Sample Size and Screening Size Trade Off in the Presence of Subgroups with Different Expected Treatment Effects

Kyle D. Rudser[†], Edward Bendert[‡], Joseph S. Koopmeiners[†]

[†]Division of Biostatistics, School of Public Health, University of Minnesota, 420 Delaware St. SE, Minneapolis, Minnesota, 55455, U.S.A.
[‡]Statistics Collaborative, 1625 Massachusetts Ave., NW; Suite 600 Washington, DC 20036, U.S.A.

Abstract

Statistical study design considerations typically focus on sample size, power, and a single population treatment effect given a fixed significance level (generally 0.05). Eligibility criteria is formulated to select the patient population of interest to be studied for which the magnitude of the treatment effect is expected to hold. In some instances researchers may expect there to be subgroups such that the treatment is expected to have the largest effect in one group while the others will exhibit an attenuated effect. Identification of these subgroups can be based on a clinical decision rule, e.g., biomarker cutoff, but may not be precise, i.e., sensitivity and specificity are not simultaneously at 100%. In the context of subgroups with different expected treatment effects, screening procedures may be adjusted to different levels of sensitivity and specificity to detect those patients in the subgroup with the greatest expected treatment effect. As a result, depending on the corresponding positive predictive value, the sample size required, power, and/or treatment effect expected to hold for the study will change. We evaluate the impact on design operating characteristics of power and sample size, and illustrate scenarios where overall trial duration may be shortened.

Keywords: Sample size; Clinical trial design; Biomarker; Heterogeneous effects; Operating characteristics

1 Introduction

In a typical study design, e.g., a new therapy for a particular patient population of interest, considerations revolve around a treatment effect to detect (i.e., alternative), variability of the statistic used, and power for a level of type I error (usually set at 0.05). Logistically, one needs to balance the sample size, treatment effect, and power. Researchers need enough patients to detect a clinically meaningful difference between treatment groups with a predetermined level of power, in a reasonable time frame. Formulas can be generated articulating the relationships between these parameters based on a single treatment effect for the overall group.

When conducting a randomized controlled trial, there are many measurements used as eligibility criteria. In some instances it may be the case that there are two distinct groups of patients screened;

one subgroup where patients are expected to exhibit the greatest treatment effect ("optimal" group) and another where patients are expected to have an attenuated treatment effect ("suboptimal" group). There may be a number of current medical conditions or general health measurements taken on the patients that can be used in a medical decision rule to suggest if they will respond well to the treatment at hand based on prior studies. One such measurement may be a biomarker: a scientific measurement used as a predictor or indicator of a biological state, e.g., C-reactive protein in blood as a predictor of cardiovascular disease.

Patients in the suboptimal group are expected to have an attenuated treatment effect relative to the optimal group, but are also under consideration for enrollment, either intentionally or inadvertently due to imprecise enrollment criteria. When the optimal and suboptimal groups can be distinguished precisely and immediately, we may choose to conduct a trial in only the optimal group. Although the sample size for the trial will be minimized, the length of time to screen enough patients to complete enrollment may be prohibitive and restricted inclusion may lessen generalizability. When the enrolled group is comprised of a mixture of optimal and suboptimal populations, we may choose to evaluate 1) the effect in the combined population, 2) the effect in either subpopulation (with control for experimentwise type I error), 3) the effect in both subpopulations separately, i.e., the alternative is an effective treatment in both, or 4) a differential effect, i.e., interaction. Each of these have different advantages/disadvantages and goals underlying them.

The primary focus here is when interest is in evaluating the effect in the combined population and screening criteria may be imprecise in identifying optimal and suboptimal groups. In this setting, there is a tradeoff in the design of the study between having a smaller group of patients with a large treatment effect (restricted enrollment criteria), or a larger group of patients with a small treatment effect (broader enrollment criteria). Investigators want to include as many patients from the optimal group as possible to minimize the sample size required to detect a significant effect between the treatment groups. However, investigators may also be faced with low prevalence of the optimal group, to the degree that the length of time needed to enroll the required number may be logistically unrealistic. In this case, researchers may be able to move forward more quickly by broadening enrollment criteria, thereby enrolling more patients of the suboptimal group. Enrolling patients from the suboptimal group would result in lower power for the study due to a smaller effect expected in a group of patients that includes some with an attenuated treatment effect. To maintain power, the sample size would need to be increased in order to counter the smaller expected overall treatment effect.

1.1 Overview

We examine the relationships between parameters of sample size, power, treatment effect (design alternative), and duration, as well as the specificity and sensitivity of different screening procedures. We also illustrate scenarios where overall trial duration may be shortened. In order to do this we first define design parameters and introduce notation in section 2, then explain simulation methods in section 3. We then illustrate the impact attenuation has on power while holding sample size constant in section 3.1, and evaluate the additional sample size required to maintain power in section 3.2. In section 3.3 we look at trial duration by examining the number of patients needed to screen in order to enroll the necessary number of patients. Lastly, we present considerations in the context of a mean variance relationship in section 4 and conclude with discussion in section 5.

2 Design Parameters and Notation

Trial operating characteristics of power and sample size can be calculated based on the following sample size formula (Emerson 2003):

$$N \quad = \quad \frac{\delta^2_{\alpha,\beta}\, V}{\Delta^{*2}}, \tag{1}$$

where V denotes the variability contributed by each sample unit to the statistic and $\delta_{\alpha,\beta}$ represents a function of the critical values for testing hypotheses with type I error α and type II error β. The symbol Δ^* is the alternative that the study has power $1-\beta$, which often represents the expected overall treatment effect for the study. This general formula determines the sample size for each arm in a trial; without loss of generality, we focus on two arm study designs here. For calculations based on a normally distributed statistic, $\delta_{\alpha,\beta} = z_{1-\alpha/2} + z_{1-\beta}$, as will be the case for results included here. A comment on the use of other quantiles, e.g., those of a t distribution, is included later.

The overall expected treatment effect will vary depending on the performance of the screening procedure for the trial. Most often, screening procedures are some form of physiologically based decision rule to differentiate between subgroups of the patient population being screened for enrollment, e.g., based on inclusion and exclusion criteria. For the setting examined here, we have optimal and suboptimal groups of patients that we, perhaps imprecisely, identify through our screening procedure, e.g., through a biomarker with a corresponding sensitivity and specificity. For ease of discussion, the remainder of this manuscript will use 'biomarker' to denote any medical decision rule to differentiate between optimal and suboptimal groups for trial enrollment.

