Paper 71-2010

Using HPMIXED with Other SAS® 9.2 Procedures to Efficiently Analyze Large Dimension Registry Data

Matthew C. Fenchel, M.S., Cincinnati Children's Hospital Medical Center, Cincinnati, OH
Gary L. McPhail, M.D., Cincinnati Children's Hospital Medical Center, Cincinnati, OH
Rhonda D. VanDyke, Ph.D., Cincinnati Children's Hospital Medical Center, Cincinnati, OH

Abstract

HPMIXED is an experimental procedure introduced in SAS/STAT® 9.2 software. Using sparse matrix techniques, PROC HPMIXED can process models with a very large number of fixed or random effects much more efficiently than the MIXED or GLIMMIX procedures. However, this initial release lacks some tools which are often needed in analyses.

With Registry data (1994-2007) from the Cystic Fibrosis Foundation, we estimate FEV_1% predicted (a standard measure of lung function) slope using a random intercept and slopes mixed model and compare parameter estimates and run-times of PROC HPMIXED with PROC MIXED and PROC GLIMMIX. We then use the SGPLOT and UNIVARIATE procedures to produce residual, normal and quantile-quantile plots from the PROC HPMIXED output. Using the covariance parameter estimates from PROC HPMIXED, we show how PROC GLIMMIX and PROC MIXED can be used to "re-run" the models (now more quickly) to produce output not yet available in PROC HPMIXED (i.e. residual diagnostics, influence diagnostics, LS-means comparisons, etc.). Finally, we use a simulated registry data set to further demonstrate diagnostic tools after obtaining covariance parameter estimates from PROC HPMIXED.

PROC HPMIXED is a powerful, consistent tool for fitting large dimension data – especially mixed modeling. Such models might not be possible using PROC MIXED or PROC GLIMMIX. The applications described in this paper can be used by someone with an intermediate knowledge of mixed models and SAS procedures, using computing resources with at least 3 GB of RAM.

Introduction

With the release of SAS/STAT® 9.2 software came a new tool for general linear mixed models – the experimental procedure HPMIXED. The purpose of PROC HPMIXED is to provide a more efficient process (both in time and memory) to fit models that use a large number of fixed or random effects – or where the data set itself is very large. Such models are often simply not practical with standard desktop computers, or even servers, using the MIXED, GLIMMIX or NLMIXED procedures. PROC HPMIXED was created to help fill that void.

(Author's note: after drafting this paper, SAS/STAT® 9.22 became available in late May 2010. With this release, PROC HPMIXED is production.[1] It also includes a number of enhancements. Except for the LSMEANS statement, none of those enhancements alter what is discussed in this paper. The changes in the LSMEANS statement are discussed later.)

The statistical and computational characteristics of PROC HPMIXED (along with an introduction to mixed modeling in general) have already been well described by Wang and Tobias (2009)[2] and will not be discussed here. Rather, we want to look at which tools are not yet available in PROC HPMIXED and demonstrate applications of other SAS procedures – including PROC MIXED and PROC GLIMMIX – that can "bridge the gap" until the time that such applications are available in PROC HPMIXED. Some of these include residual diagnostics, influence diagnostics, and least-squares means comparisons. SAS® Enterprise Guide 4.2 was used on a HP ProLiant (DL585 G2) server with four Opteron 8218 2.6 GHz dual core processors and 32 GB of RAM. (Using the models described here, no more than 3 GB of RAM was required. Also, a desktop PC can be used, as long as sufficient memory is available.)

Data and Model

Registry data – which can contain thousands of patients, with dozens of observations per patient – is a likely candidate for using PROC HPMIXED. In specific, we used 1994-2007 patient Registry data from the Cystic Fibrosis Foundation[3] (CFF). The CFF collects clinical encounter and annual information from all patients providing consent and being treated at U.S. CF care centers. We were interested in assessing whether cystic fibrosis (CF) patients who had been prescribed dornase alfa (Pulmozyme, produced by Gentech) had better one-second forced expiratory volume percentage of predicted values (FEV_1% – a standard measure of lung function) than patients (Control) who had never been prescribed dornase alfa (DA). Of particular interest was assessing the average FEV_1% slope over time for both groups (DA and Control). Additionally, it was hypothesized that the CF care centers where patients were treated would constitute a random effect. From the medical literature, it was known that fixed effects such as gender, baseline FEV$_1$%, insurance-type, testing positive for certain infections / conditions (MRSA, pseudomonas, CF-related diabetes) and pancreatic insufficiency may affect lung function. Thus, we initially modeled post FEV$_1$% (repeated measurements after baseline) against the above fixed effects, with random intercepts and slopes for patient and center.

The general linear mixed model can be expressed as follows, for the k^{th} observation from the j^{th} patient in the i^{th} CF care center:

$$Y_{ijk} = \mu + \eta_j + \delta_j + \zeta_j + \psi_{jk} + \tau_{jk} + \varphi_{jk} + \omega_{jk} + \upsilon_j + \xi_k + (\delta\xi)_k + (\eta\xi)_k + a_j + b_j\xi_{jk} + c_i + d_i\xi_{ik} + e_{ijk}$$

where

μ = Overall mean

η_j = Group (receiving dornase alfa therapy = 1)

δ_j = Male

ζ_j = Private insurance

ψ_{jk} = Positive for MRSA

τ_{jk} = Positive for Pseudomonas

φ_{jk} = Positive for CF related diabetes (CFRD)

ω_{jk} = Taking pancreatic enzymes

υ_j = Baseline FEV$_1$%

ξ_k = Time (years) since baseline

$(\delta\xi)_k$ = Time since baseline by male

$(\eta\xi)_k$ = Time since baseline by Group (dornase alfa = 1)

a_j = Patient (random intercept)

$b_j\xi_{jk}$ = Patient by time since baseline (random slope)

c_i = CF care center (random intercept)

$d_i\xi_{ik}$ = Time since baseline by CF care center (random slope)

e_{ijk} = Random residual

$$\begin{bmatrix} a_j \\ b_j \end{bmatrix} \sim iid\ N\left(\begin{bmatrix} \alpha_1 \\ \beta_1 \end{bmatrix}, G_1\right) \quad \boldsymbol{G_1} = \begin{bmatrix} \sigma_a^2 & 0 \\ 0 & \sigma_b^2 \end{bmatrix}$$

$$\begin{bmatrix} c_i \\ d_i \end{bmatrix} \sim iid\ N\left(\begin{bmatrix} \alpha_2 \\ \beta_2 \end{bmatrix}, G_2\right) \quad \boldsymbol{G_2} = \begin{bmatrix} \sigma_c^2 & 0 \\ 0 & \sigma_d^2 \end{bmatrix} \quad e_{ijk} \sim iid\ N(0, \sigma^2)$$

Run Times

To compare run-times for identical models in PROC HPMIXED, PROC GLIMMIX and PROC MIXED, a subset of 26 CF care centers (CID), comprising 1,195 patients (PID) and 16,330 observations, was used. The SAS code for each procedure is below.

```
proc hpmixed data=dna6 noclprint=200;
   where TimeDiff ne 0 and CID le 9030;
   class Gender CID Group Cul_MRSA Cul_PseudoA C_CFRD Pancreat Ins_PX;
   model Cli_FEV1_PCT = Gender|TimeDiff Group|TimeDiff BaseFEV1 Cul_MRSA
                        Cul_PseudoA C_CFRD Pancreat Ins_PX / s;
   random Intercept TimeDiff / subject = PID;
   random intercept TimeDiff / subject = CID;
   test Gender|TimeDiff Group|TimeDiff BaseFEV1 Cul_MRSA Cul_PseudoA C_CFRD
         Pancreat Ins_PX;
run;
```

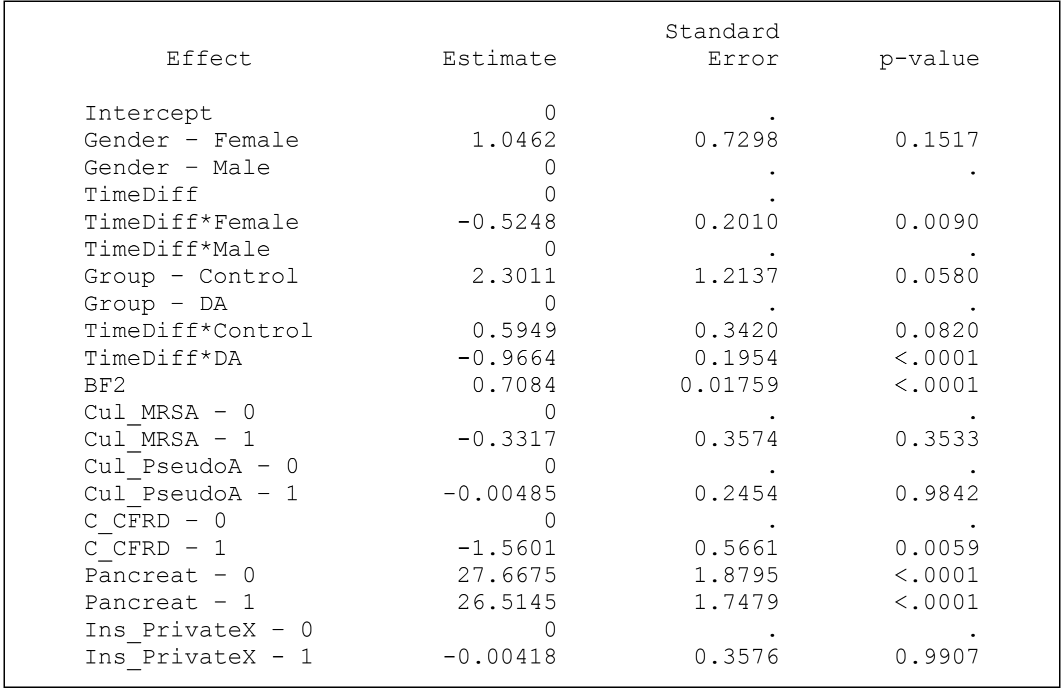
```
proc glimmix data=dna6 noclprint=200;
   where TimeDiff ne 0 and CID le 9030;
   class Gender CID Group Cul_MRSA Cul_PseudoA C_CFRD Pancreat Ins_PX;
   model Cli_FEV1_PCT = Gender|TimeDiff Group|TimeDiff BaseFEV1 Cul_MRSA
                        Cul_PseudoA C_CFRD Pancreat Ins_PX / s;
   random Intercept TimeDiff / subject = PID;
   random intercept TimeDiff / subject = CID;
run;

proc mixed data=dna6 noclprint=200;
   where TimeDiff ne 0 and CID le 9030;
   class Gender CID Group Cul_MRSA Cul_PseudoA C_CFRD Pancreat Ins_PX;
   model Cli_FEV1_PCT = Gender|TimeDiff Group|TimeDiff BaseFEV1 Cul_MRSA
                        Cul_PseudoA C_CFRD Pancreat Ins_PX / s;
   random Intercept TimeDiff / subject = PID;
   random intercept TimeDiff / subject = CID;
run;
```

