**Development of a machine learning model for predicting risk of recurrence of well-differentiated thyroid cancer – Summary Statistics R Studio**

Tracy Whitney Akinyi

2024-01-17

## Read Data

library(readr)

## Warning: package 'readr' was built under R version 4.2.3

Thyroid\_Diff <- read\_csv("Thyroid\_Diff.csv")

## Rows: 383 Columns: 17  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (16): Gender, Smoking, Hx Smoking, Hx Radiothreapy, Thyroid Function, Ph...  
## dbl (1): Age  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

View(Thyroid\_Diff)

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.2.3

## Warning: package 'ggplot2' was built under R version 4.2.3

## Warning: package 'tibble' was built under R version 4.2.3

## Warning: package 'dplyr' was built under R version 4.2.3

## Warning: package 'forcats' was built under R version 4.2.3

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ purrr 1.0.1  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.2 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.0 ✔ tidyr 1.3.0  
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the ]8;;http://conflicted.r-lib.org/conflicted package]8;; to force all conflicts to become errors

library(knitr)  
library(gtsummary)

## Warning: package 'gtsummary' was built under R version 4.2.3

## Age Category

Thyroid\_Diff <- Thyroid\_Diff %>%  
 mutate(  
 `Age Category` = dplyr::case\_when(  
 Age <= 45 ~ "<=45",  
 Age > 45 ~ ">45"  
 ),  
 `Age Category` = factor(`Age Category`, levels = c("<=45", ">45"))  
 )

## Generate Table of Summary

table1 <- Thyroid\_Diff %>%   
 select(Age,`Age Category`,Gender,Smoking,`Hx Smoking`,`Hx Radiothreapy`,`Thyroid Function`,`Physical Examination`,Adenopathy,Pathology,Focality,Risk,Tumor,Nodes,Metastasis,Stage,Response,Recurred) %>% tbl\_summary(by=Recurred,missing = "no" ,statistic = list(all\_continuous() ~ "{mean} ({sd})",  
 all\_categorical() ~ "{n} ({p}%)"),type = list(Age~"continuous",`Age Category`~"categorical",Gender~"categorical",Smoking~"categorical",`Hx Smoking`~"categorical",`Hx Radiothreapy`~"categorical",`Thyroid Function`~"categorical",`Physical Examination` ~"categorical",Adenopathy~"categorical",Pathology~"categorical",Focality~"categorical",Risk~"categorical",Tumor~"categorical",Nodes~"categorical",Metastasis~"categorical",Stage~"categorical",Response~"categorical",Recurred~"categorical" )) %>%   
 modify\_spanning\_header(c("stat\_1","stat\_2") ~ "\*\*Recurred\*\*") %>%   
 add\_p(test.args = all\_tests("fisher.test") ~ list(workspace=2e9)) %>%   
 as\_flex\_table()  
table1

