Published in final edited form as:

Int J Obes (Lond). 2017 August ; 41(8): 1306–1309. doi:10.1038/ijo.2017.93.

A quantitative analysis of statistical power identifies obesity endpoints for improved *in vivo* preclinical study design

Jangir Selimkhanov[1], **W. Clayton Thompson**[1,†], **Juen Guo**[2], **Kevin D. Hall**[2], and **Cynthia J. Musante**[1,*]

[1]Internal Medicine Research Unit, Pfizer Inc., Cambridge, Massachusetts, United States of America

[2]Laboratory of Biological Modeling, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, Maryland, United States of America

Abstract

The design of well-powered *in vivo* preclinical studies is a key element in building knowledge of disease physiology for the purpose of identifying and effectively testing potential anti-obesity drug targets. However, as a result of the complexity of the obese phenotype, there is limited understanding of the variability within and between study animals of macroscopic endpoints such as food intake and body composition. This, combined with limitations inherent in the measurement of certain endpoints, presents challenges to study design that can have significant consequences for an anti-obesity program. Here, we analyze a large, longitudinal study of mouse food intake and body composition during diet perturbation to quantify the variability and interaction of key metabolic endpoints. To demonstrate how conclusions can change as a function of study size, we show that a simulated pre-clinical study properly powered for one endpoint may lead to false conclusions based on secondary endpoints. We then propose guidelines for endpoint selection and study size estimation under different conditions to facilitate proper power calculation for a more successful *in vivo* study design.

Introduction

Obesity is a growing epidemic which is associated with millions of annual deaths worldwide [1]. Moreover, the discovery and development of anti-obesity agents aimed to treat this disease is fraught with obstacles [2]. While a potential drug candidate may fail in various stages of development, *in vivo* preclinical evaluation of promising targets and pharmacotherapies is a first step to successful clinical development. However, animal testing has significant and underappreciated challenges for analysis. Conclusions based on

[*]Corresponding author: Cynthia J. Musante (CJM); Pfizer Inc., 1 Portland Street, Cambridge, MA USA 02139; Phone: (617) 551-3140; Fax: (212) 499-3950; cynthia.j.musante@pfizer.com.

[†]Current affiliation: SAS Institute, Cary, North Carolina, United States of America

Conflict of interest: JS, WCT, and CJM were employees of Pfizer Inc. during this study. JG and KDH declare no potential conflict of interest.

Supplementary information is available at International Journal of Obesity’s website incorrectly interpreted exploratory preclinical study results may impede drug development progress into the clinic. A contributing factor to this problem is the size of many exploratory studies, which are often conducted on a small number of animals and examine a limited set of endpoints that can be difficult to measure precisely. Study designs based upon improper estimates of treatment effect size and variance may not be properly powered, and lead to significant challenges in the interpretation of results [3].

Statistical power can be defined as the probability that a test will detect an effect if the effect actually exists. For a given study design, the power associated with a statistical test (e.g., t-test for the body weight difference between an intervention group and a control group) is a function of the expected effect size, variance, and sample size. Typically, a sample size is chosen so that a desired degree of statistical power is obtained for a given effect size and variance. However, the magnitudes of the effect size and variance are rarely known *a priori*, and accurate estimation of these can be particularly difficult for systems characterized by significant feedback and interactions between multiple measured endpoints. Thus, while power calculations have been previously used to justify specific data analyses [4], a broader assessment of endpoints of interest for obesity pharmacotherapy (e.g., food intake, body composition) has yet to be undertaken. By understanding the physiological interdependence of different endpoints one can achieve a more accurate estimation of statistical power and improve the ability to correctly interpret study results.

Common macroscopic endpoints that are often used to inform preclinical anti-obesity programs include food intake (FI), body weight (BW), energy expenditure (EE), fat mass (FM), and fat-free mass (FFM) (see, e.g., [5–7]). Guo and Hall have developed and validated a mathematical model that quantitatively describes the physiological relationships between these endpoints [8, 9]. Combining this model with previously collected individual mouse FI, BW, and FM data, we use the resulting statistical model to show a cautionary example based on a simulated drug treatment study. After illustrating the difficulty of drawing correct conclusions from this improperly powered study, we propose endpoint measurements that can improve study design and lead to more accurate result interpretation.

