Reproducible study and further explorations on Bayesian dual-criterion design

Minghan Yang

February 2025

Contents

Abstract

Related to paper "Beyond p-values: A phase II dual-criterion design with statistical significance and clinical relevance" Roychoudhury et al. (2018). It addresses Bayesian dual-criterion design, sample size calculation, operational characteristics, and two cases: a randomized PoC design with time-to-event data and a single-arm PoC design with binary data. It also explores how the prior choices affect the minimum sample size.

1 Introduction

Proof-of-concept (POC) in Phase II trails is important in investigating the efficacy of an experimental drug. It will influence the decision of whether continuing or not on the development of the drug.
Dual-criterion design in frequentist and Bayesian applications are discussed.
Three generic phase II designs are reviewed:

1. Standard design
   For comparative treatment and control trails, it puts forward criteria expressed as error rates: Type I error control and power (correctly reject H_0 when it is false).
   We want to control type I error and maximize power.

Type I error: $\mathbb{P}(\text{reject } H_0 | H_0 \text{ is true}) = \alpha$

Type II error: $\mathbb{P}(\text{not reject } H_0 | H_0 \text{ is false}) = \beta$.

Limitation: statistical significant only guarantees evidence to reject "No effect", but is not sufficient for clinical perspective. Also, it always result in success or failure according to statistical significance.

Increasing the sample size increases the power for effects better than null.

2. Dual-criterion design

   Considers both the statistical significance and the effect estimate.

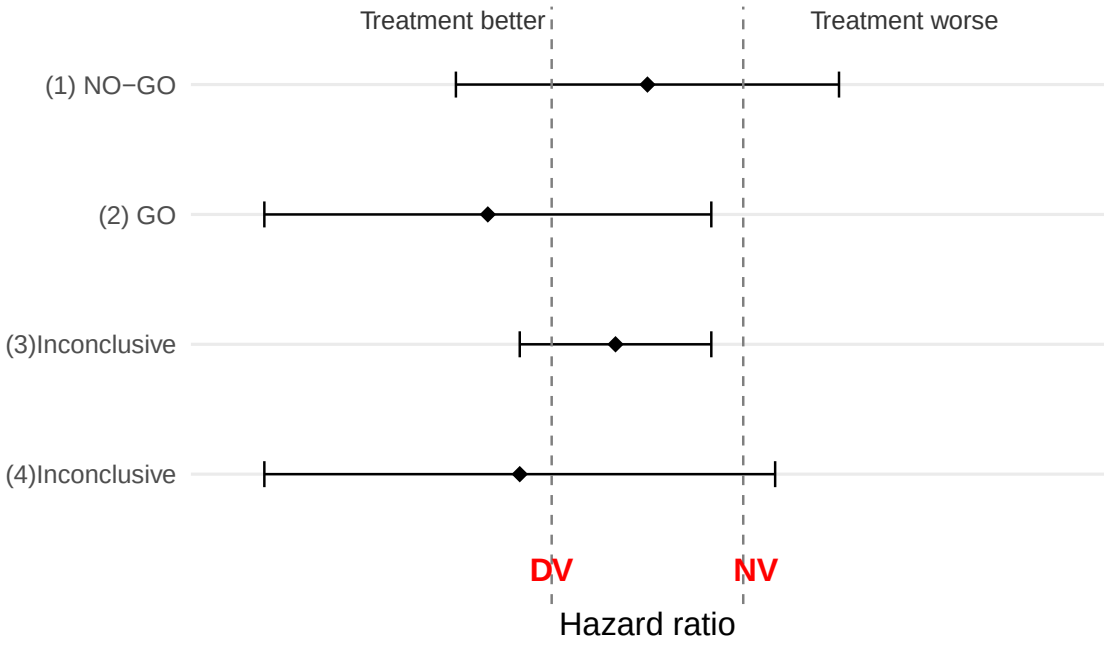
   Required inputs: type I error control (null hypothesis and type I error α) and a decision value (DV). The DV is same as the "target difference" in Fisch's paper. It is the minimal effect estimate needed for trail success (if higher than this value with moderate confidence, then a "GO" decision is made).

   By considering both criterion, we have both statistical significance and guarantees a sufficiently large effect estimate.

   The dual-criterion is more demanding, the resulting power of study is less than that of a standard design.

   Power is only increased for values superior to the DV since inferior values are clinically irrelevant.

   Decisions for dual-criterion design:



3. Precision design

   Doesn't rely on error rates. When null hypothesis or other benchmark values cannot be determined, this can be an option. It requires sufficiently precise effect estimate.

1.1 Hazard ratio and log hazard ratio

Consider hazard function in survival analysis, it describes the risk of failing. We consider hazard ratio between experimental drug and control as the outcome of interest. Hazard ratio(HR) less than 1 means the drug is better than the control. We want to have smaller hazard and hazard ratio, so that the drug is more effective.

In the paper, the log hazard ratio θ (log HR) is used instead of the hazard ratio. This could be because of the following reasons.

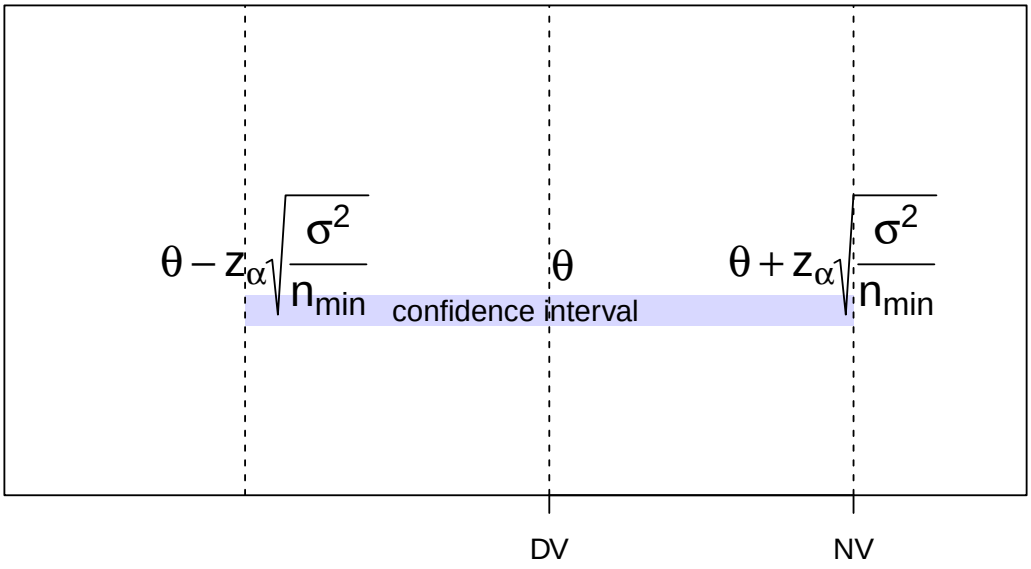
Firstly, according to the Central Limit Theorem, as the sample size increases, the distribution of the estimator (the log HR) approaches a normal distribution. Also, the HR itself is a ratio of hazard rates and is positively skewed, meaning it doesn't naturally fit a normal distribution. Taking the logarithm of the HR stabilizes its variance, transforming the skewed HR distribution into one that is more symmetric and closer to normal. Moreover, this approximate normality of log HR is useful in sample size calculation, which is discussed in the next session.

1.2 Sample size

Given the significance level α, the null value (NV) and the decision value (DV) for log HR, we can calculate the minimum sample size (for normally distributed data). Notice here that for log HR, DV $<$ NV, where DV, NV are log hazard ratios.

$$n_{\min} = \frac{\sigma^2 \times z_\alpha^2}{(\text{NV} - \text{DV})^2}$$

where σ is the outcome standard deviation, and takes the value 2 under equal randomization for the standard normal approximation to time-to-event data. z_α is the $100(1-\alpha)\%$ quantile of the standard normal distribution. $n_{\min}$ gives the minimum sample size that implies statistical significance if the effect estimate equals the DV. This value is calculated under the situation that both criterion are just satisfied. As illustrated in the below graph, when the effect estimate $\theta = \text{DV}$, and the lower bound of the confidence interval just touches the NV so that statistical significance is reached, we have the minimum sample size. Notice that when sample size equals the minimum sample size, the half-width of the confidence interval $z_\alpha\sqrt{\frac{\sigma^2}{n_{\min}}}$ equals NV-DV, so there will be no "Inconclusive" decisions. When the sample size is larger, the confidence interval becomes narrower, then an "Inconclusive" decision will occur.



1.3 Operating characteristics

The operating characteristics are the type I error and the power of the clinical trial design.
For dual-criterion designs, the power at the DV is approximately 50%, so that if the true parameter equals the DV, there is roughly equal chance that the effect estimate lies on either side of the DV. Having 50% at the DV does not mean the study is under-powered.