|  | **Recurred** | |  |
| --- | --- | --- | --- |
| **Characteristic** | **No**, N = 2751 | **Yes**, N = 1081 | **p-value**2 |
| **Age** | 38 (13) | 47 (18) | <0.001 |
| **Age Category** |  |  | <0.001 |
| <=45 | 205 (75%) | 55 (51%) |  |
| >45 | 70 (25%) | 53 (49%) |  |
| **Gender** |  |  | <0.001 |
| F | 246 (89%) | 66 (61%) |  |
| M | 29 (11%) | 42 (39%) |  |
| **Smoking** |  |  | <0.001 |
| No | 259 (94%) | 75 (69%) |  |
| Yes | 16 (5.8%) | 33 (31%) |  |
| **Hx Smoking** |  |  | 0.008 |
| No | 261 (95%) | 94 (87%) |  |
| Yes | 14 (5.1%) | 14 (13%) |  |
| **Hx Radiothreapy** |  |  | 0.002 |
| No | 274 (100%) | 102 (94%) |  |
| Yes | 1 (0.4%) | 6 (5.6%) |  |
| **Thyroid Function** |  |  | 0.3 |
| Clinical Hyperthyroidism | 17 (6.2%) | 3 (2.8%) |  |
| Clinical Hypothyroidism | 10 (3.6%) | 2 (1.9%) |  |
| Euthyroid | 234 (85%) | 98 (91%) |  |
| Subclinical Hyperthyroidism | 5 (1.8%) | 0 (0%) |  |
| Subclinical Hypothyroidism | 9 (3.3%) | 5 (4.6%) |  |
| **Physical Examination** |  |  | 0.009 |
| Diffuse goiter | 7 (2.5%) | 0 (0%) |  |
| Multinodular goiter | 88 (32%) | 52 (48%) |  |
| Normal | 5 (1.8%) | 2 (1.9%) |  |
| Single nodular goiter-left | 63 (23%) | 26 (24%) |  |
| Single nodular goiter-right | 112 (41%) | 28 (26%) |  |
| **Adenopathy** |  |  | <0.001 |
| Bilateral | 5 (1.8%) | 27 (25%) |  |
| Extensive | 0 (0%) | 7 (6.5%) |  |
| Left | 5 (1.8%) | 12 (11%) |  |
| No | 247 (90%) | 30 (28%) |  |
| Posterior | 0 (0%) | 2 (1.9%) |  |
| Right | 18 (6.5%) | 30 (28%) |  |
| **Pathology** |  |  | <0.001 |
| Follicular | 16 (5.8%) | 12 (11%) |  |
| Hurthel cell | 14 (5.1%) | 6 (5.6%) |  |
| Micropapillary | 48 (17%) | 0 (0%) |  |
| Papillary | 197 (72%) | 90 (83%) |  |
| **Focality** |  |  | <0.001 |
| Multi-Focal | 66 (24%) | 70 (65%) |  |
| Uni-Focal | 209 (76%) | 38 (35%) |  |
| **Risk** |  |  | <0.001 |
| High | 0 (0%) | 32 (30%) |  |
| Intermediate | 38 (14%) | 64 (59%) |  |
| Low | 237 (86%) | 12 (11%) |  |
| **Tumor** |  |  | <0.001 |
| T1a | 48 (17%) | 1 (0.9%) |  |
| T1b | 38 (14%) | 5 (4.6%) |  |
| T2 | 131 (48%) | 20 (19%) |  |
| T3a | 55 (20%) | 41 (38%) |  |
| T3b | 2 (0.7%) | 14 (13%) |  |
| T4a | 1 (0.4%) | 19 (18%) |  |
| T4b | 0 (0%) | 8 (7.4%) |  |
| **Nodes** |  |  | <0.001 |
| N0 | 241 (88%) | 27 (25%) |  |
| N1a | 12 (4.4%) | 10 (9.3%) |  |
| N1b | 22 (8.0%) | 71 (66%) |  |
| **Metastasis** |  |  | <0.001 |
| M0 | 275 (100%) | 90 (83%) |  |
| M1 | 0 (0%) | 18 (17%) |  |
| **Stage** |  |  | <0.001 |
| I | 268 (97%) | 65 (60%) |  |
| II | 7 (2.5%) | 25 (23%) |  |
| III | 0 (0%) | 4 (3.7%) |  |
| IVA | 0 (0%) | 3 (2.8%) |  |
| IVB | 0 (0%) | 11 (10%) |  |
| **Response** |  |  | <0.001 |
| Biochemical Incomplete | 12 (4.4%) | 11 (10%) |  |
| Excellent | 207 (75%) | 1 (0.9%) |  |
| Indeterminate | 54 (20%) | 7 (6.5%) |  |
| Structural Incomplete | 2 (0.7%) | 89 (82%) |  |
| 1Mean (SD); n (%) | | | |
| 2Wilcoxon rank sum test; Pearson's Chi-squared test; Fisher's exact test | | | |