

A Simple Modeling-free Method Provides Accurate Estimates of Sensitivity and Specificity of Longitudinal Disease Biomarkers

F. Subtil^{1, 2, 3, 4}, C. Pouteil-Noble^{2, 3, 5}, S. Toussaint^{2, 3, 5}, E. Villar^{2, 3, 5}, M. Rabilloud^{1, 2, 3, 4}

¹Hospices Civils de Lyon, Service de Biostatistiques, Lyon, France;

²Université de Lyon, Lyon, France;

³Université Lyon 1, Villeurbanne, France;

⁴CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, Equipe Biostatistique Santé, Pierre-Bénite, France;

⁵Hospices Civils de Lyon, Service de Néphrologie-Transplantation, Centre Hospitalier Lyon-Sud, Pierre-Bénite, France

Keywords

Sensitivity and specificity, prognosis, early diagnosis, longitudinal study, biological markers

Summary

Objective: To assess the time-dependent accuracy of a continuous longitudinal biomarker used as a test for early diagnosis or prognosis.

Methods: A method for accuracy assessment is proposed taking into account the marker measurement time and the delay between marker measurement and outcome. It dealt with markers having interval-censored measurements and a detection threshold. The threshold crossing times were assessed by a Bayesian method. A numerical study was conducted to test the procedures that were later applied to PCR measurements for prediction

of cytomegalovirus disease after renal transplantation.

Results: The Bayesian method corrected the bias induced by interval-censored measurements on sensitivity estimates, with corrections from 0.07 to 0.3. In the application to cytomegalovirus disease, the Bayesian method estimated the area under the ROC curve to be over 75% during the first 20 days after graft and within five days between marker measurement and disease onset. However, the accuracy decreased quickly as that delay increased and late after graft.

Conclusions: The proposed Bayesian method is easy to implement for assessing the time-dependent accuracy of a longitudinal biomarker and gives unbiased results under some conditions.

nostic value of such longitudinal clinical biomarkers has to be carefully assessed and analyzed [8, 9]. For a clinician, a biomarker is useful if it has a good discriminant accuracy and if its test becomes positive early enough to allow an efficient reaction between marker measurement and the disease clinical manifestation. Thus, the progression of a biomarker's accuracy along the delay from marker measurement and disease onset is of major interest. A marker load may also vary along the time elapsed since inclusion of a patient into a study regardless of the progression toward disease. Consequently, accuracy analyses should take into account both the marker measurement time and the delay between marker measurement and disease onset.

When a marker is measured with the disease present, it is conventional to use a ROC curve to summarize the accuracy of continuous or ordinal tests [9–12]. That curve displays the relationship between sensitivity (true-positive rate) and 1-specificity (false-positive rate) across all possible threshold values set for that test. The test accuracy is then measured by the area under the ROC curve (AUC). This area, comprised between 0 and 1, may be interpreted as the probability that the diagnostic test result in a diseased subject exceeds that result in a non-diseased one (for a complete review of classical diagnostic methods, see Pepe [3] and Zhou et al. [13]).

Recently, several methods have been proposed to assess the time-dependent accuracy of a biomarker when the measurements are repeated before disease onset [14–20]. A first approach consists in modeling semi-parametrically the time-dependent sensitivity and

Correspondence to:

Fabien Subtil
Hospices Civils de Lyon – Service de Biostatistique
162 avenue Lacassagne
69003 Lyon
France
E mail: fabien.subtil@chu-lyon.fr

Methods Inf Med 2009; 48: 299–305

doi: 10.3414/ME0583

received: July 1, 2008

accepted: December 12, 2008

prepublished: March 31, 2009

1. Introduction

Today, disease diagnosis is made not only on traditional clinical observations, but also on laboratory results; for example, fluorescence polarization, a measure of cellular functionality, is used to make the diagnosis of breast cancer [1]. Methods have been developed to use those results as diagnostic tests and to compare their accuracies [2–4]. Mo-

lecular biology has also contributed to the improvement of early diagnosis or prognosis of diseases. Recent research fields, as in genomics or proteomics, led to the development of numerous biomarkers for early diagnosis or prognosis [5, 6]. During patient follow-up, it became frequent to collect repeated measurements of a quantitative biomarker such as the CA19-9 antigen in screening for recurrence of colorectal cancer [7]. The prog-

specificity or the ROC curve itself [16, 17]; the model's validity may be checked with methods proposed by Cai and Zheng [21]. A second approach models survival conditional on the marker values [18–20]. A third approach models the marker distribution conditional on the disease status [14, 15]. In each of the previous models, effects related to marker measurement time and to the delay between marker measurement and outcome are introduced. In their comprehensive and very instructive review on the subject, Pepe et al. [22] recommended sensitivity be assessed on events that occur exactly t days after marker measurement (incident sensitivity) and not over a delay following the measurement (cumulative sensitivity). Also, they recommended specificity be evaluated in subjects with follow-up long enough to be considered as subjects who will not develop the disease (static specificity). Five out of the six above-mentioned methods [14–18, 20] use this definition of time-dependent accuracy. However, those methods require sophisticated models that are not currently available in standard statistical softwares.

