2011

A Likelihood-Based Approach to Early Stopping in Single Arm Phase II Clinical Trials

Elizabeth Garrett-Mayer

Amy E. Wahlquist
Medical University of South Carolina

MUSC Division of Biostatistics and Epidemiology Working Papers

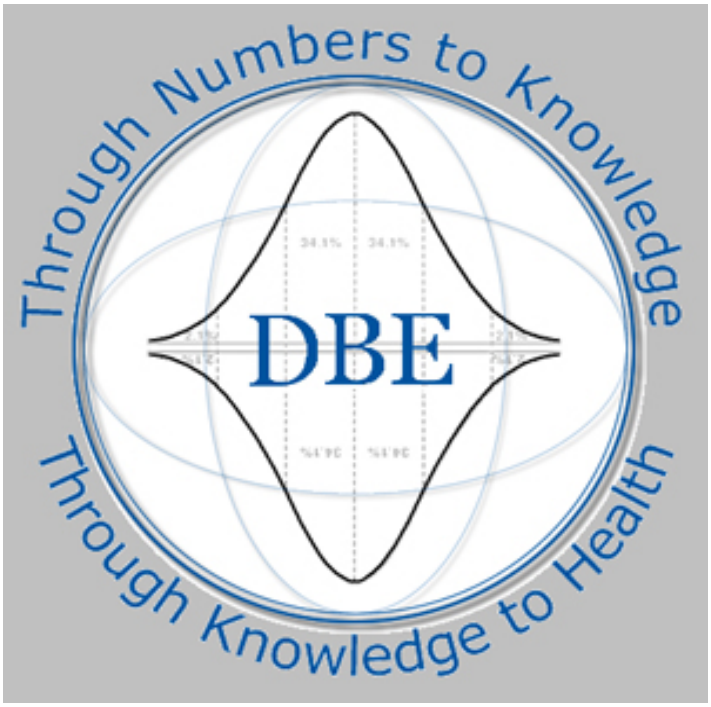
Paper Number/ Resource Identifier: 11-001

Date: 2011

MUSC Author(s): Garrett-Mayer, Elizabeth; Wahlquist, Amy E.

Paper Title: A likelihood-based approach to early stopping in single arm phase II clinical trials

Complete Author List: Garrett-Mayer, Elizabeth; Wahlquist, Amy E.



MUSC
MEDICAL UNIVERSITY of SOUTH CAROLINA

Full title: A likelihood-based approach to early stopping in single arm phase II clinical trials

Short title: Likelihood-based early stopping design phase II studies

Word Count: 6755

Elizabeth Garrett-Mayer, PhD
Amy E. Wahlquist, MS

Hollings Cancer Center
&
Department of Medicine, Division of Biostatistics and Epidemiology
Medical University of South Carolina

Corresponding Author:
Elizabeth Garrett-Mayer
86 Jonathan Lucas St., Rm. 118G
Charleston, SC 29425
843-792-7764
843-792-4233 (fax)
garrettm@musc.edu

The authors acknowledge support from the Medical University of South Carolina – Cancer Center Support Grant, Biostatistics Core (P30 CA138313) and the Hollings Cancer Center.

1. Introduction

Phase II studies in oncology have evolved over the previous several decades. Currently, the number of drugs in phase II development has increased, and patient eligibility has narrowed due to targeted agents, competing trials and curative therapies in the first-line setting. As a result of these changes, more attention needs to be focused toward conducting more efficient phase II trials. Given the increased difficulty in accruing patients to phase II studies and the ethical concern of treating patients with agents that are ineffective, there is significant motivation to stop a single arm trial early when the investigational agent shows evidence of a low response rate.

Many single arm phase II trials in oncology continue to be developed using tumor response, or another binary measure of clinical efficacy, as the primary outcome. Unlike comparative trials, which are often designed to stop early if there is convincing evidence that one treatment regimen is superior to the other, single arm studies usually only stop for futility. The most popular phase II design in oncology clinical trials has been the Simon two-stage design [1] which allows one early look to stop for futility and maintains overall type I and type II errors while increasing the sample size only modestly when compared to a single stage design. This design is easy to perform in practice and is simple to design using commonly available software and web-based programs. For any given set of study parameters (null and alternative response rates and type I and type II errors), there are many designs that can be performed with one early look. Simon defined several criteria for choosing the best design. The “optimal” design has the smallest expected sample size under the null hypothesis, while the “minimax” design has the smallest sample size at the end of the second stage of the trial [1].

The Simon two-stage design only allows stopping at one point during the trial. Using other design approaches, it is possible to consider multiple stopping times, which provide higher chances of stopping the trial if the treatment is ineffective. A related extension of the Simon two-stage design is the three-stage design [2]. Both the two- and three-stage designs mentioned thus far are frequentist in nature and measure evidence for early stopping and for rejecting hypotheses using p-values. In addition to the two- and three-stage designs mentioned, there are others proposed, mostly varying in their optimization criteria [3-6]. Due to the very nature of p-values and the problems with accumulating type I errors by repeated hypothesis testing in the frequentist setting, a number of other early stopping criteria for single arm phase II studies with binary endpoints have been proposed.

We propose a likelihood-based stopping design that follows the evidential paradigm and relies on the likelihood principle [7]. The observed data is evaluated under the null and alternative hypotheses and the resulting likelihood ratio is used for making inferences. This approach does not rely on prior information and the evidence can be quantified in a relatively simple way. Unlike frequentist designs in which type I error rates at successive interim looks can accumulate to unacceptable levels, the evidential paradigm relies on the universal bound, which limits the probability of misleading evidence in favor of the incorrect hypothesis. As a result, likelihood-based designs can look at the data early and often and terminate a trial if there is sufficiently strong evidence, and the probability of making an incorrect conclusion can be relatively small. Blume clarified some potential misconceptions about the evidential paradigm [8]. It is true that more frequent looks at the data will lead to higher chances of selecting an incorrect hypothesis in the evidential paradigm: we do not contend that our approach will lead to the same error rates as looking at the data only once or twice. The probability of misleading evidence is bounded regardless of the number of interim looks and the bound can be controlled.

There is a large literature selection on the design of single arm phase II trials [9-11]. Even so, in cancer research, the Simon two-stage design is the most commonly adopted when clinical response is the outcome of interest. Lee and Liu proposed a predictive probability design, described in more detail in section 3.2 [12] and similar to an earlier version by Herson [13]. They convincingly argue that multi-stage frequentist designs can be challenging to implement and analyze. Frequentist inferences condition on the study design, so that if an early look is implemented earlier or later than originally planned, the stopping rules and statistical properties are then undefined and inference becomes challenging. Other authors have described this in more detail and proposed some design and inference solutions [14, 15]. Bayesian and likelihood approaches do not suffer from this complication; both approaches rely on the likelihood principle which uses all data collected and does not depend on the design from which the data arose. A number of other Bayesian approaches exist in this setting but have not been widely adopted [16-24]. There are several reasons that might explain this. First, the inclusion of a prior distribution may be unacceptable to some. Second, Bayesian designs are often assumed to be more computationally challenging. For some Bayesian approaches, these assertions are not well-founded because priors can be selected that have little influence on the inferences (i.e., weak priors), and calculations for early stopping can often be done prior to the implementation of the trial. But arguably the most challenging problem to overcome by using a non-frequentist design approach is describing the resulting evidence. Medical researchers have become accustomed to the interpretation of p-values such that other forms of evidence become hard to interpret. A similar phenomenon exists with power calculations. Non-frequentist design proponents have been required to translate their design properties into type I and II errors so that the research community can interpret findings in the frequentist paradigm. The likelihood approach allows for this type of conversion due to the universal bound and by quantifying the probability of misleading and weak evidence as will be described in the following sections.

2. Methods

2.1. The likelihood approach

Historically, the dominating design approach has been frequentist, based on theories developed by Neyman & Pearson and by R.A. Fisher, and we refer to this approach as significance testing, where null and alternative hypotheses are defined, acceptable type I and II errors are chosen, and a decision is made at the end of the trial via the p-value regarding the null hypothesis [25, 26]. Royall [7] describes some of the flaws in using this approach for interpreting its results as evidence. It is beyond the scope of this manuscript to discuss these details, but chief among them is the requirement to choose between two hypotheses at the end of the trial; i.e., the choice to reject or fail to reject the null hypothesis. There is no allowance made for "weak" evidence in favor of one or the other hypothesis or even strong evidence in favor of the null hypothesis. Royall [7] and Blume [27] argue for likelihood-based approaches in clinical trial design and regard likelihood-based approaches as the *evidential* paradigm.

