Guidance on NIMH Grant Application Power Calculations

Executive Committee of the Mental Health Statistics (MHS) Section of the American Statistical Association (ASA), on behalf of the American Statistical Association

Adequately powering a study is a crucial part of experimental design, necessary - but not sufficient - to ensure scientific validity and reproducibility. It is well documented that under-powered studies are prone to both false negatives and false positives, impeding scientific progress, wasting resources, and raising ethical concerns for both human and animal research. Thus, obtaining guidance on best practices regarding sample size and power calculations is of central concern to mental health researchers and to the NIMH. The paragraphs below provide a brief outline of recommended best practices from the Executive Committee of the Mental Health Statistics (MHS) Section of the American Statistical Association (ASA). The document at this page may also be useful to refer grant applicants to, as it provides an overview of common statistical issues seen in grant proposals: http://ww2.amstat.org/misc/StatisticalIssuesInProposals.pdf.

An important issue with power calculations in NIMH proposals revolves around the extremely common practice of biomedical researchers choosing a sample size for a given grant application based on budget or feasibility considerations and then providing *post hoc* justification through power calculations that exaggerate the probability of discovering effects with the desired Type I error rate control. This results in a quandary for the statistical analyst, in that proposing an adequate sample size for the requisite power is met with resistance due to the (often accurate) belief of the applicant that available or potential resources, monetary and otherwise, will not be sufficient for the suggested target sample. This gives rise to two immediate consequences: 1) grants are submitted with overly-optimistic power calculations, thus leaving reviewers in a state of uncertainty as to the accuracy and rigor of this section in general, and 2) many funded grants that claim adequate power are not in fact adequately powered.

The solution to this conundrum is not easy, and will take concerted action by investigators and statisticians in enforcing rigorous power calculations, changes in how and which grants are funded by the NIMH and other funding agencies, and more selectivity by journals publishing results from small studies. Here, we address the aspect of increasing rigor in power calculations, with attention paid also to the challenges of power analyses and ensuring adequate power, and thus need for innovation in both of those areas.

The precise methods used for power calculations must perforce be as diverse as the number of study designs: clinical trials of many types, observational studies with one assessment time, prospective single or multi-cohort longitudinal designs, retrospective studies of institutional records, whole-genome genotyping association studies, and so forth, have vastly different designs, sample size considerations, and types of measures collected. Moreover, power calculations depend not just on study design, but also on study hypotheses and statistical analyses proposed. Thus, it is very difficult to formulate one or even a small number of power calculation methodologies that would fit all research study proposals. In addition, new and

innovative study designs may require changing how we even think about statistical power – e.g., in some study designs (such as non-experimental studies), reduction in bias may be of more import than power *per se*. We posit that further methodological work is needed in developing appropriate power calculation methods for a broad array of study designs.

Nevertheless, we believe there are certain principles that all power calculations should fulfill to be useful for evaluating the merit of a proposed research study. Below we provide a brief outline of these principles:

- Power should be computed for a range of effect sizes reflecting plausible prior evidence from similar studies using similar research designs. In this regard, it is not appropriate to choose a large effect size from one published study (which itself may be underpowered) when other studies show a range of effect sizes. Since effect sizes in the literature are known to be inflated, it is better to be conservative rather than liberal in this respect.
- If it is not possible to obtain effect sizes from prior studies, study power should be computed for a range of effect sizes including the prospect of small effects.
- If there are other parameters required for the power calculations (e.g. loss-to-follow-up rate, intervention compliance, etc.) that can be varied and which impact power, these should also be varied over a range of plausible values. In this respect, it is again desirable to choose values of the parameters from prior evidence if possible.
- The method used for power calculations should be explicitly stated and justified based on the study design and the analysis plan. It is not sufficient to merely cite the statistical software package used.
- Power calculations should reflect multiple testing adjustments necessary to control Type I error rates at the desired level.
- The definition of "effect size" utilized in the power calculations should be relevant to the proposed study, i.e., it should be based on the actual planned statistical analyses testing the study hypotheses.
- Power analysis deserves the same care and thought as the remainder of the study design and statistical analysis, and should be well-integrated with these. Therefore, it is highly recommended that a person with the requisite statistical expertise plans these calculations.
- It is appropriate that different criteria are used for judging the level of evidence required for power calculations depending on the type of study, e.g., confirmatory, exploratory, and pilot/feasibility or K awards.
- Power analyses can be quite complex and may depend on unknown high-dimensional parameters. This often necessitates simplification of power calculation methodology compared to the proposed analyses. We posit that further methodological work is needed in developing appropriate power calculation methods for many modern study designs and statistical analysis plans.

The issue of adequately powering studies to increase validity and replicability of results is important and must be addressed by all stakeholders in a concerted fashion, including funding

agencies, journals, statisticians, and investigators. Increasing the rigor and transparency of power calculations in grant applications, as recommended here, is an important step in this process. We also note that power analyses are complex and there are many nuances. We would welcome the opportunity to be further engaged in the discussions moving forward as we all work to ensure that studies funded by NIMH are done with as much rigor as possible, and lead to improvements in the public's mental health.

Wesley K. Thompson
Associate Professor
Department of Family Medicine and Public Health
Division of Biostatistics
University of California, San Diego
Council of Sections Representative,
Mental Health Statistics Section
American Statistical Association
wes.stat@gmail.com

Elizabeth A. Stuart
Associate Dean for Education
Professor
Department of Mental Health
Department of Biostatistics
Department of Health Policy and Management
Johns Hopkins Bloomberg School of Public Health
Section Chair,
Mental Health Statistics Section
American Statistical Association
estuart@jhu.edu