts to the Heagerty approach based on Bayes theorem, and uses the Kaplan-Meier or the Akritas estimator. An application to the real problem of examining possible interaction between glycaemia and a risk "genotype" on survival is presented using the Framingham database.

Analysis assessed area of chromosome 1 (from 192 to 233 cM) and evaluated the role of fasting blood glucose on survival, at 4 and 8 years of follow-up, according to the presence/absence of allele 242 (marker 23). The allele 242 showed ability in predicting survival. Kaplan-Meier and Akritas estimators provided comparable results. At 4 years of follow-up, area (SD) under ROC curve, in absence of allele 242, was 0.85 (0.09) and 0.81 (0.07) using respectively the Kaplan-Meier or Akritas estimator.

The ability showed by the time-dependent ROC curves in real data suggests that the proposed method may be valuable to detect difference in genetic subgroups. Further studies need to better clarify the usefulness of this method in other real applications.

1. INTRODUCTION

The ultimate goal of genetic association studies, which should be viewed within the larger framework of epidemiological studies of risk factor/disease associations, is to identify and map the gene(s) responsible for a given disease.

---

* *Presented at "II convegno nazionale – SISMEC" Brescia 1-4 ottobre 2003*

Criteria for identifying associations rely on detecting the alleles of the assessed candidate or marker genes preferentially transmitted with the disease. This occurrence is interpreted as an indirect proof that a still unidentified allele of the locus responsible for the disease co-segregates with the corresponding associated allele of the marker, giving rise to a haplotype. Under this assumption, the association analysis is based on the existence of linkage disequilibrium. A simple design to examine a genetic association is based on the case-control approach.

To assess the significance of a potential association, genetic association studies have traditionally used a simple chi-square analysis, evaluating whether one or more alleles of a given gene ("marker" gene(s) or "candidate" gene(s)) are more represented in one subset (e.g., cases) rather than in another (e.g., controls). In this case, the strength of association of the marker gene with the hypothetical susceptible gene for the disorder increases as physical and genetic distances decrease. We propose a new method based on time-dependent Receiver Operating Characteristic (ROC) curves. This method is simple and quite straightforward, the main limitation being its complexity when applied to family-based designs. However, its sensitivity may be satisfactory assessed using real data. Thus, we have examined data from the Framingham Heart Study (GAW13, 2002). For simplicity, we considered as case/control the presence/absence of an allele, and included in the analysis both the time of observation and potential risk factors or confounders. The relationship between fasting blood glucose and survival with respect to presence/absence of a allele of a putative candidate locus was assessed.

2. ROC CURVES

Let X denote a continuous real variable related to a diagnostic test or prognostic index, with higher values more suggestive of disease. Let D be a binary indicator of disease status. The ROC curve for X plots sensitivity associated with the dichotomized test $X > c$ versus (1 - specificity), for all possible threshold values c, i.e., the ROC curve is the monotone implicit function in [0,1], $\{(P(X > c \mid D = 0), P(X > c \mid D = 1)), c \in (-\infty, \infty)\}$. This function characterizes the diagnostic potential of a continuous test by summarizing all possible trade-offs between sensitivity and specificity. The higher the ROC curve is in the quadrant $[0,1] \times [0,1]$, the better is its capacity in discriminating diseased and non-diseased subjects.

In diagnostic medicine, ROC curves have several attractive features.

1. ROC curves describe the inherent discrimination capacity of a test without linking it to any specific threshold.

2. ROC curves are particularly useful in comparing discriminatory capacity of different diagnostic indexes. They provide a valid approach even when diagnostic indexes are on different measurement scales.
3. Area under the ROC curve equals the probability that the test result from a diseased individual exceeds the test result from a non diseased individual, and is often used to summarize the ROC curve.

A review on theory of ROC curves can be found in Thompson and Zucchini (1989) and Begg (1991).

ROC curves for continuous diagnostic tests can be estimated non-parametrically using empirical estimates of the survivor functions, $S_0(c) = P(X > c \mid D = 0)$ and $S_1(c) = P(X > c \mid D = 1)$.

3. THE HEAGERTY'S APPROACH

Disease status may be considered a time-dependent characteristic. A subject initially non diseased can die from disease during the study. Heagerty et al. (2000) outlined an approach for the estimation of ROC curves with a time-dependent disease variable, or more generally a failure time, for data obtained in prospective cohort studies. The approach is based on direct use of Bayes theorem with the Kaplan-Meier or the Akritas estimator (Akritas, 1994) .

