HHS Public Access

Author manuscript

Exp Neurol. Author manuscript; available in PMC 2016 August 01.

Published in final edited form as:

Exp Neurol. 2015 August ; 270: 82–87. doi:10.1016/j.expneurol.2015.02.024.

Statistical Considerations for Preclinical Studies

Inmaculada B. Aban, PhD and **Brandon George, PhD**

Department of Biostatistics, University of Alabama at Birmingham, Alabama, United States

Abstract

Research studies must always have proper planning, conduct, analysis and reporting in order to preserve scientific integrity. Preclinical studies, the first stage of the drug development process, are no exception to this rule. The decision to advance to clinical trials in humans rely on the results of these studies. Recent observations show that a significant number of preclinical studies lack rigor in their conduct and reporting. This paper discusses statistical aspects, such as design, sample size determination, and methods of analyses, that will help add rigor and improve the quality of preclinical studies.

Keywords

Preclinical studies; sample size; power; randomization; multiple outcomes; false positive; missing data

Introduction

The development of new therapy for a particular disease from concept to market is an extensive process that is costly in terms of time, effort and finances. The process starts with preclinical studies involving in vitro (e.g., tissue culture studies) and in vivo (animal studies) experiments in a laboratory. When the required information and results are obtained from preclinical studies, an Investigational New Drug (IND) application is submitted to the Food and Drug Administration (FDA) accompanied by the results of the preclinical studies. Researcher are allowed to conduct studies in humans only after receiving an approved IND.

Human studies start at Phase I where human volunteers are recruited with the goal of obtaining information about the side effects of the drug, and in some cases, determining the maximum tolerated dose. Phase II begins after Phase I study shows no issues with toxicity. The goal of a Phase II study is to obtain preliminary information that will show some indication of effectiveness and safety of the drug applied to the population with the disease targeted by the new therapy. After successful completion of the Phase II study, a large-scale

Corresponding author: Inmaculada B. Aban, Professor of Biostatistics, University of Alabama at Birmingham, 1530 Third Avenue South, Birmingham, AL 35294-0022, United States, Tel.: +1 205 934 2732, caban@uab.edu.

Phase III clinical trial is conducted with the goal of establishing evidence of effectiveness in a broader and larger population as well as collecting additional information about safety. Upon successful completion of Phase III, a New Drug Application (NDA) is filed to the FDA to obtain approval to market the drug. In the NDA, results from the animal studies and human studies (phases I–III) are reviewed by FDA before giving the final stamp of approval. The last phase of the drug development (Phase IV) is post-marketing surveillance. Figure 1 summarizes the different stages in drug development.

Clearly, preclinical studies being the first stage in the process play a crucial role in drug development. Unfortunately, a high proportion of these preclinical studies conducted on animals that indicated some therapeutic effect do not translate to similar results in studies in humans. This issue is mostly attributed to poor planning, conduct and reporting of most preclinical studies (see for instance Perin, 2014; Warner et al., 2014; Henderson et al., 2013; Landis et al., 2012; and Kilkenny et al., 2009) Consequently, the National Institute of Neurological Diseases calls for more rigorous reporting of these studies to raise awareness on the proper design and conduct of future preclinical studies as well as the proper interpretation of the results of completed studies (Landis et al., 2012). In line with this goal, this paper aims to review some of the basic statistical elements of clinical trials which will help researchers understand and appreciate the relevance of these concepts in the context of preclinical studies.

Study Planning and Conduct

The details of how the study will be conducted relies heavily on the question. Without a well-defined question or hypothesis, the study will most likely result in a “fishing exploration”. Given the question of interest, primary outcome can be defined and appropriate study design can be chosen. The number of primary outcomes should be kept at a minimum; the ideal case would be to have only one primary outcome. However, this may not be possible in some cases. For instance in myasthenia gravis (MG) animal studies, therapeutic effect may be reflected in different aspects such as change in strength, weight, disease severity, serum cytotoxicity and acetylcholine receptor (AChR) antibody concentration to name a few. Having multiple outcomes as contrasted to single outcome will have consequences in the sample size calculation and data analysis.