All procedures used maximum likelihood estimation (ML). The “Test” statement in PROC HPMIXED is necessary because – unlike PROC MIXED and PROC GLIMMIX – PROC HPMIXED does not automatically calculate / output the Type III tests (F tests) of fixed effects. For mixed models with many levels of fixed covariates, computing the Type III tests can greatly decrease computational efficiency.[4] (However, for the sake of an “apples-to-apples” comparison with PROC GLIMMIX and PROC MIXED, the “Test” statement was included here.) The SAS data set “DNA6” was sorted by PID prior to running any models, so that we would not have to list PID in the CLASS statement, yet still obtain correct estimates for the random patient intercepts and slopes.

PROC HPMIXED successfully ran the model in a real-time of 6.89 seconds; PROC GLIMMIX required 59 minutes, 58 seconds, while PROC MIXED needed 7 minutes and 38 seconds (as recorded by the respective SAS logs). Covariance parameter estimates between procedures were identical, as were the fit statistics (output omitted for this model). However, the parameter estimates (as given in the output) sometimes differed, for the following reason: because of the sparse matrix techniques in PROC HPMIXED, the system of linear equations (and the resulting singularities) does not always match-up with the order of effects entering the model.[4] This results in a different generalized inverse than either PROC MIXED or PROC GLIMMIX produces, which use a sweep method based on the order of the fixed effects.

Figures 1a and 1b show results from the “Solutions for Fixed Effects” tables produced by PROC HPMIXED and PROC GLIMMIX. (Results from PROC MIXED are omitted, which produced estimates in the same order as PROC GLIMMIX.)

Figure 1a – “Solutions for Fix Effects” from PROC HPMIXED (CFF Registry Data)



```
                                                 Standard
       Effect                Estimate              Error        p-value

    Intercept                       0                  .
    Gender - Female            1.0462             0.7298         0.1517
    Gender - Male                   0                  .              .
    TimeDiff                        0                  .
    TimeDiff*Female           -0.5248             0.2010         0.0090
    TimeDiff*Male                   0                  .              .
    Group - Control            2.3011             1.2137         0.0580
    Group - DA                      0                  .              .
    TimeDiff*Control           0.5949             0.3420         0.0820
    TimeDiff*DA               -0.9664             0.1954         <.0001
    BF2                        0.7084            0.01759         <.0001
    Cul_MRSA - 0                    0                  .              .
    Cul_MRSA - 1              -0.3317             0.3574         0.3533
    Cul_PseudoA - 0                 0                  .              .
    Cul_PseudoA - 1          -0.00485             0.2454         0.9842
    C_CFRD - 0                      0                  .              .
    C_CFRD - 1                -1.5601             0.5661         0.0059
    Pancreat - 0              27.6675             1.8795         <.0001
    Pancreat - 1              26.5145             1.7479         <.0001
    Ins_PrivateX - 0                0                  .              .
    Ins_PrivateX - 1         -0.00418             0.3576         0.9907
```

Figure 1b – “Solutions for Fix Effects” from PROC GLIMMIX (CFF Registry Data)

```
                                              Standard
      Effect               Estimate              Error       p-value

 Intercept                  24.6139             1.8551        <.0001
 Gender - Female             1.0462             0.7298        0.1517
 Gender - Male                    0                  .             .
 TimeDiff                   -0.9664             0.1954        0.0006
 TimeDiff*Female            -0.5248             0.2009        0.0090
 TimeDiff*Male                    0                  .             .
 Group - Control             2.3011             1.2137        0.0580
 Group - DA                       0                  .             .
 TimeDiff*Control            1.5613             0.3297        <.0001
 TimeDiff*DA                      0                  .             .
 BF2                         0.7084            0.01759        <.0001
 Cul_MRSA - 0                0.3317             0.3574        0.3533
 Cul_MRSA - 1                     0                  .             .
 Cul_PseudoA - 0           0.004815             0.2454        0.9843
 Cul_PseudoA - 1                  0                  .             .
 C_CFRD - 0                  1.5601             0.5661        0.0059
 C_CFRD - 1                       0                  .             .
 Pancreat - 0                1.1530             0.7195        0.1091
 Pancreat - 1                     0                  .             .
 Ins_PrivateX - 0          0.004224             0.3576        0.9906
 Ins_PrivateX - 1                 0                  .             .
```

P-values can also differ in a small way, as the default in PROC MIXED and PROC GLIMMIX is the containment method for this type of model, while PROC HPMIXED offers only the residual method (in addition to “NONE”).

(Although not shown here, we do want to mention that PROC HPMIXED offers one table in the output that is not offered in any form by PROC MIXED or PROC GLIMMIX – a “Descriptive Statistics” table for all continuous variables. This includes the mean, standard deviation, coefficient of variation (%), minimum value and maximum value for each variable.)

The end message here though, is the fact that PROC HPMIXED took much less time than either PROC MIXED or PROC GLIMMIX. Extrapolated to even larger, more complex data sets, the savings in time and memory would certainly be considerable.

Residual Plots

Residual scatter plots, quantile-quantile plots and normal plots are valuable in assessing the normality of the data (residuals) and the homogeneity of the residual variance. PROC HPMIXED does not yet offer such plots, nor can one use ODS graphics with PROC HPMIXED. Fortunately, though, PROC HPMIXED does offer an OUTPUT statement (also offered in PROC GLIMMIX, but not in PROC MIXED) for computing three types of residuals – RESIDUAL, PEARSON, STUDENT – and each version has two options, BLUP (conditional) and NOBLUP (marginal). Using code similar to what is given below, scatter, normal and Q-Q plots can be produced – in similar fashion to what is available directly in PROC MIXED and PROC GLIMMIX. Results are given in Figures 2a, 2b and 2c.

```
proc hpmixed data=dna6 noclprint=200;
   where TimeDiff ne 0;
   class Gender CID Group Cul_MRSA Cul_PseudoA C_CFRD Pancreat Ins_PX;
   model Cli_FEV1_PCT = Gender|TimeDiff Group|TimeDiff BaseFEV1 Cul_MRSA
                         Cul_PseudoA C_CFRD Pancreat Ins_PX / s;
   random Intercept TimeDiff / subject = PID;
   random intercept TimeDiff / subject = CID;
   output out=pr (keep=PID CID p sr) pred=p student=sr;
run;
```

```
ods graphics on;
proc sgplot data=pr2;
   scatter x=p_9004 y=sr_9004 / markerattrs = (symbol = circlefilled color=gray)
                                   name='9004' legendlabel='9004';
   scatter x=p_9005 y=sr_9005 / markerattrs = (symbol = triangle color=black)
                                   name='9005' legendlabel='9005';
   keylegend / title = "Center";
run;

proc univariate data=pr normal;
   where CID in (9004, 9005);
   Var sr;
   qqplot sr / normal(mu=est sigma=est l=2) square ;
   histogram sr / normal (mu=est sigma=est l=2);
   inset mean std min q1 median q3 max / pos = ne format = 6.2;
run;
ods graphics off;
```

(Note: Some minor data reorganization was used before running PROC SGPLOT, in order to plot separate “groups” – i.e. Centers. The DATA step is not shown.)

Figure 2a Plot of studentized residuals from patients in selected CF care centers.

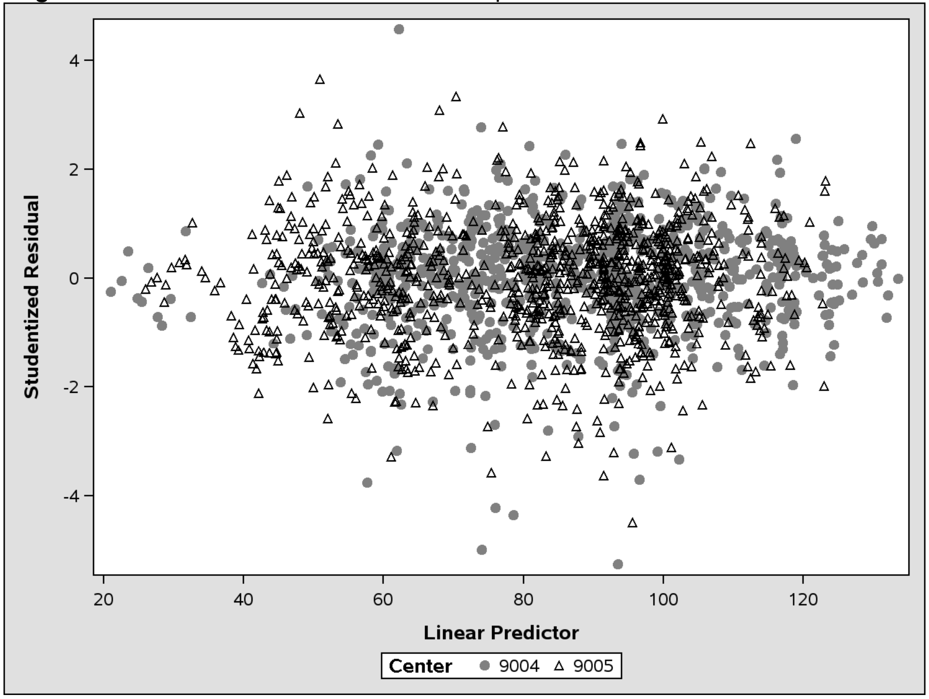


Figure 2b Normal plot of studentized residuals.

