Analyzing Medical Data to Investigate Causes of Cervical Cancer

Hank Magan

09/21/2021

Abstract

In this report, a variety of causes of cervical cancer will be examined in detail to observe their respective relationships with the disease. Different methods of diagnosis will be analyzed to determine relative effectiveness within the selected group of women. The dataset to be examined comprises a variety of information including demographics and medical records of 858 women, with ages ranging from 13-84, taken from the University of California Irvine Machine Learning Repository. In this report, multiple conclusions will be formed regarding factors which affect cervical cancer. The primary conclusion is that HPV is by far the most important and telling factor when considering cervical cancer, followed by a multitude of other STD's. This report will demonstrate these causes and support these conclusions using a variety of data modeling techniques.

1. Introduction

Cervical cancer is a type of cancer originating in the cervix. Symptoms are typically undetectable early in its early stages, however may develop into abnormal vaginal bleeding, pelvic pain, or pain during sexual intercourse. Due to the asymptomatic nature of the disease, it can be difficult to catch it in its early stages, which is unfortunately one of the best ways to combat cancer. This could be a contributor to why cervical cancer is so deadly. It is the second-most common cause of cancer in women, accounting for about 5% of total cancer cases and deaths in women. With that said, there are things that women may avoid to reduce the risk of developing cervical cancer; the main ways are to get screened regularly and to practice safe sex.

There are many different causes of cervical cancer, however *human papillomavirus*, or HPV, is by far the most common cause of death. There are dozens of types of HPV; types 16 and 18 account for 75% of cervical cancers globally, atlthough other types can still carry risk. Additionally, having multiple sexual partners, or having a sexual partner who themselves have multiple sexual partners, puts one at greater risk. Cervical cancer is also one of the many risks that come with smoking cigarettes; among HPV-infected women, current and former smokers have around two to three times the chance to develop cancer. Other common causes include but are not limited to: oral contraceptives, multiple pregnancies, and HIV.

In this report, multiple different STD's will be investigated to observe their correlation to cervical cancer. As previously discussed, HPV, is one such STD, but there are others. The variables seen below will be examined to observe their correlations with cervical cancer. Note that some of the data in this dataset is simply demographical, or may not be used when analyzing.

1. Age
2. Number of sexual partners
3. First sexual intercourse (age)
4. Num of pregnancies
5. Smoker

6. Smokes (years)
7. Packs smoked per year
8. Hormonal contraceptive user
9. Contraceptives (years)
10. Intrauterine device (IUD) user
11. IUD (years)
12. STDs
13. Num of STDs
14. Condylomatosis
15. Cervical condylomatosis
16. Vaginal condylomatosis
17. Vulvo-perineal condylomatosis
18. Syphilis
19. Pelvic inflammatory disease (PID)
20. Gential herpes
21. Molluscum contagiosum
22. AIDS
23. HIV
24. Hepatitis B
25. HPV
26. Num of diagnosis
27. Time since first diagnosis
28. Time since last diagnosis
29. Cancer diagnosis
30. Cervical intraepithelial neoplasia (CIN) diagnosis
31. HPV diagnosis
32. Cervical cancer diagnosis
33. Hinselmann
34. Schiller
35. Cytology
36. Biopsy

The last four variables are methods for detecting cervical cancer. Biopsy is a very common method of dianosing cervical cancer. A biopsy is performed via a colpscopy, or a magnified inspection of the cervix. Doctors perform this test and report results visually. This is usually a good test to identify cervical cancer, however severity is tough to pinpoint with pure visuals. A majority of this report will investigate the discussed causes of cervical cancer as they apply to this specific dataset, and is structured like so:

1. Introduction
2. Data Cleansing
3. Data Transformation
4. Human Papillomavirus (HPV) and Cervical Cancer
5. Investigating Other Common Causes
6. Detecting Cervical Cancer
7. References