Methods and Results

The general method of constructing and using a statistical model to help power a study is shown in Figure 1A. In the initial step, we generated a statistical model by fitting a mathematical model of energy balance that described mean endpoint behavior to the individual BW, FM, and FI mouse data from [8, 9] (SI Section 3). The resulting statistical model captured endpoint effect size and variance observed in the data from [8] and allowed us to estimate the contribution of inter- and intra- animal variability to each of the endpoints' variance (Figure 1B). We defined inter-animal variability as the physiological differences between animals, which do not change on a day-to-day time scale. Intra-animal variability was defined as a combination of inherent day-to-day fluctuations in certain endpoints (e.g., FI) within an individual animal as well as error associated with the measurement itself. Our statistical model indicates that intra-animal variability contributes more (chow - 77% and high-fat diet, HFD - 86%) to the variance of FI than to BW or FM, while variances of BW and FM are primarily driven by inter-animal variability (BW: chow - 86% and HFD - 92%; FM: chow - 93% and HFD - 76%).

Specifying study variables (e.g., population, treatment, size, endpoints), we are able to use the statistical model to simulate virtual animals/colonies in order to estimate endpoint effect sizes and variances (Figure 1A). To illustrate this capability, we simulated a typical 2-week drug-treatment study consisting of placebo and treatment arms, with the treatment effect assumed to reduce FI by an average of 15% (SI Section 3.5.2). We then calculated effect sizes and variances for a set of typical preclinical endpoints: changes from baseline (Δ) in BW, FM and FFM, single-day FI measurement (last day of the study), and the cumulative FI over the length of the study (SI Section 3.5.3). Using model-predicted effect sizes and variances, we calculated the number of animals (6 per group) required to properly power our simulated study for ΔBW (Figure 2A). This is the lowest number of animals that corresponds to statistical power greater than 80% and $\alpha = 0.05$, which is commonly used to determine statistical significance in exploratory preclinical studies [5–8, 10–12]. In SI Section 4.2, we adjusted for multiple testing and animal/data loss, which increases the sample size to 10. Considering other endpoints, we found that with 6 (or 10) animals per group, power calculated for ΔFM, ΔFFM, and single-day FI measurements, unlike cumulative FI, falls below the 80% threshold level (Figure 2B, SI Figure S10B). All model simulations and data fitting were performed using a commercial software package (MATLAB 2014B, MathWorks Inc., Natick, MA 2014); the associated simulation code is available in the SI.

Discussion

To illustrate the importance of powering for each endpoint of interest, we consider a common exploratory preclinical scenario based on our simulated study of a hypothetical anorectic agent. Given that ΔBW is often a primary or secondary endpoint, we can correctly power the study to detect ΔBW while reducing the number of animals required per arm. The sample size estimated based on ΔBW=−5% (a common criteria for minimal clinically significant weight loss [13]) matches the number of animals per group commonly used in preclinical studies [5–8, 10–12]. Let us assume that we observe a statistically significant difference between placebo ΔBW and treatment ΔBW at the end of the study. To better understand the mechanism driving ΔBW, we may then want to determine whether FI is a contributor to the ΔBW. As a cautionary example, we decide to collect a single-day FI measurement, which is a common approach (see, e.g., [5, 10, 11]). Our results show that, even in our simulated study where FI was the sole driver of ΔBW, we are underpowered to detect the underlying FI difference using a single-day FI measurement under these conditions (i.e., with a study sample size chosen based on expected ΔBW difference between groups). Therefore, we are unlikely to detect a statistically significant FI difference between placebo and treatment groups even when a difference exists. Given that ΔBW may be due to changes in FI or EE, one may then inaccurately surmise that the treatment has an effect on EE. This conclusion may lead to the shift in the focus of an anti-obesity program towards a potential EE mechanism of action, when, in fact, this mechanism does not exist. Similarly, if we attempt to use body composition measurements to estimate EE as described in [9], we are unlikely to find significant difference between groups since the study is underpowered to detect a difference in ΔFM and ΔFFM. This would further reduce the value of this hypothetical study and require additional resources committed to finding the mechanism of action for ΔBW. However, we can avoid these Type II (false-negative) errors associated with statistical power with proper study sample size calculation based on multiple endpoint testing and accounting for potential animal/data loss (N=25, SI Section 4.2).

