Power and Sample Size Analysis For a Prospective Clinical Trial

Xiang, Yongshi *yongshix@uci.edu*

Liu, Yiming *yiminl10@uci.edu*

Gao, Xu *xgao2@uci.edu*

Prof Miriam Bender *miriamb@hs.uci.edu*

Prof Hernando Ombao *hombao@uci.edu*

University of California, Irvine

April 18, 2017

Abstract

For a clinical trial within one or multiple similar medical care units, linear model is used to capture Gaussian outcomes (IV-Indwelling time) and logistic regression is used to model binary outcome (IV-removal due to symptom). Monte Carlo method was implemented to obtain a range of sample sizes under various assumptions such as power, effect sizes etc. For a clinical trial within multiple different medical care units, generalized linear mixed effects models are used to model both Gaussian and binary outcomes. The study aims to properly capture the effects of the clinically indicated replacement group versus the routine replacement group. The future study is assumed to be balanced, which means both the clinically indicated group and the routine group will have the same number of patients. The data from the pilot study conducted for this future clinical trial is assumed to be the actual representation of the reality.

1 Introduction

In the United States, approximately 200 million of peripheral intravenous (PIV) catheters are used every year(Rickard et al., 2012). Repeated PIV catheters insertion causes pain to the patients(Dao, 2016) and increases health care costs for both the hospitals and the patients(Rickard et al., 2012). The recommended guidelines of the Centers for Disease Control and Prevention (CDC) is that no more frequently than at 96-hour PIV catheters replacement for patients not receiving blood products. There are various studies on minimization of the routine replacement of PIV catheters in order to reduce discomfort and costs.

The School of Nursing at University of California, Irvine (UCI) is preparing to conduct a clinical trial to examine the device days (indwelling time) and indications for PIV catheter removal in 96-hour routine replacement versus clinically indicated replacement in the adult patient population(Dao, 2016). A retrospective pilot study was conducted at the UCI Medical Center over a 6-month period with 73 routine replacement (96-hour) cohort and 64 clinically indicated cohort. They found that there was no significant difference in the rate of symptoms for the clinically indicated replacement group as compared to the routine replacement group and the mean device days for the clinically indicated replacement group was significantly longer than the routine replacement group(Dao, 2016). The School of Nursing wants to generalize the results to a broader population, so Dr Bender came to us and asked us to conduct a power analysis in order to find the optimal number of patients they need in a future clinical trial, which aims to more properly capture the effects of the clinically indicated replacement group vs the routine replacement group.

The intuition of our analysis is that, since the future clinical trial aims to capture and study the effects of two different policies(routine replacement vs clinically indicated replacement) upon two primary outcomes("Indwelling time" and "removal due to symptoms"), the most important task is to detect those effects effectively. Therefore, we aim to find the optimal sample size that will maximize the probability of detecting those effects when they exist, which can be directly tied to maximizing the power of a statistical test concerning the

existence of effects. Main co-variates in our analysis include age, sex, length of stay (days) and number of IV catheters inserted into a patient.

For binary outcomes, Dang et al. (2008) proposed formulas for power and sample size calculation using generalized linear mixed model based on penalized quasi-likelihood. However, their study is based on longitudinal data with attrition over time and ours is not. Therefore, their theoretical results cannot be applied to our analysis. We believe that Monte Carlo simulation method based on the pilot study can estimate the relationship between power and sample size.

The remainder of this paper is organized into two sections. The first section is for study in one care unit or multiple SIMILAR care units, while the second section is for study in multiple DIFFERENT care units. For a single hospital, we used linear regression to model device days and used logistic regression to model removal due to symptom. For multiple different care units, we used generalized linear mixed effects model with random effects to capture the unit factor because of the large amount and variability across different units. The entirety of our analysis was performed using R.

2 Single Unit

In this section we will focus on single-unit study setting, and the general flow of this section is as follows. Firstly we will specify the overall assumptions, upon which we will ground our analysis. Secondly, we will build up models corresponding to two primary outcomes "indwelling time" and "removal due to symptom" separately. Thirdly, we will then conduct power analysis with respect to specific models and effect size. Before the commencing of formal analysis, we will reiterate the practical goal of this study here. Through this study, we aim to estimate the optimal number of patients needed that will maximize the probability of detecting the effects of two different policies upon "indwelling time" and "removal due to symptom" given the effects exist for a potential future clinical trial.

