Colorectal Neoplasia in Young Adults

Belina Eunyoung Jang[1] Rona Marie A. Lawenko[2]

Jia Belle C. Sta. Maria[2] Rial Juben De Leon[2] Wei Xu[1,3]

Anna Theresa Santiago[3]

[1] Biostatistics Division, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada.

[2] De La Salle University Medical Centre, Cavite, Philippines.

[3] Department of Biostatistics, University Health Network, Toronto, Ontario, Canada.

Abstract

Colorectal cancer (CRC) is a significant health concern globally, being the second leading cause of cancer-related deaths in men and a third in women within Canada with a growing trend among young adults worldwide. This retrospective cohort study of young adult patients in the Philippines identifies risk factors for advanced colorectal neoplasia (ACN) and proposes an age cut-off for CRC screening. Key variables are age, gender, family history of CRC, and lesion location. A multivariable logistic regression model was used to determine significant predictors of ACN. Various multiple regression models were trained using three-fold cross-validation. An age cut-off was determined using the receiver operating characteristic curve

(ROC). We identified age, gender, family history, and lesion location to be the significant predictors of ACN. The odds of ACN increased with age (OR [95% CI] = 1.10 [1.05, 1.16]), in females (OR [95% CI] = 2.62 [1.07, 6.75]), and among individuals with a family history of CRC (OR [95% CI] = 13.5 [3.12, 97.0]). Left-sided lesions such as lesions in the descending colon (OR [95% CI] = 5.78 [1.31, 29.0]) were more strongly associated with ACN than right-sided lesions in the cecum and ascending colon. Logistic regression resulted in the highest mean and median F1 score. The ROC identified 36 years as the optimal age cut-off for ACN risk stratification. These findings can inform health policy to improve early detection and provide a foundation for revising the recommended screening age for CRC in the Philippines.

Introduction

Background and Motivation

Early-onset colorectal cancer (EOCRC) is cancer in the colon or rectum that occurs among patients less than 50 years of age. Colorectal cancer (CRC) is the 2nd leading cause of death from cancer in men and the 3rd leading cause of death from cancer in women in Canada. (Canadian Cancer Society 2024) While it's commonly considered a disease of older populations, there is a growing trend of CRC cases among young adults worldwide. (Akimoto et al. 2021) Unfortunately, the literature on EOCRC in Southeast Asia, particularly in the Philippines, remains limited, making this a critical area for future research. The growing EOCRC rate raises concerns about whether the current recommended screening age of 50 years in the Philippines is sufficient, prompting calls to reevaluate the screening guidelines in light of the rising incidence of CRC among younger adults.

Advanced Colorectal Neoplasia

Advanced Colorectal Neoplasia (ACN) includes premalignant polyps and colorectal cancer (carcinoma). Premalignant polyps include adenomas, which are non-cancerous tumors with a diameter of 10 millimeters or more, sessile serrated polyps, tubular adenomas with high-grade dysplasia, as well as tubulovillous or villous adenomas. These polyps are considered to have a higher risk of progressing to colorectal cancer. Colorectal cancer (carcinoma) itself is confirmed through histopathological diagnosis. Understanding these two categories of ACN is essential to identifying and predicting risk factors of CRC. Figure 1 shows how premalignant or precancerous polyps can develop into cancer if left untreated.

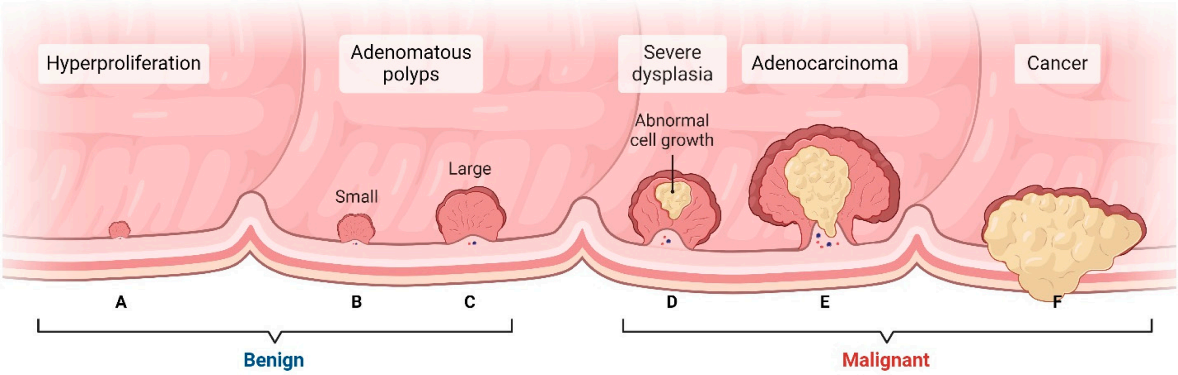


Figure 1: Adenoma-carcinoma sequence of colorectal cancer development (Gharib and Robichaud 2024)

Research Objective

We analyzed data from a retrospective cohort of young adult patients who underwent their first colonoscopy at a private academic hospital in the Philippines. The primary objectives of this study are to identify risk factors associated with advanced colorectal neoplasia, to develop predictive models to support early screening for colorectal cancer, and to propose an evidence-based age cutoff for active colorectal cancer screening. ACN, which includes premalignant polyps and colorectal cancer, was used as the primary outcome due to the small number of colorectal cancer cases in the dataset. These findings aim to inform health policies tailored to patient-specific conditions such as family history and other risk factors.

Materials and Methods

Study design and population

We conducted a retrospective cohort study of 130 young adult patients (aged 18–49) who underwent their first diagnostic or screening colonoscopy between 2018 and 2020 at a private academic hospital, De La Salle University Medical Center (DLSUMC), in the province of Cavite, Philippines. The dataset was derived from a random sample of 136 eligible patients selected by the original study team from all individuals in this age range who underwent their first colonoscopy during the study period using a probability-based, specifically stratified random sampling method (Appendix Table A1). The original population included all patients (aged $18 - 49$) who underwent their first diagnostic or screening colonoscopy between 2018 and 2020 at DLSUMC. The target sample size of 136 was determined by the original team, using OpenEpi for their primary objective of estimating the prevalence of early-onset colorectal neoplasia, assuming a hypothesized prevalence of 30%, a 95% confidence level and a 5% margin of error. The data were de-identified before being transferred for the analysis.

The aim of the study was to identify predictors of advanced colorectal neoplasia (ACN) in young adults. ACN was defined as a primary binary outcome variable including histologically confirmed colorectal cancer and high-risk premalignant polyps (e.g., $\geq$ 10mm adenomas, sessile serrated polyps, or tubular adenomas with high-grade dysplasia). Colorectal cancer (CRC) was defined separately as a secondary binary outcome variable indicating the presence of histologically confirmed colorectal cancer.

Predictor variables included patient demographic and clinical characteristics: age at colonoscopy (discrete), gender (female or male), family history of colorectal cancer (with or without family history), indication for colonoscopy (screening, diagnostic-bleeding, or diagnostic-other), location of the lesion (cecum and ascending colon, rectosigmoid colon and anorectum, descending colon, transverse colon) and type of service (private or charity).