2. Data Cleansing

Having clean data is necessary for performing any form of meaningful data analysis. The term "data cleansing" describes a wide variety of particular processes, all of which work together to transform incomplete, inaccurate, or improperly formatted data into data that is fully complete, making the dataset significantly easier to work with. Essentially, data cleansing prepares data for proper analysis and examination. As such, it is necessary before performing any sort of work on a dataset. To determine if the cervical cancer data is clean, one must examine the structure of the data, along with any other descriptive statistics. R has many commands which are useful for such a task, such as `str`, `summary`, `head`, and `tail`, as demonstrated in the code below.

```
ccData <- read.csv("cervicalCA.csv", header=TRUE)

# examine the structure of the data
str(ccData)
```

```
## 'data.frame':    858 obs. of  36 variables:
##  $ Age                                 : int  18 15 34 52 46 42 51 26 45 44 ...
##  $ Number.of.sexual.partners           : chr  "4" "1" "1" "5" ...
##  $ First.sexual.intercourse            : chr  "15" "14" "?" "16" ...
##  $ Num.of.pregnancies                  : chr  "1" "1" "1" "4" ...
##  $ Smokes                              : chr  "0" "0" "0" "1" ...
##  $ Smokes..years.                      : chr  "0" "0" "0" "37" ...
##  $ Smokes..packs.year.                 : chr  "0" "0" "0" "37" ...
##  $ Hormonal.Contraceptives             : chr  "0" "0" "0" "1" ...
##  $ Hormonal.Contraceptives..years.     : chr  "0" "0" "0" "3" ...
##  $ IUD                                 : chr  "0" "0" "0" "0" ...
##  $ IUD..years.                         : chr  "0" "0" "0" "0" ...
##  $ STDs                                : chr  "0" "0" "0" "0" ...
##  $ STDs..number.                       : chr  "0" "0" "0" "0" ...
##  $ STDs.condylomatosis                 : chr  "0" "0" "0" "0" ...
##  $ STDs.cervical.condylomatosis        : chr  "0" "0" "0" "0" ...
##  $ STDs.vaginal.condylomatosis         : chr  "0" "0" "0" "0" ...
##  $ STDs.vulvo.perineal.condylomatosis: chr  "0" "0" "0" "0" ...
```

```
##  $ STDs.syphilis                     : chr  "0" "0" "0" "0" ...
##  $ STDs.pelvic.inflammatory.disease  : chr  "0" "0" "0" "0" ...
##  $ STDs.genital.herpes               : chr  "0" "0" "0" "0" ...
##  $ STDs.molluscum.contagiosum        : chr  "0" "0" "0" "0" ...
##  $ STDs.AIDS                         : chr  "0" "0" "0" "0" ...
##  $ STDs.HIV                          : chr  "0" "0" "0" "0" ...
##  $ STDs.Hepatitis.B                  : chr  "0" "0" "0" "0" ...
##  $ STDs.HPV                          : chr  "0" "0" "0" "0" ...
##  $ STDs..Number.of.diagnosis         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ STDs..Time.since.first.diagnosis  : chr  "?" "?" "?" "?" ...
##  $ STDs..Time.since.last.diagnosis   : chr  "?" "?" "?" "?" ...
##  $ Dx.Cancer                         : int  0 0 0 1 0 0 0 0 1 0 ...
##  $ Dx.CIN                            : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Dx.HPV                            : int  0 0 0 1 0 0 0 0 1 0 ...
##  $ Dx                                : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ Hinselmann                        : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ Schiller                          : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ Citology                          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Biopsy                            : int  0 0 0 0 0 0 1 0 0 0 ...
```

```
# examine the basic descriptive statistics
summary(ccData)
```

```
##       Age        Number.of.sexual.partners First.sexual.intercourse
##  Min.   :13.00   Length:858                Length:858
##  1st Qu.:20.00   Class :character          Class :character
##  Median :25.00   Mode  :character          Mode  :character
##  Mean   :26.82
##  3rd Qu.:32.00
##  Max.   :84.00
##  Num.of.pregnancies    Smokes          Smokes..years.     Smokes..packs.year.
##  Length:858         Length:858         Length:858         Length:858
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  Hormonal.Contraceptives Hormonal.Contraceptives..years.     IUD
##  Length:858              Length:858                      Length:858
##  Class :character        Class :character                Class :character
##  Mode  :character        Mode  :character                Mode  :character
##
##
##
##  IUD..years.            STDs           STDs..number.      STDs.condylomatosis
##  Length:858         Length:858         Length:858         Length:858
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  STDs.cervical.condylomatosis STDs.vaginal.condylomatosis
##  Length:858                   Length:858
##  Class :character             Class :character
```

```
##  Mode  :character             Mode  :character
##
##
##
##  STDs.vulvo.perineal.condylomatosis STDs.syphilis
##  Length:858                         Length:858
##  Class :character                   Class :character
##  Mode  :character                   Mode  :character
##
##
##
##  STDs.pelvic.inflammatory.disease STDs.genital.herpes
##  Length:858                       Length:858
##  Class :character                 Class :character
##  Mode  :character                 Mode  :character
##
##
##
##  STDs.molluscum.contagiosum  STDs.AIDS           STDs.HIV
##  Length:858                  Length:858         Length:858
##  Class :character            Class :character   Class :character
##  Mode  :character            Mode  :character   Mode  :character
##
##
##
##  STDs.Hepatitis.B     STDs.HPV         STDs..Number.of.diagnosis
##  Length:858         Length:858         Min.   :0.00000
##  Class :character   Class :character   1st Qu.:0.00000
##  Mode  :character   Mode  :character   Median :0.00000
##                                        Mean   :0.08741
##                                        3rd Qu.:0.00000
##                                        Max.   :3.00000
##  STDs..Time.since.first.diagnosis STDs..Time.since.last.diagnosis
##  Length:858                       Length:858
##  Class :character                 Class :character
##  Mode  :character                 Mode  :character
##
##
##
##    Dx.Cancer           Dx.CIN             Dx.HPV              Dx
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.00000   Median :0.00000   Median :0.00000   Median :0.00000
##  Mean   :0.02098   Mean   :0.01049   Mean   :0.02098   Mean   :0.02797
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##    Hinselmann         Schiller          Citology           Biopsy
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000
##  Median :0.00000   Median :0.00000   Median :0.00000   Median :0.0000
##  Mean   :0.04079   Mean   :0.08625   Mean   :0.05128   Mean   :0.0641
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.0000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.0000
```

```
# examine the first few rows
head(ccData)
```

```
##   Age Number.of.sexual.partners First.sexual.intercourse Num.of.pregnancies
## 1  18                         4                       15                   1
## 2  15                         1                       14                   1
## 3  34                         1                        ?                   1
## 4  52                         5                       16                   4
## 5  46                         3                       21                   4
## 6  42                         3                       23                   2
##   Smokes Smokes..years. Smokes..packs.year. Hormonal.Contraceptives
## 1      0              0                   0                       0
## 2      0              0                   0                       0
## 3      0              0                   0                       0
## 4      1             37                  37                       1
## 5      0              0                   0                       1
## 6      0              0                   0                       0
##   Hormonal.Contraceptives..years. IUD IUD..years. STDs STDs..number.
## 1                               0   0           0    0             0
## 2                               0   0           0    0             0
## 3                               0   0           0    0             0
## 4                               3   0           0    0             0
## 5                              15   0           0    0             0
## 6                               0   0           0    0             0
##   STDs.condylomatosis STDs.cervical.condylomatosis STDs.vaginal.condylomatosis
## 1                   0                            0                          0
## 2                   0                            0                          0
## 3                   0                            0                          0
## 4                   0                            0                          0
## 5                   0                            0                          0
## 6                   0                            0                          0
##   STDs.vulvo.perineal.condylomatosis STDs.syphilis
## 1                                  0            0
## 2                                  0            0
## 3                                  0            0
## 4                                  0            0
## 5                                  0            0
## 6                                  0            0
##   STDs.pelvic.inflammatory.disease STDs.genital.herpes
## 1                                0                   0
## 2                                0                   0
## 3                                0                   0
## 4                                0                   0
## 5                                0                   0
## 6                                0                   0
##   STDs.molluscum.contagiosum STDs.AIDS STDs.HIV STDs.Hepatitis.B STDs.HPV
## 1                          0         0       0                0        0
## 2                          0         0       0                0        0
## 3                          0         0       0                0        0
## 4                          0         0       0                0        0
## 5                          0         0       0                0        0
## 6                          0         0       0                0        0
##   STDs..Number.of.diagnosis STDs..Time.since.first.diagnosis
```

```
## 1                          0                                ?
## 2                          0                                ?
## 3                          0                                ?
## 4                          0                                ?
## 5                          0                                ?
## 6                          0                                ?
##   STDs..Time.since.last.diagnosis Dx.Cancer Dx.CIN Dx.HPV Dx Hinselmann
## 1                               ?         0      0      0  0          0
## 2                               ?         0      0      0  0          0
## 3                               ?         0      0      0  0          0
## 4                               ?         1      0      1  0          0
## 5                               ?         0      0      0  0          0
## 6                               ?         0      0      0  0          0
##   Schiller Citology Biopsy
## 1        0        0      0
## 2        0        0      0
## 3        0        0      0
## 4        0        0      0
## 5        0        0      0
## 6        0        0      0
```

```
# examine the last few rows
tail(ccData)
```

```
##     Age Number.of.sexual.partners First.sexual.intercourse Num.of.pregnancies
## 853  43                         3                       17                   3
## 854  34                         3                       18                   0
## 855  32                         2                       19                   1
## 856  25                         2                       17                   0
## 857  33                         2                       24                   2
## 858  29                         2                       20                   1
##     Smokes Smokes..years. Smokes..packs.year. Hormonal.Contraceptives
## 853      0              0                   0                       1
## 854      0              0                   0                       0
## 855      0              0                   0                       1
## 856      0              0                   0                       1
## 857      0              0                   0                       1
## 858      0              0                   0                       1
##     Hormonal.Contraceptives..years. IUD IUD..years. STDs STDs..number.
## 853                               5   0           0    0             0
## 854                               0   0           0    0             0
## 855                               8   0           0    0             0
## 856                            0.08   0           0    0             0
## 857                            0.08   0           0    0             0
## 858                             0.5   0           0    0             0
##     STDs.condylomatosis STDs.cervical.condylomatosis
## 853                   0                           0
## 854                   0                           0
## 855                   0                           0
## 856                   0                           0
## 857                   0                           0
## 858                   0                           0
##     STDs.vaginal.condylomatosis STDs.vulvo.perineal.condylomatosis
## 853                           0                                 0
```

```
## 854                           0                                    0
## 855                           0                                    0
## 856                           0                                    0
## 857                           0                                    0
## 858                           0                                    0
##     STDs.syphilis STDs.pelvic.inflammatory.disease STDs.genital.herpes
## 853             0                                0                   0
## 854             0                                0                   0
## 855             0                                0                   0
## 856             0                                0                   0
## 857             0                                0                   0
## 858             0                                0                   0
##     STDs.molluscum.contagiosum STDs.AIDS STDs.HIV STDs.Hepatitis.B STDs.HPV
## 853                          0         0        0                 0        0
## 854                          0         0        0                 0        0
## 855                          0         0        0                 0        0
## 856                          0         0        0                 0        0
## 857                          0         0        0                 0        0
## 858                          0         0        0                 0        0
##     STDs..Number.of.diagnosis STDs..Time.since.first.diagnosis
## 853                         0                               ?
## 854                         0                               ?
## 855                         0                               ?
## 856                         0                               ?
## 857                         0                               ?
## 858                         0                               ?
##     STDs..Time.since.last.diagnosis Dx.Cancer Dx.CIN Dx.HPV Dx Hinselmann
## 853                               ?         0      0      0  0          0
## 854                               ?         0      0      0  0          0
## 855                               ?         0      0      0  0          0
## 856                               ?         0      0      0  0          0
## 857                               ?         0      0      0  0          0
## 858                               ?         0      0      0  0          0
##     Schiller Citology Biopsy
## 853        0        0      0
## 854        0        0      0
## 855        0        0      0
## 856        0        1      0
## 857        0        0      0
## 858        0        0      0
```

Having seen this preliminary data analysis, there are many important takeaways. Firstly, much of the descriptive statistics calculations performed using the `summary` command did not work on seemingly numerical data. A closer look at the classes identified using the `str` command makes the reason for this clear: the variables facing this problem are considered `chr` variables, representing a string. This is an issue because R cannot perform numerical calculations on strings, so these columns will need to be converted into `num`'s, or numerical data. Secondly, quite a lot of the data seems to represent Boolean values, using 0 for `FALSE` and 1 for `TRUE`. This can be made cleaner by converting these columns into actual `TRUE`'s and `FALSE`'s, which will be corrected. Thirdly, many of the variable names are poor for the purposes of programming. As such, these will be converted to lower camel case to maintain readability and make them easier to type. Finally, `head` and `tail` have revealed that the data contains many missing values. This is expected, however this dataset represents missing values as question marks. To make it easier to handle these missing values, they should be represented as `NA`. Furthermore, these `NA` values must be replaced somehow in order to draw conclusions from the data. As one can see, this data is quite "dirty", further underscoring the necessity of data cleansing.

With that said, below demonstrates the process of performing all the previously mentioned tasks.

```
# read the data in again, however replace "?" with NA
# simultaneously converts chr -> num/int
ccData <- read.csv("cervicalCA.csv", header=TRUE, na.strings=c("?"))
```

It seems that since the question marks were represented by question marks (strings), R made the columns represented by `chr`'s. Since the question marks have been eliminated, the columns return to being correctly labeled as either numeric or integer.

```
# replace NAs with 0
ccData[is.na(ccData)] <- 0
```

