Ranking method simulation study

Robin Blythe

## Background

This project is an exploratory study to determine whether ranking patients by predicted risk can yield lower costs and better outcomes than using thresholds from a hospital perspective. Thresholds, while convenient to implement, may be less cost-effective than ranking, given that a ranking system should theoretically be capable of identifying true positives first from a discriminating model. Prediction models are capable of generating rich outcome data, but dichotomising into high and low risk does not fully utilise predictions. This is especially true when there may be a large number of individuals classified as high risk, but due to time or resource constraints, not all high-risk patients can be treated.

This study compares a ranking-based approach to a threshold based approach both with and without resource constraints to determine the impact of risk dichotomisation.

## Study outline

### Setup

1. Determine a use case.
   1. In this study, we will examine the implementation of a program to identify patients at high risk of deterioration leading to potentially avoidable additional costs and poor outcomes. We can use the implementation of Escobar et al (2020) to obtain prevalence for ICU admission, a common endpoint for deterioration prediction. (Escobar et al. 2020)
   2. Clinical deterioration is characterised by several challenges including which outcome to use and, depending on the selected outcome, a low prevalence of the event. (Blythe et al. 2023) However, this has not deterred hospitals from applying clinical deterioration models, despite the lack of success in improving patient outcomes. (Blythe et al. 2022)
2. Simulate an underlying patient population using the minimum required sample size.
3. Simulate a hypothetical clinical prediction model with a given AUC and underlying prevalence.
4. Obtain costs and outcomes for each state: true positive, true negative, false positive, false negative.
5. Identify a selection of thresholds based on the study population used to generate the model.
   1. Value-optimising threshold
   2. Youden index (or other ROC-based method)
   3. PPV/NNE-based approach
6. Randomly generate new sample populations based on a range of plausible AUC and prevalence values to mimic new patients being treated with the original model.
7. Apply the thresholds from step 5, filtering the data to include only high risk patients. Sort patients in order of predicted risk (descending).
8. Compare the net monetary benefit of treating a random sample of N patients to treating a ranked list of the top N patients.

## Methods

Let’s assume that, left unassessed, a deteriorating medical inpatient will need to be rapidly transferred to the ICU, which incurs significant additional costs. A deterioration detection system can escalate that decision to an outreach team, in this case comprising of an ICU registrar, the patient’s attending, and the patient’s nurse.

False positives can lead to wasted clinician time, and false negatives can lead to costly ICU admissions. Considering that many patients will require ICU admission regardless of whether they can be successfully intervened upon, we should estimate a conservative treatment effect for assessment, rather than assuming an early intervention will necessarily be preventative.

### Set up experiment

options(scipen = 999, digits = 3)  
library(predictNMB)  
library(pROC)  
library(tidyverse)  
  
# Hypothetical model specs and event rate  
auc = 0.85  
params = 30  
p0 = 0.05  
wtp = ceiling(28033\*(1.03)^(2024 - 2018))  
  
# Obtain minimum sample size/events  
pmsamp <- pmsampsize::pmsampsize(  
 type = "b",  
 prevalence = p0,   
 cstatistic = auc,  
 parameters = params)  
  
sample\_size <- pmsamp$sample\_size  
min\_events <- ceiling(pmsamp$events)  
remove(pmsamp)

For a model with 30 parameters, an AUC of 0.85 and a prevalence of 0.05, we need a sample size of 2,848 with 143 events. We can use these requirements to generate a hypothetical study population.

### Generate simulated dataset

set.seed(888)  
# Use model specs and minimum sample size to generate a sample population  
sample\_pop <- get\_sample(  
 auc = auc,   
 n\_samples = sample\_size,   
 prevalence = p0,   
 min\_events = min\_events)  
  
sample\_pop$actual <- factor(sample\_pop$actual)  
  
# Obtain predicted probabilities  
fit <- glm(actual ~ x,   
 data = sample\_pop,   
 family = binomial()  
 )  
sample\_pop$predicted <- predict(fit, type = "response")