The Neyman-Pearson theory and the evidential paradigm are both based on the likelihood ratio, but each uses it in a different way. The Neyman-Pearson theory bases inferences on the probability that the likelihood ratio will be larger than some value, k, if the null hypothesis is true. The evidential paradigm is based on the value of the likelihood ratio itself and more specifically, on the law of likelihood:

> *If hypothesis A implies that the probability of observing some data X is $P_A(X)$, and hypothesis B implies that the probability of observing some data X is $P_B(X)$, then the observation $X = x$ is evidence supporting A over B if $P_A(x) > P_B(x)$. Further, the likelihood ratio, $P_A(x)/P_B(x)$, measures the strength of that evidence [7, 28].*

The evidential paradigm uses the likelihood function, evaluated at the observed data, to quantify evidence regarding a particular hypothesized value of a parameter (see Blume [27] for more detail on interval construction). The likelihood ratio (LR) is constructed by comparing the likelihood function, evaluated at different parameter values, based on the observed data. Suppose we observe 16 responses in 45 patients in a trial (an observed response rate of 0.36) where our null hypothesis is a response rate of 0.20, and our alternative hypothesis is a response rate of 0.40. The LR comparing the alternative to the null hypothesis is 15.7 and was obtained by taking the ratio of the likelihood function at 0.40 to the height at 0.20 (0.83/0.053 = 15.7). A standard frequentist approach would calculate the p-value, which in this case is 0.014, suggesting strong evidence against the null hypothesized response rate of 0.20.

2.2. A key difference in likelihood versus significance-testing approaches

The significance-testing paradigm and the evidential paradigm have similarities, and often the inferences reached at the end of a trial would lead to the same conclusions regarding the success of the trial. However, statistically, there is a critical philosophical difference that should be considered when deciding which approach is more sensible. In the significance-testing paradigm, inferences are based on p-values which are calculated ignoring the alternative hypothesis. The question posed is thus “Is there sufficient evidence to conclude that the null hypothesis is not true?” The evidential paradigm uses the LR which compares evidence for the null versus the alternative hypotheses. In essence, the question posed is “Which of these two hypotheses is more consistent with the observed data?”

Reconsidering our example of an observed response rate of 0.36 in 45 patients, suppose our null and alternative hypotheses were 0.20 and 0.50, respectively (instead of 0.20 and 0.40). The observed response rate of 0.36 is almost equally between 0.20 and 0.50, suggesting that the

data do not appear to strongly favor either hypothesis. The p-value remains unchanged at 0.014 because the null hypothesis is the same. The LR is now only 2.80 (LR = 0.15/0.053), implying rather weak evidence in favor of the alternative hypothesis. The evidential paradigm provides a result that is consistent with our expectations, but the significance-testing paradigm does not. This difference in inference is because they ask two different questions.

2.3. Inference in the evidential paradigm

At the end of a study, in the evidential paradigm, a LR is calculated and then interpreted as evidence. There are three possible types of evidence that are observed: (1) weak evidence, (2) strong evidence in favor of the correct hypothesis, (3) strong evidence in favor of the incorrect hypothesis. (This is a bit of a simplification because it presumes that one of the two hypotheses posed is correct.) Weak evidence arises when there is not sufficiently strong evidence in favor of either hypothesis. In theory, this can be controlled by increasing the sample size; however, if neither hypothesis is correct and the true value of the parameter lies somewhere between, then weak evidence may arise even with a relatively large sample size. Second, strong evidence in favor of the correct hypothesis can be observed. This is, of course, the goal--to obtain convincing correct evidence. Lastly, strong evidence in favor of the incorrect hypothesis is considered *misleading* evidence (similar to type I and II errors). A trial should terminate early for futility and rarely terminate when the treatment is effective. The probability of misleading evidence should be controlled through proper study design and an appropriate choice of K.

The *universal* bound states that the probability that the LR exceeds k in favor of the wrong hypothesis can be no larger than $1/k$ [29, 30]. That is, under the null hypothesis,

$$P\left(\frac{L_1}{L_0} \geq k\right) \leq 1/k$$

where L_1 is the likelihood under the alternative and L_0 is the likelihood under the null hypothesis. An even lower bound applies in some cases (e.g., difference between normal means; large sample size), but no bound has been shown for the binomial likelihood with relatively small sample size [31].

Importantly, this bound holds for a sequence of independent observations which allows repeated estimation of the LR with the bound being maintained [32]. In other words, unlike the significance-testing paradigm where multiple looks at the data are penalized by increasing type I error rates, the accumulating data in a clinical trial can be evaluated in a fully sequential fashion, and the overall rate of misleading evidence will be bounded by $1/K$. For a single arm trial with response as the outcome and $K = k$, we could estimate the LR after every patient's response (or lack of response) had been observed. We would stop the trial if the LR in favor of the null hypothesis was greater than k, and the probability that we would mistakenly stop the trial early would be less than $1/k$.

Royall proposed guidelines for "strength of evidence" with thresholds of 8 and 32 for classifying LRs into three levels of evidence [7]. When comparing hypothesis 1 (numerator) to hypothesis 2 (denominator), LRs in the ranges 1-8, 8-32 and >32 would correspond to weak, moderate and strong evidence regions in favor of hypothesis 1, respectively. However, in exploring the appropriate values of K in the phase II oncology clinical trial setting, it must be recognized that weaker evidence is usually acceptable. That is, while phase III comparative trials usually specify two-sided alpha of 0.05 or one-sided alpha of 0.025, phase II trials very often specify one-sided

alpha of 0.05 or 0.10. As a result, the suggested value of $K=8$ as a guideline for choosing a hypothesis may be too high in our application.

2.4. Likelihood-based stopping in single arm phase II studies with binary endpoints

There is strong motivation in oncology research (and other medical research) to terminate single arm studies as soon as there is convincing evidence that the treatment regimen under study will not be as effective as desired. The universal bound allows sequential estimation and evaluation of the LR where early stopping for futility can be enacted when the LR is greater than k in favor of the null hypothesis. The computational requirement for estimation is trivial: on the log scale, we would stop the trial at time t, when y_t responses had been observed out of N_t patients, if the following were true:

$$LR_t = \left(\frac{p_1}{p_0}\right)^{y_t}\left(\frac{1-p_1}{1-p_0}\right)^{N_t-y_t} < \frac{1}{k} \qquad (1)$$

Note that p_0 and p_1 are determined prior to the start of the trial so that estimating the likelihood function at time t simply involves plugging in the number of responses (y_t) and the number of patients (N_t). If LR_t is greater than or equal to $1/k$, then the trial would continue.

LR_t can be calculated after each patient's response has been observed, and the stopping criteria can be calculated prior to the start of the trial, simplifying implementation and computational needs during the trial. Similar to Simon's two-stage design, we enumerate the stopping boundaries as part of the trial protocol. Table 1 displays the stopping boundaries where the null and alternative hypothesized response rates are 0.20 and 0.40, and our

maximum sample size is chosen to be 37 assuming a fully sequential implementation of the trial where K is chosen to be 8. Each of the thresholds for y_t that are listed correspond to the scenarios under which the likelihood ratio LR_t is less than $1/K$ where $K=8$.

2.5. The choice of K

Much of the debate in the use of the evidential paradigm has focused on the "correct" choice of K. In frequentist designs, the clinical trials community is accustomed to alpha of 0.05 or 0.10 and power in the range of 0.80 to 0.90 in phase II trials. As many statisticians will attest, these choices are somewhat arbitrary, but have become the convention. As a result, we explore values of K that we would consider appropriate based on previous research but also values of K that provide similar operating characteristics (acceptance/rejection of hypotheses) to the Simon designs [7, 8]. In this exploration, we have found that separate values of K for early stopping and for making a final decision are appropriate to provide similar operating characteristics and to improve performance in terms of early stopping and decreasing expected sample size under the null hypothesis. We refer to K for early stopping as *interim K*, denoted K_i, and K for final inference (if the study reaches its maximum allowed sample size) is referred to as *end K*, denoted K_e.

The probability of misleading evidence is bounded by $\Phi(-\sqrt{2\log(k)})$ regardless of the choice of sample size [7, 8]. This bound allows us to determine the relationship between values of K and type I and II errors by equating the probability of misleading evidence to the probability of rejecting the null when it is true (or, similarly, the probability of failing to reject the null when it is false). This approach leads to the relationship between alpha and K (Figure 1) which led to the selection of $K=8$ for denoting strong evidence (i.e., it corresponds closely to two-sided alpha of

0.05) [7]. However, in our setting, testing is usually one-sided, and the alpha level is often higher than 0.05. Simon two-stage designs are commonly implemented with an alpha of 0.10 and a one-sided test, suggesting a lower value of K may be more appropriate in the phase II setting and, referring to Figure 1, $K=2.3$ may provide similar operating characteristics to a Simon two-stage design with type I and II errors of 0.10. However, there is an asymmetry in the evaluation of the hypotheses in a futility stopping design: the trial has many opportunities for selecting the null hypothesis (i.e., failing to reject the alternative), but the alternative hypothesis can only be selected at the end of the trial. Trials that do not stop early are unlikely to accept the null hypothesis at the trial's end and will instead result in weak evidence or strong evidence in favor of the alternative. Therefore, K_e should be chosen to limit the number of trials that lead to weak evidence when the alternative hypothesis is true.

