2012•2013
FACULTY OF SCIENCES
Master of Statistics: Biostatistics

Masterproef

Prevalence of high risk HPV in female sex workers in Antwerp, Belgium

Promotor :
Prof. dr. Niel HENS

Promotor :
Dr. ELKE LEURIDAN

Paul Musingila

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Biostatistics

Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

universiteit hasselt
KNOWLEDGE IN ACTION

Universiteit Hasselt | Campus Hasselt | Martelarenlaan 42 | BE-3500 Hasselt
Universiteit Hasselt | Campus Diepenbeek | Agoralaan Gebouw D | BE-3590 Diepenbeek

universiteit hasselt | Maastricht University

2012•2013
FACULTY OF SCIENCES
Master of Statistics: Biostatistics

Masterproef

Prevalence of high risk HPV in female sex workers in Antwerp, Belgium

Promotor :
Prof. dr. Niel HENS

Promotor :
Dr. ELKE LEURIDAN

Paul Musingila
Master Thesis nominated to obtain the degree of Master of Statistics , specialization Biostatistics

universiteit hasselt | Maastricht University

Dedication

I dedicate this thesis report to my wife Elizabeth for her magnificent devotion to the family and unwavering support, my daughter Angel Mutanu, my mother Mumbe, my siblings Margaret, James, Joseph, Reginah, Mwikali and Mutinda. You all have prayed, loved, advised and cared for me throughout my life. Special dedication goes to my late father Musingila for his unwavering support and words of wisdom.

Acknowledgements

First I thank the Almighty God for giving me the strength, courage and good health throughout the two years of study at Hasselt University, Belgium. It is my utmost prayer and hope that God is going to hold on for his love, kindness and support in my further studies and challenges in future.

Secondly, I would like to appreciate my external supervisor Dr. Elke Leuriden, Faculty of Medicine and Health Sciences, University of Antwerp and the team of researchers at Ghapro, Antwerp for offering me the summer internship opportunity and providing data and research questions for this project. I wish to thank especially Alice Van Goethem and Liesbeth Aerts, Master students, Department of Medicine, University of Antwerp, for providing clinical explanations and clarifications to issues surrounding the study conduct, data and the study in general. My sincere gratitude goes to my internal supervisor Prof. dr. Niel Hens, of Center for Statistics, I-BioStat, Hasselt University, for providing advice, guided discussions, material support and guidelines on statistical methods used during data analysis and preparation of this report. I also appreciate the time spent reviewing my initial drafts, providing comments and points of improvement at all times.

Lastly, I am deeply indebted to the Vlaamse Interuniversitaire Raad (VLIR-UOS) for offering me this scholarship to pursue this valuable Masters of Statistics: Biostatistics program. I acknowledge the enormous support from my fellow students and lecturers for providing an intellectually stimulating environment during the program. I also acknowledge Martine Machiels, the Program Manager for her kindness, incredible advice, and other staff at the secretariat for their continued support throughout the program.

Paul Kithikii Musingila

Diepenbeek, Belgium

September 11, 2013

Abstract

Human papillomavirus is the most common STI and is a public health concern among women. HPV infection, especially persistent infection with HR-HPV types is strongly associated with development of cervical neoplasia lesions and cancer. FSW are at a great risk of exposure to multiple HPV infections due to the nature of their work coupled with multiple sex partners.

The objectives were to determine the prevalence and risk factors associated with high-risk human papillomavirus infection among female sex workers. To determine the prevalence of HPV genotypes and possible prevention strategies, prevalence of dysplastic cervical lesions among FSW and the HR-HPV types associated with normal or abnormal cervical cytology among FSW in Antwerp, Belgium.

Cross-sectional study of FSW (n=90) conveniently sampled from the large database (n=1471 examined between June 2006 and June 2010) from Ghapro, Antwerp. Information on sociodemographic characteristics, reproductive health, sexual health, drug use, STI was collected. Multiple logistic regression accounting and not accounting for survey design were applied.

Data from 1471 and 90 FSW were studied from the large and the small database respectively. High HR-HPV prevalence (39% large and 34% small database) was found among FSW in Antwerp, Belgium. Prevalence of HPV-16/18 was 14% and 12% in the large and small database respectively. Overall HPV prevalence was 43% (large database) and 41% (small database). Overall, HR-HPV and HPV-16/18 prevalence of dysplastic cervical lesions were significantly different. HR-HPV and HPV-16/18 genotypes in the large database are highly associated with cytology abnormality; with AGC, ASC-H and HSIL all having a prevalence of 100%, LSIL (86%), ASCUS (79%) and NILM (21%). In multiple logistic regression analysis accounting and not accounting for survey design, smoking status (OR 3.045, 95% CI: 1.129 to 8.213) and ever had STI (OR 0.234, 95% CI: 0.067 to 0.815) were statistically significantly associated with HR-HPV prevalence.

High-risk HPV prevalence in FSW in Antwerp, Belgium is high and places the FSW at a higher risk of developing cervical neoplasia. HR-HPV is associated with smoking and ever diagnosed with an STI. FSW ever diagnosed with an STI before the study conduct shows a protective effect against HR-HPV infection. Also HR-HPV infection is associated with increasing abnormal cytology among FSW. Hence HPV especially HR-HPV and cervical cancer prevention strategies focusing on this group are of great importance.

Key words

Cervical cancer, Genotype, High risk, Human Papillomavirus, Lasso logistic regression, Logistic regression, Post-stratification, Prevalence, Random forests, Survey logistic regression

Table of Contents

List of Tables

List of Figures

List of Acronyms

ACOG	American College of Obstetricians and Gynecologists
ADC	Adenosquamous-Carcinoma
AGC	Atypical Glandular Cells of Undetermined Significance
AIS	Adenocarcinoma *In Situ*
AML	Algemeen Medisch Laboratorium
ASC-H	Atypical Squamous Cells where a HSIL cannot be ruled out
ASC-US	Atypical Squamous Cells of Undetermined Significance
CART	Classification And Regression Trees
CHAID	CHi-squared Automatic Interaction Detection
CIN	Cervical Intraepithelial Neoplasia
DNA	Deoxyribonucleic Acid
FSW	Female Sex Workers
Ghapro	Health Care and Support to Prostitutes
HCW	Health Care Workers
HPV	Human Papillomavirus
HSIL	High Grade Squamous Intraepithelial Lesion
IARC	International Agency for Research on Cancer
LASSO	Least Absolute Shrinkage and Selection Operator
LLR	Lasso Logistic regression
LR	Logistic Regression
LSIL	Low Grade Squamous Intraepithelial Lesion
ML	Maximum Likelihood
NILM	Negative for Intraepithelial Lesion or Malignancy
OOB	Out-of-Bag
PCR	Polymerase Chain Reaction
PMLE	Pseudo-Maximum Likelihood Estimate
RF	Random Forests
SCC	Squamous Cell Carcinoma
SL	Survey Logistic
STI	Sexually Transmitted Infections
WHO	World Health Organization

