

2021

Using Adaptive Research Design to Define the Proper Methodology to Use a Data Peek for Power: Step by Step Process

Tom Wasser
statbiz1@aol.com

Follow this and additional works at: <https://scholar.rochesterregional.org/advances>



Part of the [Other Applied Mathematics Commons](#), and the [Translational Medical Research Commons](#)



This work is licensed under a [Creative Commons Attribution-NonCommercial 4.0 International License](#)

Recommended Citation

Wasser T. Using Adaptive Research Design to Define the Proper Methodology to Use a Data Peek for Power: Step by Step Process. *Advances in Clinical Medical Research and Healthcare Delivery*. 2021; 1(3). doi: 10.53785/2769-2779.1035.

ISSN: 2769-2779

This Article is brought to you for free and open access by RocScholar. It has been accepted for inclusion in *Advances in Clinical Medical Research and Healthcare Delivery* by an authorized editor of RocScholar. For more information, please contact Advances@rochesterregional.org.

Using Adaptive Research Design to Define the Proper Methodology to Use a Data Peek for Power: Step by Step Process

Abstract

When planning or conducting research in the hospital setting, often termed Real-World Environment (RWE), therapeutic assumptions and outcomes are often different than in the Randomized Clinical Trial (RCT) where medications, devices and therapies are tested and developed. This is because RWE research has a lack of experimental control, additional confounding due to patient complications and comorbid conditions, lack of pure patient selection and compliance with therapy in the patients being treated and many other factors as well. However, when RWE experiments are conducted, sample size determination using data from the RCT is common because that is the only data that is available when the RWE research is being developed. Using RCT data to derive sample size calculations within the RWE hospital or outpatient setting, on real patients with vastly different conditions has the potential to give inaccurate results. Using newly developed Adaptive Research Designs[[chow](#)], which allow for the individual study's own data for sample size determination is a viable and highly accurate method to prevent under or over sampling in the RWE research context. This paper outlines the proper methodology to use to conduct a "Data-Peek for Power" which is a within RWE, "Adaptive" methodology to calculate sample size without risking reductions in p-values, termed 'alpha-spend'. Using a Data-Peek for Power is a method that allows for no alpha spend, free from multiple comparison, assessment of statistical power or sample size calculation. When needed it can easily be implemented and described in a research protocol or a proposal that is submitted to the IRB for review listing all relevant variables to be used with the data analysis methods a-priori.

Keywords

Power Analysis, Sample Size Calculations, Biostatistics, Research design, Interim Analysis

Creative Commons License



This work is licensed under a [Creative Commons Attribution-NonCommercial 4.0 International License](https://creativecommons.org/licenses/by-nc/4.0/)

Conflict of Interest Statement

No conflicts of interest related to this article to disclose.

Using Adaptive Research Design to Define the Proper Methodology to Use a Data Peek for Power: Step by Step Process

Introduction

Research on outcomes of particular treatments, whether hospital based, or pharmacologic/device studies, at the Phase III and IV level are often conducted to examine effectiveness of treatments as compared to standard of care. These studies can be performed to determine therapeutic or costs differences, efficacy over time, patient compliance, longitudinal treatment changes and many other reasons. Late phase studies can also be performed to demonstrate that two therapies are clinically equivalent, but that one therapy costs less or is less traumatic to patients when used in practice. As a result, statistics used in these studies can include tests for differences, non-inferiority or equivalence.^{1,2} To conduct these studies, statisticians have an arsenal of approaches to use, and standard significance levels of 0.05 are seldom applicable in application.

For example, a recent study examining two different types of ventilation used in an ICU. A study was performed to determine if a less expensive ventilator performed as well as a more expensive one which required more staff involvement and maintenance with increased cost. In this study patient outcomes were examined for equivalence with a conservative p-value of $p < 0.20$ to catch any indication of patient outcome differences. After the study it was determined that the patient outcomes were not statistically different ($p > 0.20$) and the less expensive ventilator was used. In a second example two types of forearm fracture fixation were examined (internal versus external) in a pediatric Randomized Clinical Trial (RCT). In this study, the researchers used a difference in treatment design, without specifying which method was thought to work better (2-tailed hypothesis), with a very strict p-value of 0.01 required for significance and powered the study accordingly, only willing to consider one method of fixation superior to the other if there was over-whelming effectiveness with a p-value less than 0.01. At the end of the study neither method was found to be convincingly more effective.