When dealing with Boolean values and sensitive data such as medical history, handling `NA`'s can be a tricky task. A common method of removing invalid data is replacing the missing values with the average for each missing value's respective column. In this instance, this method is not very responsible or effective. Proclaiming a patient has a disease that they do not have or has had more sexual partners than is accurate in the pursuit of cleaner data does not accurately reflect reality. With Booleans, replacing with the mean is not even possible. Regardless, the best thing to do is probably to replace all `NA`'s with zero, as shown above. This assumes that if a patient's information for a disease is missing, that they do not have the disease, as it should (same logic applies for the number of sexual partners or number of pregnancies).

```
# verify NA values have been replaced
clean <- ifelse(complete.cases(ccData) == TRUE, 1, 0)
paste("There are ", dim(ccData)[1]-sum(clean), " rows with missing data.")
```

```
## [1] "There are  0  rows with missing data."
```

```
# change rows representing booleans using integers to actual booleans, or
# "logicals" as they are called in R; values below represent indices of boolean
# variables
ccData[, c(5, 8, 10, 12, 14:25, 29:36)] <- lapply(ccData[, c(5, 8, 10, 12, 14:25, 29:36)],
                                                  as.logical)

# re-examine the structure of the data
str(ccData)
```

```
## 'data.frame':    858 obs. of  36 variables:
##  $ Age                                : int  18 15 34 52 46 42 51 26 45 44 ...
##  $ Number.of.sexual.partners          : num  4 1 1 5 3 3 3 1 1 3 ...
##  $ First.sexual.intercourse           : num  15 14 0 16 21 23 17 26 20 15 ...
##  $ Num.of.pregnancies                 : num  1 1 1 4 4 2 6 3 5 0 ...
##  $ Smokes                             : logi  FALSE FALSE FALSE TRUE FALSE FALSE ...
##  $ Smokes..years.                     : num  0 0 0 37 0 ...
##  $ Smokes..packs.year.                : num  0 0 0 37 0 0 3.4 0 0 2.8 ...
##  $ Hormonal.Contraceptives            : logi  FALSE FALSE FALSE TRUE TRUE FALSE ...
##  $ Hormonal.Contraceptives..years.    : num  0 0 0 3 15 0 0 2 0 0 ...
##  $ IUD                                : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ IUD..years.                        : num  0 0 0 0 0 0 7 7 0 0 ...
##  $ STDs                               : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ STDs..number.                      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ STDs.condylomatosis                : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
```

```
##  $ STDs.cervical.condylomatosis      : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ STDs.vaginal.condylomatosis       : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ STDs.vulvo.perineal.condylomatosis: logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ STDs.syphilis                     : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ STDs.pelvic.inflammatory.disease  : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ STDs.genital.herpes               : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ STDs.molluscum.contagiosum        : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ STDs.AIDS                         : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ STDs.HIV                          : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ STDs.Hepatitis.B                  : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ STDs.HPV                          : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ STDs..Number.of.diagnosis         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ STDs..Time.since.first.diagnosis  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ STDs..Time.since.last.diagnosis   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Dx.Cancer                         : logi  FALSE FALSE FALSE TRUE FALSE FALSE ...
##  $ Dx.CIN                            : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ Dx.HPV                            : logi  FALSE FALSE FALSE TRUE FALSE FALSE ...
##  $ Dx                                : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ Hinselmann                        : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ Schiller                          : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ Citology                          : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ Biopsy                            : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
```

```
# re-examine the basic descriptive statistics
summary(ccData)
```

```
##       Age        Number.of.sexual.partners First.sexual.intercourse
##  Min.   :13.00   Min.   : 0.000            Min.   : 0.00
##  1st Qu.:20.00   1st Qu.: 1.000            1st Qu.:15.00
##  Median :25.00   Median : 2.000            Median :17.00
##  Mean   :26.82   Mean   : 2.451            Mean   :16.86
##  3rd Qu.:32.00   3rd Qu.: 3.000            3rd Qu.:18.00
##  Max.   :84.00   Max.   :28.000            Max.   :32.00
##  Num.of.pregnancies   Smokes        Smokes..years.   Smokes..packs.year.
##  Min.   : 0.000     Mode :logical   Min.   : 0.000   Min.   : 0.0000
##  1st Qu.: 1.000     FALSE:735       1st Qu.: 0.000   1st Qu.: 0.0000
##  Median : 2.000     TRUE :123       Median : 0.000   Median : 0.0000
##  Mean   : 2.127                     Mean   : 1.201   Mean   : 0.4463
##  3rd Qu.: 3.000                     3rd Qu.: 0.000   3rd Qu.: 0.0000
##  Max.   :11.000                     Max.   :37.000   Max.   :37.0000
##  Hormonal.Contraceptives Hormonal.Contraceptives..years.    IUD
##  Mode :logical           Min.   : 0.000                  Mode :logical
##  FALSE:377               1st Qu.: 0.000                  FALSE:775
##  TRUE :481               Median : 0.250                  TRUE :83
##                          Mean   : 1.972
##                          3rd Qu.: 2.000
##                          Max.   :30.000
##   IUD..years.         STDs         STDs..number.   STDs.condylomatosis
##  Min.   : 0.0000   Mode :logical   Min.   :0.000   Mode :logical
##  1st Qu.: 0.0000   FALSE:779       1st Qu.:0.000   FALSE:814
##  Median : 0.0000   TRUE :79        Median :0.000   TRUE :44
##  Mean   : 0.4446                   Mean   :0.155
##  3rd Qu.: 0.0000                   3rd Qu.:0.000
##  Max.   :19.0000                   Max.   :4.000
```

```
##  STDs.cervical.condylomatosis STDs.vaginal.condylomatosis
##  Mode :logical                Mode :logical
##  FALSE:858                    FALSE:854
##                               TRUE :4
##
##
##
##  STDs.vulvo.perineal.condylomatosis STDs.syphilis
##  Mode :logical                      Mode :logical
##  FALSE:815                          FALSE:840
##  TRUE :43                           TRUE :18
##
##
##
##  STDs.pelvic.inflammatory.disease STDs.genital.herpes
##  Mode :logical                    Mode :logical
##  FALSE:857                        FALSE:857
##  TRUE :1                          TRUE :1
##
##
##
##  STDs.molluscum.contagiosum STDs.AIDS        STDs.HIV       STDs.Hepatitis.B
##  Mode :logical              Mode :logical   Mode :logical   Mode :logical
##  FALSE:857                  FALSE:858       FALSE:840       FALSE:857
##  TRUE :1                                    TRUE :18        TRUE :1
##
##
##
##   STDs.HPV       STDs..Number.of.diagnosis STDs..Time.since.first.diagnosis
##  Mode :logical   Min.   :0.00000           Min.   : 0.0000
##  FALSE:856       1st Qu.:0.00000           1st Qu.: 0.0000
##  TRUE :2         Median :0.00000           Median : 0.0000
##                  Mean   :0.08741           Mean   : 0.5082
##                  3rd Qu.:0.00000           3rd Qu.: 0.0000
##                  Max.   :3.00000           Max.   :22.0000
##  STDs..Time.since.last.diagnosis Dx.Cancer         Dx.CIN
##  Min.   : 0.0000                 Mode :logical   Mode :logical
##  1st Qu.: 0.0000                 FALSE:840       FALSE:849
##  Median : 0.0000                 TRUE :18        TRUE :9
##  Mean   : 0.4814
##  3rd Qu.: 0.0000
##  Max.   :22.0000
##    Dx.HPV            Dx          Hinselmann       Schiller
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:840       FALSE:834       FALSE:823       FALSE:784
##  TRUE :18        TRUE :24        TRUE :35        TRUE :74
##
##
##
##   Citology         Biopsy
##  Mode :logical   Mode :logical
##  FALSE:814       FALSE:803
##  TRUE :44        TRUE :55
##
```

```
##
##
```

With this re-examination, it is clear that the classes are accurate, as well as the descriptive statistics given by `summary`.

```
# rename columns using lowerCamelCase
names(ccData) <- c("age", "numPartners", "firstIntercourse", "numPregnancies",
                   "isSmoker", "yearsSmoking", "packsPerYear", "onContra",
                   "yearsOnContra", "onIUD", "yearsOnIUD", "hasSTDs", "numSTDs",
                   "hasCondyloma", "hasCervCondyloma", "hasVagCondyloma",
                   "hasVulvoPerinealCondy", "hasSyphilis", "hasPID", "hasGenHerpes",
                   "hasMolluscum", "hasAIDS", "hasHIV", "hasHepB", "hasHPV",
                   "numDiagnosis", "sinceFirstDiag", "sinceLastDiag", "hasCancerDiag",
                   "hasCINDiag", "hasHPVDiag", "hasDiag", "hadHinselmann", "hadSchiller",
                   "hadCytology", "hadBiopsy")
names(ccData)
```

```
##  [1] "age"                   "numPartners"           "firstIntercourse"
##  [4] "numPregnancies"        "isSmoker"              "yearsSmoking"
##  [7] "packsPerYear"          "onContra"              "yearsOnContra"
## [10] "onIUD"                 "yearsOnIUD"            "hasSTDs"
## [13] "numSTDs"               "hasCondyloma"          "hasCervCondyloma"
## [16] "hasVagCondyloma"       "hasVulvoPerinealCondy" "hasSyphilis"
## [19] "hasPID"                "hasGenHerpes"          "hasMolluscum"
## [22] "hasAIDS"               "hasHIV"                "hasHepB"
## [25] "hasHPV"                "numDiagnosis"          "sinceFirstDiag"
## [28] "sinceLastDiag"         "hasCancerDiag"         "hasCINDiag"
## [31] "hasHPVDiag"            "hasDiag"               "hadHinselmann"
## [34] "hadSchiller"           "hadCytology"           "hadBiopsy"
```

Lower camel case is largely the standard in the realm of programming, and for a reason. This formatting makes variables easier to type, while maintaining readability. Having concise names while maintaining a degree of descriptiveness is another thing that was considered when selecting names. With this, anybody would be able to quickly glance at the name of these variables and understand what it is describing.

