Experimental Neurology 270 (2015) 82–87

Contents lists available at ScienceDirect

Experimental Neurology

journal homepage: www.elsevier.com/locate/yexnr

Statistical considerations for preclinical studies

Inmaculada B. Aban *, Brandon George

Department of Biostatistics, University of Alabama at Birmingham, AL, USA

ARTICLE INFO

Article history:
Received 25 November 2014
Revised 29 January 2015
Accepted 17 February 2015
Available online 26 February 2015

Keywords:
Preclinical studies
Sample size
Power
Randomization
Multiple outcomes
False positive
Missing data

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 relies 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.

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. Researchers 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 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. Fig. 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 does 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 Perrin, 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 rely 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

* Corresponding author at: University of Alabama at Birmingham, 1530 Third Avenue South, Birmingham, AL 35294-0022, USA.
E-mail address: caban@uab.edu (I.B. Aban).

http://dx.doi.org/10.1016/j.expneurol.2015.02.024
0014-4886/

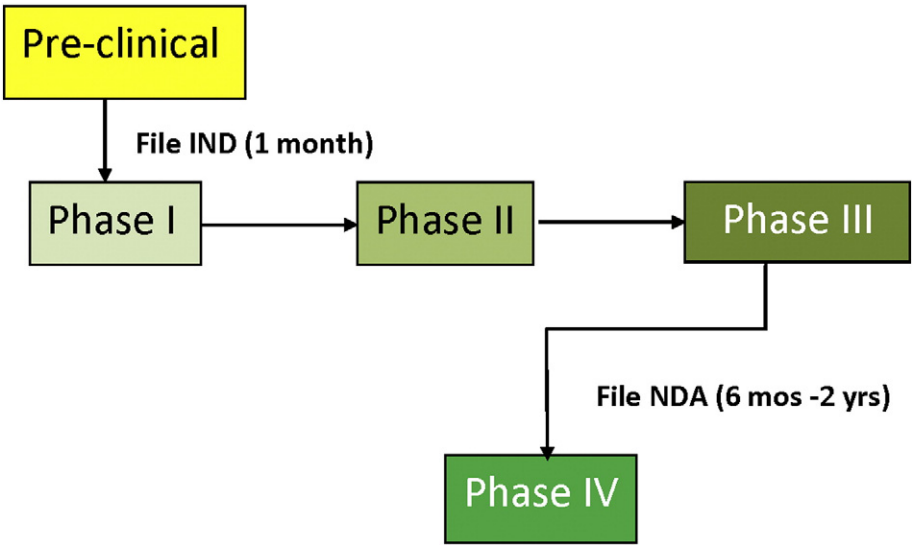

Fig. 1. Stages of drug development.

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 a 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 the 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 groups), 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 of 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 et al., 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 samples (e.g., untreated versus treated groups). For an MG study, one may be interested in the grip strength as the outcome. Assuming that 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

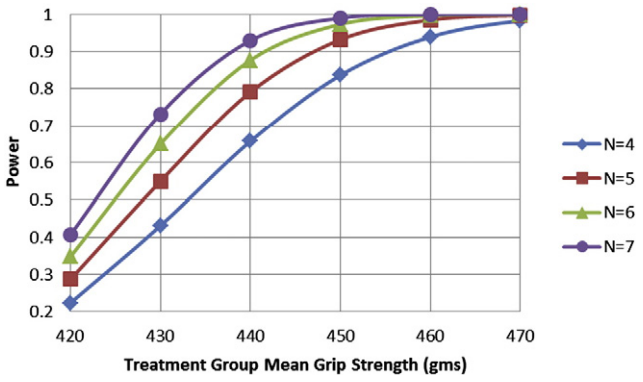

Fig. 2. Power analysis for comparing two independent groups: two-tailed 5% significance level t-test assuming a common standard deviation of 20 g and a mean of 400 g for untreated group.