**Paper CC-10**

# Comparisons of SAS® Mixed and Fixed Effects Modeling for Observed over Expected Count Outcomes in the Presence of Hierarchical or Clustered Data

Rachel E. Patzer, Emory University; Laura Plantinga, Emory University

## ABSTRACT

There are numerous SAS® modeling approaches that can be used to model an outcome of a standardized ratio measure (observed/expected counts), such as the frequently encountered Standardized Mortality Ratio (SMR). The purpose of this paper is to examine facility-level predictors associated with another standardized ratio measure---the Standardized Transplant Ratio (STR)---by comparing mixed- and fixed-effects modeling approaches in an analysis of dialysis facilities nested within 18 geographical regions of the US.

In a cross-sectional, multi-level ecologic study using the publicly available Dialysis Facility Report (2007-2010) data, we examined 4,098 dialysis facilities across the US. STRs were defined as the number of observed kidney transplants within a dialysis facility divided by the number of expected transplants, which is determined for each facility by Dialysis Facility Report, based on modeling of patient age and year. We considered the outcome both as linear (STR and log-transformed STR) and as a count (with expected counts or person-years as offsets). We utilized random effects and generalized estimating equation modeling to account for correlation of facilities within regions. We considered SAS PROC MIXED to examine fixed and random effects and PROC GLIMMIX to further examine random effects with the linear outcomes STR and log-STR. We used SAS PROC GENMOD (fixed effects) and PROC GLIMMIX (mixed effects) to examine count outcomes, using a log link and the negative binomial distribution to account for overdispersion.

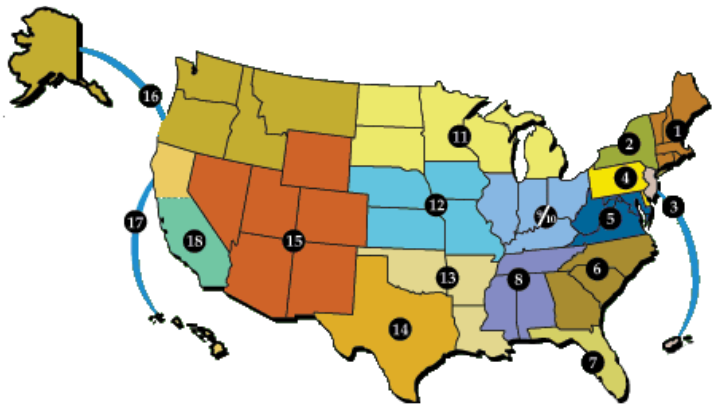
The various modeling strategies in SAS gave similar answers about the magnitude and significance of facility-level predictors. Linear mixed effects models allow for random effects at the network level, but the model assumes normality of the outcome and residual errors (which are violated), and interpretation of log-transformed STR is not intuitive. SAS PROC GENMOD with a negative binomial distribution using transplant counts as the outcome and person-years as the offset does not allow for random effects but has the advantage of expected counts not previously being modeled. Results modeling the count outcome and expected counts as the offset were similar to those modeling observed counts and person-years only, and exponentiated beta estimates are easily interpretable as change in STR associated with unit change in predictor.

## INTRODUCTION

In epidemiology, the Standardized Mortality Ratio (SMR) is a useful measure that compares the level of mortality in one population to the mortality in another population[1]. The SMR is typically defined as the ratio of the number of deaths in a specific population to the expected number of deaths in this same population. An SMR of 1.0 indicates that the number of observed deaths is equivalent to the number of expected deaths. SMRs are calculated using indirect standardization methods, where the expected number of deaths is calculated by multiplying the total number of subjects in a population (such as the U.S. population) by that populations' mortality rate. SMRs are particularly useful when adjusted rates -- such as age-specific rates -- are unstable [2]. In addition, SMRs are often preferred to standardized rate ratios in public health because of ease of interpretation of estimates.