2.1 Assumptions

In this section, we will make three major assumptions. Firstly, we assume the future clinical trial will be conducted within a single care unit, and here the "single unit" is determined by the similarity, which means even if there are multiple different hospital units in the future clinical trial, as long as the researchers have reasonable evidence to deem them similar, they can be treated as a single unit. Considering the cost and time needed to conduct a clinical trial with the participation from wide range of different hospital units, this assumption might be an adequate representation of reality.

Secondly, we assume the future study is designed to be balanced, as in both clinically indicated group and routine group will have same number of patients.

Thirdly, we assume that the data from the pilot study conducted for this future clinical trial is the actual representation of the reality. This assumption may not be appropriate, as the pilot study is only conducted in a very small scale, but given pilot data is the only source data we can rely upon, for the sake of interpretability of this analysis, we will adhere to this assumption.

2.2 Gaussian responses with only fixed effects

To measure the effects of these two different policies on "indwelling time" of IVs, we need to first establish a relationship between them, and we will use the following model to do so.

2.2.1 Model setup

The "indwelling time" of IVs (Y_i) is modeled as Gaussian distribution with mean μ_i and variance τ^2. We paremeterize the mean μ_i by

$$\mu_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + (\beta_3 + \beta_5 X_{5i}) X_{3i} + (\beta_4 + \beta_6 X_{5i}) X_{4i} + \beta_7 X_{5i},$$

where X_{1i} denotes the age of patient i, X_{2i} is an indicator with value 0 if patient i is female and 1 otherwise, X_{3i} denotes the length of stay (days) for patient i (LOS), X_{4i} denotes the number of IV catheters inserted into patient i (NIV) and X_{5i} takes value of 1 if patient i is from routine group, 0 if patient i is from clinically indicated group.

From above model, we can see the effects of different policies upon "indwelling time" are represented by β_5, β_6 and β_7. Thus we can detect those effects by testing the existence of corresponding coefficients. Now we have established a relationship between our primary outcome and variable of interest, we can take a look at the statistical tests themselves. During this entire analysis we will set the significance level α=0.05, and target power level to be 80%. As for the other crucial element of power analysis, effect size, since we have assumed the pilot data is the actual representation of reality, we will treat the estimated coefficients obtained by fitting above model on the pilot data as their actual values. The estimated values are summarized in Table 1.

Table 1: Assumed actual coefficients

β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
4.134	-0.005	0.314	0.079	-0.132	-0.070	0.174	-0.970

2.2.2 Simulation procedures

Next we will apply Monte Carlo method to estimate the relationship between the power of relevant Wald-tests and their sample size.

I. Specify relevant parameters. Set significance level α=0.05 and desired power level=0.80.

II. For a fixed sample size, generate new data based on a reasonable structure. In our case,we generate new data based on pilot data

Repeat for a number of iterations

III.A Generate new responses by inputting the newly generated data into the model described above, using the assumed actual coefficients as shown in above table

III.B Fit above linear model with newly generated data and responses to obtain a set of estimated coefficients

III.C If the relevant p-value for the Wald-test corresponding to the coefficients of interests is less than α, we deem this iteration as "Successfully reject"

IV. compute the ratio of "Successfully reject" iterations to total number of iterations, which will serve as the estimate of power under this fixed sample size