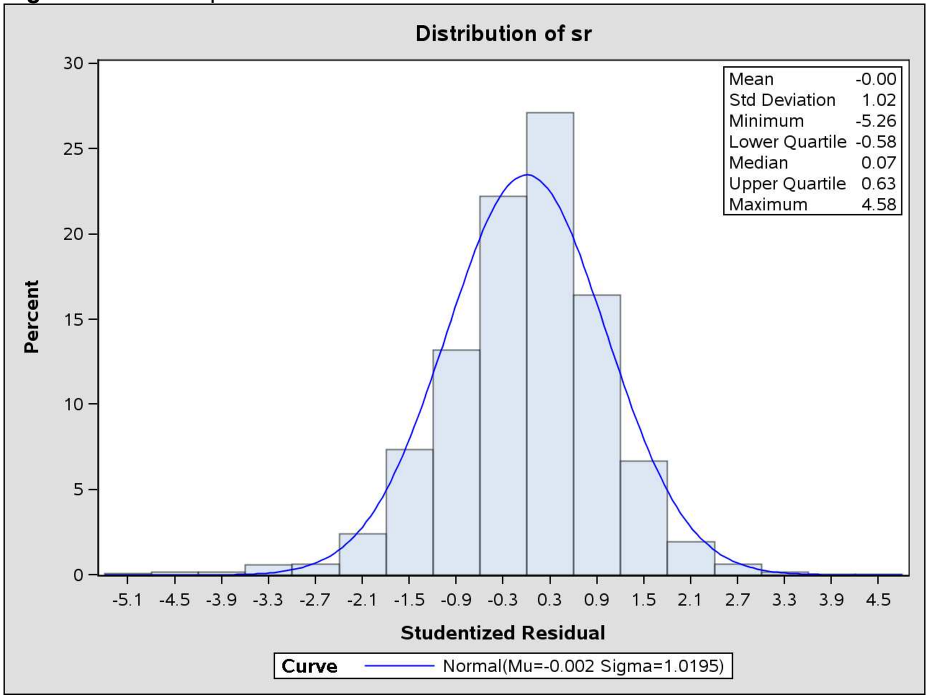
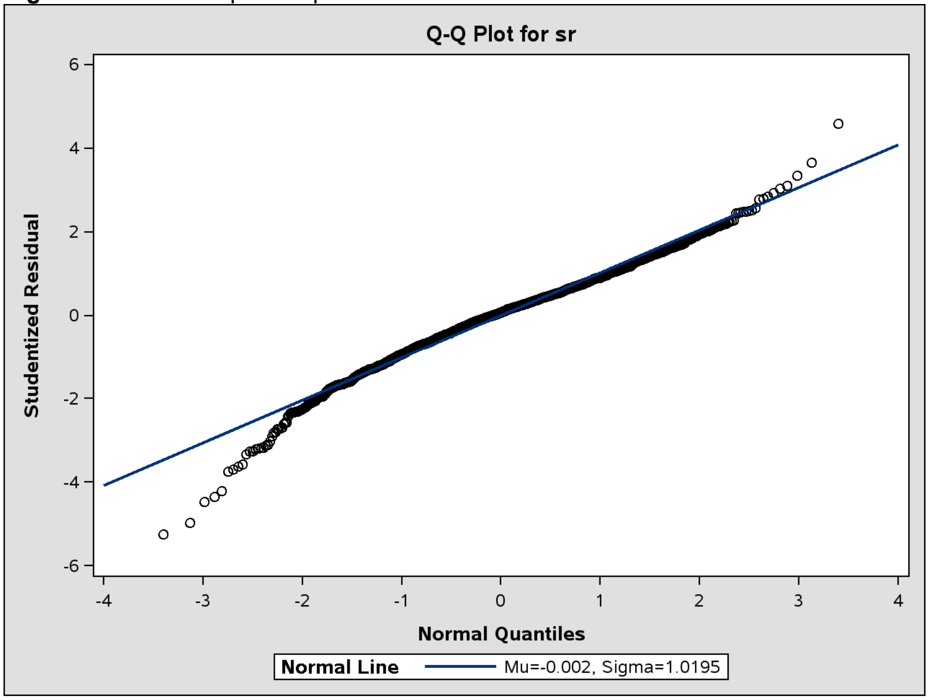


Figure 2c Quantile-quantile plot of studentized residuals.



In the OUTPUT statement, STUDENT = SR specifies outputting the marginal (BLUP is the default option), studentized residuals to the variable named "SR." The data set itself (PR) is given by the OUT option in the OUTPUT statement.

Of course, examining patient-level residuals for such a large, registry data set may not be very clear, or perhaps even of much value. Looking at subsets of patients though, may be instructive. In this case (for demonstration purposes), PROC UNIVARIATE and PROC SPLOT are limited to two CF centers.

Using PROC SPLOT and PROC UNIVARIATE in this fashion is probably the most efficient method of obtaining such plots – but it is not the only way. Using the ODS OUTPUT statement in PROC HPMIXED, one can output the covariance parameter estimates to a SAS data set and then use those estimates to run either PROC GLIMMIX or PROC MIXED. Within those procedures then, one can produce residual plots – in PROC GLIMMIX, using the PLOTS option in the PROC statement; in PROC MIXED, using the RESIDUAL option in the MODEL statement. (SAS code omitted here – see next section for using PROC HPMIXED combined with PROC MIXED or PROC GLIMMIX). Producing graphs in this fashion, however, would require much more time than simply using PROC SPLOT and PROC UNIVARIATE.

As a side note, PROC HPMIXED in this section was run using 217 CF care centers (CID), comprising 6,697 patients (PID) and 167,000+ observations. It took PROC HPMIXED 19 minutes and 45 seconds to run this analysis.

Influence Diagnostics

PROC MIXED offers a number of statistics that can help determine whether one or more observations is having an undue influence on the estimation or precision of parameter estimates, on the estimation of predicted and fitted values, or producing large changes in objective functions. The general idea is to run the full model and compute estimates, then run a reduced model after removing one or more observations (either is possible), and then compare the results based on certain test statistics. (5) This is not available in PROC HPMIXED – but using results from PROC HPMIXED, PROC MIXED can be used for these diagnostic tests.

Of course, using data (such as the CFF Registry data) that contains a large number of patients, running influence diagnostics on a patient-by-patient bases might be very time-consuming and perhaps not that beneficial. However, it might be informative to see if one or more CF care centers may be unduly influencing the estimates. Using the code below, we were able to do just that.

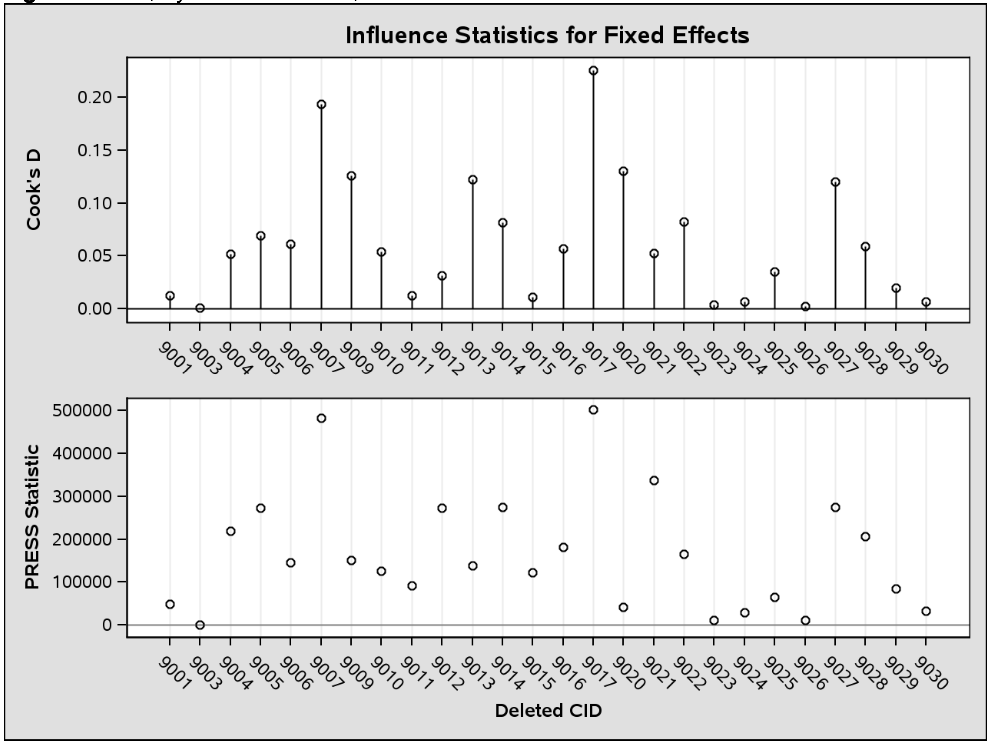
```
proc hpmixed data=dna6 noclprint=200;
   where TimeDiff ne 0 and CIE le 9030;
   class Gender CID Group Cul_MRSA Cul_PseudoA C_CFRD Pancreat Ins_PX;
   model Cli_FEV1_PCT = Gender|TimeDiff Group|TimeDiff BaseFEV1 Cul_MRSA
                        Cul_PseudoA C_CFRD Pancreat Ins_PX / s;
   random Intercept TimeDiff / subject = PID;
   random intercept TimeDiff / subject = CID;
   ods output COVParms=HPME;
run;

ods graphics on;
proc mixed data=dna6 noclprint=200;
   where TimeDiff ne 0 and CIE le 9030;
   class Gender CID Group Cul_MRSA Cul_PseudoA C_CFRD Pancreat Ins_PX;
   model Cli_FEV1_PCT = Gender|TimeDiff Group|TimeDiff BaseFEV1 Cul_MRSA
            Cul_PseudoA C_CFRD Pancreat Ins_PX / influence(effect=CID iter=0) s;
   random Intercept TimeDiff / subject = PID;
   random intercept TimeDiff / subject = CID;
   lsmeans Group / pdiff;
   estimate 'Slope: Control vs. DA' Group*TimeDiff 1 -1;
   parms / pdata=HPME HOLD = 1, 2, 3, 4, 5;
run;
ods graphics off;
```

(Due to excessive memory requirements with PROC MIXED, we can only demonstrate this code using a subset of the CF care-centers – the same subset of 26 as used previously.)

