Assignment 3: Progression of Coronary Artery Calcification in Normal Subjects

Team 1: Milan Filipovic, Vivian Truong, Lillian Chen

April 23, 2021

Contents

Exercise

Complete a Biostatistics Consulting Worksheet for the following study. Your worksheet should include: * Statement of statistical hypothesis(es) * Assessment of assumptions used in any statistical inference * Analysis of data set. (You will need to retrieve from our website one of two data files: "cac.dta" and "cac.txt".) * Interpretation of results in terms of original research hypothesis so that a non-statistician would be able to explain results to colleagues.

Biostatistics Consulting Information

Client Information

Date: 04/23/2021

Client Information:

- Name: H.C. Yoon, A. Emerick, J Hill and J Goldin
- Department: Department of Diagnostic Imaging, Kaiser Moanalua Medical Center and the Department of Radiological Sciences, UCLA
- Phone: (310) 267-8785
- Name of principal investigator: H.C. Yoon, A. Emerick, J Hill and J Goldin from the Department of Diagnostic Imaging, Kaiser Moanalua Medical Center and the Department of Radiological Sciences, UCLA

Research Hypotheses

Purpose

- Statement of research hypothesis(es):
  - To test the hypothesis that the rate of CAC progression is gender specific being greater in men than in women.
- Statement of statistical hypothesis(es):
  - The rate of CAC progression in men is greater than it is in women in asymptomatic subjects
  - Null hypothesis: The rate of change in calcium volume score (CVS) between asymptomatic men and women is statistically significantly different

Background

Background & references:

- Recent reports indicate that electron-beam computed tomography (EBCT) can document the presence and monitor the progression of atherosclerotic CAC in the general adult population as well as in those with increased cardiovascular risk (1-3).
- EBCT can measure changes in the extent of CAC in adults treated with lipid-lowering agents.
  - EBCT is fast, sensitive, and uses an electron gun instead of x-rays to scan the chest.
  - Considered low risk and uses very low amounts of radiation. *Interpretation of the clinical significance of different coronary artery calcium scores in the same patient is dependent on several factors:
  - Measurement variation
  - Expected rate of progression of coronary calcium

Descriptions:

- Study design: Retrospective study
- Population(s): Adults who experience the progression of atherosclerotic CAC
- Sample(s): 217 asymptomatic subjects who underwent at least two electron-beam computed tomography (EBCT) for detection of CAC as a part of a clinical screening program.
  - Asymptomatic defined as no history of ischemic heart disease
  - No abnormal electrocardiogram, stress test, coronary angiogram, and no prior myocardial infarction or coronary bypass surgery
- Dependent variable(s):
  - vol1
  - vol2
  - days
- Independent variable(s):
  - sex
  - age

```
##       nid sex age vol1 vol2 days
##   1:    1   0  36  110  158  694
##   2:    2   0  37    0    0  691
##   3:    3   0  44    0    0  927
##   4:    4   0  44    0    0  719
##   5:    5   0  46    0    0 1545
##  ---
## 213: 213   0  55   34  137 1057
## 214: 214   1  70  397  548  641
## 215: 215   0  63    9   13 1012
## 216: 216   1  67  282  387  714
## 217: 217   1  67   63   82  943
```

Checking Data for cleanliness

Variable Descriptions

```
# Haven stores the stata labels in as variable attributes. We can access them
# using map_chr from the purrr package
(varlabels <- cac_tble %>% map_chr(~attributes(.)$label))
```

```
# Extracting variable data using base R functions
(varlabels <- str(lapply(cac_tble, attr, "label")))
```

The variable labels are as follows:

- nid : "id subject number, 1-217"
- sex : "rf male subject, 0n/1y"
- age : "rf age when first scanned, yrs"
- vol1: "ct visit #1 CVS, mm3"
- vol2: "ct visit #2 CVS, mm3"
- days: "ct elapsed days between visits"

Note about CAC score taken from radiopaedia: CAC score of 1-112: low risk with a relative risk ratio of 1.9 (95 CI: 1.3-2.8) CAC score of 100-400: moderate risk with a relative risk ratio of 4.3 (95% CI:3.1-6.1) CAC score of 401-999: high risk with a relative risk ratio of 7.2 (95% CI:5.2-9.9) CAC score $>$ 1000 is considered very high risk with a relative risk ratio of 10.8% Source: link)

Statistical Analysis

Preliminary Analysis

```
cac_tble %>% mutate(gender = ifelse(cac_tble$sex == 1, "Men","Women"),
                    vol_change = vol2-vol1,
                    ratechange = vol_change/days) %>%
  select(-sex, -nid) %>%
  tbl_summary(by = gender, missing_text = "Missing",
              label = list(age ~ "Age First Scanned",
                           vol1 ~ "CVS at First Visit",
                           vol2 ~ "CVS at Second Visit",
```

```
                           days ~ "Days Between Visits",
                           vol_change ~ "Change in CVS",
                           ratechange ~ "Rate of Change in CVS")) %>%
  add_p(test = list(all_continuous() ~ "t.test")) %>%
  bold_labels() %>%
  bold_p() %>%
  add_overall() %>%
  modify_header(label ~ "**Participant Characteristics**")
```

```
## Table printed with `knitr::kable()`, not {gt}. Learn why at
## http://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include `message = FALSE` in code chunk header.
```