Data Preprocessing and Cleaning

Data cleaning involved inspecting values, verifying data consistency, checking for missing values, and appropriately formatting each variable. All character-type variables were standardized to proper case formatting for consistency. The numeric variable, age, was converted from character strings into numeric values. Age was then used to create an additional variable, age group, to group patients into three clinically meaningful categories: < 30, 30–39, and 40–49 years. We also coded categorical variables as factors with meaningful labels and designated reference levels. Locations of the lesion were combined when they were anatomically adjacent (e.g., “cecum” and “ascending colon” were merged into “cecum and ascending colon”). All variables included in the model had no missing data; therefore, no imputation was required.

Statistical Analysis

Descriptive Statistics

Baseline characteristics were summarized using the `table1` package in R. Continuous variables were reported as medians with ranges [minimum, maximum], and categorical variables as counts with corresponding percentages. The summary Table 1 (primary analysis) was stratified by ACN status, and Table 2 (secondary analysis) was stratified by CRC status.

Categorical Data Analysis

To identify significant predictors of ACN, we fitted a multivariable logistic regression model (glm). ACN was used as the binary outcome variable, and all other covariates were included as predictors in the model. The model is specified as follows:

$$\begin{aligned}
\text{logit}\big[P(\text{ACN}=1)\big] = {} & -4.22276 + 0.09571 \cdot \text{age} + 0.96297 \cdot 1_{\{\text{gender=Female}\}} \\
& + 2.60260 \cdot 1_{\{\text{family history=With Family History}\}} \\
& - 0.91690 \cdot 1_{\{\text{indication=Diagnostic-Bleeding}\}} \\
& - 0.67457 \cdot 1_{\{\text{indication=Diagnostic-Other}\}} \\
& + 1.55080 \cdot 1_{\{\text{location=Rectosigmoid Colon and Anorectum}\}} \\
& + 1.75394 \cdot 1_{\{\text{location=Descending Colon}\}} \\
& + 1.62418 \cdot 1_{\{\text{location=Transverse Colon}\}}
\end{aligned}$$

Adjusted odds ratios (aORs) and 95% confidence intervals (CIs) were calculated using the `gtsummary` package.

The assumptions for the logistic regression model were considered. First, the dependent variable must be binary, which was satisfied in our study since ACN and CRC were both coded as binary indicators representing the presence or absence of the corresponding condition. Second, observations should be independent, which we assumed to be true as each observation corresponded to a unique patient, and there was no known indication of dependency, such as repeated measurement or clustering, within the dataset. Third, there should be little to no multicollinearity among the independent variables. We assessed this by calculating the adjusted generalized variance inflation factor (GVIF) for the six covariates used in the model. The adjusted GVIF values ranged from 1.02 to 1.09, which are extremely low, indicating very little to no multicollinearity among the six covariates.

Secondary Analysis for Colorectal Cancer (CRC)

As a secondary analysis, we examined predictors of histologically diagnosed CRC among the study population. CRC was treated as a binary outcome variable, and the same set of covariates used in the ACN analysis was also included as predictors. Due to the small proportion of patients with CRC ($n = 29$), this analysis was considered exploratory. A multivariable logistic regression model was fitted, and adjusted odds ratios (aORs) with 95% confidence intervals (CIs) were reported. Results of this analysis are presented in the Secondary Analysis for Colorectal Cancer (CRC) subsection of the Results.

Predictive Modelling for ACN

For the predictive modelling with cross-validation, we trained four regression models, logistic, ridge, lasso, and elastic net regression, to predict the development of ACN. Then we assessed and compared predictive performances of the logistic, ridge, lasso, and elastic net regression models using three-fold cross-validation.

Optimal Age Cut-off for ACN

We used the `cutpointr` package in R to calculate the optimal age cut-off for ACN risk stratification. This package supports a variety of options for estimating cutpoints and can optimize common metrics such as the sum of sensitivity and specificity, or Youden's Index. It also allows bootstrapping, so we can estimate the variability and stability of the cutpoint. In this study, the binary outcome was the presence of ACN, and the continuous predictor was age. We applied the "maximize_metric" method and specified the Youden's Index as the metric to obtain an age value that maximizes Youden's Index. This value served as the optimal age cut-off for ACN risk stratification.

Results

The dataset originally contained 136 observations and six observations were excluded since the location of the lesion was the ileum, which is not part of the colon or rectum. Therefore, our analytic sample resulted in 130 observations.

Statistical Analysis

Descriptive Statistics

We summarized the demographic and clinical characteristics of the study population in Table 1 (stratified by ACN) and Table 2 (stratified by CRC). The cohort included a mix of patients with and without a family history of CRC, varied lesion locations, and different service types. There are two variables, age (discrete) and age group (categorical), for the same underlying information because the project focused on the effect of age, and we aimed to better describe the age distribution in the dataset. The age group classification was based on guidelines from the Philippine Society of Gastroenterologists. However, this age group classification was not used in the models we built.

Table 1: Patient Characteristics by ACN status

	ACN-negative	ACN-positive	Total
	(N=57)	(N=73)	(N=130)
Age at Colonoscopy			
Mean (SD)	31.8 (7.43)	37.5 (8.12)	35.0 (8.31)
Median [Min, Max]	31.0 [18.0, 48.0]	39.0 [21.0, 49.0]	34.0 [18.0, 49.0]
Age Group at Colonoscopy			
18-29	22 (38.6%)	16 (21.9%)	38 (29.2%)
30-39	25 (43.9%)	21 (28.8%)	46 (35.4%)
40-49	10 (17.5%)	36 (49.3%)	46 (35.4%)
Gender			
Female	19 (33.3%)	37 (50.7%)	56 (43.1%)
Male	38 (66.7%)	36 (49.3%)	74 (56.9%)
Family History			
Without Family History	55 (96.5%)	55 (75.3%)	110 (84.6%)
With Family History	2 (3.5%)	18 (24.7%)	20 (15.4%)
Indication for Colonoscopy			
Diagnostic-Bleeding	24 (42.1%)	31 (42.5%)	55 (42.3%)
Diagnostic-Other	32 (56.1%)	39 (53.4%)	71 (54.6%)
Screening	1 (1.8%)	3 (4.1%)	4 (3.1%)
location_x			
Cecum and Ascending Colon	22 (38.6%)	10 (13.7%)	32 (24.6%)
Rectosigmoid Colon and Anorectum	24 (42.1%)	41 (56.2%)	65 (50.0%)
Descending Colon	5 (8.8%)	10 (13.7%)	15 (11.5%)
Transverse Colon	6 (10.5%)	12 (16.4%)	18 (13.8%)

Data given as count (proportion in %) or median [Min, Max].

Table 2: Patient Characteristics by CRC status

	CRC-negative	CRC-positive	Total
	(N=101)	(N=29)	(N=130)
Age at Colonoscopy			
Mean (SD)	34.1 (8.17)	38.3 (8.07)	35.0 (8.31)
Median [Min, Max]	33.0 [18.0, 49.0]	41.0 [22.0, 49.0]	34.0 [18.0, 49.0]
Age Group at Colonoscopy			
18-29	31 (30.7%)	7 (24.1%)	38 (29.2%)
30-39	40 (39.6%)	6 (20.7%)	46 (35.4%)
40-49	30 (29.7%)	16 (55.2%)	46 (35.4%)
Gender			
Female	39 (38.6%)	17 (58.6%)	56 (43.1%)
Male	62 (61.4%)	12 (41.4%)	74 (56.9%)
Family History			
Without Family History	86 (85.1%)	24 (82.8%)	110 (84.6%)
With Family History	15 (14.9%)	5 (17.2%)	20 (15.4%)
Indication for Colonoscopy			
Diagnostic-Bleeding	45 (44.6%)	10 (34.5%)	55 (42.3%)
Diagnostic-Other	54 (53.5%)	17 (58.6%)	71 (54.6%)
Screening	2 (2.0%)	2 (6.9%)	4 (3.1%)
location_x			
Cecum and Ascending Colon	28 (27.7%)	4 (13.8%)	32 (24.6%)
Rectosigmoid Colon and Anorectum	48 (47.5%)	17 (58.6%)	65 (50.0%)
Descending Colon	12 (11.9%)	3 (10.3%)	15 (11.5%)
Transverse Colon	13 (12.9%)	5 (17.2%)	18 (13.8%)

Data given as count (proportion in %) or median [Min, Max].

