Paper 162-2013

Variance Partition: My mission and ambition come to fruition

Brenda Beaty and Miriam Dickinson, Colorado Health Outcomes Program, University of Colorado Anschutz Medical Campus, Denver, CO

ABSTRACT

In medical research, we are often interested in understanding the complex interplay of variables with one or more clinical outcomes. Because our bodies are always in motion, simply viewing a snapshot of data in time is sub-optimal. Longitudinal data gives us the advantage of modeling 'real-life' time-dependent variables and outcomes. This paper is an exploration of one such project. In this paper, we will first familiarize ourselves with a study of the relationship of diabetic nephropathy and blood pressure measured longitudinally. We will then explore a number of ways to model the data with the final goal of using time-varying covariates to model the trajectory of "the average patient", using complete partitioning of the variance.

INTRODUCTION

Diabetes mellitus (DM) affects more than 20 million Americans, and is expected to become ever more common given the epidemic of obesity in our country. Diabetic nephropathy, which is the damage to the kidneys' ability to filter toxins from the blood, is one serious complication of diabetes. It can affect up to 40% of individuals with diabetes, and can damage the kidneys so severely that dialysis is required. Two tests widely used to measure the extent of kidney damage are the urine albumin/creatinine ratio (ACR) and estimated glomerular filtration rate (eGFR). Two of the main ways to prevent and/or slow down kidney damage are blood pressure control and blood sugar control.

GOAL

We aim to describe within-person level associations as well as between-person level associations with the outcome over time.

STUDY POPULATION

This study used data from the hypertension registry of Denver Health, a nationally recognized, integrated safety-net health care delivery system in inner-city Denver, Colorado. Denver Health provided care to more than 140,000 persons in Denver County in 2007. The registry contains information about people with hypertension who received care at Denver Health between January 1, 2000 and December 31, 2008.

We excluded individuals who were <21 years of age at diabetes diagnosis, those with ICD-9-CM codes associated with pregnancy or delivery at any point after the first diagnosis of diabetes, those who did not have at least two measures of the outcome after diabetes diagnosis at least 180 days apart, and those who didn’t have at least one value of the important time-varying covariates prior to or simultaneous with first outcome measure to use as baseline values.

ANALYSIS VARIABLES

The main dependent variable is the urine albumin/creatinine ratio (ACR). It is an estimate of the 24-hour urine albumin excretion. When the kidneys are damaged, they start to leak and protein (albumin) passes into the urine. A *higher value* means *worse* kidney function. Due to a skewed distribution, we log transformed this measure. We used the date of the first measurement of ACR as the baseline date from which we measured time.

Important independent variables that can be used as time-varying covariates are:

- SBP (systolic blood pressure); higher is worse
- HgA1c (glycated hemoglobin); shows the average level of blood glucose (sugar) over the previous 3 months; higher is worse

Other non-time-varying Independent variables included age at baseline, duration of DM at baseline, duration of observation in the DH system, dyslipidemia (high cholesterol), race/ethnicity, insurance type (type), gender, marital status, language spoken, other vascular condition (cardiac arrhythmia, congestive heart failure, heart valve disease, peripheral vascular disease, and coronary artery disease), and eGFR (estimated Glomerular Filtration Rate, another measure of kidney function), as these have been reported to be important in other studies.

A note on logarithm syntax in SAS®:
Log (x) returns the natural (base e) logarithm and log10 (x) returns the logarithm to the base 10.

SAMPLE OF DATA SET

Study_No	Mo6	Gender	ACR	LogACR	SBP	HgA1c
1	0.00	F	14.0	2.6	106	12.6
1	3.62	F	24.0	3.2	140	7.6
1	6.44	F	8.1	2.9	131	7.3
2	0.00	M	103.3	4.6	185	7.7
2	5.36	M	50.2	3.9	188	6.3

ANALYSIS PLAN

For all models, in order to account for the longitudinal nature of these data, we used a random intercept, random slope growth curve mixed effects model with unstructured covariance using SAS PROC MIXED:.

We will construct 3 models, as follows:

Model 1 will include baseline covariates and selected interactions with time tested to see if slope differs by certain baseline characteristics.

Model 2 will model the outcome over time incorporating time-varying SBP and HgA1c.

Model 3 will model the *change* in outcome from baseline as a function of change in SBP and HgA1c, using time-varying SBP and HgA1c deviations from the individual's mean.