1. Introduction

1.1 Background

Human papillomavirus (HPV) is a Deoxyribonucleic acid (DNA) virus from the papillomavirus family that is capable of infecting humans. HPV infection is initiated by infectious HPV particles introduced into the genital tract upon sexual intercourse reaching the basal cells of the squamous cell epithelium of the cervix uteri through micro-lesions (GENTICEL, 2013). HPV is the most common genital sexually transmitted infection (STI) in the world and persistent infection with high-risk HPV types is strongly associated with high-grade cervical intraepithelial neoplasia (CIN) and cancer (CIN3+) (Baay et al., 2004; Franco et al., 2001 & Depuydt et al., 2012). The relationship between infection with HPV and both cervical cancer and genital warts has been recognized for many years (ACOG, 2005). HPV is responsible for 99.7% of cervical cancer cases and an estimated 5% of all cancers worldwide (Moscicki, 2008). Over 100 HPV genotypes have been identified with only 15 shown to be associated with cervical cancer. These HPV genotypes are divided into low, intermediate and high risk categories. The low-risk HPV types predominantly cause benign warts with HPV 6 and 11 accounting for 90% of cases, while high-risk HPV types are associated with malignant disease evolving to neoplasia and cancer (Stanley, 2008).

Worldwide, cervical cancer is the second most common cancer in women and the second most common cause of death from cancer among women aged 14 to 44 years (Castellsague et al., 2007). Despite cervical cancer screening programs in Europe, the disease still remains the second leading cause of death in women aged 15 to 44 years (Arbyn et al., 2007 & Depuydt et al., 2003). Furthermore, 0.5 million new cases are diagnosed and about 0.275 million deaths occur per year worldwide. Annually in the United States, 11,000 new cervical cancer cases and 4,000 deaths occur, and while approximately 52,000 new cervical cancer cases and 27,000 deaths occur each year in Europe (WHO-European Region) (Arbyn et al., 2007; Jemal et al,. 2009 & Stanley, 2008). With the assumption that the risk does not change and no intervention takes place, it is expected that in 2020 the incidence of cervical cancer worldwide will increase by 40% compared to 2002 (Arbyn et al., 2012; Depuydt et al., 2003). Research in the last two decades has demonstrated that infection with certain HPV genotypes is a necessary factor in developing cervical cancer (Munoz et al., 2003 & Walboomers et al., 1999).

Forty HPV genotypes are known to be sexually transmitted and infect the anogenital region, however, 15 HR-HPV genotypes: 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 68, 73 and 82 are highly associated with cervical cancer development as evidenced in epidemiologic studies (Munoz et

al., 2003; de Villiers et al., 2004 & Cogliano et al., 2005). HPV 16 and 18 are the most common and jointly present as either single or multiple infections in at least 70-75% of squamous cell carcinoma (SCC), 40-60% of its precursors and 84% of adeno- and adenosquamous-carcinoma (ADC) worldwide (Castellsague et al., 2006 & Smith et al., 2007). The other HPV genotypes account for an additional 17% of invasive cervical cancers on a global base (Bosch et al., 2008). HPV 6 and 11 are the two most common low-risk HPV types that cause genital warts on the anogenital mucosae. Research has shown that women infected with high risk HPV types and who cannot clear the viral infection, can become persistent HPV carriers, and are more exposed to the development of neoplastic lesions. Other risk factors (many related to increased risk of infection) include early age at first coitus, early first pregnancy, history of previous miscarriage, oral contraceptive use, smoking, chronic inflammation, multiple sexual partners, immunosuppressive conditions (e.g. HIV/AIDS) and persistent infection with HR-HPV types.

The recognition of the strong causal relationship between persistent cervical infection with HR-HPV types and occurrence of cervical cancer has led to the development of a series of HPV DNA or RNA tests (Hybrid Capture II (Qiagen), Cobas 4800 (Roche), Cervista HR (Hologic); Real-Time HR-HPV test (Abbott). Detection of HR-HPV DNA is used in three clinical applications for women as: (a) triage test for equivocal or mildly abnormal cytology requiring referral for diagnosis and treatment; (b) follow-up test after treatment for high-grade CIN with local ablative or excisional therapy to predict failure or cure of treatment; and (c) primary screening test, individually or simultaneously with cervical cytology for early detection of HR-HPV infection, cervical pre-cancer and cancer (Arbyn et al., 2012). For women aged over 30 years, the practice is increasingly recommended to screen for lesions either by HPV testing or by simultaneous Papanicolaou (Pap) smear and HPV testing. Women infected with HR-HPV types but negative for intraepithelial lesion or malignancy (NILM) cytology, Atypical Squamous Cells of Undetermined Significance (ASC-US) or low-grade squamous intraepithelial lesions (LSIL), have a repeat cytology and HPV testing every 6 to 12 months (watchful waiting) or are referred to colposcopy for diagnosis of possible cervical intra-epithelial neoplasia (CIN) (GENTICEL, 2013 & CDC, 2013). Cervarix and Gardasil HPV vaccines are strong weapons in HPV prevention. These safe and effective vaccines are available to protect females and males against some of the most common HPV types and related health problems. Studies of HPV vaccines were conducted among young women 9–26 years of age with the primary objective to prevent CIN. HPV vaccines have been shown to be highly efficacious against CIN associated with HPV 16 and 18 in women who were not infected at the time of immunization. However, Moscicki (2008) resume that the efficacy of Gardasil and Cervarix vaccines in already infected women is marginal. Research has demonstrated that cross-protection is found for both

vaccines, and there may be HPV types present in the vaccines with which the women did not have any contact yet. There is no evidence of protection against HPV related diseases caused by HPV types of which the woman is already infected with (ACOG, 2010). However, there is evidence of protection from diseases caused by the remaining HPV genotypes (Future II study Group, 2007). Sexually active adolescents and young women can receive either the Cervarix or Gardasil HPV vaccine although the vaccine may be less effective to individuals who have been exposed to HPV before vaccination compared to HPV naive individuals at the time of vaccination (Munoz et al., 2010 & Paavonen et al., 2009). Case–control and cross-sectional studies have reported that the prevalence of cervical HPV infection, which has been strongly and consistently associated with CIN, increases with increasing numbers of sexual partners. More so, female sex workers (FSW) are at higher risk of HPV infection and cervical cancer due to their exposure to multiple sexual partners in their occupation and the resulting exposure to multiple HPV types (Brown et al., 2010).