The ODS OUPUT statement is used in PROC HPMIXED to create a SAS data set ("HPME") that contains the covariance parameter estimates. These are then applied using the PARMS statement in PROC MIXED to "re-run" the model – but holding the covariance parameter estimates from PROC HPMIXED fixed (HOLD=), thus saving considerable time and memory. It is important to note here, that if the HOLD option is used in the PARMS statement, the ITER sub-option cannot be used in the INFLUENCE option in the MODEL statement. Otherwise, if ITER = *n*>0 then PROC MIXED will attempt to update covariance estimates even though they are being "held." No INFLUENCE statistics will be produced in this scenario.

It took PROC MIXED more than three hours to run the above code. Graphics results are shown in Figure 3.

Figure 3 Plot, by CF care center, of influence statistics for fixed effects.



It is also possible (results omitted) to run the PROC MIXED code without the HOLD option in the PARMS statement. PROC MIXED then treats the covariance estimates from PROC HPMIXED as starting values for its covariance estimations. One should then use the ITER= sub-option in the INFLUENCE option. PROC MIXED will then generate additional influence statistics – showing which observations (or CF care center in our data) are having the greatest influence on objective functions and covariance parameters. This will be demonstrated in the "Simulations" section.

Simulations

Simulation Model 1

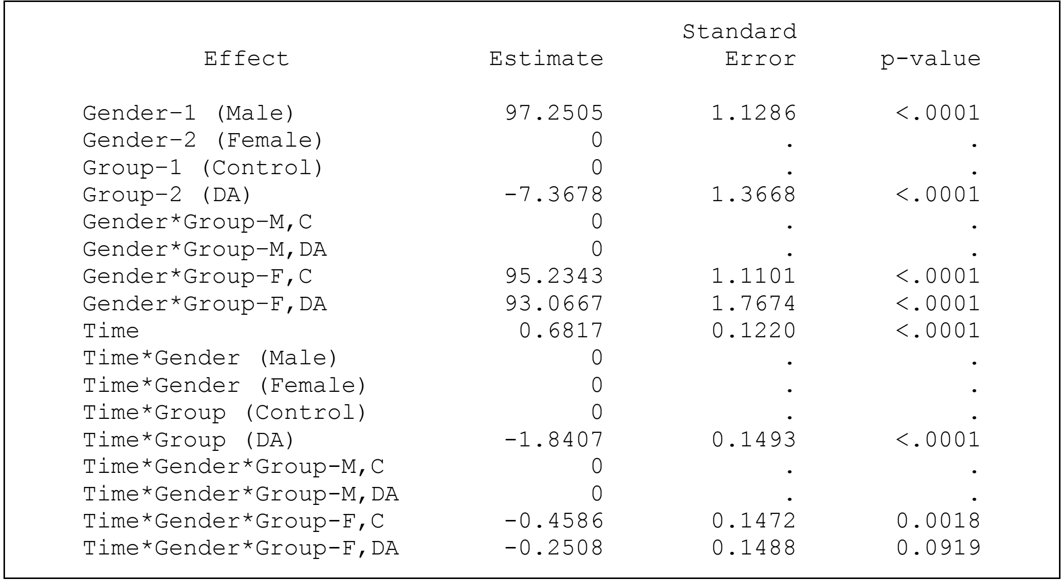
To further test the applicability of using PROC HPMIXED and PROC MIXED to run a model and produce influence diagnostics, we generated a simulated data set of 25 centers, 1,250 patients and 40 observations (quarters – FEV$_1$% is usually measured four times per year) per patient, for a total of 50,000 observations. (A description of the data set and the SAS code can be found in the Appendix.) The SAS code for both procedures follows.

```
proc hpmixed data=hp.hp_sim noclprint=200;
   class Gender Center Group;
   model FEV1 = Gender|Group|Time / noint s;
   random Intercept Time / subject = Patient s;
   random Intercept Time / subject = Center s;
   ods output SolutionR = SCRan;
   ods output COVParms=HPMEshort;
   lsmeans Gender*Group;
run;

proc mixed data=hp.hp_sim noclprint=200;
   class Gender Center Group;
   model FEV1 = Gender|Group|Time / influence(effect=Center iter=5) s noint;
   random Intercept Time / subject = Patient; *s;
   random Intercept Time / subject = Center; *s;
   lsmeans Gender*Group / pdiff;
   parms / pdata=HPMEshort;
run;
```

PROC HPMIXED needed about 13 seconds to run this simulated dataset. It took PROC MIXED (noting that it also had to produce the influence diagnostics) approximately 4.25 hours to complete.

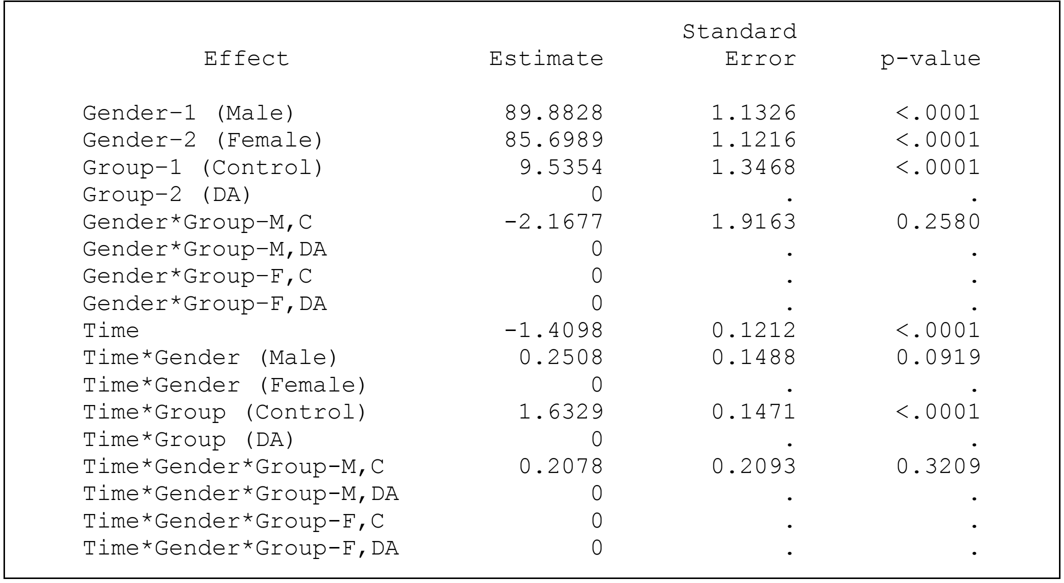
For the sake of simplicity, we did not include the covariates that were used with the CFF registry data. In addition, we used a three-way interaction of Gender, Group and Time to better demonstrate the use of the LSMEANS statement in PROC MIXED. Also important to note is that the PROC MIXED code here does not hold the parameter estimates from PROC HPMIXED fixed, but rather treats them as starting points for its own estimation. (This added to the time needed to run PROC MIXED.) This will allow us to generate additional influence diagnostics. Figures 4a and 4b show the “Solutions for Fixed Effects” from both procedures.

Figure 4a – “Solutions for Fixed Effects” from PROC HPMIXED (Simulation Model 1)



```
                                                Standard
             Effect              Estimate          Error      p-value

     Gender-1 (Male)              97.2505         1.1286       <.0001
     Gender-2 (Female)                  0              .            .
     Group-1 (Control)                  0              .            .
     Group-2 (DA)                 -7.3678         1.3668       <.0001
     Gender*Group-M,C                   0              .            .
     Gender*Group-M,DA                  0              .            .
     Gender*Group-F,C             95.2343         1.1101       <.0001
     Gender*Group-F,DA            93.0667         1.7674       <.0001
     Time                          0.6817         0.1220       <.0001
     Time*Gender (Male)                 0              .            .
     Time*Gender (Female)               0              .            .
     Time*Group (Control)               0              .            .
     Time*Group (DA)              -1.8407         0.1493       <.0001
     Time*Gender*Group-M,C              0              .            .
     Time*Gender*Group-M,DA             0              .            .
     Time*Gender*Group-F,C        -0.4586         0.1472       0.0018
     Time*Gender*Group-F,DA       -0.2508         0.1488       0.0919
```

Figure 4b – “Solutions for Fixed Effects” from PROC MIXED (Simulation Model 1)



```
                                                Standard
             Effect              Estimate          Error      p-value

     Gender-1 (Male)              89.8828         1.1326       <.0001
     Gender-2 (Female)            85.6989         1.1216       <.0001
     Group-1 (Control)             9.5354         1.3468       <.0001
     Group-2 (DA)                       0              .            .
     Gender*Group-M,C             -2.1677         1.9163       0.2580
     Gender*Group-M,DA                  0              .            .
     Gender*Group-F,C                   0              .            .
     Gender*Group-F,DA                  0              .            .
     Time                         -1.4098         0.1212       <.0001
     Time*Gender (Male)            0.2508         0.1488       0.0919
     Time*Gender (Female)               0              .            .
     Time*Group (Control)          1.6329         0.1471       <.0001
     Time*Group (DA)                    0              .            .
     Time*Gender*Group-M,C         0.2078         0.2093       0.3209
     Time*Gender*Group-M,DA             0              .            .
     Time*Gender*Group-F,C              0              .            .
     Time*Gender*Group-F,DA             0              .            .
```

As discussed earlier, the system of linear equations does not always match the order that they enter model in PROC HPMIXED, thus resulting in different output from PROC MIXED. However, for example, the estimate for the intercept for a Male in the Control group (97.2505+0+0=97.2505) from PROC HPMIXED is equal to the estimate from PROC MIXED (89.8828+9.5354-2.1677=97.2505).