Let F_{op} denote the distribution of the biomarker in the optimal group. We expect the treatment to have the greatest effect among these patients. Ideally, these patients will be plentiful as then we would be able to conduct our trial quickly, i.e., our screening procedure would enroll patients at a high rate and fewer patients would be needed to demonstrate a significant effect if these patients were enrolled exclusively because these patients are expected to have a larger treatment effect than those of the suboptimal group. In an analogous fashion, let F_{so} denote the distribution of the biomarker in the suboptimal group. Then F_{all} is the distribution of the biomarker in the entire population:

$$F_{all} \quad = \quad (prevalence)F_{op} + (1 - prevalence)F_{so}. \tag{2}$$

The cut point of the biomarker that is chosen for determining who is enrolled in the trial and who is not is an important decision. Let the screening cut point of the biomarker be denoted by c. Without loss of generality, consider patients with biomarker values greater than c will be enrolled while those with values less than c will not. For each value of c, let u denote the corresponding inverse of the cumulative distribution function F_{all}, which is the proportion of the overall population below the cut point c. For example, if $u = 0.9$ and $F_{all}^{-1}(0.9) = 2.71$, then 10% of the values of the biomarker fall above 2.71, and these 10% of patients would be enrolled in the trial. In these terms, $1 - u$ is the proportion of screened individuals enrolled, and u is the proportion of screened individuals not enrolled. For the purposes of discussion here, we presume all patients meeting enrollment criteria are enrolled, e.g., consent is obtained prior.

Each $c(u)$ has a corresponding sensitivity and specificity, and therefore a corresponding positive predictive value depending on the prevalence of the optimal group. Note that sensitivity and specificity are with respect to the trial screening procedure discriminating between patients in the optimal and suboptimal groups. Sensitivity is the probability a patient is enrolled given they are from the optimal group and specificity is the probability a patient is not enrolled given they are from the suboptimal group:

$$\begin{aligned} sensitivity(u) &= \int_{c(u)}^{\infty} f_{op}(t)dt = 1 - F_{op}(c(u)), \quad (3) \\ specificity(u) &= \int_{-\infty}^{c(u)} f_{so}(t)dt = F_{so}(c(u)). \quad (4) \end{aligned}$$

The positive predictive value (PPV) of our screening procedure is then the probability that a patient is from the optimal group, given they were enrolled:

$$PPV(u) = \frac{(sensitivity(u))prevalence}{(sensitivity(u))prevalence + (1 - specificity(u))(1 - prevalence)}. \quad (5)$$

The resulting expected treatment effect for the enrolled study population is expressed as follows:

$$\Delta^*(u) = PPV(u)\Delta_1 + (1 - PPV(u))a\Delta_1, \quad (6)$$

where the treatment effect in the optimal group is denoted as Δ_1, and the treatment effect in the suboptimal group is attenuated by a factor of a. This also will influence the variance:

$$V(u) = PPV(u)\sigma_1^2 + (1 - PPV(u))\sigma_a^2 + PPV(u)(1 - PPV(u))(1 - a)^2\Delta_1^2 \quad (7)$$

where σ_1^2 and σ_a^2 denote the variability of a sample unit from the optimal and suboptimal groups respectively.

An example of possible distributions for the optimal group, suboptimal group, and overall are presented in Figure 1. The overall distribution represents the mixture of the optimal and suboptimal group that is screened for the trial. In this hypothetical example, the distributions were arbitrarily set as $N(0, 1)$ and $N(2, 1)$ for the suboptimal and optimal groups respectively, with a prevalence of 40%.

Again, the degree of attenuation between the optimal and suboptimal groups is denoted as a scalar value a where lower values of a indicate a larger degree of attenuation. Values of $a < 0$ represent a harmful treatment effect in the suboptimal group, while $a > 1$ would reflect a scenario where the suboptimal group has a greater treatment effect than the optimal group. Therefore, we restrict attention to values between 0 and 1, e.g., if $a = 0.5$, we would expect the treatment effect in the suboptimal group to be half that of the treatment effect in the optimal group, and if $a = 0$, we would expect there to be no treatment effect in the suboptimal group. Based on these parameters we can calculate the sample size (or power) from equation 1 and may characterize them as functions of u. We can also calculate the number of patients needed to screen in order to successfully enroll the required number of patients:

$$ScN(u) = \frac{N(u)}{1-u} = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2 V(u)}{\left[\Delta^*(u)\right]^2 (1-u)} \tag{8}$$

where $\Delta^*(u)$ and $V(u)$ are as defined in equations 6 and 7 above.

3 Simulations

Throughout the simulations, α was set to 0.05 and the treatment effect for the optimal group was set to 0.4. The variance was set to 1 and will initially be independent of the mean, but scenarios where this may not be the case are discussed further in section 4. The overall treatment effect expected for the trial, $\Delta^*(u)$, depends on the inclusion criteria cut-off and is a weighted average of the optimal treatment effect and the suboptimal treatment effect, as seen in equation 6.

For evaluating trial duration, the distributions for the optimal and suboptimal groups, F_{op} and F_{so} were taken to be normal with variance 1 and different locations. It should be noted that features of the normal distribution are not central to any calculations and that choice of distribution family is arbitrary. Instead, they represent a degree to which the distributions of the optimal and suboptimal groups overlap and how far apart the "centers" of their distributions are. For PPV the key aspects are the prevalence, initially set to 0.4 and then varied, and $1 - F_{op}(c(u))$ relative to $1 - F_{so}(c(u))$. We presume without loss of generality that higher values of the biomarker are indicative of patients from the optimal group. For each value of u (or equivalently c) we can determine the corresponding sensitivity and specificity (equations 3 and 4), the positive predictive value given prevalence (equation 5) and the study effect size $\Delta^*(u)$ (equation 6). We can also, after specifying type I error α and type II error β, calculate the sample size required (equation 1) and subsequently the number of patients needed to screen in order to enroll that sample size (equation 8). In summary, in addition to α and β, we have the following variables for the calculation of sample size and number of patients needed to screen:

1. Prevalence of optimal subgroup
2. u (hence c; influenced by F_{op}, F_{so}, and prevalence of optimal group)
3. Sensitivity of screening procedure to identify optimal subgroup (dependent on u and F_{op})
4. Specificity of screening procedure to identify suboptimal subgroup (dependent on u and F_{so})
5. Effect size of optimal group
6. Degree of attenuation in suboptimal group relative to the optimal group

3.1 Power for Fixed Sample Size

The level of power will be affected by the proportion of individuals enrolled in the trial who are from optimal and suboptimal groups (decreased for any $a < 1$) when maintaining a fixed sample size. Based on a fixed sample size of $N = 50$, and power of 80% ($\Delta_1 \approx 0.396$), Figure 2 displays the power across the full range of PPV for varying degrees of attenuation. As expected, higher attenuation causes a larger decrease in power to detect a treatment effect of $\Delta^*(u)$. It is worth noting that the area of particular interest for this type of graph is that for which the PPV is greater than the prevalence. This is consistent with presuming the screening criteria used to identify those of the optimal population is at least as good as flipping a coin (which has PPV=prevalence).

