Comparing methods for eGFR Slope estimates among Adult CKD Patients in a UK Electronic Health Record System: Challenges and Opportunities

Ling Zhang1, Chengan Du1, Ruth Farmer2, Remi Poopla2, Christina Shay1
1. Boehringer Ingelheim, Ridgefield, CT, USA,
2. Boehringer Ingelheim, Bracknell, BRC, UK

MSR21

Objectives

This study aims to compare the statistical performance of linear and non-parametric models in estimating eGFR slope using real-world data, with the goal of enhancing the precision of slope estimation in the chronic kidney disease (CKD) population.

Introduction

- Estimated glomerular filtration rate (eGFR) is a measure of kidney function commonly used as an outcome in clinical trials and real-world studies.
- Estimating eGFR slope can be challenging due to high measurement variability. Although linear models are commonly used, use of non-linear models has recently gained attention by providing more accurate estimates for such complex outcomes.

Methods

- Using Clinical Practice Research Datalink AURUM data (2010-2019), patients with chronic kidney disease (CKD) (≥18 years) defined as ≥2 serum creatinine measurements <60 ml/min/1.73m² 90 to 365 days apart were included in the analyses.
- The index date was the second serum creatinine measurement date. eGFRs was calculated from serum creatinine values using the CKD-EPI Creatinine Equation (2021)[1]. Patients with ≥4 eGFR measurements over the 10-year study period with at least one year between the first and last tests were included.
- eGFR slope was estimated using the following approaches[2]:
- Linear regression
- Theil/median-based linear regression(MBLM)
- Quantile regression(RQ) with Tau = 0.5
- Generalized additive models (GAM)
- Locally Weighted Scatter-plot Smoother (LOESS) using nonparametric kernel regression with linear regression

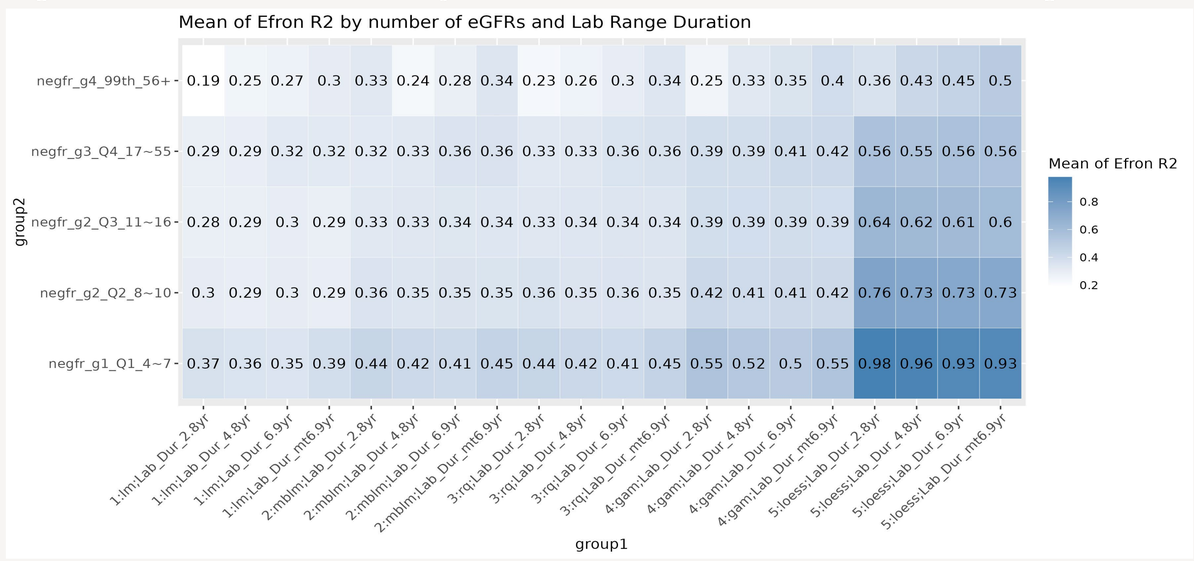
Results

Table 1. Summary of performance of slope estimates models among ~100,000 patients