Participant Characteristics	**Overall**, N = 217	**Men**, N = 114	**Women**, N = 103	**p-value**
Age First Scanned	57 (50, 65)	54 (46, 64)	61 (52, 66)	**0.001**
CVS at First Visit	21 (0, 175)	96 (8, 252)	0 (0, 42)	**0.004**
CVS at Second Visit	40 (0, 223)	117 (7, 366)	4 (0, 85)	**0.004**
Days Between Visits	701 (475, 992)	700 (471, 983)	705 (512, 1,011)	0.4
Change in CVS	5 (0, 49)	14 (0, 75)	0 (0, 17)	**0.025**
Rate of Change in CVS	0.01 (0.00, 0.07)	0.02 (0.00, 0.11)	0.00 (0.00, 0.02)	0.10

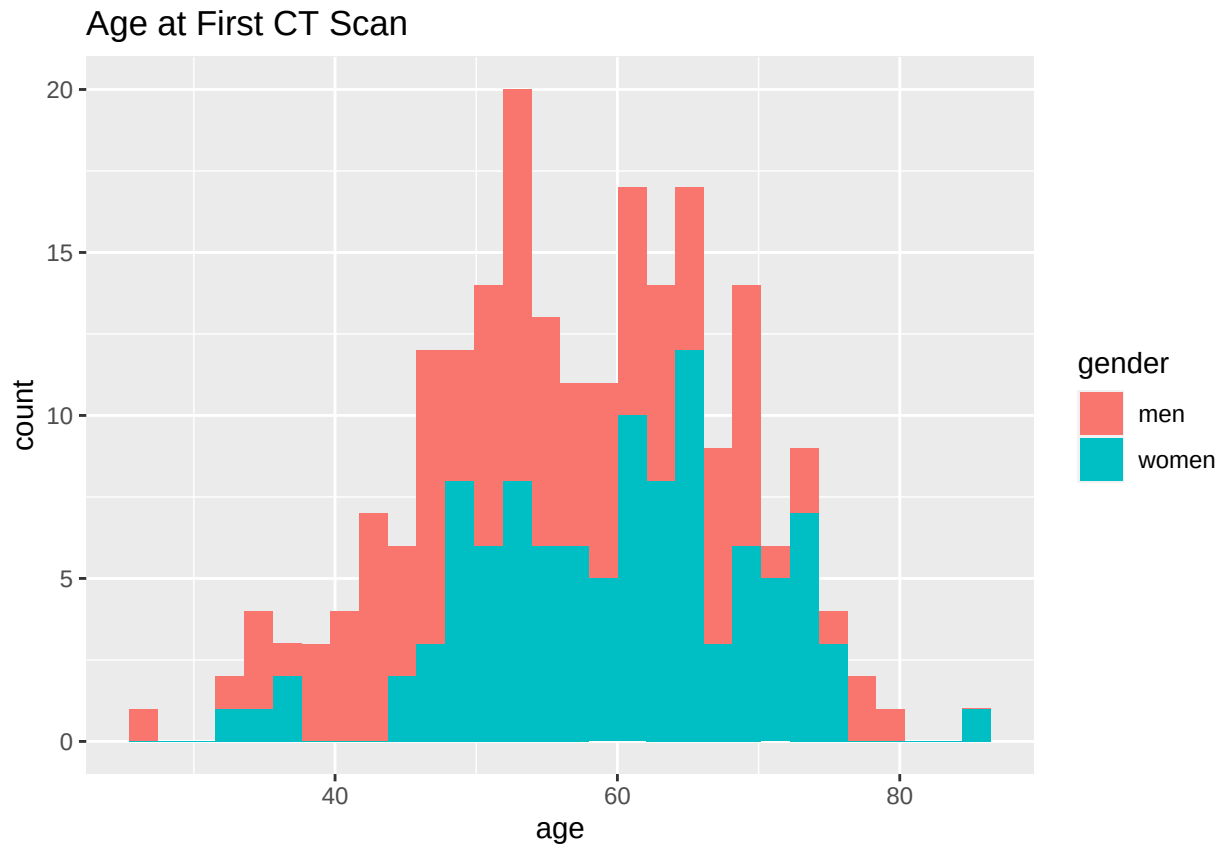
```
cac_tble <- cac_tble %>%
  mutate(gender = ifelse(cac_tble$sex == 1, "men","women"),
         vol_change = vol2-vol1,
         ratechange = vol_change / days)

summary(cac_tble)
```