The lower portion of Figure 2 summarizes the loss in power for a range of high positive predictive values between 0.75 and 1.0 and reflects a meaningful impact that can be observed with PPV not far from 1.0. For a positive predictive value of 0.8, and attenuation factor of 0.4, there is a 11.05% loss of power compared to when there is no attenuation. This is a substantial loss in power. The last column reflects no loss in power for any level of attenuation when the positive predictive value is equal to 1, i.e., when only those from the optimal group are enrolled. In analogous cases for power set to 90% and 97.5%, there is a loss in power of 9.07% and 4.72% respectively, for predictive value of 0.8 and an attenuation factor of 0.4 (data not shown).

3.2 Sample Size for Fixed Power

In order to avoid a loss in power for the expected treatment effect, the sample size needs to be adjusted (increased for any $a < 1$) depending on the proportion of individuals from the suboptimal group that are enrolled in the trial. Figure 3 presents the necessary sample size modification to maintain power over varying levels of attenuation and the full range of positive predictive values. As the positive predictive value increases, the ratio of sample sizes necessarily converges to 1. This is because larger positive predictive values correspond to a larger proportion of the enrolled patients being from the optimal group, meaning fewer patients will be required to maintain power.

It is also apparent by this graph that as the degree of attenuation increases (a becomes smaller) the number of patients required to maintain power increases dramatically. For example, if the positive predictive value is 0.5, i.e., half of enrolled patients come from the optimal group, and $a = 0$, we will need four times as many patients to maintain power. This is an expected result from the stand point that this represents the scenario where the alternative for the trial is half of that compared to when there is no attenuation. As can be seen in equation 1, the relationship between treatment effect and sample size goes by the square of the inverse, hence four times the sample size is required to maintain the power for a treatment effect half as big.

The lower portion of Figure 3 highlights the sample size inflation factor to maintain power over a range of high PPV between 0.75 and 1.0. The PPV of screening procedures does not need to be very far from 1.0 before meaningful repercussions are observed. For the situation where the attenuation factor is 0.4 and PPV is 0.80, we would need to enroll 29% more patients than if there was no attenuation at all. This is a significant increase and reflects the strong effect an attenuated group can have on the number of patients required to maintain power. The relationship of the ratio of sample sizes between no attenuation and attenuated cases will be identical for any level of power and equal to:

$$\frac{PPV\sigma_1^2 + (1 - PPV)\sigma_a^2 + PPV(1 - PPV)(1 - a)^2\Delta_1^2}{\left(PPV\sigma_1^2 + (1 - PPV)\sigma_a^2\right)\left(PPV + (1 - PPV)a\right)^2} \tag{9}$$

For each of the tabled cells in Figure 3, there is an associated expected treatment effect for the resultant enrolled study population. A high degree of attenuation, combined with a low positive predictive value may result in an expected treatment effect that is no longer clinically relevant. E.g., if the positive predictive value is 0.2, and the attenuation factor is also 0.2, the expected treatment effect for the study will only be 0.14, based on a treatment effect of 0.4 in the optimal group, which may no longer constitute a clinically meaningful difference and therefore would not be a viable trial design. This potential result needs to be kept in mind when planning a trial and considering inclusion of subgroups with an attenuated effect.

3.3 Trial Duration

The past two subsections have aimed to quantify the influence an attenuated treatment effect has on trial operating characteristics of power and sample size. Due to the dependence on the performance of the screening procedure, we have presented the results across a range of possible positive predictive values. Operationally, control over the positive predictive value of our screening procedure is realized through the cutoff value c, or equivalently u, (the proportion of the screening population with a biomarker value below c). By changing this cutoff value u, we induce changes in the sensitivity and specificity of our screening procedure. Up to this point we have focused on the required number of enrolled patients and not the number required to screen (a surrogate for the time required to complete a trial). It may, however, be the case that patients in the optimal group are scarce and choice of a cutoff with high positive predictive value is prohibitive, making broader enrollment criteria needed. If that were the case, although broader criteria will result in enrolling a larger proportion of screened patients (potentially reducing the time required to complete study enrollment), at the same time, more patients will be required to be enrolled to maintain power of the expected treatment effect (a counter influence to shortening the time required to complete the study).

For larger values of u (and necessarily c), criteria for entering the study is more stringent, and the sensitivity will decrease while the specificity will increase. That is, as the requirement to enter the study is raised, fewer patients in the optimal group will be enrolled, but there is also a corresponding affect of avoiding those in the suboptimal group. The positive predictive value increases monotonically as the u quantile increases with the minimum positive predictive value equal to the prevalence (at $u = 0$). The overall expected effect size for the study is the same as was expressed in equation 6, although we now note the positive predictive value is a function of u.

As u increases, the specificity increases, making our sample increasingly pure, causing the expected effect size to increase toward 0.4, the treatment effect in the optimal group (Figure 4(a)). As a result, increased sample size is needed depending on the value of u (Figure 4(b)).

Related to the sample size of Figure 4(b) is the number of patients needed to screen in order to enroll a given sample size (equation 8, Figure 5). As the attenuation factor a decreases from 1 to 0, there is a certain point where the minimum number of patients needed to screen is no longer at $u = 0$, i.e., including everyone screened. In order to find this value of a, we take a partial derivative with respect to u to get the first derivative in terms of the attenuation factor a given all of the other values are fixed. We then set this equation equal to zero and find the maximum value of a such that there exists a solution to the following equation:

$$\frac{\partial}{\partial u}\left[\frac{(z_{1-\alpha/2}+z_{1-\beta})^2 V(u)}{\left[\Delta^*(u)\right]^2(1-u)}\right] = 0. \tag{10}$$

For the example highlighted here, there exists a solution to this equation for values of a between 0 and 0.295 (evaluated numerically). This means if we expect the treatment effect in the suboptimal group to be at least 29.5% of the treatment effect in the optimal group, enrolling all patients regardless of their biomarker value will result in the least number of patients screened. However, in the event the treatment effect in the suboptimal group is suspected to be less than 29.5% of that in the optimal group, investigators can minimize the number screened by restricting what levels of the biomarker are allowed to be entered into the study. This can save calendar time by requiring fewer patients screened.

