PharmaSUG 2022 - Paper SA-142

Analysis of sample size calculations in clinical trial – errors, pitfalls, and conclusions

Igor Goldfarb, Ritu Karwal, Xiaohua Shu
Accenture Life Sciences

ABSTRACT

The goal of this work is to increase awareness of investigators, medical writers and statisticians working on the development, planning and conduction of the randomized clinical trial (RCT) about real importance of the statistical sections of the corresponding study protocols describing calculations of the sample size for this trial.

Every RCT is carefully and thoroughly planned from its very early stage. This plan typically includes the main objectives of the trial, primary and secondary endpoints, method of collecting the data, sample size with scientific justification, methods of handling data, statistical methods and assumptions. This plan represents a natural part of the Protocol of RCT. One of the key aspects of this plan is an estimation of the required sample size – a number of subjects to be enrolled in the RCT. Normally every protocol developed for upcoming RCT contains special section describing statistical considerations and medical justifications for the proposed number of patients/participants of the trial.

The authors reviewed a large number of protocols available on the government website clinicaltrials.gov and came to conclusion that in many cases sample size calculations are still inadequately reported, often erroneous, and based on assumptions that are frequently inaccurate. The paper illustrates how far from each other can be numbers described in the Protocol (sample size of RCT) and estimations obtained using the same assumptions (to verify replicability) that were carefully described in sections of protocols devoted to the calculation of the sample size. Such a situation raises questions about how sample size is calculated in RCT and is supposed to attract more attention to still actual need to perform a correct estimation of sample size.

INTRODUCTION

It is well known that a journey of a new invented biologic product or chemical compound from research laboratory to the authority's approval as an effective and safe medicine takes many years. Clinical trials take a significant part of this time.

It is also broadly acknowledged that the prospective study is expected to be conducted and analyzed strictly in accordance with the corresponding Protocol, which describes the background, rationale, objectives, design, methodology, statistical considerations, and organization of a prospective clinical research project. The Protocol is developed in a process of a close collaboration of principal investigators, clinicians, medical specialists, pharmaceutical developers, medical writers, biostatisticians, etc. The results of this collaborative work are typically presented in the form of the document representing the set of very detailed instructions how the prospective study is supposed to be run. No need to say that the planned study must meet all the regulatory requirements.

Conduction of clinical trial and analysis of the obtained results represent in our days a process taking place in a well-established and regulated (both on national and international levels) environment. One of many

internationally acknowledged documents regulating conduction of RCT is Guidelines for Good Clinical Practice (ICH E6(R2)).

Section 2 of these Guidelines (entitled “The Principles of ICH GCP”) presents very basic principles that any clinical research must follow to be accepted by the broad scientific community. Specifically, Subsection 2.5 clearly dictates for the prospective investigation to be scientifically reasonable: “Clinical trials should be scientifically sound, and described in a clear, detailed protocol.” Section 6 of the document (entitled “Clinical Trial Protocol and Protocol Amendment(s)”) provides some more details how the prospective protocol should be organized. The authors would like to emphasize, in particular, Subsection 6.9.2 that reads: “The number of subjects planned to be enrolled. In multicenter trials, the numbers of enrolled subjects projected for each trial site should be specified. Reason for choice of sample size, including reflections on (or calculations of) the power of the trial and clinical justification.”

It is evident that the requirement “to be scientifically sound” is rather general and covers many different aspects that are important for the planned research to be successful. The present paper is concentrated on one particular characteristic of the prospective study – how the sample size is calculated and reported in the RCT. While a full-size discussion on this subject would require a lengthy discussion (far beyond the scope of the present paper), the author focus on statistical considerations as they relate to the number of participants, or sample size, of clinical trials. In our analysis we will follow the principles listed in the ICH E9 Statistical Principles for Clinical Trials. More specifically, the authors will utilize the following statement from the Subsection 3.5: “The method by which the sample size is calculated should be given in the protocol, together with the estimates of any quantities used in the calculations (such as variances, mean values, response rates, event rates, difference to be detected).”

The present work was initiated by unexpected results of routine review of a protocol that one of Accenture’s sponsors was going to implement. Department of Clinical Statistics at Accenture Life Science took part in this review and in particular the estimation for the required sample size was examined. To the joint surprise the replicated values (based on the parameters detailed in the corresponding section of the protocol) were not even close to the numbers announced in the protocol. It was evident that the prospective study is going to be seriously underpowered, and its results will not be reliable. The outcome was double checked, but it just proved the correctness of this conclusion. Appropriate message was included in the final review and delivered to the sponsor.

The authors became intrigued and started to learn about the issues related to the estimation of sample size for prospective studies. It was immediately found that the subject is not new and even though it has been under discussion for a long time the issue is still actual. The next step was analysis of some active projects from this point of view (not included in the present paper). After that a couple of publicly available protocols (of completed and reported studies) were downloaded from https://www.clinicaltrials.gov/ and reviewed and analyzed. The obtained results were disappointing – replicated sample size estimations did not match corresponding values presented in the protocols in a relatively large portion of the researched protocols. At some point the authors decided to conduct systematic review of the situation based on the latest protocols available on https://www.clinicaltrials.gov/. The present work summarizes the main results of this activity.