For all models, we scaled systolic blood pressure, so that we could interpret the coefficient as change per 10 mm Hg instead of 1 mm Hg as follows:

```
Sys10=systolic/10;
```

We also scaled time in 6 month increments, and grand-mean centered the following continuous variables: age at baseline (BLAge), duration of DM at baseline (DxToBL), eGFR, systolic blood pressure (Sys10), and glycated hemoglobin (HgA1c). An easy SQL method of doing this is shown below.

```
proc sql;
    create table Model1 as
    select *,
    (BLAge - mean(BLAge)) as GMCBLAge,
    (DxToBL-mean(DxToBL)) as GMCDxToBL,
    (eGFR-mean(eGFR)) as GMCeGFR,
    (Sys10-mean(sys10)) as GMCSys10,
    (HgA1c-mean(HgA1c)) as GMCHgA1c
    from m1;
quit;
```

Note that for sake of brevity, we have not included all variables in code and results. However, all models are adjusted for language, time since diabetes diagnosis, presence of vascular disease, and baseline eGFR.

MODEL 1: Non-Time-Varying - Baseline variables and slope over time only

The first model uses only variables measured at or before the first ACR measurement. For each repeated outcome measure, the data have the same values as those at baseline.

This model uses all data where ACR is measured (N=3342 observations among 1304 patients)

Table 1-Data. Example listing of data for Model 1

Study_No	Mo6	Log (ACR)	SBP	GMC Sys10	Sys10	CSys10	GMC HgA1c	HgA1c	CHgA1c
1	0.00	2.6	106	13.6	10.6	-3.0	8.5	12.6	4.1
1	3.62	3.2	140	13.6	10.6	-3.0	8.5	12.6	4.1
1	6.44	2.9	131	13.6	10.6	-3.0	8.5	12.6	4.1

To build the model, we first examine each independent variable as the only variable in addition to time in the mixed model to see if it is associated with the intercept. One way to do this is by using a macro like the one below:

```
%macro c (var);
proc mixed data=model1 covtest noclprint;
  class Study_no &var;
  model LogACR=Mo6 &var /solution ddfm=kr;
  random intercept Mo6/ subject=Study_no type=un;
run;
%mend c;

%c(Gender) %c(Race), etc.
```

This macro is for categorical variables. Another one could easily be constructed for continuous data simply by leaving the '&var' out of the CLASS statement. In model building, we start by including all variables where the p value was <0.5 (we used a high cutoff to try and capture all important variables), and did a backward selection, checking at each step for evidence of confounding. We determined which of the baseline characteristics significantly affect the intercept. These variables included age, gender, race, baseline SBP and baseline HgA1c.

Next, we want to examine whether or not variables affect the slope of the model (i.e. does change over time vary by baseline characteristics such as gender, etc.?) This can be done using a slightly different macro, as shown below:

```
%macro sl (var);
proc mixed data=model1 covtest noclprint;
  class Study_no Gender Race;
  model logACR=GMCTZeroAge Gender Race GMCSys10 GMCHgA1c Mo6 &var*Mo6/solution
ddfm=kr;
  random intercept Mo6/subject=Study_no type=un;
run;
%mend sl;

%sl(Gender) %sl(Race), etc.
```

Interaction terms that were significant were retained in the model. The final model is shown below:

```
proc mixed data=model1 covtest noclprint;
  class Study_no Gender Race;
  model LogACR=GMCBLAge Gender Race GMCSys10 GMCHgA1c Mo6 GMCSys10*Mo6/solution
ddfm=kr;
  random intercept Mo6/subject=Study_no type=un;
run;
```

Table 1-Results. Results of Model 1: Estimated log (ACR) progression analysis (n=3342 observations among 1304 subjects)

Variables	Intercept Coefficient (SE)	p-value
Variables that affect the intercept		
Intercept*	3.34 (0.12)	<.0001
Age at baseline (years)	-0.011 (0.004)	0.01
Female Gender	-0.27 (0.09)	0.002
Race/Ethnicity		
African American	0.02 (0.15)	
Latino	0.19 (0.13)	
Other race	0.70 (0.27)	
White	Ref.	0.03
Previous or baseline Systolic BP (centered, per 10 points)	0.21 (0.02)	<.0001
Previous or baseline HgA1c (centered)	0.17 (0.02)	<.0001
Variables that affect Slope	Slope Coefficient (SE)	p-value
Time (per 6 months)**	-0.004 (0.008)	0.60
Baseline Systolic BP x time (per 6 months)***	-0.021 (0.004)	<.0001

* Log (ACR) at baseline when covariates=reference group
** Change in log (ACR) per 6 months from baseline when covariates=reference group
*** Additional change in log (ACR) for each 10 point increase in baseline SBP

Increasing age and being female are associated with a lower value of baseline log (ACR), while being non-white, having higher SBP at baseline, and having higher HgA1c at baseline are associated with a higher value of baseline log (ACR). There is a small, non-significant decrease (-0.004 points per 6 months) in the log (ACR) over time, and a greater decline in log (ACR) per 6 months for each 10 point increase in baseline SBP. One possible explanation could be that those with higher baseline SBP are sicker, and therefore have more 'room' for improvement in SBP control, and can therefore improve (lower) the log (ACR) more quickly over time once they enter treatment.

