Yuefu Jiang

Biostatistics Final Project: Cervical Cancer Risk Factor Analysis

Introduction

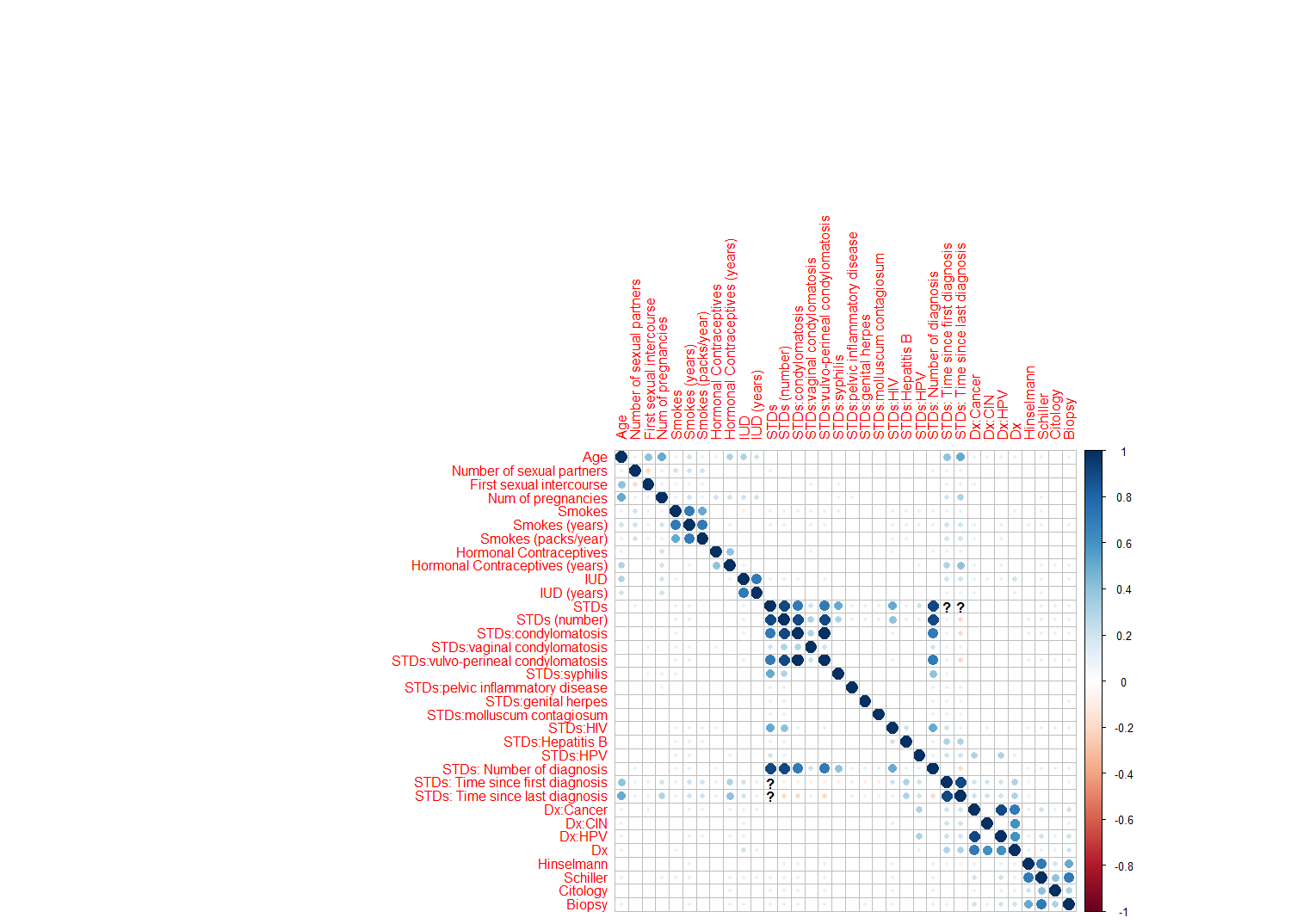
Cervical cancer was a major cause of death in United States in the last century. Although the medical practice to treat this disease has advanced during last few decades, it still has an alarming high diagnosis/death rate, especially in developing countries. Reasons behind this high mortality besides the malicious nature of cancer were believed to be low awareness of gynecological diseases as well as lack of cancer screening. In this analysis, a data set containing 36 attributes from 858 patients in Venezuela was analyzed in order to identify some plausible risk factors of cervical cancer. According to recent researches, HPV infections was one of the major risk factors contributing to the development of the disease; it was also believed that smoking, number of pregnancies as well as some methods of birth control employed could contribute to the acquirement cervical cancer. In this analysis, such factors, along with other factors will be analyzed statistically in seek of significant relations. On the other hand, most of the modern analyses were done in developed countries such as US; as this data set was collected in Venezuela, a developing country, it was plausible that some risk actors may differ from previous results gained in developed countries.