3. Results

3.1. Performance of likelihood stopping design compared to Simon's two-stage designs

To determine the performance characteristics of the likelihood stopping design (LSD), we performed simulations of trials and compared the LSD to Simon's optimal two-stage design (O2SD) and Simon's minimax two-stage design (M2SD). In our simulations, the null (H_0) is considered an ineffective level of response, and the alternative (H_1) is a response rate that is sufficiently high to warrant further study of the treatment. The following performance characteristics were considered:

- expected value of the sample size under the null hypothesis
- probability of stopping under the null hypothesis
- probability of acceptance of H_1 under H_1 (similar to power)

- probability of acceptance of H_0 under H_0 (similar to 1-alpha).

Although theoretically the properties of a design can be estimated exactly, the computational burden is significant and increases with the number of possible stopping thresholds. As a result, simulations were performed (simulating 10,000 trials for each set of conditions) to estimate performance characteristics.

A scenario is defined by its null (p_0) and alternative (p_1) response rates. We chose type I and II errors to be 10% and then identified the optimal and minimax designs for the scenario. We compared the Simon designs to likelihood designs: one with the maximum sample size of the optimal design and one with the maximum sample size of the minimax design. We are consistent with Simon's notation for his designs: N_1 is the stage 1 sample size, N is the total possible sample size, the trial stops at stage 1 if ≤ r_1 responses are seen, we fail to reject the null if ≤ r responses occur in N patients. We explored different values of K for early stopping and for final inference and compared the design operating characteristics to the Simon designs in terms of the probabilities of acceptance and rejection of hypotheses, probability of early stopping, and expected final sample sizes. Three of our design scenarios are shown here, although many more were considered.

3.1.1. Scenario 1: $p_0 = 0.20$, $p_1 = 0.40$, $K_i = 8$, $K_e = 2.3$, $N=37$ (O2SD), $N=36$ (M2SD)

The O2SD is characterized by $N_1=17$, $r_1= 3$, $N=37$, and $r=9$ and the M2SD by $N_1=22$, $r_1= 4$, $N=36$, and $r=9$. LSDs with maximum N of both 37 and 36 were considered to show comparability to the Simon designs. Table 1 shows the stopping boundaries for the likelihood designs with $K_i=8$. $K_e=2.3$ was chosen to allow comparability to a type I or type II error of 0.10. Choosing $K_i=8$ and $K_e=2.3$, we achieve similar performance under the null as the Simon designs (Figures 2A and 2B); however, the performance is not directly comparable in the sense that

there are three inferential categories in the likelihood design (strong evidence for H_0, strong evidence for H_1, and weak evidence) versus only two categories in the frequentist approach (reject or fail to reject H_0). Figure 2A displays, for the range of true response rates, the inferential probabilities for each design with a total possible sample size of 37 compared to O2SD. For a true response rate of 20%, we have strong evidence in favor of the null 91% of the time, weak evidence 4.4% of the time, and in only 4.3% of trials do we conclude that the alternative is true (i.e., reject the null). This is almost identical to the O2SD which has a 91% chance of failing to reject the null and a 9% chance of a type I error. The likelihood design performs slightly worse under $K_i = 8$ and $K_e = 2.3$ when the alternative is true. For a true response rate of 40%, the chance of correctly accepting the alternative is 84%, weak evidence is 5%, and falsely accepting the null is 11%. This is similar to the Simon design which rejects the null 90% of time and falsely fails to reject only 10% of the time.

When the true response rate is 0.30 (midway between the null and alternative), the LSD has an 88% chance of selecting one of the two hypotheses but 12% of the time will find weak evidence (LR is between 1/2.3 and 2.3). The O2SD will reject the null 54% of the time and fail to reject 46% of the time. The level of weak evidence for the O2SD may seem low when the true response rate is 0.30, but there is a relatively low threshold for K_e. Figure 2B shows the corresponding results for the minimax design, which are quite similar.

Figures 2C and 2D compare expected sample size and probability of early stopping in the likelihood and Simon designs. Under the null hypothesis, the probability of early stopping with the likelihood design is 0.82 for both N=37 and N=36, compared with 0.55 and 0.54 for the Simon optimal and minimax designs, respectively. Under the alternative, the likelihood and Simon designs have similar early stopping probabilities. One of the most attractive features of LSDs is the increased chance of early stopping under the null hypothesis which exposes fewer

patients to an ineffective therapy as demonstrated in Figure 2D where the expected sample size under the null hypothesis is 20 for the likelihood designs and 26 and 28 for the O2SD and M2SD, respectively. Under the alternative, the expected samples sizes are close: $E(N|H_1) = 35$ in the likelihood design and $E(N|H_1) = 36$ in the Simon designs, both designs continuing to the maximum sample size in the large majority of trials.

3.1.2. Scenario 2: $p_0 = 0.20$, $p_1 = 0.40$, $K_i=8$, $K_e=1$, N=37 (O2SD), N=36 (M2SD).

In the previous scenario, the chance of weak evidence could be lowered, suggesting that for comparison to the Simon designs, $K_e=2.3$ may be too high and choosing a lower value for K_e may improve relative performance.

To address this, scenario 2 implements that same early stopping rule ($K_i=8$); however, for trials reaching the final sample size, a likelihood ratio threshold is set to 1 ($K_e=1$). This is the lowest reasonable bound in the evidential paradigm and eliminates the weak evidence category. Comparing the likelihood designs in Figures 2A and 2B to those in 3A and 3B (i.e., $K_e=2.3$ vs. $K_e=1$), we see improved operating characteristics in relation to acceptance of the correct hypothesis at the null and alternative probabilities. Briefly, compared to the scenario with $N=37$ and $K_e=2.3$, choosing $K_e=1$ leads to the same chance of choosing the null when it is true, but only a relatively small increase in the chance of choosing the alternative (9% vs. 4%) when the null is true. Operating characteristics under the alternative are also improved when $K_e=1$ relative to the case when $K_e=2.3$. However, the behavior for other true response rates should be considered. In the previous scenario, when $p=0.30$, the probability of weak evidence was 0.12. With $K_e=1$, when $p=0.30$ the probability of selecting the null is 0.48, and the chance of selecting that alternative is 0.52 which are very similar to scenario 1.

Comparing the design with $K_i=8$, $K_e=1$ to the Simon designs (Figures 3A and 3B), Simon's and the likelihood designs give us almost identical properties under the null and only a relatively small difference under the alternative. Despite comparable rejection/acceptance of hypotheses, the early stopping probabilities and expected sample size (Figures 3C and 3D) vary substantially when comparing the likelihood to the Simon designs. Given that the K_i is the same as in Scenario 1, the early stopping probabilities and expected sample sizes are the same in Scenarios 1 and 2. The results in Figure 3 suggest that we can obtain properties as good as the Simon designs (in terms of rejecting/accepting/failing to reject the correct hypotheses), yet we will treat fewer patients when the null hypothesis is true.

3.1.3. Scenario 3: $p_0 = 0.05$, $p_1 = 0.20$, $K_i=8$, $K_e=1$, $N=37$ (O2SD), $N=32$ (M2SD).

A common situation arises in phase II oncology trials when a new treatment is to be tested in a patient population for which there is no curative therapy. In this case, we may assume a low null response rate of 5%, and the alternative usually ranges anywhere from 15% to 35% with a 20% response rate being a fairly common choice for the alternative. This yields an O2SD with $N_1=12$, $r_1= 0$, $N=37$, and $r=3$ and M2SD with $N_1=18$, $r_1= 1$, $N=32$, and $r=3$. The likelihood designs have few opportunities to stop in a situation with a low expected response rate. When choosing $K_i=8$, there are only three early stopping opportunities shown in Table 2: 0 responses in 13 patients, 1 response in 22 patients, and 2 responses in 31 patients. It would be expected that with fewer opportunities for stopping, the likelihood and Simon designs would be more similar in performance.