Quadrivalent HPV vaccine has been found to offer protection against cervical cancer, cervical dysplasia, vulvar or vaginal dysplasia, and genital warts associated with HPV 6, 11, 16, and 18 (Munoz et al., 2010 & CDC, 2013). According to limited published data on vaccination of FSW with an HPV vaccine, vaccination with Gardasil vaccine early on in their sexual careers may induce protection against chronic infection by quadrivalent HPV genotypes (HPV 6/11/16/18 types) and provides 98% protection against CIN and cervical cancer caused by HPV 16 and 18 (Brown et al., 2011 & Future II study, 2007). This vaccination might protect against subsequent re-infection or reactivation of HPV (Collins et al., 2003). This development would enable FSW (in Antwerp) to be potentially eligible for a preventive vaccination. Thus an analysis of the current HPV epidemiology of FSW in Antwerp, Belgium is interesting for the evaluation of a possible implementation of HPV vaccination program with the Gardasil HPV vaccine or the upcoming multivalent therapeutic HPV (6/11/16/18/31/33/45/52/58) vaccine targeting most prevalent HR-HPV types (GENTICEL, 2013).

1.2 Study Objectives

The primary objective of the cross-sectional study is to determine the prevalence of HR-HPV in FSW in Ghapro, Antwerp.

The secondary objectives of the cross-sectional study were to:

(a) To determine which HPV genotypes were present during screening,

(b) To determine the prevalence of dysplastic cervical lesions among FSW,

(c) To determine the HPV genotypes associated with an abnormal and a normal smear of the cervix and

(d) To determine factors influencing HPV infection in FSW.

2. Data

2.1 Participants

In this cross-sectional study, the data consists of HPV screening measurements from a convenience sample of 90 subjects who were selected from a large database of 1471 FSW followed-up from June 2006 to June 2010 by Ghapro (www.ghapro.be), Antwerp, Belgium. Ghapro (Health Care and Support to Prostitutes) is an association providing free and anonymous support (both medical and social care) to women and men working in prostitution in the province of Antwerp and a part of Flemish Brabant.

2.2 Databases

2.2.1 Large Database

The large database consists of 2106 FSW who were followed-up from June 2006 to June 2010 in Ghapro Antwerp, Belgium. The following FSW were excluded from the final analysis: 38 had incomplete laboratory results, 585 with repeated measurements were excluded to reduce bias, 4 were immuno-compromised (HIV positive) and 8 who were working under other (non typable) sectors. Thus 1471 FSW who had HPV genotype and cytology results were available for the final analysis. By routine anamnesis, a limited number of risk factors (age, sex industry (sector) and the origin) were available per person.

2.2.2 Small Database

The convenience sample of 99 FSW aged 18-45 years, was obtained from the large database. All the FSW in the sample voluntarily consented to participate in the survey. Nine of the FSW were excluded because of either being immuno-compromised, aged over 45 years, or had incomplete data: other (non typable) sectors of work category, 1, unknown age at first coitus, 1, and unknown ever had STI, 1. Thus, data from 90 FSW was available for the final analysis. The survey study took place from June 2009 to June 2010. The study protocol was approved by the medical ethical board of Antwerp University. The survey questionnaire (Appendix A.2) was administered by health care workers (HCW) in Ghapro. All interviews were carried out in Dutch, French, Spanish or English. Information for specific risk factors was collected from the convenience sample. The risk factors used for this survey study were supported by literature (Table A.1). HPV genotypes 6, 11, 16, 18, 31, 33, 35, 39, 45, 51, 52, 53, 56, 58, 59, 66, 67 and 68 were tested for this study (coded: 0: Absent and 1: Present). The response variable of interest was infection with any high risk HPV genotype defined as

$$hr - HPV = \begin{cases} 1 & if\ FSW\ has\ high - risk\ HPV \\ 0 & Otherwise \end{cases}$$

2.3 Data collection

All FSW received a vaginal smear with a monolayer technique of BD SurePath® during consultations at Ghapro. Cervical cells were collected by means of a CervexBrush® (Rovers, Oss, The Dutch). Immediately after extraction, the head portion of the vaginal swab was directly stored in a medium based on alcohol (AutoCyte®, Tripath Imaging Inc., Burlington, NC, USA), and the samples were sent to the laboratory.

2.3.1 Laboratory Processing

The cervical sample processing was done in the Laboratory of Clinical Pathology, campus Riatol in Antwerp. In the laboratory, a thin layer is composed based on the cervical smear. This preparation is then mechanically analysed by a BD FocalPointTM® device that classifies the samples into five categories. Preparations in category 1 have the best chance of abnormal cells. On the other hand, the probability of finding abnormal cells in the compositions of Category 5 is very minimal (Schmitt et al., 2012). The cervix cytology is then evaluated according to the Bethesda classification. The possible categories are: NILM, ASC-US, atypical glandular cells of undetermined significance (AGC), LSIL, high grade squamous intraepithelial lesion (HSIL), CIN and atypical squamous cells which cannot be excluded from a high-grade lesion (ASC-H). With the remaining cells in suspension, HPV genotypes are identified by means of a real-time polymerase chain reaction (PCR) analysis. This is done by combining DNA, which was extracted from the cervical cells, with different types of specific viral sequences (Schmitt et al., 2012 & Petter et al., 2000). The sample is examined as described by Micalessi et al (2012) for the presence of 18 different HPV genotypes (HPV6 E6, HPV11 E6, HPV16 E7, HPV18 E7, HPV31 E6, HPV33 E6, HPV35 E6, HPV39 E7, HPV45 E7 (Depuydt et all, 2012), HPV51 E7, HPV52 E7, HPV53 E6, HPV56 E7 (Depuydt et all, 2012), HPV58 E7, HPV59 E7, HPV66 E6, HPV67 and HPV68 E7) and divided into the 3 groups defined by AML.

2.3.2 Human Papillomavirus Typing

2.3.2.1 High Risk Human Papillomavirus Genotypes

The hr-HPV genotypes have a high carcinogenic potential and are referred to as oncogenic HPV types. According to the International Agency for Research on Cancer (IARC), there are 12 hr-HPV genotypes with oncogenic properties, namely types 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58 and 59. The Algemeen Medisch Laboratorium (AML) laboratory includes in this group HPV genotypes 66 and 68 (Depuydt et al., 2003).