V. Repeat **II** to **IV** until we sweep over all the sample sizes of interest

2.2.3 Results

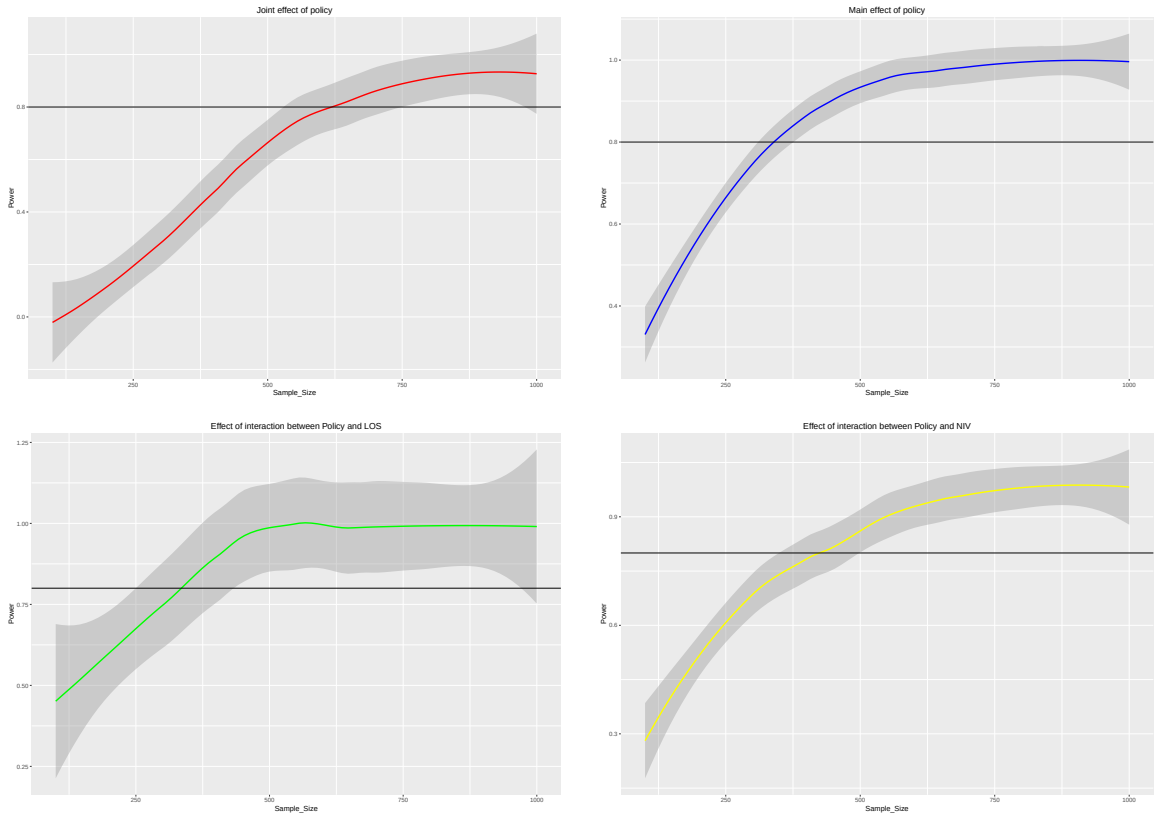

Figure 1: Top left: The estimated power of joint effect of policies vs sample size. Top right: The estimated power of main effect of policies vs sample size. Bottom left: The estimated power of effect of interaction between LOS and policies vs sample size. Bottom right: The estimated power of effect of interaction between IV and policies vs sample size.

From above power curves, under 0.05 significance level and specified effect size, we can see in order to effectively detect (0.8 detection probability when effect exists) individual effects of policies upon "indwelling time", such as main effect of policies, effect of interaction between policies and Length of stay, and effect of interaction between policies and Number of inserted IVs, under 0.05 significance level, we will need about 300 patients in each case. However, to effectively detect the three-way joint effect, it would need about 625 patients, roughly double the number needed for individual effects.

2.3 Binary responses with only fixed effects

Here, we will establish the relationship between "removal due to symptom" and different policies with following model.

2.3.1 Model setup

"removal due to symptom" (Y_i) is modeled as Bernoulli distribution (taking value of 1 if one of the IVs of patient i is removed due to certain symptom, 0 otherwise) with $\Pr(Y_i = 1) = \mu_i$. Here we will parameterize the probability μ_i by

$$\log(\frac{\mu_i}{1-\mu_i}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + (\beta_3 + \beta_5 X_{5i}) X_{3i} + (\beta_4 + \beta_6 X_{5i}) X_{4i} + \beta_7 X_{5i},$$

where X_{1i} denotes the age of patient i, X_{2i} is an indicator with value 0 if patient i is female and 1 otherwise, X_{3i} stands for the length of stay (days) for patient i (LOS), X_{4i} denotes the number of IV catheters inserted into patient i (NIV) and X_{5i} takes value of 1 if patient i is from routine group, 0 if patient i is from clinically indicated group.

Following previous assumptions, we will treat the estimated coefficients obtained by fitting above generalized linear model with pilot data as the actual coefficients, which will act as our control for effect size. Table 2 shows all the estimated coefficients.

Following the same Monte Carlo procedure as described in last section, with only differ-

Table 2: Assumed actual coefficients

β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
-2.495	-0.007	-0.112	0.157	0.859	-0.164	-0.466	2.003

ences being, within each iteration, we will now generate responses from a Bernoulli distribution instead of a normal and, we will fit a generalized linear model with our generated data instead of a linear model.