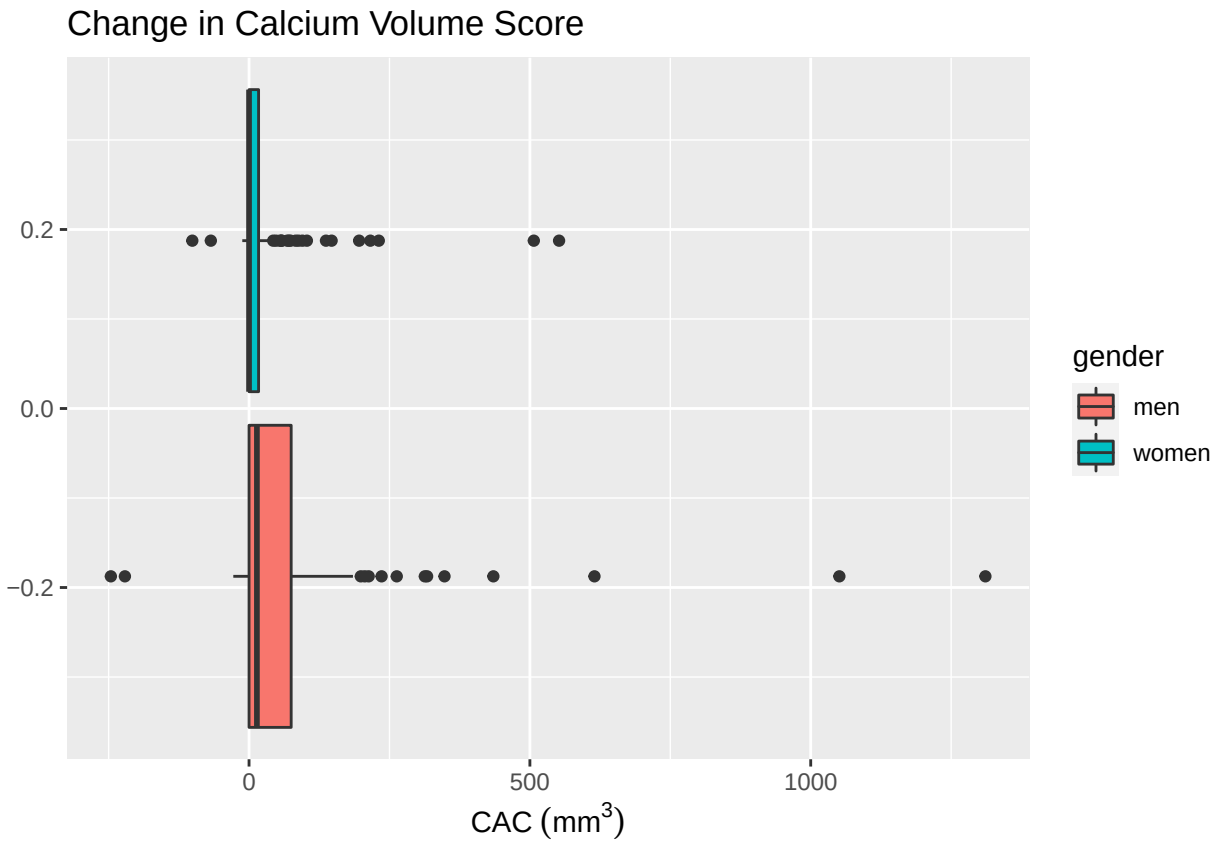
```
##       nid           sex              age             vol1
##  Min.   :  1   Min.   :0.0000   Min.   :26.00   Min.   :   0.0
##  1st Qu.: 55   1st Qu.:0.0000   1st Qu.:50.00   1st Qu.:   0.0
##  Median :109   Median :1.0000   Median :57.00   Median :  21.0
##  Mean   :109   Mean   :0.5253   Mean   :57.16   Mean   : 175.9
##  3rd Qu.:163   3rd Qu.:1.0000   3rd Qu.:65.00   3rd Qu.: 175.0
##  Max.   :217   Max.   :1.0000   Max.   :85.00   Max.   :3138.0
##       vol2             days          gender            vol_change
##  Min.   :   0.0   Min.   : 245.0   Length:217         Min.   :-246.00
##  1st Qu.:   0.0   1st Qu.: 475.0   Class :character   1st Qu.:   0.00
##  Median :  40.0   Median : 701.0   Mode  :character   Median :   5.00
##  Mean   : 227.4   Mean   : 761.5                      Mean   :  51.57
##  3rd Qu.: 223.0   3rd Qu.: 992.0                      3rd Qu.:  49.00
##  Max.   :4291.0   Max.   :1941.0                      Max.   :1311.00
##    ratechange
##  Min.   :-0.72566
##  1st Qu.: 0.00000
##  Median : 0.00675
##  Mean   : 0.07847
##  3rd Qu.: 0.07407
##  Max.   : 3.12798
```

```
cac_tble %>% ggplot(mapping = aes(x = age)) +
  geom_histogram(mapping = aes(x = age, fill = gender)) +
  labs(title = "Age at First CT Scan")
```