Often, when designing or performing this research, the exact effect of therapy and the variation of the treatments are not known. In these studies, experimentation is often done in the Real-World Environment (RWE): within hospitals, with a convenience sample of patients that are available during the time in which the researcher (often a resident physician) has access to them. These studies are far outside of the pure clinical trial setting where conditions are well controlled. Patient behavior and characteristics of RWE studies are far less controlled than the Randomized Clinical Trial (RCT) at Phases I, II and III (Table 1). As a result, the treatment effects and other statistics of interest that are published internally and externally while pharmaceutical products are being developed in the RCT setting, are not directly transferable to the RWE experimental context. The US. Food and Drug Administration understands that sample size recalculation during the research may be needed in order to properly complete a study³. Often

times in an early Phase (1) studies it is common for the FDA to ask a pharmaceutical company to add a safety group, or variable to a pharmacokinetic study which would require more subjects to fill a group or new stratification. Or on the other hand, researchers might feel during a feasibility study that a study can not be conducted as designed and decide to terminate the study with less than the number of patients thought to be needed from the original study plan. In this case, completion of data collection for those patients that are currently enrolled or completed the study would be used for analysis, and the study terminated so as to not place further patients at risk and to provide alternative treatments.

Table 1: Relevant Differences Between RCT and RWE (Pragmatic) research.

Randomized Clinical Trials (Phase I, II and III)	Real World Environment (Hospital based)
<ul style="list-style-type: none"> • Studies are in the pre-FDA approval stage and are sample based. • Patients are randomized to study groups (treatment or placebo arms). • Well controlled (lab setting) examining patients with few or no other pre-existing conditions. • Patients either have no disease (Phase 1 safety), or only the disease of interest (Phase II and III). • Data on effect size and variation are restricted to the controlled experiment and have little value outside of the clinical trial. • Studies examine efficacy. 	<ul style="list-style-type: none"> • Studies are post-FDA approval and are population based. • Typically, not randomized, but if they are they are randomized into treatment and existing treatment groups. • No lab setting, patients are in the real world and may have other illnesses or comorbid conditions. • Patients may be on multiple medications to treat their other disease. • Existing data on the RWE patients regarding effect size and variation are often not known. • Studies examine the relative effectiveness.

RWE and/or hospital-based research still does maintain some of the common needs that earlier phase studies do, especially when this research is prospective^{1-2,4}. Concerns about sample size, safety of participants are still relevant.⁵ The challenges for this type of research are finding good and reliable estimates to use for sample size calculations and study power ($1-\beta$ which is the probability of correctly rejecting the null hypothesis that a treatment effect exists).⁶ One solution which is outside of a formal interim analysis is referred to as a “Data-Peek for power”. This ‘Data-Peek

for power’ is an examination of actual study data that is collected during a study to calculate the statistics needed in order to determine what the correct sample size would be. This is an Adaptive Design approach mentioned by Rosenberg.⁶ By performing this analysis during the RWE setting, the data are then accurate to the actual experiment being conducted which avoids the problems with using published reports for research that does not fit the RWE design. A Data-Peek for Power also avoids issues related to bias because it does not require p-value adjustment (alpha spend) like other approaches would.^{7,8}

Requirements for sample size determination

Effect Size – is typically a clinically determined value. The statistician seldom identifies the effect size for the comparisons to be made. In practice the effect size is determined clinically. These concepts are defined in Table 2.