With all of this, the data has been properly cleansed. To reiterate, this process is absolutely necessary if any sort of significant meaning is to be gained from the data. One is now able to create accurate models based on the data, and in an organized fashion. After all, if the data was not cleaned, missing values and numbers represented by strings would make it impossible to do so. Once clean data has been produced, it is good practice to write it to a new file for any future use, as demonstrated below.

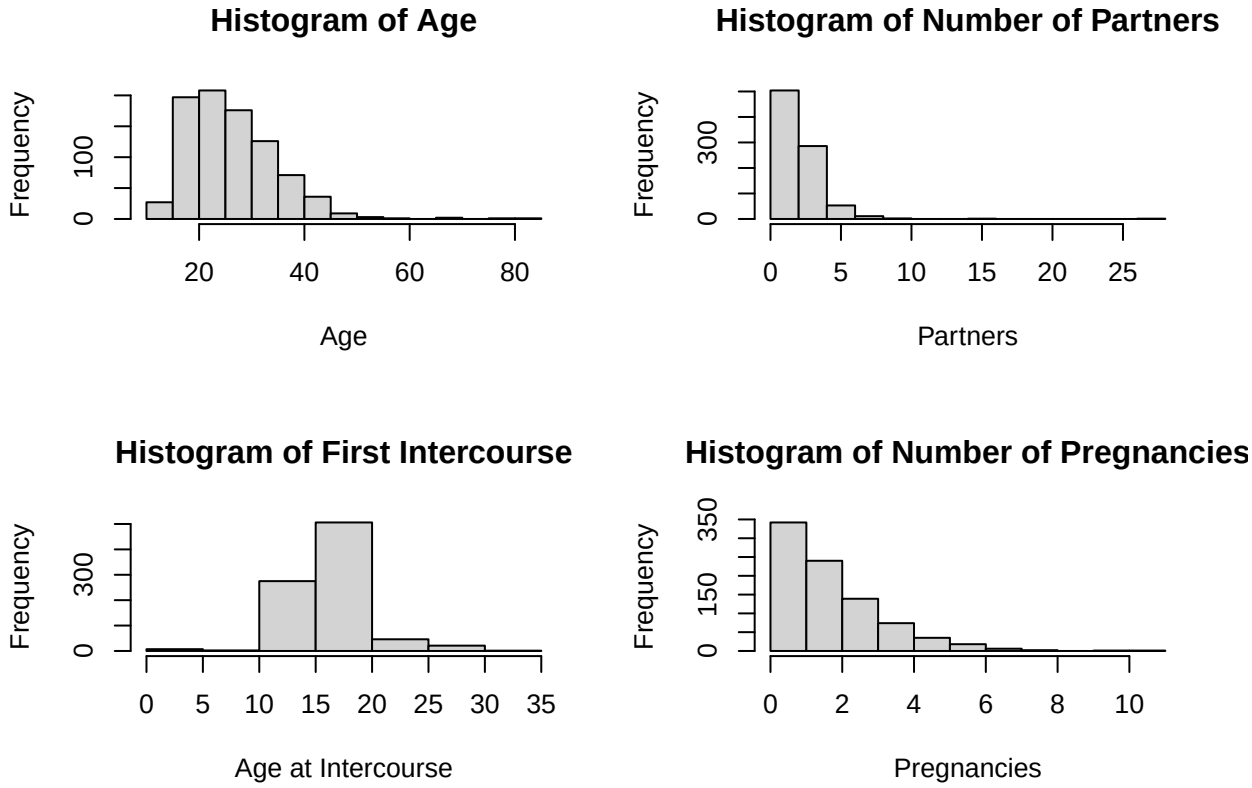
```
# write clean data to a new CSV
write.csv(ccData, "ccdataMod.csv", row.names=TRUE)
```

3. Data Transformation

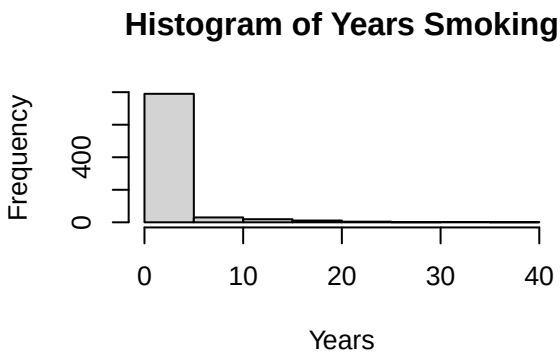
Now that the data is clean, analysis may commence. Before creating models on the clean variables, it is advised to normalize the numerical data (numerical as in representing by numbers, not the class of data built into R). This changes the values of the data to share a common scale, which will make certain algorithms and models perform better. The numerical data in this dataset includes the following: `age`, `numPartners`, `firstIntercourse`, `numPregnancies`, `yearsSmoking`, `packsPerYear`, `yearsOnContra`, `yearsOnIUD`, `numSTDs`, `numDiagnosis`, `sinceFirstDiag`, and `sinceLastDiag`. It should be noted that the

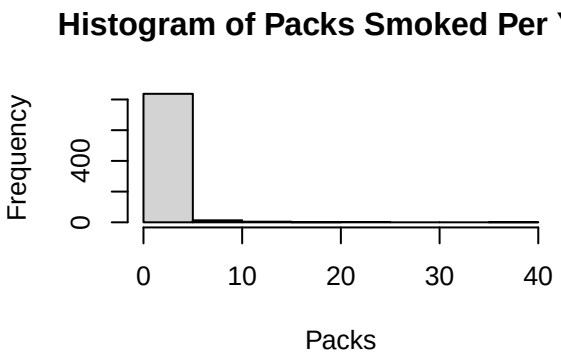
normalized data will be added as a series of new columns, as we might still want to use the original data sometime. To observe the data how it stands, the histograms of the original data can be seen below.

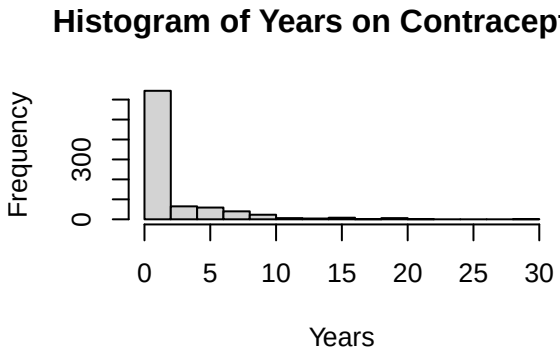
```
# put 2x2 plots per graphic
par(mfrow=c(2, 2))
hist(ccData$age, main="Histogram of Age", xlab="Age")
hist(ccData$numPartners, main="Histogram of Number of Partners", xlab="Partners")
hist(ccData$firstIntercourse, main="Histogram of First Intercourse", xlab="Age at Intercourse")
hist(ccData$numPregnancies, main="Histogram of Number of Pregnancies", xlab="Pregnancies")
```

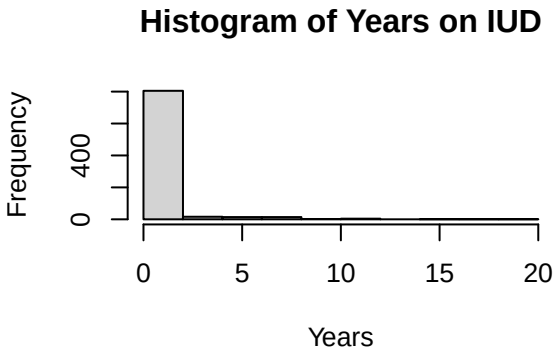



```
hist(ccData$yearsSmoking, main="Histogram of Years Smoking", xlab="Years")
hist(ccData$packsPerYear, main="Histogram of Packs Smoked Per Year", xlab="Packs")
hist(ccData$yearsOnContra, main="Histogram of Years on Contraceptives", xlab="Years")
hist(ccData$yearsOnIUD, main="Histogram of Years on IUD", xlab="Years")
```