Figure 6 displays a heatmap across u and a where the shading corresponds to the number of patients needed to screen in order to enroll the necessary number of patients to maintain power for the expected treatment effect for the study. The distribution of the optimal group and suboptimal group, prevalence, and Δ_1 remain the same as in previous sections. When a is 0.1, there is a value of $u \neq 0$ for which a minimum number of patients are required to be screened. This can be seen by following a straight line horizontally at $a = 0.1$. We see that at $(u = 0,\ a = 0.1)$ the trial would require approximately 250 patients to be screened. As u increases, the required number to screen drops at first, down to approximately 175 patients before rising again. This is a 30% decrease in patients to screen (hence time to complete enrollment for a study) by adjusting the eligibility criteria.

This plot also shows that any value of u above approximately 0.8 would be disadvantageous to select as a cut point for the biomarker from the standpoint that a very large number of patients would be needed to screen in order to obtain the necessary sample size. Referring to equation 8, it can be seen that the expression for the number of patients needed to screen has $1 - u$ in the denominator. Thus, for $u = 0.8$, on average for every person enrolled, 5 people are needed to be screened. If u is increased further to 0.9, (halving $1 - u$), the result is a doubling of the number of patients needed to screen. This explains the rapid increase in number of patients needed to screen as u approaches 1.

When considering changing enrollment criteria, researchers need to be cognizant of the resulting expected treatment effect for the study. Presumably there is a minimum scientifically relevant treatment difference, and it may be the case that characteristics of the screening procedure combined with the attenuated treatment effect of the optimal group will result in an irrelevant expected treatment effect. The two solid black lines in Figure 6 represent contours for two alternatives. The outer most line represents the points where $\Delta^* = 0.25$, and the inner most line represents the points where $\Delta^* = 0.2$. For example, if it were the case that the minimal scientifically relevant effect size is 0.2, then the inner line represents a cutoff for combinations of a and u that correspond to a scientifically viable study design. One may subsequently calculate the range of valid $c(u)$ to examine the support of the biomarker distribution, which may serve as a basis for generalizability of the patient population in the trial.

The series of Figure 7 diagrams illustrate the scenario where the centers of the optimal and suboptimal distributions become increasingly farther apart, i.e., less overlap in their underlying distributions, which represents a scenario where the decision rule (e.g., biomarker) used for screening has better discriminatory performance between the optimal and suboptimal groups. For this situation it was still assumed the prevalence $= 0.4$, $\Delta_1 = 0.4$, and the variance of the biomarker distributions was held constant at 1. Figures 7(a):7(d) indicate the corresponding number of patients needed to screen (Figure 7(b) is identical to Figure 5) while Figures 7(e):7(h) illustrate the underlying distributions. We see that as the distance increases to a situation where there is essentially complete separation, (Figure 7(d) and 7(h)), the contours representing different values of the attenuation factor a converge at a point where $u = 0.6$, and overlap each other for $u > 0.6$. This is a result of the prevalence being 0.4. As the underlying distributions become separated, any value of $u > (1 - \mathit{prevalence})$ will result in only patients from the optimal group being enrolled into the trial, hence the value of a has no impact for those values of u. For values of $u < 0.6$, the number of patients needed to screen varies more dramatically across the a contours than in the situation where the distributions overlap. This is because as u decreases beyond the lower bound of supports for the optimal group in the complete separation example, any additional patients allowed to be enrolled will only be from the suboptimal group. There is no chance we may also be including additional patients from the optimal group as well (they are all already captured). Therefore the attenuation will be more dramatic, and we will need more patients to compensate.

While Figures 7(a):7(h) examine the affect increasing the separation of the distributions has on number of patients to screen, Figures 8(a):8(h) examine changes in prevalence for values of 0.2, 0.4, 0.6, and 0.8 respectively where Figures 8(e):8(h) show underlying distributions in each of these scenarios. When there is low prevalence and dramatic attenuation for the suboptimal group, e.g., $a = 0$, the number of patients needed to recruit will dramatically increase (Figure 8(a)). However, as the prevalence increases, the impact of different attenuation factors decreases.

Between the series of plots in Figure 7 and in Figure 8 it is clear that the range of values of a for which a minimum number to screen exists for $u \neq 0$ depends on the prevalence of the optimal group and also the degree of dissociation between optimal and suboptimal biomarker distributions. The bounds on a have been evaluated numerically, which are also influenced by the power, although to a lesser degree (Figure 9 The largest value of a for which there is a minimum number to screen for $u \neq 0$ occurs when there is complete separation of biomarker distributions and when the optimal group has higher prevalence. Figure 9(b) presents these values across a range of prevalence from 5% to 95% and illustrates the power level of interest will impact the result, but not to a great degree, particularly for low prevalence.

4 Mean-Variance Relationship

To this point, we have considered scenarios in the absence of a mean-variance relationship. We now comment on the impact such a relationship can have on the previously mentioned design considerations. In the setting of a mean-variance relationship, the variance will depend on Δ^*. As such, we may adjust equation 1 as follows:

$$N = \frac{\left(z_{1-\alpha/2}\sqrt{V_0(\Delta^*)} + z_{1-\beta}\sqrt{V_1(\Delta^*)}\right)^2}{\Delta^{*2}} \tag{11}$$

where $V_0(\Delta^*)$ and $V_1(\Delta^*)$ denote the variance, as a function of Δ^*, under the null and alternative hypotheses respectively. It is useful to distinguish between a mean-variance relationship and heteroskedasticity. We recognize that a treatment may not have an impact on the summary measure of interest, e.g., the mean, yet still change the variability of our estimate of the summary measure of interest. With heteroskedasticity, the variance of the summary measure of the mixture of patients from optimal and suboptimal groups need not equal the variance among those from the optimal group alone. We will leave this aside for the discussion here as the principles are the same.

Focus will be entirely on $V_1(\Delta)$ because under the null, the optimal and suboptimal groups have the same magnitude of effect, i.e., treatment effect of zero. Based on equation 11, we can see that the effect will depend on the direction of the mean-variance relationship. If the variance is a decreasing function in the mean, then the results shown in the previous sections will be magnified as the numerator will be increasing ($V_1(\Delta^*) > V_1(\Delta_1)$) along with the denominator decreasing ($\Delta^* < \Delta_1$) in the degree of attenuation. If instead the variance is an increasing function in the mean, e.g., in the case of a Poisson distribution where a higher mean corresponds to higher variance, then there is a trade-off. The impact relative to having no attenuated effect will be muted to the degree that the attenuated variability may offset the attenuated magnitude of effect.