Table 2: Requirements for sample size determination

Effect Size	A clinical value, and is a translation of what clinical effect of the variables would be clinically significant with regard to the experiment. In other words, identify the difference in the variables that would be clinically important to the researcher or patient.
Variation	Requires the knowledge of the statistical analysis that is to be performed. Whether using the means or proportions, etc., the data are analyzed in order for the determination of the variation (usually either the standard deviation or variance for the variables of interest).
Individual participant enrollment or cluster samples	The researcher needs to examine how participants are being enrolled into the research. Individual patients being enrolled from no specific organized practices (ie., walk-ins to the ER) are not clustered however participants enrolled from a subset of available sources (cooperating medical practices) may represent cluster sampling and the calculation of the ‘intra-cluster correlation coefficient’ (ICC) may be required. Cluster sampling increases the sample size needed as compared to individual participant enrollment.
Statistical Assumptions for p-value and required Power	Typically, standard values are used. A p-value equal to 0.05 for significant results and power ($1-\beta$) of either 80% or 90%. Often times, calculations of power include both 80% and 90% for power.

Process for a Data-Peek for Power

Step 1 – *Determine the statistic that you are going to use as a part of your analysis.*

Be aware that if you are fortunate enough to find some values already published in the peer reviewed literature that these reported statistics must agree with the data analysis that you have planned as a part of your study. For example, if a study reports an incidence rate of a particular disease over patient years, yet your analysis is looking at the proportion of patients with the same disease as compared to another proportion your statistic is a comparisons of proportions and not of rates as was recorded in the article you reviewed. As a result, you might need to convert the data from the article into a proportion if that is even possible.

Step 2 – *Identify the power (sample size) formula that is appropriate to your particular statistic.*

Generally, every statistical test has a power formula that has been developed for it, or there is a way to construct a formula based on either existing formulas or the distribution that the statistic will use. There are several statistics for which power formulas do not exist but almost all of the more common statistical procedures typically used in the RWE have formulas already constructed. An example of a very basic power calculation is provided here, using the information required from Table 2: A researcher reads an article describing weight loss in a study group which he thinks might be less than he observes in his own patient population. He wishes to know how many patients he would need to study to verify this but he only has 10 patients and wants to know how many more he would need. In the article the weight loss is 17 pounds with a standard deviation of 5 pounds ($5 = \text{variation } \sigma^2$ Table 2). In his practice the 10 patients lost on average 20 pounds ($20 - 17 = 3$ which is the effect size Table 2). The researcher finds an online calculator and enters the values using the formula in Table 3 and determines that he would need 22 patients to determine that his practice patients losing 20 pounds on average would be statistically different than the article mean value of 17 pounds lost. It should be noted that there are many online calculators available to researchers to perform these exact calculations however caution should be used to make sure that the calculator used has the same formulas as the statistics that will be used for the actual study.

Table 3: Basic Sample size calculation for 80% power. One group versus population value.

$$N = \frac{\sigma^2(z_{1-\beta} + z_{1-\alpha/2})^2}{(\mu_0 - \mu_1)^2}$$

$$N = \frac{5^2(0.84 + 1.96)^2}{(17 - 20)^2}$$

$$N = 22$$

Where: μ_0 = population mean, μ_1 = mean of study population, N = sample size of study population σ = standard deviation of study population, α = probability of type I error (usually 0.05), β = probability of type II error (usually 0.2), z = critical Z value for a given α or β .

As a part of this determination of an appropriate power formula the researchers also need to determine if the participants being selected are being selected from clusters. In other words if a physician has hundreds of practices from which to abstract a sample but decides that 5-10 practices will suffice because he believes they are more accessible, have duplicate staffing, have staff that are familiar with the clinical or pragmatic trial forms or paperwork, have proven themselves to be compliant with previous research or any other reason, the rules and requirements of cluster sampling might be applicable. Donner and Klar¹² have developed alternative formulas for this cluster sample scenario. Many of these sample size calculation situations require the calculation of the ‘intra-class correlation coefficient’ or ‘intra-cluster correlation coefficient’ (ICC). A basic definition of the ICC is the amount of variation within the outcome variable (dependent variable) that is explained by how the patients are placed into groups. Meaning some groups of patients may outperform or underperform other groups within the treatment cohorts. This similarity of treatment typically would have the effect of increasing the sample size when the ICC is high (close to 1.0). If, however each study participant was to be selected from individual practice, and no other participants from that practice were selected then the ICC would equal zero and typical sample size formulas could be used.