Design and Sample Size

Design of the study, characteristics of the outcome, and the number of outcomes are some factors that affect the determination of the appropriate sample size. Researchers must carefully consider the choice of study design based on the question of interest. They must aim to use the simplest appropriate design as study design dictates the method of data analysis and interpretation, and the method of data analysis dictates the method of calculating sample size. The most popular design used is the parallel group design where different animals are used in each of the M treatment group. The simplest of this design is the case where there are only two groups, i.e, $M = 2$, for example comparing the outcome of untreated group to the outcome of the group treated with a new drug. The analysis associated with this design is typically a t-test for two independent samples when the outcome follows a normal distribution or a Fisher’s exact test (or chi-squared test for large

samples) to compare two proportions when the outcome is a binary variable (e.g., with improvement or no improvement). Increasing the number of groups to compare, say from 2 to 3, will increase the required sample size. Designs such as cross-over design, where each animal serves as their own control, will require smaller sample size than parallel design but it has other requirements that may not be feasible for some experiments (for instance, cases where animals are euthanized to obtain the outcome of interest). Having multiple primary outcomes which then result in multiple statistical testing in the data analysis stage will require larger sample size relative to a single outcome due to the required adjustments necessary to avoid inflation of the false positive error rates. When the outcome is binary (e.g., compare the proportion showing improvement between the treated and untreated group), a larger sample will be required compared to the case where the outcome is continuous (e.g., measuring actual weight or strength). Also, the case where one of the two binary outcomes is rare in both groups will require a larger sample size than a case where both possibilities are common.

When the outcome of interest is the time to occurrence of an event where methods of data analyses are based on survival analyses, power is highly dependent on the expected number of events for a given period of time in addition to the overall sample size, and the number of events that will be observed is highly dependent on the length of follow-up time. Censored observations, i.e., outcomes of subjects who did not experience the event due to drop-out or end of follow-up, are not uncommon in survival analysis studies. However, the higher percent of censoring the less amount of information is available resulting in lower power to detect a given effect size. Therefore, power can be increased while keeping the sample size and effect size constant by increasing the follow up time that will result in an increase in the expected number of observed events in that period. Note that although it may be of clinical interest to model a continuous outcome as the time for it to reach a certain cutoff point and use survival analysis methods, doing so sacrifices a large amount of statistical efficiency (e.g., loss of power) and thus should be avoided (Zucker, 2012).

We illustrate the process of sample size determination for a single continuous outcome based on a two-tailed t-test for comparing two independent sample (e.g., untreated versus treated groups). For an MG study, one may be interested in the grip strength as the outcome. Assuming grip strength (in grams) follows a normal distribution, we examine how the power changes across different scenarios. We set the significance level at 5%, the mean grip strength of the placebo group at 400 and standard deviation of 20 (common between the treatment groups). The effect size is defined as the difference between the mean grip strength of the untreated and treated group. Figure 2 shows the power for the different values of the mean of the treatment group for a given group sample size. [Calculations done using PASS 11 software (Hintze, 2011).] As the mean of the treated group increases relative to the mean of the untreated group, the effect size also increases, and the sample size necessary to achieve a fixed desired power level decreases. Another way of looking at this is that for a fixed sample size, the power increases as the effect size increases. These observations are intuitive because a larger difference is easier to detect than a smaller difference. Figure 2 may also be used to conclude that 5 rats in each group is sufficient to detect an effect size of 40 grams (i.e., an increase in expected grip strength from 400 to 440 grams) or more with power of about 80%. Although not shown, increasing the value of the

assumed common standard deviation (fixing all other parameters) will require larger sample size because a larger standard deviation implies more noise in the outcome which makes it more challenging to detect a given treatment effect. Increasing the sample size will compensate for the increase in noise and achieve the desired power. We highly recommend that researchers investigate different scenarios in a manner similar to this illustration when determining the sample size so that they are aware of what to expect from the study if an assumption about a parameter is not correct.

