A practical approach to sample size calculation for fixed populations.

Maurits C. Kaptein

Abstract

Researchers routinely compute desired sample sizes of clinical trials to control type-I and type-II errors. While for many experimental designs sample size calculations are well-known, it remains an active area of research. Work in this area focusses predominantly on controlling properties of the trial. In this paper we provide ready-to-use methods to compute sample sizes using an alternative objective, namely that of maximizing the outcome for a whole population. Considering the expected outcome of both the trial, and the resulting guideline, we formulate and numerically analyze the expected value of the entire allocation procedure. Our approach strongly relates to theoretical work presented in the 60's which demonstrated the effectiveness of allocation procedures that incorporate population sizes when planning experiments over designs that focus solely on error rates within the trial. We add to this work by a) extending to alternative designs (mean comparisons not assuming equal variances and comparisons of proportions), b) providing easy-to-use software to compute sample sizes for multiple experimental designs, and c) presenting numerical analysis that demonstrate the efficiency of the suggested approach.

Keywords: Sample size calculation, clinical trial, decision policies.

1 Introduction

Investigators should properly calculate sample sizes before the start of their randomized controlled trials (RCTs) and adequately describe the details in their published report(s)[Schulz and Grimes, 2005]. The landmark article by Freiman, Chalmers, and Smith [Freiman et al., 1978] was one of the first to highlighted the importance of sample size calculations: numerous previously reported RCTs were severely underpowered and hence their failure to identify the efficacy of the treatments under scrutiny could hardly be considered decisive evidence. Precise estimation and powerful testing are innately connected to the number of observations collected and hence a-priori sample size considerations should be an integral part of RCT planning.

Despite the fact that for many well known RCT designs (e.g., those testing for differences in means, differences in proportions, etc.) sample size calculations are well known, the accurate computation of sample sizes for complex designs is still an active area of research. Several authors have recently considered sample size computations for specific — more complex — experimental designs[Zhu et al., 2017, Cunningham and Johnson, 2016, Shan, 2018, Qiu et al., 2016]. Furthermore, researchers have recently focussed on Bayesian methods for computing sample sizes [Brakenhoff et al., 2018], and have considered the embedding of the trial within its larger context [Whitehead et al., 2016]. In all of these cases, sample size calculations aim to control the type I (false-positive) and type II (false-negative) error rates of the RCT over repeated executions of the trial given that the assumptions made regarding the population that entered the sample size calculations are accurate.

In this paper we examine an alternative objective to determining sample sizes in RCTs. We consider the RCT as merely the first stage in a two-stage treatment allocation policy that, ultimately, allocates one out of a set of competing

treatments to all individuals suffering from a specific disease (the population). The RCT, *combined with the resulting guidelines for clinical practice*, jointly decide which patient in the population receives what treatment. Given this setup, sample size calculations can be motivated by a desire to maximize the expected overall outcome over all patients in a population. This alternative objective for sample size calculation has been studied before in the 60's—a literature we discuss in section 2.2—and its optimization leads to a demonstrably more effective allocation procedure than attained when planning trial sizes solely based on error rates. We hope to contribute by reviving this idea and bringing it to clinical practice by providing an easy-to-use software package to compute sample sizes according to this criterion for various designs, and by numerically examining the differences between the standard approach and the one advocated in this work.

In the remainder of this work we first formalize the problem at hand and motivate our focus on two-stage allocation procedures (an RCT resulting in a deterministic guideline). Next, we review prior work in this area and motivate how our work contributes. In section 3 we introduce the open-source and freely available `[R]` package `ssev` that allows researchers and practitioners to easily compute optimal sample sizes for various two-group comparisons. Next, we present a number of numerical results to further illustrate the impact of changing the sample size planning objective from the trial to population; we demonstrate that for small populations our current trials are often overly large, while for large populations they are overly small. Finally, we reflect on our presented results and discuss possible future extensions.