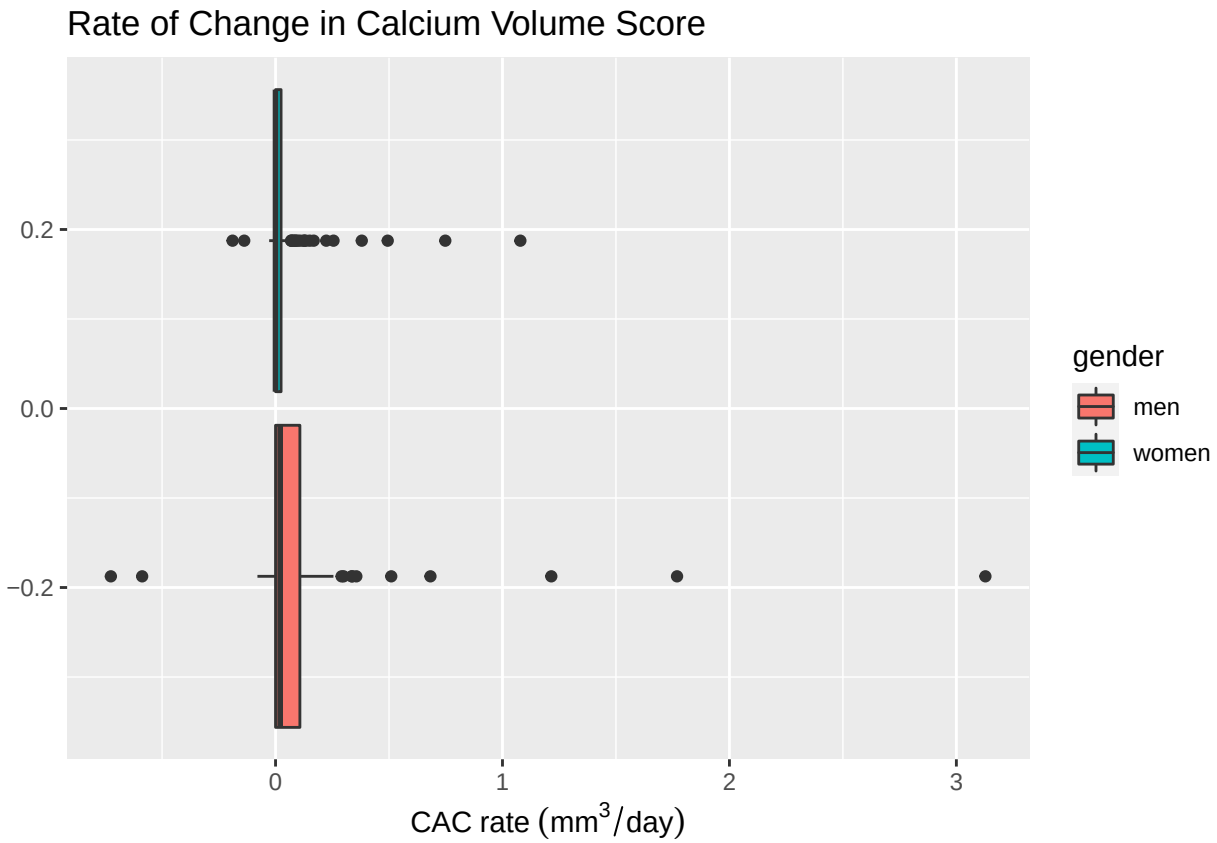
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
cac_tble %>% ggplot() +
  geom_boxplot(mapping = aes(x = vol_change, fill = gender)) +
  labs(x = expression(CAC~(mm^3)), title = "Change in Calcium Volume Score")
```



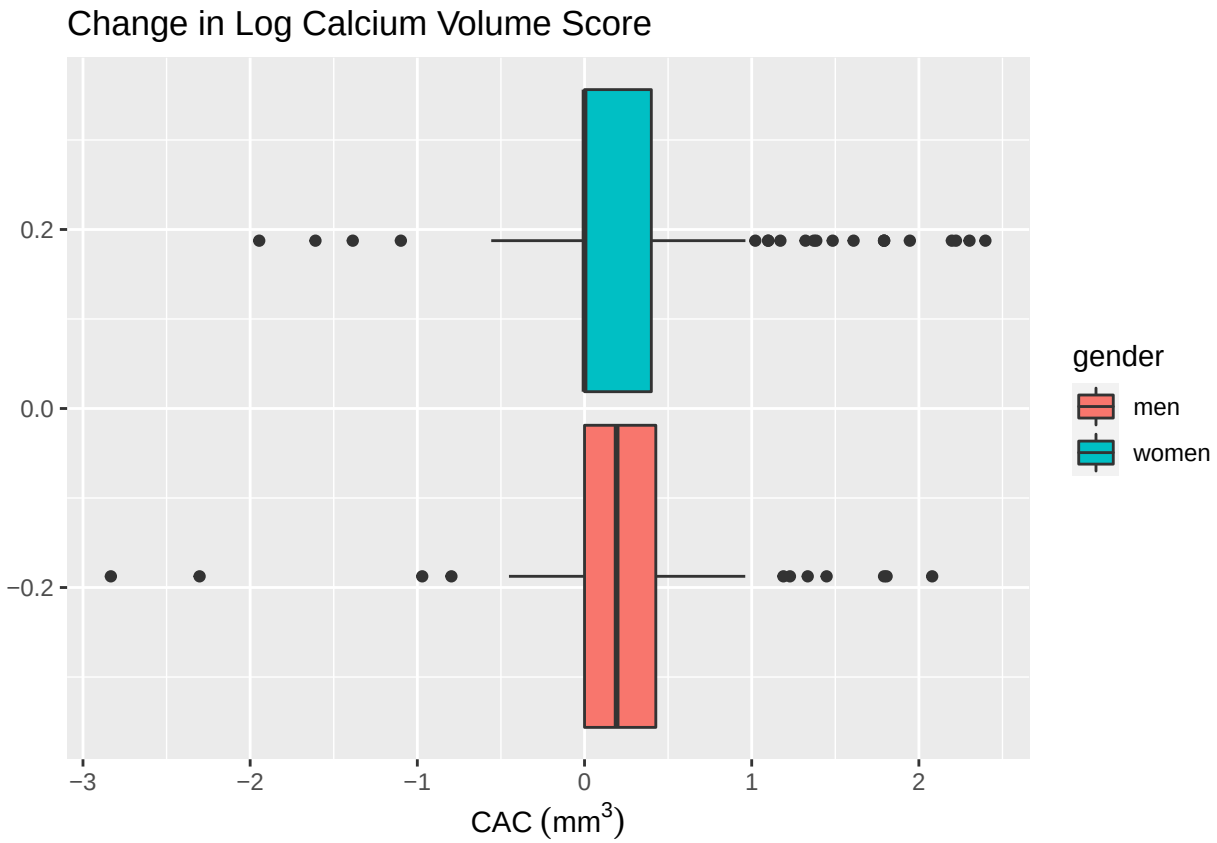