Methods

Correlation matrix was built upon all attributes to select highly correlated attributes for further analysis, excluding HIV diagnosis and cervical condylomatosis as no observation across all patients could be found. Spearman correlation was employed as many diagnostic attributes were binary. For following statistic analysis on specific attributes and cervical cancer diagnosis, general linear model was built as logistic regression was performed at alpha = .05. Pearson’s linear regression was also performed on numerical attributes and cervical cancer diagnosis. All analysis was done under R 3.5.1.

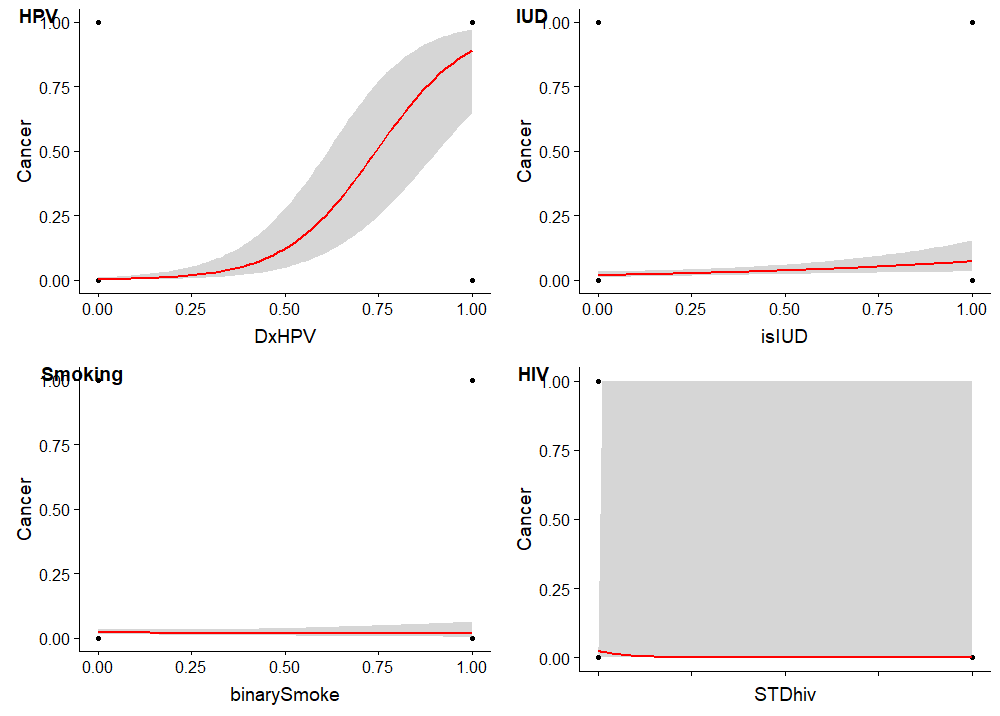
Results

From generated correlation matrix (Fig. 1), it was obvious that HPV was the major positive correlator of cervical cancer, with a r of 0.9. Weaker positive correlation was also observed in both time since first diagnosis of STD as well as time since last diagnosis of STD. (r = 0.2 for both). Other common risk factors were not observed to be significantly correlated to cervical cancer with this analysis.