Using the above illustration, suppose that we categorize the actual grip strength as at or above a certain cut-off. Thus, the primary outcome is changed from the actual grip strength to a binary "success" or "failure" based on a target grip strength. Assuming the same normal distribution for actual grip strengths in each group as described in the preceding illustration (i.e., mean of 400 grams in placebo group and 440 grams in treated group with common standard deviation of 20 grams), the sample size required to test the proportion of animals demonstrating the target grip strength depends on the cut off. If the target grip strength is defined as 430 grams, then it is expected that 6.68% of the placebo and 30.85% of the treated group will demonstrate a grip strength of at least 430 grams. To detect a difference of 24.17% (=30.85–6.68) or higher in proportion of animals demonstrating this target grip strength, using a two-tailed large-sample z-test for two independent proportions with 5% significance level and power of at least 80%, the number of animals in each group assuming equal allocation is calculated to be 37 requiring a total of 74 animals (using PASS 11). This is a case in point of binary outcomes requiring larger sample size than a continuous outcomes – compare 74 animals required for this binary outcome example to the 10 animals that we previously calculated when the primary outcome is the actual grip strength. If the target is decreased to 425 grams, the proportions of animals reaching this target change to 10.56% in the placebo and 22.33% in the treated groups. The sample size required in each group increases to 148 (a total of 296 animals) because the effect size to be detected is down to 11.77%.

Although sample size is usually associated with power, there may be preclinical studies for which the goal is to estimate the treatment effect on a particular outcome (for instance, AChR concentration level) as preliminary data for subsequent preclinical studies. In this case, sample size calculation will be based on a confidence interval around the treatment effect. Instead of power, the goal is to minimize the margin of error of the confidence interval for a fixed confidence level (typically at 95% confidence level).

As a final note for this section, researchers must avoid using rules of thumb for sample sizes. As an example, in regression modeling, a commonly used rule of thumb is to have 10 observations per predictor variable so that if one is interested in testing 5 predictor variables, 50 observations must be obtained. The problem with rules of thumb is that, in most cases, there is no theoretical justification for the rule. Referring back to the regression model example, one should define the main goal to determine the appropriate sample size. For instance, is the goal to find predictors, or is the goal to accurately predict the outcome? Sample size for regression depends on the variability of the outcome, the number of predictor variables of interest, the number of covariates to be adjusted for, and the degree of correlation among the variables (outcome, predictors and covariates) to name a few. A

researcher must work with a biostatistician in determing the best approach to determine and justify the sample size based on the specific aim of the study.

Randomization and Blinding

Another important aspect of study design for a study with the goal of comparing different groups is the process of treatment assignment. To avoid selection bias and minimize confounding of treatment effect with covariates (measured or unmeasured), treatment assignment must be done in a random manner. Consider a study that compares two treament groups. Figure 3 displays 3 different schemes (S1, S2 and S3) to assign the treatments (new drug=T, control=C). Schemes S1 and S2 are not acceptable because they follow a predictable pattern, and hence may still result in selection bias. Scheme S1 also has the problem that treatment effects will be confounded with the conditions in the laboratory during the early/later part of the study. Scheme S3 follows no predictable pattern and the treatment and control are distributed across the study period, and hence, a good treatment assignment scheme produced through randomization.

Randomization is a process of assigning treatment to subjects as they are enrolled in the study in a non-subjective and unpredictable manner. Randomization scheme must be determined ahead of time and must only be known to the staff who would not be involved in the laboratory. Laboratory staff in charge of giving the treatment will only be aware of the treatment assignment(s) at the time that the animals are ready to be treated. There are variations in randomization schemes to accommodate availability of animals. One of these variations is stratified randomization. For instance, if animals arrive in batches, researchers need to make sure that treatment types are well represented in each batch or group. Stratified randomization creates a different randomization scheme for each batch (i.e., stratum) as if each batch is a different study and will be more appropriate than the regular randomization scheme for the whole study.

In addition to avoiding or minimizing selection bias with randomization of treatment, observer bias must also be minimized. Observer bias typically occurs when the outcome of interest is measured in a more subjective manner. In MG preclinical studies, an observer bias may influence the assessment of strength or disease severity. To minimize observer bias, the person evaluating the outcome and providing dose (if not fixed ahead of time) must be blinded to the treatment group.