1.4 Reproduce Figure 1

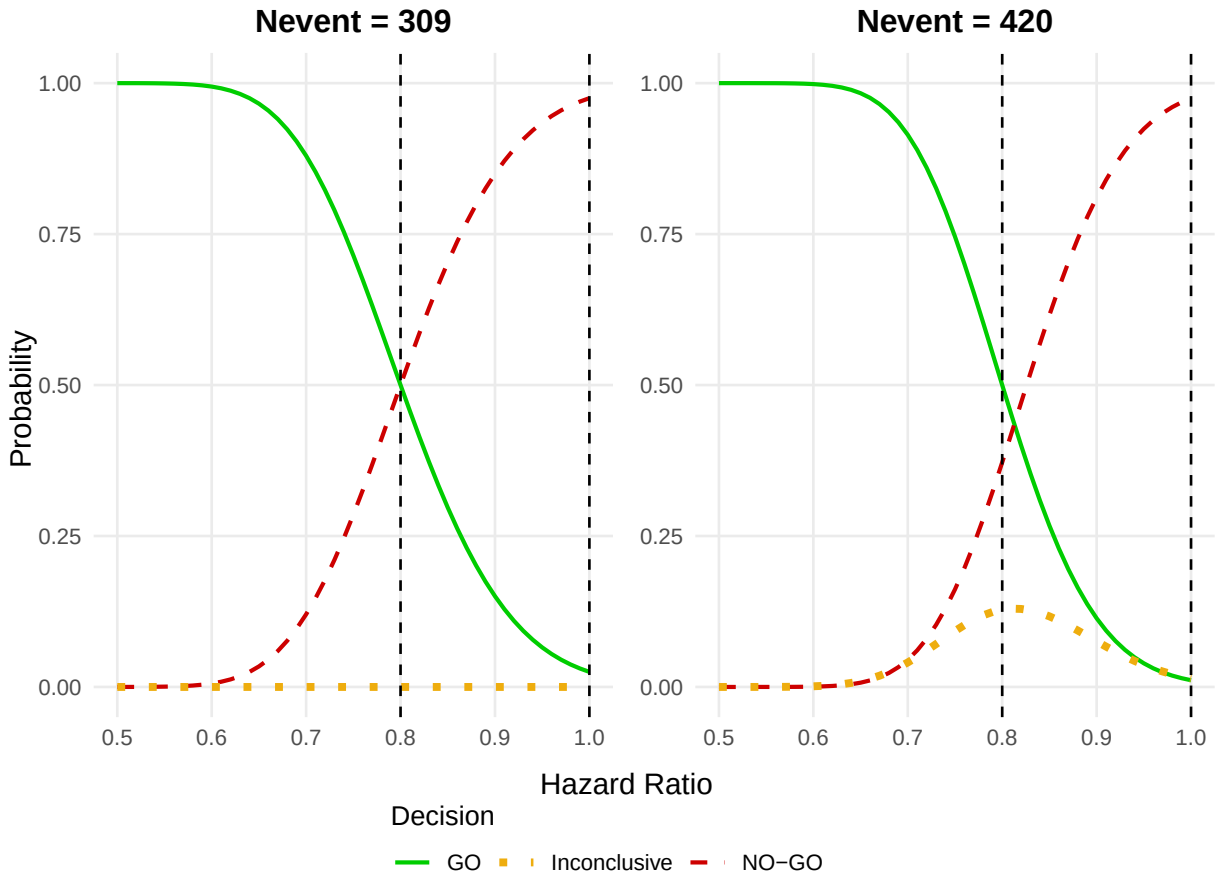
In Figure 1, the two plots illustrate the operating characteristics of dual-criterion designs with 309 and 420 events. The number 309 is the minimum sample size calculated under the example conditions $\sigma = 2$, $\alpha = 2.5\%$, NV $= \log(1)$ and DV $= \log(0.8)$ for log HR:

$$n_{\min} = \frac{2^2 \times z^2_{0.025}}{(\log 1 - \log 0.8)^2} = 308.594 \approx 309.$$

The probability of making a "GO" decision is the probability of the estimate smaller than the DV (i.e. clinical relevance) while the NV is outside the confidence interval (i.e. statistical significance). The "NO-GO" decision is made when the estimate is larger than the DV, and the NV is inside the confidence interval. The "Inconclusive" decisions are made if neither "GO" nor "NO-GO" is satisfied. The probability of "Inconclusive" decision is hence 1 minus the probability of "GO" and "NO-GO" decisions.
Below code presents the process of obtaining Figure 1 in the paper Roychoudhury et al. (2018). An extra notice here is the calculation of the cut value for statistical significance when the sample size is larger than $n_{\min}$. As we mentioned earlier, when $n > n_{\min}$, the confidence interval is shorter, so the cut value `cut.ssig` of the "NO-GO" decision should be a value between DV and NV such that $z_\alpha \sqrt{\frac{\sigma^2}{n}} =$ NV $-$ `cut.ssig`. So in this case, the `cut.ssig` $= \log(1) - z_\alpha \sqrt{\frac{\sigma^2}{n}}$.

```r
# Sequence of true hazard ratios in log scale
t.d <- log(seq(0.5, 1, 0.01))

# Left panel (n = 309)
n1 <- 309
sd1 <- sqrt((2^2) / n1) # standard deviation
cut.ssig1 <- log(0.8) # cutting point for statistical significance
cut.crel1 <- log(0.8) # cutting point for clinical relevance
pp.go1 <- pnorm(cut.crel1, t.d, sd1) # probability of GO decision
pp.ngo1 <- 1 - pnorm(cut.ssig1, t.d, sd1) # probability of NO-GO decision
pp.intd1 <- 1 - pp.go1 - pp.ngo1 # probability of inconclusive decision
df1 <- data.frame(HazardRatio = exp(t.d), GO = pp.go1,
                  NOGO = pp.ngo1, Inconclusive = pp.intd1)

# Right panel (n = 420)
n2 <- 420
sd2 <- sqrt((2^2) / n2) # standard deviation
cut.ssig2 <- log(1) - qnorm(0.975) * sqrt(2^2 / 420) # statistical significance
cut.crel2 <- log(0.8) # clinical relevance
pp.go2 <- pnorm(cut.crel2, t.d, sd2) # probability of GO decision
pp.ngo2 <- 1 - pnorm(cut.ssig2, t.d, sd2) # probability of NO-GO decision
pp.intd2 <- 1 - pp.go2 - pp.ngo2 # probability of inconclusive decision
df2 <- data.frame(HazardRatio = exp(t.d), GO = pp.go2,
                  NOGO = pp.ngo2, Inconclusive = pp.intd2)

# Define the line colors and types
line_colors <- c("GO" = "green3", "Inconclusive" = "darkgoldenrod2", "NO-GO" = "red3")
line_types <- c("GO" = "solid", "Inconclusive" = "dotted", "NO-GO" = "dashed")

# Plot for n = 309
p1 <- ggplot(df1, aes(x = HazardRatio)) +
  geom_line(aes(y = GO, color = "GO", linetype = "GO"),size=0.8) +
  geom_line(aes(y = NOGO, color = "NO-GO", linetype = "NO-GO"),size=0.8) +
  geom_line(aes(y = Inconclusive, color = "Inconclusive",
                linetype = "Inconclusive"),size=1.5) +
  geom_vline(xintercept = c(0.8, 1.0), linetype = "dashed", color = "black") +
  labs(title = "Nevent = 309") +
  scale_color_manual(values = line_colors) +
  scale_linetype_manual(values = line_types) +
  theme_minimal() +
  theme(
    legend.position = "none",
    plot.title = element_text(hjust = 0.5, face = "bold"),
```

```
    panel.grid.minor = element_blank(),
    axis.title.y = element_blank(),  # Remove individual y-labels
    axis.title.x = element_blank()
  )

# Plot for n = 420
p2 <- ggplot(df2, aes(x = HazardRatio)) +
  geom_line(aes(y = GO, color = "GO", linetype = "GO"),size=0.8) +
  geom_line(aes(y = NOGO, color = "NO-GO", linetype = "NO-GO"),size=0.8) +
  geom_line(aes(y = Inconclusive, color = "Inconclusive",
                linetype = "Inconclusive"),size=1.5) +
  geom_vline(xintercept = c(0.8, 1.0), linetype = "dashed", color = "black") +
  labs(title = "Nevent = 420") +
  scale_color_manual(values = line_colors) +
  scale_linetype_manual(values = line_types) +
  theme_minimal() +
  theme(
    legend.position = "none",
    plot.title = element_text(hjust = 0.5, face = "bold"),
    panel.grid.minor = element_blank(),
    axis.title.y = element_blank(),  # Remove individual y-labels
    axis.title.x = element_blank()
  )

# Combine the plots into a single figure without individual y-axis labels
combined_plots <- plot_grid(p1, p2, ncol = 2, align = 'hv', rel_widths = c(1, 1))

# Extract and create a shared legend
legend <- get_legend(
  p1 + theme(legend.position = "right") +
    guides(color = guide_legend(title = "Decision", nrow = 1),
           linetype = guide_legend(title = "Decision", nrow = 1)))

# Add a shared y-axis label using grid.arrange
final_plot <- grid.arrange(
  arrangeGrob(combined_plots,
              left = textGrob("Probability", rot = 90, vjust = 1.2),
              bottom = textGrob("Hazard Ratio",just = "centre")),
  legend = legend,
  ncol = 1,
  heights = c(10, 1)
)
```




```
# Print the final plot
print(final_plot)
```

2 Example 1: A randomized PoC design with time-to-event data

Randomized, double-blind, RCT. Patients were randomized equally to: (experimental drug + standard care) OR (standard care only).
Primary outcome of interest (or called "endpoint") is the progression-free survival (PFS), which is the time when the disease or cancer do not get worse. The endpoint was assessed with a *log-rank test* and *Cox regression* with treatment as a covariate.

- *log-rank test*: compare the survival distributions of two or more groups. It tests the hypothesis that there is no difference in survival (or time-to-event) between the two groups. If the log-rank test indicates a significant difference, it suggests the treatment affects how long patients live without their disease worsening.

- *Cox regression*: estimate the hazard ratio between two groups, which tells us the relative risk of disease progression in the treatment group compared to the control group. If the HR < 1, it suggests that the new treatment delays disease progression better than the control treatment.

As for the DV, HR=0.7 was deemed necessary to be clinically meaningful. Values larger than 0.7 are unsatisfactory to clearly justify further development of the drug.
So the dual-criterion is:

1. Statistical significance: one-sided p-value of log-rank test ≤ 0.1.
2. Clinical relevance: estimated HR from Cox regression ≤ 0.70.

