Code exploration

Minghan Yang

November 4, 2024

Abstract

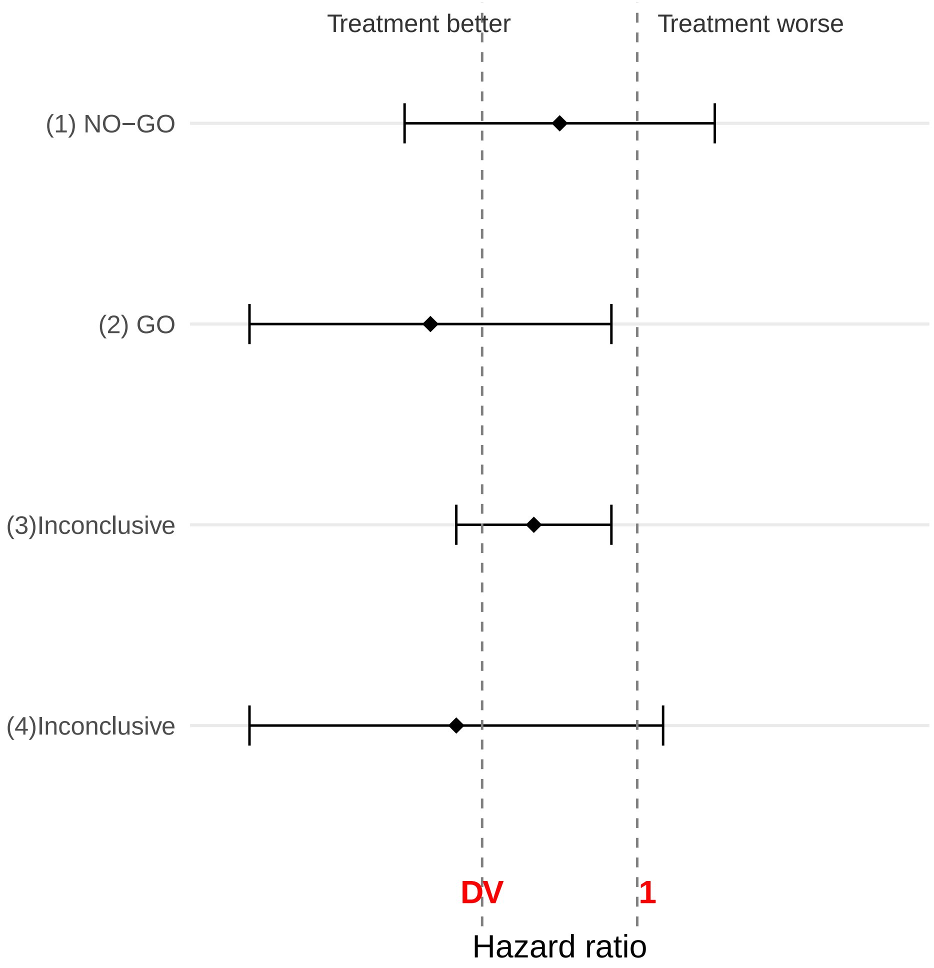
Related to paper "Beyond p-values: A phase II dual-criterion design with statistical significance and clinical relevance" Roychoudhury et al. (2018).

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 continuing the development of this 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).
   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 GO).
   By considering both, 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 compared to 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.

   Precision=$\frac{\text{TP}}{\text{TP+FP}}$. High Precision means that when the model predicts a positive outcome, it is very likely to be correct.