Considering those facts, we developed a simple method to assess the time-dependent accuracy of a longitudinal biomarker using a Bayesian approach. In agreement with the recommendations of Pepe et al., that method takes into account interval-censored measurements and, possibly, biomarkers with a detection threshold.

The first section of the present article describes the method. Numerical studies were conducted in order to compare the results obtained with and without consideration of the sparse nature of the measurements. The method is also illustrated by an analysis of data stemming from a clinical study where patients were screened by PCR measurements to predict cytomegalovirus (CMV) disease after renal transplantation.

2. Methods

2.1 Time-dependent Accuracy Definition

Heagerty and Zheng [23] have proposed several ways to integrate time into ROC analysis according to how “cases” and “controls” are defined. As recommended by Pepe et al. [22], the

incident sensitivity definition was used here [23]: cases correspond to patients who develop the disease exactly t time units after marker measurement. Thus, for a t time units delay between marker measurement and outcome, sensitivity is estimated with measurements taken exactly t time units before the outcome. Sensitivity is assessed at different delays t to assess its progression along the delay from marker measurement to outcome. In this article, a positive test is defined as a marker value higher than or equal to a certain threshold (though equal or lower values may be elsewhere considered). If $Y_i(s)$ denotes a measurement relative to patient i at time s since his inclusion into the study, and T_i the event onset time, the incident sensitivity for a delay t between marker measurement and outcome and for a threshold c may be formalized as:

$$\text{Sensitivity}(c, t) = P[Y_i(s) \geq c \mid T_i - s = t]$$

The progression of sensitivity along t reflects the test ability of early prediction of the outcome.

Controls are defined as subjects who do not develop the disease τ days after inclusion into the study, τ being a fixed delay, long enough to consider as controls patients who will probably never develop the disease. Specificity is estimated using measurements in those patients, which leads to static specificity estimates. A possible progression of specificity after inclusion may be taken into account by estimating specificity using, in the controls, the measurements taken at different periods after inclusion. For each subject of the control group, the highest measurement obtained during the period $[s_j, s_{j+1}]$ is kept, s_j and s_{j+1} denoting successive times since inclusion. The definition of specificity may be formalized as:

$$\text{Specificity}(c, \tau, s_j, s_{j+1}) = P\left[\max_{s_j \leq s < s_{j+1}} (Y_i(s)) < c \mid T_i > \tau\right]$$

2.2 Time-dependent Accuracy Estimation

Estimating incident sensitivity requires that a marker measurement be taken exactly t days before the onset of the disease in each subject who developed that disease, which is not the

case in most studies. A first method, called the crude method, consists in using for each cases the last value obtained before $T_i - t$, introducing a bias because the delay between marker measurement and $T_i - t$ might vary widely from one patient to another.

Because of measurements sparsity, a marker threshold value is often crossed between two dates; this leads to “interval-censored data” [24]. For example, for each couple of measurements, the crude method supposes that the marker value was Y_i at time t_i and Y_j at time t_j , whereas Y_j was actually reached and crossed during interval $]t_i; t_j]$. Biomarkers with a detection threshold raise similar issues. All that can be known is that the biomarker has crossed the detection threshold between two dates.

One way to deal with interval-censored measurements is to estimate the exact threshold crossing times using a Bayesian method with non-informative priors and assuming that, for a given threshold, the crossing times of all patients who crossed it follow a Weibull distribution. The Weibull distribution was chosen because it is commonly used to model times to event, in particular failure times, but other positive distributions can be used if appropriate. The moment at which each observed marker value is crossed by each patient can be estimated. Unlike the crude method, that Bayesian method uses all the information contained in interval-censored data or measurements below a detection threshold. Then, in patients who develop the disease, the most recent threshold value crossed at $T_i - t$ is used as a diagnostic test for ROC analysis. In patients who do not develop the disease, the diagnostic test used is the highest threshold value crossed between s_j and s_{j+1} obtained using the Bayesian method.

