

# Sample Size Estimation Through Simulation of a Random Coefficient Model by Using SAS

Junliang Chen<sup>1</sup>, Shannon Stock<sup>2</sup>, Chunqin Deng<sup>1</sup>

<sup>1</sup> Talecris Biotherapeutics Inc., Research Triangle Park, NC 27709

<sup>2</sup> Department of Biostatistics, Harvard University, 02115

---

## Abstract:

In chronic pulmonary diseases, the development of emphysema progresses over many years. As a result, the assessment of drug efficacy requires the observation of large numbers of patients, followed for long periods of time. Recently, lung densitometry has been studied as a potential clinical endpoint for the assessment of lung tissue loss over time in patients with emphysema. Clinical trials using lung densitometry as an endpoint are typically designed as longitudinal studies with repeated measurements at fixed time points. Since lung density measurements are generally correlated with lung volume measurements, lung volume should be included in statistical analyses as a longitudinal covariate. The clinical efficacy of treatments can be assessed by comparing the progressions of decreased lung densities through the use of a random coefficient model – a longitudinal linear mixed model with a random intercept and slope. However, the implementation of such complex statistical analyses in clinical trials makes sample size calculations difficult. In this article, an empirical approach to sample size calculations is proposed using simulated trajectories of lung densities and lung volumes. We present step-by-step details for sample size calculations using simulations, and discuss the pros and cons of this approach. SAS Macros are provided.

*Keywords:* Sample size; Lung densitometry; Longitudinal studies; CT scan; Random coefficient model

---

## Introduction:

In planning and development stage of the clinical trials, a sample size calculation is a critical step. In controlled clinical trials, sample size calculation is required to maintain specific statistical power (i.e. 80% power) for the study. Not surprisingly, there are many software packages (i.e. nQuery, PASS) which perform sample size calculation for certain statistical tests. However, such software package is not available for sample size calculation for clinical trials with complex statistical models.

One example is in chronic pulmonary diseases (such as Chronic Obstructive Pulmonary Disease – COPD) regards the development of emphysema. It is a slow progression over many years and the assessment of drug efficacy requires the observation of large numbers

of patients for a long period of time. Recently, lung densitometry (measuring the lung density through CT scan) has been studied as a potential clinical endpoint for assessing the lung tissue loss over time in patients with emphysema. The clinical trial with lung densitometry as an endpoint is typically designed as a longitudinal study with repeated measurements at fixed time intervals. Since lung density measurements are closely correlated with lung volume (inspiration level), it is important to include lung volume measurements in statistical analyses as a longitudinal covariate. Lung volume is normally measured at the same time as the lung density is measured. The clinical efficacy can be assessed by comparing the progression of lung density loss between two treatment groups using a random coefficient model – a longitudinal linear mixed model with a random intercept and slope. In planning the clinical trial with such complex statistical analyses, the calculation of the sample size required to achieve a given power to detect a specified treatment difference is an important, often complex issue.

In this article, an empirical approach is proposed to calculate the sample size by simulating trajectories of lung density and lung volume using SAS. We present step-by-step details for sample size calculation through simulation, and discuss the pros and cons of this approach. SAS Macros are provided at the end.

## Methods:

### Statistical Model Notation and Assumptions

The random coefficient model assumes the following form:

$$Y_{ij} = (\beta_0 + b_0) + \beta_1 * TRT + (\beta_2 + b_2) * TIME + \beta_3 * COV_{ij} + \beta_4 * TRT * TIME + \varepsilon_{ij} \quad (1)$$

Where  $Y_{ij}$  is the efficacy endpoint (i.e. lung density) measurement for subject  $i = 1, 2, \dots, n$ , at fixed time point  $j = 1, 2, \dots, K$ . TRT is an indicator of subject  $i$ 's treatment group (i.e. TRT=1 for active drug; TRT=0 for placebo).  $COV_{ij}$  is a longitudinal covariate (i.e. logarithm of lung volume) for subject  $i = 1, 2, \dots, n$ , at fixed time point  $j = 1, 2, \dots, K$ . Here  $b_0$  and  $b_2$  are subject-specific random effects for the intercept and slope, respectively, which are from a normal distribution with mean 0 and variance  $\sigma_0^2$  and  $\sigma_2^2$ , respectively.  $\varepsilon_{ij}$  is the random error from a normal distribution with mean 0 and variance  $\sigma^2$ . The regression parameters  $\beta_0, \beta_1, \beta_2, \beta_3$ , and  $\beta_4$  are the fixed effects for intercept, treatment, time, covariate and interaction of treatment and time respectively.