```
cac_tble %>% ggplot() +
  geom_boxplot(mapping = aes(x = ratechange, fill = gender)) +
  labs(x = expression(CAC~rate~(mm^3/day)), title = "Rate of Change in Calcium Volume Score")
```



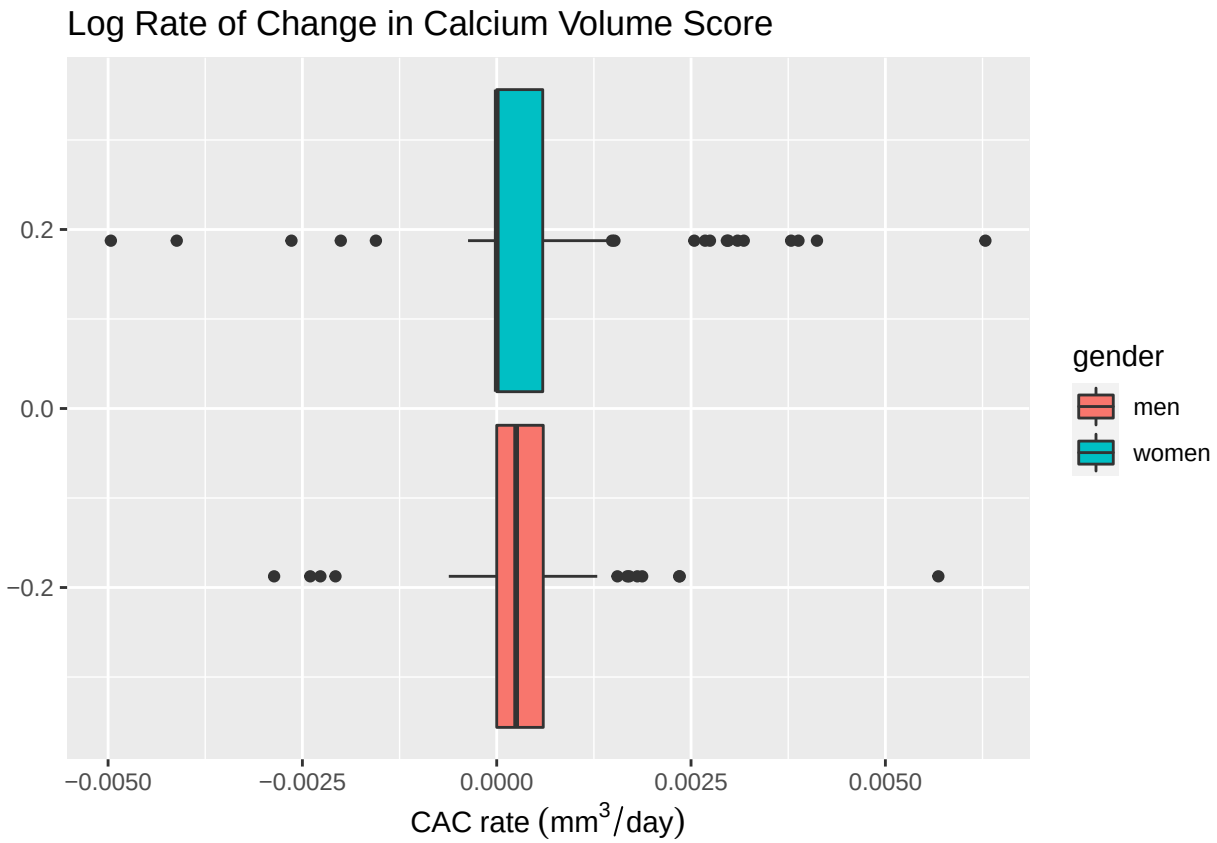
Age appears to be normally distributed, volume change appears to be right skewed, and rate of volume change is also right skewed and overdispersion at ratechange = 0. Some transformation should be applied to normalize the data, and outliers should be examined for validity. Additionally, the values for volume change and rate of volume change have a large number of 0s, so our data is zero-inflated and should be noted so that log transformations include an offset.