3. Simulation Study

3.1 Numerical Studies

Numerical studies were carried out to compare the results obtained with the crude method to those obtained with the Bayesian method. Let us consider 200 subjects who developed a given disease at time T_i , and 100 subjects who did not develop that disease. Marker measurements were considered throughout a follow-up duration that did not

Table 1

Estimated mean AUC values and sensitivities for thresholds 1, 2, 3, and 4, with their respective standard errors, obtained with the Bayesian method and the crude method over 100 simulations, for three delays between marker measurement and disease outcome

Delay	Method	AUC	Se 1	Se 2	Se 3	Se 4
2	Theoretical	0.985	0.999	0.971	0.787	0.378
	Bayesian	0.868 (0.037)	0.959 (0.027)	0.842 (0.045)	0.617 (0.056)	0.312 (0.050)
	Crude	0.697 (0.029)	0.885 (0.026)	0.610 (0.029)	0.284 (0.038)	0.085 (0.043)
4	Theoretical	0.791	0.871	0.5	0.129	0.012
	Bayesian	0.616 (0.031)	0.838 (0.025)	0.509 (0.038)	0.175 (0.048)	0.026 (0.029)
	Crude	0.458 (0.045)	0.717 (0.030)	0.299 (0.049)	0.060 (0.055)	0.012 (0.041)
6	Theoretical	0.618	0.664	0.233	0.03	0.001
	Bayesian	0.418 (0.048)	0.682 (0.037)	0.253 (0.055)	0.043 (0.051)	0.006 (0.025)
	Crude	0.342 (0.049)	0.578 (0.041)	0.176 (0.051)	0.027 (0.045)	0.007 (0.032)

True AUCs and sensitivities were estimated according to process of generation of the biomarker values. Se denotes sensitivity.

exceed 30 days. High marker values were considered indicative of disease onset. The way data were simulated is described in ► Appendix 1. The biomarker predictive ability was assessed by the crude and the Bayesian method. Sensitivity was estimated at $t = 2, 4,$ and 6 days before the outcome. Specificity was estimated only during the period $[0, 10[$ days after inclusion because, in controls, there was no trend for change of biomarker values over time. One hundred simulations were performed. The means obtained for the 100 areas under the ROC curve and for sensitivities at four threshold values (1, 2, 3, and 4) were compared to the theoretical time-dependent area under the ROC curve and sensitivity assessed according to the process of generation of the biomarker values (► Table 1).

3.2 Results

Except for the delay of six days and the threshold value 4, the Bayesian method led to higher sensitivities with differences ranging between 0.02 and 0.33. The standard errors were roughly of the same order of magnitude with the two methods. The comparisons with the theoretical results showed that, except for the delay of six days and the threshold value 4, the crude method clearly underestimated the test sensitivity and that the use of the Bayesian method corrected this underestimation. Besides, except for a delay of two days, the sensitivities obtained with the Bayesian method were close to the theoretical values with small differences ranging between -0.05 and 0.03 . The precision of threshold crossing times es-

timates depends partly on the measurement frequency. With measurements taken approximately every three days, there is a lack of information to precisely estimate the latest threshold crossed two days before the event, especially when the biomarker values increase as quickly as the onset of disease become closer in time. This explains the differences between the theoretical and the Bayesian results. A way to increase the precision of Bayesian estimates is to make more frequent measurements or to increase the number of cases.

Both the Bayesian and the crude method underestimated the specificities at low thresholds (data not shown). This was not due to the exact estimations of the thresholds crossing times but to the fact that specificity was assessed using the highest value reached in each control during a given period. The longer was the period, the highest was the bias. Hence, the choice of the period should be made with great caution.

The AUC values obtained with the Bayesian method were higher than those obtained with the crude method and corrected partly the underestimation of accuracy with the latter method. The differences between the Bayesian and the theoretical values came from underestimation of sensitivity with a delay of two days, but also and mainly from underestimation of specificity.

The Bayesian methods led to a better estimation of sensitivity, which is the aim of the present article. Underestimation of specificity came from the empirical assessment of specificity and not from the exact threshold crossing times.

4. Example: CMV Disease Prediction after Renal Transplantation

4.1 Study Description

The study involved 68 patients who had undergone kidney transplantation between January 1, 1999 and December 31, 2003, at the Centre Hospitalier Lyon Sud (Lyon, France). All were CMV-seropositive before transplantation; 46 received a CMV-positive graft and 22 a CMV-negative one. They were weekly monitored for CMV by quantitative PCR during eight weeks after transplantation, semi-monthly until the third month, then monthly until the sixth month. Because the probability of developing CMV disease six months after renal transplantation is low, patients who did not present a CMV disease after a six-month follow-up were considered disease-free.

CMV infection was defined as isolation of CMV by early or late viral culture. CMV disease was defined as the presence of the above defined CMV infection plus either: i) an association of two among the following clinical or biological signs: temperature above 38°C for at least two days, leukopenia (less than 3.5 G/L), thrombocytopenia (less than 150 G/L), abnormalities of liver enzymes (twice or more the reference levels); ii) isolated leukopenia (less than 3 G/L); or iii) tissue injury (invasive disease).