There are numerous SAS® modeling approaches that can be used to model an outcome of a standardized ratio measure (observed/expected counts), such as the frequently encountered SMR. This work is motivated by an analysis that aims to examine associations between dialysis facility characteristics and access to kidney transplantation within 18 regions across the United States. The goal of the analysis was to examine facility-level predictors associated with the Standardized Transplant Ratio (STR) - defined as the total number of observed first kidney transplants divided by the total number of expected first transplants within a dialysis facility. For our analyses, the expected number of transplants within a facility was provided in the data and was defined by a Cox model that adjusted for age and calendar year.

**Figure 1**. The 18 ESRD Network Regions for which the care of ESRD patients within dialysis facilities is overseen.



The purpose of this paper is to compare mixed- and fixed-effects modeling approaches in an analysis of dialysis facilities nested within 18 geographical regions, or End Stage Renal Disease (ESRD) Networks of the United States (**Figure 1**). The examples utilize Dialysis Facility Report (DFR) data, which are publicly available and reported annually by the University of Michigan Kidney Epidemiology and Cost Center under a contract with the Centers for Medicare and Medicaid Services (CMS).

## METHODS AND DATA SOURCES

Dialysis Facility Report (DFR) data are publicly available and reported annually by the University of Michigan Kidney Epidemiology and Cost Center under a contract with the Centers for Medicare & Medicaid Services (CMS).

In a cross-sectional, multi-level ecologic study using the publicly available Dialysis Facility Report (2007-2010) data, we examined 4,098 U.S. dialysis facilities. We considered the outcome (STR) both as linear (STR and log-transformed STR) and as a count (observed counts as outcome with expected counts or person-years as offsets). In the DFR dataset, 'strz_f' is the facility-level STR and there are a number of facility-level variables that represent aggregate demographic and clinical covariate information of patients within the facility. DFR reports facility information yearly. For simplicity, we will use year 1 (2007) data for this analysis. Example aggregate patient variables include age (agemy1_f), black race (blackmy1_f), and diabetes (diabmy1_f), and facility-level factors such as staffing (staffy1_f) and profit/non-profit status (owner_f). In addition, we may want to also consider covariates that are measured at the ESRD Network region level, such as the total number of transplant centers within a region (txctr_n). If our goal is to examine facility-level factors associated with STR, we have to first decide how to consider modeling this outcome.

How might we want to model the outcome of STR? We consider that dialysis facilities that are located within ESRD Network regions of the country may be correlated with one another, since ESRD Network regions are responsible for overseeing the quality of care among dialysis facilities within their respective regions. Thus, we utilized random effects and generalized estimating equation modeling to account for correlation of facilities within regions. We considered SAS PROC MIXED to examine fixed and random effects and PROC GLIMMIX to further examine random effects with the linear outcomes STR and log-transformed STR. We used SAS PROC GENMOD (fixed effects) and PROC GLIMMIX (mixed effects) to examine count outcomes, using a log link and the negative binomial distribution to account for overdispersion.

Table 1 summarizes several modeling strategies for STR.

| | Type | SAS Procedure | Outcome | Interpretation of β |
|---|---|---|---|---|
| **Model 1** | Linear mixed effects | PROC MIXED | STR (obtrz_f/extxz_f) | Change in STR |
| **Model 2** | Linear mixed effects | PROC MIXED | log(STR+1) | Change in log(STR+1) |
| **Model 3** | Linear mixed effects | PROC GLIMMIX | STR | Change in STR |
| **Model 4** | Linear mixed effects | PROC GLIMMIX | log(STR+1) | Change in log(STR+1) |
| **Model 5** | Poisson or Negative Binomial fixed effects | PROC GENMOD | Transplant count (offset: expected count) | log(change in STR) |
| **Model 6** | Poisson or Negative Binomial fixed effects | PROC GENMOD | Transplant count (offset: person time) | log(IRR) |
| **Model 7** | Poisson or Negative Binomial mixed effects | PROC GLIMMIX | Transplant count (offset: expected count) | log(change in STR) |
| **Model 8** | Poisson or Negative Binomial mixed effects | PROC GLIMMIX | Transplant count (offset: person time) | log(IRR) |