```
cac_tble <- cac_tble %>%
  mutate(gender = ifelse(cac_tble$sex == 1, "men","women"),
         vol_change = vol2-vol1,
         ratechange = vol_change / days,
         logvol1 = log(vol1 + 1),
         logvol2 = log(vol2 + 1),
         logdiffvol_change = logvol2 - logvol1,
         lograteschange = (logvol2-logvol1)/days)

cac_tble %>% ggplot() +
  geom_boxplot(mapping = aes(x = logdiffvol_change, fill = gender)) +
  labs(x = expression(CAC~(mm^3)), title = "Change in Log Calcium Volume Score")
```

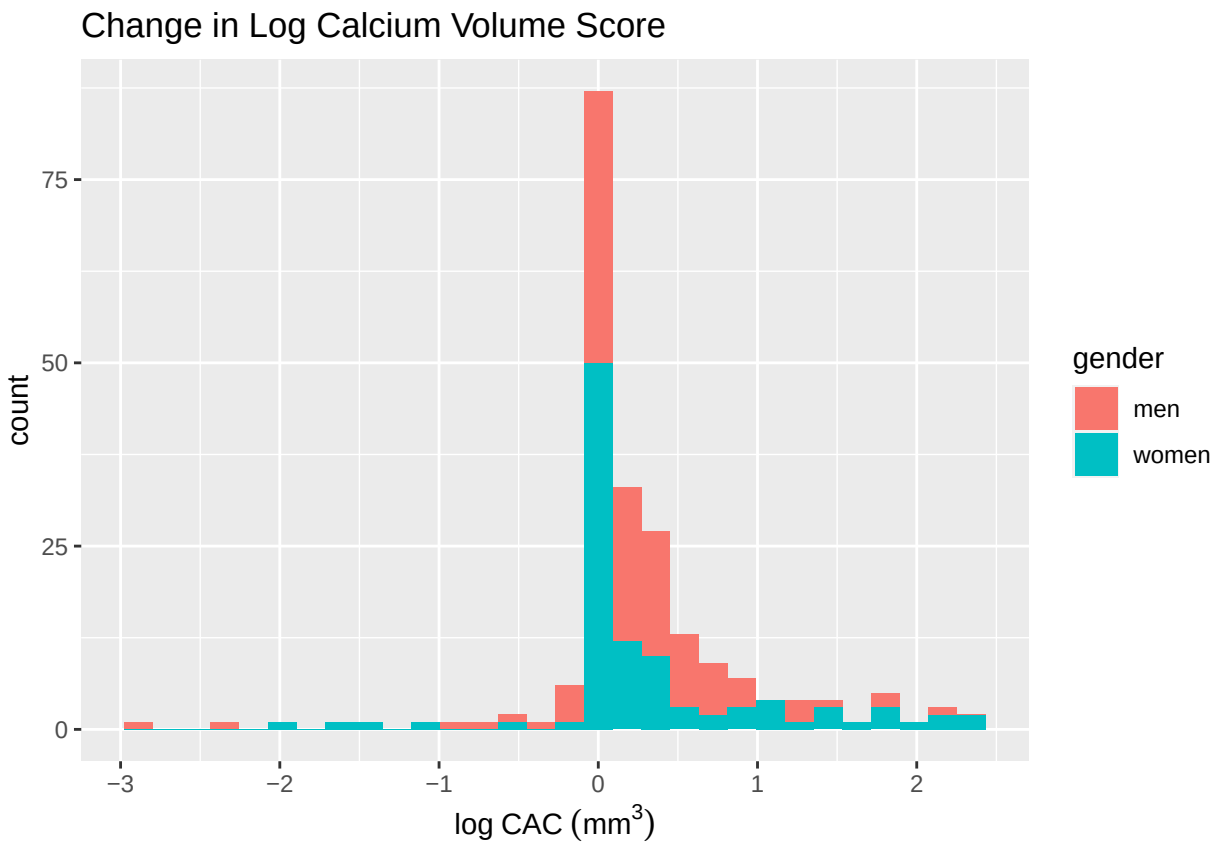