2.1 Reproduce Table 3

Here we attempt to reproduce the values in Table 3.
For the first sub-table, similar to Section 1.4, when $n > n_{\min}$, the cut value $\hat{\theta}$ (`cut.ssig`) of the "NO-GO" decision is given by $\texttt{cut.ssig} = \log(1) - z_\alpha\sqrt{\frac{\sigma^2}{n}}$. These correspond to $\hat{\theta} > 0.736 \neq 0.7$ from the paper.
For the second sub-table, $n = n_{\min}$.
For the third sub-table, it requires one-sided type I error of 0.1 and power of 0.9 for HR=0.5. So the cut value for clinical relevance is chosen to satisfy these requirements. The authors used 0.901 as the power for HR=0.5, rather than 0.9, probably because this ensures a power of at least 0.9.
For the forth sub-table, it requires one-sided type I error of 0.1 and power of 0.8 for HR=0.5. The sample size $n = 38 < n_{\min}$. The cut value of the decisions are calculated in a similar way as in Section 1.4. The cut value of $\hat{\theta}$ is slightly different from 0.659 as claimed in the paper, very possibly because of rounding.
For sub-table five, it used a type I error of 0.2 and a power of 0.9, with sample size $n = 38 < n_{\min}$. The cut value of the decisions are calculated in a similar way as in Section 1.4.

```
# minimum sample size
n.min <- ceiling((4*qnorm(0.9)^2)/(log(1)-log(0.7))^2) # n.min = 52

# a sequence of true log(HR).
t.d <- log(seq(0.5, 1, 0.1))

# Dual-criterion design: alpha=0.1, DV=log(0.7), n=70
n1 <- 70
sd1 <- sqrt((2^2)/n1) # standard deviation
cut.ssig <- log(1)-qnorm(0.9)* sqrt(2^2 / n1) # statistical significance
cut.crel <- log(0.7) # critical relevance
pp.go1 <- pnorm(cut.crel, t.d, sd1)
pp.ngo1 <- 1- pnorm(cut.ssig, t.d, sd1)
pp.intd1 <- 1 -pp.go1 - pp.ngo1

subtable1 <- matrix(data=round(c(exp(t.d), pp.go1,pp.ngo1,pp.intd1),3), ncol=4)
colnames(subtable1) <- c("True HR", "GO: $\\hat{\\theta} \\leq 0.7$",
                         "NO-GO: $\\hat{\\theta} > 0.736$","Inconclusive")

# Dual-criterion design: alpha=0.1, DV=0.7, n=52
n2 <- 52
sd2 <- sqrt((2^2)/n2)
cut.ssig <- log(0.7)
cut.crel <- log(0.7)
```

```
pp.go2 <- pnorm(cut.crel, t.d, sd2)
pp.ngo2 <- 1-pnorm(cut.ssig, t.d, sd2)
pp.intd2 <- 1 -pp.go2 - pp.ngo2

subtable2 <- matrix(data=round(c(exp(t.d), pp.go2,pp.ngo2,pp.intd2),3), ncol=4)
colnames(subtable2) <- c("True HR", "GO: $\\hat{\\theta} \\leq 0.7$",
                         "NO-GO: $\\hat{\\theta} > 0.7$","Inconclusive")

# Dual-criterion design: alpha=0.1, beta=0.1, n=55
n3 <- 55
sd3 <- sqrt((2^2)/n3)
cut.ssig <- qnorm(0.901,log(0.5),sd3)
cut.crel <- qnorm(0.901,log(0.5),sd3)
pp.go3 <- pnorm(cut.crel, t.d, sd3)
pp.ngo3 <- 1-pnorm(cut.ssig, t.d, sd3)
pp.intd3 <- 1 -pp.go3 - pp.ngo3

subtable3 <- matrix(data=round(c(exp(t.d), pp.go3,pp.ngo3,pp.intd3),3), ncol=4)
colnames(subtable3) <- c("True HR", "GO: $\\hat{\\theta} \\leq 0.708$",
                         "NO-GO: $\\hat{\\theta} > 0.708$","Inconclusive")

# Dual-criterion design: alpha=0.1, beta=0.2, n=38
n4 <- 38
sd4 <- sqrt((2^2)/n4)
cut.ssig <- log(1)-qnorm(0.9)* sqrt(2^2 / n4) # different from 0.659
cut.crel <- log(1)-qnorm(0.9)* sqrt(2^2 / n4)
pp.go4 <- pnorm(cut.crel, t.d, sd4)
pp.ngo4 <- 1-pnorm(cut.ssig, t.d, sd4)
pp.intd4 <- 1 -pp.go4 - pp.ngo4

subtable4 <- matrix(data=round(c(exp(t.d), pp.go4,pp.ngo4,pp.intd4),3), ncol=4)
colnames(subtable4) <- c("True HR", "GO: $\\hat{\\theta} \\leq 0.659$",
                         "NO-GO: $\\hat{\\theta} > 0.659$","Inconclusive")

# Dual-criterion design: alpha=0.2, beta=0.1, n=38
n5 <- 38
sd5 <- sqrt((2^2)/n5)
cut.ssig <- log(1)-qnorm(0.8)* sqrt(2^2 / n5)
cut.crel <- log(1)-qnorm(0.8)* sqrt(2^2 / n5)
pp.go5 <- pnorm(cut.crel, t.d, sd5)
pp.ngo5 <- 1-pnorm(cut.ssig, t.d, sd5)
pp.intd5 <- 1 -pp.go5 - pp.ngo5
```

```
subtable5 <- matrix(data=round(c(exp(t.d), pp.go5,pp.ngo5,pp.intd5),3), ncol=4)
colnames(subtable5) <- c("True HR", "GO: $\\hat{\\theta} \\leq 0.761$",
                         "NO-GO: $\\hat{\\theta} > 0.761$","Inconclusive")
```

1. Dual-criterion design: $\alpha = 0.1$, DV=0.7, n=70

True HR	GO: $\hat{\theta} \leq 0.7$	NO-GO: $\hat{\theta} > 0.736$	Inconclusive
0.5	0.920	0.053	0.027
0.6	0.740	0.196	0.063
0.7	0.500	0.417	0.083
0.8	0.288	0.636	0.076
0.9	0.147	0.800	0.054
1.0	0.068	0.900	0.032

2. Dual-criterion design: $\alpha = 0.1$, DV=0.7, n=52

True HR	GO: $\hat{\theta} \leq 0.7$	NO-GO: $\hat{\theta} > 0.7$	Inconclusive
0.5	0.887	0.113	0
0.6	0.711	0.289	0
0.7	0.500	0.500	0
0.8	0.315	0.685	0
0.9	0.182	0.818	0
1.0	0.099	0.901	0

3. Dual-criterion design: $\alpha = 0.1, \beta = 0.1(\theta_A = 0.5)$, n=55

True HR	GO: $\hat{\theta} \leq 0.708$	NO-GO: $\hat{\theta} > 0.708$	Inconclusive
0.5	0.901	0.099	0
0.6	0.729	0.271	0
0.7	0.516	0.484	0
0.8	0.324	0.676	0
0.9	0.186	0.814	0
1.0	0.100	0.900	0

4. Dual-criterion design: $\alpha = 0.1, \beta = 0.2(\theta_A = 0.5)$, n=38

True HR	GO: $\hat{\theta} \leq 0.659$	NO-GO: $\hat{\theta} > 0.659$	Inconclusive
0.5	0.804	0.196	0
0.6	0.615	0.385	0
0.7	0.428	0.572	0
0.8	0.276	0.724	0
0.9	0.169	0.831	0
1.0	0.100	0.900	0

5. Dual-criterion design: $\alpha = 0.2, \beta = 0.1(\theta_A = 0.5)$, n=38

True HR	GO: $\hat{\theta} \leq 0.761$	NO-GO: $\hat{\theta} > 0.761$	Inconclusive
0.5	0.902	0.098	0
0.6	0.768	0.232	0
0.7	0.602	0.398	0
0.8	0.439	0.561	0
0.9	0.303	0.697	0
1.0	0.200	0.800	0

3 Example 2: A single-arm PoC design with binary data

Experimental drug in Chinese patients with non-small-cell lung cancer.
Primary endpoint is objective response rate (ORR), which quantifies the preliminary efficacy of the experimental drug.
Because of the absence of a comparator (in single arm trials), NV is set to 7.5% rather than 0, based on a literature review and clinical discussions. Hence the prior is chosen as a minimally informative unimodal beta prior distribution Beta(0.0811, 1), which gives a mean of 0.075.
A minimum improvement of 10% is considered necessary for further development, so the DV is set to be 10%+7.5%=17.5%. Notice in this example, NV < DV.
So the dual-criterion is:

1. Bayesian statistical significance: $\mathbb{P}(\text{ORR} \geq 7.5\%|\text{data}) \geq 0.95$
2. Clinical relevance: Posterior median $\geq 17.5\%$

Null hypothesis: there is no effect of the drug, i.e. ORR=7.5%
$\mathbb{P}(\text{type I error}) = \mathbb{P}(\text{reject } H_0|H_0 \text{ is true}) = \mathbb{P}(\text{reject } H_0|\text{ORR} \leq 7.5\%)$
$\mathbb{P}(\text{type II error}) = \mathbb{P}(\text{not reject } H_0|H_0 \text{ is false}) = \mathbb{P}(\text{reject } H_0|\text{ORR} = \text{response rate})$

3.1 Reproduce Figure 2

For this dual-criterion, the minimally required sample size is 22. For $n \geq 22$, clinical relevance ensures statistical significance. To see this, we are now going to reproduce Figure 2 in the paper.
For a Beta(a,b) prior,