BACKGROUND

The current Section contains a brief description of the statistical terms and notions and incomplete review of the published articles devoted to the issue of sample size estimation and its reporting in the study protocols.

TERMS AND NOTIONS

The calculation of the correct sample size is one of the first and most important steps in study design and development of its protocol. Before starting of an actual presentation of the results let's remind – very shortly - basic statistical terms and notions one needs to understand to compute sample size for prospective study.

Study Hypotheses – Null H_0 and Alternative H_1

Consider an arbitrary pharmaceutical company testing its investigational product against control treatment (already existing on the market) to evaluate if the investigational product is better in curing subject's suffering from some specific disease as compared to control treatment. Let's formulate two basic statistical hypotheses around this hypothetical research that are typically under consideration:

- Null Hypothesis (H_0): investigational product is equal to control treatment
- Alternate Hypothesis (H_1): investigational product is better than control treatment

Type I and II Errors – Significance and Power

Let's assume that based on the results of the study it is concluded that investigational product is better than control treatment. In other words, one rejects the null hypothesis H_0 and accepts alternative one H_1. Knowing (assumption just to illustrate the example) that the investigational product in reality is equivalent (=not better) to the control treatment (H_0), one is making an error here by rejecting H_0. This is called as Type I error. Statistically it is defined as Type I error = [(Reject H_0)/(H_0 is true)]. Probability of Type I error is called as level of significance and is denoted as α.

To demonstrate an opposite case let's assume that based on the results of the study it is concluded that the investigational product is not better (=equivalent) than the control treatment. In other words, one rejects the alternative hypothesis H_1 and accepts the null one H_0. Knowing (once again - assumption just to illustrate the example) that the investigational product in reality is better than the control treatment (H_1), one is making an error here by rejecting H_1. This is called as Type II error. Statistically it is defined as Type II error = [(Reject H_1)/(H_1 is true)]. Probability of Type II error is denoted as β, but the term "power" (=1- β) is used more frequently than "type II error β".

Normally these types of errors are presented in the form of table (Table 1) for better convenience.

	The null hypothesis is true: there is no true difference (H_0 is true, H_1 is false)	The null hypothesis is false: there is a true difference (H_0 is false, H_1 is true)
Study accepts the null hypothesis H_0	Correct conclusion: specificity (1−α)	Type II error (β)
Study rejects the null hypothesis H_0	Type I error (α)	Correct conclusion: power (1− β)

Table 1. Type I and II Errors

Sample Size Calculation – Required Items

Using the standard approach for determining the appropriate sample size, the following items need to be specified to make the computation feasible:

- A primary variable(s)
- The test statistic(s)
- The null hypothesis
- The alternative (working) hypothesis at the chosen dose(s)
- Effect size

- The probability of erroneously rejecting the null hypothesis (the type I error - α)
- The probability of erroneously failing to reject the null hypothesis (the type II error - β)
- Adjustments for treatment withdrawals and protocol violations

LITERATURE OVERVIEW

In a landmark study published in 1978, Freiman *et al* (Freiman *et al*, 1978) conducted a survey of 71 negative RCT published in the medical literature. They demonstrated that reports of many RCT do not have adequate power to demonstrate clinically important differences in the therapeutic effect. On the 71 trials evaluated, 67 had less than a 90% power to detect a 25% therapeutic improvement and 50 had less than a 90% power to detect a 50% improvement.

Dimick, Diener-West and Lipsett (2001) identified 90 reports of RCT with negative results in the surgical literature between 1988 and 1998. The manual review of 1997 showed a 100% retrieval rate for their search strategy. After applying the Two One-Sided Tests Procedure, 35 reports (39%) met the criteria for demonstrating equivalency. The other 55 reports (61%) contained at least 10% absolute difference in the 90% confidence interval of Δ. Using the power calculation method, only 22 (24%) articles had a power greater than 0.80 to detect a 50% difference in therapeutic effect. Only 29% of the reports included a formal sample size calculation and these studies were more likely to demonstrate equivalency than those without a sample size estimate ($P<0.01$).

Charles *at al* (2009) have assessed quality of reporting of sample size calculation, ascertained accuracy of calculations, and determined the relevance of assumptions made when calculating sample size in randomized controlled trials. They searched MEDLINE for all primary reports of two arm parallel group RCT of superiority with a single primary outcome published in six high impact factor general medical journals between 1 January 2005 and 31 December 2006. All extra material related to design of trials (other articles, online material, online trial registration) was systematically assessed. Data extracted by use of a standardized form included parameters required for sample size calculation and corresponding data reported in results sections of articles. The authors (Charles *at al.*, 2009) checked completeness of reporting of the sample size calculation, systematically replicated the sample size calculation to assess its accuracy, then quantified discrepancies between a priori hypothesized parameters necessary for calculation and *a posteriori* estimate.