Let T_i denote the failure time, C_i the censoring time and $Z_i = \min(T_i, C_i)$ the follow-up time. Let δ_i denote a censoring indicator, $\delta_i = 1$ if $T_i \leq C_i$ and $\delta_i = 0$ if $T_i > C_i$. We resort to a counting process defined by $D_i(t) = 1$ if $T_i \leq t$, and $D_i(t) = 0$ if $T_i > t$. $D_i(t) = 1$ indicates subject i having developed failure before time t.

Let X_i indicate a covariate value for subject i. Recall that ROC curves display the relationship between a covariate X_i and a binary variable D_i, by plotting estimates of sensitivity, $P(X > c \mid D = 1)$, versus one minus the specificity, $1 - P(X \leq c \mid D = 0)$, for all possible values c. When disease status is time dependent, both sensitivity (Se) and specificity (Sp) are time-dependent functions:

$$Se(c,t) = P\{X > c \mid D(t) = 1\}$$

$$Sp(c,t) = P\{X \leq c \mid D(t) = 0\}, \quad t > 0.$$

Using these settings, we can define the corresponding ROC curve for any time t, ROC(t).

We use Bayes' theorem to rewrite specificity and sensitivity.

$$Sp(c,t) = P(X \le c \mid D(t) = 0) = \frac{P(D(t) = 0 \mid X \le c)P(X \le c)}{P(D(t) = 0)} = \frac{S(t \mid X \le c)P(X \le c)}{S(t)}$$

$$Se(c,t) = P(X > c \mid D(t) = 1) = \frac{P(D(t) = 1 \mid X > c)P(X > c)}{P(D(t) = 1)} = \frac{(1 - S(t \mid X > c))P(X > c)}{1 - S(t)},$$

where S(t) is the survival function P(T > t) and S(t | X > c) the conditional survival function, for the subset defined by X > c.

A widely used nonparametric estimate of S(t) is obtained by the Kaplan-Meier method. Let τ_n denote the subset of values of Z_i for observed events, $\delta_i = 1$. The Kaplan-Meier (KM) estimator is defined by

$$\hat{S}_{KM}(t) = \prod_{\substack{s \in \tau_n \\ s \le t}} \left(1 - \frac{\sum_j I_{(Z_j = s)} \delta_j}{\sum_j I_{(Z_j \ge s)}} \right),$$

where I is the indicator function.

A simple estimator for sensitivity and specificity at time t is provided by combining the Kaplan-Meier estimator and the empirical distribution function of the covariate X, i.e.

$$\hat{P}_{KM}(X > c \mid D(t) = 1) = \frac{(1 - \hat{S}_{KM}(t \mid X > c))(1 - \hat{F}_X(c))}{1 - \hat{S}_{KM}(t)}$$

$$\hat{P}_{KM}(X \le c \mid D(t) = 0) = \frac{\hat{S}_{KM}(t \mid X \le c)\hat{F}_X(c)}{\hat{S}_{KM}(t)},$$

where $\hat{F}_X(c) = \frac{1}{n} \sum I_{(X_i \le c)}$.

Since the KM estimator doesn't guarantee the monotonicity of sensitivity and specificity functions, the Akritas smoothed formula (Heagerty, 2000) may be a valid alternative. The Akritas estimator is

$$\hat{S}_{\lambda_n}(c,t) = \frac{1}{n} \sum_i \hat{S}_{\lambda_n}(t \mid X = X_i) I_{(X_i > c)},$$

where $\hat{S}_{\lambda_n}(t \mid X = X_i)$ is a weighted Kaplan-Meier estimator and λ_n the smoothing parameter, indicating the half-width of the inter-quantile interval, centered on X_i.

To compare different ROC curves, Heagerty et al. (2000) suggested to provide confidence intervals using bootstrap. The computational time for a single test and the need of replication over time discourage to use this procedure in large data sets, while distribution free methods may be preferable (Kesler, 2001).

4. APPLICATION TO REAL DATA

High blood glucose levels, determined by a complex interplay between pancreatic function and responsiveness to insulin, are a major indicator of type 2 diabetes (Albarrak et al, 2002). Current incomplete understanding of the molecular basis for this syndrome complicates the identification of any hypothetical etiological gene. The analysis of Meigs et al. (2002) of the Framingham Heart Study data suggests that Quantitative Trait Loci influencing glucose homeostasis may be located on chromosomes 1q and 10q. Meigs also suggested that the ~187-218 cM area on chromosome 1 seems to be the most interesting to look for type 2 diabetes susceptibility genes. Accordingly, we considered only the area of chromosome 1, in the interval 192 - 233 cM.