K_e was chosen to be 1 for comparability to the Simon designs. The acceptance/rejection of hypotheses probabilities are similar in pattern to scenario 1: under the alternative hypothesis, the likelihood designs have a lower chance of choosing the alternative as compared to the

Simon designs (Figure 4A). However, the performance of the likelihood design under the null is better in the optimal LSD than the O2SD. Figure 4B demonstrates that the M2SD and the LSD with *N=32* are almost identical. In Figures 4C and 4D, the LSDs perform better under the null in terms of expected sample sizes and early stopping probabilities, although the improvement in expected number of patients under the null is not as dramatic as in the previous scenarios given that there are fewer opportunities for stopping. The chance of early stopping under the null in the likelihood designs is much greater (84%) than in the O2SD (54%), and an even greater difference exists when comparing the M2SD (84% for the LSD vs. 40% for M2SD). Under the null, the likelihood designs have expected sample sizes of 21 and 20 with maximum possible sample sizes of *N=37* and *N=32*, respectively, while the Simon designs have expected sample sizes of 23 and 26. When comparing the likelihood design to the optimal design with the same sample size, the difference in expected sample size under the null is only 2, but when comparing the likelihood design to the minimax design, the expected difference is 6, which is substantial given that the maximum sample size is only 32. The probabilities of early stopping under the alternative in the likelihood designs are 9% and 10% for *N=37* and *N=32* as compared to 7% and 2% in the Simon designs. However, the expected sample sizes are almost identical under the alternative (Figure 4D).

3.1.4. Other situations

There are infinitely many possibilities to consider, but due to space limitations, we have only shown a few here. Additional scenarios are described in section 4. As it turns out, the larger the null hypothesized response rate, the greater advantage the likelihood design will have on the expected sample size and on the probability of early stopping when the true response rate is less than or equal to the null. This is not surprising given the increased number of opportunities for early looks at the data.

3.2. Comparison to the predictive probability approach

Comparisons have focused on the Simon design because it is most present in oncology trials. Lee and Liu's predictive probability design (PPD) may be a more natural comparison due to its increased number of looks [12]. Despite being Bayesian, the PPD is similar regarding its treatment of the alternative hypothesis to Simon designs. It relies on the probability that the response rate is larger than p_0, regardless of the proposed alternative. Specifically, early stopping is based on the probability that the treatment will be deemed efficacious (i.e. $p>p_0$) by the end of the study (if the study were to be completed) given the data observed at an interim look. Lee and Liu chose to estimate the predictive probability beginning with the 10^{th} patient in a trial and to then monitor continuously, although they also explored other non-sequential designs.

Lee and Liu compared PPDs with type I and II levels less than 0.10 to Simon designs with type I and II error levels less than 0.10 [12]. We have selected a subset of the designs proposed by Lee and Liu; specifically, the ones with the same sample sizes as the O2SD and M2SD (Table 3). The PPD shows improvements relative to the Simon designs in regards to probability of early stopping and smaller expected sample sizes under the null. We have added LSDs with the same maximum sample sizes and $K_i=8$ and $K_e=1$ to the comparisons. Table 3 shows type I errors of less than 0.10 for all LSDS and type II errors ranging from 0.091 to 0.133. Most of the type II errors are not less than 0.10, but all are negligibly different. Under the null, the expected sample size is smaller for all LSDs compared to PPDs and Simon designs. Probability of early termination (*PET0*) is smallest under the Simon designs and highest under the PPDs, with the LSDs having stopping probabilities close to but smaller than the PPDs. These results suggest that the PPDs stop early more often but stop later in the trials given that their expected sample sizes tend to be larger. We argue that with the primary goal of sparing patients ineffective

treatments, having a smaller expected sample size under the null is the more important characteristic.

3.3. Sample size calculations for designing likelihood early stopping designs.

A common approach for study design in a frequentist setting is to graph sample size versus power for a fixed level of alpha and effect size. A similar approach in the likelihood setting for single arm studies with a binary response is to plot acceptance probabilities of both null and alternative hypotheses for fixed values of K_i and K_e and fixed p_0 and p_1. An additional quantity that can be included is the probability of early stopping under the null (and/or alternative). In the previous section, we explored study designs with $K_i=8$ and with K_e values of 2.3 and 1 with null and alternative hypothesized response rates of 0.20 and 0.40. In Figure 5, four combinations of K_i and K_e are considered: (A) $K_i=8$, $K_e=8$; (B) $K_i=8$, $K_e=2.3$; (C) $K_i=8$, $K_e=1$; *(D)* $K_i=4$, $K_e=2.3$. In each panel, sample size is plotted versus the probability of acceptance of a hypothesis and early stopping under each hypothesis (i.e., $p_0 = 0.20$; $p_1 = 0.40$). Figure 5A suggests that if proposing $K_i = K_e = 8$, the sample size should be 50 to ensure at least an 80% chance of accepting the alternative when it is true and 90% chance of accepting the null when it is true. This sample size will also yield a very high probability of early stopping under the null (0.90). Decreasing the K for the final look to $K_e=2.3$, a required sample size of 38 will provide >90% and >85% chance of accepting the null and alternative hypotheses (Figure 5B), respectively, when they are true and >80% chance of early stopping under the null. If the stringency of futility stopping is decreased to $K_i=4$ with a final $K_e=2.3$ (Figure 5D), even increasing the sample size to $N=80$ will not provide a sufficiently high probability of accepting the alternative when it is true due to the high chance of early stopping with somewhat weak evidence. For large sample sizes ($N>60$) in Figure 5D, the probability of accepting the null when it is true is greater than 0.98, implying an imbalance in the likelihood of strong misleading evidence. These sample size

figures help us choose reasonable values for K_i and K_e that lead to acceptable and appealing design performance characteristics for a range of sample sizes. As an example, inspection of Figure 5D clearly demonstrates that regardless of sample size, the choice of $K_i=4$ and $K_e=2.3$ has poor operating characteristics.

4. Examples

4.1. Clinical trial in patients with non-small cell lung cancer (NSCLC).

A clinical trial has been proposed (and submitted to the NCI) to test the hypothesis that treatment of advanced stage and refractory NSCLC patients with a novel agent that is FDA-approved for multiple sclerosis will improve the tumor response rate via activation of ceramide-PP2A tumor suppressor signaling by binding/targeting I2PP2A oncoprotein leading to c-Myc degradation, telomerase inhibition, and consequent tumor suppression. The primary endpoint is disease-control rate (DCR; i.e., the proportion of patients with complete response, partial response and stable disease) at 8 weeks. Our null hypothesis is that the DCR is 0.30 and our alternative is 0.50. A likelihood stopping design has been proposed with a likelihood ratio of 8 favoring the null versus the alternative and, at the end of the study, the hypothesis is chosen which the likelihood ratio favors (i.e., $K_e = 1$). With a total possible sample size of 46, the selected design has a 94% chance of accepting the null when it is true and 87% chance of accepting the alternative when it is true. These are comparable to alpha of 0.06 and power of 87% in the O2SD. However, the expected sample size under the null is 29 in the O2SD versus 22.7 using the LSD.

4.2. Vitamin D_3 supplementation in African American adults

A randomized study has been planned in African American adults to determine (1) if 4,000 IU of Vitamin D_3 daily is safe, and (2) if supplementation is able to improve health-related measures, such as blood pressure and lipids (e.g., cholesterol levels). One-hundred and fifty patients will be randomized to the supplementation arm and 75 to the placebo arm. The safety endpoint (i.e., toxicity) is defined as the incidence of any grade II, III, or IV toxicity (as defined by CTCAE v 4.0) within the first 16 weeks of treatment in the supplementation arm. The null hypothesis is that 85% of patients will not experience a toxicity and the alternative is that 95% of patients will not have a toxicity. (This is analogous to a null toxicity rate of 15% and an alternative toxicity rate of 5%). Using a likelihood early stopping approach with these presumed hypotheses, choosing $K_i=8$ and $K_e=2.3$, the probability of accepting the null when it is true is 0.98; the probability of accepting the null when the alternative is true is 0.07; and the probability of weak evidence is <0.01 regardless of whether the null or alternative is true. If the null is true (i.e., toxicity rate of 0.15), the probability of early stopping is 0.98 and the expected sample size in the supplementation arm is only 38. Under the alternative hypothesis (i.e., toxicity rate of 0.05), the chance of early stopping is only 7% and the expected sample size is 142. If, for practical implementation, it is decided to evaluate the data after every tenth patient has reached 16 weeks of follow-up, the expected sample sizes under the null and alternative are 45 and 145, respectively; and the probabilities of accepting the null when it is true and when the alternative is true are 0.98 and 0.09, respectively. With the same total sample size, the Simon optimal design can be performed with a type I error of 3.5%, power of 95% and expected sample size under the null of 69.