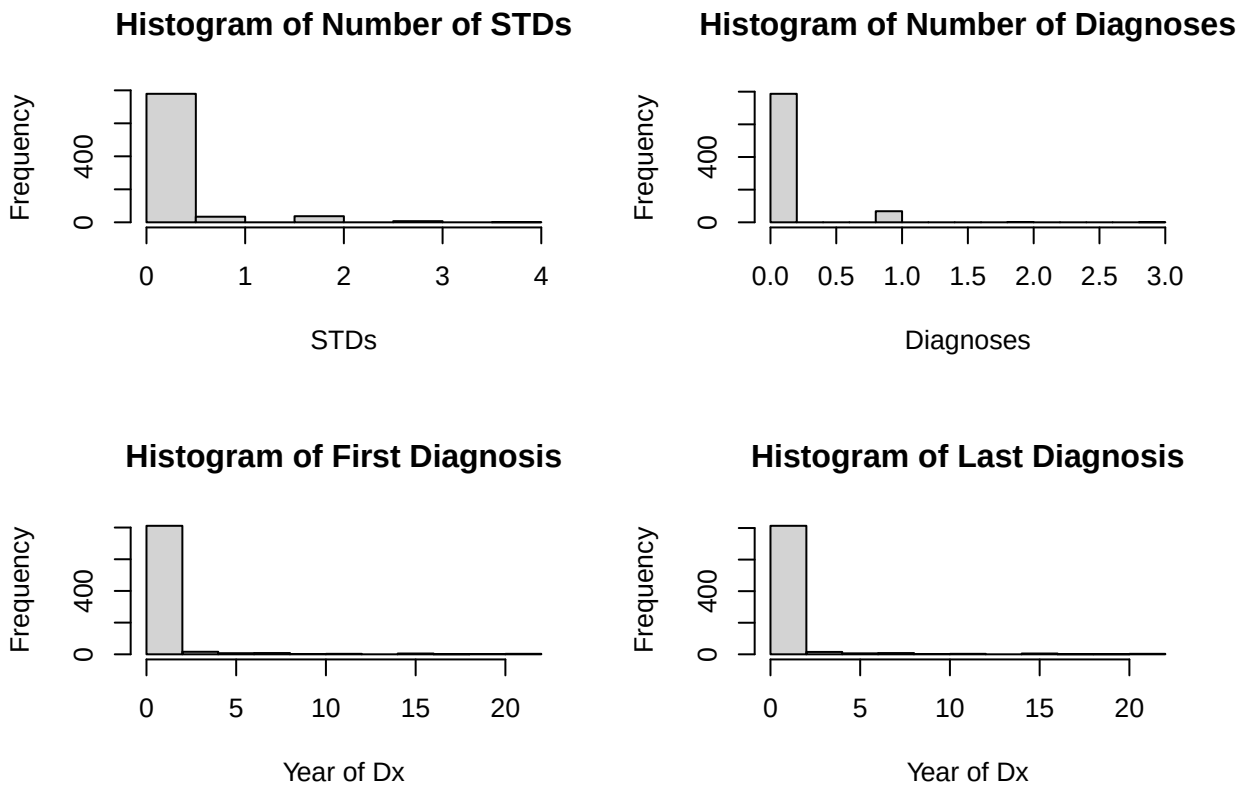









```
hist(ccData$numSTDs, main="Histogram of Number of STDs", xlab="STDs")
hist(ccData$numDiagnosis, main="Histogram of Number of Diagnoses", xlab="Diagnoses")
hist(ccData$sinceFirstDiag, main="Histogram of First Diagnosis", xlab="Year of Dx")
hist(ccData$sinceLastDiag, main="Histogram of Last Diagnosis", xlab="Year of Dx")
```



The data looks alright, however normalizing it should help to create better looking histograms. To normalize, a Rank-Z transformation will be performed on all of the numeric data. Furthermore, all of this transformed data will be stored in a separate dataset. It is common practice to simply append the new, transformed data to the existing dataset, however this dataset is already quite wide, and it would be simpler to treat it as a separate entity. The new dataset is called `ccNumeric`, and demonstrated below.

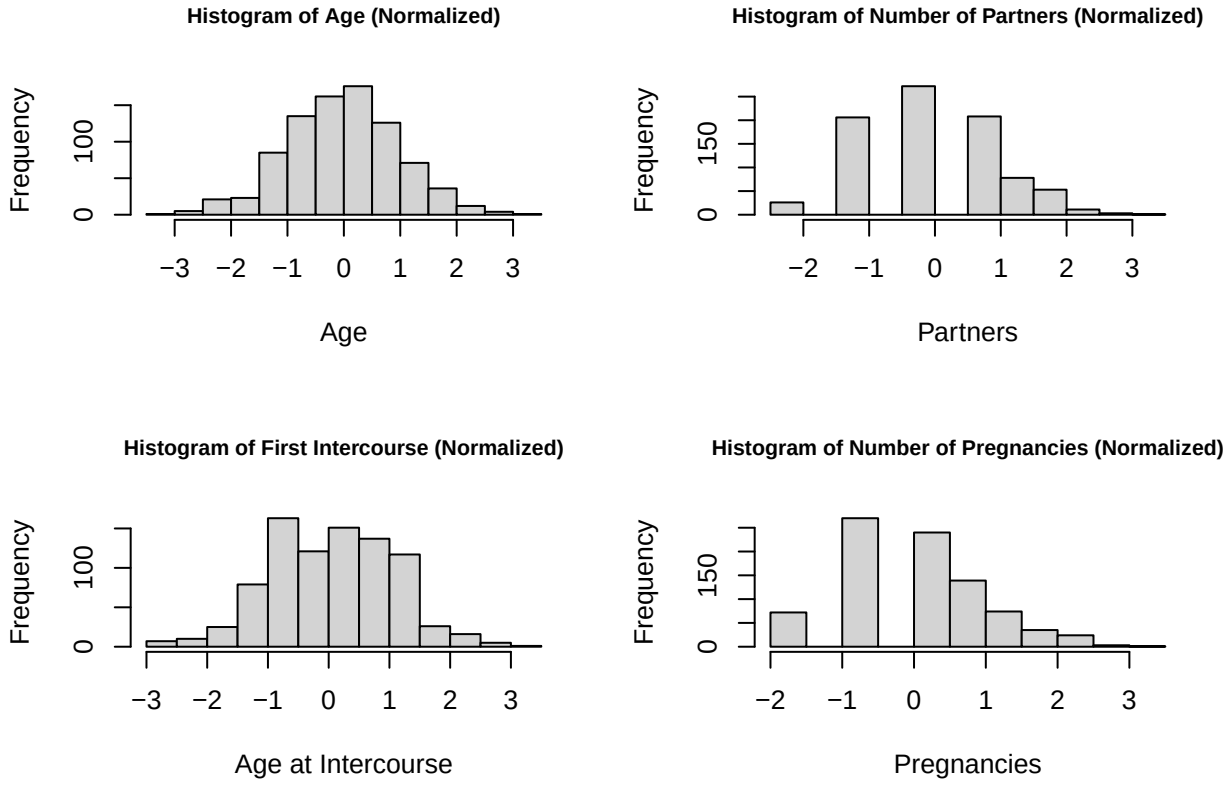
```
# select only the num/int data, then normalize it using a Rank-Z transformation
ccNumeric <- ccData %>%
             select("age", "numPartners", "firstIntercourse", "numPregnancies",
                    "yearsSmoking", "packsPerYear", "yearsOnContra", "yearsOnIUD",
                    "numSTDs", "numDiagnosis", "sinceFirstDiag", "sinceLastDiag") %>%
             lapply(rz.transform)

# append "_RZ" to each column to differentiate it further from regular data
# this doesn't follow lowerCamelCase, however keeps the names readable
names(ccNumeric) <- lapply(names(ccNumeric), function(colName){
  str_c(colName, "_RZ")
})

str(ccNumeric)
```

```
## List of 12
##  $ age_RZ             : num [1:858] -1.133 -2.058 0.868 2.429 2.083 ...
##  $ numPartners_RZ     : num [1:858] 1.15 -1.033 -1.033 1.606 0.549 ...
##  $ firstIntercourse_RZ: num [1:858] -0.718 -1.308 -2.6 -0.249 1.49 ...
##  $ numPregnancies_RZ  : num [1:858] -0.701 -0.701 -0.701 1.19 1.19 ...
```

```
##  $ yearsSmoking_RZ    : num [1:858] -0.18 -0.18 -0.18 3.04 -0.18 ...
##  $ packsPerYear_RZ    : num [1:858] -0.18 -0.18 -0.18 3.04 -0.18 ...
##  $ yearsOnContra_RZ   : num [1:858] -0.772 -0.772 -0.772 0.76 2.123 ...
##  $ yearsOnIUD_RZ      : num [1:858] -0.121 -0.121 -0.121 -0.121 -0.121 ...
##  $ numSTDs_RZ         : num [1:858] -0.116 -0.116 -0.116 -0.116 -0.116 ...
##  $ numDiagnosis_RZ    : num [1:858] -0.104 -0.104 -0.104 -0.104 -0.104 ...
##  $ sinceFirstDiag_RZ  : num [1:858] -0.104 -0.104 -0.104 -0.104 -0.104 ...
##  $ sinceLastDiag_RZ   : num [1:858] -0.104 -0.104 -0.104 -0.104 -0.104 ...
```

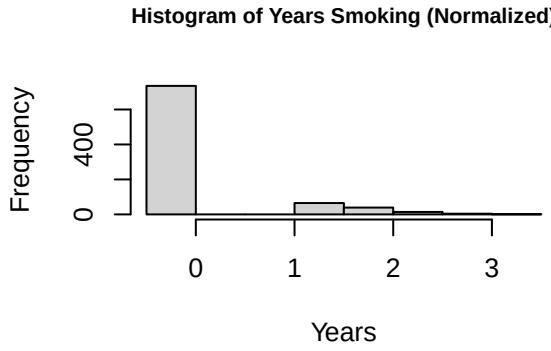
With this transformed dataset prepared, the new histograms may be examined and contrasted to their originals.