Step 3 – Determine the Effect Size, the Variation and the Statistical Assumptions needed for the Data-Peek in the form that is needed by the formula. Make sure that the values that are being used are directly applicable to the power formula that you are going to use. Some formulas may call for standard errors rather than standard deviations and you need to make sure that the values you are using are those that match the formula.

The effect size is generally not determined by the statistician but rather by the clinician. A simple perspective of this value is: What difference does the researcher think would be important clinically? For example: For an operative procedure, how much shorter would the length of stay need to be for the treatment to be considered worthwhile? What percentage of patients not having an adverse event would be an improvement over an existing adverse event rate? Quantifying all of these questions into mean differences, or percent differences are examples of effect size.

Step 4 – *Construct a matrix of possible sample size values based upon the Data-Peek for Power results.* This matrix will have the measure of variation used (either the standard deviation or the standard error) on one axis and the estimates of effect size on the other. Each of these variables will consist of the actual value derived from the RWE data as well as several other estimates both above and below these values. This method then creates a possible range of values to use in sample size estimation.

Step 5 - *Select a sample size goal from the table.* This table calculated in Step 4 can then be used to select a sample size that the researcher thinks will best represent the study and patients that might be available to him/her during the course of their experiment.

Example:

A researcher is attempting to determine if a new therapy for insulin delivery is more effective or less effective than a standard injection-based insulin delivery method in newly diagnosed elderly diabetics with other comorbid conditions. In this example both methods of insulin delivery are FDA approved but have never been studied in an elderly population that is already suffering with many other disease types so there is no applicable data for a proper power calculation.

The researcher elects to use a Difference in Difference design and collects the data on the first 30 subjects that were enrolled into each insulin delivery method. The researcher observes a difference between groups of 0.26 (effect size). And a calculated standard deviation of 0.636 for group 1 and 0.760 for Group 2. He pools these values and uses 0.698 for the power calculation. The researcher then selects some other values both above and below these estimates as defined in Step 4 above. (Please see the matrix used in Table 4). For this example, the researcher is using the standard alpha level of 0.05 ($p < 0.05$) for significance and a power to obtain of 80% ($1 - \beta = 0.80$). Note: Typically, alpha levels are 0.05 and power values are either 80% or 90%.

Table 4: Sample size required from the Data-Peek for Power*.

		Range of Effect Size Based on Observed Data			
		0.20	0.25	Actual ES=0.26	0.30
Range of Standard Deviations Based on Observed Data	0.5	99	63		44
	0.6	142	91		63
	0.698 (Actual SD calculated from a data peek for power at 1/3 participant enrollment)			113	
	0.7	193	124		86

*Sample size values in the shaded portion of this table are per group.

Using these values, the researcher then calculates all combinations of sample sizes as shown in Table 4, and determines that the actual sample size he/she needs is $n=113$ for each group ($n=226$ total). He/she can also see that the range of sample sizes that he would need ‘per-group’ would be $n=193$ on the high end, in the situation where the ES is equal to 0.20 and the SD = 0.70, and $n=44$ on the low end in the situation where the ES is equal to 0.30 and the SD is equal to 0.50. Based on this Data Peek for Power the researcher decides to continue enrolling patients into the study beyond the 30 he/she had until enrollment of 113 patients per group is achieved.