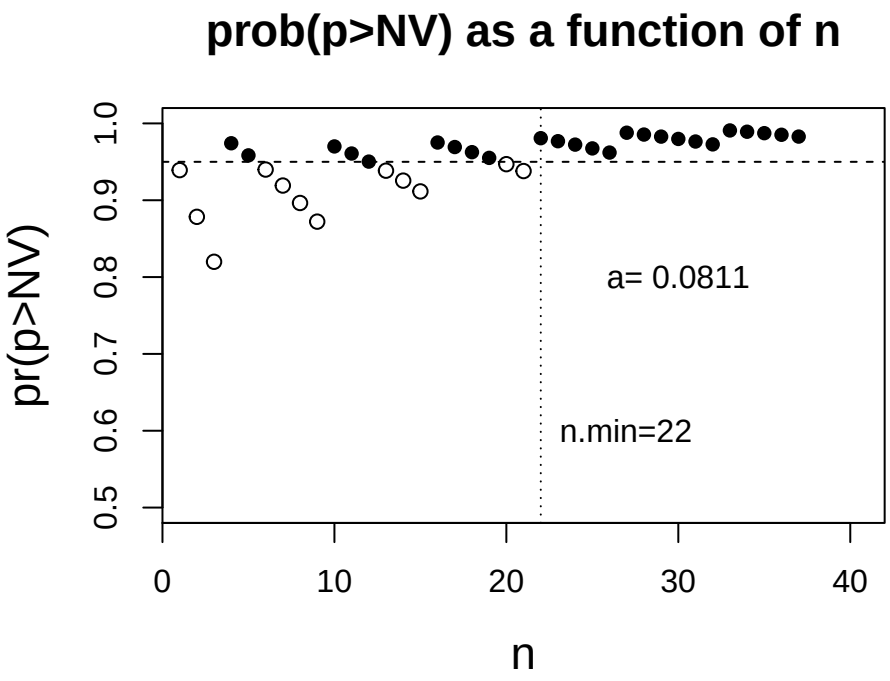
$$\text{Beta}(a, b) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a, b)}$$

with binomial sampling (according to the paper), if we observe a number of successes (r) and failures ($n - r$), then the posterior is Beta($a + r$, $b + n - r$).
In the below code, we first find the required numbers of responders (`rr`) to satisfy clinical relevance under different sample sizes (n), then we find the minimal sample size n such that clinical relevance guarantees Bayesian statistical significance.
The minimal sample size is 22. Final sample size is taken to be 25.

```r
#This function calculates the minimal sample size
# for Bayesian DC design with binary endpoint
SS_DC_BayesBin <- function(
  null.value, # null.value
  p.positive  = 0.95, # required probability to be better than the null
  dec.value,  # decision value
  post.est = c("median","mean")[1], # median by default
  a=1, b=1,    # prior specification, a Beta(a,b) prior
  n.max = 1000 # numbers of iterations
){
  NV = null.value
  DV = dec.value
  n <- 1:n.max

  # required number of responders to satisfy DV criterion (posterior mean)
  if (post.est == "mean") { # if estimate posterior mean
    rr = rep(NA,max(n)) # a vector to store rr (required responders)
      for ( nn in n) { # n iterations
      r = 0:nn # different numbers of responders
      est1 = (a+r)/(a+b+nn) # posterior estimate based on different success numbers
      rr[nn] = r[est1>=DV][1] # minimum r s.t. DV criterion is fulfilled
      }
    print(cbind(rr,n)) # required responders for each sample size n
    est = (a+rr)/(a+b+n) # posterior estimate of mean
  }

  # required number of responders to satisfy DV criterion (posterior median)
  if (post.est == "median") {
    rr = rep(NA,max(n))
    for ( nn in n) {
      r = 0:nn
      est1= qbeta(0.5,a+r,b+nn-r) # the median of beta posterior
      rr[nn] = r[est1>=DV][1] # minimum r s.t. posterior median >= DV
    }
    print(cbind(rr,n))
    est = qbeta(0.5,a+rr,b+n-rr)
  }

  p.obs = rr/n

  # find n such that posterior prob >= p.positive (= 0.95)
  # in other words, find n such that statistical significance is guaranteed.
  post.crit <- 1 - pbeta(NV,a+rr,b+n-rr) # posterior probability of being greater than NV.
```

```
  post.crit.index = post.crit >= p.positive # index=TRUE if the probability >= 0.95

  # find n such that for all m>=n, statistical criterion is fulfilled
  find.n =  cbind( n, rev(post.crit.index),
                   cumsum(rev(post.crit.index))==n,
                   # true if all m>=n fulfill stats criterion
                   rev(n))
  colnames(find.n) = c("index","okay","all.next.okay","n")

  n.min=NULL

  if (any(find.n[,3]==1) ) {
    n.okay = find.n[find.n[,3]==1,,drop=FALSE] # take all the rows with all.next.okay==1
    n.min = n.okay[nrow(n.okay),4] # take the n in the last row (find minimum n)
    n.min.15 = n.min+15
    est=round(est,3)
    ss.table = data.frame( n=n[1:n.min.15],
                           # sample size from 1 to n.min+10
                           r=rr[1:n.min.15],
                           # number of responders needed for DV criterion
                           p.obs=p.obs[1:n.min.15],
                           est = est[1:n.min.15],
                           # posterior estimate of median (by default)
                           post.p.positive=post.crit[1:n.min.15],
                           # posterior probability of a positive effect
                           okay=post.crit.index[1:n.min.15]
                           # TRUE if statistical significance is satisfied
                           )}

  if (is.null(n.min))
    stop("No sample size satisfies required posterior
         probability criterion pr(p>NV): try higher n.max")

plot( 1:(n.min+15), post.crit[1:(n.min+15)], type="p", xaxs="i", xlab="n",
        ylab="pr(p>NV)",xlim = c(0,n.min+20), ylim=c(0.5,1),
        lwd=1, pch = ifelse(post.crit.index,16,1),
        main="prob(p>NV) as a function of n", cex.lab=1.4, cex.main=1.4)
  abline(h=p.positive, lty=2)
  abline(v=n.min, lty=3)
  text(1.05*n.min, 0.6, paste("n.min=",n.min, sep=''),adj=0)
  text(30, 0.8, paste("a=",round(a,4)))

  return(list(ss.table=ss.table,n.min=n.min))}
```

```r
#Example
nB <-  SS_DC_BayesBin(null.value = 0.075, p.positive = 0.95,
                      dec.value = 0.175, a = 0.0811, b=1,
                      post.est = "median")
```



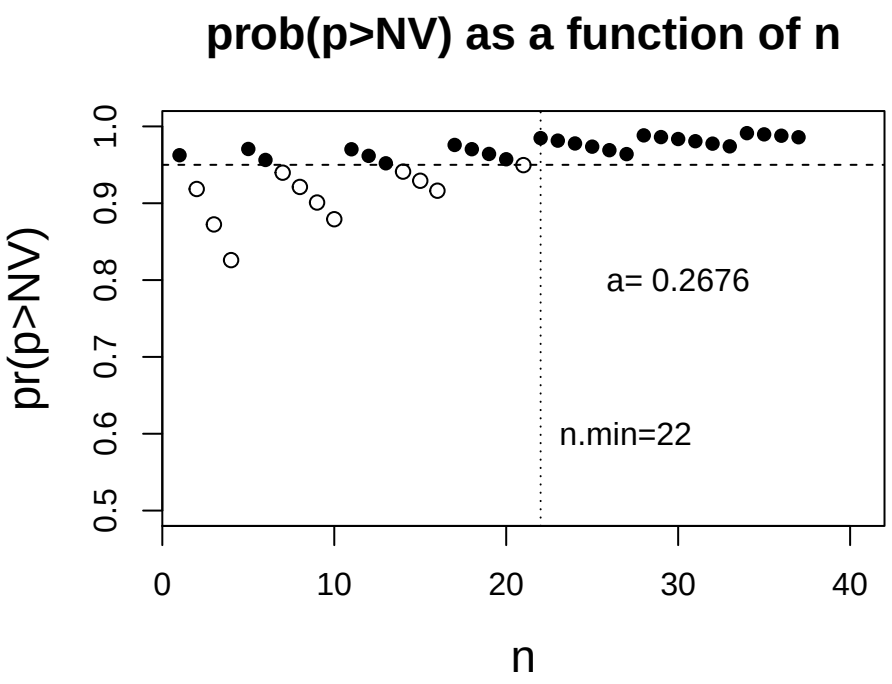
```
# returns a table of minimal sample size
kable(nB$ss.table)
```

n	r	p.obs	est	post.p.positive	okay
1	1	1.0000000	0.527	0.9392105	FALSE
2	1	0.5000000	0.316	0.8784199	FALSE
3	1	0.3333333	0.224	0.8199084	FALSE
4	2	0.5000000	0.397	0.9741575	TRUE
5	2	0.4000000	0.324	0.9584199	TRUE
6	2	0.3333333	0.274	0.9399281	FALSE
7	2	0.2857143	0.237	0.9191247	FALSE
8	2	0.2500000	0.209	0.8964143	FALSE
9	2	0.2222222	0.187	0.8721628	FALSE
10	3	0.3000000	0.264	0.9700500	TRUE
11	3	0.2727273	0.241	0.9607987	TRUE
12	3	0.2500000	0.222	0.9502624	TRUE
13	3	0.2307692	0.205	0.9384881	FALSE
14	3	0.2142857	0.191	0.9255363	FALSE
15	3	0.2000000	0.179	0.9114783	FALSE
16	4	0.2500000	0.229	0.9751277	TRUE
17	4	0.2352941	0.215	0.9692226	TRUE
18	4	0.2222222	0.204	0.9625583	TRUE
19	4	0.2105263	0.193	0.9551275	TRUE
20	4	0.2000000	0.184	0.9469304	FALSE
21	4	0.1904762	0.175	0.9379739	FALSE
22	5	0.2272727	0.212	0.9807693	TRUE
23	5	0.2173913	0.203	0.9768319	TRUE
24	5	0.2083333	0.195	0.9724075	TRUE
25	5	0.2000000	0.187	0.9674799	TRUE
26	5	0.1923077	0.180	0.9620360	TRUE
27	6	0.2222222	0.210	0.9878835	TRUE
28	6	0.2142857	0.202	0.9854972	TRUE
29	6	0.2068966	0.196	0.9828022	TRUE
30	6	0.2000000	0.189	0.9797817	TRUE
31	6	0.1935484	0.183	0.9764197	TRUE
32	6	0.1875000	0.177	0.9727022	TRUE
33	7	0.2121212	0.202	0.9908432	TRUE
34	7	0.2058824	0.196	0.9891762	TRUE
35	7	0.2000000	0.191	0.9872993	TRUE
36	7	0.1944444	0.185	0.9851992	TRUE
37	7	0.1891892	0.180	0.9828627	TRUE

Additional Table: Minimal sample size required for the analysis.