To effectively combat misinterpretation of data due to study power, we can use model-predicted endpoint effect size, variance, and source of variability to identify optimal measurements *a priori* for a particular study. For example, since FI variance is primarily driven by day-to-day variability and measurement error (Figure 1B), the use of paired (baseline vs. final value) single-day FI measurements does not offer significant reduction in sample size compared to other endpoints (SI Figure S12). This is due in part to the previously noted technical difficulty of detecting small differences in daily FI, which can result in significant ΔBW [12]. However, cumulative FI has much lower variance and, if measured, can reduce the number of animals required to power a study to detect the differences in FI (Figure 2B, SI Section 4.3 and Figure S10). We can further extend this approach beyond FI in our hypothetical study, with the model-based simulations providing the basis for evaluation of more complex endpoints, study designs, and statistical tests (e.g., multiple study arms, functional data analyses), to increase the probability of study success (Figure 1A) [14,15].

The ultimate goal of using a statistical model such as the one presented here is to improve *in vivo* preclinical study design and to reduce misinterpretation of study results, which will enhance the likelihood of program success. Conversely, ignoring statistical power of a given endpoint, and making a scientific claim based solely on its statistical significance is likely to lead to a false conclusion. As a warning, we used a hypothetical study to illustrate potential problems that can arise when statistical power is ignored in an exploratory preclinical study design setting. Analysis of previous mouse data indicates that BW and cumulative FI offer more statistical power than FM, FFM, or single-day FI. For study designs and endpoints outside the scope of our model (SI Section 4.5), the key to success is careful evaluation of each endpoint of interest and its statistical power. Furthermore, provided alternative ways of measuring a given response (e.g. cumulative vs. single-day FI), it is important to select an appropriate endpoint measurement that improves its statistical power.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding: Pfizer Inc. supported the research of JS, WCT, and CJM. Data used in this research was from a study sponsored by the Intramural Research Program of the NIH, NIDDK.

Data used in this research was from a study sponsored by the Intramural Research Program of the NIH, NIDDK.

References

1. Finucane MM, Stevens GA, Cowan MJ, Danaei G, Lin JK, Paciorek CJ, et al. National, regional, and global trends in body-mass index since 1980: systematic analysis of health examination surveys and epidemiological studies with 960 country-years and 9·1 million participants. The Lancet. 2011 Feb; 377(9765):557–67.
2. Dietrich MO, Horvath TL. Limitations in anti-obesity drug development: the critical role of hunger-promoting neurons. Nature Reviews Drug Discovery. 2012 Sep 1; 11(9):675–91. [PubMed: 22858652]
3. Tsang R, Colley L, Lynd LD. Inadequate statistical power to detect clinically significant differences in adverse event rates in randomized controlled trials. Journal of clinical epidemiology. 2009 Jun 30; 62(6):609–16. [PubMed: 19013761]
4. Ravussin Y, Gutman R, LeDuc CA, Leibel RL. Estimating energy expenditure in mice using an energy balance technique. International journal of obesity. 2013 Mar 1; 37(3):399–403. [PubMed: 22751256]
5. Morton GJ, Thatcher BS, Reidelberger RD, Ogimoto K, Wolden-Hanson T, Baskin DG, Schwartz MW, Blevins JE. Peripheral oxytocin suppresses food intake and causes weight loss in diet-induced obese rats. American Journal of Physiology-Endocrinology and Metabolism. 2012 Jan 1; 302(1):E134–44. [PubMed: 22008455]
6. Wagner JD, Zhang L, Kavanagh K, Ward GM, Chin JE, Hadcock JR, Auerbach BJ, Harwood HJ. A selective cannabinoid-1 receptor antagonist, PF-95453, reduces body weight and body fat to a greater extent than pair-fed controls in obese monkeys. Journal of Pharmacology and Experimental Therapeutics. 2010 Oct 1; 335(1):103–13. [PubMed: 20605903]
7. Ravussin Y, LeDuc CA, Watanabe K, Mueller BR, Skowronski A, Rosenbaum M, Leibel RL. Effects of chronic leptin infusion on subsequent body weight and composition in mice: Can body weight set point be reset? Molecular metabolism. 2014 Jul 31; 3(4):432–40. [PubMed: 24944902]
8. Guo J, Hall KD. Estimating the continuous-time dynamics of energy and fat metabolism in mice. PLoS Comput Biol. 2009 Sep 1.5(9):e1000511. [PubMed: 19763167]
9. Guo J, Hall KD. Predicting changes of body weight, body fat, energy expenditure and metabolic fuel selection in C57BL/6 mice. PLoS One. 2011 Jan 5.6(1):e15961. [PubMed: 21246038]
10. Lin B, Koibuchi N, Hasegawa Y, Sueta D, Toyama K, Uekawa K, et al. Glycemic control with empagliflozin, a novel selective SGLT2 inhibitor, ameliorates cardiovascular injury and cognitive dysfunction in obese and type 2 diabetic mice. Cardiovasc Diabetol. 2014 Oct 26; 13(1):2215–5.
11. Doyon C, Denis RG, Baraboi ED, Samson P, Lalonde J, Deshaies Y, et al. Effects of Rimonabant (SR141716) on Fasting-Induced Hypothalamic-Pituitary-Adrenal Axis and Neuronal Activation in Lean and Obese Zucker Rats. Diabetes. 2006 Nov 27; 55(12):3403–10. [PubMed: 17130486]
12. Guo J, Jou W, Gavrilova O, Hall KD. Persistent Diet-Induced Obesity in Male C57BL/6 Mice Resulting from Temporary Obesigenic Diets. PLoS ONE. 2009 Apr 29; 4(4):e5370–9. [PubMed: 19401758]
13. Stevens J, Truesdale KP, McClain JE, Cai J. The definition of weight maintenance. Int J Obes. 2006; 30:391–9.
14. Wang JL, Chiou JM, Mueller HG. Review of functional data analysis. Annual Review of Statistics and its Application. 2015 Jun.3:257–295.
15. Senn, SS. Statistical issues in drug development. John Wiley & Sons; 2008 Feb 28.