Table 4c shows the fixed values that were used as starting values in generating the simulated data, compared with the estimates from PROC HPMIXED (after some basic addition). The estimates from PROC HPMIXED are identical to those estimates from PROC MIXED. The estimates are very close to the starting values, with some slight over-estimations and some slight under-estimations.

Figure 4c – Comparisons between starting values for simulated data and PROC HPMIXED estimates (Simulation Model 1)

```
              Effect          Starting Value      Estimated

     Male-Control                         97        97.2505
     Male-DA                              89        89.8827
     Female-Control                       95        95.2343
     Female-DA                            87        85.6989
     Time*(Male-Control)                0.64         0.6817
     Time*(Male-DA)                    -1.05        -1.1590
     Time*(Female-Control)              0.25         0.2231
     Time*(Female-DA)                  -1.46        -1.4098
```

LSMEANS Comparisons

As discussed in the introduction, the experimental release of PROC HPMIXED in SAS/STAT® 9.2 does not offer the possibility of estimating differences in the least-squares means with the LSMEANS statement. That is now possible in the production version of PROC HPMIXED in SAS/STAT® 9.22. However, even the production release does not offer an ADJUST option for multiple comparison adjustments. Again, using the covariance parameter estimates from PROC HPMIXED in PROC MIXED, the differences can be estimated, including implementing a multiple comparison adjustment. Please see the previous SAS code for PROC MIXED, the relevant section of which is reproduced below:

```
proc mixed data=dna6 noclprint=200;
    .
    lsmeans Gender*Group / pdiff adjust=Tukey;
    .
run;
```

The estimates from PROC HPMIXED and PROC MIXED for the least squares means were identical. Results from PROC MIXED for the least squares means estimates and estimated differences are given in Figures 5a and 5b.

Figure 5a – "Least Squares Means" from PROC MIXED (Simulation Model 1)

Effect	Gender	Group	Estimate	Standard Error	p-value
Gender*Group	Male	Control	100.57	1.2616	<.0001
Gender*Group	Male	DA	84.2328	1.2662	<.0001
Gender*Group	Female	Control	96.3220	1.2411	<.0001
Gender*Group	Female	DA	78.8262	1.2540	<.0001

Figure 5b – "Differences of Least Squares Means" from PROC MIXED (Simulation Model 1)

Effect	Gender / Group	_Gender / _Group	Estimate	Standard Error	p-value	Adjusted p-value
Gender*Group	Male-C	Male-DA	16.3412	1.5252	<.0001	<.0001
Gender*Group	Male-C	Female-C	4.2521	1.5040	0.0047	0.0243
Gender*Group	Male-C	Female-DA	21.7479	1.5176	<.0001	<.0001
Gender*Group	Male-DA	Female-C	-12.0892	1.5108	<.0001	<.0001
Gender*Group	Male-DA	Female-DA	5.4066	1.5203	0.0004	0.0021
Gender*Group	Female-C	Female-DA	17.4958	1.5031	<.0001	<.0001

Influence Diagnostics

By way of review, influence diagnostics are measures of the effect of removing one or more observations from the model, fitting the model again, and then evaluating the differences in fit statistics, fixed effect estimates and variance/covariance parameter estimates between the main model (containing all observations) and the "smaller" model (all observations, except for those that were removed). In our analysis, we were interested in the influence of each Center – i.e. what were the changes in key parameter estimates when all observations from a given Center were removed.

Figures 6a and 6b show graphically the results of the influence diagnostics. (The tables with actual numerical results are omitted here). While there are certainly some notable differences among the 25 Centers, no one or two centers stand out as being exceptional by their absence/presence. As one tool in evaluating the distribution of the simulated data, we conclude that our simulation did a reasonable job of randomly allocating Patients to Centers, and random effects estimates to Patients.

Figure 6a: Plot of "Restricted Likelihood Distance" from PROC MIXED (Simulation 1)

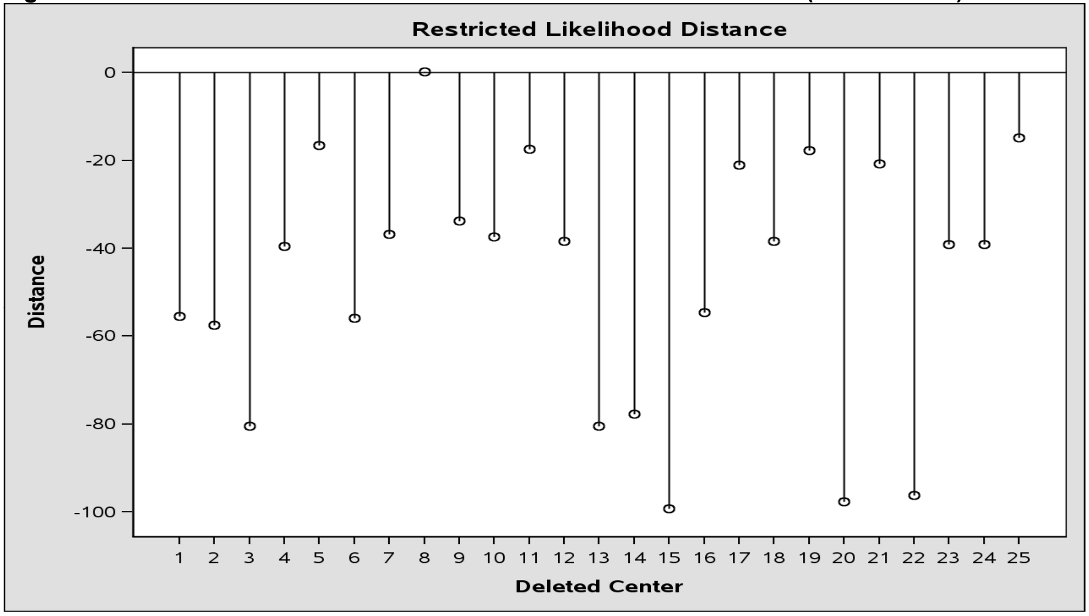
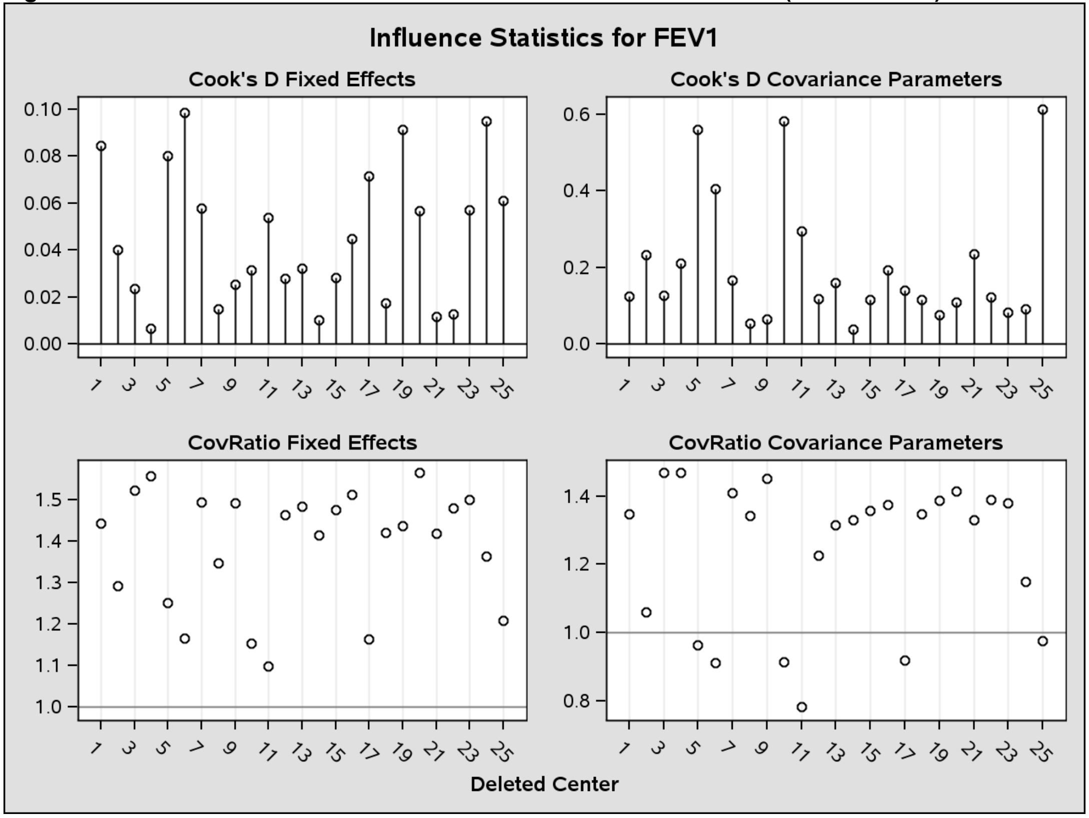


Figure 6b: Plots of additional Influence Statistics from PROC MIXED (Simulation 1)



For additional information on the purpose and use of each influence statistic, see Schabenberger (2004).[6]

Simulation Model 2

Approximately 2.2% of the simulated observations were outside of the regular range of FEV_1% values (typically 20 – 140). The PROC HPMIXED and PROC MIXED models shown above were also run without these “outliers” and the results (Figure 7) were similar to what is reported above. The least squares means estimates (omitted) were also nearly identical to Simulation Model 1.