Now we explore how the prior choice affect the minimum sample size. Firstly, it might be more reasonable to choose the prior with median 0.075, rather than with mean 0.075. So we try Beta(0.2676,1) as the prior, as it has median 0.075. The below codes and plot explore how this affects the minimum sample size. From the plot we can spot some slight difference compared to the previous plot, but their difference is not significant.

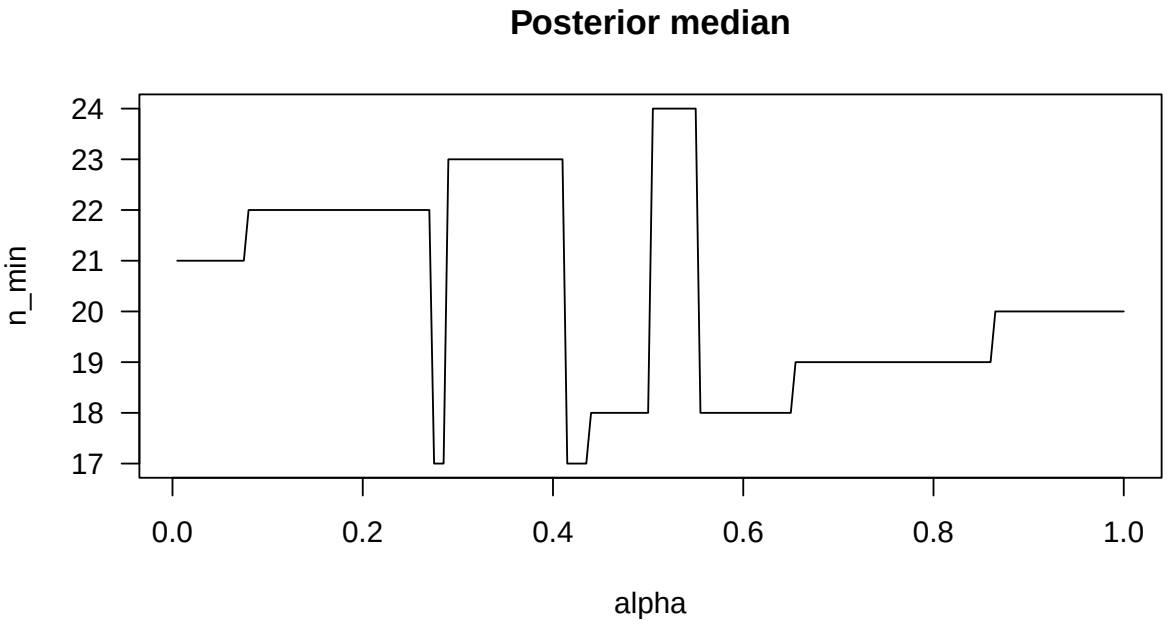
```r
# First try beta prior with median 0.075, rather than mean 0.075
# Target median
median_target <- 0.075
beta_param <- 1

# Function to find alpha such that the median of Beta(alpha, beta=1) is 0.075
find_alpha <- function(alpha) {
  abs(pbeta(median_target, alpha, beta_param) - 0.5)
}

# Use optimization to find alpha
alpha_result <- optimize(find_alpha, c(0.001, 10))$minimum
alpha_result # 0.2676044

nM <-  SS_DC_BayesBin(null.value = 0.075, p.positive = 0.95,
                      dec.value = 0.175, a = alpha_result, b=1,
                      post.est = "median")
```

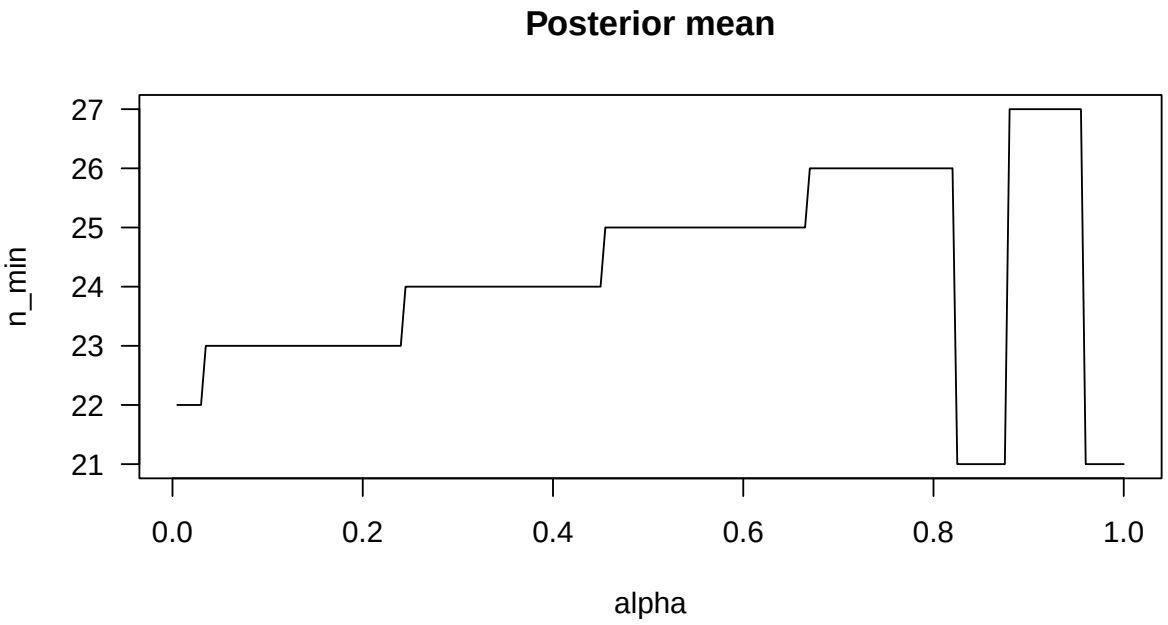


```r
nM$n.min # is still 22
```

Additionally, we explore how the minimum sample size changes according to different values of a in the Beta(a, b) prior. The below plots give the relationship between the minimal sample size and the choices of a, with the median and mean being the posterior estimate, respectively. We can see from the plots that there are jumps of minimal sample sizes around some values of a, especially in the first plot. When the posterior estimate is the mean (the second plot), the jumping pattern only seem to occur for a values closer to 1. This might suggest that the posterior mean gives more stability and less sensitivity of minimal sample size with respect to prior choices. For both cases, we obtain the values of a at these jumping points and included relevant plots in the Appendix.

```r
# Next plot the minimum sample size VS a (in Beta prior)
alpha_vector <- seq(0.005,1,0.005)

# posterior median
n_min_vector_median <- rep(0, length(alpha_vector))
for (i in 1:length(alpha_vector)){
  a <- alpha_vector[i]
  n <- SS_DC_BayesBin(null.value = 0.075, p.positive = 0.95,
                      dec.value = 0.175, a = a, b=1,
                      post.est = "median")
  n_min_vector_median[i] <- n$n.min[[1]]
}
# posterior mean
n_min_vector_mean <- rep(0, length(alpha_vector))
for (i in 1:length(alpha_vector)){
  a <- alpha_vector[i]
  n <- SS_DC_BayesBin(null.value = 0.075, p.positive = 0.95,
                      dec.value = 0.175, a = a, b=1,
                      post.est = "mean")
  n_min_vector_mean[i] <- n$n.min[[1]]
}
```





3.2 Reproduce Table 4

Table 4 results show that this dual-criterion design is a three-outcome design with desirable properties.

Introduction of the three-outcome design Sargent et al. (2001).