The PCR method had a detection threshold of 200 copies/mL; 321 measurements out of 494 fell below this threshold. Those left-

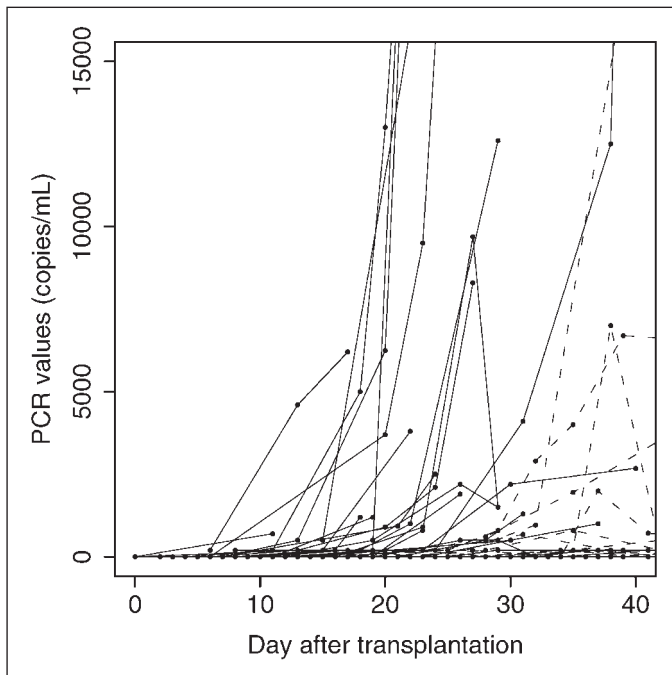


Fig. 1 PCR measurements for cases (solid lines) and controls (dotted lines) versus measurement day after transplantation. x and y scales have been truncated.

censored measurements were given value 0. Forty-three subjects developed a CMV disease with transplantation-to-disease quartiles 21, 25, and 31 days, respectively. The quartiles relative to the number of measurements in those patients were 3, 4, and 5 measurements, respectively. Most patients who developed a CMV disease had an earlier sharp increase in the viral load (► Fig. 1). The viral load of the 25 subjects who did not develop the disease remained generally low;

however, six of them had a slight increase starting from the 20th day, followed by a decrease starting about the 30th day, then a return to the initial level. This may strongly influence the diagnostic test specificity. However, during the first 30 days, the variability between measurements in subjects who did not develop the disease remained very low.

Specificity was estimated at four periods after transplantation, p_1 to p_4 : $[0; 10[$, $[10; 20[$, $[20; 30[$, and $[20; 30[$ days, respectively, with

measurements in 25 patients. Sensitivity was estimated at $t = 0, 5$, and 10 days before the outcome, with measurements in 43 patients. Threshold crossing times were estimated using the Bayesian method. The model was fitted using WinBUGS software package [25]; its corresponding code is given in ► Appendix 2. ROC curves were then constructed with those sensitivity and specificity estimates. There was a large gap between thresholds 0 and 200 on ROC curves, although there was no information on other in-between thresholds. Therefore, only the partial area above threshold 200 was estimated [26]. The obtained values were transformed in values between 0 and 1, as proposed by McClish [27]. The confidence intervals (CI) for AUC values and the standard errors (SE) for sensitivity and specificity were assessed by bootstrap, based on 1000 samples.

4.2 Results

For a fixed delay between marker measurement and disease onset, the ROC curves corresponding to the first 10 days p_1 and 10–20 days p_2 after transplantation were very close (► Fig. 2). Regarding the two later periods p_3 and p_4 , the ROC curve was as much close to the diagonal as the period was late after transplantation. For each period during which specificity was estimated, the ROC curves were all the more close to the diagonal that the delay between marker measurement and