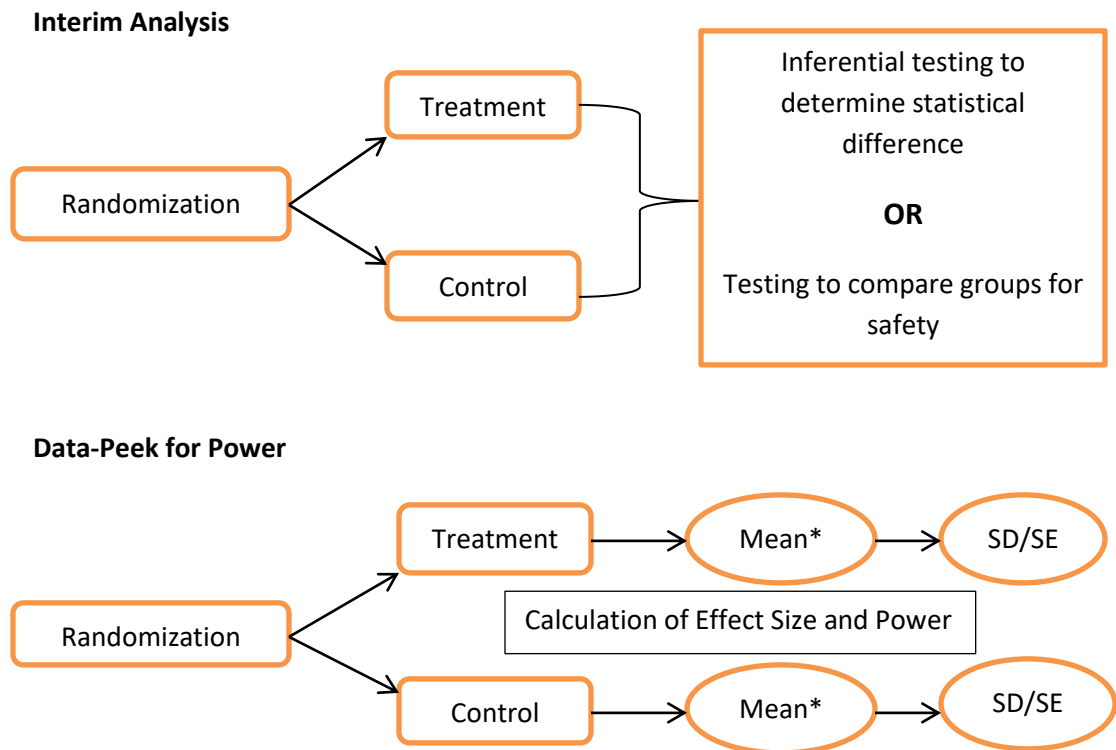
Differences between Interim Analysis and Data-Peek for Power

It is important to note that a Data-Peek for Power is not an Interim Analysis, and is entirely different from the type of analysis that a Data Safety Monitoring Board (DSMB) would suggest for patient safety or for the early termination of a study (Figure 1). In both an Interim Analysis and a DSMB required analysis significance tests are performed to determine if there is a treatment effect in the data. If there is a statistically significant difference the Interim Analysis or the DSMB might decide to stop or terminate the study early or at some other time.¹⁰ These analyses are performed with actual inferential based statistics and apply p-values to the tests based on the value of a distribution (F, t or χ^2 etc.). As a result of these analyses a study might be stopped if a treatment effect is already present or if safety concern is present.¹¹

In a Data-Peek for Power there is no inferential testing performed at all, and actually significance testing and the application of p-values is avoided. Rather as illustrated in Figure 1, the effect size is determined typically by comparing the means or proportions from the sample and then standard deviations or standard errors are generated from the values of the statistic of interest be they mean, median, proportion or percentage among others.¹²

Using these values, the sample size can then be determined in order to properly power the study at values typically of 80 or 90 percent.¹³ Using this method, the p-value for the existing data use to perform the Data-Peek for Power is never calculated. It is also typically calculated by individuals outside of the study which prevents the team conducting the study what the actual values (ES and SD/SE's) actually are. The only information that is sent back to the study team is the sample size that is needed to complete the study, Figure 1.¹⁴

Figure 1: Example of the differences between Interim Analysis and Data-Peek for Power.



*Calculation of the sample mean or some other statistic such as proportion, percent, etc.