Figure 7 – “Solutions for Fixed Effects” from PROC HPMIXED (Simulation Model 2)

```
                                                Standard
         Effect                Estimate            Error      p-value

Gender-1 (Male)                 97.3067           1.1181       <.0001
Gender-2 (Female)                     0                .            .
Group-1 (Control)                     0                .            .
Group-2 (DA)                    -7.4688           1.3526       <.0001
Gender*Group-M,C                      0                .            .
Gender*Group-M,DA                     0                .            .
Gender*Group-F,C                95.0398           1.1009       <.0001
Gender*Group-F,DA               93.1097           1.7498       <.0001
Time                             0.5586           0.1175       <.0001
Time*Gender (Male)                    0                .            .
Time*Gender (Female)                  0                .            .
Time*Group (Control)                  0                .            .
Time*Group (DA)                 -1.7189           0.1439       <.0001
Time*Gender*Group-M,C                 0                .            .
Time*Gender*Group-M,DA                0                .            .
Time*Gender*Group-F,C           -0.4094           0.1422       0.0040
Time*Gender*Group-F,DA          -0.2395           0.1436       0.0954
```

Simulation Model 3

For the third simulation model, we used the same data as in the previous two models with one exception: fifty patients (4% of the total number in the data set) were randomly chosen at one Center (also chosen at random) to exhibit a much greater annual decline in FEV_1%. This was accomplished by subtracting 7 points (-7.0) from the fixed, starting slopes each of these fifty were given (based on the random assignments of Gender and Group) in the data simulation. The objective here was to determine how the influence diagnostics in PROC MIXED would identify this “outlier” effect.

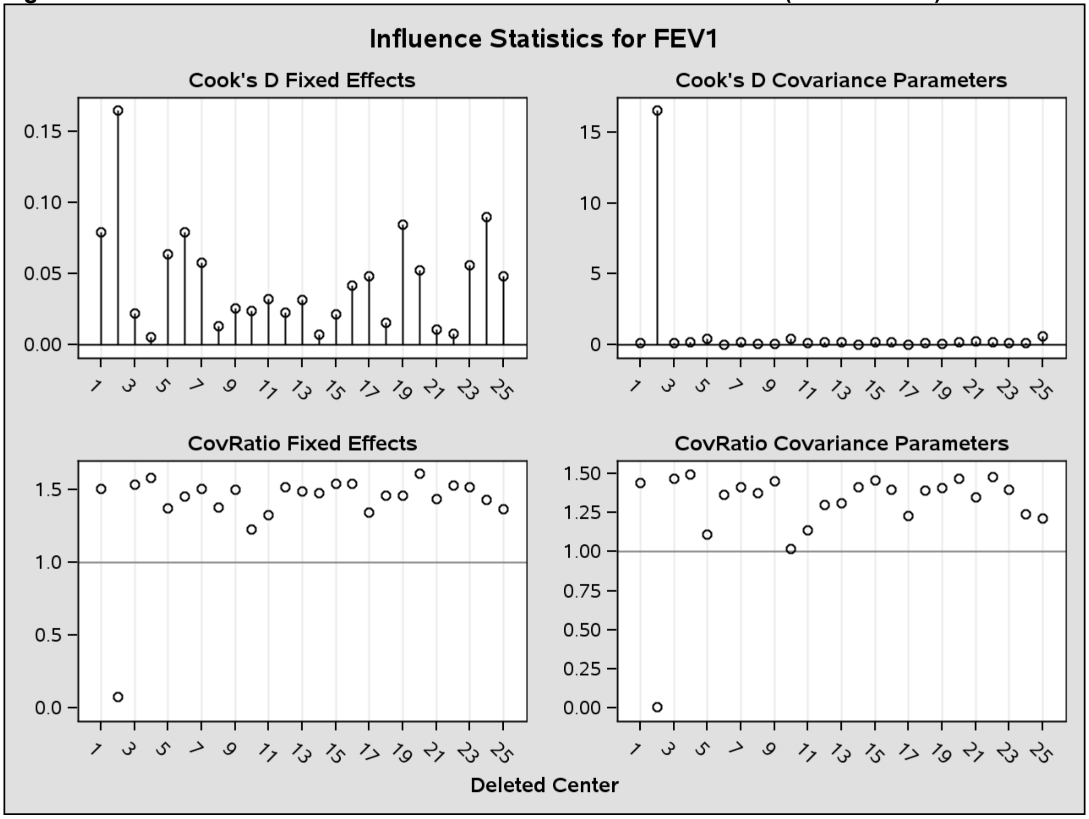
Figure 8a shows the parameter estimates for this model. Most of the estimates are very close to those from Simulation Model 1 – except, of course, the estimate for Time. (In Model 1, the estimate for Time = 0.6817.)

Figure 8a – “Solutions for Fixed Effects” from PROC HPMIXED (Simulation Model 3)

```
                                                Standard
         Effect                Estimate            Error      p-value

Gender-1 (Male)                 97.2485           1.1280       <.0001
Gender-2 (Female)                     0                .            .
Group-1 (Control)                     0                .            .
Group-2 (DA)                    -7.3673           1.3667       <.0001
Gender*Group-M,C                      0                .            .
Gender*Group-M,DA                     0                .            .
Gender*Group-F,C                95.2307           1.1095       <.0001
Gender*Group-F,DA               93.0686           1.7669       <.0001
Time                             0.4601           0.2851       0.1065
Time*Gender (Male)                    0                .            .
Time*Gender (Female)                  0                .            .
Time*Group (Control)                  0                .            .
Time*Group (DA)                 -1.8366           0.1571       <.0001
Time*Gender*Group-M,C                 0                .            .
Time*Gender*Group-M,DA                0                .            .
Time*Gender*Group-F,C           -0.4631           0.1547       0.0028
Time*Gender*Group-F,DA          -0.3090           0.1567       0.0486
```

As expected, Center 2 exhibited noticeably different influence diagnostics from the rest of the Centers (Figure 8b). Not shown is the "Restricted Likelihood Distance" plot. The restricted likelihood distance for Center 2 was more than 219. The next highest Center (10) was 0.67.

Figure 8b: Plots of additional Influence Statistics from PROC MIXED (Simulation 3)



While there is obviously statistical and programming rationale (as mentioned previously) for analyzing an exceptional subset of the data, there is also a medical justification. McPhail, et. al. (2009)[7] showed that a local, unique strain of *Achromobacterxylosoxidans* was strongly associated with a much greater rate of decline in FEV$_1$%. This or other center-specific factors that cannot be anticipated in the experimental design and analysis (especially considering the low percentage of patients affected, compared with the entire CF population) could possibly be initially detected through influence diagnostics, and then possibly accounted for with an additional covariate (Simulation Model 4).

If we include the ESTIMATES sub-option in the INFLUENCE option (in the PROC MIXED MODEL statement), we can also obtain "Fixed Effects Deletion Estimates" and "Covariance Parameter Deletion Estimates" which further delineate where the presence or absence of Center 2 may be having the most influence. The two plots (of three) of most interest to us are shown in Figures 8c and 8d.

Again, as anticipated, Center 2 is having the greatest influence on fixed effect estimate for Time, and on the covariance parameter estimates for the patient slope (Time) and center slope (Time).

Figure 8c: Fixed Effects Deletion Estimates (Simulation 3)

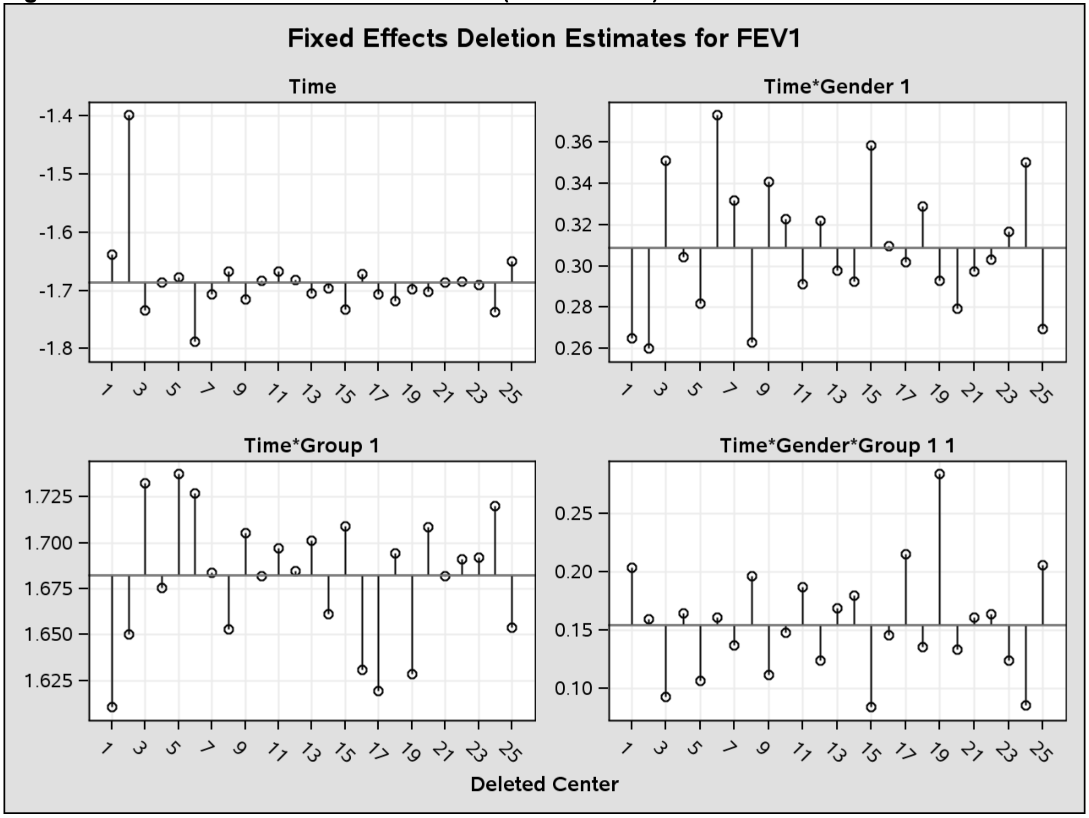
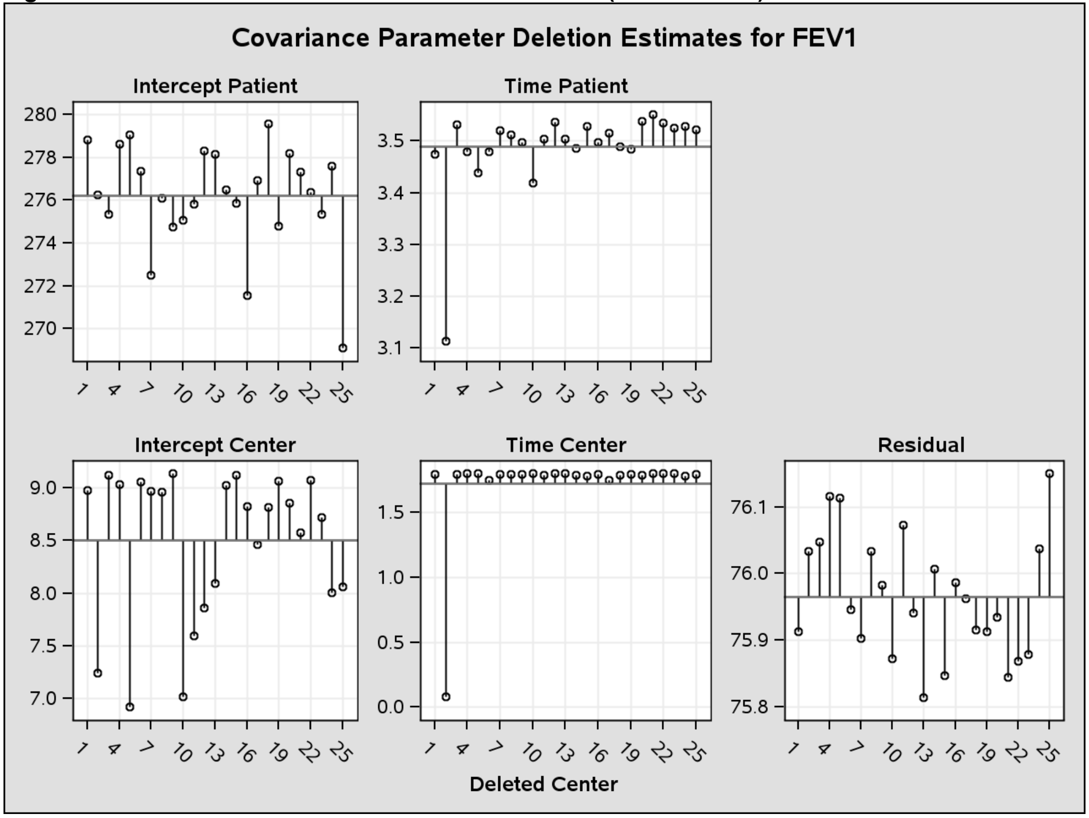