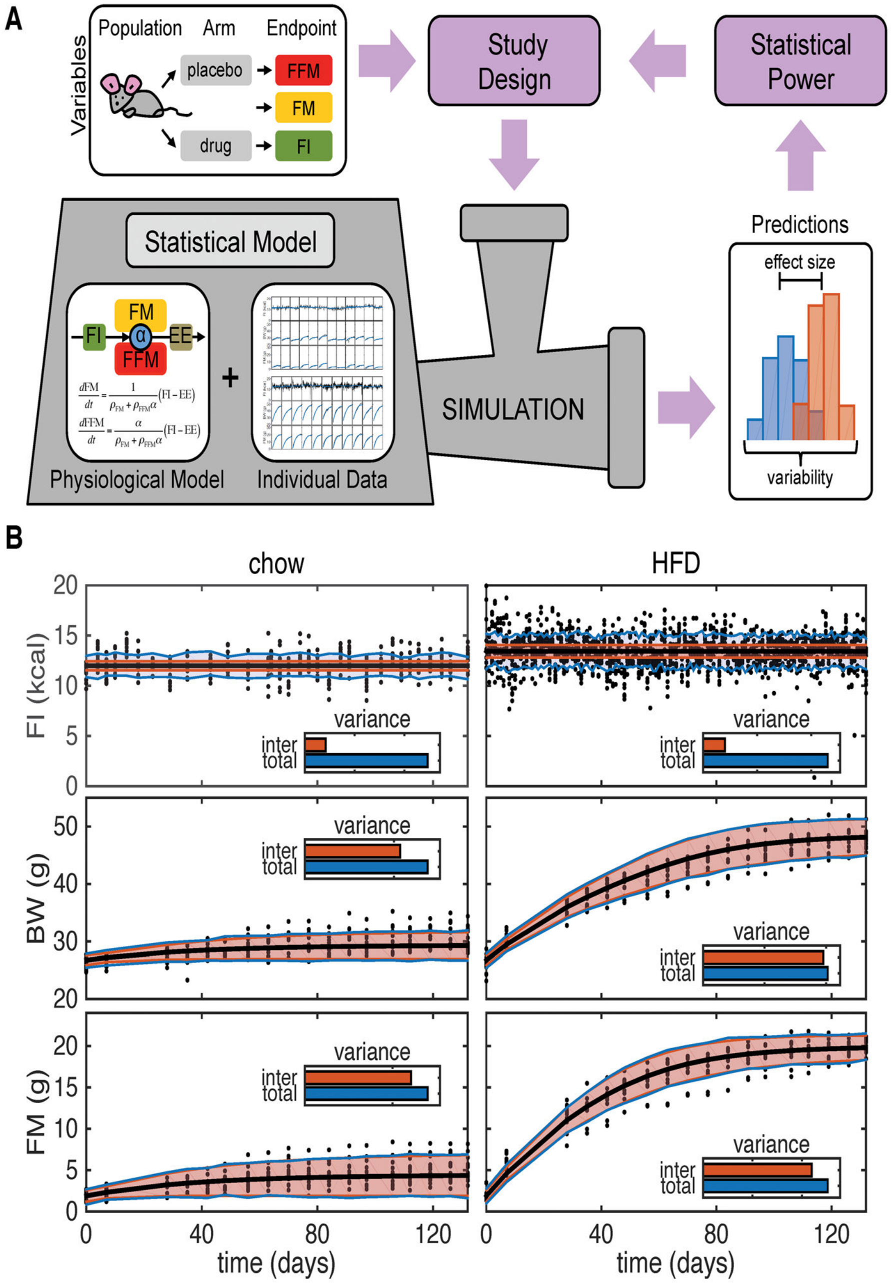

Figure 1. Statistical model identifies sources of endpoint variance and helps estimate statistical power through model simulation