2 Problem formalization and relations to the RCT

The general problem we consider can be phrased in the language of potential outcomes [Rubin, 2005, Rubin, 2004]. Consider $i = 1, \ldots, N$ patients in population P, each with potential outcome $y_i(k)$ for treatment $k = 1, \ldots, K$. We are interested in evaluating the performance of different treatment allocation policies π that allocate, for each patient i in the population, one of the K treatments. Specifically, we are interested in the performance of a subset of all possible treatment allocation policies that we coin *two-stage allocation policies*:

1. In *Stage I* a number of patients n (where often $n \ll N$) is randomly selected from the population, and we randomly assign one of the K treatments to each of these patients. Thus, the probability that a patient selected in this stage receives treatment k is $p_k^I = \frac{1}{K}$. Note that in the remainder of this article we will use the notation $n(k)$ and $\bar{y}(k)$ for the sample size and sample mean computed over all patients who received treatment k and we will use $y_i(\cdot)$ to denote the observed value for unit i irrespective of the treatment received.

2. In *Stage II* we use the data collected in *Stage I* to select one of the k treatments using some decision procedure δ, and we subsequently subscribe the selected treatment $k = k^*$ to the remaining $N - n$ patients in P. Thus, in stage two we have $p_k^{II} = 1$ if $k = k^*$ and $p_k = 0$ otherwise. In practice this is done by including treatment k^* into our guidelines.

We are interested in the performance of these two-stage allocation policies in terms of its expected outcome per unit when executed in a population of size

N. Thus, we are interested in:

$$\begin{aligned}
\mathbb{E}(\pi_N) &= \frac{\mathbb{E}[\sum_{i=1}^{N} y_i(\cdot)]}{N} \\
&= \frac{\sum_{i=1}^{N}\sum_{k=1}^{K} p_k^{(i)} y_i(k)}{N} \\
&= \frac{\sum_{i=1}^{n}\sum_{k=1}^{K} p_k^{I} y_i(k)}{N} + \frac{\sum_{i=(n+1)}^{N}\sum_{k=1}^{K} p_k^{II} y_i(k)}{N} \\
&= \frac{\sum_{i=1}^{n}\sum_{k=1}^{K} \frac{1}{K} y_i(k)}{N} + \frac{\sum_{i=(n+1)}^{N}\sum_{k=1}^{K} \Pr(k=k^*) y_i(k)}{N} \qquad (1)
\end{aligned}$$

where the expectation is over the random sampling and allocation in *Stage I* and possibly over a random component of the decision procedure δ in *Stage II* that determines the probability that a specific treatment k is selected. In the second line of Equation 1 we use $p_k^{(i)}$ to denote the probability that treatment k is selected for patient i, while in the third line we split up the expectation value of the experiment and the resulting guideline using p_k^I and p_k^{II} respectively since within each stage p_k is a constant. In the last line these probabilities are provided: $p_k^I = \frac{1}{K}$, and $p_k^{II} = \Pr(k = k^*)$ which, with slight abuse of notation, denotes the probability that a specific treatment is selected for inclusion into the guidelines $k = k^*$. Note that for a given population P of size N, when considering a fixed number of treatments K, the value of $\mathbb{E}(\pi_N)$ depends on the choice of n and the specification of $Pr(k = k^*)$, i.e., the probability the decision procedure δ selects treatment k. Hence, in this setting for a given population, $\mathbb{E}(\pi_N) = f(n, \delta)$. Ultimately, we are interested in finding n, given the current approach to δ, such that $\mathbb{E}(\pi_N)$ is maximized.

2.1 Completing the two-stage approach using current RCT practice.

The two-stage allocation policy defined above provides a simplified formalization of our current practice of testing treatments using RCTs. *Stage I* encompasses the RCT itself, and subsequently *Stage II* encompasses the decision to, based on the RCTs results, adopt one of the K treatments [Kaptein, 2018]. The formalization is simplified as we do not consider the common practice of putting prospective treatments k through several rounds of testing [Spiegelhalter et al., 2004, Sedgwick, 2011]. Our conceptual treatment can however easily be extended to such a situation situation as Eq. 1 would still hold but would need to be partitioned into more than two stages. Furthermore, our formalization is simplified in the sense that we do not consider the—relatively common—situation in which new treatments are developed over time, and thus are not available for a subset(s) of patients at some points in time (assuming the patients are treated sequentially) [Robbins, 1985]. Finally, we assume that the population size N is known; this assumption will never be exactly met, but often reasonable estimate can be made in many cases in which for specific diseases incidence rates are known [Dye et al., 1999, Feigin et al., 2003].