Data Analysis

Parametric versus non-parametric tests

T-tests for paired data and two independent samples as well as analysis of variance model (ANOVA) are common statistical methods used in studies. These tests are known as parametric tests because it is assumed that the continuous outcome being analyzed follow a particular parametric distribution, typically the normal distribution. In addition, these tests compare the mean outcomes of the different groups. Even if the outcome does not follow a normal distribution, these tests are still applicable (in an approximate sense) as long as the sample size is large enough (as a result of the Central Limit Theorem in statistics). In some cases, where the outcome distribution is skewed instead of symmetric, applying a

transformation (such as taking the logarithm) to the outcome may result in the transformed data following a normal distribution. Consequently, standard methods such as t-test and ANOVA, may be used on the transformed data.

In studies where the sample sizes are small (say less than 10 in each group), the assumption that outcome variable follow a normal distribution must be checked carefully. If there is evidence of violation of this assumption or if there are outliers (extreme observations) which can easily affect the sample mean, the above methods may no longer be applicable and may provide misleading results. In these cases, non-parametric tests must be considered. The commonly used non-parametric tests are the Wilcoxon rank sum test and the Mann-Whitney test which are the non-parametric analogs of t-tests for comparing two groups and Wilcoxon signed rank test for paired data. Kruskall-Wallis test is the analog of the ANOVA test for comparing more than two independent groups.

Nonparametric tests do not assume a particular form of the distribution as long as it is continuous. These tests compare the medians of the groups instead of the means and use the rank of the observations instead of the actual observed values in the calculation of the test statistic. Therefore, non-parametric methods are more robust to outliers. The trade-off of using non-parametric tests is some loss of power because ranks are used instead of the actual observed values.

Outcomes using scales (discrete ordinal categories), although numerical, should not be analyzed using methods that assume a normal distribution (such as the t-test or ANOVA). An example is the widely accepted assessment of MG disease severity in animal studies where: 1= can grip but cannot lift the lid of a cage; 2 = unable to grip cage lid; 3 = unable to grip and has hind-limb paralysis; and 4 = moribund (see for instance Piddlesden et al., 1996; Soltys, 2009). When the goal is to compare two or more independent groups based on this ordinal outcome, Mantel-Haenszel test is an appropriate test. In the special case where only two groups are being compared, Mantel-Haenszel test is equivalent to the Mann-Whitney test. When modeling is desired to add more predictor variables and/or adjust for covariates, the cumulative logit model is preferred over the conventional regression model. It is important to note that most of these methods for analyzing categorical data are based on the assumption of large sample. If the study has a small sample size compared to the number of predictor variables, such as in preclinical studies, methods based on exact distribution must be used. For a binary outcome, conditional logistic model is recommended for small sample studies. [For more details about the analysis of categorical outcomes and conditional logistic model, see Agresti, 2007.]

The value of p-values

Recall that in statistical hypothesis testing, there are two types of errors that can be committed. Type I error is incorrectly concluding that there is a treatment effect (i.e., false positive) while type II error is incorrectly concluding that there is no treatment effect (i.e., false negative). Procedures used to conduct statistical test of hypotheses are based on the assumption that the null hypothesis of no treatment is true unless there is evidence from the data collected to reject this null hypothesis.

P-values are commonly used to report study results and make conclusions. However, a large number of researchers do not completely understand the concept of p-values resulting in the misuse of this quantity. P-values quantify the probability of committing a type I error (false positive): the smaller the p-value, the more unlikely it is for the null hypothesis (say, no treatment effect) to be true in light of the data that will be collected. A large p-value does not support the null hypothesis. If one fails to get a significant result, it is not correct to conclude that there is no treatment effect because the lack of significant result may either be due to lack of power (sample size not large enough) to detect the treatment effect or that indeed there is no treatment effect.

P-values do not measure the success of the study. It is a function of the sample size, and simply increasing the sample size will make the p-value small enough to reject the null hypothesis and conclude a significant treatment effect. This effect may be signficant statistically but not clinically. It is therefore important to report confidence intervals for treatment effect with p-values (significant or not) because confidence intervals provide information on the magnitude of the treatment effect and the amount of error associated with the estimate given by the width of the interval. [For other discussions on the misconceptions about p-values, see Goodman, 2008.]