We applied time-dependent ROC curves (Heagerty, 2000) to the Framingham Heart Study data. In order to assess the sensitivity of this approach, we analyzed the survival status (D) using the fasting blood glucose levels as time dependent covariate (X), according to the presence/absence of a allele located in a putative susceptible chromosomal region for type 2 diabetes.

The original Framingham database consisted of 4692 persons, 2348 males and 2344 females, 1213 belonging to original cohort, 1672 to offspring cohort and 1807 founders. Since it would be difficult to consider the familial structure in the analysis, and persons were neither in "original" cohort nor in "offspring" cohort, we excluded from the analysis the non-informative subjects, i.e. persons who never underwent any control. After excluding these subjects we worked on a database of 2885 persons, 1409 males and 1476 females, 1213 belonging to original cohort and 1672 to offspring cohort, and 719 dead.

The population was categorized in two subgroups, according to presence/ absence about a specific "risk" allele in the region of chromosome 1. In particular, our analysis of the Framingham population pointed to the allele 242 of marker number 23 as potential survival predictor related to blood glucose. Six hundred and thirty-nine subjects having missing information about the allele were not included

in the following analyses. The number of subjects can differ over time because the covariate measurement may be missing and the number of subjects at the follow-up time may vary.

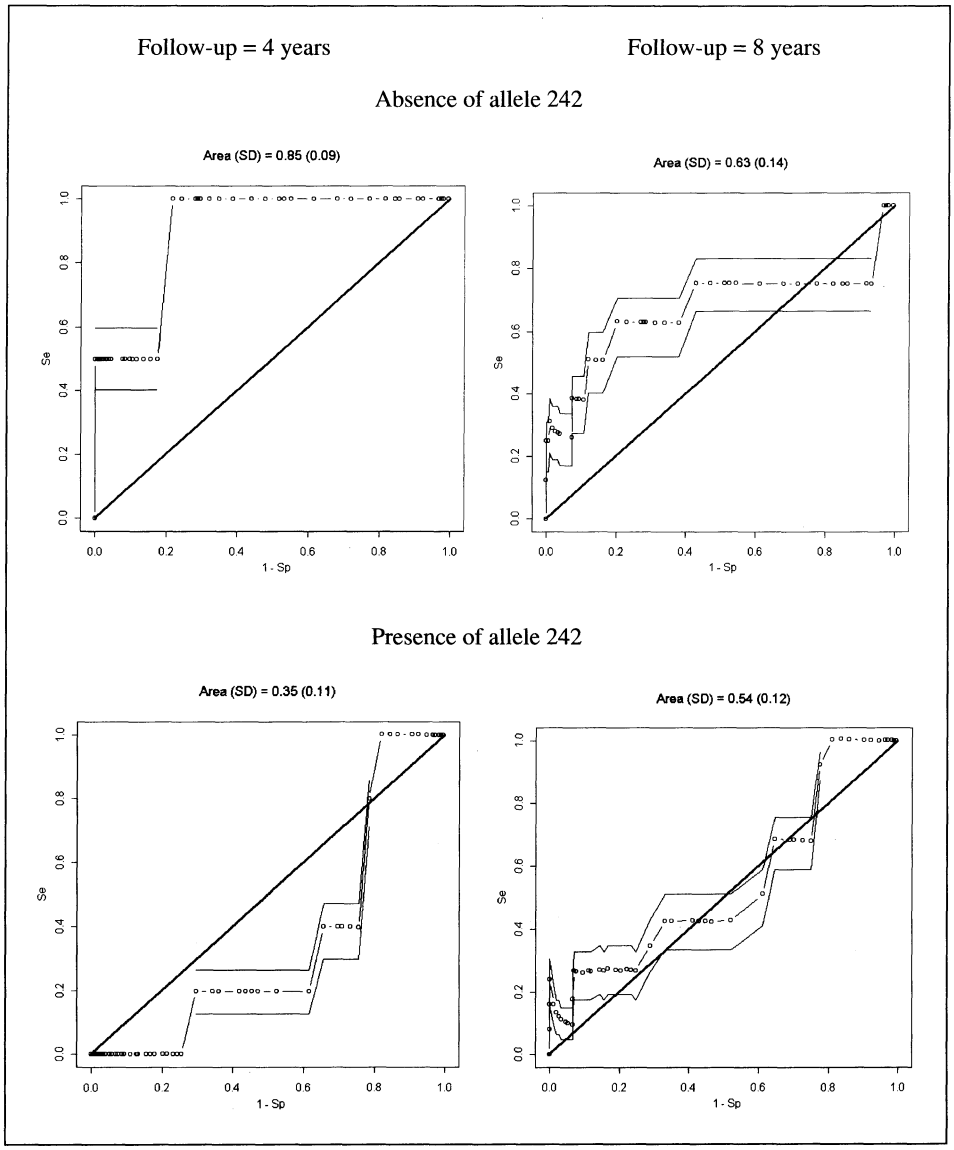

Fig. 1: ROC analysis of the role of fasting blood glucose levels on survival, according to presence/absence of the allele 242 (marker 23), using Kaplan-Meier method.

Figure 1 and 2 show results of the analysis, at 4 and 8 years of follow-up, according to the absence/presence of the allele 242 (marker 23), using, respectively, Kaplan-Meier or Akritas estimator.

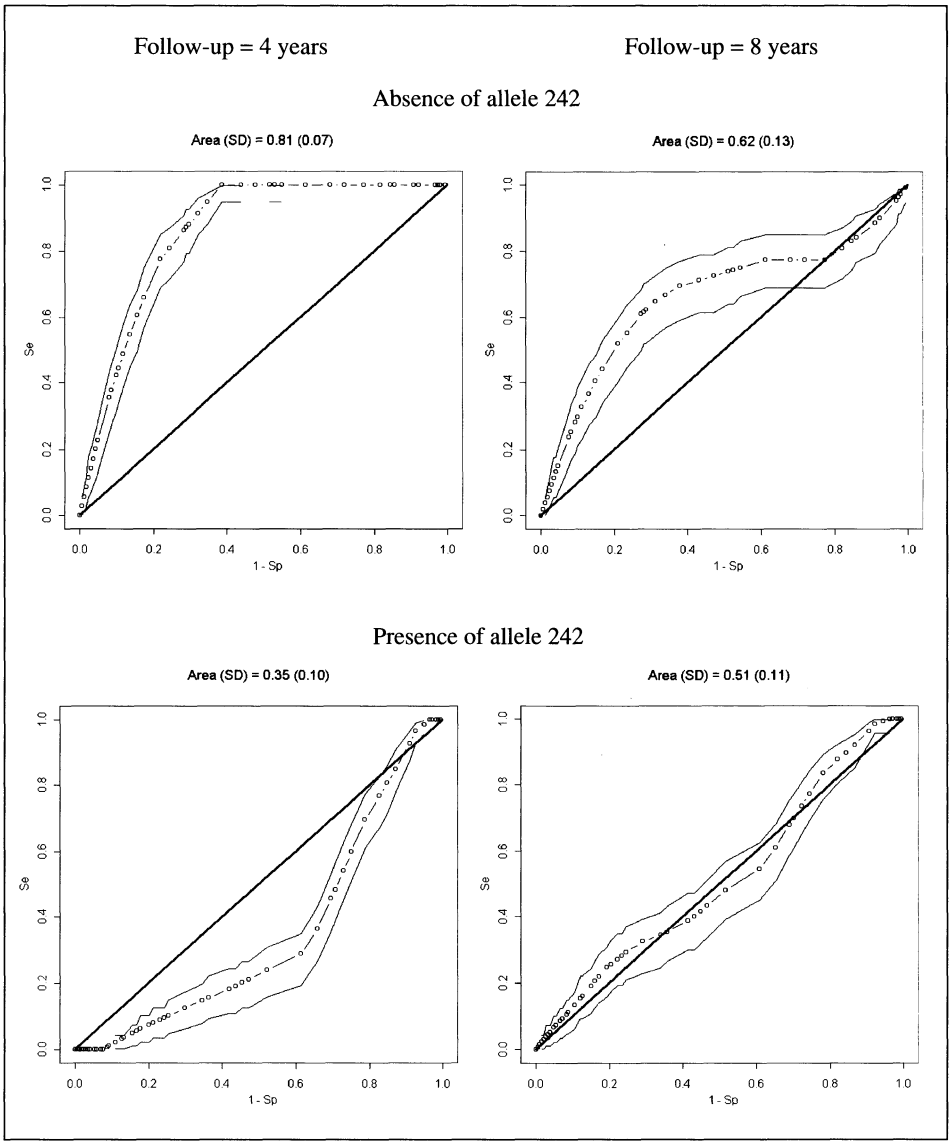

Fig. 2: ROC analysis of the role of fasting blood glucose levels on survival, according to presence/ absence of the allele 242 (marker 23), using Akritas method.