5. Discussion

The evidential paradigm provides a natural framework for early stopping in single arm phase II studies using the likelihood principle and taking advantage of the universal bound which limits

the probability of misleading evidence. The examples shown in Section 3 illustrate that it is possible to achieve similar acceptance and rejection of proposed hypotheses while improving upon the early stopping probabilities under the null hypothesis. Implementation of two different thresholds for evidence - one threshold for early stopping and another threshold for hypothesis selection at the trial's end allow us to control the probability of weak and misleading evidence while increasing the chance of early stopping under the null. Choices of $K_i=8$ and $K_e=1$ or 2.3 appear to be reasonable choices. The combination of $K_i=8$ and $K_e=8$ yielded weak evidence relatively frequently. The combination of $K_i=4$ and $K_e=2.3$ caused an imbalance in the likelihood of acceptance of the null when it is true (high probability) versus the alternative when it is true (low probability). This is due to the asymmetry in the design (which can stop early only for futility) and to the low threshold for early stopping resulting in 20% or more of trials stopping early when the treatment is effective.

Although it was not emphasized in our illustrative examples, Figure 2 shows that when the true response rate is even lower than the null response rate, the chance of early stopping will be even higher. This is also true of the Simon designs, but because of the increased number of looks in the likelihood approach, the expected sample size comparisons between the Simon designs and the likelihood designs tend to more strongly favor the likelihood designs.

Single arm phase II clinical trials with binary endpoints can be designed allowing for more frequent looks at the data while preserving operating characteristics (i.e., probability of accepting the correct hypothesis) and improving others (i.e., probability of stopping early under the null hypothesis). Fewer patients are exposed to ineffective agents, and resources are preserved by stopping futile trials earlier. Comparisons to the PPD and Simon designs suggest that LSDs allow earlier stopping (although not necessarily more frequent stopping) and subject fewer patients to ineffective therapies with little or no increase in error rates. The methods

proposed have been developed into an R library, allowing users to compare the likelihood-based design to the Simon designs, and to design future trials using sample size graphics.

Acknowledgements: The authors acknowledge support from the Medical University of South Carolina – Cancer Center Support Grant, Biostatistics Core (P30 CA138313) and the Hollings Cancer Center.

References

1. Simon, R., *Optimal two-stage designs for phase II clinical trials.* Control Clin Trials, 1989. **10**(1): p. 1-10.
2. Ensign, L.G., et al., *An optimal three-stage design for phase II clinical trials.* Stat Med, 1994. **13**(17): p. 1727-36.
3. Fleming, T.R., *One-sample multiple testing procedure for phase II clinical trials.* Biometrics, 1982. **38**(1): p. 143-51.
4. Jung, S.H., M. Carey, and K.M. Kim, *Graphical search for two-stage designs for phase II clinical trials.* Control Clin Trials, 2001. **22**(4): p. 367-72.
5. Jung, S.H. and K.M. Kim, *On the estimation of the binomial probability in multistage clinical trials.* Stat Med, 2004. **23**(6): p. 881-96.
6. Jung, S.H., et al., *Admissible two-stage designs for phase II cancer clinical trials.* Stat Med, 2004. **23**(4): p. 561-9.
7. Royall, R., *Statistical Evidence*. 1997, Boca Raton, FL: Chapman & Hall/CRC.
8. Blume, J.D., *How often likelihood ratios are misleading in sequential trials.* Communications in Statistics - Theory and Methods, 2008. **37**(8): p. 1193-1206.
9. Thall, P.F. and R.M. Simon, *Recent developments in the design of phase II clinical trials.* Cancer Treat Res, 1995. **75**: p. 49-71.
10. Kramar, A., D. Potvin, and C. Hill, *Multistage designs for phase II clinical trials: statistical issues in cancer research.* Br J Cancer, 1996. **74**(8): p. 1317-20.
11. Gehan, E.A., *Update on planning of phase II clinical trials.* Drugs Exp Clin Res, 1986. **12**(1-3): p. 43-50.
12. Lee, J.J. and D.D. Liu, *A predictive probability design for phase II cancer clinical trials.* Clin Trials, 2008. **5**(2): p. 93-106.
13. Herson, J., *Predictive probability early termination plans for phase II clinical trials.* Biometrics, 1979. **35**(4): p. 775-83.
14. Herndon, J.E., 2nd, *A design alternative for two-stage, phase II, multicenter cancer clinical trials.* Control Clin Trials, 1998. **19**(5): p. 440-50.
15. Chen, T.T. and T.H. Ng, *Optimal flexible designs in phase II clinical trials.* Stat Med, 1998. **17**(20): p. 2301-12.
16. Thall, P.F. and R. Simon, *A Bayesian approach to establishing sample size and monitoring criteria for phase II clinical trials.* Control Clin Trials, 1994. **15**(6): p. 463-81.
17. Sylvester, R.J., *A bayesian approach to the design of phase II clinical trials.* Biometrics, 1988. **44**(3): p. 823-36.
18. Heitjan, D.F., *Bayesian interim analysis of phase II cancer clinical trials.* Stat Med, 1997. **16**(16): p. 1791-802.
19. Banerjee, A. and A.A. Tsiatis, *Adaptive two-stage designs in phase II clinical trials.* Stat Med, 2006. **25**(19): p. 3382-95.
20. Ding, M., G.L. Rosner, and P. Muller, *Bayesian optimal design for phase II screening trials.* Biometrics, 2008. **64**(3): p. 886-94.
21. Sambucini, V., *A Bayesian predictive two-stage design for phase II clinical trials.* Stat Med, 2008. **27**(8): p. 1199-224.
22. Tan, S.B. and D. Machin, *Bayesian two-stage designs for phase II clinical trials.* Stat Med, 2006. **25**(19): p. 3407-8.
23. Thall, P.F., et al., *Hierarchical Bayesian approaches to phase II trials in diseases with multiple subtypes.* Stat Med, 2003. **22**(5): p. 763-80.
24. Zohar, S., S. Teramukai, and Y. Zhou, *Bayesian design and conduct of phase II single-arm clinical trials with binary outcomes: a tutorial.* Contemp Clin Trials, 2008. **29**(4): p. 608-16.

25. Neyman, J. and E.S. Pearson, *On the Problem of the Most Efficient Tests of Statistical Hypotheses.* Phil. Trans. R. Soc., Series A, 1933. **231**: p. 289-337.
26. Fisher, R., *The Logic of Inductive Inference.* Journal of the Royal Statistical Society, 1935. **98**: p. 39-54.
27. Blume, J., *Tutorial in Biostatistics: Likelihood methods for measuring statistical evidence.* Statistics in Medicine, 2002. **21**: p. 2563-2599.
28. Hacking, I., *Logic of Statistical Inference*. 1965, Cambridge: Cambridge University Press.
29. Birnbaum, A., *On the foundations of statistical inference (with discussion).* Journal of the American Statistical Association, 1962. **53**: p. 259-326.
30. Smith, C., *The Detection of Linkage in Human Genetics.* Journal of the Royal Statistical Society, Series B, 1953. **15**(153-192).
31. Royall, R. and T.-S. Tsou, *Interpreting statistical evidence by using imperfect models: robust adjusted likelihood functions.* Journal of the Royal Statistical Society, B, 2003. **65**: p. 391-404.
32. Robbins, H., *Statistical methods related to the law of the iterated logarithm.* Annals of Mathematical Statistics, 1970. **41**: p. 1397-1409.

Table 1: Early stopping thresholds for $p_0 = 0.20$, $p_1 = 0.40$ with $K_i=8$ and a maximum sample size of N=37. Stopping occurs if $1/LR_t < 0.125$

Number of responses (y_t)	Number of patients (N_t)	Estimated response rate	$1/LR_t$
0	8	0	0.10
1	11	0.09	0.11
2	15	0.13	0.095
3	18	0.17	0.11
4	21	0.19	0.12
5	25	0.20	0.10
6	28	0.21	0.11
7	32	0.22	0.096
8	35	0.23	0.11

Table 2: Stopping thresholds for $p_0 = 0.05$, $p_1 = 0.20$ with $K_i=8$ and a maximum sample size of N=37. Stopping occurs if $1/LR_t < 0.125$

Number of responses (y_t)	Number of patients (N_t)	Estimated response rate	$1/LR_t$
0	13	0	0.11
1	22	0.045	0.11
2	31	0.065	0.11

Table 3: Comparison of three designs, based on probability of early termination under the null hypothesis (PET0), expected value of the sample size under the null (E(N0)), and the final decision rule for accepting the null (or failing to reject the null). If there are r or fewer responses in N patients, the null will be accepted (or fail to be rejected). For each null and alternative hypothesis pair, the first row represents the sample size selected based on the Simon Minimax Design with $\alpha = \beta = 0.10$. The second row corresponds to the sample size for the Simon Optimal Design with $\alpha = \beta = 0.10$. The PPD and LSD were derived using the same maximum sample size.