2.3.2.2 Intermediate Risk Human Papillomavirus Genotypes

The IARC describes intermediate risk HPV genotypes: 26, 53, 66, 67, 68, 70, 73 and 82. The AML laboratory does not include HPV genotypes 66 and 68 in this group.

2.3.2.3 Low-Risk Human Papillomavirus Genotypes

The low risk HPV types have low carcinogenic potential; they can cause benign or low grade epithelium in the cervix and condylomata accuminata (IARC). IARC describes HPV types 6, 7, 11, 13, 30, 32, 34, 40, 42, 43, 44, 54, 61, 62, 69, 71, 72, 74, 81, 83, 84, 85, 86, 87, 89, 90, 91, 97, 102, 106, 114 as low risk (Depuydt et al., 2006). Only the types 6 and 11 were tested by the AML laboratory (Petter et al., 2000 & Depuydt et al., 2012).

In this analysis, the classification of the AML laboratory was used to determine which HPV genotype were to be considered as low, intermediate or high risk types.

3. Statistical Methodology

3.1 Exploratory Data Analysis

Exploratory data analysis is a fundamental tool carried out to gain insight into the data. The tools considered in this report include descriptive statistics, cross-tabulations, univariate logistic regression analysis and graphical representations. The aim was to study the relationship between the predictor variables and the outcome of interest. These relationships were further investigated using formal statistical methods and models.

3.2 Tree-Based Methods for Variable Selection

In data mining, exploratory data analysis employs wide variety of graphical and statistical methods in order to identify the most relevant variables and determine the complexity and/or the general nature of models that can be taken into account in model building. Recursive partitioning is a fundamental tool as it helps in exploring the structure of a data set. It also develops easy ways to visualize decision rules for predicting either a categorical or continuous outcome. Tree-based methods partition the feature space into a set of rectangles, and then fit a simple model in each one. They are conceptually simple yet powerful (Hastie et al., 2009). In this study, Classification and Regression Trees (CART) modeling for tree-based methods: classification trees and random forests were reviewed. However, a random forest was considered for this analysis over classification trees because of its robustness in variable selection.

Random Forests

Random forests (RF) as Breiman (2001) stated is a machine learning algorithm that fits many classification or regression tree models to random subsets of the input data and uses the combined result (the forest) for prediction. RF are particularly well-suited to small sample and many predictor variable problems even in the presence of complex interactions (Strobl et al. 2009b). In this way, RF are able to better examine the contribution and behaviour that each predictor variable has, even when one predictor's effect would usually be over-shadowed by more significant competitors (www.stanford.edu, 2013). RF use out-of-bag (OOB) samples to measure prediction accuracy. RF provides variable importance measure through permutation test, which could be of interest in reducing the number of variables in a statistical analysis. RFs provide proximity measures between observations resulting in outlier detection. RF produces better predictions than the results of one classification tree because they are robust (Strobl et al. 2008). While using random forests, variables are considered informative and important if their variable importance measure is above the absolute

value of the lowest negative-scoring variable (Strobl et al., (2009a, 2009b)). "The rationale for this rule of thumb is that the importance of irrelevant variables varies randomly around zero" (Strobl et al., (2009a, 2009b)). The results from random forests and conditional variable importance were verified via multiple random forest runs starting with different seeds and sufficiently large *ntree* values to insure the robustness and stability of results (Strobl et al., (2009a, 2009b)). The final RF models are obtained by aggregating number of trees as base classifiers, with 5 variables tried at each split so as to obtain robust estimates. Where, this choice of number of trees provides robust estimates since the OOB error rate must have stabilized. The plot of variables ordered by the variable importance measure is used in order to enhance our understanding regarding the relationship between predictor variables and the response. The function *cforest* from the *party* R package (Hothorn et al., 2006) was used over the *randomForest* R package (Liaw & Wiener, 2002). This was because the *party* package is robust in unbiased variable selection, correlated predictors conditional permutation importance measure and variable importance measure for predictor variables of different types (categorical or continuous), which is key to reliable prediction and interpretability in both individual trees and forests (www.stanford.edu, 2013).

3.3 Logistic Regression

Logistic regression (LR) is used to describe the association between a categorical response (binary, ordinal or nominal) variable and one or a set of predictor (continuous or categorical) variables. However, each type of categorical variables requires different techniques to model its relationship with the predictor variables. LR models play an important role in medical research for binary response data, providing data analytic tools for understanding the importance of different predictors. Moreso, LR can be used to classify new individuals and especially where one is interested in estimating the class probabilities, for use in risk screening. The logistic model (Cox, 1970 & Agresti, 2002) is often used when the outcome Y is binary. In the logistic model, the mean response $p(\boldsymbol{X}) = \Pr[Y = 1|\, \boldsymbol{X} = \boldsymbol{x}] = 1 - \Pr[Y = 0|\, \boldsymbol{X} = \boldsymbol{x}]$ for an individual with covariate vector $\boldsymbol{X}$ is

$$p(\boldsymbol{X}) = \Pr[Y = 1|\boldsymbol{X}] = \frac{exp(\boldsymbol{\beta X})}{1 + \exp(\boldsymbol{\beta X})}$$

This is mathematically equivalent to the log odds, called the logit, with the linear relationship

$$logit\ [p(\boldsymbol{X})] = \log\left(\frac{p(\boldsymbol{X})}{1 - p(\boldsymbol{X})}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \ldots + \beta_p x_p$$

where β_j is the maximum likelihood parameter estimate, $\boldsymbol{X}$ is the vector of predictors

Logistic regression employs various link functions such as logit, probit and complementary log-log. The link function for logistic regression is given by

$$g\left(p(\boldsymbol{X})\right) = \log\left(\frac{p(\boldsymbol{X})}{1-p(\boldsymbol{X})}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \ldots + \beta_p x_p$$

Logistic regression models are usually fitted by maximum likelihood, using the conditional likelihood of Y given $\boldsymbol{X}$. The parameters of $Pr(Y = 1|\boldsymbol{X})$ are fitted by maximizing the conditional likelihood – the logistic probabilities, $Pr(Y = 1|\boldsymbol{X})$ (Hastie et al., 2009). The log-likelihood for N observations can then be written as

$$l(\beta) = \sum_{i=1}^{N} \{y_i \beta^T x_i - \log(1 - e^{\beta^T x_i})\}$$

Assessing individually the relationship between the response variable and the possible important predictors provides a preliminary idea of the king of relationship which might be non-linear, linear or quadratic. This approach is attained graphically by plotting the univariate logistic regression fitted probabilities against the predictor variable (Agresti, 2002). Hosmer and Lemeshow (2000) and Agresti (2002) also propose that during univariate logistic regression, covariates with $p < 0.25$ should be considered as possible candidates for multiple logistic regression modelling.