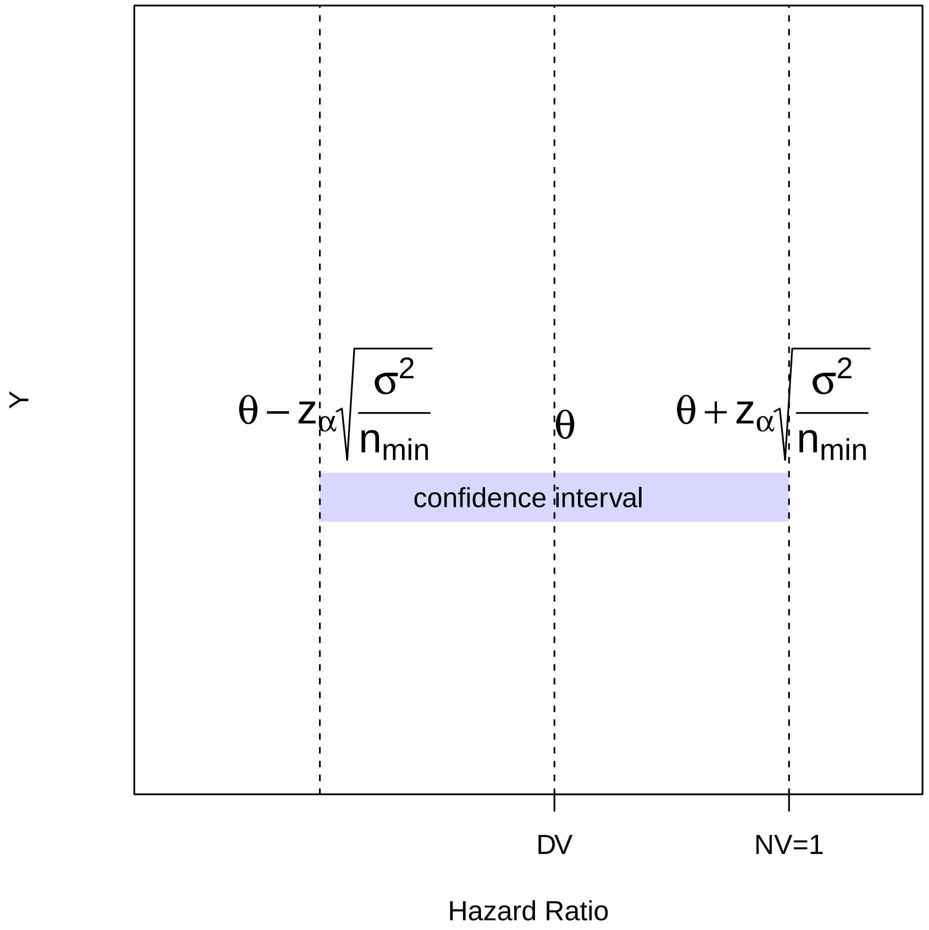
Consider hazard function in survival analysis, it describes the risk of failing. We consider hazard ratio between experimental drug and control as the outcome of interest. Hazard ratio(HR) less than 1 means the drug is better than control. We want to reduce hazard and hazard ratio.

Sample size:

Given the significance level α, the null value (NV), and the decision value (DV), we can calculate the minimum sample size (for normally distributed data):

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

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. The z_α is the $100(1-\alpha)\%$ quantile of the standard normal distribution. It is 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, there can only be GO or NO-GO decisions.


Y
$\theta - z_{\alpha}\sqrt{\frac{\sigma^2}{n_{min}}}$
θ
$\theta + z_{\alpha}\sqrt{\frac{\sigma^2}{n_{min}}}$
confidence interval
DV
NV=1
Hazard Ratio

Operating characteristics:

The operating characteristics are the type I error and 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.