A design with three possible outcomes includes an additional region, called the inconclusive region, between these two regions. This requires two cut-off points, r and s, with the following decision rule

If $Y \leq r$, reject H_1

If $r < Y < s$, inconclusive
If $s \leq Y$, reject H_0
Define the following parameters

$$\alpha = \max_{p\in(0,p_0]} \mathbb{P}(\text{reject } H_0|H_0 \text{ true}) = \mathbb{P}(Y \geq s|p = p_0)$$
$$\beta = \max_{p\in[p_1,1)} \mathbb{P}(\text{reject } H_1|H_1 \text{ true}) = \mathbb{P}(Y \leq r|p = p_1)$$
$$\eta = \max_{p\in(0,p_0]} \mathbb{P}(\text{reject } H_1|H_0 \text{ true}) = \mathbb{P}(Y \leq r|p = p_0)$$
$$\pi = \max_{p\in[p_1,1)} \mathbb{P}(\text{reject } H_0|H_1 \text{ true}) = \mathbb{P}(Y \geq s|p = p_1)$$

where α is the maximum probability of making a wrong decision by rejecting the null hypothesis when it is true, and η is the minimum probability of making the decision to reject alternative hypothesis when the null is true. In contrast to the standard two-outcome design, $\eta + \alpha < 1$ due to the presence of the inconclusive region. The same interpretation holds true for π and β relative to the alternative hypothesis.
Define $\lambda := \mathbb{P}(r < Y < s|p = p_0)$ and $\delta := \mathbb{P}(r < Y < s|p = p_1)$, then we have

$$\eta + \lambda + \alpha = 1, \qquad \beta + \delta + \pi = 1$$

In the paper "Beyond p-values: A phase II dual-criterion design with statistical significance and clinical relevance", they consider a three-outcome design with H_0 : ORR $\leq 7.5\%$, H_1 : ORR $\geq 27.5\%$, $\alpha = 0.05$, $\beta = 0.1$, $\eta = 0.8$, $\pi = 0.9$, and the minimal sample size is 27.
The below code reproduce Table 4. The first design is the situation when sample size is the minimum sample size, at which the inconclusive case does not happen. The second design has larger sample size ($n = 36$), where an inconclusive case will happen. For a "GO" decision, we need the number of responders to be larger or equal to 7 according to Additional Table in the sample size discussion, so that the clinical relevance guarantees the statistical significance. For "NO-GO" decision, the maximum number of responders are chosen when the statistical significance is just not satisfied. The third design is the three-outcome design, where the cut values of numbers of responders are chosen based on the values of η and α.

```
# a vector of true ORR values
ORR.t <- c(0.075, 0.125, 0.175, 0.225, 0.275)
a = 0.0811
b=1

# Dual-criterion design: alpha=0.05, DV=0.175, n=25
n1 <- 25 # is the minimum sample size
DV <- 0.175
NV <- 0.075
alpha <- 0.05

# required number of responders for statistical stignificance
```

```r
cut.ssig <- nB$ss.table$r[nB$ss.table$n==n1]
# required number of responders for clinical relevance
cut.crel <- nB$ss.table$r[nB$ss.table$n==n1]

# calculate probabilities
p_GO1 <- round(1-pbinom(cut.crel-1, n1, prob=ORR.t, lower.tail = T),3)
p_NOGO1 <- round(pbinom(cut.ssig-1, n1, prob=ORR.t, lower.tail = T),3)
p_inconcl1 <- round(1- p_GO1 - p_NOGO1,3)

subtable_1 <- matrix(data=c(ORR.t*100, round(c(p_GO1,p_NOGO1,p_inconcl1),3)), ncol=4)
colnames(subtable_1) <- c("True ORR(\\%)", "GO: $r \\geq 5$",
                          "NO-GO: $r < 5$","Inconclusive")

# Dual-criterion design: alpha=0.05, DV=0.175, n=36
n2 <- 36
DV <- 0.175
NV <- 0.075
alpha <- 0.05

# required number of responders for clinical relevance
cut.crel <- nB$ss.table$r[nB$ss.table$n==n2]
# find cut.ssig: required number of responders for statistical significance
r <- seq(0,10,1)
crit <- 1 - pbeta(NV,a+r,b+n2-r) # posterior probability of being less than NV.
crit.index = crit < 1-alpha
cut.ssig <- max(r[crit.index])

# calculate probabilities
p_GO2 <- round(1-pbinom(cut.crel-1, n2, prob=ORR.t, lower.tail = T),3)
p_NOGO2 <- round(pbinom(cut.ssig, n2, prob=ORR.t, lower.tail = T),3)
p_inconcl2 <- round(1- p_GO2 - p_NOGO2,3)

subtable_2 <- matrix(data=c(ORR.t*100, round(c(p_GO2,p_NOGO2,p_inconcl2),3)), ncol=4)
colnames(subtable_2) <- c("True ORR(\\%)", "GO: $r \\geq 7$",
                          "NO-GO: $r \\leq 5$","Inconclusive: r = 6")

# Three-outsome design: n=27, H0:ORR<0.075, H1:ORR>=0.275,
# alpha=0.05, beta=0.1, eta=0.8, pi=0.9
n3 <- 27 # is the minimum sample size under three-criterion design
eta <- 0.8
alpha <- 0.05

# find s and r for three-outcome design
```

```
s <- qbinom(alpha, n3, prob=0.075, lower.tail = F)+1
r <- qbinom(eta, n3, prob=0.075, lower.tail = T)

# calculate probabilities
p_GO3 <- round(1-pbinom(s-1, n3, prob=ORR.t, lower.tail = T),3) # Y >= s
p_NOGO3 <- round(pbinom(r, n3, prob=ORR.t, lower.tail = T),3) # Y <= r
p_inconcl3 <- round(1- p_GO3 - p_NOGO3, 3)

subtable_3 <- matrix(data=c(ORR.t*100, round(c(p_GO3,p_NOGO3,p_inconcl3),3)), ncol=4)
colnames(subtable_3) <- c("True ORR(\\%)", "GO: $r \\geq 5$",
                          "NO-GO: $r \\leq 3$","Inconclusive: r = 4")
```

1. Dual-criterion design: $\alpha = 0.1$, DV=0.7, n=70

True ORR(%)	GO: $r \geq 5$	NO-GO: $r < 5$	Inconclusive
7.5	0.036	0.964	0
12.5	0.195	0.805	0
17.5	0.451	0.549	0
22.5	0.693	0.307	0
27.5	0.858	0.142	0

2. Dual-criterion design: $\alpha = 0.1$, DV=0.7, n=52

True ORR(%)	GO: $r \geq 7$	NO-GO: $r \leq 5$	Inconclusive: r = 6
7.5	0.016	0.950	0.034
12.5	0.156	0.709	0.135
17.5	0.446	0.380	0.174
22.5	0.731	0.149	0.120
27.5	0.902	0.044	0.054

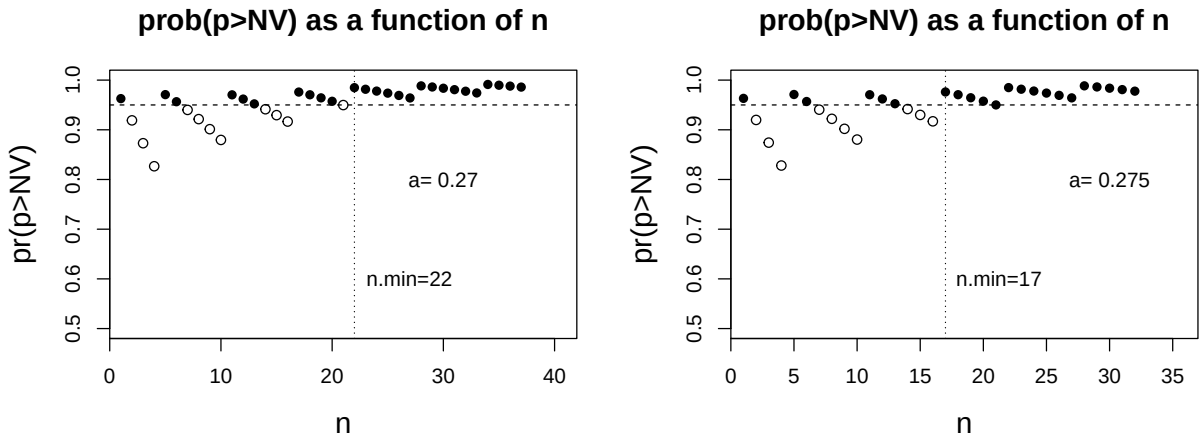
3. Three-outcome design: $n = 27$

True ORR(%)	GO: $r \geq 5$	NO-GO: $r \leq 3$	Inconclusive: r = 4
7.5	0.048	0.860	0.092
12.5	0.243	0.558	0.199
17.5	0.523	0.280	0.197
22.5	0.759	0.113	0.128
27.5	0.901	0.038	0.061

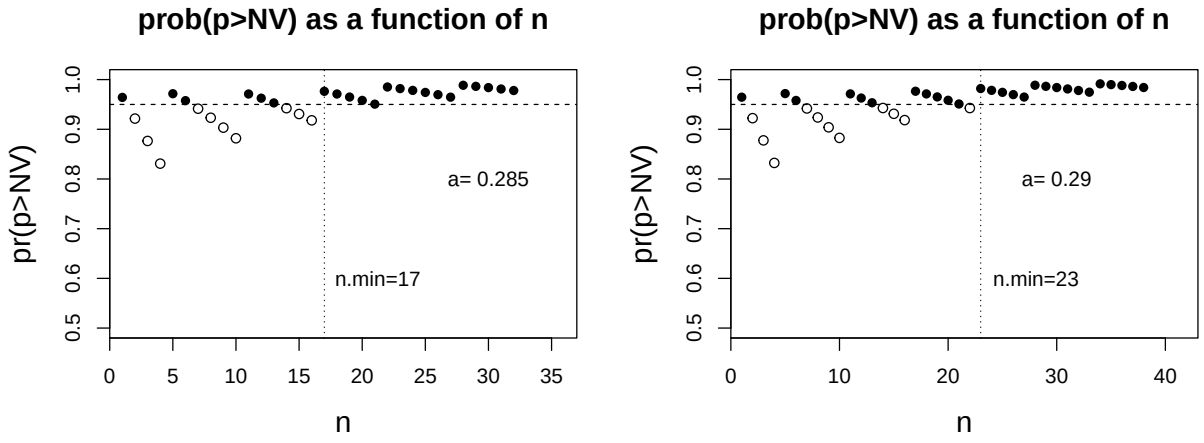
Appendix

Posterior median

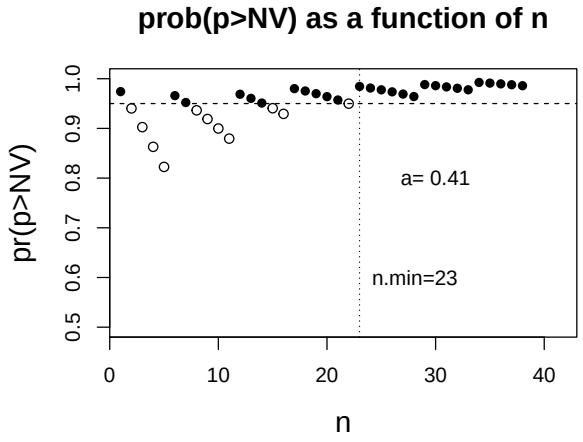
```
# First get the values of a in Beta prior where the jumps in minimal sample size occurs.
(a_vector <- alpha_vector[c(54,55,57,58,82,83,87,88, 100, 101, 110, 111, 172,173)])

par(mfrow=c(1,2))
n1 <- SS_DC_BayesBin(null.value = 0.075, p.positive = 0.95, dec.value = 0.175,
                a = a_vector[1], b=1, post.est = "median")
n2 <- SS_DC_BayesBin(null.value = 0.075, p.positive = 0.95, dec.value = 0.175,
                a = a_vector[2], b=1, post.est = "median")
```