MODEL 2: Time-Varying Systolic BP and HgA1c

For models 2 and 3, we can only use data where ACR, SBP and HgA1c are measured on the same day (N=2134 observations among 923 patients), because we want the values of these variables to be measured at the same time as ACR.

Systolic BP and HgA1c are still centered at the grand mean, and are measured *at the same time* as ACR, as shown in the example data below.

Table 2-Data. Example listing of data for Model 2

Study_No	Mo6	Log (ACR)	SBP	GMCSys10	Sys10	CSys10	GMCHgA1c	HgA1c	CHgA1c
1	0.00	2.6	106	13.6	10.6	-3.0	8.5	12.6	4.1
1	3.62	3.2	140	13.6	14.0	0.4	8.5	7.6	-0.9
1	6.44	2.9	131	13.6	13.1	-0.5	8.5	7.3	-1.2

```
proc mixed data=model2 covtest noclprint;
  class Study_No Gender Race;
  model LogACR=GMCBLAge Gender Race GMCSys10 GMCHgA1c Mo6/solution ddfm=kr;
  random intercept Mo6/ subject=Study_No type=un;
run;
```

How is this code different from Model 1? The only code we change is to delete GMCSys10*Mo6, since now we use GMCSys10 as a time-varying covariate. Otherwise, it is the data that change, not the code.

Table 2-Results. Results of Model 2: Estimated log (ACR) progression analysis
(n=2134 observations among 923 subjects)

Variables	Intercept Coefficient (SE)	p-value
Variables that affect the intercept		
Intercept*	3.29 (0.15)	<.0001
Age at baseline (years)	-0.012 (0.005)	0.01
Female Gender	-0.20 (0.10)	0.04
Race/Ethnicity		
African American	-0.02 (0.18)	
Latino	0.12 (0.16)	
Other race	0.59 (0.31)	
White	Ref.	0.20
Variables that affect Slope	Slope Coefficient** (SE)	p-value
Systolic BP (per 10 points)	0.18 (0.01)	<.0001
HgA1c (per point)	0.10 (0.01)	<.0001
Time (per 6 months)***	0.014 (0.009)	0.14

Many of the same associations with the intercept are the same as in Model 1: increasing age and female gender are still associated with a lower baseline log (ACR) value, and 'Other' race is associated with a higher baseline log (ACR) value. In terms of the slope, we see that higher measures of SBP and HgA1c (at the same time of ACR measurement) are strongly associated with higher log (ACR) over time. This is more informative than Model 1, as we now know that these three measurements (ACR, SBP, and HgA1c) track together over time.

MODEL 3: Fully partitioned model using change from baseline in ACR as outcome and partitioned Time-Varying Systolic BP and HgA1c without baseline covariates

When we model log (ACR) over time, we assume that a person's SBP and HgA1c are related to their log (ACR) at the same point in time. It might also be plausible that they are related to the *change* in log (ACR) from baseline at that time point.

We explore both the within-person effect of SBP and HgA1c as well as the between-person effect of SBP and HgA1c by separating each term into two variables: a person-specific mean value (non-time-varying, accounting for between-person effects), and the person's *deviation* from their own mean value at a certain point in time (time-varying, accounting for within-person effects). In other words, we model the association of a person's *change* in ACR from baseline as a function of their *deviation* in SBP and HgA1c from their person-specific means for those variables.

We also create a change in log (ACR) variable by subtracting each person's baseline log (ACR) from the other measures. This gives the change in log (ACR) from baseline within the individual.