2.3.2 Results

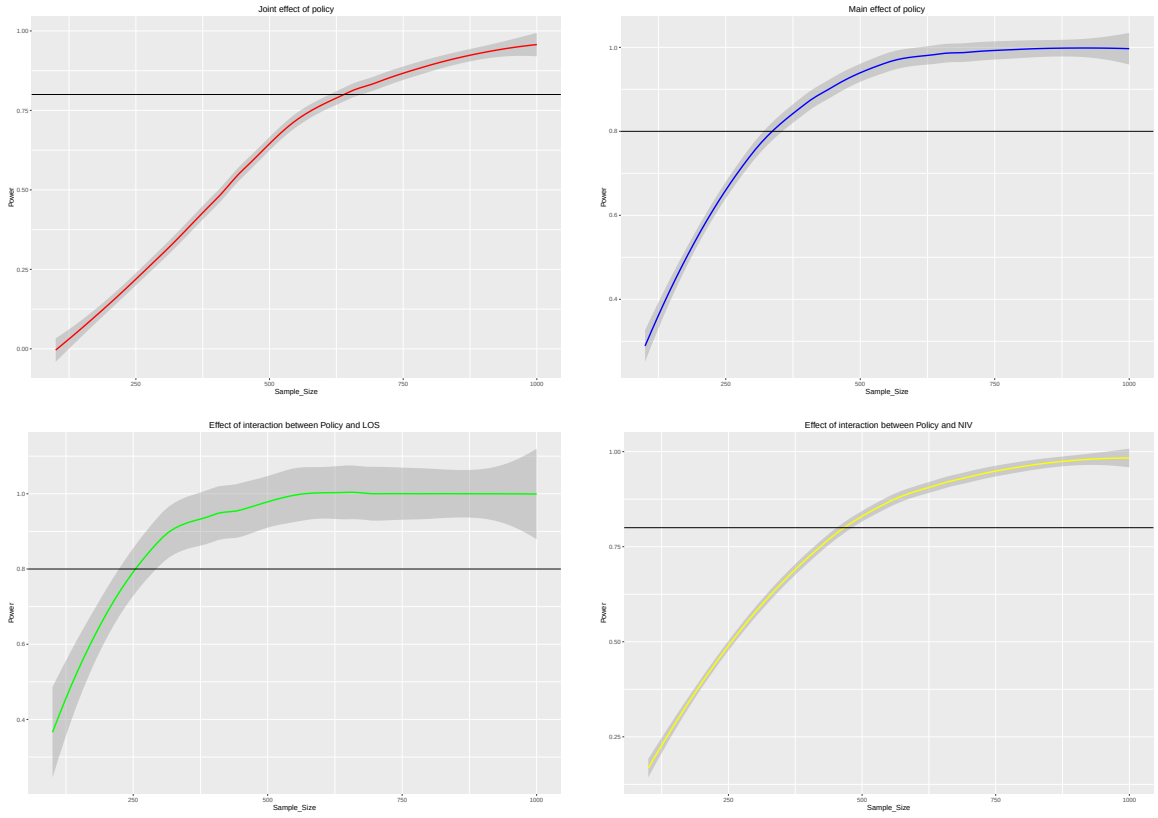

Figure 2: Top left: The estimated power of joint effect of policies vs sample size. Top right: The estimated power of main effect of policies vs sample size. Bottom left: The estimated power of effect of interaction between LOS and policies vs sample size. Bottom right: The estimated power of effect of interaction between IV and policies vs sample size.

From above power curves, under 0.05 significance level and specified effect size, we can see, to effectively detect the main effect individually, we would need about 300 patients, and

to effectively detect the effects of two interactions individually, we would need about 250 and 400 patients in respective cases. Furthermore, the three-way joint effect requires about 625 patients to reach our desired power level.

2.4 Discussion

In general, I would say, given that the future clinical trial, designed to be balanced, will be conducted in one or multiple similar medical units, and our pilot data is a fair representation of the reality, 600 to 650 patients are the optimal number of patients needed to effectively capture the effects we intend to measure in the future clinical trial.

3 Multiple Units

In this section, we will generalize the existing model to capture the variability across multiple different unites. Following the similar strategies in Section 2, we will build up a framework that model the "indwelling time" and "removal because of symptoms" separately. To present our simulation results, this section is organized as follows. Section 3.2 is mainly focusing on modeling the "indwelling time". We will report the association between the power of particular predictor and sample size. Suggested sample size for nursing policy makers will be provided at the end. Section3.3 is devoted to capturing "removal because of symptoms". The analysis strategies will be similar to Section 3.2.