**Table 1. Alternate Fixed and Mixed Modeling Strategies for STR outcome**

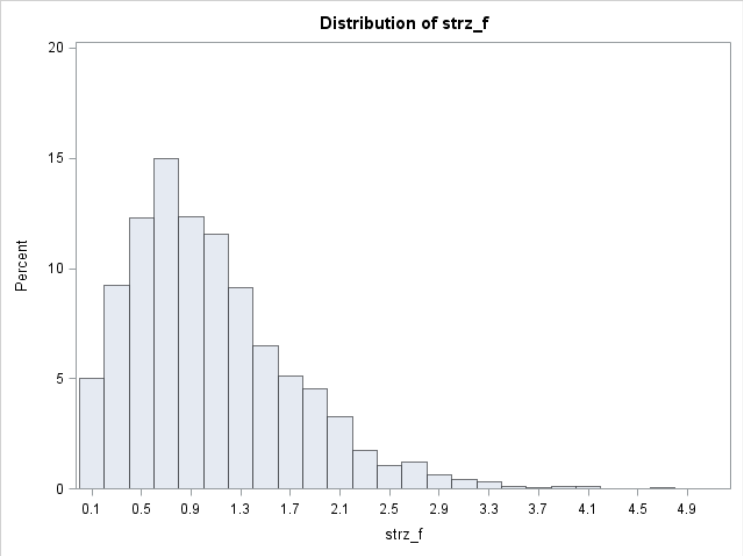
IRR, incidence rate ratio.

## RESULTS

We first consider the most simple model, which considers STR (strz_f from the DFR dataset) as a continuous outcome. We first examine the distribution of the data to determine whether a linear regression model is a good fit.

In addition to examining plots of the data by the various predictors to examine trends, we will also test linearity modeling assumptions including whether observations are: 1) independent and identically distributed observations, 2) normality of the error distribution, and 3) homoscedasticity, or constant variance of errors.

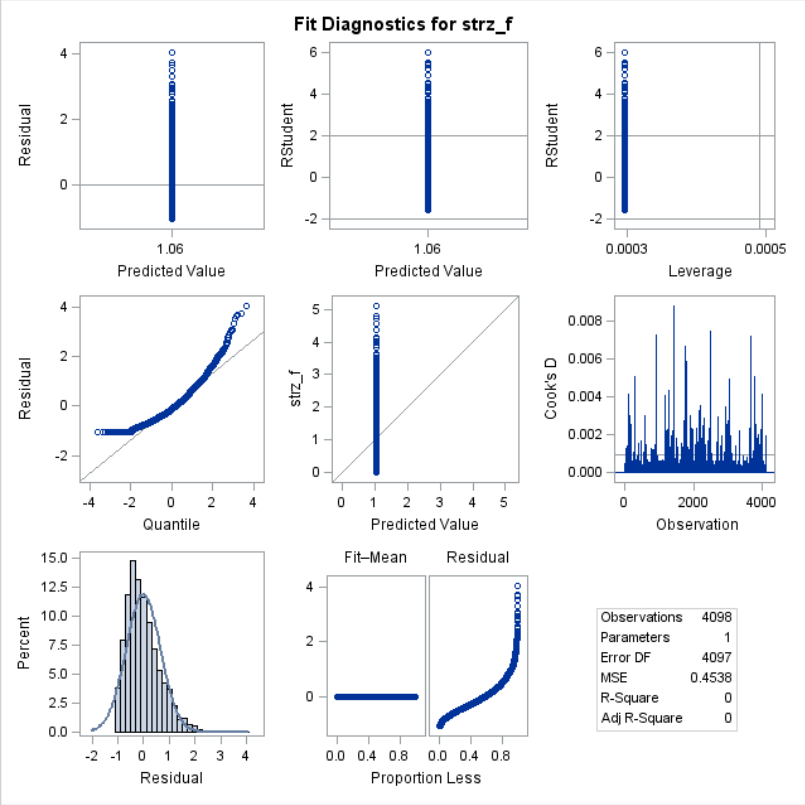
```
proc univariate data= dfr1 normal plot;
       var strz_f;
       qqplot ;
       histogram;
run;
```