		Simon Two-Stage Design $\alpha = \beta = 0.10$				Predictive Probability Design $\alpha = \beta = 0.10$				Likelihood Stopping Design $K_i = 8$, $K_e = 1$			
	Max N*	PET_0	$E(N_0)$	α	β	PET_0	$E(N_0)$	α	β	PET_0	$E(N_0)$	α	β
0.10 vs. 0.30	25	0.52	20.4	0.095	0.097	0.79	20.0	0.096	0.095	0.73	15.8	0.091	0.125
	35	0.66	19.8	0.098	0.099	0.86	22.0	0.062	0.099	0.87	17.1	0.046	0.114
0.20 vs. 0.40	36	0.46	28.3	0.086	0.098	0.86	27.7	0.088	0.094	0.82	20.4	0.082	0.125
	37	0.55	26.0	0.095	0.097	0.85	25.1	0.100	0.084	0.82	20.4	0.089	0.109
0.30 vs. 0.50	39	0.37	35.0	0.094	0.100	0.86	32.5	0.096	0.083	0.79	22.1	0.078	0.129
	46	0.67	29.9	0.097	0.095	0.88	33.5	0.081	0.088	0.85	22.9	0.058	0.127
0.40 vs. 0.60	41	0.55	33.8	0.095	0.099	0.87	31.1	0.096	0.098	0.78	24.0	0.090	0.118
	46	0.56	30.2	0.095	0.100	0.88	32.1	0.091	0.093	0.81	24.8	0.058	0.133
0.50 vs. 0.70	39	0.50	31.0	0.098	0.099	0.87	29.2	0.100	0.095	0.80	21.4	0.084	0.118
	45	0.67	29.0	0.096	0.098	0.88	25.4	0.100	0.091	0.85	22.3	0.054	0.126
0.60 vs. 0.80	35	0.82	28.5	0.097	0.100	0.89	25.5	0.090	0.089	0.82	17.8	0.093	0.113
	38	0.47	25.4	0.097	0.096	0.88	21.6	0.099	0.081	0.85	18.4	0.088	0.105
0.70 vs. 0.90	25	0.55	20.0	0.091	0.092	0.89	16.4	0.091	0.098	0.80	13.0	0.083	0.123
	28	0.54	17.8	0.099	0.090	0.88	15.7	0.100	0.077	0.84	13.4	0.088	0.091

* Max N refers to the maximum sample size the trial would enroll using the Minimax Design sample size and the Optimal Design sample size.

Figure Captions:

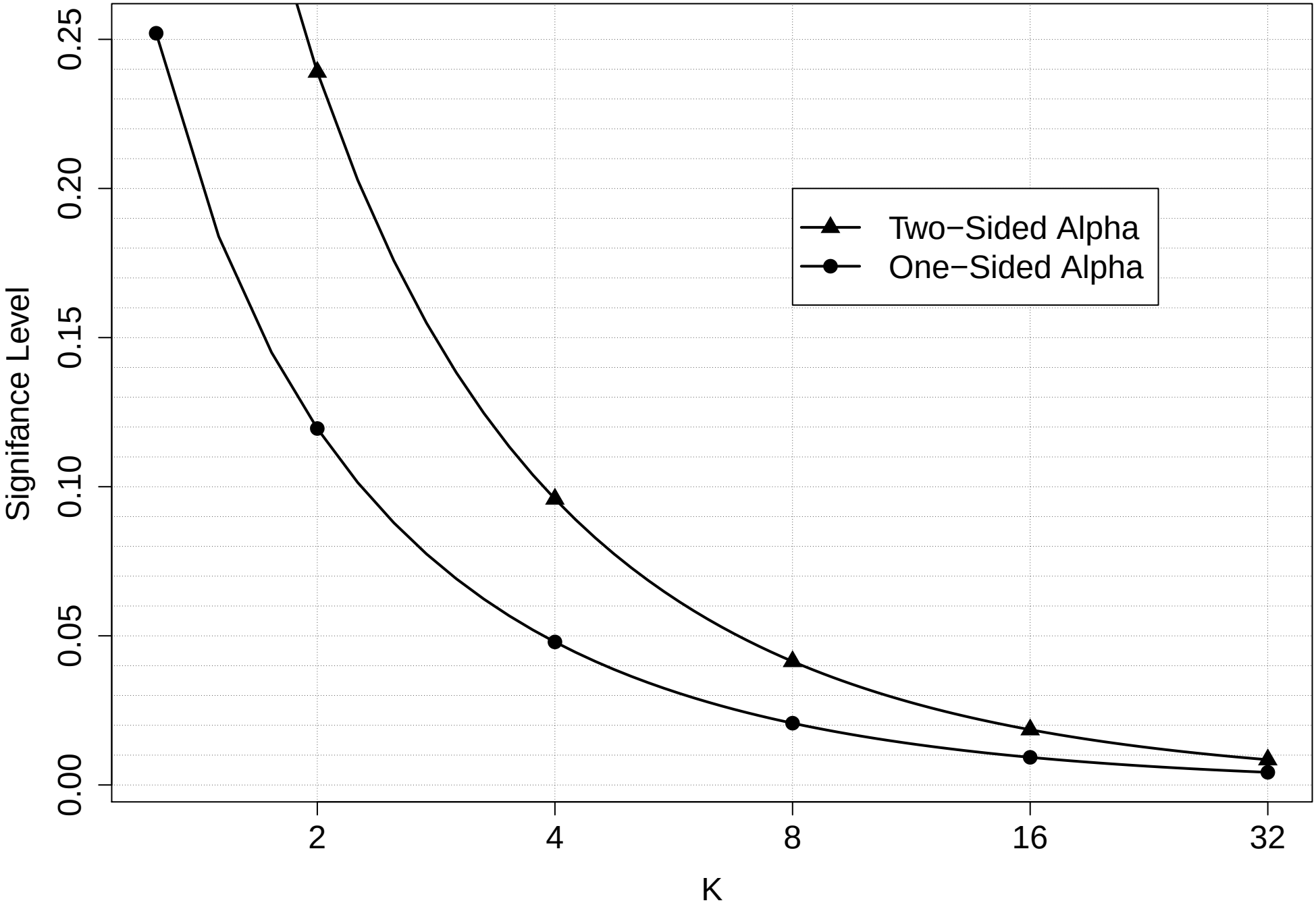
Figure 1: The relationship between K and the significance level (alpha) for one- and two-sided hypothesis testing based on the limiting frequency of observing strong misleading evidence. Specifically, the probability of strong misleading evidence (i.e., a type I or II error in the frequentist paradigm) is limited by $\Phi(-\sqrt{2\log(k)})$ in the evidential paradigm.

