Time-Dependent ROC Analysis for Early Detection of End-Stage Renal Disease (ESRD) Using Baseline Glomerular Filtration Rate

Nan Hu
University of Utah

1. Introduction

Biomarkers are important tools for early detection of various kinds of diseases. For example, biomarkers have been developed to detect cancers before the onset of clinical disease (Pepe et al., 2001). A receiver operating characteristic (ROC) curve is commonly used to assess the accuracy of a biomarker in distinguishing between diseased and non-diseased patients (Zhou, Obuchowski and McClish 2002). Traditionally, ROC analysis deals with dichotomous diagnostic outcomes such as determining whether a disease is present or absent at a cross-sectional time point. For prognostic tests, however, we often deal with disease status that changes over time. A traditional ROC analysis for dichotomous disease status is not adequate under this situation because the disease outcome of a subject is not fixed. To accommodate time-dependent disease outcomes, a time-dependent ROC curve has been proposed (Heagerty and Zheng 2004) to assess the prognostic accuracy of a biomarker to distinguish between subjects with and without the disease event over time.

A time-dependent ROC curve are often used for comparing the prognostic accuracies of a large number of candidate biomarkers for disease-related events such as death from cancers, HIV infections, and kidney failures. Methods to evaluate the prognostic value of a biomarker other than a time-dependent ROC curve exist. Examples include the extended $R2$ approach proposed by O'Quigley and Xu (2001) that measures the variation explained by a time-to-event model, and rank-based correlation coefficient approaches such as concordance, Kendall's tau (Kendall 1938), and Spearman's correlation coefficients. However, the difficulty for these approaches is that the categorizations of biomarker values are often arbitrary and may not yield categories that are comparable for different biomarkers. For this reason, the time-dependent ROC curves is more attractive since it can provide a common scale for the comparison of accuracy among different biomarkers.

As the motivation, we describe our study question using a randomized clinical trail, namely the African American study of kidney disease and hypertension (AASK). This study enrolled 1094 African Americans with the following conditions: (1) hypertension; (2) 18 to 70 years of age; (3) a original glomerular filtration rate (GFR) of 20 to 65 ml/min per 1.73 $m2$; and (4) no other apparent cause of renal insufficiency other than hypertension. Study participants were randomized to a usual mean arterial pressure (MAP) goal of 102 to 107 mmHg or a low MAP goal of <92 mmHg, and to initial treatment with one of three anti-hypertensive study drugs: a sustained-release β-blocker (metoprolol), an angiotensin converting enzyme inhibitor (ACEI, ramipril), or a dihydropyridine calcium channel blocker (amlodipine). The primary end point of the study is the time to end stage renal disease (ESRD), which was subject to censoring. The three major biomarkers are estimated GFR (eGFR), iothalamate GFR (iGFR) and urinary protein to creatinine ratio (UPCR) measured at baseline. Here our goal is to evaluate the performance of the above biomarkers for the prognosis of ESRD events in a future time from baseline. To obtain the goal, it's natural that we establish a time-dependent ROC curves to evaluate the ability of the above biomarkers for the prognosis of ESRD events.

Other clinical risk factors, such as treatment and hypertension status, may affect the prognosis accuracy. Hence, it would be important to consider and adjust for these factors in constructing the ROC for the biomarkers.

2. Notations and Definitions

2.1. Notations

Here, we consider the scenario where a biomarker has a continuous distribution and is measured only at baseline. Let Ti and Ci denote the event time and censoring time, respectively, for a subject i ($i = 1, \ldots, n$), and let $Di(t)$ denote the binary disease outcome for the subject i at time t from baseline. Define $Vi = \min(Ti, Ci)$, where min indicates the minimum of its arguments, and define $\Delta i = 1(Ti \leq Ci)$ where 1(.) is the indicator function, having value 1 when the condition of its argument is satisfied and value 0 otherwise. Let Yi be the baseline biomarker and $\mathbf{X}i$ be the vector of covariates for subject i. For notational convenience, we define $\mathbf{Z}i = (\mathbf{X}i, Yi)$ for each i.

2.2. Time-dependent True Positive Rate

In the classic setting with binary outcomes, the true positive rate is the proportion of cases who have positive test results. When the outcome is a failure time, however, there are different ways of defining cases, resulting in different kinds of true positive rates being defined. Here we present two definitions of the time-dependent true positive rate (TPR), which has been given in the literature such as Heagerty and Zheng (2004). By using the Baye's rule, these TPRs can be expressed by the ratio of the integrals of the conditional survivor function, as shown in equations (1) and (2).

2.2.1. Cumulative True Positive Rate, TPR_C:

$$TPR_C(y;t,\mathrm{x}) = Pr(Y > y|T \leq t, \mathbf{X} = \mathrm{x}) = \frac{\int_y^\infty \{1 - S(t|\mathrm{x},u)\} dP(Y \leq u|\mathrm{x})}{\int_{-\infty}^\infty \{1 - S(t|\mathrm{x},u)\} dP(Y \leq u|\mathrm{x})}, \quad (1)$$

where $S(t|\mathbf{x}, u)$ is the survival function at event time $T = t$, conditional on covariates $\mathbf{X} = \mathbf{x}$ and biomarker value $Y = u$. Here, $TPRC$ evaluates the capacity of biomarker, Y, for detecting events occurring up to a time point t. Using this definition, we are interested in predicting the disease prevalence of the study cohort at a given time.

