Power Calculations Mendel-UPenn

December 16, 2024

Study design

The primary purpose of the Mendel.ai study is to establish whether the predictions on oncology clinical trial eligibility made by a human+AI collaboration are non-inferior to the predictions made by humans alone. The predictions for trial eligibility are made by assessing eligibility across multiple clinical categories (ex. tumor stage, cancer biomarkers).

Primary Outcome

The primary outcome for this study is the proportion of categories correctly classified with respect to its eligibility status. This outcome is measured for each patient.

Hypothesis

Non-inferiority is established by comparing the differences in the mean of the primary outcome between the two comparison arms with a pre-defined non-inferiority margin, Δ.

To this end, the Mendel trial tests the following one-sided hypotheses:

$H_0 : \mu_a - \mu_h \leq -\Delta$ (Null hypothesis)

$H_a : \mu_a - \mu_h > -\Delta$ (Alternative hypothesis)

Parameter	Description
μ_a	Mean proportion of categories correctly classified by human+ai
μ_h	Mean proportion of categories correctly classified by human+ai
$\Delta > 0$	Non-inferiority margin, defined by researcher (ex. .05, .10, .15)

Method

The purpose of these calculations is to establish a minimum sample size required for a well-powered non-inferiority study.

Mean proportion correct

Let Y_i represent the proportion of correctly classified criteria for the ith patient, and S_{ij} denote the classification accuracy result of the jth (j=1..m) criterion for the ith (i=1..n) patient. Then,

S_{ij} ~ Bernoulli(E(S_{ij})), where $E(S_{ij}) = P(S_{ij} = 1)$ = probability recorded observation is correct for category j

$$S_{ij} = \begin{cases} 1 & \text{if correct,} \\ 0 & \text{if not correct.} \end{cases}$$

Then $Y_i = \sum_{j=1}^{m} \frac{S_{ij}}{m}$

Let Z_{ij} denote the classification result of the jth criterion for the ith patient, where

Z_{ij} ~ Bernoulli(E(Z_{ij})), where $E(Z_{ij}) = P(Z_{ij} = 1)$ = probability recorded observation is marked eligible for category j

$$Z_{ij} = \begin{cases} 1 & \text{if marked eligible,} \\ 0 & \text{if marked not eligible.} \end{cases}$$

And let T_{ij} denote the true classification of the jth criterion for the ith patient, where

T_{ij} ~ Bernoulli(E(T_{ij})), where $E(T_{ij}) = P(T_{ij} = 1)$ = probability recorded observation is truly eligible for category j (event rate)

$$T_{ij} = \begin{cases} 1 & \text{if truly eligible,} \\ 0 & \text{if truly not eligible.} \end{cases}$$

Then:

$$\mu = E(Y_i) = E(\sum_{j=1}^{m} \frac{S_{ij}}{m}) = \frac{1}{m} \sum_{j=1}^{m} E(S_{ij}) = \frac{1}{m} \sum_{j=1}^{m} P(S_{ij} = 1)$$

where

$$P(S_{ij} = 1) = P(Z_{ij} = 1, T_{ij} = 1) + P(Z_{ij} = 0, T_{ij} = 0)$$

= P(Z_{ij} = 1 | T_{ij} = 1) P(T_{ij} = 1) + P(Z_{ij} = 0 | T_{ij} = 0) P(T_{ij} = 0) =
$$TPR_j * p_j + (1 - FPR_j) * (1 - p_j) = TPR_j * p_j + TNR_j * (1 - p_j)$$
$$= \frac{TP_j + TN_j}{TP_j + TN_j + FP_j + FN_j}$$

and

$$Var(Y_i) = Var(\frac{1}{m} \sum_j S_{ij}) = \frac{1}{m^2} \sum_j Var(S_{ij}) = \frac{1}{m^2} \sum_j [P(S_{ij} = 1)(1 - P(S_{ij} = 1))]$$

We calculate μ, a constant mean proportion of correctly classified criterion, for both the human+ai collaboration arm, μ_a = E(Y_i^a), and the human arm, μ_h = E(Y_i^h). We assume independence in the evaluation of different individuals and different categories.