Figure 2 (A) shows the percentage of ACN-positive and ACN-negative cases for each age group. Among patients aged 18 to 29 years, the proportion of ACN-negative cases was 15.8% greater than that of ACN-positive cases. This difference got smaller to 8.6% in the 30–39 age group. In contrast, in the 40–49 age group, the proportion of ACN-positive cases becomes 56.6% greater than that of ACN-negative cases.

Figure 2 (B) shows the percentage of CRC-positive and CRC-negative cases for each age group within the ACN-positive group. Since this is within the ACN-positive group, CRC-negative represents high-risk premalignant polyps (precancer), and CRC-positive represents histologically confirmed CRC. This figure illustrates the age breakdown of patients with ACN and with known CRC. We can see that CRC-positive is still the highest in the 40–49 age group among the patients with ACN.

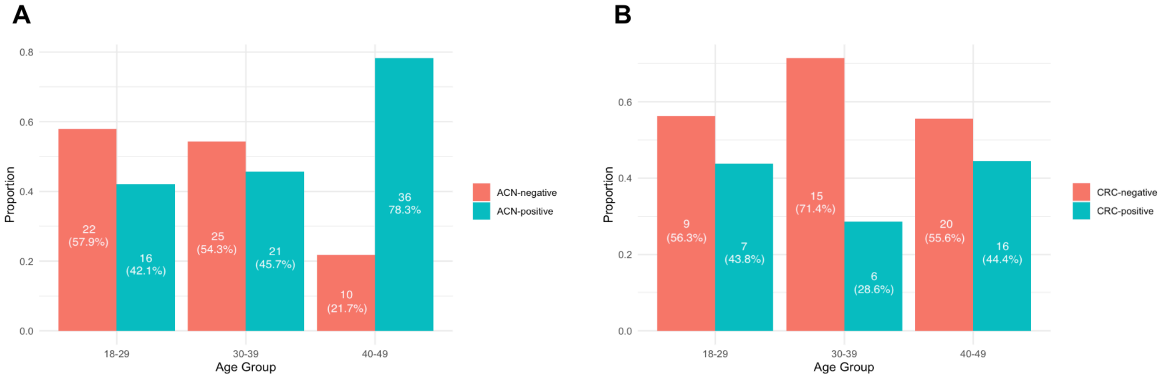
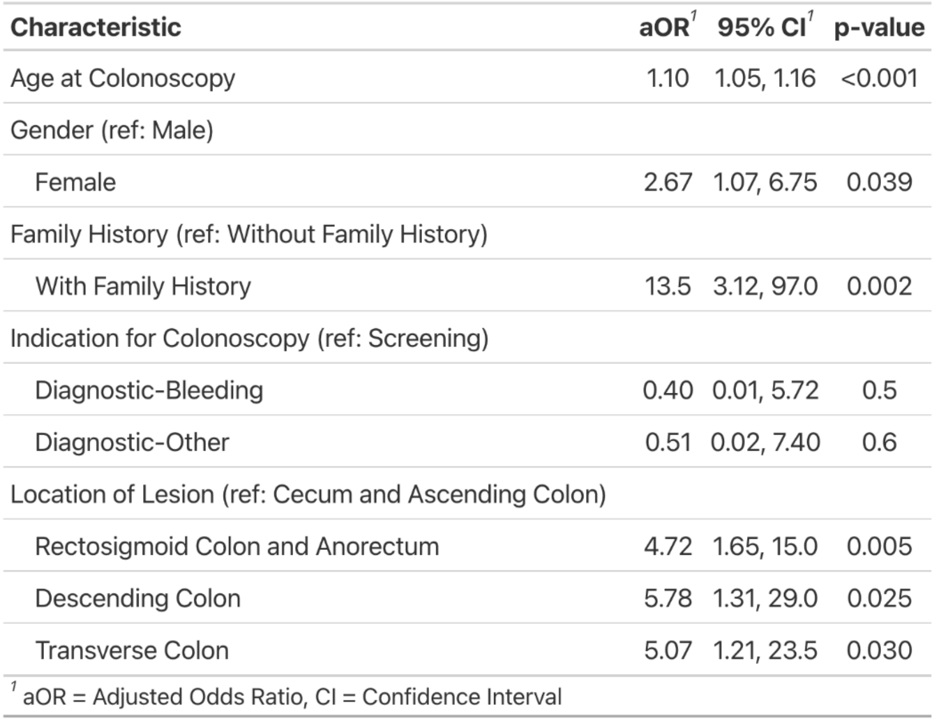


Figure 2: Proportion of advanced colorectal neoplasia (A), and proportion of colorectal carcinoma within ACN-positive cases (B) by age group

Categorical Data Analysis

Table 3 presents the results of the multivariable logistic regression model for ACN. Significant predictors of ACN included age, gender, family history of CRC, and lesion location. First, we can see that the odds of ACN increase when you are older, when you are a female, and with a positive family history. Additionally, as you can see from Figure 3, the reference level of location, which is the cecum and ascending colon, is located on the right side of the abdomen. And the odds ratios (ORs) of all the other left side locations with respect to the right side (the reference level) are well above 1. So the odds of ACN increase when the location of the lesion is on the left side. This finding is consistent with the regional literature and aligns with the clinician's experience that left-sided colorectal cancers are more common than right-sided colorectal cancers.

Table 3: Multivariable logistic regression model for ACN



Characteristic	**aOR**[1]	**95% CI**[1]	**p-value**
Age at Colonoscopy	1.10	1.05, 1.16	<0.001
Gender (ref: Male)			
Female	2.67	1.07, 6.75	0.039
Family History (ref: Without Family History)			
With Family History	13.5	3.12, 97.0	0.002
Indication for Colonoscopy (ref: Screening)			
Diagnostic-Bleeding	0.40	0.01, 5.72	0.5
Diagnostic-Other	0.51	0.02, 7.40	0.6
Location of Lesion (ref: Cecum and Ascending Colon)			
Rectosigmoid Colon and Anorectum	4.72	1.65, 15.0	0.005
Descending Colon	5.78	1.31, 29.0	0.025
Transverse Colon	5.07	1.21, 23.5	0.030