Addressing multiple outcomes and multiple testing

As previously stated, effects of therapy for MG are typically measured in multiple outcomes. Until there is an agreement on a single outcome of interest, researchers need to be aware of the proper analyses of these outcomes. The main issue in the analysis of multiple outcomes is the testing of hypotheses associated with each of the outcomes. Multiple testing runs a high risk of inflating the type I error rate (i.e., the probability of false positives). In statistical hypothesis testing, if a single hypothesis of interest is that the new treatment is better than no treatment as measured by a defined outcome, the standard rule for concluding evidence of a treatment effect is when p-value associated with the test statistic for that particular outcome is less than 0.05.

It is implied in this analysis that at the planning stage of the study, the type I error rate was set at 5% which means that the researcher is tolerating at most 5% probability of a false positive. When the researcher is interested in testing treatment effects based on two distinct outcomes, the typical approach is to conduct two separate tests, for instance t-tests, using the same rule of p-value< 0.05 to determine significance in each test. Assuming that the two outcomes are independent, the probability of concluding a false positive (FP) in at least one outcome is

$$\begin{aligned} Prob(\text{FP in at least 1 outcome}) &= 1-[Prob(\text{TP in outcome 1})Prob(\text{TP in outcome 2})] \\ &= 1-(1-0.05)^2 = 0.0975. \end{aligned}$$

where *TP* denotes true positive. Although we set the level of significance of each test at 5%, combined results from the two tests inflated the type I error rate to 9.75%. When outcomes

are correlated, the inflation may be more or less than 9.75% depending on the degree and direction of the correlation but there is still inflation in the type I error rate.

Even when there is only a single outcome, multiple testing may still occur if the single outcome is repeatedly collected over time. Consider the study of measuring grip strength weekly over 6 weeks for treated versus untreated groups. Some reseachers will compare the grip strength of the two groups at each week by using a t-test resulting in a total of six tests performed. In addition to the inflation of the type I error rate, separate t-tests ignores the correlation among the repeated readings which contain information that may help increase the power of the test.

How then can we handle this issue of multiple testing? We present three relatively simple methods of handling multiple testing by adjusting the significance level. Let k be the number of outcomes to be tested and the desired significance level set at 5%. The goal of these methods is to adjust the rule to determine significance so that the resulting probability of getting at least 1 false positive in testing k outcomes is no more than 5%. The simplest adjustment is known as the Bonferroni adjustment, and it states that only outcomes with p-values $< 0.05/k$ are considered to show significant treatment effects. The problem with this method is that it is too conservative, i.e., the true probability of false positive when using this method is much lower than 5%. Hence, there is a higher likelihood of missing true positive effects, thus, lowering the power of the the study.

A modification of the Bonferroni method that alleviates the issue of being too conservative is the Bonferroni-Holm step-down (Holm, 1979). The procedure is as follows.

1. Arrange the p-values corresponding to the null hypotheses $H_{(1)}, H_{(2)}, \ldots, H_{(k)}$ for the k outcomes in ascending order, i.e., $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(k)}$.
2. If $p_{(1)} < \alpha/k$, then reject $H_{(1)}$. Otherwise, conclude that none of the k outcomes showed significant treatment effect.
3. If $p_{(2)} < \alpha/(k-2)$, then $H_{(2)}$. Otherwise, stop and conclude that only the outcome reject associated with $H_{(1)}$ is significant.
4. Continue this process until a stopping rule is encountered or all k outcomes have been considered.

The last method is the Benjamini-Hochberg step-up method (Benjamini and Hochberg, 1995) which adjusts the tests with the goal of controlling the false discovery rate (FDR) in contrasts with the two previous methods which controls the false positive rate. FDR is defined as the expected proportion of false positives among the “discoveries” which in our case are the outcomes that resulted in significant treatment effect. The procedure is as follows.

1. Arrange p-values in ascending order: $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(k)}$ corresponding to the null hypotheses $H_{(1)}, H_{(2)}, \ldots, H_{(k)}$, respectively.
2. If $p_{(k)} < \alpha\left(\frac{k}{k}\right) = \alpha$, then reject k hypothesis and conclude all outcomes show significant all treatment effects.