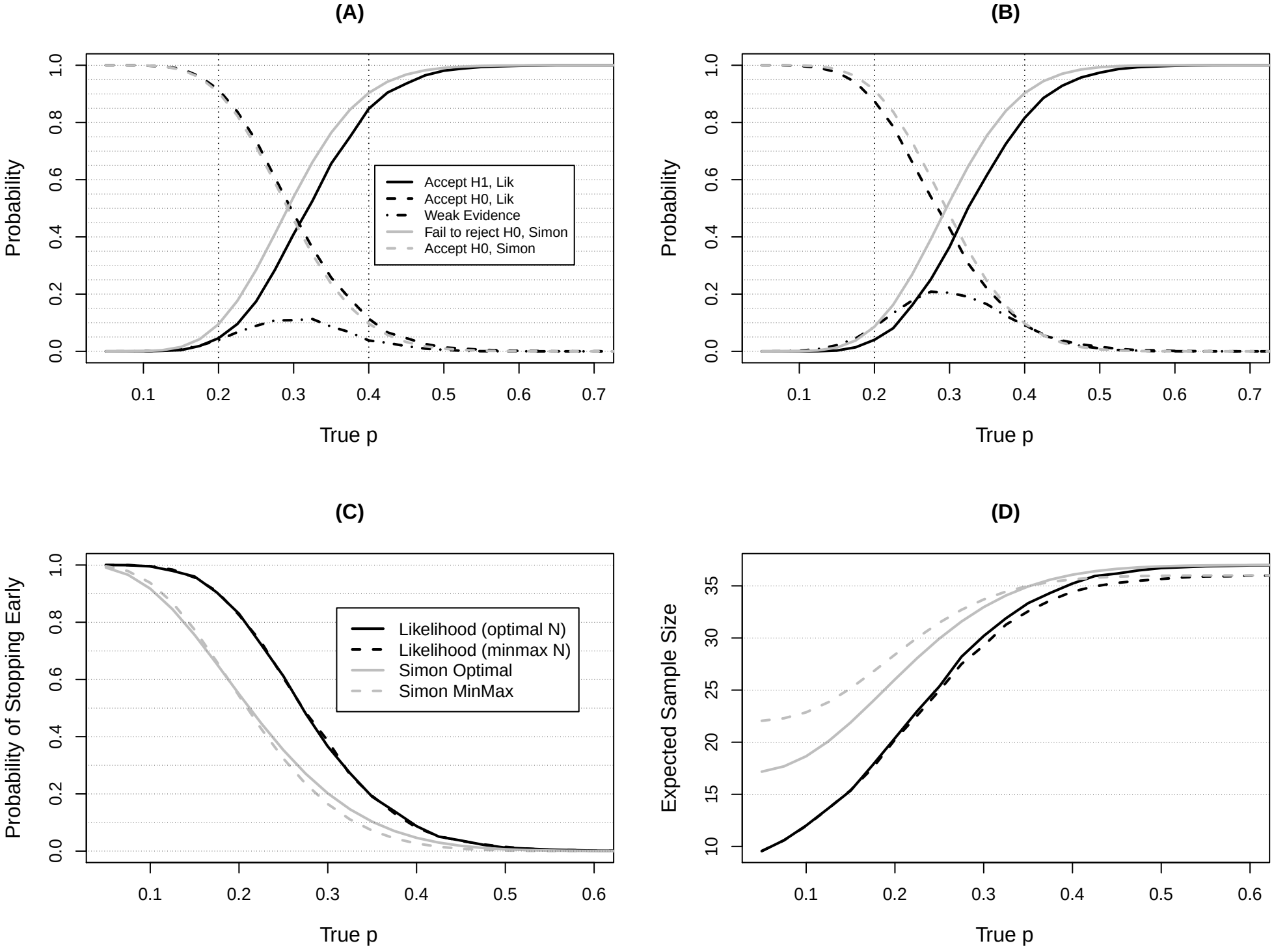
Figure 2: Comparison of operating characteristics of Simon designs and LSDs assuming H_0*: p=0.20,* H_1*: p=0.40,* K_i*=8 and* K_e*=2.3*. (A) Probabilities of accepting or rejecting null and alternative hypotheses and weak evidence based on true response rates; Simon *optimal* design versus LSD (*max N=37*). (B) Probabilities of accepting or rejecting null and alternative hypotheses and weak evidence based on true response rates; Simon *mimimax* design versus LSD (*max N=36*). (C) Probability of early stopping versus true response rates for Simon designs and LSDs. (D) Expected sample size versus true response rates for Simon designs and LSDs.

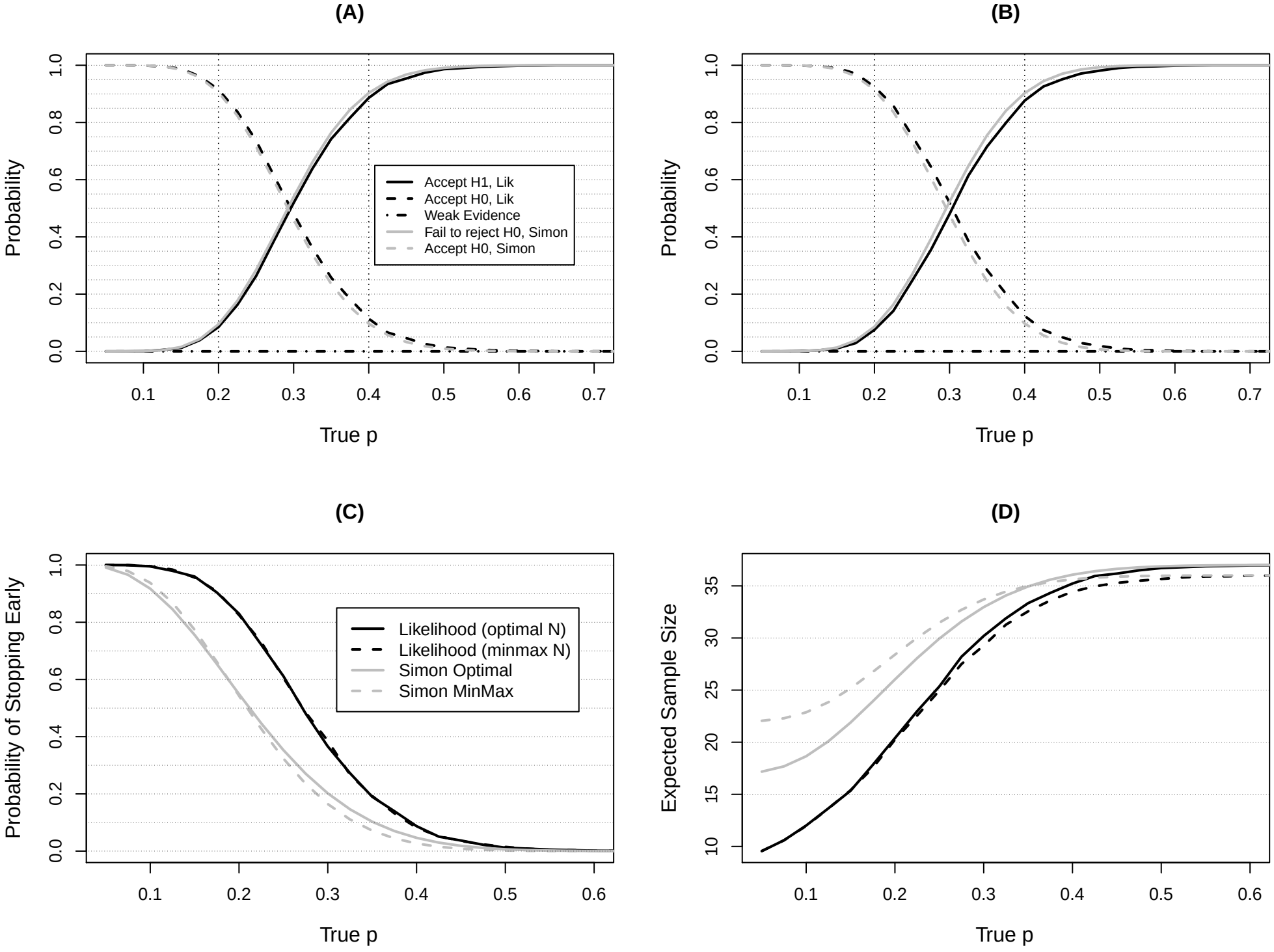
Figure 3: Comparison of operating characteristics of Simon designs and LSDs assuming H_0*: p=0.20,* H_1*: p=0.40,* K_i*=8 and* K_e*=1*. (A) Probabilities of accepting or rejecting null and alternative hypotheses and weak evidence based on true response rates; Simon *optimal* design versus LSD (*max N=37*). (B) Probabilities of accepting or rejecting null and alternative hypotheses and weak evidence based on true response rates; Simon *mimimax* design versus LSD (*max N=36*). (C) Probability of early stopping versus true response rates for Simon designs and LSDs. (D) Expected sample size versus true response rates for Simon designs and LSDs.