[1] aOR = Adjusted Odds Ratio, CI = Confidence Interval

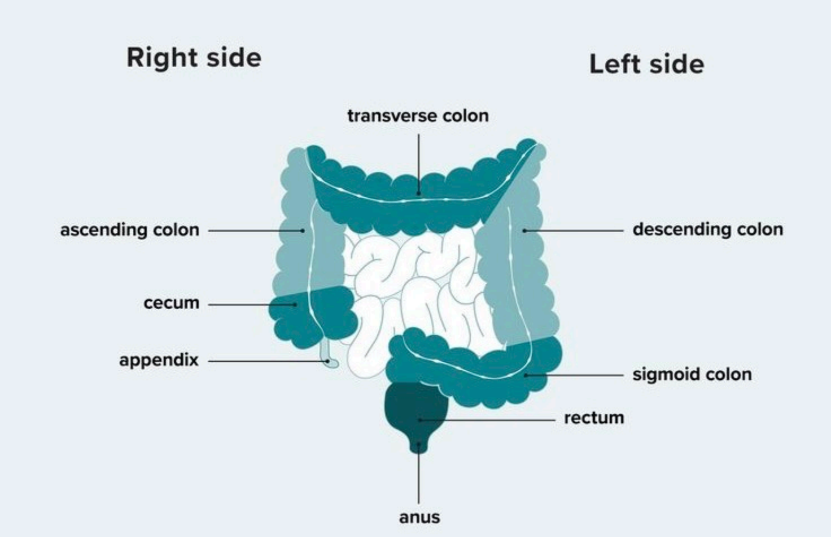
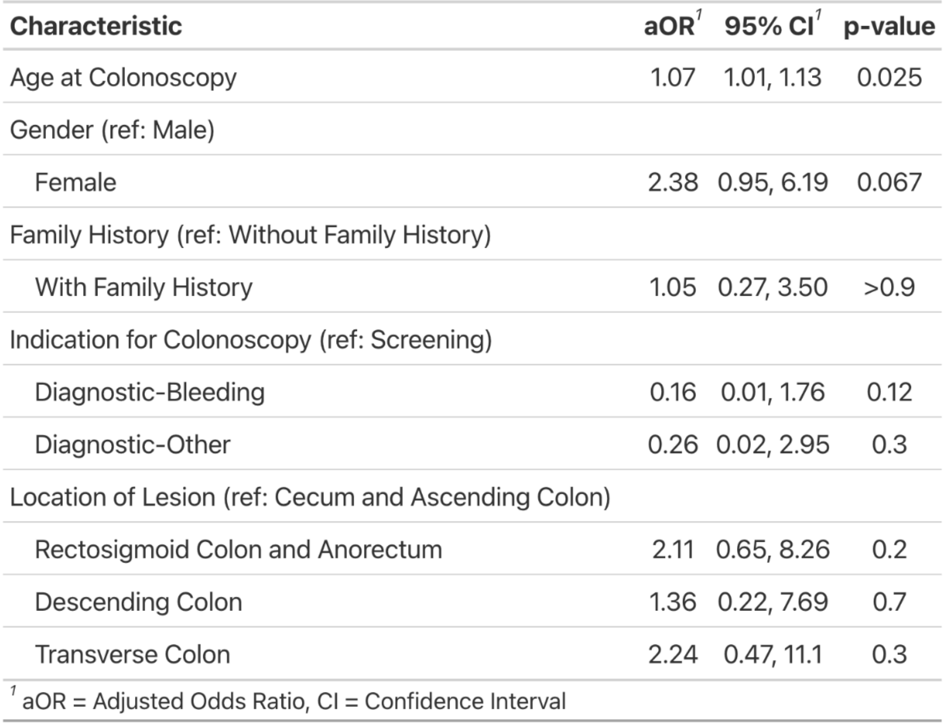


Figure 3: Locations of right- and left-sided colon cancer (Yetman 2024)

Secondary Analysis for Colorectal Cancer (CRC)

Table 4 presents the results of the multivariable logistic regression model for CRC. The direction of associations for age, gender and lesion location was consistent with those observed in the primary ACN analysis, with older age, female gender, a positive family history and right-sided lesions associated with higher odds of CRC positivity. However, most covariates did not have statistically significant differences between their levels, likely due to the smaller number of CRC-positive cases (n=29), which reduced statistical power despite trends similar to those seen in Table 1.

Table 4: Multivariable logistic regression model for CRC



Characteristic	aOR[1]	95% CI[1]	p-value
Age at Colonoscopy	1.07	1.01, 1.13	0.025
Gender (ref: Male)			
Female	2.38	0.95, 6.19	0.067
Family History (ref: Without Family History)			
With Family History	1.05	0.27, 3.50	>0.9
Indication for Colonoscopy (ref: Screening)			
Diagnostic-Bleeding	0.16	0.01, 1.76	0.12
Diagnostic-Other	0.26	0.02, 2.95	0.3
Location of Lesion (ref: Cecum and Ascending Colon)			
Rectosigmoid Colon and Anorectum	2.11	0.65, 8.26	0.2
Descending Colon	1.36	0.22, 7.69	0.7
Transverse Colon	2.24	0.47, 11.1	0.3

[1] aOR = Adjusted Odds Ratio, CI = Confidence Interval

Predictive Modelling for ACN

We compared the performance of four predictive models namely, logistic, ridge, lasso and elastic net regression, in predicting the presence of ACN. The box plots further illustrate the consistency and reliability of each model's performance in predicting ACN. While all four models showed very similar performance, Figure 4 (D) (box plots of F1 Score) shows that logistic and ridge regression models had the highest median F1 scores, indicating their effectiveness in balancing between precision and recall (sensitivity). The logistic regression also had the highest mean F1 score.

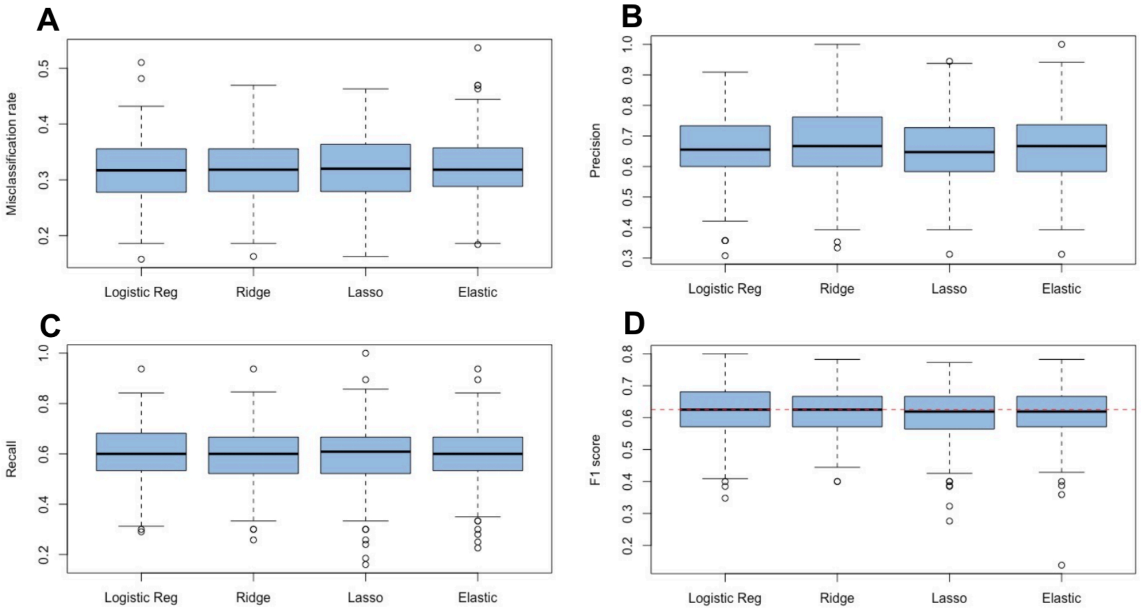


Figure 4: Box plots of misclassification rate (A), precision (B), recall (C), and F1 score (D) for four classification models

Optimal Age Cut-off for ACN

The top left panel of Figure 5 shows the count of ACN-positive and ACN-negative cases with corresponding age value. It also shows the calculated optimal age cut-off of 36 years marked by a thin vertical line. The top right panel displays the ROC curve, which plots sensitivity (true positive rate) against 1-specificity (false positive rate) at various cutoff points. The dot represents the performance at the optimal age cut-off of 36 years. The bottom left panel shows the bootstrap distribution of optimal cutpoints, which illustrates the variability and robustness of the selected threshold. The bottom right panel shows the bootstrap distribution of Youden's Index (out-of-bag estimates), providing an assessment of the stability of the optimal cutpoint's performance across bootstrap samples.