2.2.2. Incident True Positive Rate, TPR_I :

$$TPR_I(y;t,\mathrm{x}) = Pr(Y > y|T = t, \mathbf{X} = \mathrm{x}) = \frac{\int_y^\infty f(t|\mathrm{x},u) dP(Y \leq u|\mathrm{x})}{\int_{-\infty}^\infty f(t|\mathrm{x},u) dP(Y \leq u|\mathrm{x})}, \quad (2)$$

where $f(t|\mathbf{x}, u)$ is the probability density function for $T = t$ conditional on covariates $\mathbf{X} = \mathbf{x}$ and biomarker value $Y = u$. Here, $TPRI$ evaluates the capacity of the biomarker, Y, for detecting events that occur right at the time point t. Using this definition, we are more interested in the disease incidence of the study cohort at a given time.

2.3. Time Dependent False Positive Rate

Similar to the time-dependent true positive rate, there are different types of time-dependent false positive rates according to the definition of controls. Next, we give two kinds of the false positive rate (FPR), which also has been given in previous literature (Heagerty and Zheng 2005). By using the Baye's rule, these FPRs can also be expressed by the ratios of the integrals of the conditional survivor functions.

2.3.1. Dynamic False Positive Rate, FPRD:

$$FPR_D(y;t,\mathrm{x}) = Pr(Y > y|T > t, \mathbf{X} = \mathrm{x}) = \frac{\int_y^\infty S(t|\mathrm{x},u)dP(Y \le u|\mathrm{x})}{\int_{-\infty}^\infty S(t|\mathrm{x},u)dP(Y \le u|\mathrm{x})}, \qquad (3)$$

Under this definition, the time defining the controls is dynamic. Usually this dynamic time is the time lags after the measurement of biomarker for subjects in the study. Controls are those who have not developed the disease at these time lags.

2.3.2. Static False Positive Rate, FPRS:

$$FPR_S(y;t^*,\mathrm{x}) = Pr(Y > y|T > t^*, \mathbf{X} = \mathrm{x})\frac{\int_y^\infty S(t^*|\mathrm{x},u)dP(Y \le u|\mathrm{x})}{\int_{-\infty}^\infty S(t^*|\mathrm{x},u)dP(Y \le u|\mathrm{x})}. \qquad (4)$$

Under this definition, controls are subjects who are disease free in time period, (0, $t*$), where $t*$ is a fixed time point. For static FPR, we usually need to prespecify the time point $t*$ such that the controls can be described as a reference group of the study. The problem of *FPRS* is that the time defining controls, $t*$, is not the same as, or related to, the time defining cases in the corresponding TPR.

2.4. Time-dependent ROC Curves

After choosing the time-dependent TPR and FPR, we can define the time-dependent ROC curves. The cumulative/dynamic ROC curve (ROC**C/D**), the incident/static ROC curve (ROC**I/S**), and the incident/dynamic ROC curve (ROC**I/D**) are all meaningful time-dependent ROC curves, which are compound functions of the time-dependent TPR and FPR defined in the Sections 2.2 and 2.3 (Heagerty and Zheng 2005; Cai et al. 2006). The definitions for the above time-dependent covariate-specific ROC curves can be expressed as follows:

$$\mathrm{ROC}^{\mathbf{C/D}}(v;t,\mathrm{x}) = \mathrm{TPR}_{\mathbf{C}}\left\{\mathrm{FPR}_{\mathbf{D}}^{-1}(v;t,x);t,\mathrm{x}\right\}, \qquad (5)$$

$$\mathrm{ROC}^{\mathbf{I/S}}(v;t,\mathrm{x}) = \mathrm{TPR}_{\mathbf{I}}\left\{\mathrm{FPR}_{\mathbf{S}}^{-1}(v;t,x);t,\mathrm{x}\right\}, \qquad (6)$$

$$\mathrm{ROC}^{\mathbf{I/D}}(v;t,\mathrm{x}) = \mathrm{TPR}_{\mathbf{I}}\left\{\mathrm{FPR}_{\mathbf{D}}^{-1}(v;t,x);t,\mathrm{x}\right\}, \qquad (7)$$

where v denotes the time-dependent FPR, t (satisfying $t > 0$) denotes the time lag from baseline, and $\mathbf{x}$ denotes a certain realization of covariates $\mathbf{X}$. Here, $\mathrm{ROC}^{\mathbf{C/D}}$, $\mathrm{ROC}^{\mathbf{I/S}}$, $\mathrm{ROC}^{\mathbf{I/S}}$ denote the corresponding cumulative/dynamic, incident/static, and incident/dynamic true positive rate, respectively. Let $y*$ be the threshold value that yields

v, i.e., FPR$D(y^*; t, \mathbf{x}) = P(Y > y^* |T > t, \mathbf{x}) = v$ or FPR$S(y^*; t^*, \mathbf{x}) = P(Y > y^*|T > t^*, \mathbf{x}) = v$. Then the time-dependent ROC is the true positive (or sensitivity) obtained using the threshold value y^* such that:

$$\mathrm{ROC}^{\cdot/\cdot}(y^*; t, \mathrm{x}) = v$$

where ROC$^{\cdot/\cdot}$ denotes the three types of time-dependent ROC curves defined above.