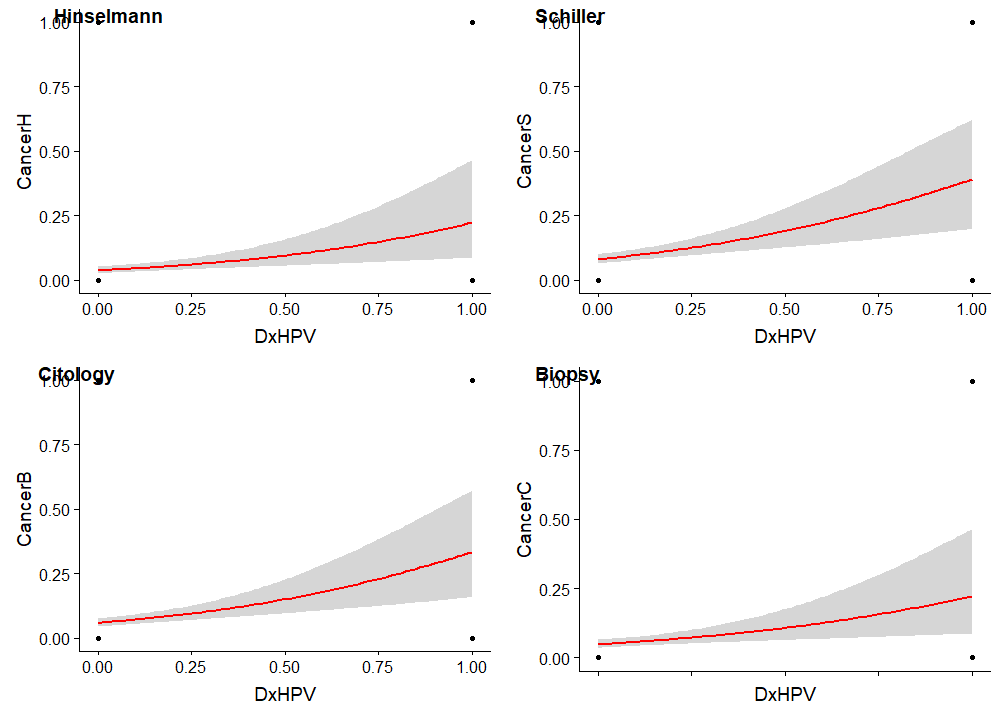
**Figure 1.** *Correlation matrix on all attributes. Size of the dot suggests the potency of correlation, as color of dots suggests direction of correlation. (Red dot for negative correlation and blue dot for positive correlation.)*

Following analysis of logistic regression of four risk factors suggested by previous research as well as the correlational analysis. HPV diagnosis, IUD usage, Smoking and HIV diagnosis were selected for logistic regression against cervical cancer diagnosis (Fig 2). Apparently, HPV diagnosis was the most significant contributor to cervical cancer, as P-Value of 3.51e-15 suggests a potent effect. Analysis regarding to IUD usage also returns to be significant, as a P-value of 0.00531 suggest significant contribution toward cervical cancer. One fact to note is that the estimated coefficient was very small (~1.5 logits), the effect is weak but significant. Smoking and HIV diagnosis both returned to be non-significant regarding to cervical cancer diagnosis.

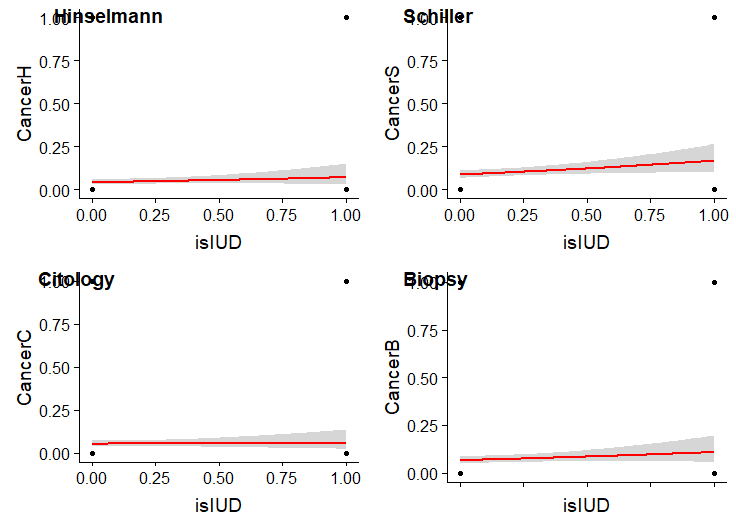


**Figure 2.** *Logistic models of HPV diagnosis, IUD usage, Smoking and HIV diagnosis against cancer. HPV diagnosis and IUD usage against cancer were both statistically significant (P = 3.51e-15 (\*\*\*\*) and P = 0.00531(\*\*)) while both other results were non-significant (P > 0.05).*

Final analysis was done on diagnostic method of cervical cancer (Hinselmann, Schiller, Citology and Biopsy) in relation to HPV diagnosis and IUD usage. Result of HPV against diagnostic methods shows that there were no significant differences in diagnostic method with patients with HPV, as all 4 analysis show similar P-values, with Cytology method show a lower P-value. Result of IUD usage against diagnostic methods shows that there much discrepancies among diagnostic methods of cervical cancers against IUD usage, as only Schiller’s test displayed significant relationship in this case (P-value = 0.022(\*)).



**Figure 3.** *Different diagnostic methods of cervical cancer against HPV diagnosis. All P-values are very significant (\*\*\*\*) except Citology with P = 0.00312 (\*\*).*



**Figure 4**. *Different diagnostic methods of cervical cancer against IUD usage. Only Schiller was significant (P = 0.022(\*)).*

Discussion

Among the course of the analysis, the correlational matrix has suggested some plausible correlations judging by the potency of the effect, which only leads to HPV infection.