4.1 Binomial Proportions

In the case of using a difference of two proportions, we have $\Delta^* = p_1^* - p_0$, $V_0 = 2p_0(1-p_0)$ and $V_1(\Delta^*) = p_0(1-p_0) + p_1^*(1-p_1^*)$ to be used in equation 11. In this setting, $V_1(\Delta)$ can be increasing or decreasing in Δ depending on p_0 and p_1^*. For example, if $p_0 = 0.1$ and $p_1 = 0.3$, attenuation from a suboptimal group will cause the variance to decrease. Alternatively, for $p_1 = 0.9$, and $p_0 = 0.7$, attenuation will cause the variance to increase. This is because the maximum variance of a Bernoulli random variable occurs at $p = 0.5$.

Instead of looking at an outcome of a difference of two proportions, it may be be the case that investigators are instead interested in an odds ratio or relative risk. Here also, depending on the values of the proportions the investigators use in planning their study, the variance may not be changing similarly in all scenarios. Overall, if the variance decreases with larger alternatives (increases with larger attenuation), the effect attenuation has as described in previous sections will be magnified. If instead the variance increases with larger alternatives (decreases with larger attenuation), then the degree to which the variance decreases along with the magnitude of effect will impact the extent to which these may offset each other and diminish the effect described in previous sections.

5 Discussion

In the presence of a patient population with subgroups that have different treatment effects, eligibility criteria and screening procedures designed to target one group over the others influence sample size, power, difference to detect, and number of patients needed to screen. It is important researchers know how many patients they need to recruit: not enough patients will lead to lack of evidence, and too many patients will be inefficient. We evaluated the impact of the design operating characteristics of power and sample size, and illustrated scenarios where overall trial duration may be shortened. Allowing patients with a suboptimal treatment effect to enroll will lower power to detect a particular alternative for a given sample size with higher attenuation resulting in a larger loss in power. In order to remediate this, a larger sample size will be required. Furthermore, depending on the degree of attenuation in the suboptimal group and positive predictive value of screening procedures, the minimum number of patients to screen can be evaluated and has been shown not to be trivial. For scenarios where the subgroup is expected to experience a moderately attenuated treatment effect, the number of patients required to screen was the least when all screened patients were enrolled. This was not the case for larger degrees of attenuation. Results will depend on the underlying distributions of the subgroups and prevalence of the optimal group. As such, evaluation for each specific context is warranted and may be conducted following the outline presented here. As discussed in sections 3.2 and 3.3, care needs to be exercised to be sure inclusion of a larger proportion of patients with an expected attenuated effect will not result in a study design for an effect that is not meaningful scientifically.

It should be noted that the sample size formula used throughout is quite general and pertains to any normally distributed statistic, e.g., estimates for log odds ratios or regression coefficients (Emerson 2003). As often the case in analyses and hence sample size/power estimates, asymptotic arguments are made for the use of such formulas without reliance on normally distributed data. At times, some may choose to use other distributional quantiles, e.g., t quantiles. This will lead to quantitative differences from the results presented here in some places, e.g., loss in power will be of a different magnitude, but qualitatively the results would be the same.

There is a tradeoff between allowing more patients into the trial resulting in a diminished treatment effect expected to be observed, or allowing only those with the greatest treatment effect and having a smaller sample size. Changing the patient population may also impact the generalizability and/or safety profile. The risk-benefit ratio as part of the study design process cannot be ignored, especially in patients who are likely to experience an attenuated treatment effect. If the safety profile is identical for both the optimal and suboptimal groups, the risk-benefit trade-off is necessarily different because the suboptimal group is expected to have an attenuated effect. As such, it may no longer be ethically appropriate for these patients to be enrolled, either absolutely or relative to another treatment available that is potentially more efficacious, thereby presenting a better risk-benefit. These considerations are needed to be evaluated on a trail by trial basis.

For a study design with a survival endpoint, statistical information is generally tied to the number of events observed rather than the number of patients. In this case, there is an additional aspect to the tradeoff between number of patients enrolled and calendar time because statistical information is not immediately obtained with the enrollment of an individual. If many patients were enrolled immediately, it may affect our ability to assess time-varying effects, such as whether there is greater magnitude of effect earlier or after some delay. This survival endpoint scenario is more complex and warrants further investigation.

The focus of discussion here has been on fixed sample designs. We recognize that often the economic and/or ethical influences will impact the design of trials and in some cases demand a sequential analysis to meet those needs, e.g., through group sequential designs (Jennison and Turnbull 2000; Emerson, Kittelson, and Gillen 2007) or adaptive designs (Tsiatis and Mehta 2003; Jennison and Turnbull 2006). For these designs, it is often the case that the study will be powered for the minimal clinically meaningful alternative and by design will stop early when the observed effect is sufficiently large at an interim analysis. For the scenario presented here, with optimal and suboptimal subgroups, the powered alternative would be unchanged (an attenuated expected treatment effect does not change what is minimally clinically meaningful) but the trial operating characteristics will be different because at the same interim analysis where a trial among only patients from the optimal group would stop early for efficacy (or not stop for futility), the trial that incorporates a larger proportion from the suboptimal group may not stop for efficacy (or stop for futility) due to an attenuated treatment effect at that time. This deserves more consideration as well, but is beyond the scope here.

References

Emerson, S. S. (2003). S+seqtrial technical overview. *Technical Report, Insightful Corporation, Seattle, Washington*.

Emerson, S. S., J. M. Kittelson, and D. L. Gillen (2007). Frequentist evaluation of group sequential clinical trial designs. *Statistics in Medicine 26*, 5047–5080.

Jennison, C. and B. W. Turnbull (2000). *Group Sequential Methods With Applications to Clinical Trials*. CRC Press.

Jennison, C. and B. W. Turnbull (2006). Adaptive and non-adaptive group sequential tests. *Biometrika 93*(1), 1–21.

Tsiatis, A. A. and C. R. Mehta (2003). On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika 90*, 367–378.