The authors (Charles *at al.,* 2009) found that out of the 215 selected articles, 10 (5%) did not report any sample size calculation and 92 (43%) did not report all the required parameters. The difference between the sample size reported in the article and the replicated sample size calculation was greater than 10%in 47 (30%) of the 157 reports that gave enough data to recalculate the sample size. The difference between the assumptions for the control group and the observed data was greater than 30% in 31% (n=45) of articles and greater than 50% in 17% (n=24). Only 73 trials (34%) reported all data required to calculate the sample size had an accurate calculation and used accurate assumptions for the control group.

Charles *at al.* (2009) concluded that sample size calculation is inadequately reported, often erroneous, and based on assumptions that are frequently inaccurate. Such a situation raises questions about how sample size is calculated in RCT.

Abdulatif *et al.* (2015) have evaluated the pitfalls in reporting sample size calculation in RCT published in the 10 highest impact factor anesthesia journals.

Superiority RCTs published in 2013 were identified and checked for the basic components required for sample size calculation and replication. The difference between the reported and replicated sample size was estimated. The sources used for estimating the expected effect size (Δ) were identified, and the difference between the expected and observed effect sizes (Δ gap) was estimated.

The authors (Abdulatif *et al.,* 2015) enrolled and analyzed 194 RCTs. Sample size calculation was reported in 91.7% of studies. Replication of sample size calculation was possible in 80.3% of studies. The original and replicated sample sizes were identical in 67.8% of studies. The difference between the

replicated and reported sample sizes exceeded 10% in 28.7% of studies. The expected and observed effect sizes were comparable in RCT with positive outcomes (P=0.1). Studies with negative outcome tended to overestimate the effect size (Δ gap 42%, 95% confidence interval 32–51%), P<0.001. Post hoc power of negative studies was 20.2% (95% confidence interval 13.4–27.1%). Studies using data derived from pilot studies for sample size calculation were associated with the smallest Δ gaps (P=0.008).

The authors (Abdulatif *et al.*, 2015) concluded that sample size calculation is frequently reported in anesthesia journals, but the details of basic elements for calculation are not consistently provided. In almost one-third of RCTs, the reported and replicated sample sizes were not identical and the assumptions for the expected effect size and variance were not supported by relevant literature or pilot studies.

One can easily conclude from the incomplete literature review above that the issue with correct estimation of the required number of subjects for a prospective study and appropriate reporting of the details of this estimation in the statistical section of a protocol did not lose its actuality in our days despite decades of discussion on different forums, in various formats and within unparallel professional communities.

ANALYSIS AND REVIEW OF ACTUAL STUDIES' DATA

Accenture Life Sciences has extensive experience in various aspects of conduction of clinical trials (e.g., development of prospective study design, computation of corresponding sample size, authorization of statistical section of a protocol, development of CRF, building study database, collection different types of medical data, analysis of collected clinical information, processing and presentation of derived data, etc.). For the purposes of the research displayed in the present paper the authors used the protocols of completed studies available at the database https://www.clinicaltrials.gov/. This repository contains privately and publicly funded clinical studies conducted around the world. As of March 2022, the database included more than 406,000 studies conducted in all 50 states and in more than 220 countries.

nQuery

There are many competing tools on the market in the field of statistical analysis – SAS, IBM SPSS, R studio, STATA, JMP, GNU-Octave, to name the few. Most of them are full spectrum packages oriented for the wide range of potential tasks – from calculation of averages to multiple imputations. Attention that is devoted to study design and sample size calculation in these all-purpose packages is minimal and in some cases is just absent. On the contrary, nQuery is intentionally oriented to resolve the tasks from this list – from sample size calculation for classical double-blind RCT with two treatment groups to development of trials with flexible design and multiple interim analyses (α-spending approach). nQuery is used by the Department of Clinical Statistics at Accenture Life Sciences as the main tool for the design of prospective studies and computation of corresponding samples sizes.

As it was already mentioned above all computations presented in the current work were performed using nQuery as the main tool.

RESULTS OF THE ANALYSIS – GENERAL DESCRIPTION

The authors investigated 128 actual protocols from full spectrum of therapeutic areas available on the website https://www.clinicaltrials.gov/. The main criteria for the protocol to be chosen for analysis was a presence in it a separate section devoted to the calculation of the sample size that was used as a target for recruitment of patients. Some of the protocols described the studies which included two or three stages and a couple of such individual stages had their own calculation of the planned sample size. In this situation the authors modelled every one of these stages separately. In total 132 different sections

containing details of sample size computation were thoroughly examined. To avoid any confusion between actual protocols and the distinct stages of some of them having separate sample size calculations the authors will refer to these individual stages as to protocols.

Despite careful preliminary screening the authors found that 10 of the 132 selected protocols do not contain an actual calculation of the sample size that can be used for analysis and replication (despite the presence of the statistical section devoted to this subject). Out of 122 remaining protocols 17 do not contain enough information allowing us to conduct replication analysis. Finally, only 105 protocols allowed the authors to examine themselves and to compare replicated numbers with one declared in the protocols. Corresponding counts for general conclusions that were made after meticulous analysis of every one of the protocols are displayed in Table 2.