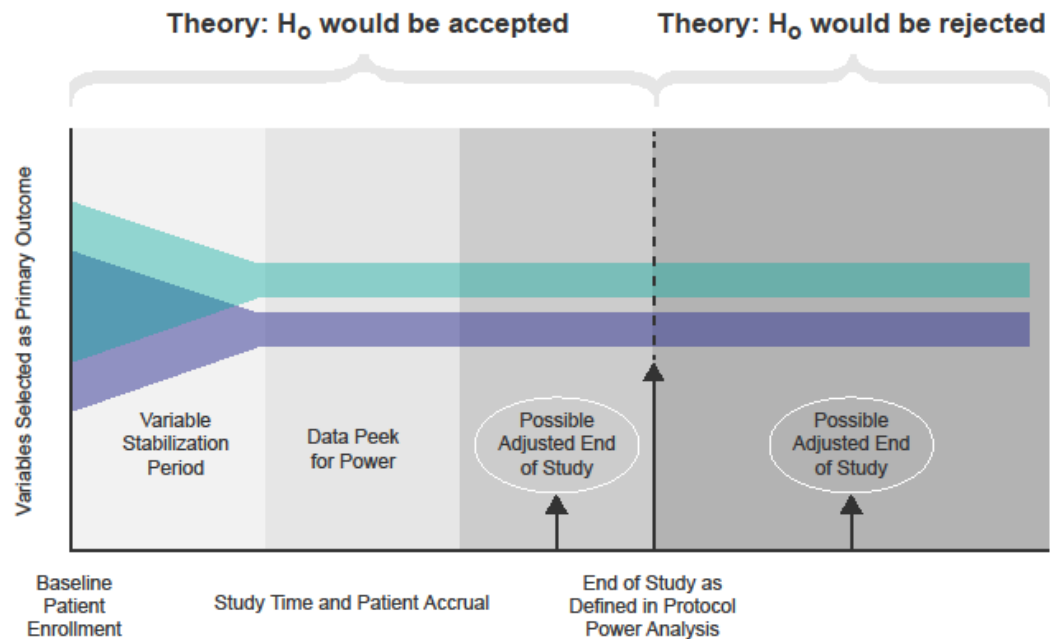
Why there is not an alpha spend for the data peek for power.

What is the alpha spend? There is some debate in statistical terms as to if an alpha spend correction is required or not. The basic theory behind alpha spend is that researchers performing a lot of tests, given the assumption of statistics with an alpha level of 0.05 or 5% would mean that one out of every 20 inferential tests applied in a study would be statistically significant by chance alone. To correct that rate, various corrections to this 'alpha spending' would be required to keep the occurrence of a statistically significant finding due only to chance as low as possible. There are many ways to do this, the most common is the Bonferroni correction where a researcher would adjust the p-value (alpha level) by dividing the significance level by the number of tests being performed. For example if the researcher is conducting seven (7) statistical tests then the p-value of 0.05 would be divided by 7 and that resulting 'alpha adjusted' p-value would be used to determine significance for each test: $(0.05/7=0.0071)$. In this example the p-value of 0.0071 would be the value where statistical significance is judged. Values below 0.0071 would be statistically significant where values above 0.0071 would be considered to be not significant.

Putting aside the issue that corrections for multiple comparisons are even required in the first place, the reason that there is not an 'alpha-spend' requiring a correction for multiple comparisons with the data peek for power is that the expectation of the null hypothesis at an interim analysis is different.¹² You are confirming your sample size calculation that a difference in groups will occur at 'X' time or with 'Y' sample size; therefore you are NOT expecting a difference would be present at the time of the data peek but rather the assumptions of group means or difference in proportions, or slope of survival curves, would be that the samples are in fact the same at the point of the data peek for power (Figure 2). At the end of the study you would expect to Reject H_0 , but in validating the sample characteristics before the expectation of the power analysis is achieved (or sample size acquired) your expectation is that if tested (which it will not be) H_0 would be accepted.

Much as a chef may taste the progress of making a fine sauce, he does not expect to be tasting the final product but rather testing the assumptions of taste at that point in the process; determining if another sprinkle of salt a pad of butter is needed. So it is with the statistician performing a data peek for power. The statistician is testing the progress of the study as is defined in the protocol. Are samples accruing as planned, standard deviations settling in to a predefined value, determining if the proportions of patients having an adverse event are consistent, etc. But this examination is always made under the assumption that while the research is ongoing, there would be no significance at the point in time that the data peek for power is performed, and if tested, H_0 would be accepted, Figure 2.

Figure 2: Illustration as to why there is a conceptual difference in alpha level expenditure in the data peek for power.



When do I conduct the data peek for power?