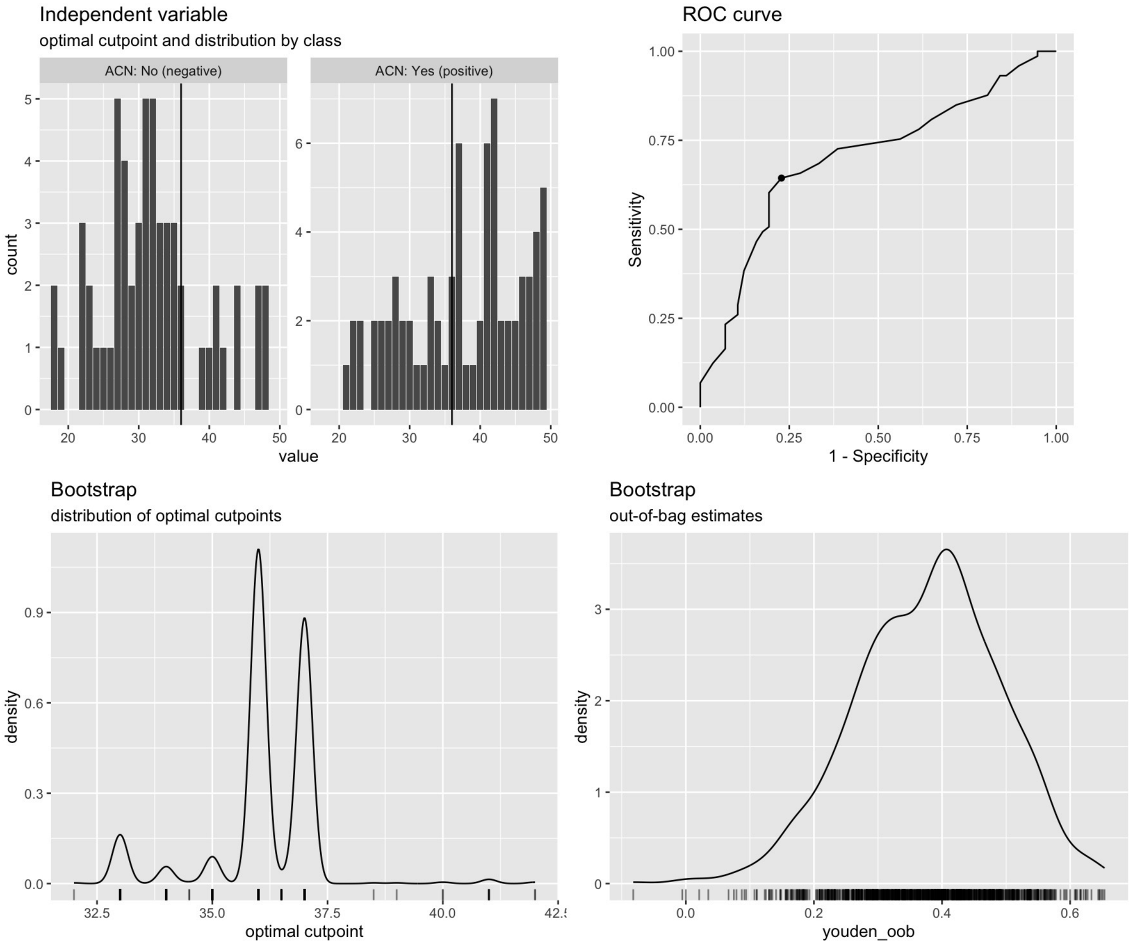


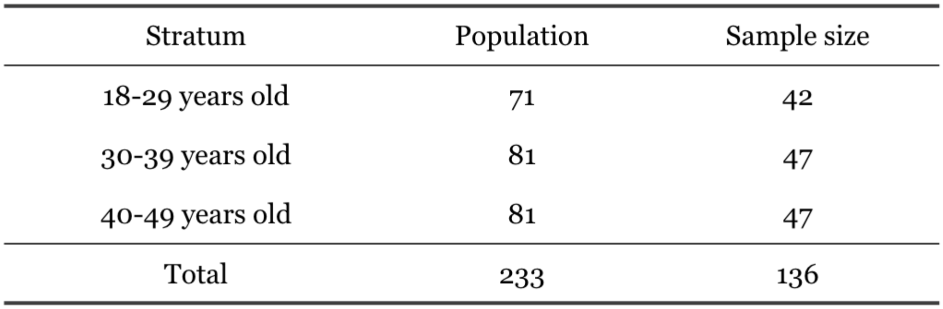
Figure 5: Optimal age cutpoint determination using cutpointr

Discussion

Our findings align well with the global literature, which shows a stepwise increase in neoplasia with age and a strong influence of family history. (Trivedi et al. 2022) Interestingly, while CRC is generally more common in males, our data showed a higher odds of ACN in females, which is consistent with regional trends reported in Southeast Asia. (Danpanichkul et al. 2023) Because our sample was collected from a single private hospital with patients possibly largely from higher socioeconomic backgrounds, the findings may not be fully generalizable to the broader population. Additionally, due to the small number of CRC-positive cases (n = 29), we used ACN as the primary outcome variable. Our results suggest that current screening guidelines, which typically recommend the start of screening at age 50, may not be fully optimal for young adults in Southeast Asia. Therefore, incorporating the identified risk factors and the proposed age cut-off into screening strategies could greatly help inform future screening policies and enhance early detection, improving patient prognosis.

Appendix

Table A1: Population distribution and stratified random sample sizes by age group in the original Philippine study



Stratum	Population	Sample size
18-29 years old	71	42
30-39 years old	81	47
40-49 years old	81	47
Total	233	136