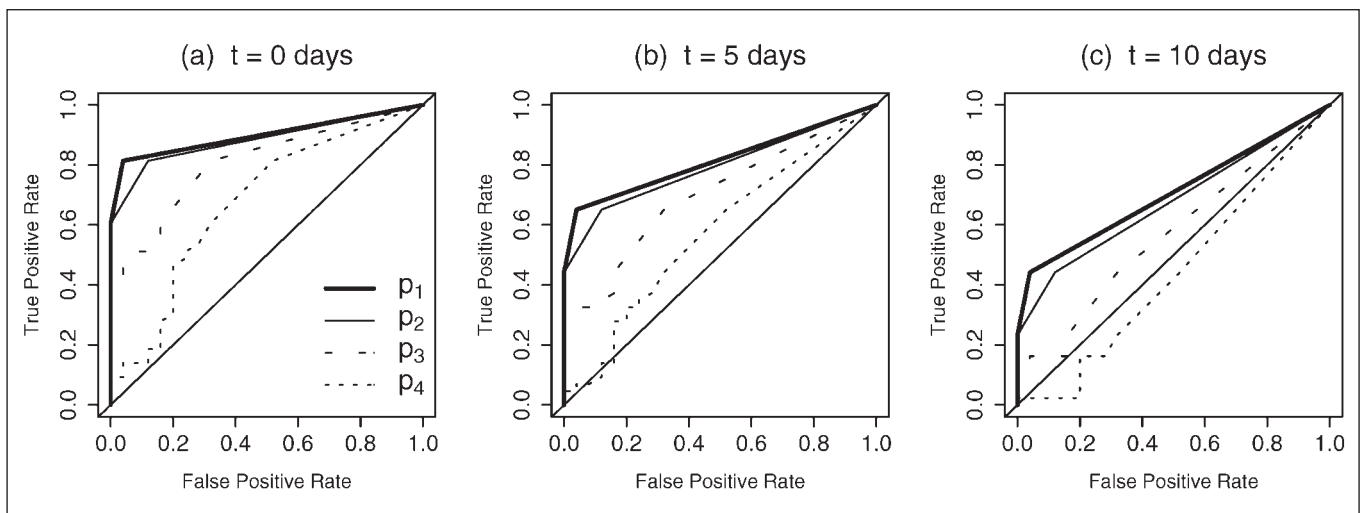


Fig. 2 ROC curves estimated at three delays between marker measurement and disease onset ($t = 0, 5$, and 10 days) and during four periods after transplantation for specificity: $p_1 = [0; 10[$, $p_2 = [10; 20[$, $p_3 = [20; 30[$, and $p_4 = [20; 30[$ days

disease onset increased. AUC estimates in ►Table 2 show that the test accuracy was good during the two first periods after graft and at 0- and 5-day delay between test and disease onset ($t = 0$ and $t = 5$). The AUC was then over 75% but it decreased quickly as the period and the delay increased. The AUC decrease with the advance of the period was linked to a decrease of specificity in late periods; thus, specificity depended on the period after graft. The decrease of the AUC along the delay between marker measurement and disease onset was linked to a decrease of sensitivity. The discriminant ability was not significantly greater than 0.5 neither in the third period p_3 with $t = 10$ nor in the fourth period p_4 with $t = 5$ or $t = 10$ (value 0.5 lies within the 95% confidence interval).

At the specific threshold of 200, the sensitivity was above 80% for $t = 0$, but lower than 50% at $t = 10$ (►Table 3). This threshold was associated with a good specificity during the two first periods p_1 and p_2 , but that specificity decreased quickly to less than 50% during the fourth period.

5. Discussion

The Bayesian method to estimate the exact threshold-crossing times described in this article allows estimating incident sensitivity and static specificity of a longitudinal biomarker. The numerical studies showed that the crude method underestimated sensitivity in the case of interval-censored measurements whereas, under some conditions, the Bayesian method corrected that bias.

In the application, quantitative PCR seemed reliable to predict CMV disease within five days preceding the disease onset and within the first 20 days after transplantation. Before that fifth day, the test sensitivity decreased quickly with the increasing delay between marker measurement and disease onset and the test specificity decreased quickly after the 20th day after transplantation. To our knowledge, this is the first study on early diagnosis of CMV disease that took into account the progression of accuracy with both the marker measurement time and the delay between marker measurement and the disease clinical detection. This was found crucial and explained the differences that exist in the literature about quantitative PCR accuracy,

Table 2 Partial AUC values (95% confidence interval) estimated at three delays between marker measurement and disease onset and during four periods for specificity

Period after graft (days)	Delay between test and disease onset (days)		
	0	5	10
[0; 10[0.852 (0.783; 0.907)	0.769 (0.694; 0.833)	0.662 (0.591; 0.721)
[10; 20[0.845 (0.780; 0.906)	0.759 (0.684; 0.833)	0.647 (0.574; 0.717)
[20; 30[0.757 (0.661; 0.844)	0.669 (0.569; 0.761)	0.550 (0.349; 0.642)
[20; 30[0.634 (0.509; 0.759)	0.555 (0.356; 0.678)	0.344 (0.157; 0.555)

where the delay or the measurement period changes from one study to another [28–30].