Over-dispersion in LR can be as a result of missing covariates and/or interaction terms, negligence of non-linear effects, wrong link function, existence of large outliers and binary data or small cell values. However, these challenges can be excluded through EDA and regression diagnostics. Hence, over-dispersion can be explained by variation among success probabilities or correlation between the binary responses. Since our response variable is a binary with positive response being that the FSW has at least one HR-HPV infection, a multiple logistic regression was plausible for describing the relationship between the predictors and probability of at least one HR-HPV infection. Deviance and Hosmer–Lemeshow test were used to assess the final model goodness of fit.

3.4 Survey Logistic Regression

Survey design building blocks for probability samples, that is, simple random sampling, stratification, and multistage cluster sampling, were all developed with the goal of minimizing the survey cost while controlling the uncertainty associated with key estimates (Pfeffermann & Rao, 2009). Applying classical statistical methods to survey data without accounting for survey design features can lead to erroneous inferences to the finite sample due to serious underestimation of standard errors of parameter estimates and associated confidence interval coverage rates, as well as inflated test levels and misleading model diagnostics (Agresti, 2002). For correct statistical inferences of survey samples, the SURVEYLOGISTIC (SL) procedure (SAS Institute, 2013) which provides logistic regression analysis of survey data was considered. This procedure incorporates

complex survey sample designs, including designs with stratification, post-stratification, clustering, and unequal weighting. The link function for LR is the logit (Nelder and Wedderburn, 1972) defined

$$g\big(\pi(x)\big) = logit\big(\pi(x)\big) = ln\langle\frac{\pi(x)}{1-\pi(x)}\rangle = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \ldots + \beta_p x_p$$

The SL procedure uses various link functions: the cumulative logit function (clogit), the generalized logit function (glogit), the probit function (probit), and the complementary log-log function (clogolog). The regression parameters are estimated by maximum likelihood with either Newton-Raphson or Fisher Scoring algorithm (Agresti, 2002) by maximizing the following weighted pseudo-likelihood function:

$$PL(\boldsymbol{\beta}|X) = \prod_{i=1}^{n} \{\pi(x_i)^{y_i}.[1-\pi(x_i)]^{1-y_i}\}^{w_i}$$

where $\boldsymbol{\pi}(\boldsymbol{x_i}) = \exp(\boldsymbol{x_i}\boldsymbol{\beta})/[1+\exp(\boldsymbol{x_i}\boldsymbol{\beta})]$, w_i sampling weights, X vector of covariates and y_i the original binary response. This model provides appropriate parameter estimates and standard errors (Clogg and Eliason, 1987). Variances of the regression parameters and odds ratios are computed by using either the Taylor series (linearization) method or replication (resampling) methods to estimate sampling errors of estimators based on complex sample designs (Binder, 1983; Särndal et al., 1992; Wolter, 2007 & Rao et al., 1992). This method was considered for this analysis to account for survey design which in turn enables generalization of the survey inferences.

Weighting in Survey Data

Survey design can be done prior to the conduct of the survey or after the survey has already been done. Post-stratification, raking ratio estimation, generalized regression estimation, and calibration are forms of post-survey weight adjustments that may be employed to improve the precision and accuracy of survey estimates (Pfeffermann and Rao, 2009). Post-stratification improves the quality of sample survey estimates by incorporating known information on the full survey population borrowing strength from data sources external to the sample. For example, post-stratification make use of auxiliary population information or adjusts for nonresponse in some way. Thus, post-stratification adjusts the sampling weights so that the estimated population group sizes are correct, as they would be in stratified sampling. The criteria used to select variables for forming post-strata include: (1) variables such as age, gender, and region that define post-strata for which accurate population control totals are available from external sources; (2) post-stratification variables that are highly correlated with key survey variables; and (3) variables that may be predictive of non-coverage in the sample frame (Chambers and Skinner, 2003 & Molenberghs, 2012). Based on the advice from the Doctors in charge of the survey study, age group variable with six categories was used as the post-stratification variable. In this survey study, post-stratification was employed by sampling n_l

individuals from population stratum l $(l = 1, .., L)$ containing N_l individuals. The post-stratification weight applied to each respondent in this survey study was calculated as

$$W_{ps,i} = \frac{1/p_{s,l}}{\sum_l 1/p_{s,l}} n_{s,l}$$

where $p_{s,i} = n_{s,l}/N_{p,l}$, $n_{s,l}$ sample observations in strata l and $N_{p,l}$ population observations in stratum l.

3.5 Lasso Logistic Regression

The Lasso (least absolute shrinkage and selection operator) has become a popular shrinkage and variable selection method for regression models (Tibshirani, 1996), though originally was proposed for ordinary least squares regression models. The reason for its popularity is because it doesn't suffer much from high variability as exhibited by subset selection approach. The Lasso logistic regression (LLR) model is defined as

$$logit(\pi_i) = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} \quad (i = 1, .., n \; and \; j = 1, \dots, p)$$

Where $\pi_i = P(Y_i = 1|\boldsymbol{X}_i', \boldsymbol{\beta}) = 1/(1 + \exp(-\boldsymbol{\beta}^T \boldsymbol{X}_i'))$, β_j is the regression parameter, $\boldsymbol{X}_i' = (1, \boldsymbol{X}_i)^T$ to include intercept parameter β_0, $\boldsymbol{\beta} = (\beta_0, \beta_1, .., \beta_p)^T$) vector of parameter estimates and $\boldsymbol{X}_i$ is the vector of predictors. The two vectors $\boldsymbol{X}_i'$ and $\boldsymbol{\beta}$ are of length p+1.

The lasso estimator is then defined as

$$\hat{\beta}_\lambda = \underset{\beta}{arg\,min} \{ \sum_i^n \left(y_i - \left(\beta_0 + \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 \right) + \lambda \sum_{j=1}^{p} |\beta_j| \}; \quad \lambda \geq 0$$

