Annals of Epidemiology xxx (2018) 1–8

Contents lists available at ScienceDirect

Annals of Epidemiology

journal homepage: www.annalsofepidemiology.org

Original article

Time-varying coefficient of determination to quantify the explanatory power of biomarkers on longitudinal GFR among children with chronic kidney disease

Derek K. Ng, PhD [a,*], Anthony A. Portale, MD [b], Susan L. Furth, MD PhD [c,d], Bradley A. Warady, MD [e], Alvaro Muñoz, PhD [a]

[a] Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD
[b] Division of Nephrology, Department of Pediatrics, University of California, San Francisco, CA
[c] Division of Nephrology, Department of Pediatrics, Children's Hospital of Philadelphia, Philadelphia, PA
[d] Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania, Philadelphia, PA
[e] Division of Nephrology, Department of Pediatrics, Children's Mercy Kansas City, Kansas City, MO

ARTICLE INFO

Article history:
Received 23 June 2017
Accepted 7 May 2018
Available online xxx

Keywords:
Epidemiologic methods
Coefficient of determination
Chronic kidney disease
Pediatrics
Glomerular filtration rate
Proteinuria
Fibroblast growth factor

ABSTRACT

Purpose: Coefficients of determination (R^2) for continuous longitudinal data are typically reported as time constant, if they are reported at all. The widely used mixed model with random intercepts and slopes yields the total outcome variance as a time-varying function. We propose a generalized and intuitive approach based on this variance function to estimate the time-varying predictive power (R^2) of a variable on outcome levels and changes.
Methods: Using longitudinal estimated glomerular filtration rate (eGFR) from the Chronic Kidney Disease in Children Study, linear mixed models characterized the R^2 for two chronic kidney disease (CKD) risk factors measured at baseline: a traditional marker (proteinuria) and a novel marker (fibroblast growth factor 23 [FGF23]).
Results: Time-varying R^2 divulged different disease processes by risk factor and diagnoses. Among children with glomerular CKD, time-varying R^2 for proteinuria had significant upward trends, suggesting increasing power to predict eGFR change, but crossed with FGF23, which was higher up to 2.5 years from baseline. In contrast, among those with nonglomerular CKD, proteinuria explained more than FGF23 at all times, and time-varying R^2 for each risk factor was not substantially different from time-constant estimates.
Conclusions: Proteinuria and FGF23 explained substantial eGFR variability over time. Time-varying R^2 can characterize predictive roles of risk factors on disease progression, overcome limitations of time-constant estimates, and are easily derived from mixed effects models.

Background

Clinical epidemiologists are often interested in the prospective predictive power of biomarkers. Determining how much variability of an outcome is explained by the level of a predictive biomarker over time provides an important epidemiological characterization of disease progression. Several methods have been proposed that summarize predictive power by the coefficient of determination (referred to as R^2) in the setting of longitudinal data: typically this is a single fixed value that describes the proportion of variability of the outcome of interest over time explained by an exposure [1–8]. However, there is a reasonable biological expectation that in some settings, the predictive power of a biomarker on an outcome is time-varying and the assumption that R^2 is constant over time may not be appropriate. To our knowledge, few studies have quantified this type of dynamic relationship.

Linear mixed effects models are commonly used as conventional tools to parametrically characterize longitudinal changes of a continuous outcome and how these are modified by an exposure [9–14]. In addition to providing estimates of outcome levels and changes, the variance components serve to characterize the

* Corresponding author. Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe Street Room E7642, Baltimore, MD 21205.
E-mail address: dng@jhu.edu (D.K. Ng).

https://doi.org/10.1016/j.annepidem.2018.05.002
1047-2797/

behavior of the outcome variance over time, enabling the calculation of R^2. For the simpler case of random intercepts only, it is well known that the variance of the outcome is simply the sum of the between individuals' variance and the within individuals' variance. Hence, the associated R^2 is time constant [2,5,7]. In contrast, in the more common case of allowing random intercepts and slopes, the variance of the outcome is time-varying and so is the associated R^2.

The purpose of the present study was to construct and compare time-fixed and time-varying R^2 values in the context of pediatric chronic kidney disease (CKD) progression. Methodologically, we show cases where the time-constant R^2 oversimplifies the behavior of the data and we discuss the need to enrich models to ensure that the proposed model of random intercepts and slopes suffice. Epidemiologically, we aimed to compare separately and combined the baseline levels of urine protein (a marker of kidney injury [15–20]) and baseline levels of fibroblast growth factor 23 ([FGF23]; a bone-derived hormone and a novel CKD risk factor associated with mineral metabolism [21–23]). Urine protein is a ubiquitous clinical measurement; in contrast, FGF23 is currently not a clinical biomarker. Finally, we sought to demonstrate how heterogeneity of effects by underlying CKD diagnoses can improve epidemiologic understanding of these conditions. This applied example seeks to highlight the utility of time-varying explained variability, the parameters of which are easily obtained in widely used linear mixed models.