R code

```r
# Load necessary libraries
library(tidyverse)
library(janitor)
library(here)
library(texreg)
library(multgee)
library(table1)
library(readxl)
library(dplyr)
library(flextable)
library(glmnet)
library(Matrix)
library(sjPlot)
library(caret)

library(simplecolors)

# Table 1

## Data preparation for Table 1

# load the excel data
original_data <- read_xlsx(here("data/original",
 ↪  "Copy_Statistical-and-Analytical-Test-V4.xlsx"), sheet=1)
original_data <- data_frame(original_data)

# drop first row (contains explanation for each variable)
cleaned_data <- original_data[-1,]
cleaned_data <- clean_names(cleaned_data)

lapply(cleaned_data, unique)
```

```
# remove rows where location_x = "ILEUM" - 6 rows
cleaned_data <- cleaned_data %>% filter(location_x != "ILEUM")

cleaned_data$location_x <- str_to_title(cleaned_data$location_x)

# change type to numeric
cleaned_data$year <- as.numeric(cleaned_data$year)
cleaned_data$age_x <- as.numeric(cleaned_data$age_x)

# rename the variable acn_y_n_y to acn_y
cleaned_data <- cleaned_data %>% rename(acn_y = acn_y_n_y)

cleaned_data

#Labeling the data & grouping

# predictors (Y)
cleaned_data$acn_y <- factor(cleaned_data$acn_y, levels=c("Y","N"),
 ↪  labels=c("ACN-positive","ACN-negative"))

# below fixed glm ??
cleaned_data$acn_y <- relevel(cleaned_data$acn_y, ref="ACN-negative")

cleaned_data$premalignant_y <- factor(cleaned_data$premalignant_y,
 ↪  levels=c("Y","N"),
                                      labels=c("Yes","No"))
cleaned_data$crc_y <- factor(cleaned_data$crc_y, levels=c("Y","N"),
 ↪  labels=c("Yes","No"))

# covariates (X)
cleaned_data$gender_x <- factor(cleaned_data$gender_x,
 ↪  levels=c("FEMALE","MALE"), labels=c("Female","Male"))

cleaned_data$family_history_x <- factor(cleaned_data$family_history_x,
 ↪  levels=c("With Family History","No Family History"), labels=c("With
 ↪  Family History","Without Family History"))

cleaned_data$family_history_x <- relevel(cleaned_data$family_history_x,
 ↪  ref="Without Family History")

cleaned_data$indication_x <- factor(cleaned_data$indication_x,
                                    levels=c("DO","DB","S"),
                                    labels=c(
                                      "Diagnostic-Bleeding",
                                      "Diagnostic-Other",
                                      "Screening"))

cleaned_data$location_x <- as.factor(cleaned_data$location_x)

# others
cleaned_data$service <- as.factor(cleaned_data$service)
```

```r
# Age at Colonoscopy To be grouped into 18-29, 30-39, 40-49
cleaned_data$age_x_group <- cut(
  cleaned_data$age_x,
  breaks = c(18, 30, 40, 50),
  labels = c("18-29", "30-39", "40-49"),
  right = FALSE
)

label(cleaned_data$report_id) <- "Subject ID"
label(cleaned_data$year) <- "Year of Colonoscopy"
label(cleaned_data$age_x) <- "Age at Colonoscopy"
label(cleaned_data$age_x_group) <- "Age Group at Colonoscopy"
label(cleaned_data$gender_x) <- "Gender"
label(cleaned_data$service) <- "Service"
label(cleaned_data$family_history_x) <- "Family History"
label(cleaned_data$indication_x) <- "Indication for Colonoscopy"
label(cleaned_data$location_x) <- "Location of Lesion"
label(cleaned_data$acn_y) <- "Advanced Colorectal Neoplasm"
label(cleaned_data$premalignant_y) <- "Premalignant Polyp"
label(cleaned_data$crc_y) <- "Colorectal Carcinoma"

# combine location_x values into groups
cleaned_data$location_x <- cleaned_data$location_x %>% fct_collapse(
  "Cecum and Ascending Colon" = c("Ascending Colon", "Cecum"),
  "Transverse Colon" = "Transverse Colon",
  "Descending Colon" = "Descending Colon",
  "Rectosigmoid Colon and Anorectum" = c("Anorectum", "Rectosigmoid",
   ↪  "Sigmoid", "Rectum")
)

cleaned_data$location_x <- relevel(cleaned_data$location_x, ref="Cecum and
 ↪  Ascending Colon")

cleaned_data

# count the number of levels for acn_y grouped by crc_y
cleaned_data %>% group_by(crc_y, acn_y) %>% summarise(n=n())


## Table 1-A with total

caption  <- "Patient Characteristics by ACN status"
footnote <- "Data given as count (proportion in %) or median [Min, Max]."

# with total & no p-values
table1_1 <- table1( ~
 ↪  age_x+age_x_group+gender_x+family_history_x+indication_x+location_x |
 ↪  acn_y, data = cleaned_data, overall=c(Left="Total"), caption=caption,
 ↪  footnote=footnote)
# save the table1 as jpg
save_as_image(t1flex(table1_1), here("output", "table1-a.png"))
```

```
save_as_html(t1flex(table1_1), path=here("output", "table1-a.html"))
save(table1_1, file = "table1_a.RData")

library(magrittr)
library(flextable)

# Table 1 for crc_y

cleaned_data$crc_y <- factor(cleaned_data$crc_y, levels=c("Yes","No"),
 ↪  labels=c("CRC-positive","CRC-negative"))
cleaned_data$crc_y <- relevel(cleaned_data$crc_y, ref="CRC-negative")

caption <- "Patient Characteristics by CRC status"

table1_2 <- # with total & no p-values
  table1( ~
   ↪  age_x+age_x_group+gender_x+family_history_x+indication_x+location_x |
   ↪  crc_y, data = cleaned_data, overall=c(Left="Total"), caption=caption,
   ↪  footnote=footnote)

table1_2

# save the table1 as jpg
save_as_image(t1flex(table1_2), here("output", "table1-b.png"))
save_as_html(t1flex(table1_2), path=here("output", "table1-b.html"))
save(table1_2, file = "table1_b.RData")
```

```
# bar graph (yes vs no count for each age group)

cleaned_data %>%
  ggplot(aes(x=age_x_group, fill=acn_y)) +
  geom_bar(position="dodge") +
  labs(title="Graph 1. Advanced Colorectal Neoplasm by Age Group", x="Age
   ↪  Group", y="Count") +
  theme_minimal()

cleaned_data %>%
  ggplot(aes(x=age_x_group, fill=acn_y)) +
  geom_bar(position="fill") +
  labs(title="Graph 1. Advanced Colorectal Neoplasm by Age Group", x="Age
   ↪  Group", y="Count") +
  theme_minimal()

# Calculate proportions
proportions_data <- cleaned_data %>%
  group_by(age_x_group, acn_y) %>%
  summarise(count = n(), .groups = "drop") %>%
  group_by(age_x_group) %>%
  mutate(proportion = count / sum(count))

# Create the plot
proportions_data %>%
  ggplot(aes(x = age_x_group, y = proportion, fill = acn_y)) +
  geom_col(position = "dodge") +
  labs(title = "Graph 1. Advanced Colorectal Neoplasm by Age Group",
       x = "Age Group",
       y = "Proportion") +
  theme_minimal()

# bar graph (yes vs no count for each age group for ACN: Yes)

cleaned_data %>%
  filter(acn_y == "ACN-positive") %>%
  ggplot(aes(x=age_x_group, fill=crc_y)) +
  geom_bar(position="dodge") +
  labs(title="Graph 2. Colorectal Carcinoma by Age Group within ACN: Yes
   ↪  (positive) Group", x="Age Group", y="Count") +
  theme_minimal()

cleaned_data %>%
  filter(acn_y == "ACN-positive") %>%
  ggplot(aes(x=age_x_group, fill=crc_y)) +
  geom_bar(position="fill") +
  labs(title="Graph 2. Colorectal Carcinoma by Age Group within ACN: Yes
   ↪  (positive) Group", x="Age Group", y="Count") +
  theme_minimal()

# Calculate proportions
```

```
proportions_data2 <- cleaned_data %>%
  filter(acn_y == "ACN-positive") %>%
  group_by(age_x_group, crc_y) %>%
  summarise(count = n(), .groups = "drop") %>%
  group_by(age_x_group) %>%
  mutate(proportion = count / sum(count))

# Create the plot
proportions_data2 %>%
  ggplot(aes(x = age_x_group, y = proportion, fill = crc_y)) +
  geom_col(position = "dodge") +
  labs(title = "Graph 2. Colorectal Carcinoma by Age Group within ACN: Yes
   ↪  (positive) Group",
       x = "Age Group",
       y = "Proportion") +
  theme_minimal()

# number of CRC patients in each age group

# bar graph (ACN yes count for each age group) just yes

cleaned_data %>%
  filter(crc_y == "CRC-positive") %>%
  ggplot(aes(x=age_x_group, fill=crc_y)) +
  geom_bar(position="dodge") +
  labs(title="Graph 3. Colorectal Carcinoma by Age Group", x="Age Group",
   ↪  y="Count") +
  theme_minimal()

# models (logistic regression)

## table2: naive glm (without cross validation)

# set reference levels
cleaned_data$family_history_x <- relevel(cleaned_data$family_history_x,
 ↪  ref="Without Family History")

cleaned_data$indication_x <- relevel(cleaned_data$indication_x,
 ↪  ref="Screening")

cleaned_data$gender_x <- relevel(cleaned_data$gender_x, ref="Male")

cleaned_data2 <- cleaned_data %>% dplyr::select(-c(premalignant_y, crc_y,
 ↪  report_id, year, age_x_group, service))

glm(acn_y~., data=cleaned_data2, family="binomial")

glm(acn_y~., data=cleaned_data2, family="binomial") %>% sjPlot::tab_model()
summary(glm(acn_y~., data=cleaned_data2, family="binomial"))