The use of the highest biomarker value from each control during a given period may lead to an underestimation of specificity; this bias is conservative because we are sure that the true biomarker accuracy is not smaller than the one estimated. There is no consensus throughout the literature on the way to estimate specificity empirically with repeated marker measurements. Our choice was partly motivated by Murtaugh [31], who also kept the highest marker value from each control to estimate specificity. He compared these results to those obtained keeping the average marker value from each control, but the differences were slight. Emir et al. [32, 33], then Slate and Turnbull [15] proposed another way to assess static specificity without modeling it. At a specific threshold, the specificity

Table 3 Estimated sensitivities and specificities (standard error) for quantitative PCR, the threshold being 200 copies/mL

Sensitivity	
Delay between test and disease onset (days)	
0	0.814 (0.063)
5	0.651 (0.073)
10	0.442 (0.077)
Specificity	
Period after transplantation (days)	
[0; 10[0.960 (0.040)
[10; 20[0.880 (0.066)
[20; 30[0.680 (0.091)
[20; 30[0.480 (0.100)

with each control was estimated by the proportion of negative tests; then the global specificity was defined as the average of all individual specificities, possibly weighted by the number of measurements per subject. The possible bias of this method was not analyzed; the underestimation might be smaller than the one stemming from Murtaugh's method; however, both methods should lead to similar results when estimation periods are short, with few measurements by subject. All those methods could be used after estimation of the threshold-crossing times. A third method would be to model specificity; but then, the bias would depend on the validity of the model assumptions. Certainly, there is still a lot of work to do about estimation of specificity with repeated measurements along time.

One contribution of this article is the assessment of specificity over different periods. This is relevant when specificity progresses along time after inclusion.

The exact estimation of the threshold-crossing times relies on the assumption that, for a specific threshold, the crossing times follow a Weibull distribution. This distribution is commonly used to model failure time data; this is the case of parametric regression for interval-censored data [34–37]. Lindsey [35] compared the results obtained from nine different distributions (including the Weibull, the log-normal, and the gamma distributions) and concluded that, except for heavily interval-censored data, the results may change with the distributional assumptions. However, in the above CMV study, the use of a log-normal distribution led to results, and especially ROC curves, which were almost identical to those obtained with a Weibull distribution.

Other forms than incident and static have been proposed for sensitivity and specificity

[23]; for example, estimating the cumulative sensitivity using the measurements taken during the t days preceding the outcome and not exactly t days before the outcome. However, cumulative sensitivity estimates depend on the time to disease distribution conditional on the marker measurement time and, thus, do not simply reflect biomarker sensitivity. In the concept of dynamic specificity, the controls are the patients who do not develop the disease during the t days following a measurement. However, in our study, the patients developed CMV diseases rapidly after transplantation. Among the subjects whose viral load increased during the few days before disease onset, some developed the disease very soon after t days following a measurement; these would therefore be considered as controls, inducing a high estimate of the false-positive rate and, thus, an underestimation of the real specificity. Thus, the incident sensitivity/static specificity definition of accuracy is, to our opinion, the best way to integrate the concept of time in ROC analysis. As stated by Pepe et al. [22], this should be used in most studies.

Compared to previous methods [15–20], the one proposed here is really easy to implement using standard statistical softwares (the code for Bayesian computations under WinBUGS is given in ►Appendix 2). Moreover, there is no need to define and select a model for biomarker progression, sensitivity, specificity, the ROC curve, or the survival conditional to biomarker values; hence, the method can be very quickly adapted to other settings. Despite the need for a complex modeling phase, the method proposed by Cai et al. [17] remains appealing, but it requires large datasets because each biomarker value for which sensitivity or specificity is estimated adds a new parameter to the model; however, biomarker development studies do not always include a high number of patients. Anyway, our method imposes a restriction: it requires control follow-ups be long enough to assume they are real controls, i.e., the method does not allow so far for censoring, but it may be improved to deal with censored data using ideas similar to those proposed by Cai et al. [17]. The next step of our research would be to analyze the effect of the delay between measurements on accuracy estimates when that delay depends on the last measurement value. Within the context of longi-

tudinal biomarker modeling, Shardell and Miller [38], then Liu et al. [39] have directly addressed this problem.

We hope our simple method will help statisticians undertake complete and precise analyses of longitudinal biomarkers accuracy taking into account the marker measurement time and the delay between marker measurement and outcome. In most studies, this is essential.

Acknowledgments

The authors are grateful to Dr J. Iwaz, PhD, scientific advisor, for his helpful comments on the manuscript.