To closely relate our two-stage formalization to existing RCT practice, we have to specify the decision rule δ and our choice of the sample size n; indeed, in our current practice these are intimately related. Our decision rule δ is — despite much modern work advocating other approaches [Sedgwick, 2011] — often based on the practice of null hypothesis significance testing: we specify a null hypothesis H_0, and we specify acceptable levels of α and β, the probabilities of making a type I or type II error respectively [Schulz and Grimes, 2005]. Next, we make a statement about a meaningful alternative hypothesis (e.g., the effect size of interest). Given choices for each of these we can, in many situ-

ations, compute the minimal sample size n that controls the error rates given that our assumptions regarding the hypotheses involved are correct. Next, after conducting the trial of size n it is standard practice to compute a p-value and if $p < \alpha$ we reject the null hypothesis and accept the alternative. In practice rejecting the null hypothesis often leads researchers to select the treatment with the highest mean outcome during the trial (thus $k^* = \arg\max_k \bar{y}(k)$) while not rejecting the null often leads researchers to select the current status-quo.[1] Depending on the study design and the choice of α the probability of rejecting H_0 and the probability of selecting treatments k if H_a is accepted are readily provided by standard power calculations. Jointly this completes the specification of the decision procedure δ and hence the specification of p_k^I and p_k^{II} necessary to evaluate Eq.1.

From the analysis above it is clear that in our current practice $\mathbb{E}(\pi_N)$ is defined by our choice of α, β, and our assumptions regarding H_0 and H_a (or the effect size): these jointly define δ and n. However, note that this is not a necessity; even if we stick close to current practice by performing a null-hypothesis significance test we could relax our focus on controlling error rates and rather focus on maximizing $\mathbb{E}(\pi_N)$. A simple method to generate alternative two-stage treatment allocation policies that is very close to current practice would be to keep our standard level of α, keep our standard decision procedure, but determine n such that $\mathbb{E}(\pi_N)$. This can be done by adding to the current assumptions (e.g., H_0 and some estimate of the effect size) an informed estimate of N, the population size. Once all of these are known, we can, for many different designs, evaluate Eq.**??** and select n such that $\mathbb{E}(\pi)$ is maximized. When doing

[1] In our numerical analysis below we assume $\Pr(k = k^*) = c = \frac{1}{K}$ in such cases. This default choice is motivated by the idea that prior to the study, all k arms are equally likely to be superior and hence a random choice after a failed trial seems reasonable. However, in many situations this choice might not be reasonable; e.g., it is unlikely that a placebo is adapted after a failed trial. In such cases one might want to change the `ties` parameter in the `ssev` package (see Section 3).

so the power, $1 - \beta$, will *follow* from the procedure. This is the approach implemented in the package `ssev` we present below.

2.2 Prior work and a motivation for two-stage approaches

Surely, others must have considered treatment allocation policies that maximize the expected outcome of the full allocation procedure as opposed to controlling type I and type II errors within the trial? There is actually a very large literature that considers the analysis of different treatment allocation procedures and indeed focusses on the overall outcome of the procedure (often called *reward* in this literature). This literature on the multi-armed-bandit (MAB) problem—which formalizes the decision problem we described above as a problem in which, sequentially, a gambler selects different arms of a slot-machine, each with a potentially different pay-off, such that she maximizes her rewards—is too large to properly review; we refer the interested reader to Robins [Robbins, 1985] or Gittins, Glazebrook and Weber [Gittins et al., 2011].