With this model, baseline covariates are unimportant, because we model only the change from baseline.

```
* CREATE PERSON-CENTERED SYSTOLIC BP AND HGA1C;
proc means data=model3 noprint;
  class study_no;
  var Sys10 Hga1c;
  output out=person1 mean=PCSys10 PCHgA1c;
run;

data person2 (drop=_type_ _freq_); *N=923;
  set person1;
  * DELETE FIRST RECORD WITH GROUP MEANS;
  if study_no=. then delete;
run;
```

Here is how one record looks in data set 'person2':

Study_No	PCSys10	PCHgA1c
1	12.6	9.2

```
data model3step2; *N=2134;
  merge model3 two;
  by study_no;
  PSys10Dev=Sys10-PCSys10;
  PHgA1cDev=HgA1c-PCHgA1c;
run;
```

Table 3a-Data. Example listing of data for Model 3

Study_No	Mo6	Sys 10	PCSys10	PSys10 Dev	HgA1c	PCHgA1c	PHgA1c Dev
1	0.00	10.6	12.6	-2.0	12.6	9.2	3.4
1	3.62	14.0	12.6	1.4	7.6	9.2	-1.6
1	6.44	13.1	12.6	0.5	7.3	9.2	-1.9

```
data model3final;
  set model3step2;
  retain first;
  by study_no;
  if first.study_no then do; first=logacr; diff=0; end;
  else diff=logacr-first;
run;
```

Table 3b-Data. Example listing of data for Model 3

Study_No	Mo6	LogACR	Diff	Sys 10	PCSys10	PSys10 Dev	HgA1c	PCHgA1c	PHgA1c Dev
1	0.00	2.6	0.0	10.6	12.6	-2.0	12.6	9.2	3.4
1	3.62	3.2	0.5	14.0	12.6	1.4	7.6	9.2	-1.6
1	6.44	2.9	0.3	13.1	12.6	0.5	7.3	9.2	-1.9

```
proc sort data=model3fin; by study_no mo6; run;

proc mixed data=model3fin covtest noclprint;
  class study_no;
  model diff=mo6 PCSys10 PSys10Dev PCHgA1c pHgA1cDev/solution ddfm=kr;
  random intercept mo6/ subject=study_no type=un;
run;
```

Table 3-Results. Estimated log (ACR) progression analysis
(n=2134 observations among 923 subjects)

Variables	Intercept Coefficient (SE)	p-value
Variables that affect the intercept		
Intercept*	-0.06 (0.10)	0.30
Mean person-specific Systolic BP (per 10 points)	0.003 (0.007)	0.68
Mean person-specific HgA1c (per point)	0.002 (0.006)	0.78
Variables that affect Slope	Slope Coefficient (SE)	p-value
Time (Per 6 Months)**	0.004 (0.011)	0.75
Time-varying difference in Systolic BP from mean person-specific Systolic BP (per 10 points)	0.08 (0.01)	<.0001
Time-varying difference in HgA1c from mean person-specific HgA1c (per point)	0.02 (0.01)	0.002

* Population-based change in log ACR from baseline to next measurement
** Population-based change in log ACR from baseline per 6 months

Coefficients of the intercepts and slopes are *change* in estimated log (ACR) from baseline as a function of a person's mean SBP and HgA1c and time-specific *deviations* from those means.

This is the most informative model of all. From this model, we can say that mean values of SBP and HgA1c are not significantly associated with the change in log (ACR). It is the *time-specific deviation* from that person's mean value that has a much stronger association with the change in log (ACR).

The implication is that no matter what a person's overall mean SBP and HgA1c, what matters most are the *trajectories* of SBP and HgA1c.
The interpretation is that the between-subjects effects are not significantly associated with *change* in log (ACR), and that the within-person deviation from their mean SBP at any given time is strongly associated with (in the same direction) their change in log (ACR) at that point in time. Thus, improvement in SBP and HgA1c are associated with improvement in kidney function, as measured by log(ACR); worsening of SBP and HbA1c are associated with greater decline in kidney function, as measured by log(ACR).

CONCLUSION

Taking advantage of longitudinal data can yield much more specific information about the effects of independent variables on the dependent variable. Here we were able to show that a person's mean blood pressure and glycated hemoglobin are not nearly as important in determining change in kidney function from baseline as variations from those means over time.

ACKNOWLEDGEMENTS

The author gratefully acknowledges Kirk Lafler for his kind help and thoughtful review.

RECOMMENDED READING

- Hedeker D, Gibbons R. Longitudinal Data Analysis. 2006, Hoboken, Hew Jersey: Wiley & Sons, pp 69-76.
- Neuhaus JM, Kalbfleisch JD. Between- and within cluster covariate effects in the analysis of clustered data. Biometrics. 1998 Jun;54(2):638-45.
- Fairclough DL: Design and analysis of quality of life studies in clinical trials. New York, Chapman & Hall/CRC, 2002.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Brenda Beaty, MSPH
Colorado Health Outcomes Program
University of Colorado Denver
Mail Stop F443
13199 E. Montview Ave., Suite 300
Aurora, CO 80045-0508
Work Phone: (303) 724-1076
E-mail: Brenda.Beaty@ucdenver.edu