3.1 Assumptions

Patients are likely to be residing on multiple care units in the hospital in the course of a hospitalization. It is still possible that multiple units during a hospitalization will impact on the quality of health care (Kanak et al., 2008). During the past few decades, such effects have been widely study. Kanak et al. (2008) implemented general linear model to investigate the effect of hospitalization on various units. Their work mainly focus on the effect on selected

nursing treatments, resource use, and clinical outcomes. Bacon and Mark (2009) studied the effects of patient characteristics and satisfaction on nursing unit structure. Inspired by the aforementioned results, we strongly believe that the effect from various units on either "indwelling time" or "removal because of symptoms" is non-negligible. Due to the large amount and variability across units, we will introduce linear mixed effect model with the random effects capturing the unit factor. From existing studies in the literature, there is no substantial results in regarding to this field. Our methods, to this end, serve as an extension and frontier strategy that takes multiple units into account. It could serve as an extension of the framework in Section 2. In addition, the results from this model can also be utilized when future studies aim to a more diverse and broad scale. Particularly, on the modeling side, the fixed effects include patients's age, gender, length of stay days (LOS), number of IVs and the treatment groups (routine group or clinically indicated group). As for the random effects, we assume patients' baseline varies across different units and thus we use a random effects across units to capture this discrepancy.

In summary, we will generalize the framework in Section 2 to linear mixed effect model to capture the variability and impact from multiple units. Following similar arguments from Section 2, we will concentrate on "indwelling time" and "removal because of symptoms" separately as dependent variables. The other covariates include age, gender, ethnicity, length of stay (LOS), number of IVs inserted into patient (NIV) and group.

3.2 Gaussian responses with linear mixed effects

3.2.1 Model setup

In this section, we will concentrate on modeling "indwelling time". Following the discussion in Section 3.1, the model is summarized as follows

$$
\begin{aligned}
&Y_{ij}|\mu_{ij} \sim N(\mu_{ij}, \tau^2),\\
&\mu_{ij} = \beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \beta_3 X_{3ij} + \beta_4 X_{4ij} + \beta_5 X_{5ij} + \gamma_i,\\
&\gamma_i \sim N(0, \sigma^2),
\end{aligned}
$$

where X_{1ij} denotes the age of patient j at unit i, X_{2ij} denotes the gender that 0 if the patient is female and 1 otherwise, X_{3ij} denotes the length of stay days (LOS), X_{4ij} denotes the number of IVs inserted into the patient, X_{5ij} is the group indicator where 1 means the patient is from routine group and 0 from clinically indicated group. γ_i models the random effect from different units. Specifically, it captures the stochastic discrepancy between patients reside in various units. Patients's baseline situation varies randomly across different units.

3.2.2 Simulation procedures

We utilized Monte Carlo method to estimate the power with respect to different sample size. Note that the effective size in this study is obtained from pilot analysis.

The procedure are summarized as follows

I. Obtain estimates $\hat{\beta}, x_{ij}$ from pilot analysis

II. For every fixed sample size

Repeat

II.A Generate simulated dataset

II.B Fit GLMM to obtain point estimates of fixed effect

II.C Conduct hypothesis tests

III. Obtain Monte Carlo estimates of the power for this particular sample size

IV. Repeat the previous steps until we sweep over all the sample sizes of interest.

3.2.3 Results

We implemented the method discussed in Section 3.2.2 to estimate the power corresponding to variables of age, gender, LOS, IV and Group factors. Figure 3 shows the relationship between power and sample size with respect to the underlying variables. We could clearly observe a non-descending pattern and flatted out as the sample size grows large. Figure 4 shows the power versus sample size on variable "group". Compared to the other variables, "group" is likely to gain enough power at a relatively small sample size.