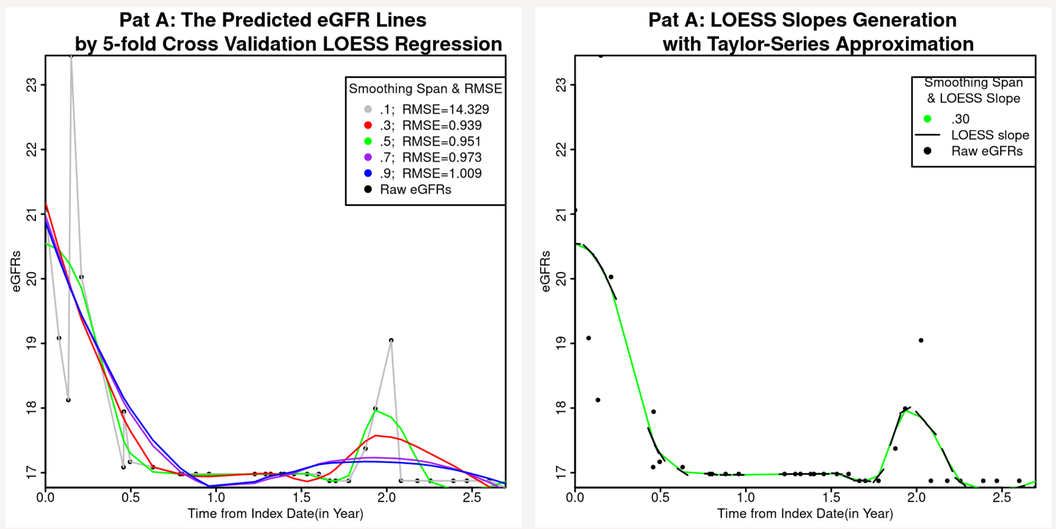
Model Full Name	Results	RMSE[a]	NRMSE mean[b]	Efron R^2	CV Prcnt[c]
Linear Regression*	Mean (SD)	4.96, 2.71	0.1, 0.05	0.32, 0.28	9.93, 5.49
	Median (IQR)	4.51 (3.1-6.31)	0.09 (0.06-0.13)	0.26 (0.07-0.53)	9.06 (6.3-12.6)
Theil/Median Based Linear Regression*	Mean (SD)	6, 7.43	0.12, 0.16	-0.47, 31.85	12.08, 16.36
	Median (IQR)	5.14 (3.53-7.31)	0.1 (0.07-0.14)	0.07 (-0.14-0.39)	10.3 (7.17-14.5)
Quantile Regression*	Mean (SD)	5.34, 2.95	0.11, 0.06	0.2, 0.36	10.68, 5.99
	Median (IQR)	4.82 (3.34-6.74)	0.1(0.07-0.13)	0.15(-0.03-0.46)	9.67(6.8-13.4)
Generalized additive model	Mean (SD)	4.36, 2.54	0.09 ,0.05	0.44, 0.3	8.72, 5.10
	Median (IQR)	3.99 (2.64-5.66)	0.08 (0.05-0.11)	0.43 (0.18-0.69)	8.04 (5.38-11.2)
Locally-weighted polynomial regression (LOESS)	Mean (SD)	2.74, 2.38	0.06, 0.05	0.73, 0.26	5.53, 4.81
	Median (IQR)	2.55 (0.47-4.15)	0.05 (0.01-0.08)	0.78 (0.54-0.99)	5.27 (0.92-8.38)

*Linear regression-based nonparametric methods, as linear extended models, can generate slopes in the results.
GAM and, especially LOESS regression provided better to good description of eGFR's trends, but the model didn't provide any formula or estimates.
[a]Root Mean Squared Error
[b]Mean of Normalized Root Mean Squared Error
[c]Percentage of SD/Mean