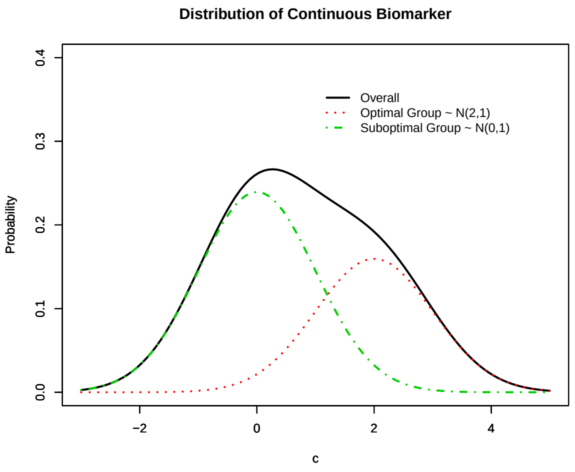
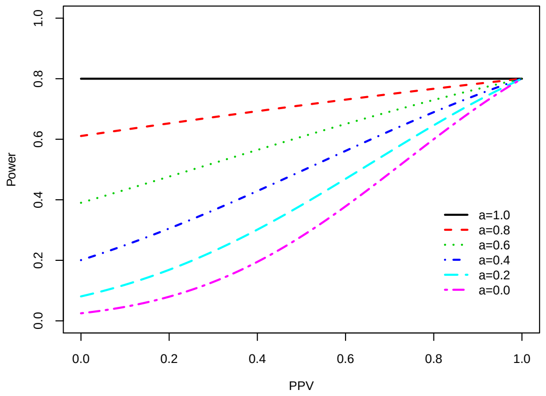
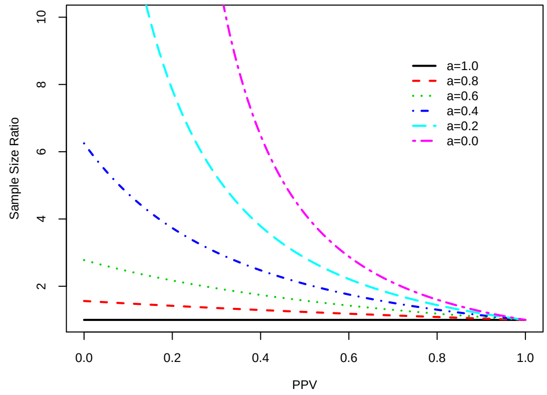


Figure 1: Example of possible distributions of a biomarker in a patient population with optimal and suboptimal groups. Prevalence of the optimal group here is 40%.



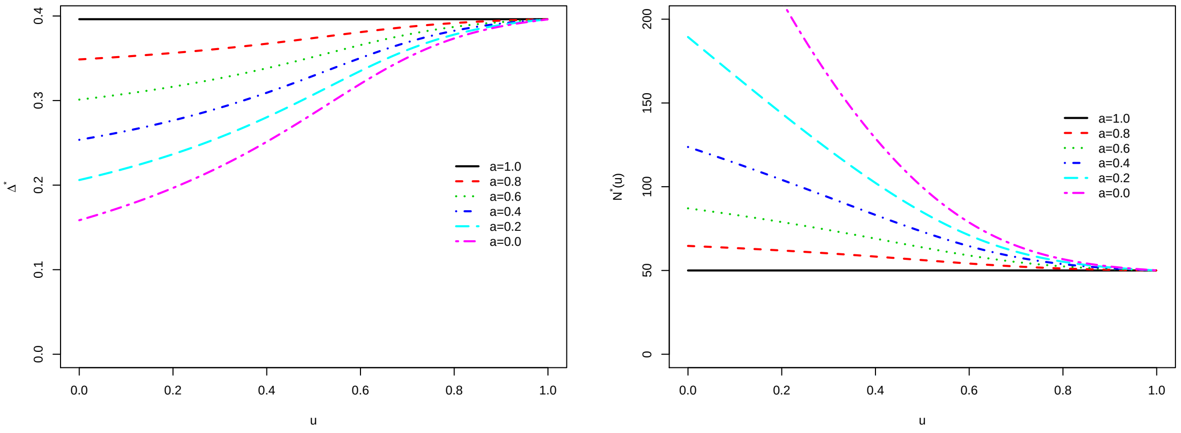
Difference in Power	0.75	0.80	0.85	0.90	0.95	1.00
a=0.8	4.20	3.32	2.47	1.63	0.81	0.00
a=0.6	8.93	7.02	5.16	3.38	1.65	0.00
a=0.4	14.13	11.05	8.08	5.24	2.54	0.00
a=0.2	19.72	15.39	11.21	7.23	3.48	0.00
a=0.0	25.58	19.99	14.54	9.33	4.46	0.00

Figure 2: Impact of attenuation on power for fixed sample size ($N = 50$). Tabled values below plot represent the difference in power from no attenuation ($a = 1$, power = 80%) for PPV between 0.75 and 1.0.



	0.75	0.80	0.85	0.90	0.95	1.00
a=0.8	1.11	1.09	1.06	1.04	1.02	1.00
a=0.6	1.24	1.19	1.14	1.09	1.04	1.00
a=0.4	1.40	1.30	1.22	1.14	1.07	1.00
a=0.2	1.59	1.44	1.31	1.19	1.09	1.00
a=0.0	1.83	1.60	1.41	1.25	1.12	1.00

Figure 3: Ratio of sample sizes required to maintain power (relative to no attenuation: a=1.0) for the expected treatment effect in the study depending on the degree of attenuation and positive predictive value (PPV) of screening procedures. Δ_1 and prevalence are 0.4. Tabled values below plot represent the ratios for PPV between 0.75 and 1.0.



(a) Δ^*

(b) N^*

Figure 4: Impact of attenuation on Δ^* and sample size (N^*) needed to overcome enrollment of patients with attenuated treatment effect and keep power constant, with power equal to 80% across values of u. For the example here $F_{op} \sim N(2,1)$, $F_{so} \sim N(0,1)$, prevalence $= 0.4$, and $\Delta_1 = 0.4$.

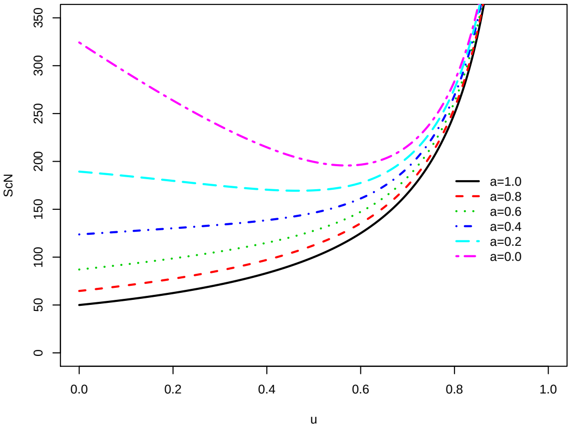


Figure 5: Impact of attenuation on number of patients needed to screen, with power equal to 80% and with $F_{op} \sim N(2,1)$, $F_{so} \sim N(0,1)$, prevalence $= 0.4$, and $\Delta_1 = 0.4$.