Where $Y \in \mathbb{R}^{\boldsymbol{n}}$, an $n \times p$ design matrix $\boldsymbol{X}$, $\boldsymbol{\beta}$ is the vector of parameters, $\sum_{j=1}^{p} |\beta_j|$ is the l_1 lasso penalty, λ is the *tuning parameter* which may be selected by the user or calculated via methods such as cross-validation, generalized cross-validation and a variant of Stein's unbiased estimate of risk (Tibshirani, 1996). Cross-validated log-likelihood was used in this study to compare the predictive ability of different values of λ. λ is chosen in a way to minimize the estimate of the prediction error. Tibshirani (1996) notes that $|\beta_j|$ is proportional to the negative log-density of a Laplace (double-exponential) distribution with mean zero and the probability density function of a Laplace-distributed random variable β_j with mean μ and variance $2/\lambda^2$ is

$$f(\beta_j) = \frac{\lambda}{2} \exp(-\lambda|\beta_j - \mu)$$

In LLR, dummies were formed for the categorical predictor variables by ensuring that input predictors are coded as factors in the data set which in turn uses R functions *contr.none* for unordered and *contr.diff* for ordered categorical variables from *penalized* R package (Goeman, 2009). LLR is robust to numerical instability in the parameter estimations (Caster, 2007) and hence

its use in this study for variable selection. Lasso regression is robust where there is large number of covariates in the model structure like in this study. Lasso is a continuous process for subset selection and identifies predictors which are most strongly associated with the outcome of interest. Some coefficients which have minimal association with the outcome variable are shrunk towards zero (Hastie et al., 2009). The Lasso encourages sparseness, setting most small coefficients to zero, due to the penalty function's sharp peak at zero. As λ increases, all the coefficients shrink, each one ultimately becoming zero (Hastie et al., 2009). The predictor variables which are below the chosen λ are set to zero and hence are considered as not strongly related to the response. LLR is implemented in R packages such as *penalized* (Goeman, 2009), *glmnet* (Friedman et al., 2010), *grplasso* (Meier et al., 2008). However, for this analysis *penalized* R package was considered.

3.6 Boosting

Boosting is a powerful machine learning meta-algorithm for reducing bias in supervised learning. This method was originally designed for classification problems and later adapted to regression problems (Hastie et al., 2009). This method consists of iteratively learning weak classifiers with respect to a distribution and adding them to a final strong classifier producing a powerful committee. For a binary output variable, Y, and a vector of predictor variables $\boldsymbol{X}$, a classifier $G(X)$ produces a prediction taking one of the two response values. The error rate on the training sample is

$$\overline{err} = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq G(x_i))$$

and the expected error rate on future predictions is $E_{XY}I(Y \neq G(X))$ (Hastie et al., 2009). A weak classifier is one whose error rate is only slightly better than random guessing (Hastie et al., 2009). Boosting sequentially apply the weak classification algorithm repeatedly modifying the data, producing weak classifiers $G_m(x)\ m = 1,2,\ldots,M$. This algorithm initializes the observation weights, fits a classifier to the learning data set using weights and computes the misclassification error (typically <0.5 (0.5 for random guessing) to terminate the classifier generation process) (Hens, 2013). Reweighting occurs after a weak learner is added so that the misclassified ones by the classifier gain weight and those that are classified correctly lose weight. Boosting generates multiple models or classifiers for prediction or classification, and also to derive weights (initial weight, $w_i = 1/n$) to combine the predictions from those models into a single prediction or predicted classification through a weighted majority vote to produce the final prediction (www.obgyn.cam.ac.uk, 2013; Hastie et al., 2009). The final prediction is given by

$$G(x) = sign\left(\sum_{m=1}^{M} \alpha_m\, G_m(x)\right), \quad \alpha_m > 1$$

here $\alpha_1, .., \alpha_M$ are calculated by the boosting algorithm and weight the contribution of each respective $G_m(x)$, thus, give higher influence to the more accurate classifiers in the sequence (Hastie et al., 2009). Boosting fits a 'basis functions' model with a forward stagewise approximation with exponential loss (Hastie et al., 2009; Hens, 2013). Boosting algorithms result in prediction rules that have the same interpretation as common statistical model fits which is a key merit over machine learning techniques such as random forests (Breiman, 2001) that result in non-interpretable 'black-box' predictions. Therefore, boosting results from test data set were used to compare predictions obtained from ordinary LR. R packages AdaB*oost* (Freund & Schapire, 1997) and *adabag* (Friedman et al., 200) are used in boosting. For this analysis *penalized* R package was considered.

3.7 Software

Data manipulation, exploratory data analysis and statistical analysis were performed using R (Version 3.0.1) and Statistical Analysis Software (SAS, version 9.3). All statistical tests were conducted at $\alpha = 0.05$ level of significance. Also where necessary, 95% confidence intervals were computed.

4. Results

4.1 Exploratory Data Analysis

4.1.1 Small Database

The sample had 90 FSW; 74 (82%) originated from Europe, 12 (13%) from Africa (Sub-Sahara), 3 (3%) from America and 1 (1%) from South-East Asia. Infection with HR-HPV types was not associated with the region of origin, (p=0.7860). Majority of FSW 26 (29%) were aged 26-30 years, 21 (23%) were aged 21-25 years, 14 (13%) were aged 31-35 years, 14 (13%) were older than 40 years, 10 (11%) were aged 36-40 years and younger than 21 years 7 (8%). Majority of the FSW were ≤30 years 54 (60%). Infection with HR-HPV types was not associated with age groups (p=0.6508). Forty five (50%) FSW were working from windows (red light district), 21 (23%) from private house/massage parlors/home/SM-studio and 13 (14%) from bars/bars and windows. Infection with HR-HPV types was not associated with current sector of work (p=0.3405). Twenty two FSW (24%) had a past self-reported history of ever had **STI** at the survey study point. That is, ever diagnosed with chlamydia trachomatis (50%), Neisseria gonorrhoea (27%), herpes simplex (9%), crabs (9%) or trichomonas vaginalis (5%). There was borderline association between infection with HR-HPV types and FSW who ever had STI (STI prevalence = 24%) and those who never had STI (76%) (p=0.0648). Most of the FSW 57 (63%) started their sexual debut aged between 16-20 years, 26 (29%) younger than 16 years and 7 (8%) older than 20 years. HR-HPV infection was not associated with age at first coitus (p=0.1887). Over half of the FSW, 58 (64%), had no children, 16 (18%) had one child, 11 (12%) had more than one child and 5 (6%) had unknown number of children. Infection with HR-HPV types was not associated with parity (p=0.8891). A high proportion of FSW 44 (49%) used hormonal contraceptives, 34 (39%) condoms only, 8 (9%) used IUD/TL and 4 (4%) reported unknown contraceptive method. Infection with HR-HPV types was not associated with contraceptive use (p=0.1117). **Smoking** (p=0.0329) and **condom use** for each sex technique with customer (p=0.0320) were both associated with HR-HPV infection. Infection with HR-HPV types was not associated with time in prostitution (years), cervical microbiology, number of private partners in the last 12 months, condom use with private partners, the types of STI diagnosed in the past, number of customers per day, infection with gonorrhoea at screening, infection with chlamydia at screening, drug use in the past and present (Table 4-1).