Each ROC curve displays at different follow-up times (years): for all threshold levels c of blood glucose: (Se) the estimated proportion of subjects surviving at different follow up times having blood glucose > c, and (1-Sp) the estimated proportion of subjects surviving at different follow up times having blood glucose ≤ c. The upper and lower continuous bands represent the 90th and 10th percentile, respectively. A ROC curve above (below) the diagonal represents an increasing (decreasing) survival, as function of increasing values of the blood glucose.

At 4 years of follow-up, area (SD) under the ROC curve, in absence of allele 242, was 0.85 (0.09) and 0.81 (0.07) when calculated by the Kaplan-Meier or Akritas estimators respectively. At 8 years of follow-up the corresponding values were 0.35 (0.11) and 0.35 (0.10) respectively.

5. CONCLUSIONS

Time-dependent ROC curves may be profitably used in real situations to detect difference in genetic subgroups. Attractiveness of the method may include its semplicity and ability in measuring interaction between a disease status and a prognostic factor. Sensitivity and 1-specificity, as conditional survival in diseased and not diseased subjects, allows for a direct "visual" comparison between genetic subgroups. Moreover, the present findings suggest that in real applications both Kaplan-Meier and Akritas estimators provide comparable results. Akritas estimator provides a reliable theoretical approach and more appealing graphical representation. However, Kaplan-Meier estimator might be preferable in pratical applications because it is simpler to implement and moreover requires shorter computational time. Further studies are needed to better clarify the usefulness of the method in other real data applications.

REFERENCES

AKRITAS MG (1994) Nearest neighbor estimation of a bivariate distribution under random censoring. *Annals of Statistic*, **22**, 1299-1327.

ALBARRAK AI, LUZIO SD, CHASSIN LJ, PLAYLE RA, OWENS DR AND HOVOROK (2002) Associations of glucose control with insulin sensitivity and pancreatic beta-cell responsiveness in newly presenting type diabetes. *Journal of Clinical Endocrinology and Metabolism*, **87**, 198-203.

BEGG CB (1991) Advances in statistical methodology for diagnostic medicine in the 1980's. *Statistics in Medicine*, **10**, 1887-1895.

GAW13 – Genetic Analysis Workshop 13 November 11-14, New Orleans, LA, USA 2002: data kindly made available within this framework.

HEAGERTY PJ, LUMLEY T, PEPE MS (2000) Time–dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*, **56**, 337-344.

KESTLER HA (2001) ROC with confidence - a Perl program for receiver operator characteristic curves. *Computer Methods and Programs in Biomedicine*, **64**, 133-136.

MEIGS JB, PANHUYSEN CI, MYERS RH, WILSON PW, CUPPLES LA (2002) A genome-wide scan for loci linked to plasma levels of glucose and HbA (1c) in a community-based sample of Caucasian pedigrees: The Framingham Offspring Study. *Diabetes*, **51**, 833-840.

THOMPSON ML, ZUCCHINI W (1989) On the statistical analysis of ROC curves. *Statistics in Medicine*, **8**, 1277-1290.

CURVE ROC DIPENDENTI DAL TEMPO IN SOTTOGRUPPI GENETICAMENTE DETERMINATI

Riassunto

Lo scopo principale degli studi di associazione genetica è identificare i geni responsabili di una data malattia. Questo articolo discute un metodo statistico nuovo e semplice, basato sulle curve Receiver Operating Characteristic (ROC) dipendenti dal tempo, per valuare l'esistenza di associazioni genetiche. Seguendo l'approccio di Heagerty l'analisi è effettuata tramite lo stimatore di Kaplan-Meier o di Akritas. Il metodo è applicato al problema reale della possibile associazione tra glicemia basale e genotipi di "rischio" per la sopravivenza usando i dati dello studio di Framingham.

In particolare è esaminata un' area del cromosoma 1 (da 192 a 233 cM), ed è valutata la sopravvivenza a 4 e 8 anni di follow up rispetto alla presenza dell' allele 242 del marker 23. I valori ricavati dagli stimatori di Kaplan-Meier e di Akritas sono comparabili. A 4 anni di follow up l' area (SD) sottesa dalla curva ROC, in assenza dell' allele 242, è uguale a 0.85 (0.09) e 0.81 (0.07) rispettivamente. I risultati suggeriscono che il metodo proposto potrebbe essere utile per lo studio di associazioni genetiche.