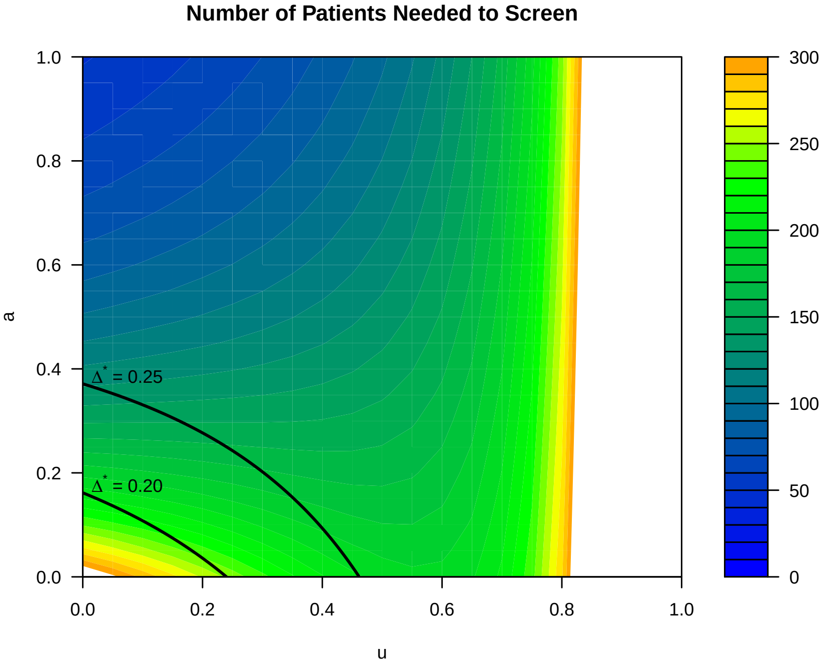
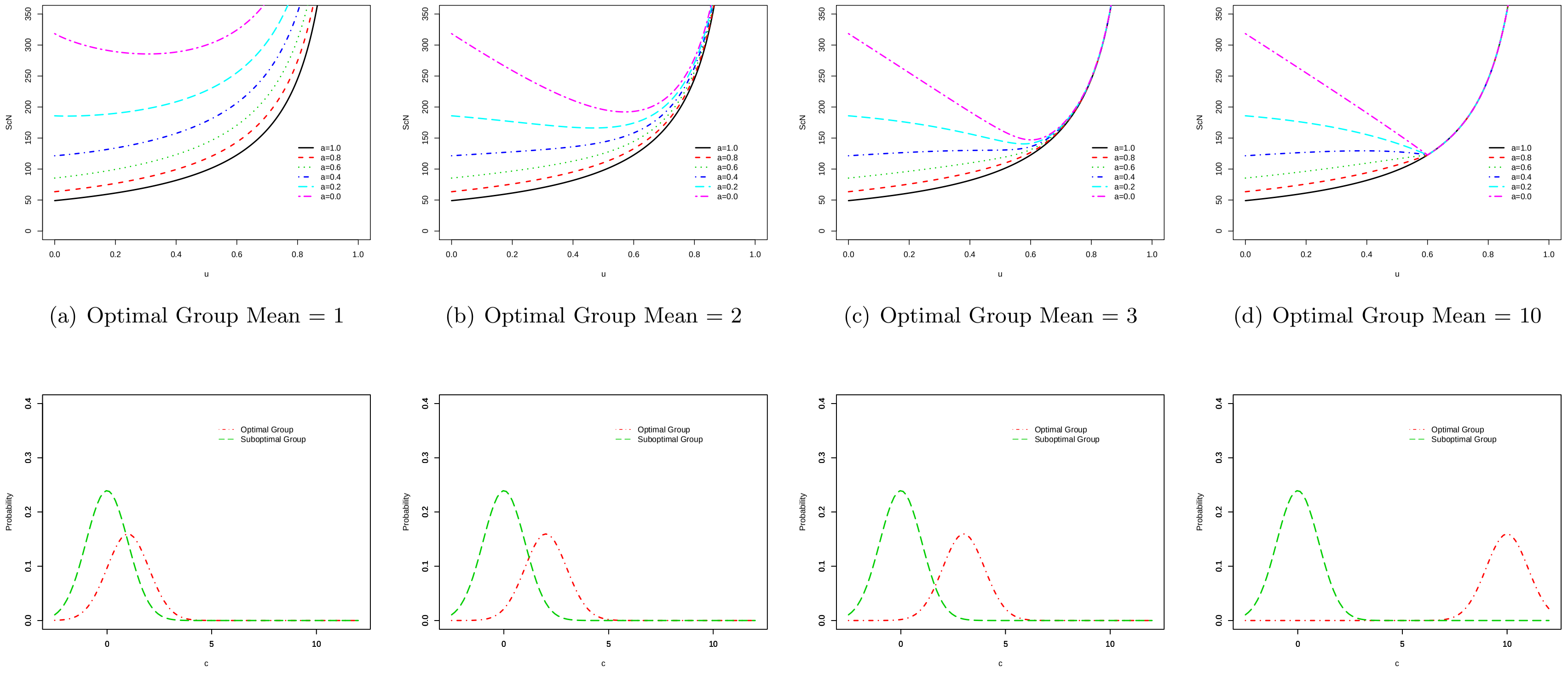


Figure 6: Elevation map depicting number of patients needed to screen while varying a and u with $F_{op} \sim N(2,1)$, $F_{so} \sim N(0,1)$, prevalence $= 0.4$, and $\Delta_1 = 0.4$. The horizontal line indicates the value of a wherein all of the values above the line have a minimum number of patients to screen equal to u=0 and all values below have a minimum with $u > 0$.