Conclusion	Count
Able to replicate	81
No sample size calculations performed	10
Not able to replicate	24
Not enough information provided to be able to replicate	17
Grand Total (reviewed protocols)	**132**

Table 2. General conclusions made after protocol's analysis.

These 105 protocols were using different statistical tests and methods to estimate the sample size for prospective studies, to name the few – t-test, chi-square, Simon design, etc. The distribution of the methods is presented in Table **3**. It can be easily concluded that t-test, two-stage Phase II Simon design, Chi-square and paired t-test are leading the list of popular statistical methods with 31 (29.6%), 20 (19.0%), 16 (15.2%), and 13 (12.4%) cases, correspondingly.

Statistical test/approach	Count
ANOVA	1(1.0%)
Binomial	8(7.6%)
Chi-square	16(15.2%)
Correlation	1(1.0%)
Fisher	4(3.8%)
Fleming	1(1.0%)
Green and Dahlberg	1(1.0%)
Paired t-test	13(12.4%)
Simon	20(19.0%)
Single Stage Phase II Design	1(1.0%)
Survival (Cox-Regression)	1(1.0%)
Survival (Log-Rank)	5(4.8%)
t-test	31(29.6%)
Two One-sided Test (TOST)	2(1.9%)
Grand Total	**105**

Table 3. Distribution of statistical methods applied in reviewed protocols

These 105 protocols allow us to perform recalculation process for the corresponding sample sizes. It means that every one of the protocols contains adequate amount of input data that can be plugged into independent replication computation. This group of the protocols represents a final set of the data that allow an arbitrary reader to review the sample sizes declared in the corresponding statistical sections of the protocols. Table 4 shows that the authors were able to replicate the sample size values in 81 (77.1%) cases, whereas in other 24 (22.9%) the difference between replicated value and the protocol's one was greater than 10%. The largest distinctions (>50%) were observed in 15 studies (62.5% - total 24 mismatches were used as a denominator for calculation of percentages for various grades of differences).

Replication results vs protocol Grade of mismatch	**Count N (%)**
Don't Match	24 (22.9%)
11-25%	3 (12.5%)
26-50%	6 (25%)
>50%	15 (62.5%)
Match	81(77.1%)
Grand Total	**105**

Table 4. Summary of comparison of data presented in protocols and replication results

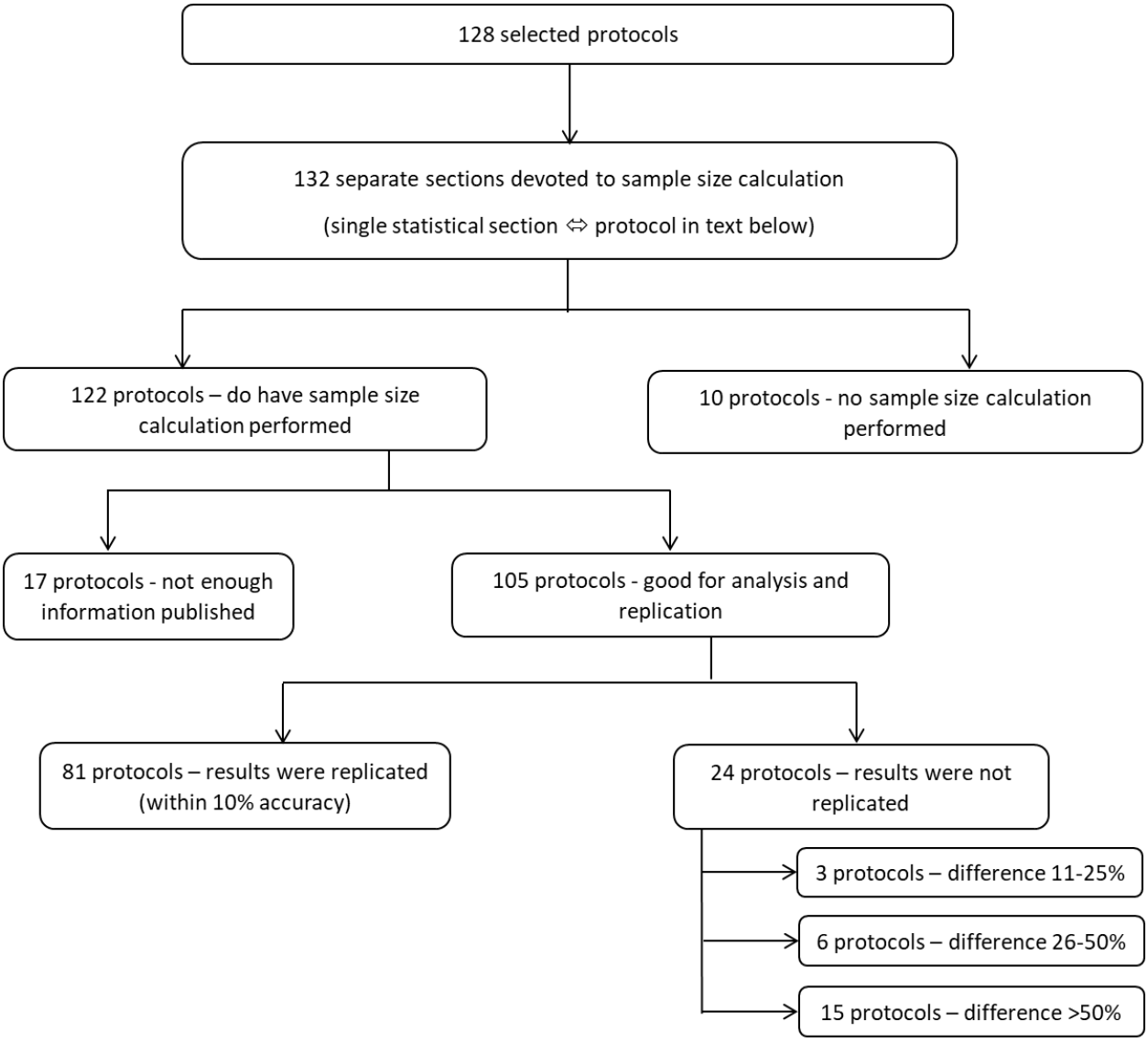
Flowchart presented in Display 1 contains all the details about the workflow of the review of available protocols and analysis of sample sizes used in the corresponding studies.