In the decades that the MAB problem has been studied, we have been able to bound the expected rewards of distinct policies [Bubeck et al., 2012], and we have developed allocation policies that are asymptotically optimal [Whittle, 1980, Auer et al., 2002]. We have also connected this mostly theoretical literature directly to our practice in clinical trials [Bartroff et al., 2012]. However, the literature on the MAB problem has primarily focussed on allocation policies other than the two-stage policies since any two-stage procedure is provably suboptimal [Bubeck et al., 2012]: optimal solutions to the MAB problem effectively balance exploration (learning the effects of each treatment) and exploitation (selecting the best treatment). Optimal allocation policies smoothly balance these two objectives by—effectively—decreasing $p_k^{(i)}$ smoothly from $\frac{1}{K}$ to 0 for all $k \neq k^*$ as i increases. The exact rate of the decrease depends on the observed data

and the structure of the problem, but any optimal policy will have a smooth decrease as opposed to the step-wise decrease we see in two-stages policies. Effectively, two-stage policies first explore (when $i \leq n$) and subsequently move to exploitation (when $i > n$). This sudden change from exploitation to exploration does not yield an optimal reward, and hence two-stage policies (coined ϵ-first in the MAB literature [Tran-Thanh et al., 2010]), are not considered particularly interesting.

However, despite the fact that they are not (asymptotically) optimal, two-stage treatment allocation policies have a practical benefits over alternative allocation policies that constantly change $p_k^{(i)}$. The two-stage policy is clearly separated into a trial in which all possible treatments are considered, and the subsequent guideline stage in which only one specific treatment needs to be considered. This makes that after the trial we can inform medical professionals of the results of the trial and they do not need to consider alternatives. We can inform patients of the "best" treatment without needing to resort to complex explanations to justify changing probabilities for each patient. And, finally, we can distribute a single treatment (e.g., a medication) to all treatment locations, as opposed to distributing all possible treatments for the (often unlikely) event that a treatments is selected by the policy. These practical benefits of two-stage policies over smooth allocation policies have resulted in a slow uptake of smooth policies in practice[Kaptein, 2018]. Therefore, we focus specifically on two-stage allocation policies and study alternative methods of determining n; the main parameter that drives the step from exploration to exploitation.

Notably, even when focussing solely on two-stage decision procedures that are close to current practice, this work is not the first in its kind: in the 60's a body of theoretical work emerged studying the required sample size when aiming to maximize the expected outcome when choosing between treatments. Initially

work focussed on choosing between two treatments from normal populations with variances known [Colton, 1963]. The work was quickly extended to allowing for multiple stages [Colton, 1965], or multiple treatments [Dunnett, 1960]. Researchers also examined fully sequential allocation [Anscombe, 1963, Cornfield et al., 1969]; an approach closer to the MAB literature. The analysis was further extended to alternative decision rules such as play the winner[Zelen, 1969] and to dichotomous outcomes [Canner, 1970]. These all works convincingly demonstrate the effectiveness gains of including the population size in computations of the sample size, a message we also demonstrate in this work. We deviate from this prior work by focussing more strongly on current RCT practice (i.e., by including a null-hypothesis significance test within the decision procedure a case not included in these prior analyses[2]) and by providing easy to use software to compute sample sizes for comparisons of two treatments.

3 An easy to use [R] package for sample size computation

Instead of focussing on an analytical treatment of different two-stage decision procedures as has been done in prior work [Colton, 1963, Dunnett, 1960], we focus on creating easy-to-use software to compute sample sizes for practical RCT designs while staying close to the current null-hypothesis testing practice. Here we present the `ssev [R]` package that allows researchers to include population sizes in their RCT planning when setting up comparisons between two groups (i.e., $K = 2$) when comparing means (using t-tests with equal variances assumed or not assumed) or proportions.

[2]Prior work mostly uses $k^* = \arg\max_k \bar{y}(k)$; we stay closer to current RCT practice by chosing $k^* = \arg\max_k \bar{y}(k)$ only when H_0 is rejected.

```
Two independent samples t-test (equal variances assumed).

                    Sample size RCT = 64
           Expected mean reward RCT = 0.4503131
                Sample size optimal = 261
      Expected mean reward optimal = 0.4997161

Percentage gain (optimal over RCT): = 10.97081

Sig level: 0.05, power (RCT): 0.8, population size (optimal): 5e+05
NOTE: n is number in *each* group.
```

Figure 1: Example output of the `ssev` package.