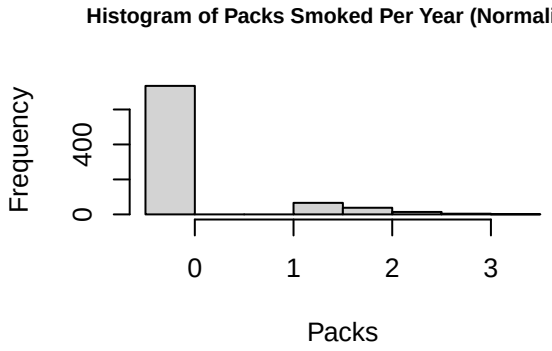
```
# put 2x2 plots per graphic; reduce main font size to prevent truncation of title
par(mfrow=c(2, 2))
hist(ccNumeric$age_RZ, main="Histogram of Age (Normalized)", xlab="Age", cex.main=0.85)
hist(ccNumeric$numPartners_RZ, main="Histogram of Number of Partners (Normalized)",
     xlab="Partners", cex.main=0.85)
hist(ccNumeric$firstIntercourse_RZ, main="Histogram of First Intercourse (Normalized)",
     xlab="Age at Intercourse", cex.main=0.85)
hist(ccNumeric$numPregnancies_RZ, main="Histogram of Number of Pregnancies (Normalized)",
     xlab="Pregnancies", cex.main=0.85)
```

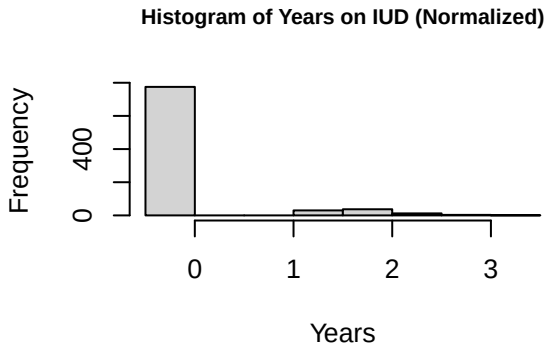



```
hist(ccNumeric$yearsSmoking_RZ, main="Histogram of Years Smoking (Normalized)",
     xlab="Years", cex.main=0.85)
hist(ccNumeric$packsPerYear_RZ, main="Histogram of Packs Smoked Per Year (Normalized)",
```

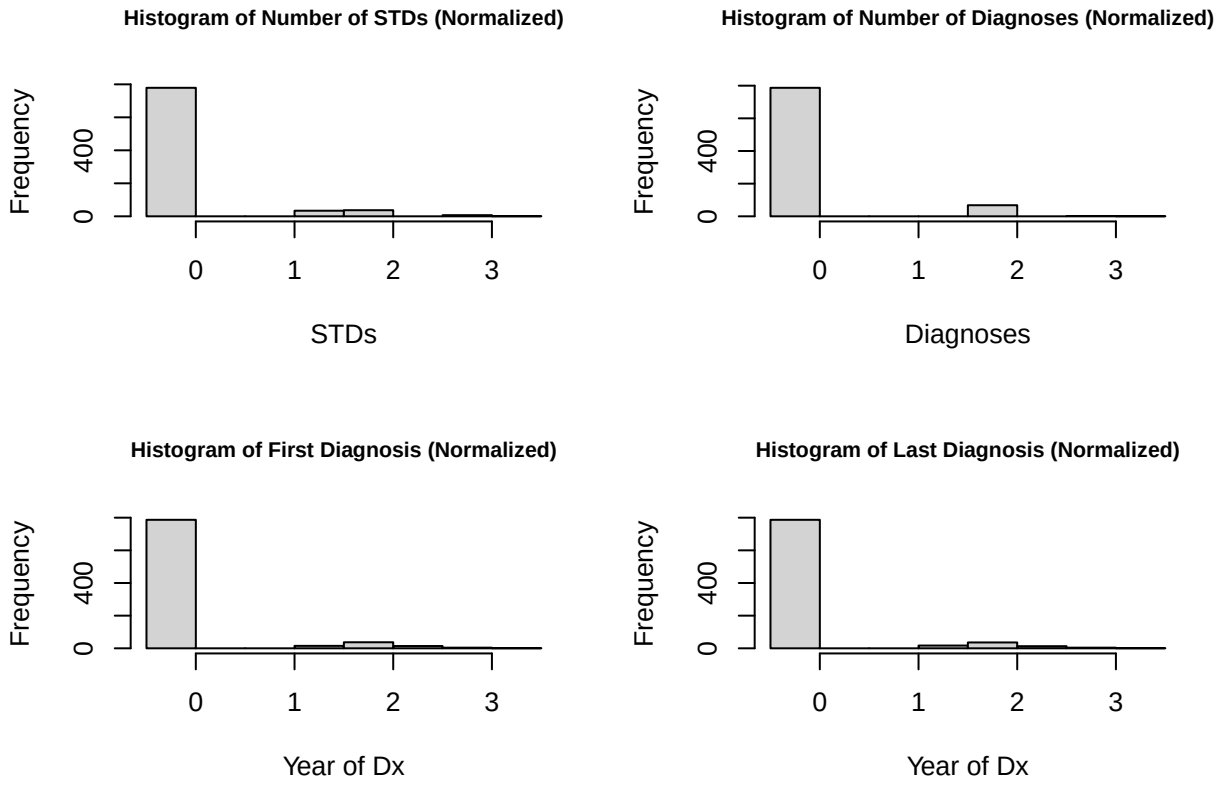
```
     xlab="Packs", cex.main=0.85)
hist(ccNumeric$yearsOnContra_RZ, main="Histogram of Years on Contraceptives (Normalized)",
     xlab="Years", cex.main=0.85)
hist(ccNumeric$yearsOnIUD_RZ, main="Histogram of Years on IUD (Normalized)",
     xlab="Years", cex.main=0.85)
```








```
hist(ccNumeric$numSTDs_RZ, main="Histogram of Number of STDs (Normalized)",
     xlab="STDs", cex.main=0.85)
hist(ccNumeric$numDiagnosis_RZ, main="Histogram of Number of Diagnoses (Normalized)",
     xlab="Diagnoses", cex.main=0.85)
hist(ccNumeric$sinceFirstDiag_RZ, main="Histogram of First Diagnosis (Normalized)",
     xlab="Year of Dx", cex.main=0.85)
hist(ccNumeric$sinceLastDiag_RZ, main="Histogram of Last Diagnosis (Normalized)",
     xlab="Year of Dx", cex.main=0.85)
```



This new, transformed data appears to be more normally distributed (as intended), however it should be noted that many of the histograms remained heavily skewed. This can be attributed to the overwhelming amount of zeroes that can be found in many of these columns, such as in the number of STDs, or years spent smoking. These conditions and behaviors are relatively rare, and thus the data is greatly skewed.

4. Human Papillomavirus (HPV) and Cervical Cancer

As discussed briefly in the introduction (1), *human papillomavirus*, or HPV, is the leading cause of cervical cancer in women, accounting for nearly 90% of cases of cervical cancer in women. This figure is rather striking, and naturally should be looked at in-depth in this report.

Not only does HPV account for a large majority of cervical cancer cases, but it is the most common viral infection of the reproductive tract. While cervical cancer is a female-specific cancer, HPV is an infection that can manifest itself in both men and women. Because of this (and the frequency of the STD), practicing safe sex is the most effective method of reducing the risk of receiving the disease. Interestingly, many types of HPV are practically harmless, however certain cases greatly increase one's risk for a variety of problems, such as cervical cancer.

When a women gets infected with HPV, pre-cancerous lesions form in the cervix. As aforementioned, many cases of HPV are relatively harmless, so these lesions often clear up without the need for intervention. With that said, sometimes they do not—when this happens, and the infection is chronic, these lesions begin to develop into invasive cancer. It usually takes at least fifteen years for cancer to develop from an HPV infection, however it can occur more quickly in women with weakened immune systems. The HPV vaccine protects against the two types which are most likely to develop into cancer, however there is still a risk of getting a dangerous type of HPV even with the vaccine.

To investigate the correlation between cervical cancer and HPV, or more accurately causality, BIC models

will be used. A BIC (Bayesian information criterion) model is used to test the likelihood that a certain variable caused another, so it can be used to determine whether HPV has a palpable effect on cancer rates. The code below demonstrates this model:

```
# BIC model to test if HPV causes cervical cancer
self_BIC <- BIC(lm(ccData$hasDiag~1)) # -644.6145
cause_BIC <- BIC(lm(ccData$hasDiag~ccData$hasHPVDiag)) # -1047.82
paste("Difference of", self_BIC-cause_BIC)
```

```
## [1] "Difference of 403.205132512408"
```

When using BIC models, it is usually fair to say that variable x causes variable y if the value of y~1 is at least ten points higher than the value of y~x. In the case of these two variables, the difference was over 400. It is absolutely certain that a diagnosis of HPV increases the likelihood of getting diagnosed with cervical cancer. Of course, this simply data analysis just adds to the extensive research that has already been conducted, but it further demonstrates the correlation from a statistical perspective.