SIMON DESIGN – UNIQUENESS OF REPLICATION OUTCOMES

The authors would like to make a special note regarding the uniqueness of the replication results for Phase II studies built on 2-stage Simon design. Remind, Simon's Two-Stage Design is a special type of design developed specifically for Phase II clinical trials. It is currently one of the most common multi-stage designs used in Phase IIa clinical trials. The Simon two-stage design is an exact design which allows flexibility regarding the null and alternative hypotheses while also allowing stopping for futility. The objective of the two-stage design is to try and establish whether the proportion of responses is sufficiently high to recommend this drug to go to the next step in the clinical trial phase, Phase IIB.

The matter is that all 20 statistical sections demonstrating estimations of the planned sample size for Phase 2 Simon design were successfully replicated with 100% accuracy (sic!). It simply means that in all 20 cases the authors were able to get the numbers for corresponding sample sizes exactly coinciding with the values declared in the original protocols. As it was already mentioned the conclusion "Able to replicate" was made if the distinction between the protocol's numbers and replicated ones did not exceed 10%. The fact that the authors observed 100% accuracy for Simon's design witnesses about very robust algorithm standing behind this type of design. This conclusion is straightforward – there are no reasons to believe that all the statisticians who prepared corresponding statistical section for these 20 protocols were using nQuery. Different people exploit different tools (see nQuery) for the purpose of sample size determination and different tools are expected to provide the researchers with probably close, but not exactly coinciding numbers. Therefore, it would be very natural to observe slight differences between the protocol's numbers and replicated ones, but it is not what happened. Hence, one may conclude that the algorithm is so robust that being implemented by various statistical packages it gives identical results.

To finalize the discussion about Simon's design it would be worthwhile to note also that one of the protocols (NCT02324543) used 2-stage Phase II design according to Green and Dahlberg approach. The authors were not able to identify calculation algorithm in nQuery that is following this approach and Simon's design was applied instead. It turned out that the obtained results are very close to those announced in the protocol – the total required number of patients was 32 (versus 30 in the protocol) and number of subjects for the first stage was 12 (versus 15 in the protocol). Thanks to the fact that the difference between total numbers lies within 10% interval (32 vs 30) we assigned this case to the group "Able to replicate" (even though the method is different from the one cited in the protocol).



Display 1. Flowchart of reviewed protocols

MISMATCHES - EXAMPLES

Simon's design was really unique – not all sample sizes were replicable. To illustrate how sample sizes declared in the protocols can be different from what follows from the calculations let's analyze in details a couple of examples.

Example 1. Study "A Pilot Trial to Assess Low-Intensity Ultrasound in Osteoarthritis (PLUS+OA)" (ClinicalTrials.gov Identifier: NCT02034409). The study results were posted to the website in October 2021.

According to the Protocol of the study the expected change in the primary variable (48-week cMFCtTH change) was expected to be 33 μm with standard deviation 152 μm (full details of the calculation are displayed in Display 2).

II.D.4 **Sample Size Calculation**

cMFCtTh Change at 48-Weeks

For the 48-week cMFCtTH change outcome measure the minimally clinically significant difference is 33 μm. With a difference in cMFCtTh measurement of 33 μm, σ=152 μm standard deviation, k=2 groups, and P=0.90 probability of correct selection we use Table 1 of Gibbons et al. 1979 [92] to obtain τ=1.8124, which we use in the equation to estimate the per-group sample size: $n=\sigma^2(\tau/\delta^*)^2=72$ or a total of 144. With a total sample size of 144 the probability of correctly selecting the group with the greatest true cMFCtTh change at 48-weeks is slightly above 0.90 if the true group difference in cMFCtTh change at 48-weeks is at least 33 μm. If the assumed standard deviation of 152 μm turns out to be conservative, realized standard deviations of 196 μm and 237 μm will provide us with probabilities of correct selection of 0.85 and 0.80 respectively.