```
cac_tble %>% ggplot() +
  geom_boxplot(mapping = aes(x = logratechange, fill = gender)) +
  labs(x = expression(CAC~rate~(mm^3/day)), title = "Log Rate of Change in Calcium Volume Score")
```



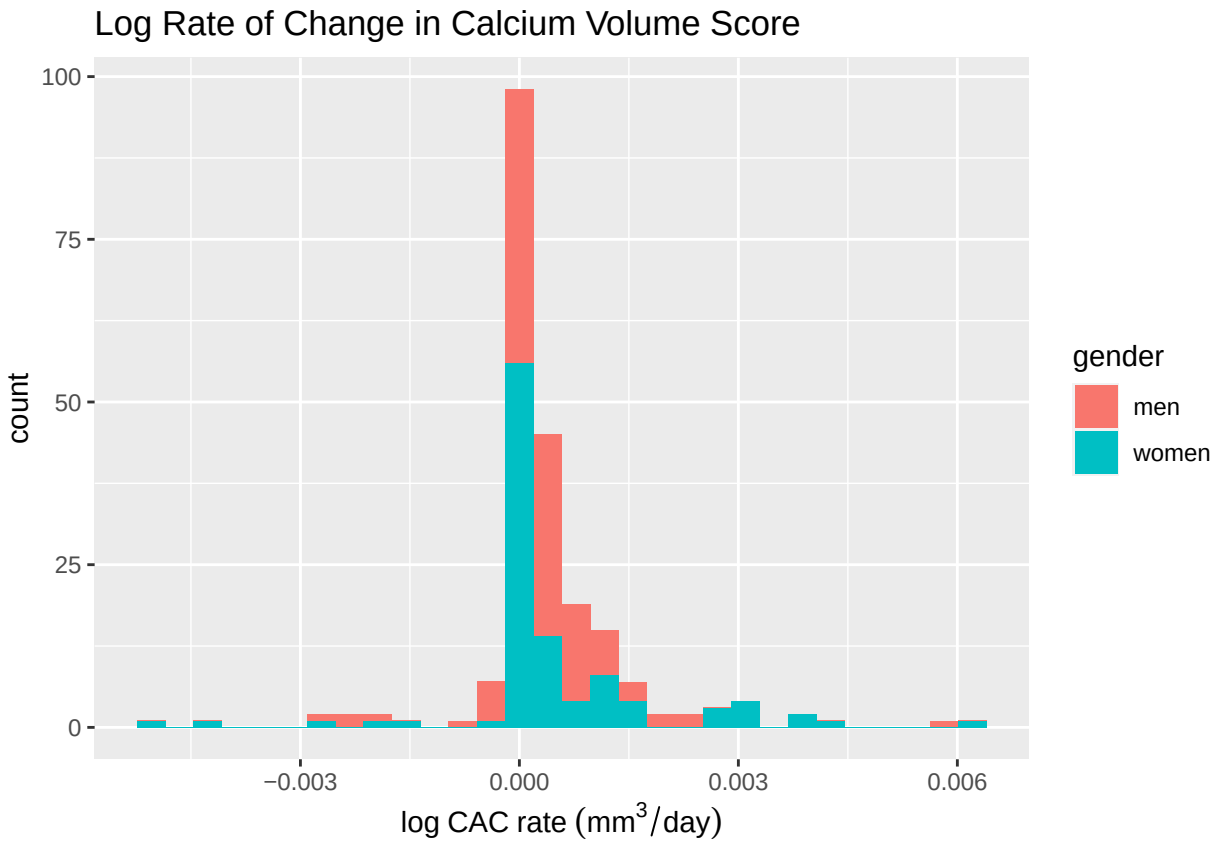

```
cac_tble %>% ggplot() +
  geom_histogram(mapping = aes(x = logdiffvol_change, fill = gender)) +
  labs(x = expression(log~CAC~(mm^3)), title = "Change in Log Calcium Volume Score")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
cac_tble %>% ggplot() +
  geom_histogram(mapping = aes(x = logratechange, fill = gender)) +
  labs(x = expression(log~CAC~rate~(mm^3/day)), title = "Log Rate of Change in Calcium Volume Score")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The data is still extremely zero-inflated (containing a lot of zeros) and non-normal. It seems like the best course of action is not a linear regression, since despite transformations the distribution of the outcome is not normally distributed. Since there is no transformation that can be applied for us to meet normality assumptions of parametric tests, a possible route to take is to do a nonparametric test on the rate of volume change between men and women.