**Output 1** shows the distribution of the outcome variable STR ('strz_f')

```
proc reg data= dfr1;
       model strz_f = / dw spec;
       output out=resids r=res ;
run;
```

From **Output 1** and **Output 2**, we can see that the distribution of strz_f is left-skewed because there are a number of facilities that have an STR of zero. The Q-Q plot suggests that the error terms of the STR are not normally distributed.




**Output 2** shows the fit diagnostics for the outcome variable 'strz_f'.

There are a number of potential approaches to model the STR outcome, which are summarized in Table 1. These include considering STR as a continuous outcome or as a count outcome with a Poisson or Negative Binomial Distribution. In SAS, linear mixed models estimation methods can be used, such as PROC MIXED, PROC GENMOD, or PROC GLIMMIX.

In order to account for the potential correlation of facilities within ESRD Networks, we could use the mixed linear model procedure PROC MIXED, which uses maximum likelihood estimation (MLE) methods to estimate the model coefficients and variances (**Model 1**).

```
proc mixed data=dfr1 covtest;
      class network;
      model strz_f = / solution;
      random intercept / sub=network;
run;
```

We may want to allow the intercept to vary by considering a random intercept. The SUB=option specifies the cluster level at the Network region. We consider in the model statement covariates that are on the facility level (level 1) and the ESRD Network level (level 2; *e.g.*, txctr_n), using the random statement with the SUB=network_n to account for potential clustering of facilities within ESRD Network regions. We consider a random intercept for ESRD Network region so that we can make inference on facilities within regions.

```
proc mixed data=dfr1 covtest;
      class network;
      model strz_f = blackmy1_f diabmy1_f staffy1_f owner_f
            txctr_n / solution cl;
      random intercept / sub=network_n;
run;
```

This model converges. However, Figure 2 suggests the normality assumptions and some linearity assumptions are not met. One strategy to address the skewed distribution is to transform the STR, such as with a log transformation (**Model 2**). Because zero is a meaningful value of the outcome STR, we considered the outcome as ln(STR+1). We run the same model as above, but with the transformed outcome.

```
proc mixed data=work.dfr1 covtest;
      class network;
      model logstr = blackmy1_f diabmy1_f staffy1_f owner_f
            txctr_n / solution cl;
      random intercept / sub=network_n;
run;
```

The model statement specifies the fixed effects and the random statement specifies the random effects. We specify that the subject, or cluster, is the ESRD Network region (network_n). A repeated statement could also be used here to specify the variance-covariance structure of the errors. This model gives estimates that are somewhat difficult to interpret (change in the ln(STR+1)).

We may also want to consider modeling continuous STR or ln(STR + 1) using PROC GLIMMIX, since PROC GLIMMIX allows a non-normal response distribution of the outcome whereas PROC MIXED requires normality of the response variable. PROC GLIMMIX has very similar syntax to PROC MIXED. **Model 3** uses a non-transformed STR.

```
proc glimmix data=work.dfr1 covtest;
      model logstr = blackmy1_f diabmy1_f staffy1_f owner_f
            txctr_n / solution;
      random intercept / sub=network;
run;
```

In **Model 4**, we consider using the log-transformed STR to improve on model fit.

```
proc glimmix data=work.dfr1;
      model logstr = blackmy1_f diabmy1_f staffy1_f owner_f
            txctr_n /  solution;
            random intercept / subject=network_n;
run;
```