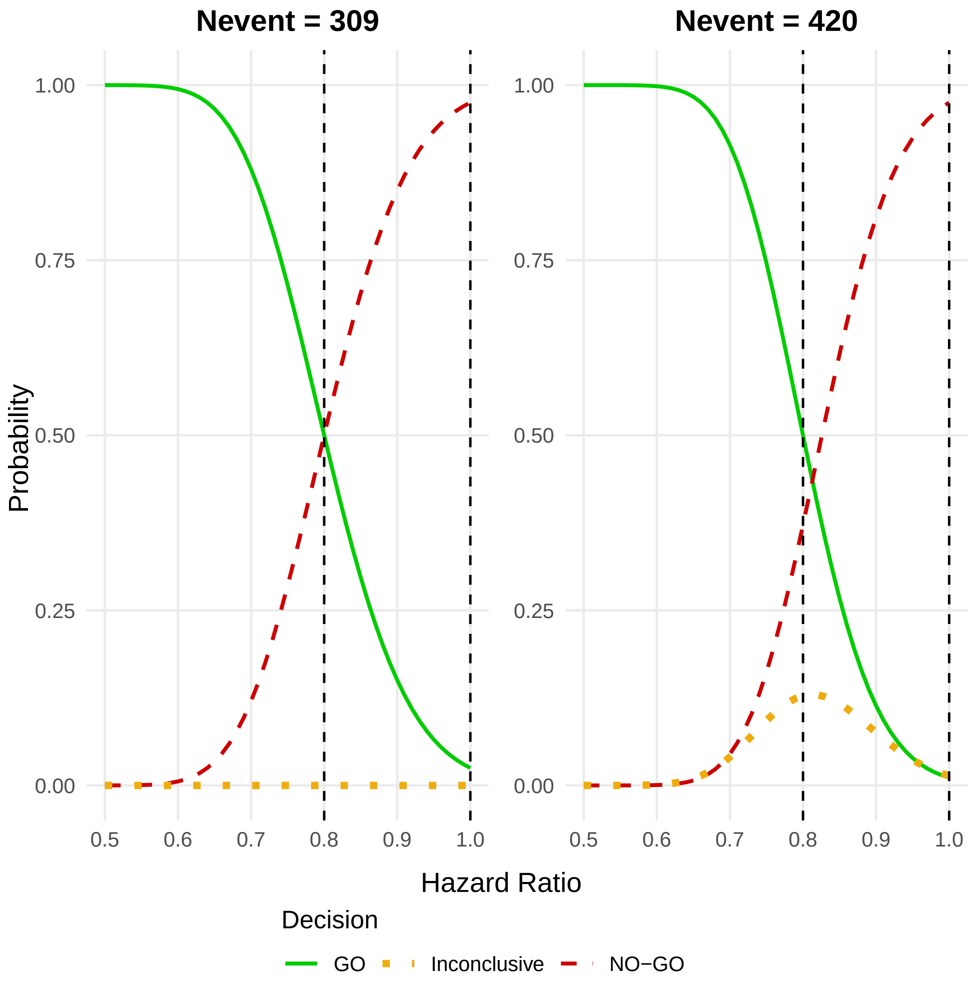
1.1 Reproduce Figure 1

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 calcualted under the example conditions $\sigma = 2$, $\alpha = 2.5\%$, log hazard ratios NV = log(1), DV = log(0.8).

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

The probability of making a "GO" decision is the probability of the estimate smaller than the decision value and

```
> # Load necessary libraries
> library(ggplot2)
> library(cowplot)
> library(gridExtra)
> library(grid)
> # 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)
> cut.ssig1 <- log(0.8)
> cut.crel1 <- log(0.8)
> pp.go1 <- pnorm(cut.crel1, t.d, sd1)
> pp.ngo1 <- 1 - pnorm(cut.ssig1, t.d, sd1)
> pp.intd1 <- 1 - pp.go1 - pp.ngo1
> df1 <- data.frame(HazardRatio = exp(t.d), GO = pp.go1, NOGO = pp.ngo1, Inconclusive = pp.int
> # Right panel (n = 420)
> n2 <- 420
> sd2 <- sqrt((2^2) / n2)
> cut.ssig2 <- log(1) - qnorm(0.975) * sqrt(2^2 / 420)
> cut.crel2 <- log(0.8)
> pp.go2 <- pnorm(cut.crel2, t.d, sd2)
> pp.ngo2 <- 1 - pnorm(cut.ssig2, t.d, sd2)
> pp.intd2 <- 1 - pp.go2 - pp.ngo2
> df2 <- data.frame(HazardRatio = exp(t.d), GO = pp.go2, NOGO = pp.ngo2, Inconclusive = pp.int
> # 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

Attempt to reproduce the values in Table 3. The main problem is to find the correct value for σ. There seem to be no detailed information about the choice of value for σ. We chose it to be 2 according to page 455 in the paper.
Also, the calculation of $\hat{\theta} = 0.736, 0.708, 0.659, 0.761$ should be looked into when the correct σ is decided. The current understanding is that these values are calculated when the estimate $\hat{\theta} >$ DV and the confidence interval just touches 1, i.e. NO-GO is implied.