Mann-Whitney U Test

The Mann-Whitney U Test may be able to test the null hypothesis H_0: There is no difference in the rate of volume change between men and women in the study population. The idea of the Mann-Whitney U Test is that we treat men and women as independent samples containing information on the rate of volume change, and we are comparing them on the rate of volume change. We use this test when the data is not normally distributed, which appears to be the case here after our data visualization conducted above.

```
cac_men <- cac_tble %>%
  filter(gender == "men")
cac_women <- cac_tble %>%
  filter(gender == "women")

# Mann-Whitney U test - the statements below yield the same result
# but are different ways of writing it
wilcox.test(cac_men$ratechange, cac_women$ratechange, paired = F)
```

```
##
##  Wilcoxon rank sum test with continuity correction
```

```
##
## data:  cac_men$ratechange and cac_women$ratechange
## W = 7092, p-value = 0.007229
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(cac_tble$ratechange~cac_tble$gender)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  cac_tble$ratechange by cac_tble$gender
## W = 7092, p-value = 0.007229
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(cac_men$logratechange, cac_women$logratechange, paired = F)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  cac_men$logratechange and cac_women$logratechange
## W = 6165, p-value = 0.5183
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(cac_tble$logratechange~cac_tble$gender)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  cac_tble$logratechange by cac_tble$gender
## W = 6165, p-value = 0.5183
## alternative hypothesis: true location shift is not equal to 0
```

```
median(cac_men$ratechange)
```

```
## [1] 0.02027831
```

```
median(cac_women$ratechange)
```

```
## [1] 0
```

```
exp(median(cac_men$logratechange))
```

```
## [1] 1.00025
```

```
exp(median(cac_women$logratechange))
```

```
## [1] 1
```

The Mann-Whitney U test showed that there was a significant difference (W = 7092, p = .007) in the rate of volume change between men and women enrolled in the study. The median rate of volume change for men was 0.020 mm^3/day, and the median rate of volume change for women was 0 mm^3/day.

However, this test includes the 24 individuals that had a negative rate (decrease in CVS score), which means this test may not adequately answer the researchers' questions since the researchers are asking about the increase in rate of change, and researchers may choose to exclude individuals that did not experience an increase.

Concluding Remarks

Limitations:

The dataset is limited and does not include additional risk factors for CAC (smoking, diet, etc). Additionally, there is only one follow-up measurement conducted for CVS. Potential confounding from the purpose of the CT scan may also affect the measured CVS, since patients were receiving CT scans for unrelated purposes and we are not sure of the definition of "unrelated", since there could be risk factors that the patient had that were not available in this dataset.

Consulting Recommendations:

When consulting the researchers of this study, validity of outliers should be addressed. In addressing non-normality and zero inflation, researchers may want to consider selecting a different pool of participants who may have CAC growth. Alternatively, scores can be converted into a categorically based risk system such as the existing defined classification system for CAC scores (see classification mentioned above under 'Background').

```
sessionInfo()
```

```
## R version 4.0.4 (2021-02-15)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19042)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] faraway_1.0.7     sjlabelled_1.1.7  gtsummary_1.3.7   data.table_1.14.0
##  [5] forcats_0.5.1     stringr_1.4.0     dplyr_1.0.5       purrr_0.3.4
##  [9] readr_1.4.0       tidyr_1.1.3       tibble_3.1.0      ggplot2_3.3.3
## [13] tidyverse_1.3.0   haven_2.3.1
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.6          lubridate_1.7.10    lattice_0.20-41
##  [4] assertthat_0.2.1    digest_0.6.27       utf8_1.2.1
```

```
##  [7] R6_2.5.0           cellranger_1.1.0    backports_1.2.1
## [10] reprex_2.0.0       evaluate_0.14       highr_0.8
## [13] httr_1.4.2         pillar_1.6.0        rlang_0.4.10
## [16] readxl_1.3.1       minqa_1.2.4         rstudioapi_0.13
## [19] nloptr_1.2.2.2     Matrix_1.3-2        rmarkdown_2.7
## [22] labeling_0.4.2     splines_4.0.4       statmod_1.4.35
## [25] lme4_1.1-26        munsell_0.5.0       broom_0.7.6
## [28] compiler_4.0.4     modelr_0.1.8        xfun_0.22
## [31] pkgconfig_2.0.3    htmltools_0.5.1.1   insight_0.13.2
## [34] tidyselect_1.1.0   fansi_0.4.2         crayon_1.4.1
## [37] dbplyr_2.1.1       withr_2.4.1         MASS_7.3-53.1
## [40] grid_4.0.4         nlme_3.1-152        jsonlite_1.7.2
## [43] gtable_0.3.0       lifecycle_1.0.0     DBI_1.1.1
## [46] magrittr_2.0.1     scales_1.1.1        cli_2.4.0
## [49] stringi_1.5.3      farver_2.1.0        broom.helpers_1.3.0
## [52] fs_1.5.0           xml2_1.3.2          ellipsis_0.3.1
## [55] generics_0.1.0     vctrs_0.3.6         boot_1.3-26
## [58] tools_4.0.4        glue_1.4.2          hms_1.0.0
## [61] survival_3.2-7     yaml_2.2.1          colorspace_2.0-0
## [64] gt_0.2.2           rvest_1.0.0         knitr_1.31
## [67] usethis_2.0.1
```