Figure 8d: Covariance Parameter Deletion Estimates (Simulation 3)



Simulation Model 4

In Simulation Model 4 then, we essentially duplicated the analysis from Model 3 – except that a Time*InfectionX fixed effect was added to the model (with 1= infected and 0 = not infected for InfectionX). Figure 9a shows the "Solutions for Fixed Effects" table. With the addition of the covariate, the rest of the results are very close to Simulation Model 1 (Figure 4a).

Figure 9a – "Solutions for Fixed Effects" from PROC HPMIXED (Simulation Model 4)

```
                                                Standard
          Effect                  Estimate         Error      p-value

     Gender-1 (Male)               97.2497        1.1289       <.0001
     Gender-2 (Female)                   0             .            .
     Group-1 (Control)                   0             .            .
     Group-2 (DA)                  -7.3662        1.3666       <.0001
     Gender*Group-M,C                    0             .            .
     Gender*Group-M,DA                   0             .            .
     Gender*Group-F,C              95.2338        1.1105       <.0001
     Gender*Group-F,DA             93.0652        1.7675       <.0001
     Time                           0.7012        0.1232       <.0001
     Time*Gender (Male)                  0             .            .
     Time*Gender (Female)                0             .            .
     Time*Group (Control)                0             .            .
     Time*Group (DA)               -1.8465        0.1493       <.0001
     Time*Gender*Group-M,C               0             .            .
     Time*Gender*Group-M,DA              0             .            .
     Time*Gender*Group-F,C         -0.4596        0.1472       0.0018
     Time*Gender*Group-F,DA        -0.2460        0.1488       0.0983
     Time*InfectionX               -7.4741        0.3872       <.0001
```

The plots for the overall influence diagnostics are omitted for this model. It should be noted though, that there was a loss of rank for Center 2, and the overall influence diagnostics could not be computed for this center. However, this was not the case for the "Fixed Effects Deletion Estimates" and "Covariance Parameter Deletion Estimates", which are shown in Figures 9b, 9c and 9d. The results show that Center 2 is no longer having an undue influence on parameter or covariance estimates, especially those involving Time.

Figure 9b: Fixed Effects Deletion Estimates, Plot 1 (Simulation 4)

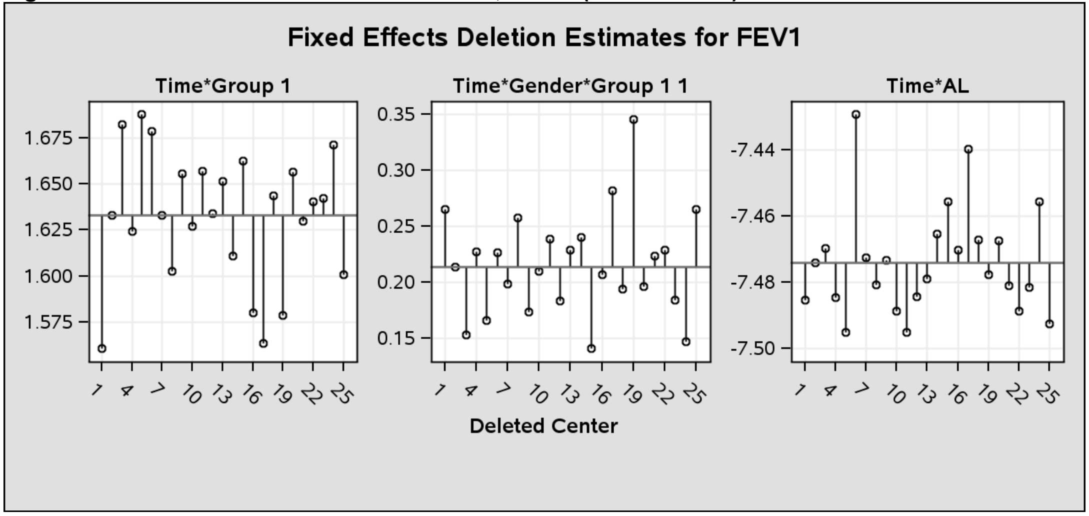


Figure 9c: Fixed Effects Deletion Estimates, Plot 2 (Simulation 4)

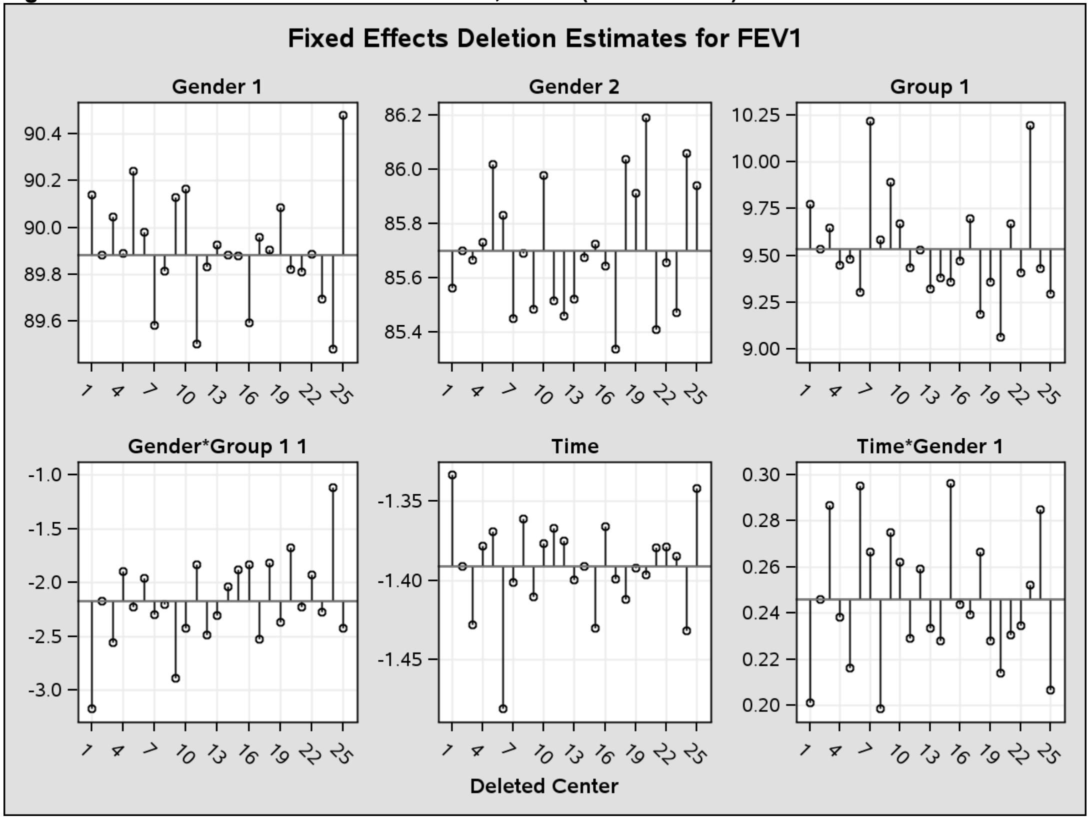
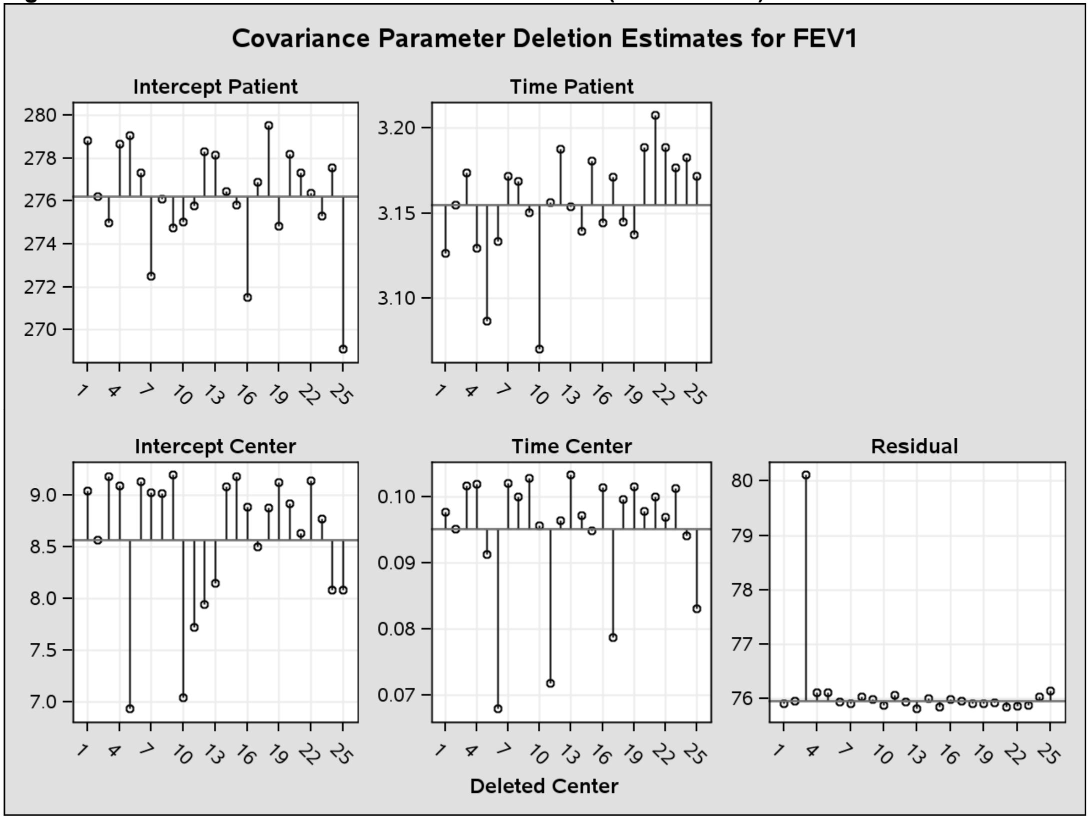