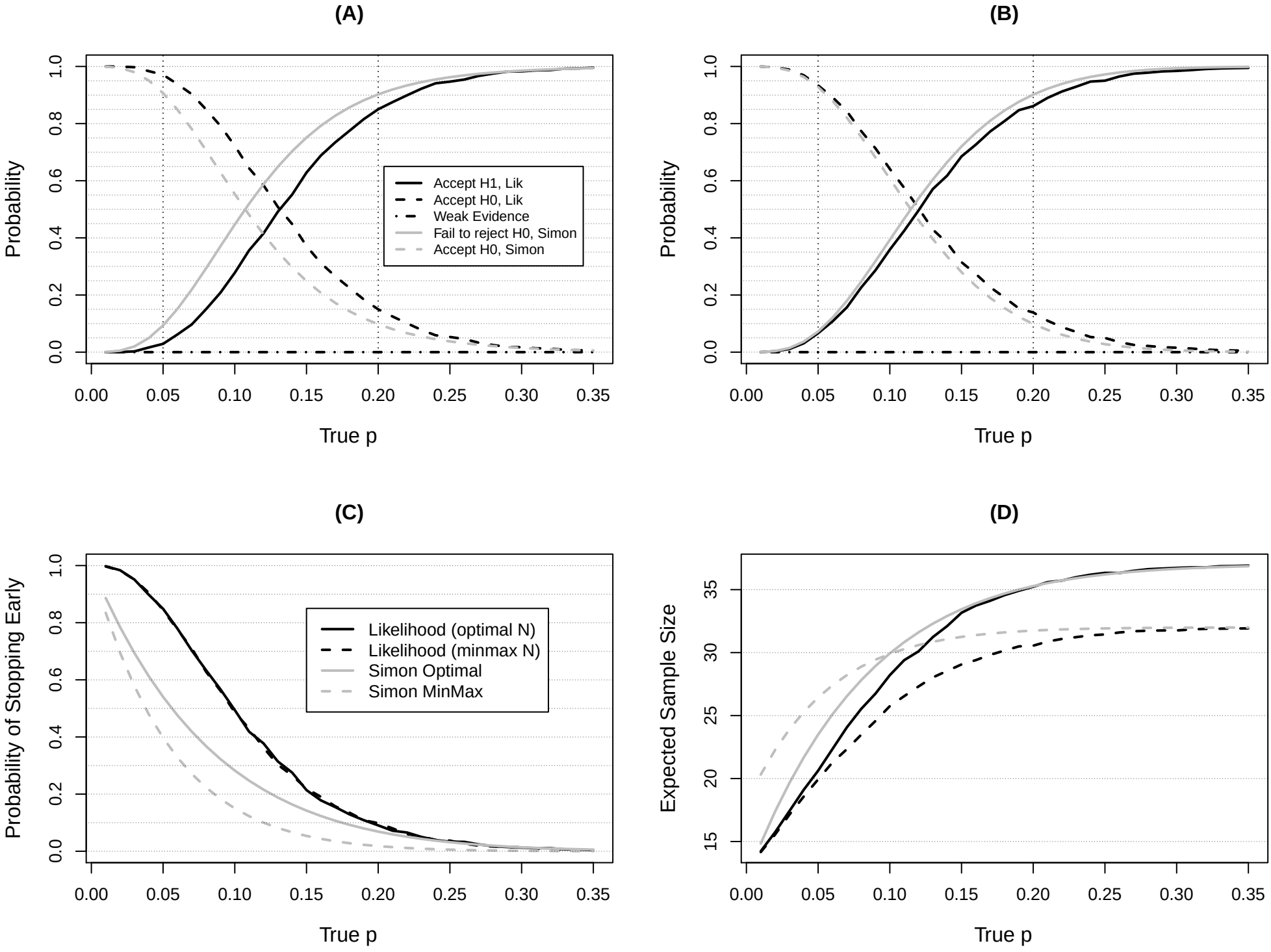
Figure 4: Comparison of operating characteristics of Simon designs and LSDs assuming H_0*: p=0.05,* H_1*: p=0.20,* K_i*=8* and K_e*=1*. (A) Probabilities of accepting or rejecting null and alternative hypotheses and weak evidence based on true response rates; Simon *optimal* design versus LSD (*max N=37*). (B) Probabilities of accepting or rejecting null and alternative hypotheses and weak evidence based on true response rates; Simon *mimimax* design versus LSD (*max N=32*). (C) Probability of early stopping versus true response rates for Simon designs and LSDs. (D) Expected sample size versus true response rates for Simon designs and LSDs.

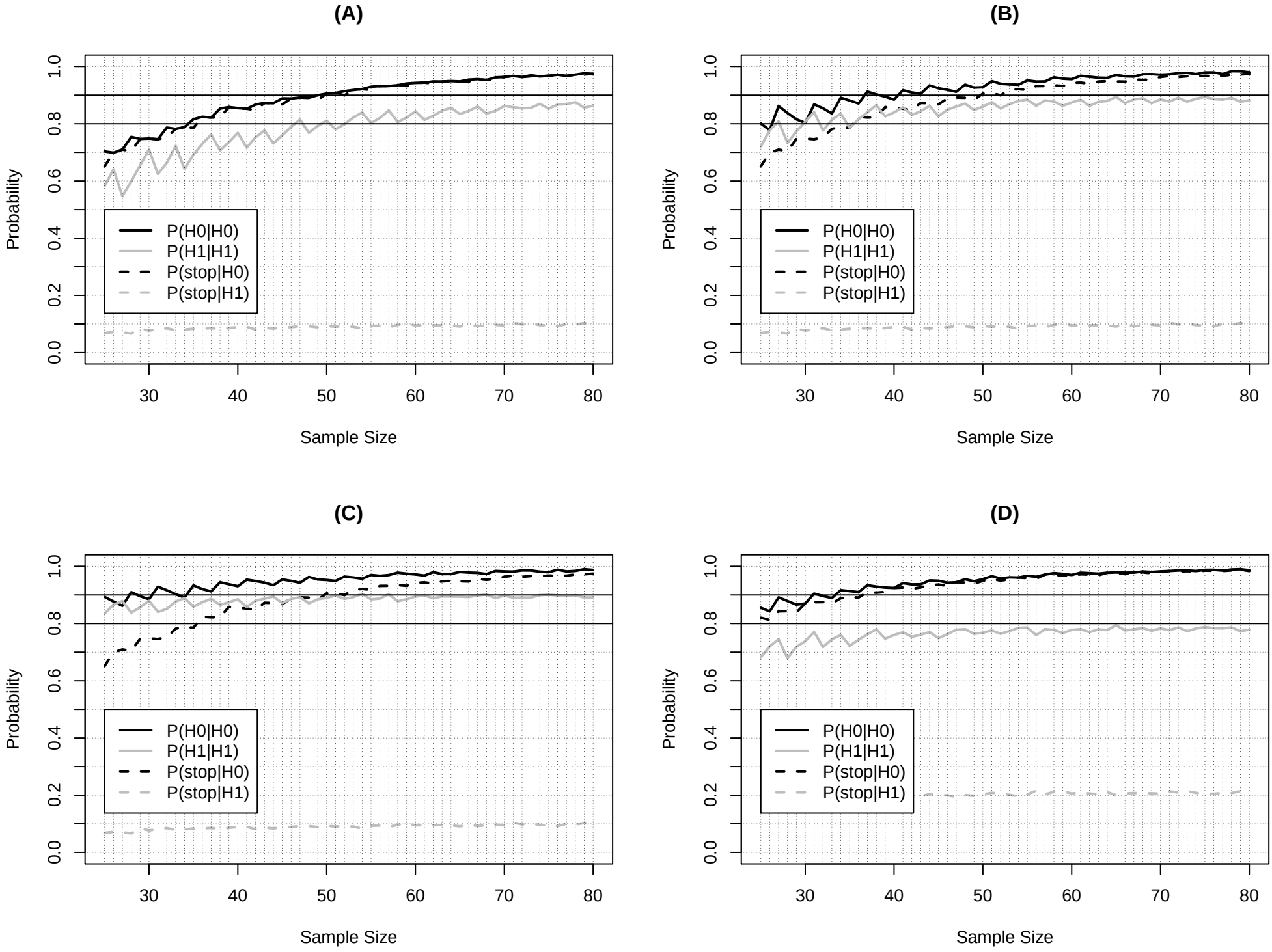
Figure 5: Sample size plots for various choices of K_i and K_e assuming H_0*: p=0.20,* H_1*: p=0.40* for LSDs. $P(H_i|\ H_i)$ = probability of accepting H_i when it is true, $P(stop|\ H_i)$ is the probability of early stopping under H_i. (A) K_i*=8* and K_e*=8*; (B) K_i*=8* and K_e*=2.3*.; (C) K_i*=8* and K_e*=1*; (D) K_i*=4* and K_e*=2.3*.


Signifance Level
0.00
0.05
0.10
0.15
0.20
0.25
2
4
8
16
32
K
Two−Sided Alpha
One−Sided Alpha


(A)
Probability
True p
Accept H1, Lik
Accept H0, Lik
Weak Evidence
Fail to reject H0, Simon
Accept H0, Simon
(B)
Probability
True p
(C)
Probability of Stopping Early
True p
Likelihood (optimal N)
Likelihood (minmax N)
Simon Optimal
Simon MinMax
(D)
Expected Sample Size
True p


(A)
(B)
(C)
(D)
Probability
True p
Accept H1, Lik
Accept H0, Lik
Weak Evidence
Fail to reject H0, Simon
Accept H0, Simon
Probability of Stopping Early
Likelihood (optimal N)
Likelihood (minmax N)
Simon Optimal
Simon MinMax
Expected Sample Size


(A)
Probability
True p
Accept H1, Lik
Accept H0, Lik
Weak Evidence
Fail to reject H0, Simon
Accept H0, Simon
(B)
Probability
True p
(C)
Probability of Stopping Early
True p
Likelihood (optimal N)
Likelihood (minmax N)
Simon Optimal
Simon MinMax
(D)
Expected Sample Size
True p


(A)
(B)
(C)
(D)
Probability
Sample Size
0.0
0.2
0.4
0.6
0.8
1.0
30
40
50
60
70
80
P(H0|H0)
P(H1|H1)
P(stop|H0)
P(stop|H1)