4.1.2 Large Database

In the large database, a total of 1471 FSW were followed up; 854 (58%) originated from Europe, 375 (25%) from Africa (Sub-Sahara), 92 (6%) from South-East Asia, 86 (6%) from America, 40 (3%)

from Eastern Mediterranean, 18 (1%) from unknown origin and the rest from West Pacific Ocean. Infection with HR-HPV types was associated with region of origin (p=0.0004). Majority of the FSW 432 (29%) were aged 21-25 years, 370 (25%) were aged 26-30 years, 213 (14%) were aged 31-35 years, 168 (11%) were younger than 21 years, 156 (11%) were older than 40 years and 132 (9%) were aged 36-40 years. Majority of the young FSW were ≤ 30 years (66%) accounting for 73% of HR-HPV prevalence. HR-HPV infection was association with age groups (p<0.0001). Most of the FSW 647 (44%) were working from windows (red light district), 454 (31%) were working from private house/massage parlours/home/SM-studio, 183 (12%) were working from street or African café, 151 (10%) from bar/bar + window, 25 (2%) from escort and the rest from unprecedented sectors. Infection with HR-HPV was associated with current sector of work (p= 0.0048) (Table 4-1).

Table 4-1: Socio-demographic and other risk factor characteristics of FSW According to HR-HPV Infection for the Large and Small Databases

	Large Database					Small database				
	Total (%)	**Pos (%)**	**OR(95% CI)**	**p-Value**	**Trend**	**Total (%)**	**Pos (%)**	**OR(95% CI)**	**p-value**	**Trend**
Variables	**1471(100)**	**566(38)**				**90(100)**	**31(34)**			
Age Group (years)										
<21	168(11)	106(19)	1	<0.0001	<0.0001	7(8)	4(13)	1	0.6726	0.066
21-25	432(29)	183(32)	0.430(0.298,0.621)			21(23)	9(29)	0.563(0.100,3.168)		
26-30	370(25)	122(22)	0.288(0.197,0.421)			26(29)	8(26)	0.333(0.060,1.849)		
31-35	213(14)	71(13)	0.293(0.192,0.447)			13(14)	4(13)	0.333(0.050,2.239)		
36-40	132(9)	43(8)	0.283(0.175,0.457)			10(11)	3(10)	0.321(0.043,2.417)		
> 40	156(11)	41(7)	0.209(0.130,0.335)			13(14)	3(10)	0.225(0.031,1.623)		
Region of Origin										
Uknown	18(1)	8(1)	1.742(0.671,4.527)	0.0004					0.768	
Africa (Sub-Sahara)	375(25)	118(21)	1			12(13)	3(10)	1		
America	86(6)	27(5)	0.997(0.602,1.651)			3(3)	1(4)	1.500(0.098,23.069)		
South-East Asia	92(6)	33(6)	1.218(0.755,1.966)			1(1)	0(0)	*		
Europe	854(58)	370(65)	1.665(1.288,2.152)			74(82)	27(87)	1.723(0.429,6.917)		
Eastern Mediterranean	40(3)	8(1)	0.544(0.243,1.218)							
West Pacific Ocean	6(0)	2(0)	*							
Current work sector										
Bar, Bar + Window	151(10)	65(11)	1	0.0048		13(14)	6(19)	1	0.3405	
Private house, Massage parlors, Home, SM-studio	454(31)	199(35)	1.033(0.712,1.497)			21(23)	10(32)	1.061(0.265,4.243)		
Window (red light district)	647(44)	234(41)	0.750(0.523,1.074)			45(50)	12(39)	0.424(0.119,1.518)		
Escort	25(2)	12(2)	1.221(0.523,2.852)			6(7)	1(3)	0.233(0.021,2.593)		
Street/African women Cafe	183(12)	53(9)	0.539(0.343,0.849)			5(6)	2(6)	0.778(0.096,6.322)		
Unprecedented	11(0)	3(0)	0.496(0.127,1.944)							
Number of Children (Parity)										
0						58(64)	21(68)	1	0.8891	0.2798
1						16(18)	5(16)	0.801(0.245,2.619)		

>1	11(12)	4(13)	1.007(0.264,3.845)		
Unknown	5(6)	1(3)	0.440(0.046,4.203)		
Smoking status					
No	40(44)	9(29)	1	0.0329	
Yes	50(56)	22(71)	2.706 (1.069,6.851)		
Contraception use					
Only condom	34(38)	7(23)	1	0.1117	
Hormonal Contraceptive	44(49)	20(65)	3.214(1.157,8.926)		
IUD/TL	8(9)	2(6)	1.286(0.212,7.804)		
Unknown	4(4)	2(6)	3.857(0.459,32.424)		
Condom use for each technique with customers					
Always	83(92)	26(84)	1	0.032	
Unknown	7(8)	5(16)	5.481(0.997,30.127)		
Age of first intercourse (years)					
<16	26(29)	11(35)	1	0.1887	0.5071
16-20	57(63)	16(52)	0.532(0.202,1.403)		
>20	7(8)	4(13)	1.817(0.336,9.820)		
Unknown					
Number of private partners (in the last 12 months)					
0	12(13)	4(13)	1	0.9933	0.503
1	52(58)	18(58)	1.059(0.280,4.001)		
2 to 5	21(23)	7(23)	1.000(0.222,4.502)		
>5	5(6)	2(6)	1.333(0.155,11.498)		
Condom use with private partners					
No	55(61)	20(65)	1	0.6311	
Yes	31(34)	9(29)	0.716(0.277,1.852)		
Unknown	4(4)	2(6)	1.750(0.229,13.398)		
STI in the past					
No	68(76)	27(87)	1	0.0648	
Yes	22(24)	4(13)	0.337(0.103,1.106)		

Which STI?					
Neisseria Gonorrhoea	6(27)	0(0)	*	0.2846	
Chlamydia trachomatis	11(50)	2(50)			
Herpes Simplex	2(9)	1(25)			
Crabs	2(9)	1(25)			
Trichomonas Vaginalis	1(5)	0(0)	*		
Time in prostitution (Years)					
<2	17(19)	8(26)	1	0.1038	0.4912
2	16(18)	8(26)	1.125(0.287,4.412)		
3 to 4	22(24)	6(19)	0.422(0.111,1.606)		
5 to 9	17(19)	2(6)	0.150(0.026,0.868)		
>9	11(12)	3(10)	0.422(0.082,2.160)		
Unknown	7(8)	4(13)	1.500(0.254,8.844)		
Number of customers per day					
0 to 2	18(20)	7(23)	1	0.5735	0.5094
3 to 5	46(51)	17(55)	0.921(0.300,2.826)		
6 to 10	17(19)	3(10)	0.337(0.070,1.612)		
>10	4(4)	2(6)	1.571(0.178,13.860)		
Unknown	5(6)	2(6)	1.048(0.138,7.934)		
Drug use in the past					
No	53(59)	17(55)	1	0.8396	
Yes	34(38)	13(42)	1.311(0.533,3.226)		
Unknown	3(3)	1(3)	1.059(0.090,12.503)		
Drug use in the present					
No	76(84)	26(84)	1	0.9893	
Yes	11(12)	4(13)	1.099(0.295,4.100)		
Unknown	3(3)	1(3)	0.962(0.083,11.107)		
Gonorrhoea at screening					
Negative	87(97)	31(100)	1	0.2016	
Unknown	3(3)	0(0)	*		