All of the time-dependent ROC curves defined above can be used to evaluate and compare the accuracy of biomarkers in classifying subjects based on their times to disease event after adjusting for covariates. The cumulative/dynamic ROC curve is useful in distinguishing subjects failing by a given time from those failing after that time; the incident/dynamic ROC curve is useful in distinguishing subjects failing at a given time from those failing after that time; the incident/static ROC curve is useful in distinguishing subjects failing at one time point from those failing after a fixed time point; In this paper, we will focus on cumulative/dynamic, and incident/static ROC curves.

3. Time-dependent ROC Model

From the definition of time-dependent TPR and FPR in (1)-(4) and the time-dependent ROC curves in (5)-(7), we can see that both the conditional survival function for the disease time and the conditional distribution of the biomarker must be estimated in order to obtain the time-dependent TPRs, FPRs, and ROC curves.

In this section, the semi-parametric location model for biomarkers and the non-parametric time-to-event model are introduced. The estimation procedure is described in Section 4. The survival function of the disease time conditional on biomarker and covariates can be estimated from the former model together with the regression parameters and the transformation function. The biomarker distribution conditional on covariates can be obtained from the latter model.

3.1. Model of Biomarkers

Here, we assume that biomarker Y depends on covariates $\mathbf{X}$ through the following semi-parametric model:

$$Y_i = \alpha_0^{\mathrm{T}} \mathbf{X}_i + \epsilon_i^*, \qquad (8)$$

where ϵ^* is the random error with unknown distribution. Then, the distribution function of the biomarker Y can be be given by:

$$Pr(Y \leq y|X) = H^*(y - \alpha_0^{\mathrm{T}} \mathbf{X}), \qquad (9)$$

where $H*$(.) is the cumulative distribution function (CDF) of $\epsilon *$, and α0 is the vector of parameters for the linear regression of covariates $\mathbf{X}$ on biomarker Y . This semi-parametric location model was also adopted in Song and Zhou (2008).

3.2. Model of Event Time

We consider the typical setting of a censored time-to-event outcome where the data is subject to right censoring. We use the notations given in Section 2.1 and suppose there are n subjects in the dataset. One important assumption we make is that, given $\mathbf{Z}i$, the event time Ti and the censoring time Ci are independent. We assume that the event time Ti depends on biomarker and covariates through the following nonparametric transformation model:

$$G(T_i) = \beta^{\mathrm{T}}\mathbf{Z_i} + \epsilon_i \ , \qquad (10)$$

where $G(.)$ is an unknown monotone increasing transformation function, and ϵ is the unobserved random error with mean zero and variance 1. Since ϵ and C are unobserved, there is a well-known identification problem. Given the observed variables Δi, Z and T, Equation (10) continues to hold if G, ϵ and $C*$ are replaced by αG, $\alpha\epsilon$ and $\alpha C*$, for any constant αG. Therefore, the scale and location normalizations are needed. As shown by Gorgenes and Horowitz (1996), and Ichimura (1993), identification of β up to scale requires that Z has at least one component that is continuously distributed conditional on the others and whose β-coefficient is nonzero. Without loss of generality, we let the first element of vector **X** have coefficient 1 and other elements have coefficients vector θ, i.e., βT = (1, θT). For location normalization, we let $\Lambda(t0) = 0$, where $t0$ is some constant.
Since the time to disease event, Ti, is subject to censoring, it is not always observed. However, Vi and Δi are always observed. Thus, the model (10) can be expressed in terms of Vi and Δi as follow:

$$G(V_i) = \min[G(T_i), G(C_i)] = \min[\beta^{\mathrm{T}}\mathbf{Z}_i + \epsilon_i, C_i^*] \ , \qquad (11)$$

where $C^*\, i = G(Ci)$, the transformation of the censoring variable.

3.3. Time-Dependent ROC Curve
Based on different kinds of the time-dependent TPR and FPR given by (1)–(4), the time-dependent ROC curve can be obtained by using equations (5)–(7).

4. Application in Kidney Disease Study - AASK

As an application, we consider the data from AASK randomized clinical trial with a factorialdesign. For comparison purpose, we used the our time-dependent ROC estimator, ROC$^{C/D}$, to three different baseline biomarkers (eGFR, iGFR and UPCR) at select t (1, 2, 3, and 4 years since baseline). Table 7.1 shows the baseline summary statistics of AASK cohort by the two baseline blood pressure groups(BP=L and BP=M). Figures 1 and 2 present the covariate-specific time-dependent ROC$^{C/D}$ curves (with 95% confidence intervals) of the above three biomarkers and their corresponding AUCs at the four time points listed above.