Here we assume that the clinical benefits of a given treatment can be assessed quantitatively by comparing the slopes of lung density trajectories for the two treatment groups. This quantity is captured by  $\beta_4$ .

### Sample Size Estimation Using Simulations

In model (1),  $\beta_4$  is typically our interest, which is the difference in slope of time between two treatment groups (active vs. placebo). There is no direct mathematical formula to calculate the sample size for a given statistical power (i.e. 80%) to test the null hypothesis:  $\beta_4=0$  with a specified type I error (i.e.  $\alpha=0.05$ ) for such model in (1). One approach to calculate the sample size for a given power is through the simulation.

Assume we know the parameters ( $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$ , and  $\sigma_0^2, \sigma_2^2, \sigma^2$ ) from either history data, previous clinical trials or meaningful clinical differences we want to test, the study design in terms of number of time points ( $K$ ) and fixed time intervals (TIME), and the longitudinal covariate  $COV_{ij}$ . For a fixed equal sample size  $n$  for each treatment, the trajectories of efficacy measurement  $Y_{ij}$  (i.e. lung density) for the  $n$  subjects can be simulated through model (1) for each treatment group. Then, perform a statistical test on  $\beta_4=0$  by using the SAS Proc MIXED on the simulated data set, and record whether the p-value  $< 0.05$ . The sample code to perform the test is as follow:

```
proc mixed data = data;
  class id trt;
  model y = trt time trt*time cov / solution;
  random intercept time/ subject = id type = un;
run;
```

For the fixed sample size  $n$  per treatment group, simulate  $M$  (i.e.  $M=1000$ ) times and the proportion of significance test on  $\beta_4=0$  among the total  $M$  simulations is the statistical power for the sample size  $n$  per treatment group. Then, adjust the sample size  $n$  to achieve desirable statistical power.

In order to simulate the trajectories of  $Y_{ij}$ , it is necessary to simulate the trajectories of longitudinal covariate  $COV_{ij}$ . Similarly, assume  $COV_{ij}$  is from a linear model regressing against time with a random intercept:

$$COV_{ij} = (\gamma_0 + r_0) + \gamma_1 TIME + \varepsilon_{ij} \quad (2)$$

Where  $\gamma_0$  and  $\gamma_1$  are the fixed intercept and slope respectively;  $r_0$  and  $\varepsilon_{ij}$  are from a normal distribution with mean 0 and variance  $\delta_0^2$  and  $\delta_1^2$ , respectively. If we know the parameters ( $\gamma_0, \gamma_1, \delta_0^2, \delta_1^2$ ) from history data or previous clinical trials for the study population, it will be simple to simulate the trajectories of the longitudinal covariate  $COV_{ij}$  by using SAS random generating functions.

In detail, a sample size can be determined for the models above through the following steps:

1. Obtain the pre-specified parameters through either history data, previous clinical trials or meaningful clinical difference to be tested from clinicians
2. Specify a desired statistical power (i.e. 80%) and a type-1 error rate (i.e. 5%)
3. Simulate trajectories of efficacy measurement (i.e. lung density) and longitudinal covariate (i.e. logarithm of lung volume) for a fixed sample size ( $n$ ) of subjects within each treatment arm
  - a. Trajectories of longitudinal covariate (i.e. logarithm of lung volume) are simulated through model (2)
  - b. Trajectories of efficacy measurement (i.e. lung density) are simulated through model (1)

4. Perform the statistical test on  $\beta_4=0$  based on the simulated data set. Record whether a p-value  $< 0.05$  was obtained
5. Repeat steps 3 and 4 M (i.e. M=1000) times and calculate the statistical power for the fixed sample size
6. Repeat steps 3 - 5 for various values of  $n$ . Stop when desired statistical power is obtained