### Assign costs and outcomes to 2 x 2 table

#### Parameters

* Additional costs due to deterioration
  + The additional utilisation cost of deterioration is ~𝒩(14134, 686) (Curtis et al. 2021)
  + This cost is in addition to a clinical ICU assessment (see below)
* Cost of clinical time per minute based on ICU outreach registrar, the resident medical officer, and the patient’s nurse (Bohingamu Mudiyanselage et al. 2024)
  + ICU outreach registrar hourly: 69.91/60 \* (1.03)^(2024 - 2016) = 1.48
  + Resident medical officer hourly: 50.79/60 \* (1.03)^(2024 - 2016) = 1.07
  + Nurse hourly: 30.47 \* (1.03)^(2024 - 2016) = 0.64
  + Altogether: $3.19/minute (fixed)
* Duration of clinical assessment (using MET call time) ~𝚪(110.314, 0.172) (Bellomo et al. 2003)
  + Note that this was converted to a Gamma distribution based on a mean duration of 19 minutes and a standard error of 18/√99 = 1.809 minutes using ShinyPrior (White and Blythe 2023)
* Effect of alert on ICU admission rate (hazard ratio) ~𝒩(0.910, 0.036) (Escobar et al. 2020)
  + Note: must be taken as (1 - HR) for sampler
* QALYs lost from deterioration episode ~𝒩(0.03, 0.04) (Holmes et al. 2024)
* Opportunity cost of a positive alert = probability patient can be successfully treated \* underlying prevalence of event \* cost of outcome avoided if successfully treated (i.e., if the clinicians were doing something more productive with their time)

##### Fixed

fx\_nmb <- get\_nmb\_sampler(  
 # Cost of ICU admission  
 outcome\_cost = 14134,  
 # Willingness to pay per QALY  
 wtp = wtp,  
 # QALYs lost due to deterioration event  
 qalys\_lost = 0.03,  
 # Cost of an evaluation = (Clinician time cost \* duration of MET) + (Opportunity cost = chance of successful intervention \* outcome cost \* underlying p0)  
 high\_risk\_group\_treatment\_cost = (3.19 \* 19) + ((1 - 0.910) \* 14134 \* p0),  
 # Chance of successful intervention  
 high\_risk\_group\_treatment\_effect = 1 - 0.910  
)

##### Stochastic

# Need to make a wrapper as the same estimates should be used for multiple cells  
sampler <- function() {  
 cost\_outcome <- rnorm(1, 14134, 686)  
 eff\_outcome <- 1 - rnorm(1, 0.910, 0.036)  
   
 fx\_nmb\_sampler <- get\_nmb\_sampler(  
 outcome\_cost = cost\_outcome,  
 wtp = wtp,  
 qalys\_lost = function() rnorm(1, 0.03, 0.04),  
 high\_risk\_group\_treatment\_cost =  
 function() rgamma(1, shape = 110.314, scale = 0.172) \* 3.19 +  
 (p0 \* eff\_outcome \* cost\_outcome),  
 high\_risk\_group\_treatment\_effect = eff\_outcome  
 )  
   
 fx\_nmb\_sampler()  
}

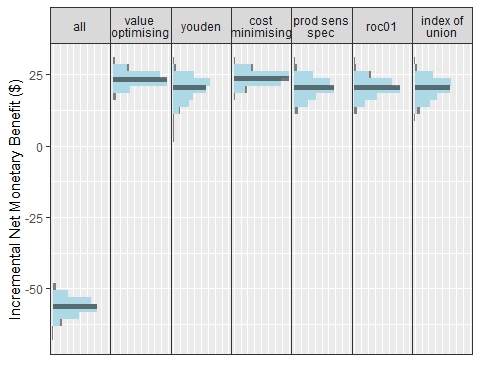
## Derive cutpoints

### ROC-curve and NMB-based cutpoints

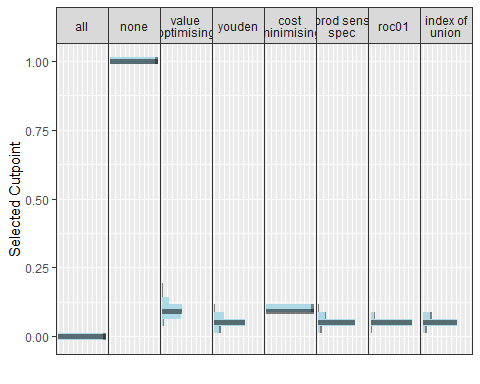
nmb\_simulation <- do\_nmb\_sim(  
 sample\_size = sample\_size,  
 n\_sims = 500,  
 n\_valid = 10000,  
 sim\_auc = auc,  
 event\_rate = p0,  
 fx\_nmb\_training = fx\_nmb,  
 fx\_nmb\_evaluation = fx\_nmb  
)  
summary(nmb\_simulation)