There are only slight changes in ROC curves (and AUC) of the three biomarker across the six combinations of the two risk factors (BP and Drug). However, the figures show that the performances of eGFR and iGFR for prognosis of ESRD event are similar across the four pre-specified time points. Among the three biomarkers, eGFR and iGFR have consistently better overall performance than that of UPCR. For detection of early ESRD events at 12 months, UPCR has better sensitivities for specificities higher than 80% (or, equivalently, false positive rate lower than 20%). We also plot the covariate-specific area under the time-dependent ROC curve (AUC) as a function of time in this example (we customized the time range from 6 months top 60 months). Figures 3 and 4 show the time-dependent AUCs for the six different combinations of the baseline covariates (BP and DRUG). For eGFR and iGFR, the AUCs are consistently increasing from about 0.80 at 6 months to about 0.95 at 60 months. Besides, There is a slight ”U” shape of the AUC(t) curve of UPCR, with the bottom of AUC at 0.65 around 24 months. However, we can find only tiny differences in AUC(t) curves among different covariates combinations.

References

Appel, L.J., Middleton, J., Miller, E.R., Lipkowitz, M., and Norris, K. et al. (2003) The Rationale and Design of the AASK Cohort Study. *Journal of American Society of Nephrology* **14**, S166-S172.

Cai, T., Pepe. M.S. Lumley, T. and Jenny, N.S. (2006) The sensitivity and specificity of markers for event times. *Biostatistics* **7**, 182-197.

Cai, T. and Pepe. M.S. (2002) Semiparametric Receiver Operating Characteristic Analysis to Evaluate Biomarkers for Disease. *Journal of American Statistical Association* **97**, 1099-1107.

Cai, Z. Fan, J. and Li, R. (2000) Efficient estimation and inferences for varying-coefficient models. *Journal of American Statistical Association* **95**, 888-902.

Chen, Y.Q., Hu, N. and Cheng, S. (2011) On a proportional odds model with time-varying covariates. *Journal of American Statistical Association*, In press.

Cheng, S.C., Wei, L.J., and Ying, Z. (1995) Analysis of transformation models with censored data. *Biometrika* **82**, 835-845.

Fan, J, and Huang, T. (2005) profile likelihood inferences on semiparametric varyingcoefficient partially linear models. *Bernoulli* **11**, 1031-1057.

Fan, J., Huang, L. and Zhou, Y.(2006) Local partial-likelihood estimation for lifetime data. *The Annals of Statistics* **34**, 290-325.

Gray, R.J. (1992) Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association* **87**, 942-951.

Gorgens, T. and Hoeowitz, J. (1999) Semiparametric estimation of a censored regression model with an unknown transformation of the dependent variable. *Journal of Econometrics* **90**, 159-191.

Grambsch, P.M. and Therneau, T. M. (1994) Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* **81**, 515-526.

Hastie, T. and Tibshirani, R. (1990) General Additive Models. *Chapman and Hall*

Hastie, T. and Tibshirani, R. (1993) Varying-coefficient models. *Journal of Royal Statistical Society, B.* **55**, 757-796.

Heagerty, P.J., Lumley, T. and Pepe, M.S. (2000) Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometric* **56**, 337-344.
Heagerty, P.J. and Zheng, Y. (2005) Survival model predictive accuracy and ROC curves. *Biometrics* **98**, 409-417.

Horowitz, J.L. (1996) Semiparametric estimation of a regression model with an unknown transformation of dependent variable. *Econometrika* **64**, 103-137.

Horowitz, J.L. (1998) Semiparametric Methods in Econometrics, *Springer,* New York.

Hsieh, F. (2001) On the heterscedastic hazards regression models: theory and applicatrion.*Journal of Royal Statistical Society, B* **63**, 63-79.

Kendall, M. (1938) A New Measure of Rank Correlation. *Biometrika* **30**, 8189.

Khan, S. and Tamer, E. (2007) Partial rank estimation of duration models with general forms of censoring. *Journal of Econometrics* **136**, 251280.

Mazumdar, S., Berhane, Z., Weissfeld, L., Begley, P.R., and Reynolds, C.F. (2002) Survival models with time-varying coefficients: A flexible approach to the analysis of Psychiatric survival data. *General Psychopharmacology* **36**, 84-91.

O'Quigley, J. and Frandre, P. (1994) Predictive capacity of proportiaon hazards regression. *National Academy of Science* **91**, 2310-2314.

O' Quigley, J. and Xu, R. (2001) Explained variation in proportional hazards regression in Handbook of Statistics in Clinical Oncology, Editor: J. Crowley, *Marcel Dekker,* New York.

Pakes, A., Pollard, D. (1989) Simulation and the asymptotics of optimization estimators. *Econometrica* **57**, 1027-1057.

Pepe, M.S., Etzioni, R., Feng, Z.D., Potter, J., Thompson, M., Thornquist, M., Winget, M. and Yasui, Y. (2001) Phases of biomarker development for early detection of cancer. *Journal of the National Cancer Institute* **93**, 10541061.