Display 2. Sample size calculation (screenshot from the Protocol for the study NCT02034409)

The authors made several attempts to replicate the results of this sample size determination using the parameters presented in the Protocol. As one can easily see the Protocol's description does not contain value of α that was used in the calculation. Therefore, two most popular values (α=0.05 and α=0.1) for this type of calculations were used to cover all possible options. The calculated values of the sample size are 447 and 365 per group, correspondingly (refer to Display 3 to see nQuery screenshot with the results). The formulation proposed by nQuery for a hypothetical study protocol (based on the entered values) reads as "A sample size of 447 in each group will have 90% power to detect a difference in means of 33 assuming that the common standard deviation is 152 using a two-group t-test with a 5% two-sided significance level.".

No need to say that both values are quite different from n=72 per group required according to the Protocol. In other words, the original study was essentially underpowered: to detect the announced difference of 33 μm with standard deviation 152 μm (presuming 90% power and α=0.05) the study needs many more (more than 8 times more!) patients to be recruited.

MTT0-1 / Two Group t-test of Equal Means

	1	2
Test Significance Level, α	0.100	0.050
1 or 2 Sided	2	2
Group 1 Mean, μ_1		
Group 2 Mean, μ_2		
Difference in Means, $\mu_1 - \mu_2$	33.000	33.000
Common Standard Deviation, σ	152.000	152.000
Effect Size, $\delta=\lvert\mu_1-\mu_2\rvert/\sigma$	0.217	0.217
Power (%)	90	90
n per Group	365	447

Display 3. Replication results for study NCT02034409 (nQuery screenshot)

Example 2. Study "Phase 1/2 Study of Carfilzomib for the Prevention of Relapse and Graft-versus-host Disease in Allogeneic Hematopoietic Cell Transplantation for High-risk Hematologic Malignancies" (ClinicalTrials.gov Identifier: NCT02145403). The study results were posted to the website in January 2020, last update was posted in January 2022.

According to the Protocol of the study the expected improved probability of event within 1 year was estimated as 0.65 (whereas historical probability for the same time period was 0.85). Presuming 80% power and α=0.05 the Protocol estimates the required number of patients for the second stage of the study as 35 (full details of the calculation are displayed in Display 4).

10.2 SAMPLE SIZE CONSIDERATIONS

Based on the University of Michigan experience on fludarabine-based related donor allo-HCT, the event-free survival at 1 year was 15% [unpublished data]. Event is defined as relapse/progression, grade III-IV acute GVHD or chronic GVHD requiring systemic therapy.

We expect that adding carfilzomib to a standard regimen will increase EFS at 1 year from 15% to 35% and 35 subjects need to be evaluated for primary efficacy endpoint, with α error of 0.05 and statistical power of 80%. Thus, we will enroll 37 patients in the optimal dose used in the phase II part, assuming 2 patients will not be evaluable for endpoint analysis.

$$N = \frac{\left(Z_\alpha\sqrt{p_0(1-p_0)} + Z_\beta\sqrt{p_1(1-p_1)}\right)^2}{(p_0-p_1)^2}$$

N = Number required
p_0 = Historical probability of event within 1 year = (1-0.15) = 0.85
p_1 = Desired probability of event within 1 year = (1-0.35) = 0.65
Z_α = 1.96 for 5% significance level
Z_β = 0.84 for 80% power

For phase 1 part, up to 18 additional subjects may be needed for possibility of 6 subjects per cohort. Thus, we plan to enroll a maximum 55 patients (18 max Phase I, 37 Phase II).

Display 4. Sample size calculation (screenshot from the Protocol for the study NCT02145403)

The authors applied several available algorithms to replicate the results of this sample size determination using the parameters presented in the Protocol. The Section 10.2 of the Protocol (Display 4) clearly and explicitly displays all the required parameters for computation. This set of parameters is plugged in in the column #1. The calculated value of the sample size for these parameters is 64 subjects, which exceeds by approximately 82% the number declared in the Protocol (35). The proposed by nQuery formulation for a hypothetical study protocol (based on the entered values in the column #1) reads as "When the sample size in each group is 64, with a Total number of events required, E, of 90, a 0.05 level two-sided log-rank test for equality of survival curves will have 80% power to detect the difference between a Group 1

proportion π_1 at time t of 0.35 and a Group 2 proportion π_2 at time t of 0.15 (a constant hazard ratio of 0.553); this assumes no dropouts before time t..".

To examine how sensitive the obtained sample size is to the parameters that were plugged in, the authors conducted additional computation. Column #2 in Display 5 contains the results of sample size modeling for 1-sided test (all other parameters were kept the same as in column #1), and column #3 displays the results of sample size modeling for α=0.1 test (all other parameters were kept the same as in column #1). The results are equal to each other (50 subjects) demonstrate that the calculation is rather sensitive to the input parameters. Nevertheless, the obtained outcome (50) exceeds the by approximately 41% the number declared in the Protocol (35).