# A tibble: 8 × 3  
 method median `95% CI`   
 <chr> <dbl> <chr>   
1 all -812. -863.3 to -755.1  
2 cost minimising -732. -786.5 to -672.9  
3 index of union -736. -789.7 to -677.8  
4 none -755. -812.2 to -693.3  
5 prod sens spec -736. -789.6 to -677.8  
6 roc01 -736. -789.6 to -677.4  
7 value optimising -733. -786.9 to -673.3  
8 youden -736. -790.6 to -679.4

autoplot(nmb\_simulation, what = "inb", inb\_ref\_col = "none") + theme\_sim()



autoplot(nmb\_simulation, what = "cutpoints") + theme\_sim()



cutpoint\_youden <- median(nmb\_simulation$df\_thresholds$youden)  
cutpoint\_nmb <- median(nmb\_simulation$df\_thresholds$value\_optimising)  
remove(nmb\_simulation)

### PPV-based threshold

A recent consideration with regards to implementing thresholds for risk stratification has been to limit the number of alerts based on what clinicians will tolerate, using the positive predictive value (PPV) to determine number needed to evaluate (NNE).(Romero-Brufau et al. 2015) For example, a model with a PPV of 0.1 corresponds to a rate of 1 positive per 10 alerts, or a NNE of 10. This approach can be considered in addition to other, ROC or value-based methods. (Parsons et al. 2023)

roc\_curve <- roc(response = sample\_pop$actual, predictor = sample\_pop$predicted)  
ppv <- coords(  
 roc\_curve,   
 x = "all",   
 input = "threshold",   
 ret = c("threshold", "ppv")  
 )  
  
# Maximum tolerable number of false positives per true positive (example)  
nne = 4  
  
# NNE = 1/PPV  
ppv$nne <- 1/ppv$ppv  
cutpoint\_nne <- min(ppv$threshold[ppv$nne == nne], na.rm = T)  
remove(roc\_curve, ppv, sample\_pop)

## Obtain predictions for variety of external validation scenarios

By repeating the sample population simulation step for a selection of prevalence and AUC values, we can mimic an external validation study by fitting the original model to these external settings.

combs <- expand.grid(  
 auc = c(0.65, 0.75, 0.85, 0.95),  
 p0 = c(0.01, 0.05, 0.10)  
 )  
  
sims <- map2(.x = combs$auc, .y = combs$p0,   
 \(x, y) cbind(  
 get\_sample(auc = x,   
 n\_samples = 1000,   
 prevalence = y),  
 auc = x,  
 p0 = y)  
)  
remove(combs)

### Apply cutpoints to classify predictions

We can now simulate an external validation study under a variety of conditions based on a range of AUC and prevalence values. This mimics a scenario in which we take the original model, developed under the conditions of AUC = 0.85 and p0 of 0.05, and apply the predictions to other datasets, also created using the get\_sample() function.

source("./utils.R")  
  
# The 'sims' object can then be turned into a set of predictions and classifications (see utils.R)  
  
df\_sims <- obtain\_class(sims)  
remove(sims)

### Obtain NMB of each strategy

We now have a set of randomly drawn datasets, predictions, and classifications (treat or do not treat). However, as the thresholds were drawn using all the data, they may not be suitable when extrapolating to other scenarios, for example a future patient population. In practice, we may often see that a threshold can identify many patients at risk, but we can only see a small number of these patients based on resource constraints (e.g., clinical time). We can use the risk predictions to prioritise patients (ranking), or we can just randomly select positive cases.

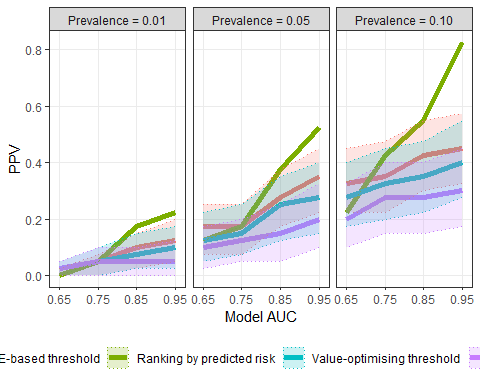
We can simulate this by taking random draws of e.g. 40 patients with a positive prediction for each threshold method. We can then compare this to a ranking approach where we sort by predicted risk and take the top 40.