Pollard, D. (1984) Convergence of Stochastic Processes. *Springer*, New York.
Schemper, M. and Henderson, R. (2000) Predoctive accuracy variation in Cox regression. *Biometrics* **56**, 249-255.

Schoenfeld, D. (1982) Partial residual for the proportional regression model. *Biometrika* **69**, 239-241.

Sherman, R.P. (1994) Maximal inequalities for degenerate U-processes with applications to optimization estimators. *Annals of Statistics* **22**, 439-459.

Slate, E.H. and Turnbull, B.W. (2000) Statistical models for longitudinal biomarkers of disease onset. *Statistics in Medicine* **19**, 617-637.

Song, X., Ma, S., Huang, J. and Zhou, X.H. (2006) A semiparametric approach for the nonparametric transformation survival model with multiple covariates. *Biostatistics*

8, 197-211.

Song, X. and Zhou, X-H (2008) A semiparametric approach for the covariate specific ROC curve with survival outcome *Statistica Sinica* **18**, 947-965.

Tian, L., Zucker, D. and Wei, L.J. (2005) On the cox model with time-varying coefficients. *Journal of American Statistical Association* **100**, 172-183.

Tsiatis, A.A. and Davidian, M. (2004) Joint modelling of longitudinal and time-to-event data: An overview. *Statistica Sinica* **14**, 809-834.

Wasserman, L. (2006) All of nonparametric statistics, *Springer*. New York.

Zeng, D. and Lin, D.Y. (2006) Efficient estimation of semiparametric transformation models for counting process. *Biometrika* **93**, 627-640.

Zeng, D. and Lin, D.Y. (2007) Maximum likelihood estimation in semiparametric regression models with censored data. *Journal of Royal Statistical Society, B* **69**, 1-30.

Zheng, Y. and Heagerty, P.J. (2004) Semiparametric estimation of time-dependent ROC curves for longitudinal marker data. *Biostatistics* **5**, 615-632.

Zhou, X-H, Obuchowski, N.A., and McClish D.K. (2003) Statistical Method in Diagnostic Medicine, *Wiley*, New York.

Zhou, X-H, Lin, H., and Johnson, E. (2008) Nonparametric heteroscedastic transformation regression models for skewed with an application to health care cost. *Journal of Royal Statistical Society, B* **70,** 1029-1047.