Code for finding vector of probability of success P(S_ij = 1), expectation of Y_ij, and variance of Y_ij

```
p.s <- function(TPR, FPR, event_rate){TPR*event_rate + (1-FPR)*(1-event_rate)} #finding probability of success/correct for category j if TPR, FPR are given for category j
mu <- function(p.s_vector){(1/length(p.s_vector))*sum(p.s_vector)} #input is the vector of probabilities of correct for all categories j=1..m
var <- function(p.s_vector){(1/(length(p.s_vector))^2)*sum(p.s_vector*(1-p.s_vector))}#input is the vector of probabilities of correct for all categories j=1..m
```

Test statistic

When the Human and Human+AI arms consist of different group of patients, the calculation of sample size can be performed using standard formula for testing equivalence of two proportions. Assuming both arms recruit n patients, the test statistic is formed by:

$T_n = \frac{(\mu_a - \mu_h) - (-\Delta)}{\sqrt{Var(D)/n}} = \frac{(\mu_a - \mu_h) + \Delta}{\sqrt{Var(D)/n}}$, where D = $\mu_a - \mu_h$ and Var(D) are calculated below:

Paired design

In this study design, the human and human+AI collaboration arms have the same sample population leading to paired data.

D = $\frac{1}{n} \sum_{i=1}^{n} Y_i^a - Y_i^h$

Var(D) = $Var(\frac{1}{n} \sum_{i=1}^{n} Y_i^a - Y_i^h) = \frac{1}{n^2} \sum_i Var(Y_i^a - Y_i^h) = \frac{1}{n}[Var(Y_i^a) + Var(Y_i^h) - 2Cov(Y_i^a, Y_i^h)]$

Let $Var(D') = Var(Y_i^a) + Var(Y_i^h) - 2Cov(Y_i^a, Y_i^h)$, where

$Cov(Y_i^a, Y_i^h) = \rho\sqrt{Var(Y_i^a)Var(Y_i^h)} = E(Y_i^a Y_i^h) - E(Y_i^a)E(Y_i^h) = E(Y_i^a Y_i^h) - \mu_a \mu_h$

Alternative approach, concordance rate different between criterion:

$Cov(Y_i^a, Y_i^h) = Cov(\frac{1}{m} \sum_j S_{ij}^a, \frac{1}{m} \sum_j S_{ij}^h) = \frac{1}{m^2} \sum_j Cov(S_{ij}^a, S_{ij}^h)$, where we assume independence between individuals and between criterion and

$Cov(S_{ij}^a, S_{ij}^h) = E(S_{ij}^a S_{ij}^h) - E(S_{ij}^a)E(S_{ij}^h) = P(S_{ij}^a = 1, S_{ij}^h = 1) - P(S_{ij}^a = 1)P(S_{ij}^h = 1)$, where

$P(S_{ij}^a = 1, S_{ij}^h = 1)$ = $P(Z_{ij}^a = 1, Z_{ij}^h = 1, T_{ij} = 1) + P(Z_{ij}^a = 0, Z_{ij}^h = 0, T_{ij} = 0)$

= $P(Z_{ij}^a = 1 | Z_{ij}^h = 1, T_{ij} = 1)P(Z_{ij}^h = 1 | T_{ij} = 1)P(T_{ij} = 1) + P(Z_{ij}^a = 0 | Z_{ij}^h = 0, T_{ij} = 0)P(Z_{ij}^h = 0 | T_{ij} = 0)P(T_{ij} = 0)$

=$P(Z_{ij}^a = 1 | Z_{ij}^h = 1, T_{ij} = 1)TPR_j^h p_j + P(Z_{ij}^a = 0 | Z_{ij}^h = 0, T_{ij} = 0)TNR_j^h(1 - p_j)$

We may assume $P(Z_{ij}^a = 1 | Z_{ij}^h = 1, T_{ij} = 1)$ = $P(Z_{ij}^a = 0 | Z_{ij}^h = 0, T_{ij} = 0)$ = 1

Then,

$Cov(Y_i^a, Y_i^h) = \frac{1}{m^2} \sum_j Cov(S_{ij}^a, S_{ij}^h) = \frac{1}{m^2} \sum_j [P(S_{ij}^a = 1, S_{ij}^h = 1) - P(S_{ij}^a = 1)P(S_{ij}^h = 1)]$