Method:

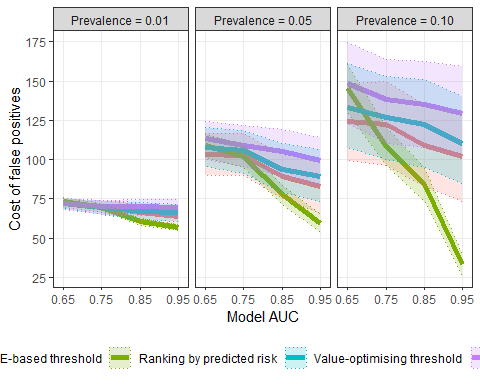
1. For each AUC and prevalence, take 40 random draws for positive cases from each threshold method (NMB, NNE, Youden)
2. Take the top 40 patients by predicted risk.
3. Compare the positive predictive value by assigning outcomes (TP, FP, TN, FN) to each patient
4. Repeat steps 1:3 B times for uncertainty estimates

Note: The issue this whole time was that by identifying more false positives, it actually makes the model look way worse. True positives are way more expensive than false positives, so the method was actually incentivising finding false positives. We actually want the false positives to be true negatives. Using NMB was not the right approach here.

n\_samples = 40  
sim\_results <- list()  
for (i in 1:1000){  
 df\_youden <- obtain\_sample(df\_sims, "class\_youden", n\_samples)  
 df\_nne <- obtain\_sample(df\_sims, "class\_nne", n\_samples)  
 df\_val\_opt <- obtain\_sample(df\_sims, "class\_val\_opt", n\_samples)  
   
 df\_rank <- df\_sims |>  
 group\_by(auc, p0) |>  
 arrange(desc(predicted)) |>  
 slice\_head(n = n\_samples) |>  
 mutate(Method = "ranking") |>  
 select(Method, auc, p0, predicted, actual)   
   
 iteration <- do.call(rbind, list(df\_youden, df\_nne, df\_val\_opt, df\_rank)) |>  
 group\_by(Method, auc, p0) |>  
 rowwise() |>  
 mutate(Outcome = ifelse(actual == 1, "TP", "FP"),  
 Cost = ifelse(  
 Outcome == "FP",  
 rgamma(1, shape = 110.314, scale = 0.172) \* 3.19 +   
 (p0 \* (1 - rnorm(1, 0.910, 0.036)) \* rnorm(1, 14134, 686)),  
 0),  
 iter = i)  
   
 sim\_results[[i]] <- iteration  
 remove(df\_youden, df\_nne, df\_val\_opt, df\_rank, iteration)  
}  
  
results <- do.call(rbind, sim\_results)  
results$TP\_rate <- ifelse(results$Outcome == "TP", 1, 0)  
  
results <- results |>  
 group\_by(Method, auc, p0, iter) |>  
 summarise(PPV = mean(TP\_rate),  
 Cost = mean(Cost)) |>  
 group\_by(Method, auc, p0) |>  
 summarise(PPV\_median = median(PPV),  
 PPV\_low = quantile(PPV, 0.025),  
 PPV\_high = quantile(PPV, 0.975),  
 Cost\_median = median(Cost),  
 Cost\_low = quantile(Cost, 0.025),  
 Cost\_high = quantile(Cost, 0.975)) |>  
 rename(PPV = PPV\_median,  
 Cost = Cost\_median) |>  
 mutate(Method = case\_when(  
 Method == "nne" ~ "NNE-based threshold",  
 Method == "val\_opt" ~ "Value-optimising threshold",  
 Method == "youden" ~ "Youden index threshold",  
 Method == "ranking" ~ "Ranking by predicted risk"  
 ))  
  
p0\_values <- c(`0.01` = "Prevalence = 0.01",  
 `0.05` = "Prevalence = 0.05",  
 `0.1` = "Prevalence = 0.10")  
  
p\_PPV <- results |>   
 ggplot(aes(x = auc, y = PPV, colour = Method, fill = Method))  
  
p\_PPV +  
 geom\_line(linewidth = 2) +  
 geom\_ribbon(aes(ymin = PPV\_low, ymax = PPV\_high),   
 alpha = 0.2, linetype = "dotted"  
 ) +  
 facet\_wrap(vars(p0), labeller = as\_labeller(p0\_values)) +  
 theme\_bw() +  
 theme(panel.grid.minor = element\_blank(),  
 legend.position = "bottom") +  
 scale\_x\_continuous(limits = c(0.64, 0.96),  
 breaks = seq(0.65, 0.95, 0.1),   
 name = "Model AUC")



ggsave(filename = "PPV\_results.png", height = 6, width = 10)  
  
p\_costs <- results |>  
 ggplot(aes(x = auc, y = Cost, colour = Method, fill = Method))  
  