**Models 3** and **4** use generalized linear modeling through PROC GLIMMIX. This approach allows for both fixed and random effects modeling, while also accounting for the potential correlation of facilities within ESRD Network regions. Because our outcome of STR has already been modeled, this implies that there is unknown error in our outcome that we may not have accounted for in this modeling approach. There are several additional modeling strategies that we can use that may provide an easier interpretation or one with more public health significance. Instead of considering the STR as the modeled outcome, we could consider the components of the STR, such as the transplant count as the outcome and the offset the expected transplant count. Poisson regression may be appropriate for these rate data. Here the rate could be defined several ways, depending on the interpretation of the model coefficients that we want. The numerator of the rate is the observed transplant count (obtrz_f), and the denominator could either be the expected first transplant count (extxz_f), such that the exponentiated coefficient represents change in STR; or the denominator could be person-years at risk for first transplant (txyz_f), such that the exponentiated coefficient represents change in the incidence rate ratio (IRR) of transplant.

```
data count;
      set dfr1;
      log_cnt = log(extxz_f); *Expected first transplant count;
      log_pt = log(txyz_f); *Person years;
run;
```

In SAS, one approach to model a Poisson distribution is to use a generalized linear modeling approach, such as PROC GENMOD, which allows for fixed effects, but not random effects. The link function is necessary to ensure that the model is linear. Note that the default link function for the Poisson distribution is the log function.

```
proc genmod data= count;
      model obtrz_f = / dist=poisson offset=log_cnt
run;
```

There is some evidence for overdispersion of the STR residuals, meaning that the observed variance is larger than the assumed variance of the STR. In these data, the overdispersion is due to 2.4% of facilities with an STR of zero. One way to examine whether the data are over- or under-dispersed is to look at the scale factor; *e.g.*, Pearson/DF, in the model diagnostics, where a value of 1.0 suggests no dispersion, <1 suggests underdispersion, and > 1 suggests overdispersion. Here, the deviance value of 2.9966 suggests that the data are overdispersed (**Output 3**).

**Output 3. Model Goodness of Fit Output**

| Criteria For Assessing Goodness Of Fit | | | |
|---|---|---|---|
| Criterion | DF | Value | Value/DF |
| Deviance | 4092 | 12262.2701 | 2.9966 |
| Scaled Deviance | 4092 | 12262.2701 | 2.9966 |
| Pearson Chi-Square | 4092 | 16060.2481 | 3.9248 |
| Scaled Pearson X2 | 4092 | 16060.2481 | 3.9248 |

We could try to adjust for overdispersion in the above model by adding the scale parameter. We would do this by adding the option SCALE=PEARSON, and examining fit statistics (scaled deviance and scaled pearson chi-square statistics) in the SAS output. For these data, when we adjust for the overdispersion, the scaled deviance is now 0.7635, which is an improvement.

If we find the overdispersion is still a problem, another alternative approach is to consider a negative binomial distribution model (**Model 5**). Here the new deviance (Value/DF) is 1.0593, which is a much better fit to the data.

```
proc genmod data= count;
      model obtrz_f = / dist=nb link=log offset=log_cnt
run;
```

We consider model covariates using the observed transplant count as the outcome, and the log of the expected transplant count as an offset in the model. This will allow us to interpret the beta coefficients as the log(change in STR).

```
proc genmod data=work.count;
      class network_n (ref="1");
      model obtrz_f = blackmy1_f diabmy1_f staffy1_f owner_f
            txctr_n /dist=poisson link=log offset=log_cnt diagnostics
                     obstats;
      repeated subject=network_n/ ;
      output out=data p=pred;
run;
```

We can further assess model fit by outputting the data and plotting the residuals, or by examining diagnostics. Note that the repeated statement specifies the covariance structure of the clustered responses for GEE model fitting, and the subject=network_n line will give robust standard errors for the model coefficients.