BIC models can be run with multiple variables, too. For example, it is said that smoking increases the risk of developing cervical cancer, especially if the smoker has HPV. To test this, a BIC model can be run with the extra consideration of cigarette use to determine if there is any significant difference.

```
# BIC model to test if HPV and smoking causes cervical cancer
self_BIC <- BIC(lm(ccData$hasDiag~1)) # -644.6145
cause_BIC <- BIC(lm(ccData$hasDiag~ccData$hasHPVDiag+ccData$isSmoker)) # -1048.967
paste("Difference of", self_BIC-cause_BIC)
```

```
## [1] "Difference of 404.352619079144"
```

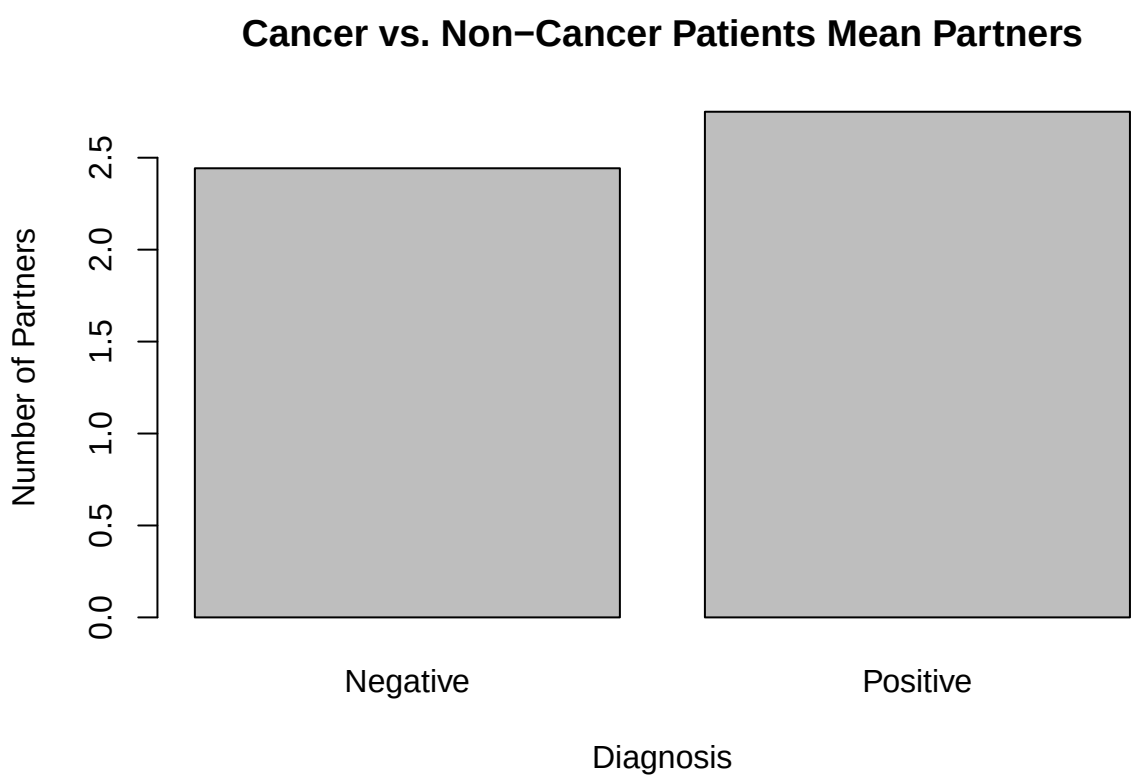
The difference only increased by around one. While this suggests a very slight additional influence, it is not significantly different. This likely occurred due to a limited data sample. It is important to understand that while analysis may suggest a certain conclusion (such as smoking has relatively little effect on the development of cervical cancer from HPV), the data can be misleading and should not be blindly followed. The large amount of research opposing this conclusion should make that obvious. That is not to say that the data is bad—it is simply one instance in which it goes against what is expected.

5. Investigating Other Common Causes

Although HPV accounts for a large majority of cervical cancer cases, other variables are linked with the disease. For example, as briefly mentioned, the use of contraceptives has been associated with the development of cervical cancer; long-term use of contraceptives (5+ years) can as much as triple risk of cervical cancer. This could be because HPV and other STD's spread via sexual intercourse, and women taking contraceptives likely actively have sex. Even having multiple pregnancies potentially increases risk as well. To measure some of these statistics, a mixture of basic bar plots and BIC models can be used, as demonstrated below.

```
# filter data into two new sets; patients with and without cervical cancer
diagData <- ccData %>% filter(ccData$hasDiag==TRUE)
nonDiagData <- ccData %>% filter(ccData$hasDiag==FALSE)

# test number of partners
avgPartnersDiag <- mean(diagData$numPartners)
avgPartnersNonDiag <- mean(nonDiagData$numPartners)
barplot(c(avgPartnersNonDiag, avgPartnersDiag),
```

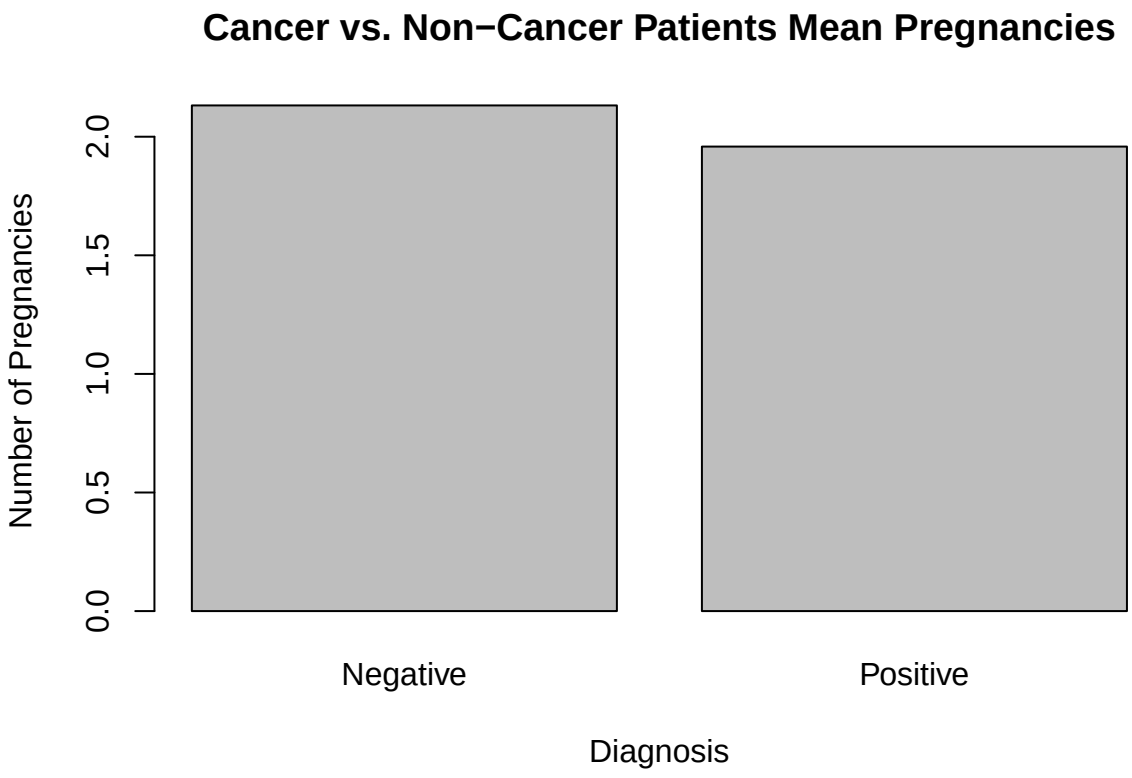
```
        xlab="Diagnosis", ylab="Number of Partners",
        names.arg=c("Negative", "Positive"),
        main="Cancer vs. Non-Cancer Patients Mean Partners")
```




```
self_BIC <- BIC(lm(ccData$hasDiag~1)) # -644.6145
cause_BIC <- BIC(lm(ccData$hasDiag~ccData$numPartners)) # -638.626
paste("Difference of", self_BIC-cause_BIC) # insignificant
```

```
## [1] "Difference of -5.9885008022884"
```

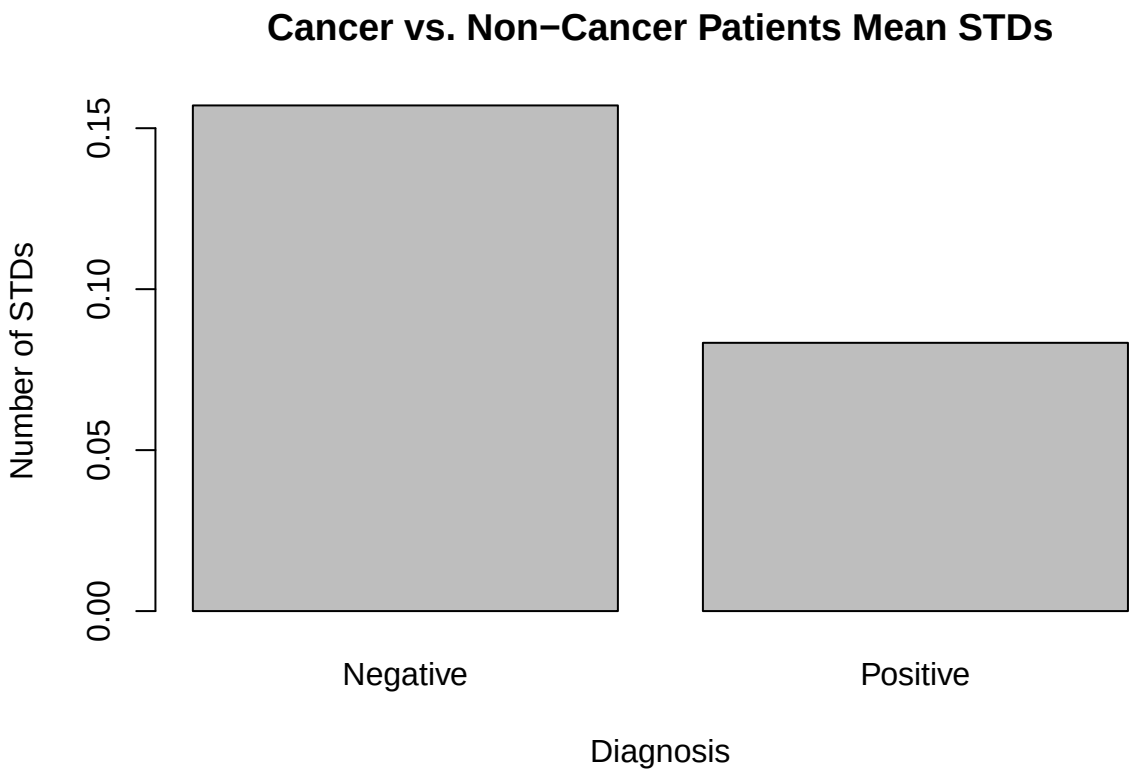
```
# test number of pregnancies
avgPregDiag <- mean(diagData$numPregnancies)
avgPregNonDiag <- mean(nonDiagData$numPregnancies)
barplot(c(avgPregNonDiag, avgPregDiag),
        xlab="Diagnosis", ylab="Number of Pregnancies",
        names.arg=c("Negative", "Positive"),
        main="Cancer vs. Non-Cancer Patients Mean Pregnancies")
```




```
self_BIC <- BIC(lm(ccData$hasDiag~1)) # -644.6145
cause_BIC <- BIC(lm(ccData$hasDiag~ccData$numPregnancies)) # -638.1692
paste("Difference of", self_BIC-cause_BIC) # insignificant
```

```
## [1] "Difference of -6.44520804039257"
```