3. If $p_{(k-1)} < \alpha(\frac{k-1}{k})$, then reject $H_{(1)}, \ldots, H_{(k-1)}$ and conclude that outcomes associated with these hypotheses show significant treatment effects.

4. Continue this process until a stopping rule is encountered or all k outcomes have been considered.

We illustrate these methods using a hypothetical example. Suppose that there are $k = 6$ outcomes and the desired overall significance level is set at 5%. The ordered observed p-values and the adjusted cut-off points for each method are given in Table 1. Based on the Bonferroni method, the outcome associated with the smallest p-value, $p_{(1)} = 0.001$, is the only significant outcome. When using Bonferroni-Holm method, the outcomes associated with the smallest four p-values (i.e., $p_{(1)}$, $p_{(2)}$, $p_{(3)}$ and $p_{(4)}$) are significant. Finally, using the Benjamini-Hochberg method, all outcomes except the outcome associated with the largest p-value are significant.

Given the necessary adjustments in the cut-offs to determine significance when there are multiple outcomes, these adjustments must also be considered in the planning stage of the study when appropriate sample size is being determined. A simple and straightforward way to adjust for the multiple outcomes is to use the Bonferroni method where one modifies the significance levels accordingly. Referring to the above example, when 6 primary outcomes are of interest and the level of significance for the combined outcomes is set at 5%, then the level of significance used to determine the effect size associated with each individual outcome for a fixed power and sample size should be 0.083%, resulting in the effect size needing to be larger to achieve the same power than the effect size needed based on an unadjusted 5% significance level.

There are other methods that analyzed multiple outcomes but these are more complex than the adjustment methods such as those described above. An example of this method is the multivariate version of ANOVA known as MANOVA. This method simultaneously tests the outcomes, controls the overall type I error and considers the correlation among the outcomes in the analysis. However, as in ANOVA, MANOVA assumes the outcomes are at least approximately normally distributed. If some primary outcomes are continuous while others are binary, MANOVA will not be appropriate.

For longitudinal data where a single outcome is measured repeatedly over time, methods such as mixed models may be utilized. In mixed models, one may either specify a working correlation matrix or use random effects to account for the correlation among repeated observations from the same subject. These models have the capability of comparing the trajectories that describe the progression of the outcomes over time between the groups of interest. Appropriate specification of the trajectories and correlation structure will help increase the power of the test. MANOVA may also be used but assumes a balanced case. In our example, if the grip strength for a particular animal is available except for week 4, then this animal will be excluded from the analysis. Mixed models can handle cases with missing data but assume normal outcomes. For outcomes that do not follow a normal distribution (such as counts or binary outcomes) measured over time, models based on generalized estimating equations may be used. (See Albert, 1999, for a tutorial on methods of analyzing longitudinal data applied to clinical trials.)

Missing data

Missing data is not uncommon in preclinical studies and should not be ignored. Circumstances possibly explaining the missing data must be clearly documented, investigated and understood, if possible, to determine if the missingness is related to treatment, animal model or other experimental factors. The information gathered in this investigation must be included in the final report as this will aid in the interpretation of the results (e.g., state as a limitation of the study). In addition, it will help in the planning of future studies by avoiding the circumstances that led to the missing data.

How should missing data be handled in the analyses? Bias may be introduced if only animals with complete data are used. Multiple imputation is one method to handle missing data. In simple terms, it is a method of obtaining a value for the missing data through simulation based on assumptions about the mechanism of missingness and the probability model associated with the outcome with missing data (see for instance Little and Rubin, 2002). Researchers must consult with a statistician to verify that multiple imputation may be appropriately used to address missing data as well as how it can be implemented. In cases where partial information may be available such as the case when the outcome variable is measured at different time points and data are missing at some time points, then one may use models that accommodate partial information. Examples of these models are mixed models, as discussed previously, as well as time-to-event or survival models.

Conclusions

Preclinical studies are an essential step in the drug development process, as they provide the initial assessment for the effectiveness of a new therapy. Like any study, preclinical studies should be properly designed in order to answer the research question at hand. The statistical analysis of the study results should be planned concurrently with the study design as the two are inextricably linked; the two must be matched in order for the results to be useful.