Methods

Study population

The chronic kidney disease in children (CKiD) cohort

The CKiD study is an ongoing pediatric cohort study of CKD who were enrolled between aged 1 and 16 years at 54 pediatric nephrology centers from the United States and Canada. Eligibility included a diagnosis of CKD and an estimated glomerular filtration rate (eGFR) between 30 and 90 mL/min|1.73 m^2 [24] with 891 children were enrolled before April 2014. Briefly, clinical and demographic information, including biological specimens for immediate analysis and for repository storage, were collected at annual study visits. All biological specimens were analyzed at the CKiD central laboratory, with the exception of plasma C-terminal FGF23, which was measured using repository specimens at the University of California, San Francisco [23]. Full details of the study protocol have been previously published [25].

Outcome

At annual study visits, the primary outcome was eGFR based on the 2012 CKiD equation, which utilized serum creatinine, cystatin c, blood urea nitrogen, as well as sex and height data [24]. Participants who contributed at least two visits with eGFR data were included.

Exposures

The time origin for this analysis was the first visit when the two predictors of interest were measured: proteinuria (i.e., urine protein to creatinine ratio, mgP/mgCr) and plasma FGF23 (RU/ml). This first visit was considered the baseline or index visit and the primary objective was to determine how these two markers explained the variability of eGFR at the index visit and variability of trajectories. The interassay and intra-assay coefficients of variation for creatinine were 2.1% and 0.8%, respectively, and for proteinuria they were 3.8% and 5.2%, respectively (Roche Diagnostics, Indianapolis, IN). For FGF23, the interassay and intra-assay coefficients of variation were 11.5% and 5.7%, respectively (Immutopics Int., San Clemente, CA).

All analyses were stratified by underlying CKD diagnosis, classified as either nonglomerular (largely congenital kidney abnormalities) or glomerular (mostly noncongenital diseases) [26]. Previous studies have described the heterogeneity of disease progression by the glomerular and nonglomerular classifications [11,14,19,27–29]. In addition, those with glomerular diagnoses were enrolled with higher eGFR than children with nonglomerular CKD.

Statistical methods

Four linear mixed models (random intercepts and slopes) were fit: the null (or reference) model included only time as the independent variable; the second model included baseline proteinuria as the independent variable; the third model included baseline FGF23 as the independent variable; and the fourth model included both baseline proteinuria and FGF23 as independent variables, allowing both to modify the intercepts and the effects of time. To make the outcome closer to a Gaussian distribution, eGFR was converted to the natural log scale, whereas proteinuria and FGF23 were $\log_2$ transformed so that the regression coefficients could be interpreted as the effect of doubling the levels of each of the predictors.

At the j^{th} visit of the i^{th} participant occurring t_{ij} years from baseline, the mixed models were of the form:

$$\log(eGFR_{ij}) = \alpha Z + a_i + (\beta Z + b_i)t_{ij} + e_{ij}, \quad (1)$$

where $Z = [1]$, $Z = [1,\ log_2(proteinuria_{i0})]$, $Z = [1,\ log_2(FGF23_{i0})]$, and $Z = [1,\ log_2(proteinuria_{i0}),\ log_2(FGF23_{i0})]$ for models 1 to 4, respectively, and the number of α and β coefficients for each corresponding model were 2, 4, 4, and 6.

The between-subject deviations from the population intercept and slope were distributed as

$$\begin{pmatrix} a_i \\ b_i \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}\right]$$

The within-subject deviations were assumed to be independent from the between-subject deviations and distributed as

$$e_{ij} \sim N\left(0,\ \sigma^2\right)$$

Therefore, the total variance of eGFR (in the log scale) at t years from baseline is determined by

$$Var(\log(eGFR)) = \sigma_1^2 + 2\sigma_{12}t + \sigma_2^2 t^2 + \sigma^2 \quad (2)$$

Biomarkers as independent variables may explain (reduce) any of the components of the total variance (from the between individual variability of initial values, σ_1^2, and slopes, σ_2^2, to the within individual variance, σ^2). Separate comparisons of the three components of variance may yield the inadmissible result of negative R^2 and fail to incorporate the expected covariance between the random intercepts and slopes (σ_{12}). Combining the components to describe the total variance of the outcome over time overcomes these limitations.

The R^2 corresponds to the ratios of total variance from models 2 to 4 to that of model 1 (i.e., null). In particular, for the fourth model with the two biomarkers, this is formally expressed as

$$R^2 = \left[1 - \frac{Var[\log(eGFR)]_{Z=(1,\ log_2(proteinuria),\ log_2(FGF23))}}{Var[\log(eGFR)]_{Z=(1)}}\right] \times 100 \quad (3)$$

The R^2 values are the proportions of total eGFR variability explained by a risk factor(s) that is not explained by passage of time alone.

Please cite this article in press as: Ng DK, et al., Time-varying coefficient of determination to quantify the explanatory power of biomarkers on longitudinal GFR among children with chronic kidney disease, Annals of Epidemiology (2018), https://doi.org/10.1016/j.annepidem.2018.05.002