Figure 9d: Covariance Parameter Deletion Estimates (Simulation 4)



Future Work

PROC HPMIXED still lacks some tools necessary for exploring model fit, and for specific comparisons. One possible direction would be to program a variety of residual plots (perhaps similar to PROC GLIMMIX) that could be requested for subsets of the data, assuming the data set has a large number of levels for subjects or other effects. Combined with the inclusion of ODS graphics, this would be a useful addition in exploring model adequacy quickly.

Related to this would be adding options for multiple comparison adjustments in the ESTIMATE, CONTRAST and LSMEANS statements. No multiple comparison adjustments are currently available in these statements, although the ESTIMATE and LSMEANS statements do offer an ALPHA= option in creating confidence intervals, which can obviously be employed if multiple comparisons are being used. In the meantime, one can use PROC MIXED in conjunction with PROC HPMIXED (in a similar fashion as we demonstrated with the LSMEANS statement) to run such statements with multiple comparison adjustments.

Ultimately, of course, influence diagnostics would be a valuable addition. With the speed of PROC HPMIXED, undoubtedly the calculation of influence diagnostics would be much faster than in PROC MIXED.

Conclusion

PROC HPMIXED is a valuable tool in analyzing very large data sets, especially those with a large number of fixed and/or random effects. The savings in time when analyzing large-scale registry data from the CFF was considerable. PROC HPMIXED offers the necessary tests for fixed effects, least squares means and (in the production version) comparison of least squares means. Although not discussed here, it also offers the following statements that are often needed in mixed model analyses – ESTIMATE, CONTRAST, EFFECT and NLOPTIONS – with similar syntax as would be used in PROC MIXED and/or PROC GLIMMIX.

Although currently PROC HPMIXED lacks some tools to explore residual and influence diagnostics, those can be duplicated through the use of other SAS procedures. For residual diagnostics, using PROC UNIVARIATE and PROC SGPLOT (or PROC GPLOT) does not take much additional time in either programming or execution. While using PROC MIXED for influence diagnostics still may take a few hours (depending upon the size of the data set), it is possible to use this procedure and take advantage of the information provided.

References

1. SAS/STAT® 9.22 User's Guide (What's New in SAS/STAT® 9.22 – The HPMIXED Procedure), Second Edition. Copyright 2009 SAS Institute Inc.

2. Wang, T and Tobias, R. (2009) "All the Cows in Canada: Massive MIXED Modeling with the HPMIXED Procedure in SAS 9.2." SAS Institute Inc. 2009. *Proceedings of the SAS® Global Forum 2009 Conference*. Cary, NC: SAS Institute Inc. Paper 256-2009.

3. Cystic Fibrosis Foundation. Cystic Fibrosis Foundation Patient Registry: 1994-2007 clinical and annual data. Bethesda, MD: 2009.

4. SAS/STAT® 9.2 User's Guide (The HPMIXED Procedure), Second Edition. Copyright 2009 SAS Institute Inc.

5. SAS/STAT® 9.2 User's Guide (The MIXED Procedure), Second Edition. Copyright 2009 SAS Institute Inc.

6. Schabenberger, O. (2004) "Mixed Model Influence Diagnostics" *Proceedings of the 29th Annual SAS Users Group International Conference*. Cary, NC: SAS Institute Inc. Paper 189-29.

7. McPhail, Gary MD; VanDyke, Rhonda PhD; Fenchel, Matthew MS; LiPuma, John J. MD; Joseph, Patricia M. MD (2009) "An Update on Clinical Outcomes Associated with a Clonal Strain of Achromobacterxylosoxidans." North American Cystic Fibrosis Conference; October 2009; Minneapolis, Minnesota.

Appendix

The following SAS program was used to generate the main simulated data set. The data set contained 1,250 patients, who were randomly assigned Gender, Center (1 through 25) and Group (Control or DA). Based on a person's Gender and Group, fixed effects for the intercept and slope (Time) were assigned, estimated from CFF registry data. Since we were using a random intercepts and slopes model – for both patient and center – residuals for these random effects were generated using separate (and different) normal distributions. Finally, a random error term for each observation was also generated from a distinct normal distribution. Forty (40) FEV_1% values per patient were calculated from the fixed effects estimates and the random residuals (five terms).

```
ods graphics on;

/* This section randomly assigns a Group, Gender and Center to each patient. */

   %let nCID = 25;                      /*  Number of Centers  */
   %let nPID = %eval(&nCID*50);         /*  Total number of patients = 1250  */

data PatientSim;

   array CID(&nPID);
   array Grp(&nPID);
   array Sex(&nPID);

   do i = 1 to &nPID;
      Grp(i) = 1 + int(2*ranuni(160372));     /* 1 = Control, 2 = DA  */
      Sex(i) = 1 + int(2*ranuni(160372));     /* 1 = Male, 2 = Female */
      CID(i) = 1 + int(&nCID*ranuni(160372)); /* Centers: 1 through nCID (1250) */
      end;

   do j = 1 to &nPID;
      Patient = j;
      Group = Grp{Patient};
      Gender = Sex{Patient};
      Center = CID{Patient};


/* This section assigns a fixed baseline FEV1 (intercept) and fixed FEV1 slope to
   each patient, based on gender and group.  These fixed values were estimated
   from the CFF registry data. */

   if Gender = 1 and Group = 1 then do; BaseFEV1 = 97; FixSlope = 0.64; end;
   if Gender = 1 and Group = 2 then do; BaseFEV1 = 89; FixSlope = -1.05; end;
   if Gender = 2 and Group = 1 then do; BaseFEV1 = 95; FixSlope = 0.25; end;
   if Gender = 2 and Group = 2 then do; BaseFEV1 = 87; FixSlope = -1.46; end;


/* This section assigns a randomly generated patient-intercept residual and
   patient-slope residual to each patient. The estimates for the variance were
   obtained from the CFF registry data. */

   RPI = 17*rannor(160372);     /* Random patient intercept residual ~ N(0,17^2) */
   RPS = 1.75*rannor(160372);   /* Random patient slope residual ~ N(0,1.75^2) */

   output;
   end;
   keep Patient Group Gender Center BaseFEV1 FixSlope RPI RPS RRes;
run;


/* This section assigns a randomly generated center-intercept residual and
   center-slope residual to each center. The estimates for the variance were
   obtained from the CFF registry data. */

data CenterSim;
   do Center = 1 to &nCID;
   RCI = 3.4*rannor(160372);    /* Random center intercept residual ~ N(0,3.4^2) */
   RCS = 0.51*rannor(160372);   /* Random center slope residual ~ N(0,0.51^2) */
   output;
   end;
run;
```

```
/*  Sorting and merging data sets.  */

proc sort data=patientsim;
   by Center;
run;

data AllSim;
   merge patientsim centersim;
   by Center;
run;

proc sort data=AllSim;
   by Patient;
run;


/* This section creates 40 observations per patient (4 per year, 10 years). Each
   observation has a randomly generated residual assigned to it. FEV1 values are
   calculated from the fixed effects, random deviations based on patient and
   center, and random residuals (error terms).  */

data hp.hp_sim; retain Patient FEV1;
   set allsim;
   do Time = 0 to 9.75 by 0.25;
   RRes = 8.7*rannor(160372);                    /* Random error ~ N(0,8.7^2) */
   FEV1 = (BaseFEV1 + RPI + RCI) + Time*(FixSlope + RPS + RCS) + RRes;
   output;
   end;
run;

proc sort data=hp.hp_sim;
   by Patient Time;
run;

/*  End of simulation programming.  */
```

Contact Information

Comments and questions are welcome. The author may be contacted at:

Matthew Fenchel
Division of Biostatistics and Epidemiology
Cincinnati Children's Hospital Medical Center
MLC 5041, 3333 Burnet Avenue
Cincinnati, OH 45229-3039

E-mail: Matthew.Fenchel@cchmc.org