Additionally, we could also model the transplant count using the offset of person-time to change the interpretation of the beta coefficients to the log(IRR) (**Model 6**).

```
proc genmod data=work.count;
      class network_n (ref="1");
      model obtrz_f = blackmy1_f diabmy1_f staffy1_f owner_f
            txctr_n/dist=poisson link=log offset=log_pt diagnostics
                     obstats;
      repeated subject=network_n/ ;
      output out=data p=pred;
run;
```

Additionally, if we would like the flexibility of incorporating both fixed and random effects into our model, an approach such as PROC GLIMMIX, is ideal. Random effects are important to incorporate if the levels represent a sample of a population (*e.g.*, if we were examining a subsample of dialysis facilities rather than all U.S. dialysis facilities), and if we wanted to make inference on all dialysis facilities.

The GLIMMIX and MIXED procedures are similar. Note the default link function for both Poisson and Negative binomial models is the log. As before, the offset specifies the exposure time, which in these data could potentially be the expected value of the transplant count or the person-time associated with the outcome, depending on the preferred interpretation of the model coefficients and tolerance for unknown error in expected counts. **Model 7** considers a Poisson distribution with the observed transplant count as the outcome and the log of the expected first transplant count as the offset.

```
proc glimmix data= count;
      model obtrz_f = blackmy1_f diabmy1_f staffy1_f owner_f
            txctr_n / dist=poisson link=log offset=log_cnt solution;
            random intercept / subject=network_n;
run;
```

Here, the model statement specifies the dependent variable and the fixed effects of the model, and the random statement defines the random effects. We may also want to consider using a RANDOM _RESIDUAL_ statement in PROC GLIMMMIX if we would also like to specify R-side random effects.

We could also consider modeling the offset as the log of the person-time at risk for a first transplant (log_pt), as above (**Model 8**)

```
proc glimmix data= count;
      model obtrz_f = blackmy1_f diabmy1_f staffy1_f owner_f
            txctr_n/ dist=nb link=log offset=log_pt solution;
            random intercept / subject=network_n;
run;
```

We still may want to consider a modeling strategy that examines count variables, but with these data we do not have the option of using GEE since our models did not converge. Table 4 reports the associations of several facility-level variables with the facility-level STR using the various modeling approaches. Of note, most of the model coefficients and p-values are similar across models, with the exception of profit status (a dichotomous variable). Additionally, the linear modeled STR (**Model 1**), which we reported a poor fit of the data to the linear STR outcome, and log-transformed STR (**Model 2**) have different p-values that could result in different conclusions if we consider p<0.05 as statistically significant. In **Models 1-2**, we would conclude that the number of transplant centers within an ESRD Network was not a significant predictor of facility-level STR, but **Models 3-8** would support a positive association between the number of transplant centers and a higher STR.

Table 4 summarizes beta coefficients for the various modeling approaches (**Models 1-8**). Note these models only include a few select covariates in the models for simplicity.

**Table 4. Summary of model coefficients for Models 1-8**

| | **Model 1** | | **Model 2** | | **Model 3** | | **Model 4** | |
|---|---|---|---|---|---|---|---|---|
| | β | p-value | β | p-value | β | p-value | β | p-value |
| Blackmy1_f | -0.008 | <0.01 | -0.004 | <0.01 | -0.008 | <0.01 | -0.004 | <0.01 |
| Diabmy1_f | -0.006 | <0.01 | -0.003 | <0.01 | -0.006 | <0.01 | -0.003 | <0.01 |
| Staffy1_f | 0.002 | 0.04 | 0.002 | <0.01 | 0.002 | 0.03 | 0.002 | <0.01 |
| Owner_f | -0.071 | 0.01 | -0.026 | 0.04 | -0.057 | 0.03 | -0.020 | 0.10 |
| Txctr_n | 0.051 | 0.09 | 0.022 | 0.10 | 0.066 | <0.01 | 0.027 | <0.01 |