References

1. Blokh D, Zurgil N, Stambler I, Afrimzon E, Shafran Y, Korech E, Sandbank J, Deutsch M. An information-theoretical model for breast cancer detection. *Methods Inf Med* 2008; 47: 322–327.
2. Benish WA. The use of information graphs to evaluate and compare diagnostic tests. *Methods Inf Med* 2002; 41: 114–118.
3. Pepe MS. The statistical evaluation of medical tests for classification and prediction. Oxford: Oxford University Press; 2003.
4. Sakai S, Kobayashi K, Nakamura J, Toyabe S, Akazawa K. Accuracy in the diagnostic prediction of acute appendicitis based on the Bayesian network model. *Methods Inf Med* 2007; 46: 723–726.
5. Maojo V, Martin-Sanchez F. Bioinformatics: towards new directions for public health. *Methods Inf Med* 2004; 43: 208–214.
6. Goebel G, Muller HM, Fiegl H, Widschwendter M. Gene methylation data – a new challenge for bioinformaticians? *Methods Inf Med* 2005; 44: 516–519.
7. Liska V, Holubec LJ, Treska V, Skalicky T, Sutnar A, Kormunda S, Pesta M, Finek J, Rousarova M, Topolcan O. Dynamics of serum levels of tumour markers and prognosis of recurrence and survival after liver surgery for colorectal liver metastases. *Anticancer Res* 2007; 27: 2861–2864.
8. Roy HK, Khandekar JD. Biomarkers for the Early Detection of Cancer: An Inflammatory Concept. *Arch Intern Med* 2007; 167: 1822–1824.
9. Ransohoff DF. Rules of evidence for cancer molecular-marker discovery and validation. *Nat Rev Cancer* 2004; 4: 309–314.
10. Hanley JA. Receiver operating characteristics ROC methodology: The state of the art. *Crit Rev Diag Imag* 1989; 29: 307–335.
11. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin Chem* 1993; 39: 561–577.
12. Pepe MS. Receiver operating characteristic methodology. *J Am Stat Ass* 2000; 95: 308–311.
13. Zhou X-H, McClish DK, Obuchowski NA. Statistical methods in diagnostic medicine. New York: Wiley; 2002.
14. Etzioni R, Pepe M, Longton G, Chengcheng Hu, Goodman G. Incorporating the time dimension in receiver operating characteristic curves: A case study of prostate cancer. *Med Decis Making* 1999; 19: 242–251.
15. Slate EH, Turnbull BW. Statistical models for longitudinal biomarkers of disease onset. *Stat Med* 2000; 19: 617–637.
16. Zheng Y, Heagerty PJ. Semiparametric estimation of time-dependent ROC curves for longitudinal marker data. *Biostatistics* 2004; 5: 615–632.
17. Cai T, Pepe MS, Zheng Y, Lumley T, Jenny NS. The sensitivity and specificity of markers for event times. *Biostatistics* 2006; 7: 182–197.
18. Zheng Y, Heagerty PJ. Prospective accuracy for longitudinal markers. *Biometrics* 2007; 2: 332–341.
19. Cai T, Cheng S. Robust combination of multiple diagnostic tests for classifying censored event times. *Biostatistics* 2008; 9: 216–233.
20. Song X, Zhou X-H. A semiparametric approach for the covariate specific ROC curve with survival outcome. *Stat Sinca* 2008; 18: 947–966.
21. Cai TZ, Yingye. Model checking for ROC regression analysis. *Biometrics* 2007; 63: 152–163.
22. 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: 86–113.
23. Heagerty PJ, Zheng Y. Survival model predictive accuracy and ROC curves. *Biometrics* 2005; 61: 92–105.
24. Lindsey JC, Ryan LM. Tutorial in biostatistics: methods for interval-censored data. *Stat Med* 1998; 17: 219–238.
25. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Stat Comput* 2000; 10: 325–337.
26. Zhang DD, Zhou X-H, Freeman DH, Freeman JL. A non-parametric method for the comparison of partial areas under ROC curves and its application to large health care data sets. *Stat Med* 2002; 21: 701–715.
27. McClish DK. Analyzing a portion of the ROC curve. *Med Decis Making* 1989; 9: 190–195.
28. Naumnik B, Malyszko J, Chyczewski L, Kovalchuk O, Mysliwiec M. Comparison of serology assays and polymerase chain reaction for the monitoring of active cytomegalovirus infection in renal transplant recipients. *Transplant Proc* 2007; 39: 2748–2750.
29. Mhiri L, Kaabi B, Houimel M, Arrouji Z, Slim A. Comparison of pp65 antigenemia, quantitative PCR and DNA hybrid capture for detection of cytomegalovirus in transplant recipients and AIDS patients. *J Virol Methods* 2007; 143: 23–28.
30. Madi N, Al-Nakib W, Mustafa AS, Saeed T, Pacsa A, Nampoory MR. Detection and monitoring of cytomegalovirus infection in renal transplant patients by quantitative real-time PCR. *Med Princ Pract* 2007; 16: 268–273.
31. Murtaugh PA. ROC curves with multiple marker measurements. *Biometrics* 1995; 51: 1514–1522.
32. Emir B, Wieand S, Su JQ, Cha S. Analysis of repeated markers used to predict progression of cancer. *Stat Med* 1998; 17: 2563–2578.
33. Emir B, Wieand S, Jung S-H, Ying Z. Comparison of diagnostic markers with repeated measurements: a non-parametric ROC curve approach. *Stat Med* 2000; 19: 511–523.