#table2
table2 <- glm(acn_y~., data=cleaned_data2, family="binomial") %>%
 ↪  gtsummary::tbl_regression(exponentiate=T)
```

```
# save as html
table2 %>%
  gtsummary::as_gt() %>%
  gt::gtsave(here("output", "table2.html"))

# naive glm for crc_y (entire dataset)

# naive glm for crc_y
cleaned_data3 <- cleaned_data %>% dplyr::select(-c(premalignant_y, acn_y,
 ↪  report_id, year, age_x_group, service))

table3 <- glm(crc_y~., data=cleaned_data3, family="binomial") %>%
 ↪  gtsummary::tbl_regression(exponentiate=T)

# save as html
table3 %>%
  gtsummary::as_gt() %>%
  gt::gtsave(here("output", "table3.html"))

## table 4: naive glm (without cross validation) for crc_y within acn_y yes

table4_data <- cleaned_data %>%filter(acn_y=="ACN-positive") %>%
 ↪  dplyr::select(-c(premalignant_y, acn_y, report_id, year, age_x_group,
 ↪  service))
glm(crc_y~., data=table4_data, family="binomial") %>%
 ↪  gtsummary::tbl_regression(exponentiate=T)
```

```
## preparing data for training and testing

cleaned_data2 <- as.data.frame(cleaned_data2)
# changes it back to binary
cleaned_data2$acn_y <- factor(cleaned_data2$acn_y,
 ↪  levels=c("ACN-negative","ACN-positive"), labels=c(0,1))

# reference at 0: No
#cleaned_data2$gender_x <- factor(cleaned_data2$gender_x,
 ↪  levels=c("Female","Male"),labels=c(0,1))

## multicollinearity check

# check multicollinearity using VIF
library(car)

# fit the logistic regression model
fit <- glm(acn_y ~ age_x + gender_x + family_history_x + indication_x +
 ↪  location_x,
           data = cleaned_data2,
           family = binomial)

# calculate VIF values
vif_values <- vif(fit)
print(vif_values)
```

```r
## Model Training and Testing

# cross-validation
nfold  = 3 #5-folds
split = cleaned_data2['acn_y']

# split data into acn_y=1 and acn_y=0
# This ensures each fold has about the same proportion of resistance vs
 ↪  non-resistance.
data.c = cleaned_data2[split==1,] #data where acn_y=1
data.nc = cleaned_data2[split==0,] #data where acn_y=0

# sample the index
#set.seed(0)
idx.c = sample(1:nfold, sum(split==1), replace=T) # acn_y=1 positive
idx.nc = sample(1:nfold, sum(split==0), replace=T) # acn_y=0 negative

## Train and Test of model

cv.accu = function(nfold, data.c, data.nc, idx.c, idx.nc, returntype)
{
  for(ii in 1:nfold){
    cM1 = cM2 = cM3 = cM4 = matrix(0, 2, 2)

    data.train = rbind(data.c[idx.c!=ii,], data.nc[idx.nc!=ii,])
    data.test  = rbind(data.c[idx.c==ii,], data.nc[idx.nc==ii,])

    # logistic regression for acn_y
    model1 = glm(acn_y~., data=data.train, family="binomial")
    pred1 = try(predict(model1, newdata=data.test, type="response"))
    if(class(pred1)=="try-error") next

    # ridge
    # data for glmnet
    x_train = model.matrix(acn_y ~ ., data.train)
    y_train = data.train$acn_y
    x_test  = model.matrix(acn_y ~ ., data.test)
    y_test  = data.test$acn_y

    #cat("N fold Iteration:", ii, "\n")
    #cat("Train rows:", nrow(x_train), "\n")
    #cat("Test rows:", nrow(x_test), "\n")


    # Ridge regression binomial??
    model2 = glmnet(x_train, y_train, family="binomial", alpha=0)
    ridge.cv.out = cv.glmnet(x_train, y_train, family="binomial", alpha=0)
    bestlam = ridge.cv.out$lambda.min
    pred2 = predict(model2, s=bestlam, newx=x_test, type = "response")

    # Lasso
```

```r
model3 <- glmnet(x_train, y_train, family="binomial", alpha = 1)
lasso.cv.out <- cv.glmnet(x_train, y_train, family="binomial", alpha = 1)
bestlam <- lasso.cv.out$lambda.min
pred3=predict(model3, s=bestlam, newx=x_test, type="response")

#cat("pred: ",nrow(pred3),"\n")


# elastic net - not working 3 more rows than others??
model4 <- glmnet(x_train, y_train, family="binomial", alpha = 0.5)
elastic.cv.out <- cv.glmnet(x_train, y_train, family="binomial", alpha = 0.5)
bestlam <- elastic.cv.out$lambda.min
pred4 <- predict(model4, s=bestlam, newx=x_test, type="response")

tryCatch({
  # Update all confusion matrices in a single block
  cM1 = cM1 + table(pred1 > 0.5, data.test[, "acn_y"])
  cM2 = cM2 + table(pred2 > 0.5, data.test[, "acn_y"])
  cM3 = cM3 + table(pred3 > 0.5, data.test[, "acn_y"])
  cM4 = cM4 + table(pred4 > 0.5, data.test[, "acn_y"])
}, error = function(e) {

  # Log a warning and skip all updates for this iteration
  warning(sprintf("Skipping all cM updates for iteration %d due to error: %s", ii, e$message))
  })

}

# Compare accuracy->prec,recall,f1,misclassification of the models
accu1 = sum(diag(cM1))/sum(cM1)
accu2 = sum(diag(cM2))/sum(cM2)
accu3 = sum(diag(cM3))/sum(cM3)
accu4 = sum(diag(cM4))/sum(cM4)

# Compare precision
prec1 = cM1[1,1]/(cM1[1,1]+cM1[1,2])
prec2 = cM2[1,1]/(cM2[1,1]+cM2[1,2])
prec3 = cM3[1,1]/(cM3[1,1]+cM3[1,2])
prec4 = cM4[1,1]/(cM4[1,1]+cM4[1,2])

# Compare recall
recal1 = cM1[1,1]/(cM1[1,1]+cM1[2,1])
recal2 = cM2[1,1]/(cM2[1,1]+cM2[2,1])
recal3 = cM3[1,1]/(cM3[1,1]+cM3[2,1])
recal4 = cM4[1,1]/(cM4[1,1]+cM4[2,1])

# Compare f1
f1_1 = (2*prec1*recal1)/(prec1+recal1)
f1_2 = (2*prec2*recal2)/(prec2+recal2)
f1_3 = (2*prec3*recal3)/(prec3+recal3)
```

```
  f1_4 = (2*prec4*recal4)/(prec4+recal4)

  # Compare misclassification
  misc1 = 1 - accu1
  misc2 = 1 - accu2
  misc3 = 1 - accu3
  misc4 = 1 - accu4

  if (returntype == 1){
    return(c(misc1, misc2, misc3, misc4))
  }
  if (returntype == 2){
    return(c(prec1, prec2, prec3, prec4))
  }
  if (returntype == 3){
    return(c(recal1, recal2, recal3, recal4))
  }
  if (returntype == 4){
    return(c(f1_1, f1_2, f1_3, f1_4))
  }
}

set.seed(0)
nrep = 200
models <- c("Logistic Reg", "Ridge", "Lasso","Elastic")
modeln <- length(models)
```

```
idxmat.up = replicate(nrep, sample(1:nfold, sum(split==1), replace=T))
idxmat.down = replicate(nrep, sample(1:nfold, sum(split==0), replace=T))

# Need separate matrix for misclassification, prec, recall, f1
miscmat= matrix(NA, nrep, modeln)
precmat= matrix(NA, nrep, modeln)
recalmat= matrix(NA, nrep, modeln)
f1_mat= matrix(NA, nrep, modeln)

# add other models later
colnames(miscmat) = models
colnames(precmat) = models
colnames(recalmat) = models
colnames(f1_mat) = models

# for each repeat (each row -> each repeat)
for(jj in 1:nrep){
  cat("misc repeat:", jj)
  miscmat[jj, ]= cv.accu(nfold, data.c, data.nc, idx.c=idxmat.up[,jj],
 ↪  idx.nc=idxmat.down[,jj], 1)
}

for(jj in 1:nrep){
  cat("prec repeat:", jj)
  precmat[jj, ]= cv.accu(nfold, data.c, data.nc, idx.c=idxmat.up[,jj],
 ↪  idx.nc=idxmat.down[,jj], 2)}
```

```
for(jj in 1:nrep){
  cat("recal repeat:", jj)
  recalmat[jj, ]= cv.accu(nfold, data.c, data.nc, idx.c=idxmat.up[,jj],
 ↪  idx.nc=idxmat.down[,jj], 3)}

for(jj in 1:nrep){
  cat("f1 repeat:", jj)
  f1_mat[jj, ]= cv.accu(nfold, data.c, data.nc, idx.c=idxmat.up[,jj],
 ↪  idx.nc=idxmat.down[,jj], 4)}



library(ggpubr)
f1_mat2 <- as.data.frame(f1_mat)

#mean of each column
f1_log <- median(f1_mat2$`Logistic Reg`, na.rm = TRUE) #highest for 3fold
 ↪  200rep
median(f1_mat2$Ridge, na.rm = TRUE) #highest for 3fold 200rep
median(f1_mat2$Lasso, na.rm = TRUE)
median(f1_mat2$Elastic, na.rm = TRUE)

mean(f1_mat2$`Logistic Reg`, na.rm = TRUE) #highest for 3fold 200rep
mean(f1_mat2$Ridge, na.rm = TRUE)
mean(f1_mat2$Lasso, na.rm = TRUE)
mean(f1_mat2$Elastic, na.rm = TRUE)
```

```
# plot misclassification rate / precision/ recall / F1 score
plot1 <- boxplot(miscmat, ylab="Misclassification rate", col=c("#97bade"))
plot2 <- boxplot(precmat, ylab="Precision", col=c("#97bade"))
plot3 <- boxplot(recalmat, ylab="Recall", col=c("#97bade"))
# mark f1_log as red dotted line
plot4 <- boxplot(f1_mat, ylab="F1 score", col=c("#97bade")) #abline(h=f1_log,
 ↪  col="red", lty=2)

# Combine plots
#combined_plot <- ggarrange(plot1, plot2, plot3, plot4, ncol=2,
 ↪  nrow=2,common.legend = TRUE)

# Best Cutoff Value for ROC Curve
library(cutoff)
cutoff::roc(score=cleaned_data2$age_x,class=cleaned_data2$acn_y)

# using cutpointr
library(cutpointr)
cutpointr(data=cleaned_data2, x=age_x,class=acn_y)

cleaned_data4 <- cleaned_data %>% dplyr::select(-c(premalignant_y, report_id,
 ↪  year, age_x_group, service))

library(cutpointr)
library(fANCOVA)
```

```r
opt_cut <- cutpointr(data=cleaned_data4, x=age_x,class=acn_y,
 ↪  method=maximize_metric, metric=youden, boot_runs=1000)
plot(opt_cut)
plot_metric(opt_cut)
opt_cut

opt_cut2 <- cutpointr(data=cleaned_data4, x=age_x,class=crc_y,
 ↪  method=maximize_metric, metric=youden, boot_runs=1000)
plot(opt_cut2)
opt_cut2
```