```
> library(kableExtra)
> n.min <- (4*qnorm(0.9)^2)/(log(1)-log(0.7))^2
> # a sequence of true HR.
> t.d <- log(seq(0.5, 1, 0.1))
> # Dual-criterion design: alpha=0.1, DV=0.7, n=70
> n1 <- 70
> sd1 <- sqrt((2^2)/n1) #sigma assumed to be 2?
> #sqrt((1.723522^2)/n1)
> # cut.ssig <- 0-qnorm(0.9)* sqrt(2^2 / 70)
> cut.ssig <- log(exp(0-qnorm(0.9)* sqrt(2^2 / n1))) # statistical significance NV=1
> 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)
> # 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)
> # Dual-criterion design: alpha=0.1, beta=0.1, n=55
> n3 <- 55
> sd3 <- sqrt((2^2)/n3)
> cut.ssig <- log(0.708)
> cut.crel <- log(0.708)
> cut.ssig <- qnorm(0.9,log(0.5),sd3) # different from 0.708
> cut.crel <- qnorm(0.9,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)
> # Dual-criterion design: alpha=0.1, beta=0.2, n=38
> n4 <- 38
> sd4 <- sqrt((2^2)/n4)
> cut.ssig <- log(exp(0-qnorm(0.9)* sqrt(2^2 / n4))) # different from 0.659
> cut.crel <- log(exp(0-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)
> # Dual-criterion design: alpha=0.2, beta=0.1, n=38
> n5 <- 38
> sd5 <- sqrt((2^2)/n5)
> cut.ssig <- log(exp(0-qnorm(0.8)* sqrt(2^2 / n5)))
> # How is the power guaranteed to be 0.1? This is only using alpha=0.2.
> cut.crel <- log(exp(0-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)
> # Combine subtables by rows
> combined_table <- rbind(subtable1, subtable2, subtable3, subtable4, subtable5)
> # Convert to data frame for better kable support
> combined_table <- as.data.frame(combined_table)
> kable(combined_table, format = "latex", align = "c", booktabs = TRUE,
+       col.names = c("True HR", "GO", "NO-GO", "Inconclusive")) %>%
+   kable_styling(full_width = FALSE, position = "center",
+                 latex_options = c("hold_position","scale_down")) %>%
+   add_header_above(c("Reproduced Table 3" = 4)) %>%
+   group_rows("Subtable 1", 1, 6) %>%
+   group_rows("Subtable 2", 7, 12) %>%
+   group_rows("Subtable 3", 13, 18) %>%
+   group_rows("Subtable 4", 19, 24) %>%
+   group_rows("Subtable 5", 25, 30)
```

Reproduced Table 3

True HR	GO	NO-GO	Inconclusive
Subtable 1			
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
Subtable 2			
0.5	0.887	0.113	0.000
0.6	0.711	0.289	0.000
0.7	0.500	0.500	0.000
0.8	0.315	0.685	0.000
0.9	0.182	0.818	0.000
1.0	0.099	0.901	0.000
Subtable 3			
0.5	0.900	0.100	0.000
0.6	0.728	0.272	0.000
0.7	0.514	0.486	0.000
0.8	0.322	0.678	0.000
0.9	0.185	0.815	0.000
1.0	0.099	0.901	0.000
Subtable 4			
0.5	0.804	0.196	0.000
0.6	0.615	0.385	0.000
0.7	0.428	0.572	0.000
0.8	0.276	0.724	0.000
0.9	0.169	0.831	0.000
1.0	0.100	0.900	0.000
Subtable 5			
0.5	0.902	0.098	0.000
0.6	0.768	0.232	0.000
0.7	0.602	0.398	0.000
0.8	0.439	0.561	0.000
0.9	0.303	0.697	0.000
1.0	0.200	0.800	0.000

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.
Prior: minimally informative unimodal beta prior distribution $Beta(0.0811, 1)$, which has mean 0.75.
NV is set to 7.5% rather than 0, because of the absence of a comparator (in single arm trials).
DV is set to be 10%+7.5%=17.5%.
So the dual-criterion is:

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

The minimal sample size was 22. Final sample size 25.
Null hypothesis: there is no effect of the drug, i.e. ORR=7.5%
$\mathbb{P}(\text{type I error}) = \mathbb{P}(\text{reject } H_0 | H_0 \text{ is true}) = \mathbb{P}(\text{reject } H_0 | ORR \leq 7.5\%)$
$\mathbb{P}(\text{type II error}) = \mathbb{P}(\text{not reject } H_0 | H_0 \text{ is false}) = \mathbb{P}(\text{reject } H_0 | ORR = \text{response rate})$
Table 4 results show that this dual-criterion design is a three-outcome design with desirable properties.

3.1 Reproduce Figure 2

3.2 Reproduce Table 4

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.

Computational details

```
> cat(paste(Sys.time(), Sys.timezone(), "\n"))
```

```
2024-11-04 00:05:03.439443 Europe/Zurich
```

```
> sessionInfo()
```

```
R version 4.4.0 (2024-04-24)
Platform: aarch64-apple-darwin20
Running under: macOS Sonoma 14.4

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;  LA

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] kableExtra_1.4.0 gridExtra_2.3    cowplot_1.1.3    latex2exp_0.9.6
[5] ggplot2_3.5.1    knitr_1.48

loaded via a namespace (and not attached):
 [1] gtable_0.3.5      dplyr_1.1.4       compiler_4.4.0    tidyselect_1.2.1
 [5] xml2_1.3.6        stringr_1.5.1     systemfonts_1.1.0 scales_1.3.0
 [9] fastmap_1.2.0     R6_2.5.1          labeling_0.4.3    generics_0.1.3
[13] tibble_3.2.1      munsell_0.5.1     svglite_2.1.3     pillar_1.9.0
[17] rlang_1.1.4       utf8_1.2.4        stringi_1.8.4     xfun_0.44
[21] viridisLite_0.4.2 cli_3.6.2         withr_3.0.0       magrittr_2.0.3
[25] digest_0.6.35     rstudioapi_0.16.0 lifecycle_1.0.4   vctrs_0.6.5
[29] evaluate_0.24.0   glue_1.7.0        farver_2.1.2      fansi_1.0.6
[33] colorspace_2.1-0  rmarkdown_2.27    tools_4.4.0       pkgconfig_2.0.3
[37] htmltools_0.5.8.1
```