34. Odell PM, Anderson KM, D'Agostino RB. Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time model. *Biometrics* 1992; 48: 951–959.

35. Lindsey JK. A study of interval censoring in parametric regression models. *Lifetime Data Anal* 1998; 4: 329–354.

36. Collet D. *Modelling Survival Data in Medical Research*. London: Chapman and Hall; 2003.

37. Sparling YH, Younes N, Lachin JM, Bautista OM. Parametric survival models for interval-censored data with time-dependent covariates. *Biostatistics* 2006; 7: 599–614.

38. Shardell M, Miller RR. Weighted estimating equations for longitudinal studies with death and non-monotone missing time-dependent covariates and outcomes. *Stat Med* 2008; 27: 1008–1025.

39. Liu L, Huang X, O'Quigley J. Analysis of Longitudinal Data in the Presence of Informative Observational Times and a Dependent Terminal Event, with Application to Medical Cost Data. *Biometrics* 2008; 64: 950–958.

Appendix 1

1. Generation of the Simulated Data

1.1 Notation

i = subject index; k = k th marker measurement; s_{ik} = time of the k th measurement for the i th subject; Δ_{ik} = delay between the k th measurement and the diagnosis time for the i th subject

1.2 Sampling Times (s_{ik})

Patients should have a biomarker measurement every three days for 30 days after inclusion into the study; but, actually, the measurement is often delayed. Generate:

$$s_{ik} = 3k + \varepsilon_{ik}, k = 0, \dots, 9$$

$$\varepsilon_{ik} = \begin{cases} \text{uniform}(1, 2.95) & \text{if } k = 0, \\ \text{uniform}(0, 2.95) & \text{if } k > 0. \end{cases}$$

1.3 Time of Diagnosis

The time of diagnosis was generated as follows:

$T_i \sim \text{uniform}(15, 20)$ with probability 0.4
 $T_i \sim \text{uniform}(20, 30)$ with probability 0.6

1.4 Biomarker Values

For Controls

Throughout each simulation, controls have their own biomarker value normally distributed with mean 1 and variance 0.25; for each measurement, an error is added that follows a normal distribution with mean 0 and variance 0.49.

For Cases

In cases, biomarker values are generated as for controls up to eight days before diagnosis; for later measurements, an extra term is added:

$$\exp(2 - (0.5 + \delta_i)\Delta_{ik})$$

δ_i corresponds to patients' specific biomarker increase with time between marker measurement and diagnosis. It follows a normal distribution, with mean 0 and variance 0.0025.

Measurements taken after the time of diagnosis are removed.

2. Calculation of the Theoretical AUC Values

When biomarkers follow normal distributions in the diseased and non-diseased populations (respectively $N(\mu_D, \sigma_D^2)$ and $N(\mu_{\bar{D}}, \sigma_{\bar{D}}^2)$), Pepe et al. [3] showed that the AUC for the ROC curve is given by

$$\Phi\left(\frac{a}{\sqrt{1+b^2}}\right)$$

where $a = (\mu_D - \mu_{\bar{D}})/\sigma_D$, $b = \sigma_{\bar{D}}/\sigma_D$, and Φ denotes the standard normal cumulative function.

According to the process of generation of biomarker values, during each period, measurements in control subjects follow a normal distribution with mean 1 and variance 0.25 ± 0.49 .

In cases, for a delay Δ between the marker measurement and the diagnosis time, the

biomarker values follow a normal distribution with mean

$$1 + \exp(0.5(4 - \Delta))$$

and variance

$$\exp(4 - \Delta) \times \text{Var}(\exp(-\delta \times \Delta))$$

where δ follows a normal distribution with mean 0 and variance 0.0025.

For small delays Δ , the variance may be approximated using the delta-method; for our applications, the variance was estimated using 10^7 random values stemming from a normal distribution with mean 0 and variance 0.0025.

Those results allow us to calculate the theoretical AUC for each period and delay between marker measurement and the onset of disease.

Appendix 2

The WinBUGS code for estimating the exact threshold-crossing time (paragraph ROC curve analysis).

```
model
{
  for(i in 1:N) ## N corresponds to the
  number of crossings
  {
    crossing_time[i]~dweib(r,mue)I
    (left[i],right[i])
    ## left[i] corresponds to the date of
    last PCR measurement whose result was
    inferior to the threshold
    ## right[i] corresponds to the date
    of first PCR measurement whose result
    was superior or equal to the thresh-
    old
  }
  r~dgamma(1.0E-3, 1.0E-3)
  mue<-exp(mu)
  mu~dnorm(0,0.000001)
}
```