The `ssev` package is available on `CRAN`, and is easily installed using the following `[R]` commands:

```
install.packages("ssev")
library(ssev)
```

After installing the package the `compute_sample_size` function is available to compute sample sizes that maximize the expected outcome of the two-stage approach described below for various cases. For example, a call to

```
compute_sample_size(means=c(0,.5), sds=1, N=500000)
```

computes the sample size when comparing two means which are expected to differ by $\frac{1}{2}$, assuming equal variances, $\sigma^2 = 1$ (i.e., Cohen's $d = \frac{1}{2}$) and a population size of $N = 500000$. The call provides the output presented in Figure 1 which shows that using conventional power calculations (with default choices $\alpha = .05$ and $1-\beta = .8$) the traditional RCT would require a sample size of 64 per group, while in this case a sample size that maximizes the expected outcome $\mathbb{E}(\pi_N)$ of a two-stage procedure would require a sample size of 261 per group. When choosing this larger sample size, the expected mean reward of the two-stage procedure over the full population would increase by more than 10%. Table 1 details the arguments to the `compute_sample_size` function.

The `ssev` package computes the desired optimal sample sizes using numerical optimization routines in combination with standard power calculations provided

`means`	A vector of length 2 containing the (assumed) means of the two groups in the case of continuous outcomes.
`sds`	A vector containing the (assumed) standard deviations of the two groups. When only one element is supplied equal variances are assumed.
`proportions`	A vector of length 2 containing the (assumed) proportions of the two groups in the case of dichotomous outcomes.
`N`	Estimated population size.
`power`	Desired power for the classical RCT (i.e. $1 - \beta$).
`sig.level`	Significance level of the test used (i.e., α).
`ties`	Probability of choosing the first group in case of a tie (i.e., in case H_0 is not rejected).
`.verbose`	Whether or not verbose output should be provided, default FALSE.
`...`	further arguments passed on to or from other methods.

Table 1: Arguments for the `ssev` package to compute sample sizes

in earlier `[R]` packages (e.g., the `MESS` and `pwr` packages). The implementation is relatively straightforward: for each design a simple utility function to compute the expected value of the complete two stage procedure as a function of the sample size n is created which implements Equation 1. Computing the expected value of the RCT is straightforward for all designs included in the package (mean comparisons assuming equal or unequal variances, proportion comparisons), but the probabilities of rejecting H_0, and subsequently the probability of selecting one of the $K = 2$ arms given that H_0 is rejected, differ; these are however readily provided using standard power calculation packages. Numerical optimization is then used to evaluate the expected value function for the desired design for values $2 \leq n \leq N$ and select the value of n that maximizes the expected outcome.

4 Numerical analysis when comparing 2 groups

To gain additional understanding of the effectiveness and efficiency of our proposed method we present a number of numerical evaluations. First, we examine

the differences in effectiveness–in terms of expected outcomes—and sample size between the common RCT procedure and our proposed approach. Next, we examine how under- and over-estimates of the population size N affect the computed sample size n.

4.1 Efficiency over current RCT practice

Table 2 presents the difference in expected outcomes—in terms of relative gains—between the common RCT and the method outlined in this paper. We examine three differences in means $d \in \{.2, .5, .8\}$ assuming either equal variances $\sigma_1^2 = \sigma_2^2 = 1$ or unequal variances $\sigma_1^2 = \sigma_2^2 = 9$ and three differences in proportions $p \in \{.1, .2, .3\}$ for different population sizes $N \in \{10^2, 10^3, \ldots, 10^8\}$. It is clear from the table that in all cases, the optimal sample size leads to a higher expected outcome, $\mathbb{E}(\pi_N)$, than current RCT practice with relative differences often exceeding 10%.