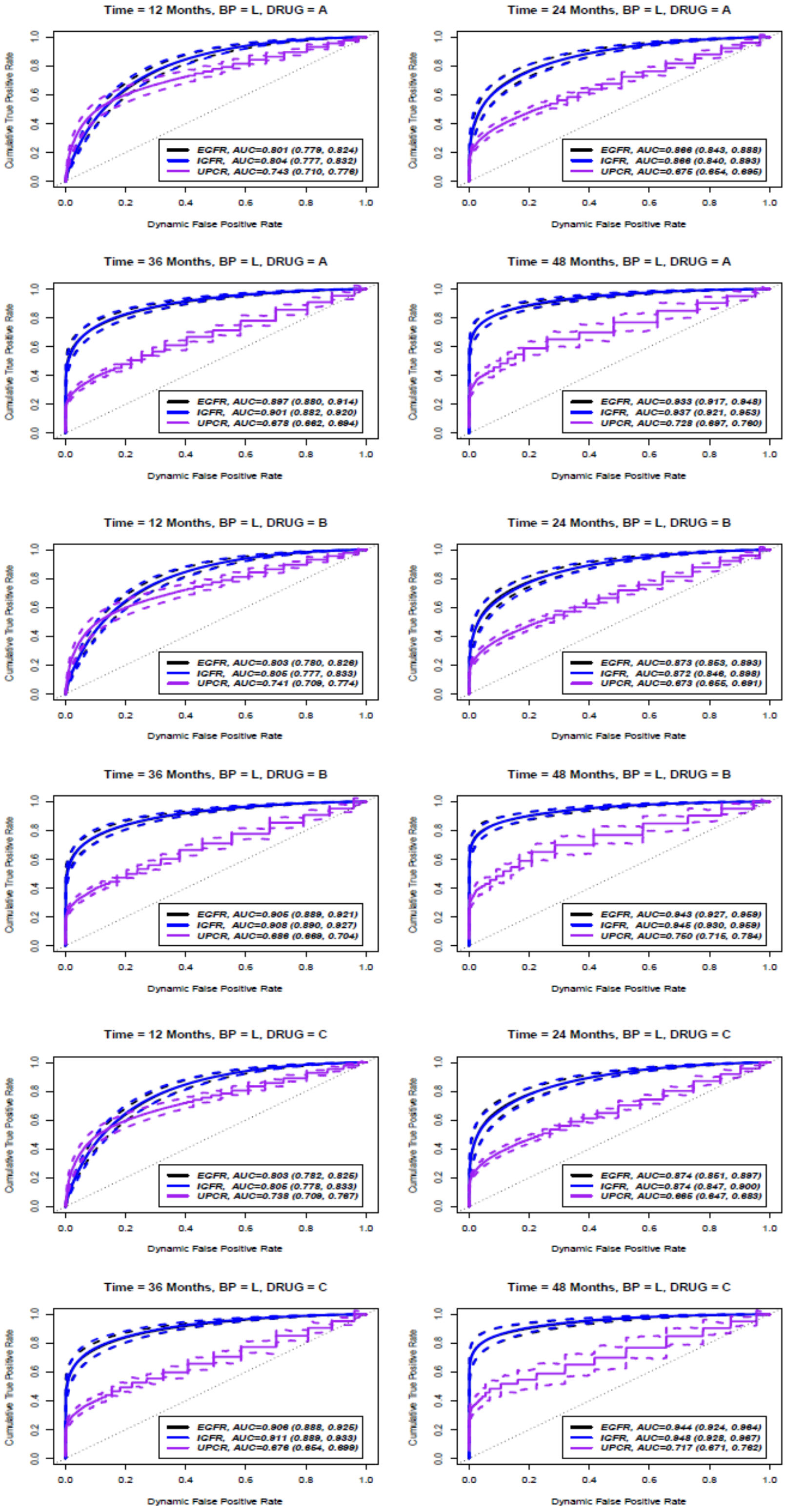


Fig. 1. Time-dependent $\mathrm{ROC}^{C/D}$ Curves of eGFR, iGFR and UPCR for bp $=$ L in AASK Trial by 1, 2, 3, and 4 Years after Enrollment

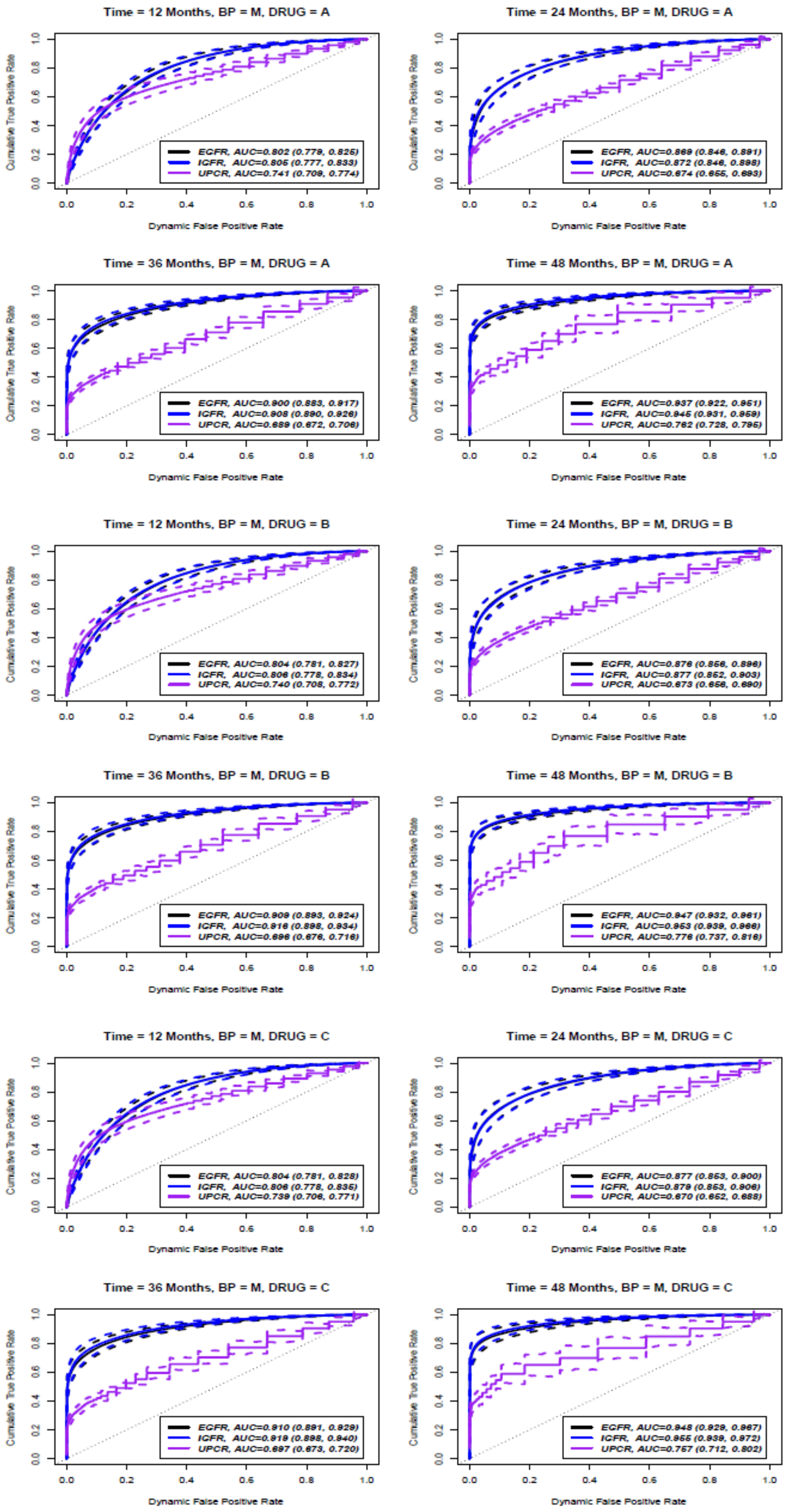


Fig. 2. Time-dependent $\mathrm{ROC}^{C/D}$ Curves of eGFR, iGFR and UPCR for bp $=$ M in AASK Trial by 1, 2, 3, and 4 Years after Enrollment