The fast answer to this question is that this decision depends on several factors. 1. The type of study you are conducting, 2. The period of time you need to wait to allow variables to stabilize within the design you are using, 3. The sample size calculation that was derived before the study was conducted. 4. The time that is required for participant involvement in the trial, whether it be a pragmatic trial or a randomized clinical trial.

In practice the real time to conduct the Data peek for power is when the researcher feels the data are stable enough to be analyzed, and the Effect Size can be calculated. This can be generally 1/3 of the way through the planned experiment (when 1/3 of the patients have completed and an ES can be calculated. Or when the researcher feels the rules of Central Limit theorem have been satisfied which is generally after each group or single group has 25-35 participants that have completed the study.

Can more than one Data Peek for Power be conducted within the same research project?

Yes. Why is this possible? The first reason is because since there are no inferential statistics applied to the data there is no alpha spend at any occurrence. This would not be true for an interim analysis where statistical assessments are conducted and actual p-values are calculated. Secondly, as a result of the first Data Peek for Power it might be determined that the calculated values for effect size or variation within the data have not yet stabilized for the samples. Upon seeing this, the researcher might decide to re-perform the same analysis after another 10 or 20 patients have completed the study to get a better idea as to what stable data might look like.

Who performs the data analysis for the Data Peek for Power?

It is best if the analysis can be done by a third party either inside or outside the organization. This prevents the study team from being tempted to take the values used for the Data Peek for Power and using them in one of many on-line calculators where the actual p-value can be determined. Conducting the analysis outside of the team involved in the research would avoid the temptation to request a p-value which would by rule require a correction for multiple comparisons at end point of the study or some other alpha-spend technique.

References

1. Chow S, Chang M. Adaptive Design methods in Clinical Trials – a review. *Orphanet J Rare Dis*. 2008;3(11):1-13 doi:10.1186/1750-1172-3-11
2. LaCaze A, Duffull S. Estimating risk from underpowered, but statistically significant studies. *J Clin Pharm Ther*. 2011;36:637-641 doi:10.1111/j.1365-2710.2010.01222
3. Food and Drug Administration. Guideline for the format and content of the clinical and statistical sections of new drug applications. FDA, U.S. Department of Health and Human Services. Rockville, MA, U.S.A. 1988
4. Hung J, Cui L, Wang S, et al. Adaptive statistical analysis following sample size modification based on interim review of effect Size. *J Biopharm Stat*. 2005;15:693-706: DOI: 10.1081/BIP-200062855
5. Moss AJ, Zareba W, Hall WJ, et al. Prophylactic implantation of a defibrillator in patients with myocardial infarction and reduced ejection fraction. *NEJM*. 2002;346:12:877–883.
6. Rosenberg M, ed. *The agile approach to adaptive research design*. John Wiley & Sons; 2010.
7. DeMets D, Lan G. Interim analysis: the alpha spending function approach. *Stat Med*. 1994;13:1341-1352.
8. Armitage, P., McPherson, C.K., Rowe, B.C. Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society*. 1969;132:235–244.
9. Donner A, Klar N, eds. *Design and analysis of cluster randomization trials in health research*. John Wiley & Sons; 2000.
10. Dallas M. Accounting for interim safety monitoring of and adverse event upon termination of a clinical trial. *J Biophar Stat*. 2008;18:631-638. DOI: 10.1080/10543400802071311
11. Tharmanathan P, Calvert M, Hampton J, et al. The use of interim data and data monitoring committee recommendations in randomized controlled trial reports: frequency, implications and potential sources of bias. *BMC Med Res Methodol*. 2008;8:12-20. doi:10.1186/1471-2288-8-12
12. Rothman K. No adjustments needed for multiple comparisons. *Epidemiology*, 1990;1(1):43-46.
13. Landau S, Stahl D. Sample size and power calculations for medical studies by simulation when closed form expressions are not available. *Stat Methods Med Res*. 2012;22(3):324-345. DOI: 10.1177/0962280212439578
14. Skovlund E. Interim analysis of survival data in cancer clinical trials. *Acta Oncol*. 1998;37(7/8):645-650.