Code to get Var(D')

```
#p11 represents P(Z_ij^a=1 | Z_ij^h=1, T_ij = 1) = probability AI + human collaboration is correct
for observation i if human was correct for that observation
#p00 represents P(Z_ij^a-0 | Z_ij^h=0, T_ij = 1) = probability AI + human collaboration is incorre
ct for observation i if human was incorrect for that observation
var_paired <- function(p11, p00, p.s_vector_1, p.s_vector_0, event_rate,fpr_0, tpr_0){ # if probab
ility of concordance is known and provided
  cov <- (1/length(p.s_vector_1))^2*sum((p11*tpr_0*event_rate + p00*(1-fpr_0)*(1-event_rate)) - (
p.s_vector_0*p.s_vector_1))
  return(var(p.s_vector_1) + var(p.s_vector_0) - 2*cov)
}
```

Calculating sample size for a paired design

n = $\frac{(Z_{1-\beta} + Z_{1-\alpha})^2 Var(D')}{((\mu_a - \mu_h) + \Delta)^2}$, where β is Type II error, $1 - \beta$ is the desired power, α is Type-I error, and $Var(D')$ is calculated for paired or unpaired designs as specified above

```
find_n <- function(tpr_h, tpr_a, fpr_h, fpr_a, event_rate, p11, p00, delta,power, alpha, paired =
TRUE){ #r = allocation ratio ; delta = margin of noninferiority (delta > 0)
  #calculate probability of success S_ij for human and human/ai arms
  p.s_h <- p.s(tpr_h, fpr_h, event_rate)
  p.s_a <- p.s(tpr_a, fpr_a, event_rate)

  #calculating Y_ij expectation and variance for human and human/ai arms
  mu_h <- mu(p.s_h)
  mu_a <- mu(p.s_a)

  var_h <- var(p.s_h)
  var_a <- var(p.s_a)

  #calculating variance of test statistic
  if(paired == TRUE ){
    var_d_prime <- var_paired(p11, p00, p.s_a, p.s_h, event_rate, fpr_h, tpr_h)
  }
  else{
    var_d_prime <- var_h + var_a #can also use the function var_unpaired above to calculate
  }

  #calculating sample size
  delta_adjusted <- delta*(mu_h+mu_a)/2
  numerator_n <- ((qnorm(power) + qnorm(1-alpha))^2)*var_d_prime
  denominator_n <- ((mu_a - mu_h) + delta_adjusted)^2
  n <- numerator_n/denominator_n

  return(n)
}
```

Calculating n with mendel data

Loading and processing data

```
library(ggplot2)
data <- read.csv("~/Desktop/Desktop - ceb-bios-363/PHD/Mendel_Sample_Size/mendel_2.csv")
names(data)[6] <- "tpr_a"
names(data)[9] <- "event_rate"
data$tnr_h <- 1 #assumption we are making
data$tnr_a <- 1 #assumption
data$fpr_h <- 1-data$tnr_h
data$fpr_a <- 1-data$tnr_a
print(data)
```

```
##    human_alone_f1 h_ai_f1 cohort_size event_size     tpr_h     tpr_a
## 1            0.90    0.95       40-50      30-40 0.8181818 0.9047619
## 2            0.70    0.85       40-50      30-40 0.5384615 0.7391304
## 3            0.85    0.95       40-50      30-40 0.7391304 0.9047619
## 4            0.80    0.85       40-50      20-40 0.6666667 0.7391304
## 5            0.85    0.90       40-50      20-30 0.7391304 0.8181818
## 6            0.85    0.90       40-50      20-30 0.7391304 0.8181818
## 7            0.70    0.90       40-50      20-30 0.5384615 0.8181818
## 8            0.90    0.85       40-50      10-20 0.8181818 0.7391304
## 9            0.70    0.75       40-50      10-20 0.5384615 0.6000000
## 10           0.70    0.80       40-50      20-40 0.5384615 0.6666667
##    event_rate_min event_rate_max event_rate tnr_h tnr_a fpr_h fpr_a
## 1             0.6           1.00      0.800     1     1     0     0
## 2             0.6           1.00      0.800     1     1     0     0
## 3             0.6           1.00      0.800     1     1     0     0
## 4             0.4           1.00      0.700     1     1     0     0
## 5             0.4           0.75      0.575     1     1     0     0
## 6             0.4           0.75      0.575     1     1     0     0
## 7             0.4           0.75      0.575     1     1     0     0
## 8             0.2           0.50      0.350     1     1     0     0
## 9             0.2           0.50      0.350     1     1     0     0
## 10            0.4           1.00      0.700     1     1     0     0
```