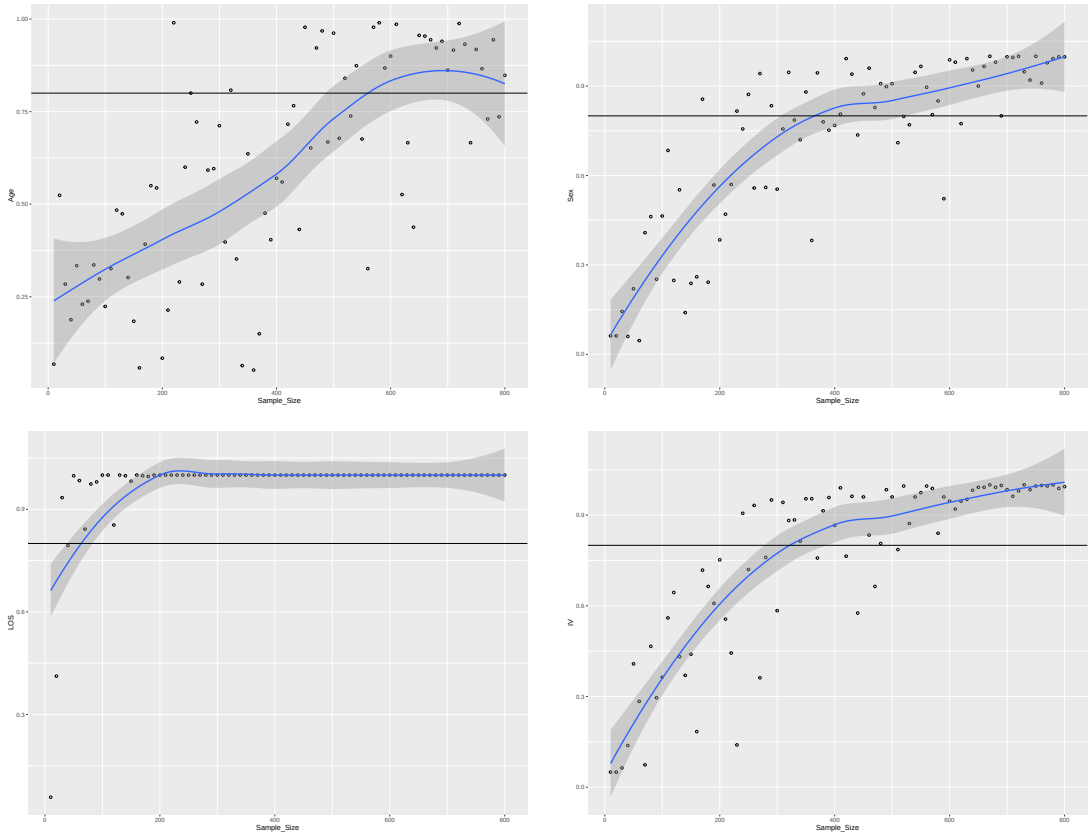

Figure 3: Top left: The estimated power of age effect versus sample size. Top right: The estimated power of sex effect versus sample size. Bottom left: The estimated power of LOS effect versus sample size. Bottom right: The estimated power of IV effect versus sample size.

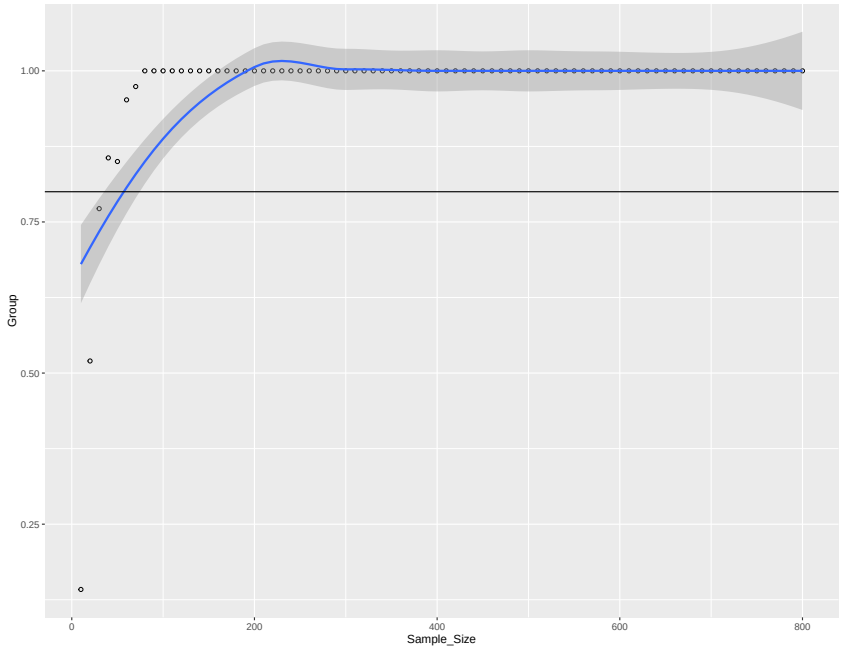

Figure 4: The estimated power of group effect versus sample size.