STT0-1 / Log-Rank Test of Survival in Two Groups Followed for Fixed Time,

	1	2	3
Test Significance Level, α	0.050	0.050	0.100
1 or 2 Sided Test?	2	1	2
Group 1 Proportion π_1 at Time t	0.350	0.350	0.350
Group 2 Proportion π_2 at Time t	0.150	0.150	0.150
Hazard Ratio, $h=\ln(\pi_1)/\ln(\pi_2)$	0.553	0.553	0.553
Power (%)	80	80	80
Sample Size per Group, n	**64**	**50**	**50**
Total Number of Events Required, E	90	71	71

Display 5. Replication results for study NCT02145403 (nQuery screenshot)

Example 3. Study "Role of Pegylated Interferon in Combination with DAAs to Cure Hepatitis C As Soon As Possible - Hepatitis C (ASAP-C)" (ClinicalTrials.gov Identifier: NCT03480932). The study results were posted to the website in May 2021.

According to the Protocol of the study the expected improvement was estimated as 25% (70% vs 95%). Assuming two-sided test and α=0.05 the Protocol concludes that each arm required 50 people to reach the power >80% (full details of the calculation are displayed in Display 4).

c. Statistical plan including sample size justification and interim data analysis.

All analyses will be conducted as intention to treat. We will use a chi-squared test of proportions to compare the proportions that achieve SVR12. Pairwise comparisons with correction for multiple comparisons will be conducted to directly compare Arms 1 and 2, Arms 1 and 3 and Arms 2 and 3. For secondary outcomes, chi squared tests of proportions will be used to compare categorical outcomes and a Wilcoxon rank-sum test will be used to compare adherence.

The primary objective of this study is to test the feasibility of DOT for HCV treatment among HCV-infected persons with substance use history in a resource limited setting. An additional objective is to generate pilot data to plan a larger powered study. We hypothesize that SVR12 will be higher in Arm 1 compared to Arms 2 and 3. Assuming a two-sided alpha=0.05, 50 people in each arm and 70% achieving SVR12 in Arm 3, we will have >80% power to detect a difference of 25% (e.g., 95% completion in the SOF+DAC+PEG).

No interim analyses will be conducted as with 150 participants, it will not be possible to observe statistical differences prior to the completion of the trial.

Display 6. Sample size calculation (screenshot from the Protocol for study NTC03480932)

This Protocol is an example of good reporting of all the details required for the calculation of the sample size – proportion of SVR12 that is used as a control value (70%), value of expected improvement (25%), exact name of the statistical test (chi-squared test of proportions), one or two-sided (two-side), value of α

(α=0.05), desired power (>80%). It is a pleasure to replicate someone's computation under this set of well-defined conditions. Display 7 contains the results of the replication. Once can see that for sequence of power values of 80%, 85%, 90% the algorithm provides corresponding values of patients per group – 36, 40, 47 (columns #1, #2, #3, correspondingly. Results in the column #4 are obtained by reverse engineering – value of 50 subjects (as defined in the Protocol) was plugged in and the power (92.05%) was computed. Once can conclude that the study was slightly overpowered – it could use 40 patients per group and meet the condition presented in the Protocol (>80%).

PTT0-1 / Two Group χ^2 Test of Equal Proportions (Odds Ratio = 1) (Equal n's)

	1	2	3	4
Test Significance Level, α	0.050	0.050	0.050	0.050
1 or 2 Sided Test?	2	2	2	2
Group 1 Proportion, π_1	0.700	0.700	0.700	0.700
Group 2 Proportion, π_2	0.950	0.950	0.950	0.950
Odds Ratio, $\Psi=\pi_2(1-\pi_1)/(\pi_1(1-\pi_2))$	8.143	8.143	8.143	8.143
Power (%)	80	85	90	92.05
Sample Size per Group, n	**36**	**40**	**47**	**50**

Display 7. Replication results for study NTC03480932 (nQuery screenshot)

DISCUSSION

The present research allows to conclude that the issue with correct estimation and appropriate reporting of the sample size for clinical studies is still alive and represent an actual problem – almost a quarter (22.9%) of the trials that contained enough information for replication did not match the recalculated values.

This result can be compared with the outcome obtained by Charles *et al*. (2009). Charles and his team searched MEDLINE (six high impact factor general medical journals between January 1, 2005, and December 31, 2006) for all primary reports of two arms parallel groups RCT of superiority with a single primary outcome. They reported that the difference between the sample size calculation was greater than 10% in 47 (30%) of the 157 reports that gave enough data to recalculate the sample size. The result presented in this paper (22.9% for the same parameter) is slightly better but scientific community could expect better adherence to well-known practice for sample size estimation.

Note that Abdulatif *et al*. (2015) restricted their investigation by the papers published in 10 high impact anesthesia journals, while authors of the present work did not apply any condition on the therapeutic area of the study. Therefore, the direct comparison should be conducted with this major difference in mind. Nevertheless, Abdulatif *et al*. (2015) reported that the difference between the replicated and reported sample sizes exceeded 10% in 28.7% of studies (compare to 22.9% in the present paper).