	Design	d	10^2	10^3	10^4	10^5	10^6	10^7	10^8
1	Eq. Var.	0.2	5.072	11.969	5.189	10.066	10.935	11.057	11.073
2		0.5	20.578	3.163	9.539	10.811	10.994	11.018	11.021
3		0.8	2.050	6.455	9.977	10.560	10.640	10.650	10.651
4	Uneq. Var.	0.2	1.975	8.444	0.259	7.494	10.545	11.033	11.099
5		0.5	5.750	5.319	6.014	10.237	10.961	11.061	11.074
6		0.8	11.544	0.440	8.441	10.583	10.910	10.953	10.959
7	Prop.	0.1	0.439	1.704	0.359	0.909	1.018	1.034	1.036
8		0.2	3.638	0.064	1.350	1.719	1.776	1.784	1.785
9		0.3	4.363	0.744	1.939	2.178	2.212	2.217	2.217

Table 2: Gain of the optimal procedure over common RCT practice in relative percentages.

Table 3 provides further details: the table shows the differences in the size of a single group (i.e., $n/2$) between the common RCT and the optimal scheme suggested in this paper. It is clear that for small population sizes RCTs often require too large sample sizes (borrowing a term from the MAB literature, in these cases the RCT over-explores), while for large populations the sample sizes

selected using common power calculations are too low (in these cases these studies over-exploit and hence too often choose the wrong treatment to end up in the subsequent guideline).

	Design	d	10^2	10^3	10^4	10^5	10^6	10^7	10^8
1	Eq. Var.	0.2	29	178	-303	-708	-1064	-1400	-1724
2		0.5	26	-34	-101	-159	-213	-266	-316
3		0.8	6	-25	-49	-71	-92	-112	-131
4	Uneq. Var.	0.2	31	285	193	-2169	-4093	-5836	-7493
5		0.5	28	109	-277	-595	-878	-1146	-1404
6		0.8	26	-20	-161	-278	-386	-489	-588
7	Prop.	0.1	11	200	-207	-570	-897	-1209	-1511
8		0.2	29	-13	-105	-186	-262	-335	-406
9		0.3	20	-20	-56	-88	-118	-148	-177

Table 3: Difference in sample size between the choice that maximizes the expected outcome and the traditional RCT. Reported is $n_{rct} - n_{optimal}$; thus, positive entries indicate that the RCT would select a larger sample than the optimal procedure. Clearly, for large populations (e.g., $N > 10^5$) our current RCTs are often too small.

4.2 Robustness to population size estimation.

As a final comparison to gain additional insight into the proposed procedure Table 4 provides the difference in the number of subjects in each group for a trial comparing two means with equal variances ($\sigma^2 = 1$) and different effect-sizes $d \in \{.2, .5, .8\}$ when the size of the population N is over-estimated or under-estimated by 10%. Thus, the first entry of 1 in Table 4 indicates that when the population size of 10^2 is under-estimated by 10% (i.e., it is estimated at 90), versus when it is over estimated by 10% (i.e., at 110) the optimal sample size differs by only one unit per group in this case. Clearly, as sample sizes increase, the effect of a (proportional) error in estimating the sample size increase and the estimated group size is more variable. In the RCT case, in which the difference between the two over- and under-estimation does not depend on the population size N, the results are 160, 26, and 10 respectively. This indicates that for small

population sizes the proposed optimal procedure is less sensitive to erroneous estimates of the population size than the RCT is. For larger sample sizes the optimal procedure becomes more variable to errors in estimating the sample size: this is however easily explained as for large populations the potential benefits of additional experimentation (e.g., a larger n) steadily increase.

	Design	d	10^2	10^3	10^4	10^5	10^6	10^7	10^8
4	Optimal	.2	1	15	202	382	532	672	806
5		.5	1	25	56	80	103	125	145
6		.8	3	15	26	35	44	53	60

Table 4: Comparison of optimal sample sizes in terms of number of subjects per group for varying population sizes.