(A) The general method of using a mathematical model to help power a study. Statistical model based on a physiological model fit to individual data is used to simulate study design to estimate endpoint statistical power, which can then inform the study design. (B) Model fit to mean FI shows that most of FI variability (blue) comes from day-to-day intra-animal variability (red). In contrast, BW and FM variability arises mostly from inter-animal variability. Shaded regions show +/- standard deviation around the mean model trajectory, while the inserts show the total variance (blue) and variance derived from inter-animal variability (red). Distributions generated from 1000 simulations.

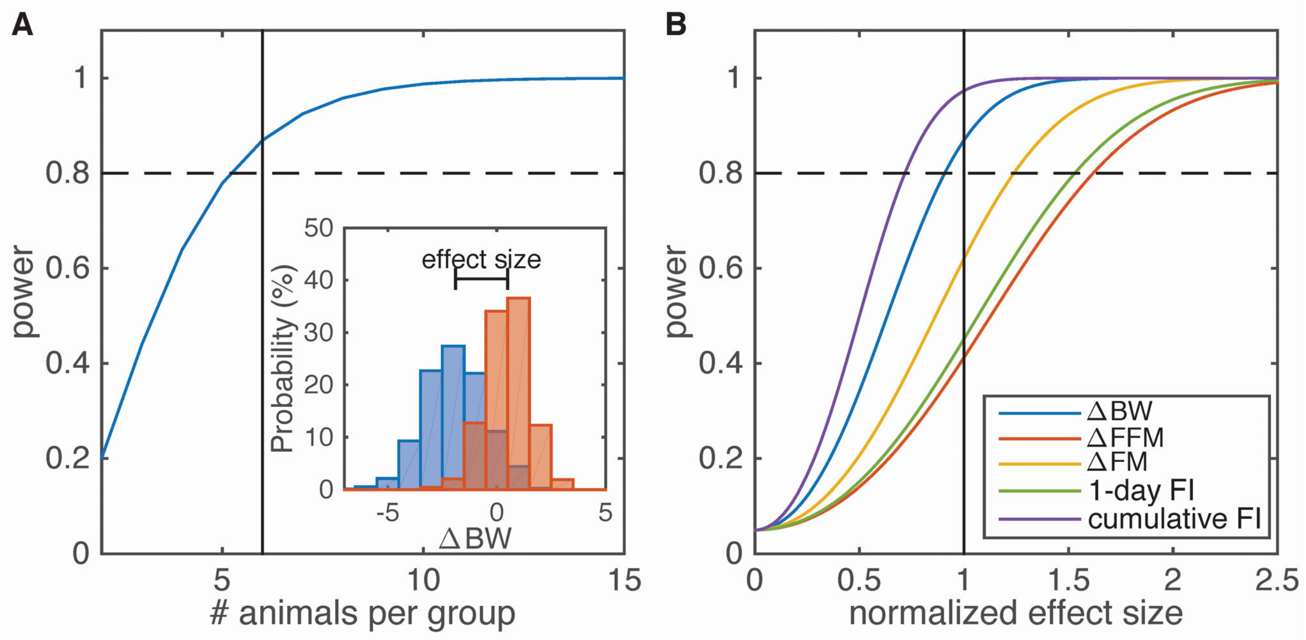

Figure 2. Study powered for body weight is underpowered for fat mass, fat-free mass, and single-day, but not cumulative food intake

(A) Statistical power calculation (α = 0.05) shows that a minimum of six animals per group (N = 6) are required to achieve power that surpasses 80% threshold (dashed line), based on model-predicted ΔBW effect size between treated (blue) and untreated (red) groups (inset). (B) With N = 6, the predicted effect size (normalized to 1, solid vertical black line) for ΔFFM (red), ΔFM (yellow), and single-day FI (green), unlike cumulative FI (purple), does not reach 80% threshold. Power calculated from 1000 simulations per group.