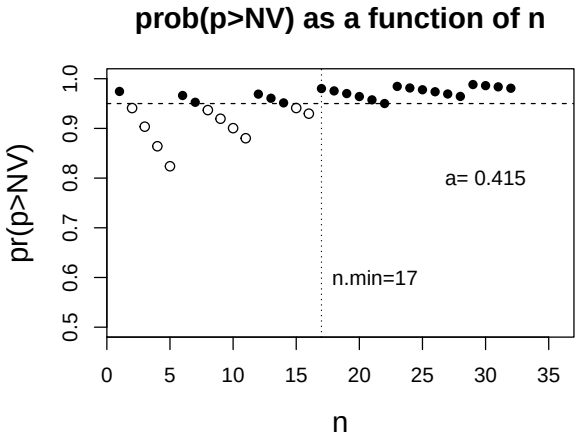



```
n3 <- SS_DC_BayesBin(null.value = 0.075, p.positive = 0.95, dec.value = 0.175,
                a = a_vector[3], b=1, post.est = "median")
n4 <- SS_DC_BayesBin(null.value = 0.075, p.positive = 0.95, dec.value = 0.175,
                a = a_vector[4], b=1, post.est = "median")
```

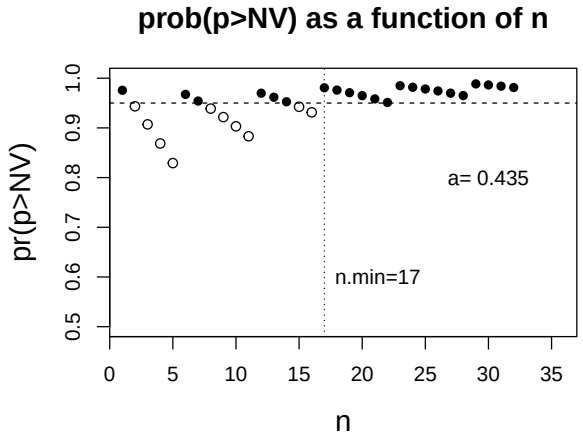


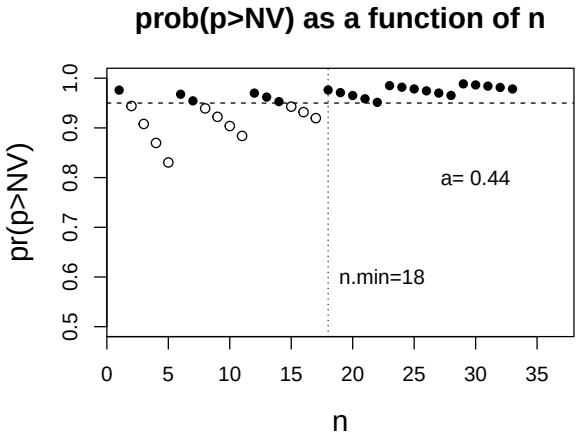
```
n5 <- SS_DC_BayesBin(null.value = 0.075, p.positive = 0.95, dec.value = 0.175,
                a = a_vector[5], b=1, post.est = "median")
n6 <- SS_DC_BayesBin(null.value = 0.075, p.positive = 0.95, dec.value = 0.175,
                a = a_vector[6], b=1, post.est = "median")
```



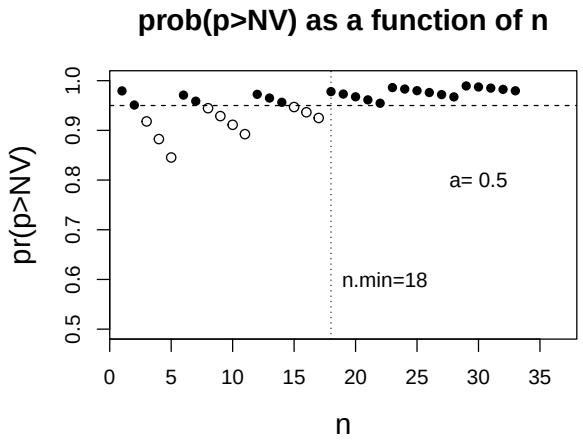


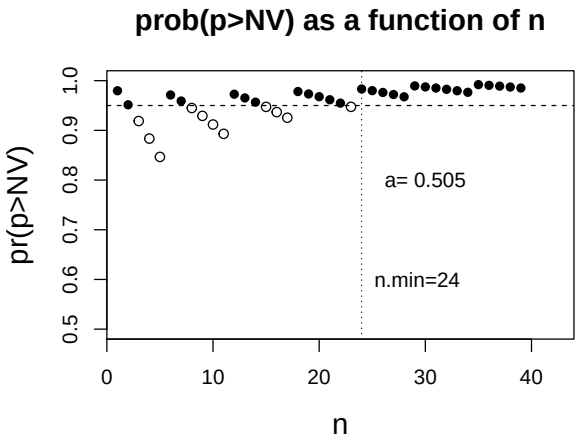

```
n7 <- SS_DC_BayesBin(null.value = 0.075, p.positive = 0.95, dec.value = 0.175,
                a = a_vector[7], b=1, post.est = "median")
n8 <- SS_DC_BayesBin(null.value = 0.075, p.positive = 0.95, dec.value = 0.175,
                a = a_vector[8], b=1, post.est = "median")
```



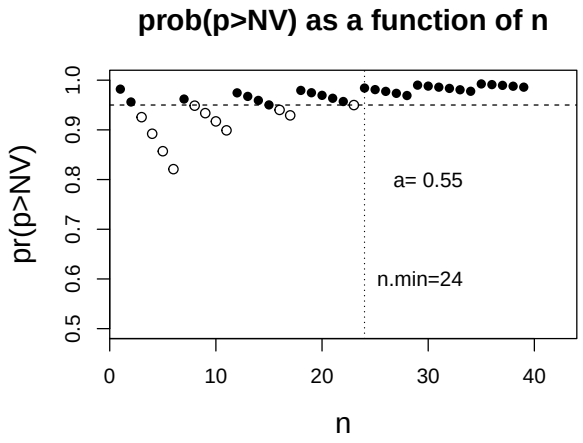


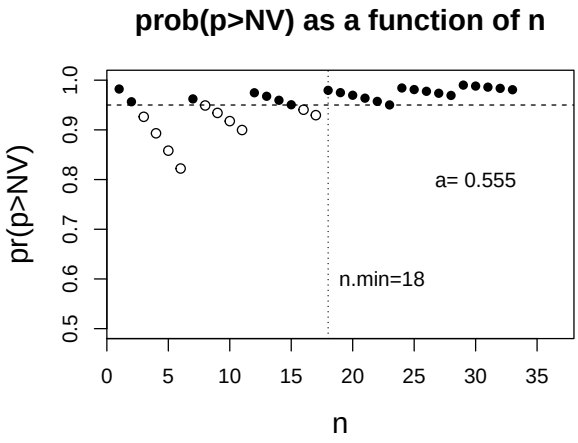

```
n9 <- SS_DC_BayesBin(null.value = 0.075, p.positive = 0.95, dec.value = 0.175,
                a = a_vector[9], b=1, post.est = "median")
n10 <- SS_DC_BayesBin(null.value = 0.075, p.positive = 0.95, dec.value = 0.175,
                a = a_vector[10], b=1, post.est = "median")
```



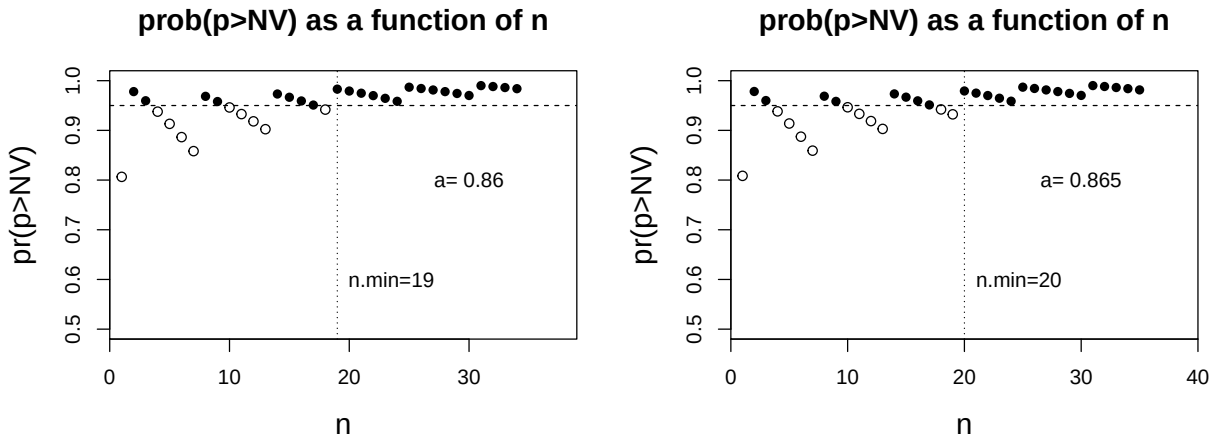



```
n11 <- SS_DC_BayesBin(null.value = 0.075, p.positive = 0.95, dec.value = 0.175,
                a = a_vector[11], b=1, post.est = "median")
n12 <- SS_DC_BayesBin(null.value = 0.075, p.positive = 0.95, dec.value = 0.175,
                a = a_vector[12], b=1, post.est = "median")
```