5 Conclusions and discussion

In this paper we discussed an alternative approach to computing sample sizes in randomized clinical trials and we have provided easy-to-use software package to carry out the procedure. The approach we suggest here considers the trial as merely the first stage of the larger process of allocating treatments to patients which can be split up into two distinct stages: first we learn about the effectiveness of treatments during the trial, and subsequently we select and administer the treatment that was most successful in the trial to the remaining patients by including it in our clinical guidelines. We have motivated that the expected outcome of these two-stage allocation policies depends on the choice of sample size n, and the decision procedure δ that is used when moving from stage Stage I to Stage II. In the current planning of RCTs we often focus on properties of the first stage (in terms of type I and type II errors), and because of this n is fixed for a given decision procedure δ. We suggest relaxing our fixation on the properties of the trial, and subsequently changing the decision procedure δ, such that we can freely choose a sample size n that maximizes the expected

outcome over the full two stage procedure. Admittedly, doing so introduces a need for informed estimates of the population size N when planning a trial. This seems cumbersome as it is something we are not generally used to. However, we would be tempted to argue that for many diseases incidence and prevalence rates—which would allow us to make informed estimates of N—are available.

A lot of prior work has considered alternatives allocation schemes compared to the traditional RCT; we have provided pointers to both the MAB literature—in which fully adaptive allocation schemes are discussed—as well as to earlier results demonstrating the effectiveness of the two stage approach we propose here [Colton, 1963]. We are well aware that the two-stage approach we examine in this work does not actually *maximize* the expected outcome of the sequential allocation of treatments over all units: more flexible allocation policies that constantly change $p_k^{(i)}$ can achieve a higher outcome. However, we believe that two-stage approaches have sufficient practical benefits to, in some cases, be preferred over more flexible sequential allocation procedures [Auer et al., 2002]. Hence, optimizing two-stage allocation policies provides a useful addition to the current literature. Our contribution is primarily of an applied nature; we build on earlier ideas to provide an easy-to-use software package that allows for the computation of optimal sample sizes for a number of common RCT designs.

The current paper also numerically examined the differences between current RCT practice and our suggested approach. Qualitatively, the main results are intuitive: For small populations we need smaller samples, while for larger populations we need larger samples, to maximize our expected outcome. Furthermore, a willingness to make assumptions regarding N improves our robustness to choices of the clinically meaningful effect-size of the treatment d when N is small. However, we have left a number of avenues unexplored: first of all we restricted ourselves to merely varying β; as also α is inherently arbitrary we

might wish to also vary α when computing n in a two-stage allocation policy. Also, despite setting up the problem for arbitrary choice of K, the package `ssev` currently handles only a choice of $K = 2$; we feel this is a meaningful contribution but future work should extend the implemented methods to including more complex designs. Finally, in our treatment of the problem we currently only focus on the direct outcomes and we do not include possible differences in costs between the two stages (the trial might be more expensive to carry out than the guidelines), or plausible variable costs during the second stage: these are welcome extensions to explore in future work. However, for now we hope the current work at the very least inspires those planning out RCTs to consider alternatives to standard power calculations advocated in many introductory text books; easily available alternatives that are close to current practice might provide an accessible step in the direction of more flexible trial planning and sample size computation.

References

[Anscombe, 1963] Anscombe, F. (1963). Sequential medical trials. *Journal of the American Statistical Association*, 58(302):365–383.

[Auer et al., 2002] Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256.

[Bartroff et al., 2012] Bartroff, J., Lai, T. L., and Shih, M.-C. (2012). *Sequential experimentation in clinical trials: design and analysis*, volume 298. Springer Science & Business Media.

[Brakenhoff et al., 2018] Brakenhoff, T., Roes, K., and Nikolakopoulos, S. (2018). Bayesian sample size re-estimation using power priors. *Statistical methods in medical research*, page 0962280218772315.

[Bubeck et al., 2012] Bubeck, S., Cesa-Bianchi, N., et al. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122.

[Canner, 1970] Canner, P. L. (1970). Selecting one of two treatments when the responses are dichotomous. *Journal of the American Statistical Association*, 65(329):293–306.

[Colton, 1963] Colton, T. (1963). A model for selecting one of two medical treatments. *Journal of the American Statistical Association*, 58(302):388–400.

[Colton, 1965] Colton, T. (1965). A two-stage model for selecting one of two treatments. *Biometrics*, 21(1):169–180.