One key design consideration is the number of subjects to use in the study. This number depends on the expected effect size, the variability of the outcome, and the nature of the outcome (continuous or categorical, bell-shaped or skewed). The nature of the outcome also infuences the statistical test used. Randomization and blinding should also be used in the allocation and administration, respectively, of the new treatment in order to reduce bias in the study and increase the validity of the results. If there are multiple outcomes or single outcome measured repeatedly over time, methods such as adjustment to the significance threshold or models for longitudinal data are necessary to control the false positive error rate and to maximize the information provided by the data. Finally, missing data should be described in reports and accounted for in the analyses.

Acknowledgments

Funding for I. Aban provided by NINDS 5U01NS42685-03 (Thymectomy in Non-Thymomatous MG Patients on Prednisone) and funding for B. George provided by NHLBI Training Grant T32HL079888.

References

Agresti, A. An Introduction to Categorical Data Analysis. 2. John Wiley and Sons; NJ: 2007.

Albert PS. Longitudinal Data Analysis (Repeated Measures) in Clinical Trials. Statistics in Medicine. 1999; 18:1707–1732. [PubMed: 10407239]

Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B. 1995; 57:289–300.

Goodman S. A dirty dozen: twelve p-value misconceptions. Seminars in Hematology. 2008; 45:135–140. [PubMed: 18582619]

Henderson VC, Kimmelman J, Fergusson D, Grimshaw JM, Hackam DG. Threats to validity in the design and conduct of preclinical efficacy studies: a systematic review of guidelines for in vivo animal experiments. PLoS Med. 2013; 10(7):e1001489.10.1371/journal.pmed.1001489 [PubMed: 23935460]

Hintze, J. PASS 11. NCSS, LLC; Kaysville, Utah, USA: 2011. www.ncss.com

Holm S. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics. 1979; 6:65–70.

Kilkenny C, Parsons N, Kadyszewski E, Festing MFW, Cuthill IC, et al. Survey of the Quality of Experimental Design, Statistical Analysis and Reporting of Research Using Animals. PLoS ONE. 2009; 4(11):e7824.10.1371/journal.pone.0007824 [PubMed: 19956596]

Landis SC, Amara SG, Asadullah K, et al. A call for transparent reporting to optimize the predictive value of preclinical research. Nature. 2012; 490(7419):187–191. [PubMed: 23060188]

Little, RJA.; Rubin, DB. Statistical analysis with missing data. 2. New York: Wiley & Sons; 2002.

Piddlesden SJ, Jiang S, Levin JL, et al. Soluble complement receptor 1 (sCR1) protects against experimental autoimmune myasthenia gravis. J Neuroimmunol. 1996; 71:173177.

Soltys J, Kusner L, Young A, Richmonds C, Hatala D, Gong B, Shanmugavel V, Kaminski H. Novel Complement Inhibitor Limits Severity of Experimentally Myasthenia Gravis. Ann Neurol. 2009; 65(1):6775.

Warner DS, James ML, Laskowitz DT, Wijdicks EF. Translational research in acute central nervous system injury: lessons learned and the future. JAMA Neurol (71). 2014 Oct 1.(10):1311–8. [PubMed: 25111291]

Zucker DM, Manor O, Gubman Y. Power comparison of summary measure, mixed model, and survival analysis methods for analysis of repeated-measures trials. Journal of Biopharmeceutical Statistics. 2012; 22:519–534.