Figure 1: Model Performance by number of eGFRs and data collection range



Figures 2 & 3: Optimal LOESS model assessment and slopes set generation



Summary

- Among ~100,000/99,999 patients from the CKD cohort with 1.3 million eGFR measures, mean and standard deviation(SD)) of two measures of eGFRs data per patient are as follows:
  - The mean (SD) number of eGFR records was 13.25 (10.52) and the mean follow-up time for lab data collection range (in years) was 4.89 (2.40).
- The Pearson correlation coefficients of slope estimates by linear and extended models:
  - Linear vs. MBLM: 0.74.
  - Linear vs. Quantile Regression: 0.92
  - MBLM vs. Quantile Regression: 0.70
- The model performance (Efron R^2 and RMSE) are presented in Table 1:
  - The LOESS model (general span = 0.75) achieved the highest accuracy, with median Efron R^2: 0.78 (IQR: 0.54–0.99) and mean RMSE: 2.74 (SD: 2.38).
  - Comparative Efron R^2 values and RMSE for other models:
    - Linear: 0.26 (IQR: 0.07–0.53), RMSE: 4.96 (SD: 2.71)
    - GAM: 0.43 (IQR: 0.18–0.69), RMSE: 4.36 (SD: 2.54)
    - MBLM: 0.07 (IQR: 0–0.39), RMSE: 6.00 (SD: 7.43)
    - Quantile Regression: 0.15 (IQR: 0–0.46), RMSE: 5.34 (SD: 2.95)
- Efron R^2 estimates show the highest Efron R^2 across different models when number of eGFRs is between 4-7 and collection date range within 2 to 4 years, which is matched to clinical observation (Figure 1)
- Using per patient's data, optimal LOESS span length could be assessed by cross validation[3]. Taking Patient A (with 34 eGFRs for 2.6 years) as example, optimal LOESS span length is 0.3 by 5-fold cross validation with the lowest RMSE value (Figure 2).
- Using optimal LOESS model, the predicted eGFR at each time point was produced. And slopes at each time point were generated by Taylor-Series approximation method (Figure 3).
- Overall, LOESS model with patient-specific tuning provides the best accuracy, though at a higher computational cost, especially in large datasets.

Conclusions

Using data from CPRD Aurum, the LOESS model demonstrated high accuracy in eGFR slope estimation in CKD patients with ≥4 eGFR results based, outperforming linear and extended linear methods.

LOESS regression could include covariates in the model, which are gradually used in complex biological outcomes[4]. This could be explored further in later study.

Disclosure Statement
- The study was funded by Boehringer Ingelheim
- The authors meet criteria for authorship as recommended by the ICMJE

References
1. Pottel H, Björk J, Courbebaisse M, et al. Development and validation of a modified full age spectrum creatinine-based equation to estimate glomerular filtration rate: A cross-sectional analysis of pooled data. Annals of Internal Medicine. 2021;174(2):183–191. doi:10.7326/M20-4366
2. Mangiafico S. Summary and Analysis of Extension Program Evaluation in R. 2016. Retrieved from https://rcompanion.org/handbook/F_12.html
3. Steven P. Sanderson II, Unveiling the Magic of LOESS Regression in R: A Step-by-Step Guide with mtcars. Retrieved from https://www.r-bloggers.com/2023/11/unveiling-the-magic-of-loess-regression-in-r-a-step-by-step-guide-with-mtcars/
4. Xie L, Liu J, Wang X, Liu B, Li J, Li J, Wu H. Role of dietary inflammatory index in the association of NT-proBNP with all-cause and cardiovascular mortality in NHANES 1999-2004. Sci Rep. 2024 Aug 28;14(1):19978. doi: 10.1038/s41598-024-70506-3.

Contact Information :
Ling Zhang - Ling_3.zhang@boehringer-Ingelheim.com

Boehringer Ingelheim