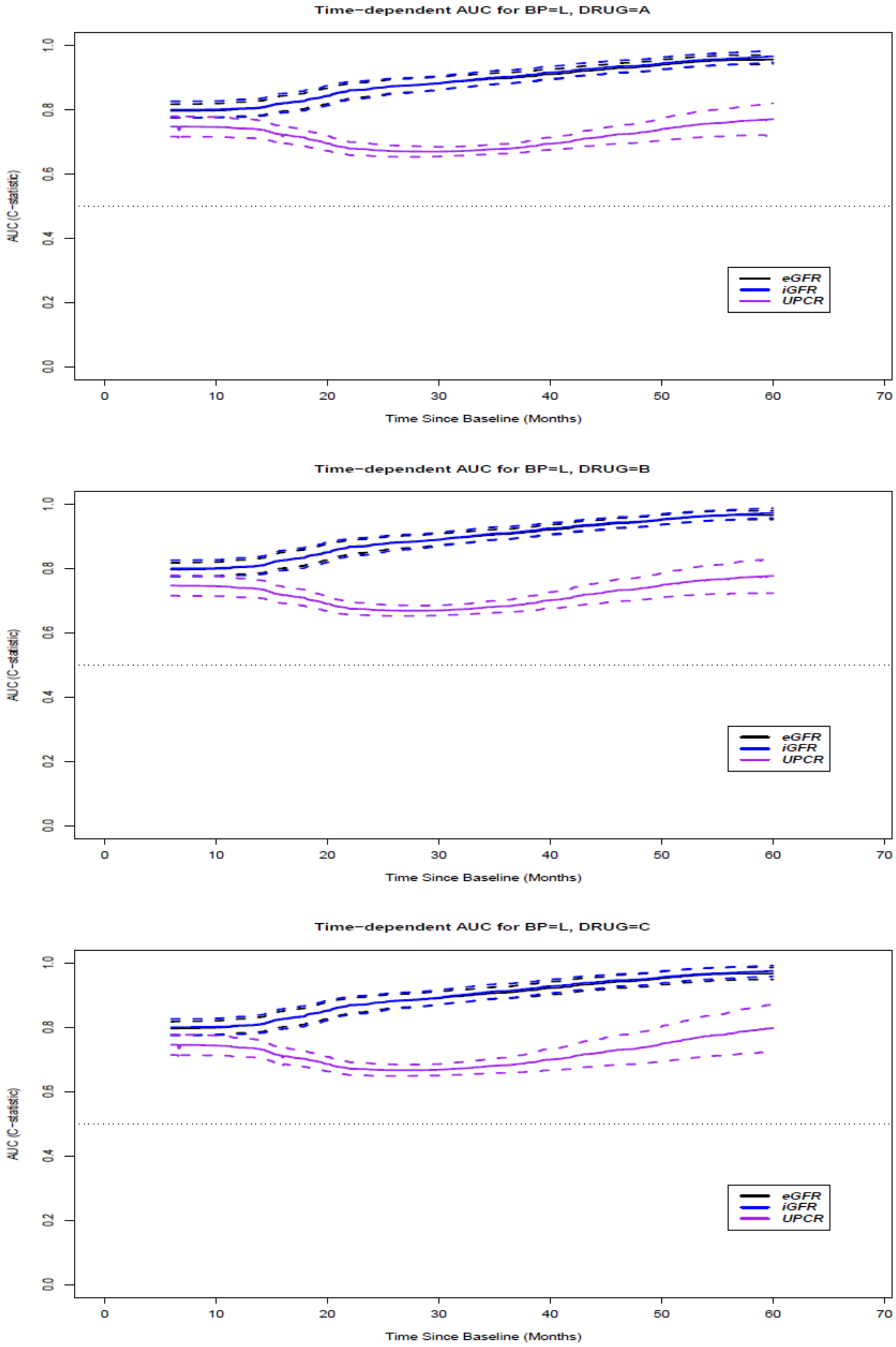
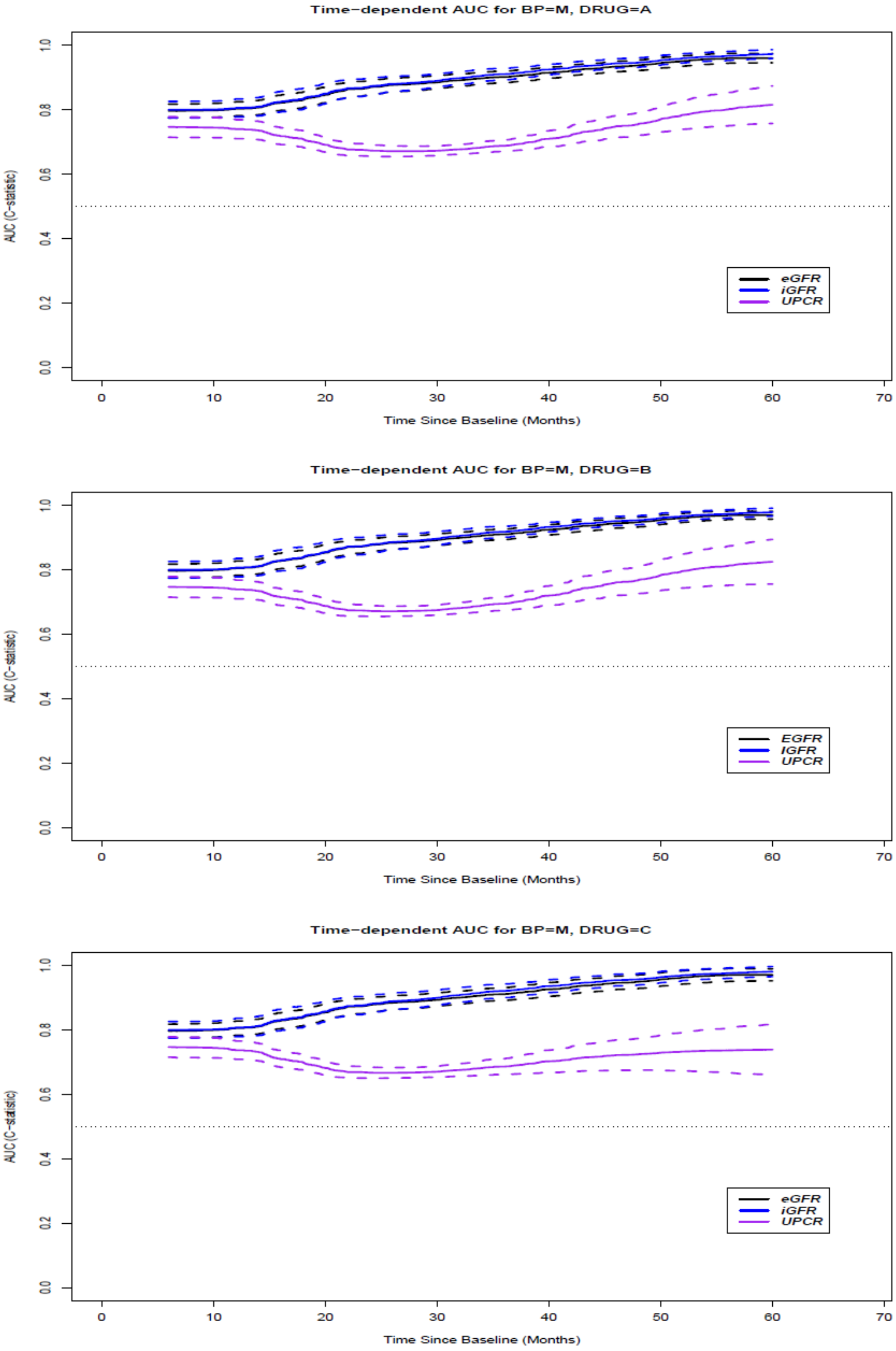