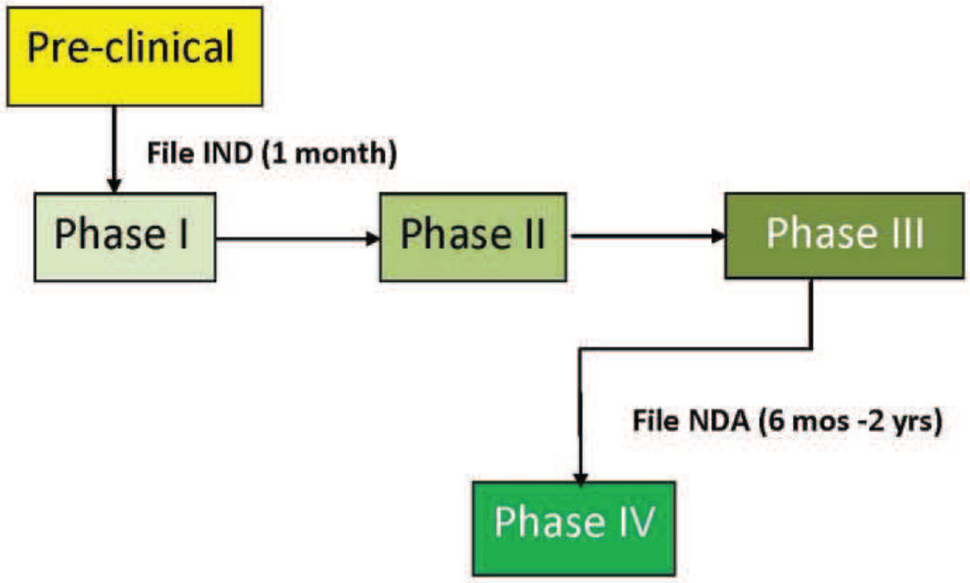

Figure 1.
Stages of Drug Development

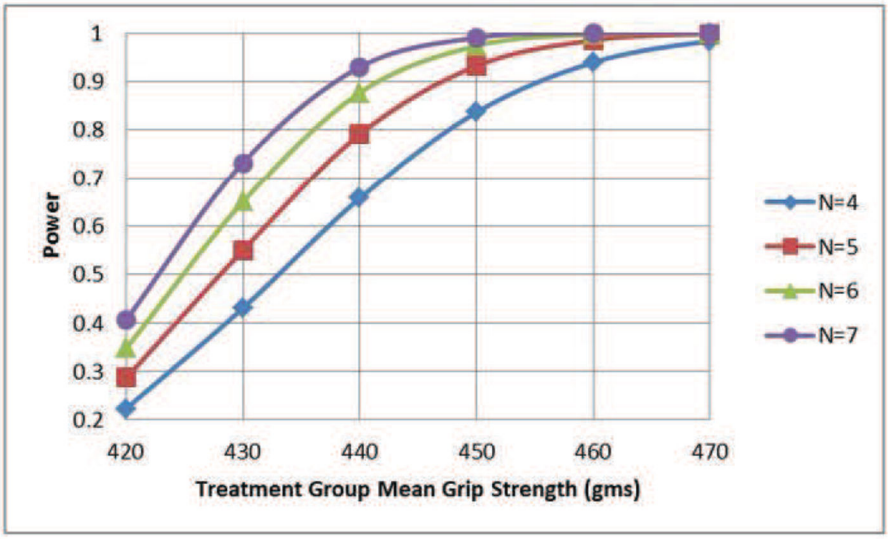

Figure 2.

Power Analysis for comparing two independent groups: two-tailed 5% significance level t-test assuming a common standard deviation of 20 grams and a mean of 400 grams for untreated group

Scheme	SUBJECT NUMBER																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
S1	T	T	T	T	T	T	T	T	T	T	C	C	C	C	C	C	C	C	C	C	NOT GOOD
S2	C	T	C	T	C	T	C	T	C	T	C	T	C	T	C	T	C	T	C	T	NOT GOOD
S3	T	T	T	C	T	C	C	T	T	T	T	C	C	C	T	C	C	C	C	T	GOOD

Figure 3.
Sample treatment assignment schemes

Table 1

Illustration of methods to adjust for multiple testing

p-value	Bonferroni	Bonferroni-Holm	Benjamini-Hoch
$p_{(1)} = 0.001$	0.00833	0.00833	0.0083
$p_{(2)} = 0.009$	0.00833	0.01	0.0167
$p_{(3)} = 0.01$	0.00833	0.0125	0.025
$p_{(4)} = 0.02$	0.00833	0.0167	0.0333
$p_{(5)} = 0.04$	0.00833	0.025	0.0417
$p_{(6)} = 0.34$	0.00833	0.05	0.05