[Cornfield et al., 1969] Cornfield, J., Halperin, M., and Greenhouse, S. W. (1969). An adaptive procedure for sequential clinical trials. *Journal of the American Statistical Association*, 64(327):759–770.

[Cunningham and Johnson, 2016] Cunningham, T. D. and Johnson, R. E. (2016). Design effects for sample size computation in three-level designs. *Statistical methods in medical research*, 25(2):505–519.

[Dunnett, 1960] Dunnett, C. W. (1960). On selecting the largest of k normal population means. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–40.

[Dye et al., 1999] Dye, C., Scheele, S., Dolin, P., Pathania, V., Raviglione, M. C., et al. (1999). Global burden of tuberculosis: estimated incidence, prevalence, and mortality by country. *Jama*, 282(7):677–686.

[Feigin et al., 2003] Feigin, V. L., Lawes, C. M., Bennett, D. A., and Anderson, C. S. (2003). Stroke epidemiology: a review of population-based studies of incidence, prevalence, and case-fatality in the late 20th century. *The Lancet Neurology*, 2(1):43–53.

[Freiman et al., 1978] Freiman, J. A., Chalmers, T. C., Smith Jr, H., and Kuebler, R. R. (1978). The importance of beta, the type ii error and sample size in the design and interpretation of the randomized control trial: Survey of 71 negative trials. *New England Journal of Medicine*, 299(13):690–694.

[Gittins et al., 2011] Gittins, J., Glazebrook, K., and Weber, R. (2011). *Multi-armed bandit allocation indices*. John Wiley & Sons.

[Kaptein, 2018] Kaptein, M. C. (2018). *Computational Personalization: Data science methods for personalized health*.

[Qiu et al., 2016] Qiu, S.-F., Poon, W.-Y., and Tang, M.-L. (2016). Sample size determination for disease prevalence studies with partially validated data. *Statistical methods in medical research*, 25(1):37–63.

[Robbins, 1985] Robbins, H. (1985). Some aspects of the sequential design of experiments. In *Herbert Robbins Selected Papers*, pages 169–177. Springer.

[Rubin, 2004] Rubin, D. B. (2004). Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics*, 31(2):161–170.

[Rubin, 2005] Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331.

[Schulz and Grimes, 2005] Schulz, K. F. and Grimes, D. A. (2005). Sample size calculations in randomised trials: mandatory and mystical. *The Lancet*, 365(9467):1348–1353.

[Sedgwick, 2011] Sedgwick, P. (2011). Phases of clinical trials. *BMJ: British Medical Journal (Online)*, 343.

[Shan, 2018] Shan, G. (2018). Sample size calculation for agreement between two raters with binary endpoints using exact tests. *Statistical methods in medical research*, 27(7):2132–2141.

[Spiegelhalter et al., 2004] Spiegelhalter, D. J., Abrams, K. R., and Myles, J. P. (2004). *Bayesian approaches to clinical trials and health-care evaluation*, volume 13. John Wiley & Sons.

[Tran-Thanh et al., 2010] Tran-Thanh, L., Chapman, A., Cote, E. M. d., Rogers, A., and Jennings, N. R. (2010). ε-first policies for budget-limited multi-armed bandits. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, pages 1211–1216. AAAI Press.

[Whitehead et al., 2016] Whitehead, A. L., Julious, S. A., Cooper, C. L., and Campbell, M. J. (2016). Estimating the sample size for a pilot randomised trial to minimise the overall trial sample size for the external pilot and main trial for a continuous outcome variable. *Statistical methods in medical research*, 25(3):1057–1073.

[Whittle, 1980] Whittle, P. (1980). Multi-armed bandits and the gittins index. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 143–149.

[Zelen, 1969] Zelen, M. (1969). Play the winner rule and the controlled clinical trial. *Journal of the American Statistical Association*, 64(325):131–146.

[Zhu et al., 2017] Zhu, H., Zhang, S., and Ahn, C. (2017). Sample size considerations for split-mouth design. *Statistical methods in medical research*, 26(6):2543–2551.