(a) Optimal Group Mean = 1

(b) Optimal Group Mean = 2

(c) Optimal Group Mean = 3

(d) Optimal Group Mean = 10

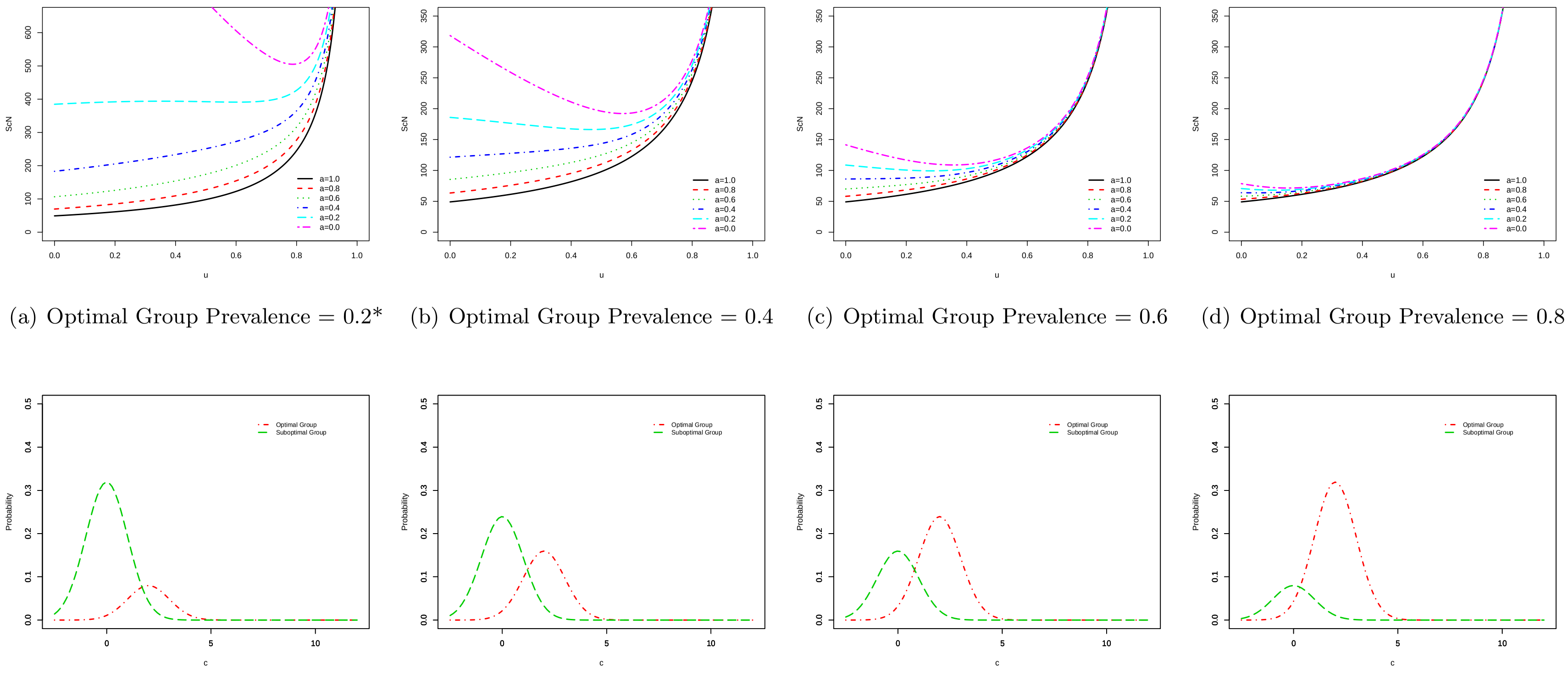
(e) Optimal Group Mean = 1

(f) Optimal Group Mean = 2

(g) Optimal Group Mean = 3

(h) Optimal Group Mean = 10

Figure 7: Impact of dissociation of marker distributions between optimal and suboptimal groups on the number of patients needed to screen (ScN). Number of patients needed to screen is in subfigures 7(a):7(d) with corresponding plot of marker densities in subfigures 7(e):7(h).



(a) Optimal Group Prevalence = 0.2*

(b) Optimal Group Prevalence = 0.4

(c) Optimal Group Prevalence = 0.6

(d) Optimal Group Prevalence = 0.8

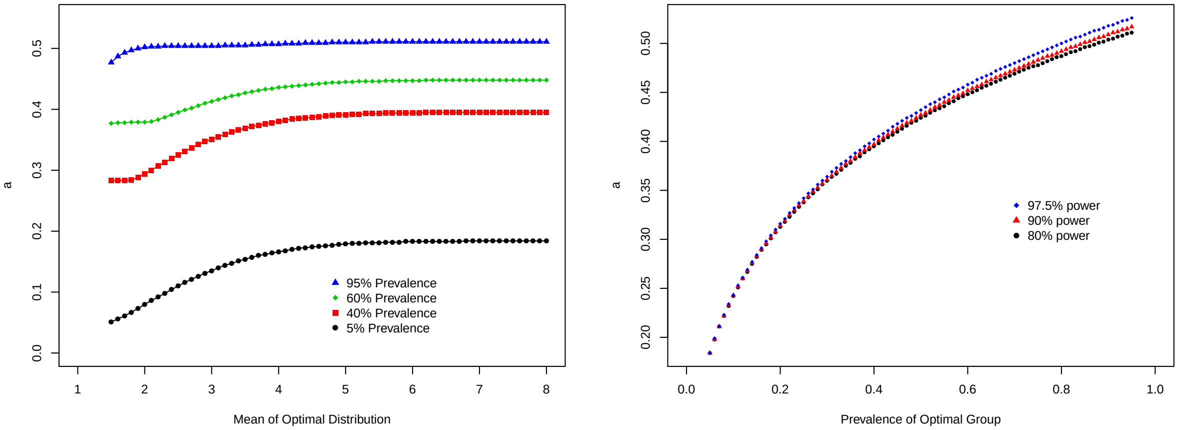
(e) Optimal Group Prevalence = 0.2

(f) Optimal Group Prevalence = 0.4

(g) Optimal Group Prevalence = 0.6

(h) Optimal Group Prevalence = 0.8

Figure 8: Impact of prevalence of optimal group on the number of patients needed to screen (ScN). Number of patients needed to screen is in subfigures 8(a):8(d) with corresponding plot of marker densities in subfigures 8(e):8(h). * Note in plot (a) a larger range for the y-axis was needed.



(a) Separation of Subgroup Biomarker Distributions

(b) Largest Attenuation Factor Across Prevalence of Optimal Distribution

Figure 9: Impact of the degree of separation of subgroup biomarker distributions and prevalence for power of 80%, 90% and 97.5% on the values of attenuation factor a for which the minimum number of screened patients is not at $u = 0$.