| | **Model 5** | | **Model 6** | | **Model 7** | | **Model 8** | |
|---|---|---|---|---|---|---|---|---|
| | β | p-value | β | p-value | β | p-value | β | p-value |
| Blackmy1_f | -0.01 | <0.01 | -0.012 | <0.01 | -0.013 | <0.01 | -0.013 | <0.01 |
| Diabmy1_f | -0.01 | <0.01 | -0.011 | <0.01 | -0.012 | <0.01 | -0.012 | <0.01 |
| Staffy1_f | -0.04 | <0.01 | -0.040 | <0.01 | -0.039 | <0.01 | -0.039 | <0.01 |
| Owner_f | -0.36 | <0.01 | -0.400 | <0.01 | -0.310 | <0.01 | -0.310 | <0.01 |
| Txctr_n | 0.07 | <0.01 | 0.070 | <0.01 | 0.085 | <0.01 | 0.085 | <0.01 |

## CONCLUSION

There are various modeling strategies to use when modeling standardized ratios, such as the standardized mortality ratio or the standardized transplant ratio. The various modeling approaches in SAS gave similar answers about the magnitude and significance of facility-level predictors in our example.

Table 5 summarizes the strengths and limitations of the modeling strategies for STR.

| Model | Advantages | Disadvantages |
|---|---|---|
| **Model 1** | Can be used for models with random effects and for data with correlated errors | Outcome is previously modeled; assumptions of normality of outcome and residual errors violated |
| **Model 2** | Can be used for models with random effects and for data with correlated errors | Outcome is previously modeled; beta coefficient estimates not easily interpretable |
| **Model 3** | Can be used for models with random effects and for data with correlated errors, no need for normality assumption of outcome | Outcome is previously modeled; assumptions of normality of residual errors violated |
| **Model 4** | Can be used for models with random effects and for data with correlated errors, no need for normality assumption of outcome | Outcome is previously modeled; beta coefficient estimates not easily interpretable |
| **Model 5** | Can be used for data with correlated errors | No random effects in GEE models |
| **Model 6** | Can be used for data with correlated errors | No random effects in GEE models |
| **Model 7** | Can be used for models with random effects and for data with correlated errors | Offset is previously modeled |
| **Model 8** | Can be used for models with random effects and for data with correlated errors | Loss of standardized ratio interpretability |

Linear mixed effects models allow for random effects at the network level, but the model assumes normality of the outcome and residual errors (which are violated), and interpretation of log-transformed STR is not intuitive. SAS PROC GENMOD with a negative binomial distribution using transplant counts as the outcome and person-years as the offset does not allow for random effects, but has the advantage of expected counts not previously being modeled. Modeling results using the count outcome and expected counts as the offset were similar to those using observed counts and person-years only and exponentiated beta estimates are easily interpretable as change in STR associated with unit change in predictor. Various modeling options should be considered and compared when the outcome of interest a standardized ratios.

## REFERENCES

Liddell FD. The development of cohort studies in epidemiology: a review. *J Clin Epidemiol.* 1988;41(12):1217-1237.

Goldman DA, Brender JD. Are standardized mortality ratios valid for public health data analysis? *Statistics in medicine.* Apr 30 2000;19(8):1081-1088.

## ACKNOWLEDGMENTS

We would like to thank the University of Michigan Kidney Epidemiology and Cost Center for their assistance with data acquisition. R.E.P. was supported in part by grants from the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number ULl TR000454 and KL2TR000455. R.E.P. and L.P. are both supported in part by R24MD008077 through the National Institute on Minority Health and Health Disparities. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

## RECOMMENDED READING

- *Statistical Analysis with the GLIMMIX Procedure*
- *Mixed Models Analyses using SAS®*
- *Generalized Linear and Nonlinear Models for Correlated Data: Theory and Applications using SAS®*

## CONTACT INFORMATION

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


Name: Rachel Patzer, PhD, MPH
Enterprise: Emory University School of Medicine, Department of Surgery
Address: 101 Woodruff Circle, 5006 Woodruff Memorial Research Building
City, State ZIP: Atlanta, GA
Work Phone: 404-727-6047
Fax: 404-727-3660
E-mail: rpatzer@emory.edu