Fig. 3. Time-dependent AUC of eGFR, iGFR and UPCR for bp $=$ L in AASK Trial



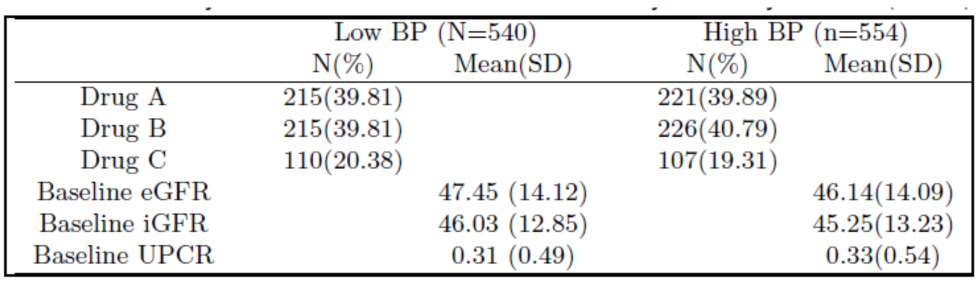
Time−dependent AUC for BP=M, DRUG=A

Time−dependent AUC for BP=M, DRUG=B

Time−dependent AUC for BP=M, DRUG=C

Fig. 4. Time-dependent AUC of eGFR, iGFR and UPCR for bp $=$ M in AASK Trial

Table 1: Patient Characteristic of AASK Trial



	Low BP (N=540)		High BP (n=554)	
	N(%)	Mean(SD)	N(%)	Mean(SD)
Drug A	215(39.81)		221(39.89)	
Drug B	215(39.81)		226(40.79)	
Drug C	110(20.38)		107(19.31)	
Baseline eGFR		47.45 (14.12)		46.14(14.09)
Baseline iGFR		46.03 (12.85)		45.25(13.23)
Baseline UPCR		0.31 (0.49)		0.33(0.54)