Table 3: Minimum sample size to obtain 80% power

Variable	Sample size
Age	580
Sex	350
LOS	80
IV	330
Group	85

To summarize the simulation results, Table 3 shows the recommended sample size that corresponds to each of the variables of interest. Similar to the results from the previous analysis, "group" and "LOS" variables require the least among all. Due to the large amount of levels, variable "Age" needs the most sample size.

3.3 Binary responses with linear mixed effects

3.3.1 Model setup

In this section, we focus on modeling the variable "removal because of symptoms", the model is summarized as follows

$$
\begin{aligned}
&Y_{ij}|\eta_{ij} \sim \text{Bernoulli}(g(\eta_{ij})),\\
&g(\eta_{ij}) = \frac{1}{1+\exp(-\eta_{ij})},\\
&\eta_{ij} = \beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \beta_3 X_{3ij} + \beta_4 X_{4ij} + \beta_5 X_{5ij} + \gamma_i,\\
&\gamma_i \sim N(0, \sigma^2),
\end{aligned}
$$

where X_{1ij} denotes the age of patient j at unit i, X_{2ij} denotes the gender that 0 if the patient is female and 1 otherwise, X_{3ij} denotes the length of stay days (LOS), X_{4ij} denotes the number of IVs inserted into the patient, X_{5ij} is the group indicator where 1 means the patient is from routine group and 0 from clinically indicated group. γ_i models the random effect from different units in a similar way shown in Section 3.2.1.

3.3.2 Results

Following the similar procedures shown in Section 3.2.2, we obtained the estimated power from simulations. Compared to the results from Gaussian case, the binary model needs more sample size to gain power. From Figure 5 and 6, it can be found that even when the sample size gets close to 3000, "sex" variable still could not gain enough power.

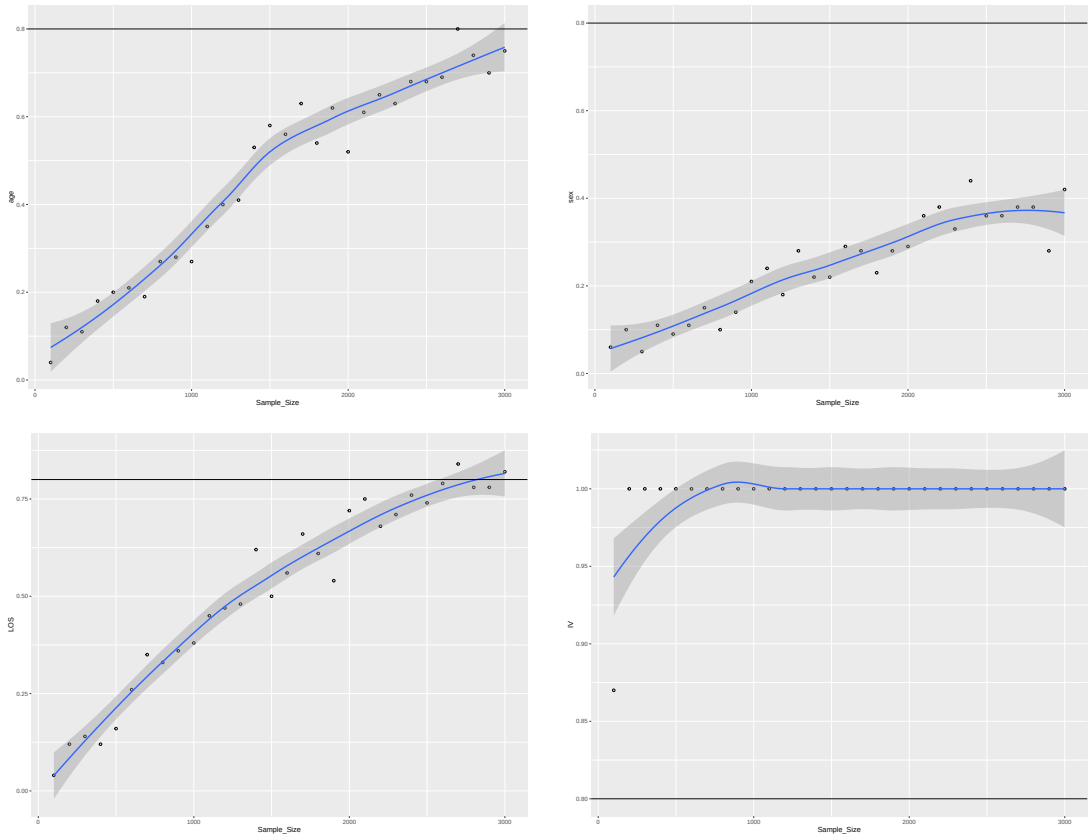

Figure 5: Top left: The estimated power of age effect versus sample size. Top right: The estimated power of sex effect versus sample size. Bottom left: The estimated power of LOS effect versus sample size. Bottom right: The estimated power of IV effect versus sample size.