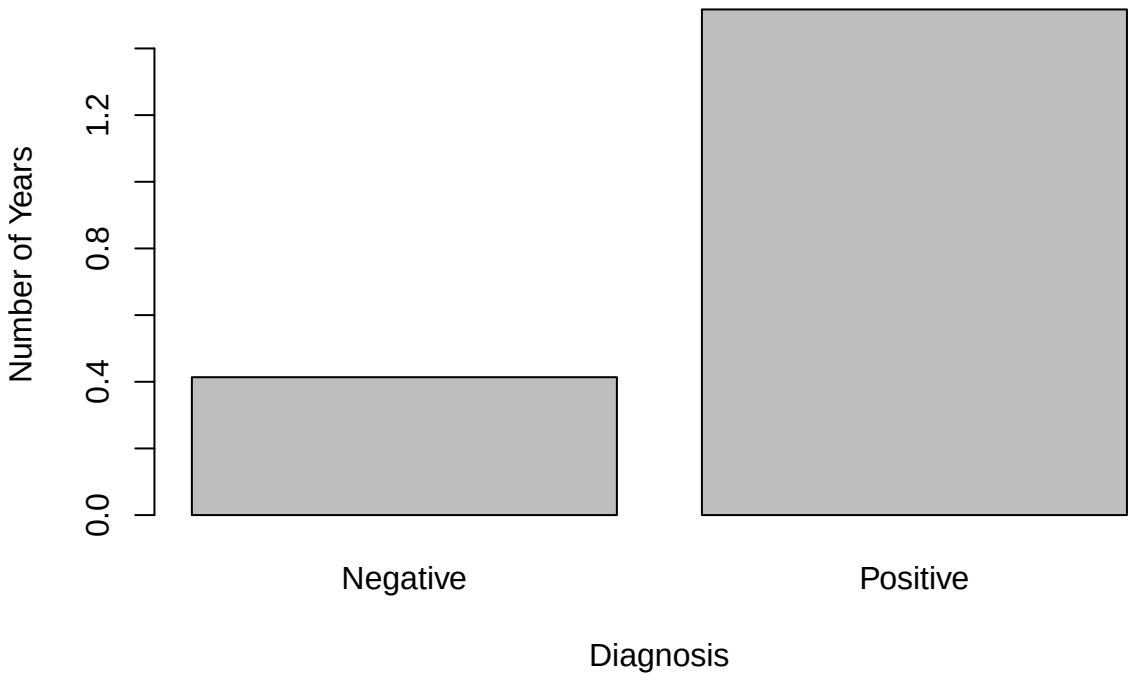
```
# test number of STDs
avgSTDsDiag <- mean(diagData$numSTDs)
avgSTDsNonDiag <- mean(nonDiagData$numSTDs)
barplot(c(avgSTDsNonDiag, avgSTDsDiag),
        xlab="Diagnosis", ylab="Number of STDs",
        names.arg=c("Negative", "Positive"),
        main="Cancer vs. Non-Cancer Patients Mean STDs")
```




```
self_BIC <- BIC(lm(ccData$hasDiag~1)) # -644.6145
cause_BIC <- BIC(lm(ccData$hasDiag~ccData$numSTDs)) # -638.3128
paste("Difference of", self_BIC-cause_BIC) # insignificant
```

```
## [1] "Difference of -6.30170080785786"
```

```
# test number of years on IUD
avgYrsIUDDiag <- mean(diagData$yearsOnIUD)
avgYrsIUDNonDiag <- mean(nonDiagData$yearsOnIUD)
barplot(c(avgYrsIUDNonDiag, avgYrsIUDDiag),
        xlab="Diagnosis", ylab="Number of Years",
        names.arg=c("Negative", "Positive"),
        main="Cancer vs. Non-Cancer Patients Mean Years on IUD")
```




```
self_BIC <- BIC(lm(ccData$hasDiag~1)) # -644.6145
cause_BIC <- BIC(lm(ccData$hasDiag~ccData$yearsOnIUD)) # -646.5421
paste("Difference of", self_BIC-cause_BIC) # mostly insignificant
```

```
## [1] "Difference of 1.92767286515561"
```

```
# test collective effect of many STDs
self_BIC <- BIC(lm(ccData$hasDiag~1)) # -644.6145
cause_BIC <- BIC(lm(ccData$hasDiag~ccData$hasCondyloma+ccData$hasAIDS+ccData$hasCINDiag
                                  +ccData$hasGenHerpes+ccData$hasHepB+ccData$hasMolluscum
                                  +ccData$hasPID+ccData$hasSyphilis+ccData$hasVagCondyloma
                                  +ccData$hasVulvoPerinealCondy)) # -979.2616
paste("Difference of", self_BIC-cause_BIC) # significant
```

```
## [1] "Difference of 334.64710494403"
```

None of these variables experienced a decrease of ten or more (with the exception of the multiple regression, which will be talked about shortly). In fact, many of them actually increased; the only one to have somewhat of a correlation was the years on IUD, and it was deemed insignificant by the BIC model. With a bigger, better dataset, maybe correlations would be more easily seen, however many of these causes did not have much of an effect on the diagnosis. Although these factors have been identified as significant through outside research, it is tough to observe the same correlations in this set of data. With that said, the (frankly huge) multiple regression involving a wide variety of STD's identified a significant amount of causality, with a difference of over 300. This supports the conclusion that STD's are the main cause of cervical cancer, which was seen in part with the causality observed in HPV. The other factors do not seem to play a significant role, at least within this particular dataset.

6. Detecting Cervical Cancer

There a couple different ways to detect cervical cancer. Biopsies seem to be the most widely used, and involve using a colposcopy to observe the cervix. From there, doctors can both visually identify cervical cancer and take a sample for testing. In cytology, a similar approach is used, however if sample tissue is recovered, it is usually much less tissue to be examined on a cellular level. In Schiller's test, an iodine solution is applied to the cervix; if the tissue turns brown, then there is no cancer present. Otherwise, abnormal areas (such as early cancer) is likely. With Hinselmann's test, acetic acid was applied to the cervix. When applied, lesions would become visible, and diagnoses could be made. The BIC values, testing causality of each test for a cancer diagnosis, are shown in Table 1.

```
# create table using BIC models of each test
diagBIC <- BIC(lm(ccData$hasDiag~1)) # -644.6145
biopsyBIC <- c(diagBIC-BIC(lm(ccData$hasDiag~ccData$hadBiopsy)))
schillerBIC <- c(diagBIC-BIC(lm(ccData$hasDiag~ccData$hadSchiller)))
cytologyBIC <-  c(diagBIC-BIC(lm(ccData$hasDiag~ccData$hadCytology)))
hinselmanBIC <- c(diagBIC-BIC(lm(ccData$hasDiag~ccData$hadHinselmann)))
testTable <- data.frame(biopsyBIC, schillerBIC, cytologyBIC, hinselmanBIC)
kable(testTable, caption="Relative effectiveness of each type of test using BIC models.
                          Recall differences of ten or higher are significant",
                 col.names=c("Biopsy", "Schiller", "Cytology", "Hinselman"))
```

Table 1: Relative effectiveness of each type of test using BIC models. Recall differences of ten or higher are significant

Biopsy	Schiller	Cytology	Hinselman
14.82716	1.687919	0.0287019	-2.26844

Based on this, it seems that the only test with significant correlation is the biopsy. This could be because biopsy is more reliable, although it could just as easily be because of chance considering the relatively low causality displayed by the BIC model.

7. References

1. https://en.wikipedia.org/wiki/Cervical_cancer#Epidemiology
2. https://www.who.int/news-room/fact-sheets/detail/human-papillomavirus-(hpv)-and-cervical-cancer
3. https://en.wikipedia.org/wiki/Schiller%27s_test
4. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3279084/