References

Akimoto, Naohiko et al. 2021. "Rising Incidence of Early-Onset Colorectal Cancer – a Call to Action." *Nature Reviews Clinical Oncology* 18 (4): 230–43. https://doi.org/10.1038/s41571-020-00445-1.

Canadian Cancer Society. 2024. "Colorectal Cancer Statistics." https://cancer.ca/en/cancer-information/cancer-types/colorectal/statistics#:~:text=Colorectal%20cancer%20is%20expected%20to,until%20actual%20data%20become%20available.

Danpanichkul, Pojsakorn, Pinyada Moolkaew, Yatawee Kanjanakot, Natchaya Polpichai, Aunchalee Jaroenlapnopparat, Donghee Kim, Frank J. Lukens, et al. 2023. "Rising Incidence and Impact of Early-Onset Colorectal Cancer in the Asia-Pacific with Higher Mortality in Females from Southeast Asia: A Global Burden Analysis from 2010 to 2019." *Journal of Gastroenterology and Hepatology* 38 (12): 2053–60. https://doi.org/10.1111/jgh.16331.

Gharib, Ehsan, and Gilles A. Robichaud. 2024. "From Crypts to Cancer: A Holistic Perspective on Colorectal Carcinogenesis and Therapeutic Strategies." *International Journal of Molecular Sciences* 25 (17). https://doi.org/10.3390/ijms25179463.

Trivedi, Parth D et al. 2022. "Prevalence and Predictors of Young-Onset Colorectal Neoplasia: Insights from a Nationally Representative Colonoscopy Registry." *Gastroenterology* 162 (4): 1136–1146.e5. https://doi.org/10.1053/j.gastro.2021.12.285.

Yetman, Daniel. 2024. "Right-Sided Colon Cancer: Symptoms and Outlook." https://www.healthline.com/health/colorectal-cancer/colon-cancer-right-side-pain.