## Results:

### Example of a Simulation

Assume there are two treatment groups (active vs. placebo) in a study design. The efficacy endpoint along with the longitudinal covariate will be measured at  $K=4$  time points at baseline, 1 year, 2 years and 3 years. All corresponding parameters specified in model (1) and (2) could be obtained either through history data, previous clinical trials or meaningful clinical difference to be tested from clinicians. For purpose of simulation, they are randomly selected and specified as below:

$$\beta_0 = 150, \beta_1 = 5, \beta_2 = -1.8, \beta_3 = -57, \beta_4 = 0.7, \text{ and } \sigma_0^2 = 280, \sigma_2^2 = 0.4, \sigma^2 = 5;$$

$$\gamma_0 = 2, \gamma_1 = 0.0007, \delta_0^2 = 0.05, \delta_1^2 = 0.0016.$$

The summary of statistical power for a given sample size per treatment based on M = 1000 simulated data sets is listed below:

N per treatment	Statistical Power (%)
30	62.4
40	76.9
<b>45</b>	<b>79.9</b>
50	84.4
60	91.3

Therefore, a sample size 45 per treatment arm has an estimated statistical 80% power to detect the treatment slope difference of 0.7 in a random coefficient model for the study design above.

## Conclusions and Discussion:

As described above, it is possible to perform sample size calculations for a random coefficient model using simulation techniques and SAS. It is also straightforward to extend the simulation frame to other linear mixed models (LMM) or generalized linear mixed models (GLMM). Other extensions to settings involving multiple treatment groups (i.e. treatment groups greater than 2), unequal sample size among treatment groups (i.e. 2:1 for active vs. placebo) can be implemented. For an active-controlled trial, it is usually interest to test non-inferiority of test drug compared to active-control. The simulation frame can be applied for such non-inferiority test by calculating the confidence interval for the parameter tested in the statistical model and comparing the

lower limit or upper limit of the confidence interval to the pre-specified equivalence margin in the Step 4.

Other study design parameters such as number of repeated measurements ( $K$ ) of efficacy endpoint and the duration of the fixed time intervals (*time*) also affect the sample size estimation. Greater number of repeated measurements of efficacy endpoint for the fixed study duration will increase the statistical power. However, it might increase the difficulty and cost of the study depending on the efficacy endpoint. The number of repeated measurement of efficacy endpoint and duration of the fixed time intervals should be determined within the clinical research team upon the constraints such as the difficulty of efficacy endpoint measurement, cost and duration of the clinical trial.

In practice, it is rarely the case that all subjects have the complete data for all visits in the study because of missing certain study visits, drop out or other reasons. Since our simulation framework assumes there are no missing observations, we recommend that the implemented sample size for the designed trial include more subjects than the number estimated from the simulation. In most cases an increase of 5% or 10% should suffice, but depending on the characteristics of the designed trial such as the study population, difficulty of study procedure, difficulty of study measurement etc to cause the subject's drop out or missing of study measurements. The appropriate percentage could vary.

**Contact Information:**

Junliang Chen, Ph.D  
Talecris Biotherapeutics, Inc  
4101 Research Commons  
79 T.W. Alexander Drive  
RTP, NC 27709