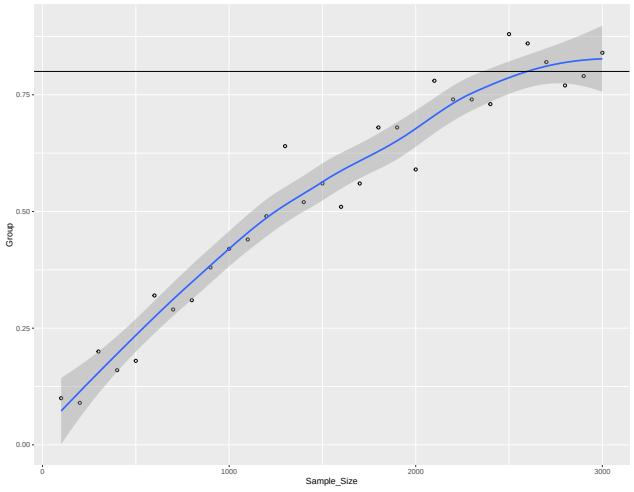

Figure 6: The estimated power of group effect versus sample size.

Table 4: Minimum sample size to obtain 80% power

Variable	Sample size
Age	> 3000
Sex	$>> 3000$
LOS	2800
IV	500
Group	2600

To summarize the results for binary case, we present Table 4 that shows the recommended sample size corresponding to each of the variables. "IV" variable needs the least size in comparison with "Age" and "Sex" that requires unreasonably large sample sizes.

4 Conclusions

According to the aforementioned simulation results, our recommendations for the sample size are summarized as follows: (1.) With the assumption that the variance of "indwelling time" is between 2.00 to 3.00, the effect size of the difference between the true mean of the populations to be 0.30 to 0.45, power to be 80% and probability of type I error to be 5%, the required sample size is between 600-650; (2.) Under the same assumption given in (1.), if we move further to capture the variability across multiple care units, the sample size is suggested to be about 2600; (3.) If the effect size is below 0.30 (or above 0.45) in further study, we would suggest to increase the sample size (or decrease the sample size) in order to obtain the power of 80% under type I error rate of 5%. As the effect size closes to 0, the increment of required sample size is dramatic. (4.) If the variance goes below 2.00 (or above 3.00), we would suggest to decrease the sample size (or increase the sample size) to achieve the power of 80% under type I error rate of 5%.

Due to the limited pilot data and insufficient approximation of reality, there are still

some potential opportunities to improve the analysis presented in this report. For example, as more data come in, the variance of "indwelling time" and other prespecified parameters in the assumptions will get closer to the real world and thus the recommended sample sizes will become more compelling. Following the same procedure proposed in this report, the recommended sample sizes will be obtained without much efforts, should more information and data come into play in the future. In summary, this study not only provides suggested sample sizes for the future clinical trial, but also builds a pipeline of estimating sample sizes in a more general framework.

References

Bacon, C. T. and B. Mark (2009). Organizational effects on patient satisfaction in hospital medical-surgical units. *The Journal of nursing administration 39*(5), 220.

Dang, Q., S. Mazumdar, and P. R. Houck (2008). Sample size and power calculations based on generalized linear mixed models with correlated binary outcomes. *Computer methods and programs in biomedicine 91*(2), 122–127.

Dao, L.-A. T. (2016). Comparison of peripherally inserted intravenous catheter complication prevalence: Before and after changing a 96-hour routine replacement standard.

Kanak, M. F., M. Titler, L. Shever, Q. Fei, J. Dochterman, and D. M. Picone (2008). The effects of hospitalization on multiple units. *Applied Nursing Research 21*(1), 15–22.

Rickard, C. M., W. M. C. Webster, J., N. Marsh, M. R. McGrail, . French, V., and A. McClymont (2012). Routine versus clinically indicated replacement of peripheral intravenous catheters: a randomised controlled equivalence trial. *The Lancet 380*(9847), 1066–1074.