p\_costs +  
 geom\_line(linewidth = 2) +  
 geom\_ribbon(aes(ymin = Cost\_low, ymax = Cost\_high),   
 alpha = 0.2, linetype = "dotted"  
 ) +  
 facet\_wrap(vars(p0), labeller = as\_labeller(p0\_values)) +  
 theme\_bw() +  
 theme(panel.grid.minor = element\_blank(),  
 legend.position = "bottom") +  
 scale\_x\_continuous(limits = c(0.64, 0.96),   
 breaks = seq(0.65, 0.95, 0.1),   
 name = "Model AUC") +  
 scale\_y\_continuous(breaks = seq(25, 200, 25),   
 name = "Cost of false positives")



ggsave(filename = "Cost\_results.png", height = 6, width = 10)

Bellomo, Rinaldo, Donna Goldsmith, Shigehiko Uchino, Jonathan Buckmaster, Graeme K Hart, Helen Opdam, William Silvester, Laurie Doolan, and Geoffrey Gutteridge. 2003. “A Prospective Before-and-After Trial of a Medical Emergency Team.” *Medical Journal of Australia* 179 (6): 283–87. <https://doi.org/10.5694/j.1326-5377.2003.tb05548.x>.

Blythe, Robin, Rex Parsons, Adrian G. Barnett, Steven M. McPhail, and Nicole M. White. 2023. “Vital Signs-Based Deterioration Prediction Model Assumptions Can Lead to Losses in Prediction Performance.” *Journal of Clinical Epidemiology* 159 (July): 106–15. <https://doi.org/10.1016/j.jclinepi.2023.05.020>.

Blythe, Robin, Rex Parsons, Nicole M White, David Cook, and Steven McPhail. 2022. “A Scoping Review of Real-Time Automated Clinical Deterioration Alerts and Evidence of Impacts on Hospitalised Patient Outcomes.” *BMJ Quality & Safety* 31 (10): 725–34. <https://doi.org/10.1136/bmjqs-2021-014527>.

Bohingamu Mudiyanselage, Shalika, Julie Considine, Alison M. Hutchinson, Imogen Mitchell, Mohammadreza Mohebbi, Jennifer J. Watts, and Tracey K. Bucknall. 2024. “An Economic Evaluation of the Prioritising Responses Of Nurses To Deteriorating Patient Observations (PRONTO) Clinical Trial.” *Resuscitation* 201 (August): 110272. <https://doi.org/10.1016/j.resuscitation.2024.110272>.

Curtis, Kate, Prabhu Sivabalan, David S. Bedford, Julie Considine, Alfa D’Amato, Nada Shepherd, Tiana Elphick, Ramon Z. Shaban, and Margaret Fry. 2021. “Treatments Costs Associated with Inpatient Clinical Deterioration.” *Resuscitation* 166 (September): 49–54. <https://doi.org/10.1016/j.resuscitation.2021.07.022>.

Escobar, Gabriel J., Vincent X. Liu, Alejandro Schuler, Brian Lawson, John D. Greene, and Patricia Kipnis. 2020. “Automated Identification of Adults at Risk for In-Hospital Clinical Deterioration.” *New England Journal of Medicine* 383 (20): 1951–60. <https://doi.org/10.1056/nejmsa2001090>.

Holmes, Emily, Huw Lloyd Williams, Dyfrig Hughes, Elke Naujokat, Bernd Duller, and Christian P. Subbe. 2024. “A Model-Based Cost-Utility Analysis of an Automated Notification System for Deteriorating Patients on General Wards.” Edited by Dominic Luke Thorrington. *PLOS ONE* 19 (5): e0301643. <https://doi.org/10.1371/journal.pone.0301643>.

Parsons, Rex, Robin Blythe, Susanna M Cramb, and Steven M McPhail. 2023. “Integrating Economic Considerations into Cutpoint Selection May Help Align Clinical Decision Support Toward Value-Based Healthcare.” *Journal of the American Medical Informatics Association* 30 (6): 1103–13. <https://doi.org/10.1093/jamia/ocad042>.

Romero-Brufau, Santiago, Jeanne M. Huddleston, Gabriel J. Escobar, and Mark Liebow. 2015. “Why the C-Statistic Is Not Informative to Evaluate Early Warning Scores and What Metrics to Use.” *Critical Care* 19 (1). <https://doi.org/10.1186/s13054-015-0999-1>.

White, Nicole, and Robin Blythe. 2023. “ShinyPrior: A Tool for Estimating Probability Distributions Using Published Evidence.” <http://dx.doi.org/10.31219/osf.io/zf62e>.