Paired design, changing event rate, concordance, and tnr

Power = .8

alpha = .05

delta = .05

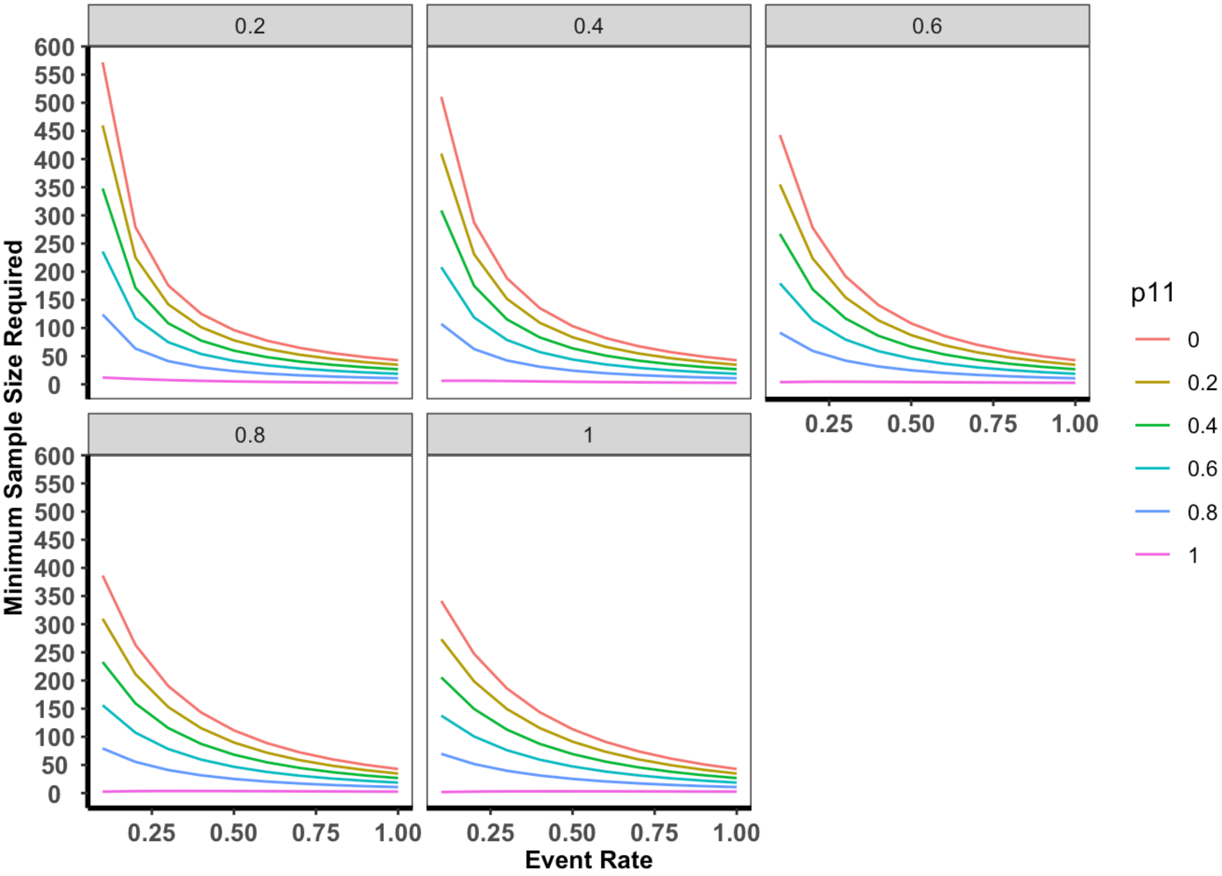
$P(Z_{ij}^a = 1 | Z_{ij}^h = 1, T_{ij} = 1)$ = $P(Z_{ij}^a = 0 | Z_{ij}^h = 0, T_{ij} = 0)$ = 0.0, 0.2, 0.4, 0.6, 0.8, 1.0

TNR = .4, .6, .8, 1.0

event_rate = 0.10, 0.20,.3, 0.40, .5, 0.60, .7, 0.80, .9, 1.0

```
## Warning: The `size` argument of `element_line()` is deprecated as of ggplot2 3.4.0.
## ℹ Please use the `linewidth` argument instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## quartz_off_screen
##                 2
```


0.2
0.4
0.6
0.8
1
Minimum Sample Size Required
600
550
500
450
400
350
300
250
200
150
100
50
0
Event Rate
0.25
0.50
0.75
1.00
p11
0
0.2
0.4
0.6
0.8
1