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

## Appendix:

```
/******  
This code can be used to determine the sample size needed to achieve  
80% power to detect the treatment slope difference of 0.7 after  
adjusting for the covariate.  
*****/  
%macro SampleSize(n);  
  %do j = &n %to &n;  
    /*Placebo Group*/  
    Data one;  
      trt=0;  
      beta0=150; beta1=5; beta2=-1.8; beta3=-57; beta4=0.7; sigma0=16.7;  
      sigma2=0.63; sigma=2.24;  
      gamma0=2; gamma1=0.0007; delta0=0.224; delta1=0.04;  
      do id = 1 to &n;  
        b0 = sigma0*rannor(64327);  
        b2 = sigma2*rannor(24681);  
        r0 = delta0*rannor(13465);  
        do time = 0, 1, 2, 3;  
          cov = (gamma0 + r0) + gamma1*time + delta1*rannor(3458);  
          y = (beta0 + b0) + beta1*trt + (beta2 + b2)*time + beta3*cov +  
              beta4*trt*time + sigma*rannor(4762);  
          output;  
        end;  
      end;  
    run;  
  
    /*Active*/  
    Data two;  
      trt=1;  
      beta0=150; beta1=5; beta2=-1.8; beta3=-57; beta4=0.7; sigma0=16.7;  
      sigma2=0.63; sigma=2.24;  
      gamma0=2; gamma1=0.0007; delta0=0.224; delta1=0.04;  
      do id = &n + 1 to 2*&n;  
        b0 = sigma0*rannor(6427);  
        b2 = sigma2*rannor(4681);  
        r0 = delta0*rannor(1365);  
        do time = 0, 1, 2, 3;  
          cov = (gamma0 + r0) + gamma1*time + delta1*rannor(458);  
          y = (beta0 + b0) + beta1*trt + (beta2 + b2)*time + beta3*cov +  
              beta4*trt*time + sigma*rannor(762);  
          output;  
        end;  
      end;  
    run;  
  
    /*concatinate*/  
    Data comb;  
      set one two;  
    run;  
  
    *Mixed model;  
    proc mixed data = comb method = reml;
```

```

class id trt;
model y = trt time trt*time cov/ solution;
random intercept time/ type = un subject = id ;
ods output Tests3 = unTemp;
run;

data un;
  set unTemp;
  iteration = 1;
run;

*Iteration macro;
%macro loop(iterations);
  %do i = 2 %to &iterations;
/*Placebo Group*/
Data one;
  trt=0;
  beta0=150; beta1=5; beta2=-1.8; beta3=-57; beta4=0.7; sigma0=16.7;
  sigma2=0.63; sigma=2.24;
  gamma0=2; gammal=0.0007; delta0=0.224; deltal=0.04;
  do id = 1 to &n;
    b0 = sigma0*rannor(64327+526*&i);
    b2 = sigma2*rannor(24681-3*&i);
    r0 = delta0*rannor(13465+55*&i);
    do time = 0, 1, 2, 3;
      cov = (gamma0 + r0) + gammal*time + deltal*rannor(3458+23*&i);
      y = (beta0 + b0) + beta1*trt + (beta2 + b2)*time + beta3*cov +
        beta4*trt*time + sigma*rannor(4762+8*&i);
      output;
    end;
  end;
run;

/*Active*/
Data two;
  trt=1;
  beta0=150; beta1=5; beta2=-1.8; beta3=-57; beta4=0.7; sigma0=16.7;
  sigma2=0.63; sigma=2.24;
  gamma0=2; gammal=0.0007; delta0=0.224; deltal=0.04;
  do id = &n + 1 to 2*&n;
    b0 = sigma0*rannor(6427+12*&i);
    b2 = sigma2*rannor(2481+5*&i);
    r0 = delta0*rannor(1345+45*&i);
    do time = 0, 1, 2, 3;
      cov = (gamma0 + r0) + gammal*time + deltal*rannor(358+33*&i);
      y = (beta0 + b0) + beta1*trt + (beta2 + b2)*time + beta3*cov +
        beta4*trt*time + sigma*rannor(462+62*&i);
      output;
    end;
  end;
run;

/*concatinate*/
Data comb;
  set one two;
run;

```

```

proc mixed data = comb method = reml;
  class id trt;
  model y = trt time trt*time cov/ solution;
  random intercept time/ type = un subject = id ;
  ods output Tests3 = unTemp;
run;

*un;
data unTemp;
  set unTemp;
  iteration = &i;
run;
data un;
  set un unTemp;
run;
%end;
%mend;

%loop(1000);

*Calculate power;
Data power;
  set un;
  if effect = 'time*trt';
run;
Data power;
  set power;
  if probf <=0.05 then sig = 1;
  if probf>0.05 then sig = 0;
run;

proc freq data = power ;
  table sig;
  ods output OneWayFreqs = significance&n;
run;
%end;
%mend;

%sampleSize(30);
%sampleSize(40);
%sampleSize(45);
%sampleSize(50);
%sampleSize(60);

```