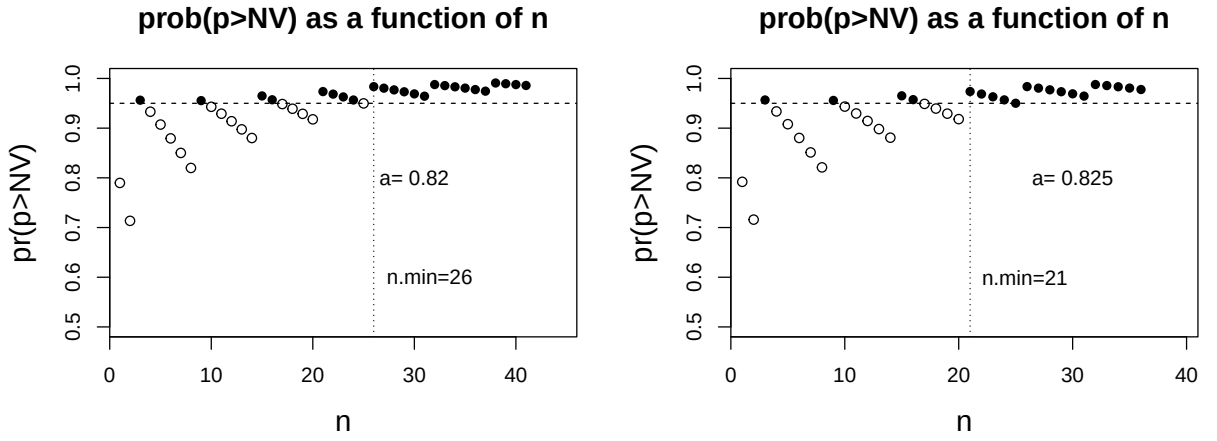



```
n13 <- SS_DC_BayesBin(null.value = 0.075, p.positive = 0.95, dec.value = 0.175,
                a = a_vector[13], b=1, post.est = "median")
n14 <- SS_DC_BayesBin(null.value = 0.075, p.positive = 0.95, dec.value = 0.175,
                a = a_vector[14], b=1, post.est = "median")
```

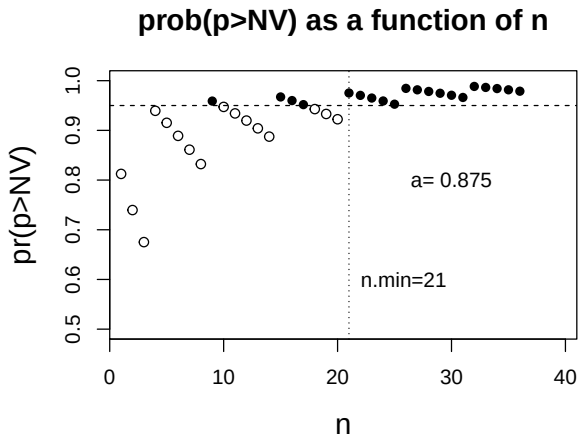


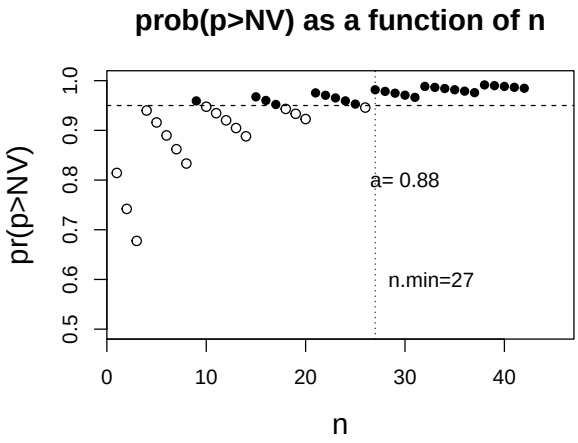
Posterior mean

```
par(mfrow=c(1,2))
a_vector_mean <- alpha_vector[c(164,165,175,176,191,192)]
n1 <- SS_DC_BayesBin(null.value = 0.075, p.positive = 0.95, dec.value = 0.175,
                a = a_vector_mean[1], b=1, post.est = "mean")
n2 <- SS_DC_BayesBin(null.value = 0.075, p.positive = 0.95, dec.value = 0.175,
                a = a_vector_mean[2], b=1, post.est = "mean")
```

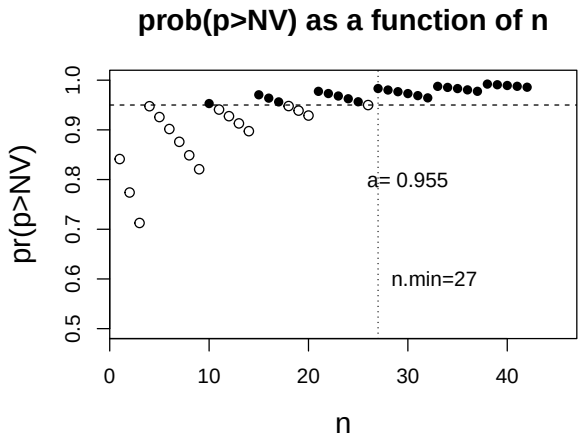


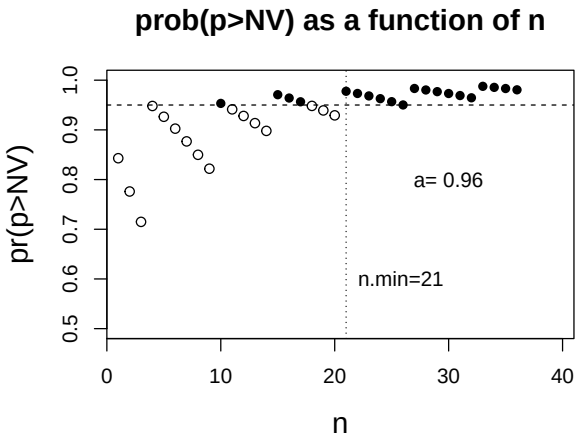

```
n3 <- SS_DC_BayesBin(null.value = 0.075, p.positive = 0.95, dec.value = 0.175,
                a = a_vector_mean[3], b=1, post.est = "mean")
n4 <- SS_DC_BayesBin(null.value = 0.075, p.positive = 0.95, dec.value = 0.175,
                a = a_vector_mean[4], b=1, post.est = "mean")
```






```
n5 <- SS_DC_BayesBin(null.value = 0.075, p.positive = 0.95, dec.value = 0.175,
                a = a_vector_mean[5], b=1, post.est = "mean")
n6 <- SS_DC_BayesBin(null.value = 0.075, p.positive = 0.95, dec.value = 0.175,
                a = a_vector_mean[6], b=1, post.est = "mean")
```





References

Roychoudhury, S., Scheuer, N., and Neuenschwander, B. (2018). Beyond p-values: A phase II dual-criterion design with statistical significance and clinical relevance. *Clinical trials (London, England)*, 15(5):452–461.

Sargent, D. J., Chan, V., and Goldberg, R. M. (2001). A three-outcome design for phase ii clinical trials. *Controlled Clinical Trials*, 22(2):117–125.

Computational details

```
## 2025-11-21 15:06:21.263698 Europe/Zurich
## R version 4.4.0 (2024-04-24)
## Platform: aarch64-apple-darwin20
## Running under: macOS 15.6.1
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRlapack.dylib;
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## time zone: Europe/Zurich
## tzcode source: internal
##
## attached base packages:
## [1] grid      stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
## [1] latex2exp_0.9.6  kableExtra_1.4.0 gridExtra_2.3    cowplot_1.1.3
## [5] ggplot2_3.5.2    knitr_1.50
##
## loaded via a namespace (and not attached):
##  [1] gtable_0.3.6      highr_0.11        dplyr_1.1.4       compiler_4.4.0
##  [5] tidyselect_1.2.1  xml2_1.3.8        stringr_1.5.1     systemfonts_1.2.2
##  [9] scales_1.3.0      fastmap_1.2.0     R6_2.6.1          labeling_0.4.3
## [13] generics_0.1.3    tibble_3.2.1      munsell_0.5.1     svglite_2.1.3
## [17] pillar_1.10.2     rlang_1.1.6       stringi_1.8.7     xfun_0.52
## [21] viridisLite_0.4.2 cli_3.6.5         withr_3.0.2       magrittr_2.0.3
## [25] digest_0.6.37     rstudioapi_0.17.1 lifecycle_1.0.4   vctrs_0.6.5
## [29] evaluate_1.0.3    glue_1.8.0        farver_2.1.2      colorspace_2.1-1
## [33] rmarkdown_2.29    tools_4.4.0       pkgconfig_2.0.3   htmltools_0.5.8.1
```