Ordering the results obtained by Abdulatif *et al*. (2009), by Charles *et al*. (2015), and by the authors of this paper in time, one can observe a clear tendency of unhurried improvement: 2009 – 30%, 2015 – 28.7%, 2022 – 22.9% (all percentages here describe a relative number of publications/RCT where difference between reported and replicated values of sample sizes is greater than 10%). Therefore, in addition to the main conclusion that the situation with computing and reporting of the sample size is still not perfect, an detected tendency provide us with some cautious optimism – situation is slowly improving with time.

It is well known that a way of a synthesized compound from research laboratory to the FDA approval as an effective and safe drug takes many years. Clinical trials take significant part of this time. Data collected during a clinical trial are cleaned, reviewed, verified, reconciled, fixed (if necessary), and, finally,

analyzed, processed, summarized, and displayed in the form of TLF. Every one of these steps takes its own time, no error is permitted in this sequence and all team members share the common task of reducing the total time required for drug approval. The combined efforts of the whole team can be compromised if an estimation of the sample size was performed inappropriately, and the study is actually seriously underpowered (as can be seen from the examples above).

The goal of this paper is to attract attention of the professionals working in the pharmaceutical industry to a necessity of correct computation and detailed report of sample size determination in the study protocols. This is of vital importance for pivotal RCT, because underpowered trial can disrupt, delay, or even stop a development of a promising treatment that many patients are waiting for.

CONCLUSIONS

To recap the discussion of the situation with computing and reporting of the sample size in RCT it would be worthwhile to summarize the conducted analysis and performed modelling (replication):

1. Literature review and analysis of the recent protocols demonstrate that a situation with correct calculation and appropriate reporting of sample size is still not perfect, and the issue is still on the agenda for researchers in the pharmaceutical industry.
2. The details of sample size calculation presented in the protocol in many cases are inadequate to allow independent replication of the numbers used as a target for patient recruitment.
3. Independent review of the reported sample sizes shows that in many cases the study was underpowered, and its conclusions may be questionable.
4. An importance of reporting of effect size is underestimated and/or misunderstood. In many cases the reported information contains standard deviation, but expected difference is missing, or vice versa.
5. An importance of reporting of statistical method/test that was used for sample size determination is also underestimated. Only a real minority of the protocols that were analyzed contained explicit reference to the method applied.
6. Design of a prospective study (including estimation of the required sample size) and consequent development of the statistical section of a protocol should be performed by experienced statistician equipped with appropriate knowledge and tools.

REFERENCES

1. Freiman, J.A., Chalmers, T.C., Smith, H., *et al*. (1978). “The importance of beta, the Type II error and sample size in the design and interpretation of the randomized controlled trial” *New England Journal of Medicine* 299:690-694.
2. Dimick, J.B., Diener-West, M., Lipsett, P.A. 2001. “Negative results of randomized clinical trials published in the surgical literature” *Arch. Surg*. 136:796-800.
3. Charles P., Giraudeau, B., Dechartres, A., Baron, G., Ravaud, P. 2009. “Reporting of sample size calculation in randomized controlled trials: review” *BMJ* 338:b1732

4. Abdulatif, M., Mukhtar, A., and Obayah, G. 2015. “Pitfalls in reporting sample size calculation in randomized controlled trials published in leading anaesthesia journals: a systematic review” *British Journal of Anaesthesia*, 115 (5): 699–707.
5. ICH E6(R2) Good Clinical Practice Guidance (https://database.ich.org/sites/default/files/E6_R2_Addendum.pdf)

   ICH E9 Statistical Principles for Clinical Trials (https://database.ich.org/sites/default/files/E9_Guideline.pdf)
6. Clinical Trial Design & Sample Size Calculation Mistakes to Avoid (https://blog.statsols.com/common-clinical-trial-design-sample-size-calculation-mistakes-to-avoid)
7. Study “A Pilot Trial to Assess Low-Intensity Ultrasound in Osteoarthritis (PLUS+OA)” (ClinicalTrials.gov Identifier: NCT02034409)
8. Study “Phase 1/2 Study of Carfilzomib for the Prevention of Relapse and GVHD in Allo-HCT for Hematologic Malignancies” (ClinicalTrials.gov Identifier: NCT02145403)
9. Study “Study of Gemcitabine/Taxotere/Xeloda (GTX) in Combination With Cisplatin and Irinotecan in Subjects With Metastatic Pancreatic Cancer” (ClinicalTrials.gov Identifier: NCT02324543)
10. Study “Role of pegylated Interferon in combination with daas to cure Hepatitis C as soon as possible - Hepatitis C (ASAP-C)” (ClinicalTrials.gov Identifier: NCT03480932)

ACKNOWLEDGMENTS

The authors are very thankful to upper management of Accenture Life Sciences for their constant support of this work.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Please feel free to contact anyone of the authors at:

Igor Goldfarb
Accenture Life Sciences
Igor.goldfarb@accenture.com

Ritu Karwal
Accenture Life Sciences
Ritu.Karwal@accenture.com

Xiaohua Shu
Accenture Life Sciences
Xiaohua.Shu@accenture.com