Chlamydia at screening				
Negative	81(90)	28(90)	1	0.8927
Positive	5(6)	2(6)	1.262(0.199,8.000)	
Unknown	4(4)	1(3)	0.631(0.063,6.351)	
Cervix microbiology				
Negative	64(71)	18(58)	1	0.141
Gardnerella	12(13)	6(19)	2.555(0.728,8.972)	
Candida, Actinomyces, Trichomonas	14(16)	7(23)	2.555(0.785,8.324)	

** OR (95% CI) values not indicated because of sparseness*

4.2 High-Risk Human Papillomavirus Prevalence

High HR-HPV prevalence of 34% and 38% were observed from the small and large database respectively. This was an indication of high sexual exposure to HR-HPV types despite the high reported consistent condom use (92%) in the survey study. There was statistical significance difference for overall HR-HPV prevalence in the current sectors of work (p=0.0004) and regions of origin (p=0.0048) in the large database. We observed a significant decreasing trend in HR-HPV prevalence with increasing age in years in the large database (Table 4-1).

4.3 Prevalence of HPV Genotypes at Screening

From the survey study, the overall HPV prevalence was 41%, HR-HPV genotypes was 34%, HPV-16/18 was12% and multiple HR-HPV genotypes 14%. While from the large database, the prevalence of overall HPV genotypes was 43%, HR-HPV genotypes was 38% with HPV-16/18 accounting for 36% and multiple HR-HPV genotypes accounting for 41% of the HR-HPV infection (Table 4-2). There was no statistical significant difference between HR-HPV prevalence between the large (38%) and small (34%) database (χ_1^2=0.4257, p=0.5141) Table 4-2. Also we found that there was no statistical significant difference between multiple HR-HPV prevalence between the large (16%) and small (14%) database (χ_1^2=0.0234, p=0.8785). The dominant HR-HPV prevalence from the survey study was from genotypes HPV-31 (14%), HPV-16 (9%) and HPV-59 (6%). From the large database, the dominant HR-HPV types prevalence was from genotypes HPV-16 (11%), HPV-31 (7%) and HPV-52 (6%) (Table 4-2).

Table 4-2: HPV Genotype Prevalence and Distribution among Female Sex Workers

HPV Subtype	Large Database		Small Database		
	HPV	HPV Positive	HPV	HPV Positive	Test of
	Total (%) 631(43)	No (%) 631(100)	Total (%) 37(41)	No (%) 37(100)	Prevalence Difference p-value
Low Risk					
HPV6	32(2)	32(5)	0(0)	0(0)	0.2946
HPV11	7(1)	7(1)	1(1)	1(3)	0.9530
High-Risk					
HPV16	159(11)	159(25)	8(9)	8(22)	0.6918
HPV16/18	203(14)	203(32)	11(12)	11(30)	0.7913
HPV18	51(3)	51(8)	4(4)	4(11)	0.8464
HPV31	109(7)	109(17)	12(13)	12(32)	0.0662
HPV33	31(2)	31(5)	1(1)	1(3)	0.7915
HPV35	55(4)	55(9)	4(4)	4(11)	0.9553
HPV39	80(5)	80(13)	2(2)	2(5)	0.2782
HPV45	25(2)	25(4)	2(2)	2(5)	1.0000
HPV51	77(5)	77(12)	3(3)	3(8)	0.5838
HPV52	91(6)	91(14)	3(3)	3(8)	0.3809
HPV56	60(4)	60(10)	3(3)	3(8)	0.9418
HPV58	50(3)	50(8)	3(3)	3(8)	1.0000
HPV59	37(3)	37(6)	6(7)	6(16)	0.0450
HPV66	57(4)	57(9)	3(3)	3(8)	1.0000
HPV68	8(1)	8(1)	0(0)	0(0)	1.0000
Intermediate-Risk					
HPV53	66(5)	66(10)	4(4)	4(11)	1.0000
HPV67	62(4)	62(10)	3(3)	3(8)	0.8929
HPV	631(43)	631(100)	37(41)	37(100)	0.6676
HR-HPV	566(39)	566(90)	31(34)	31(84)	0.5141
Multiple HR-HPV	230(16)	230(36)	13(14)	13(35)	0.8785
HRIR-HPV			35(39)	35(95)	

4.4 Prevalence of Dysplastic Cervical Lesions

Distribution of the cervical cytology results using Bethesda classification in relation with their overall HPV, HR-HPV and HPV types 16/18 prevalence's were studied. Majority of the FSW in the survey study 72 (80%) tested NILM and tested positive for HR-HPV types 16, 18, 31, 35, 39, 45, 51, 52, 56, 58 and 59, accounting for 21% of all HR-HPV prevalence. ASCUS 9 (10%) with prevalent oncogenic HPV types 16, 18, 31, 33, 56 and 58 detected accounting for 78% of all HR-HPV prevalence. LSIL 9 (10%) with HR-HPV types 16, 18, 31, 35, 39, 45, 51, 52 and 59 detected accounting for 100% of all HR-HPV prevalence. There was no diagnosis of HSIL, ASC-H, or AGC among the FSW in the survey study. Overall in the survey study, HR-HPV and HPV-16/18 prevalence of dysplastic cervical lesions were statistically significantly different in the two databases. An increasing trend in prevalence was observed with increasing cervical cytology abnormality for FSW with HR-HPV infection ($p<0.001$) and HPV 16/18 ($p<0.001$). However, from

the large database, we observe that HR-HPV and HPV-16/18 genotypes are highly associated with abnormal cytology with AGC, ASC-H and HSIL all having a prevalence of 100%, LSIL (86%), ASCUS (79%) and NILM (21%). An increasing trend in prevalence was observed with increasing cervical cytology abnormality for HR-HPV and HPV 16/18 genotypes (Table 4-3).

Table 4-3: Table of Prevalence of Dysplasia Cervical Lesions

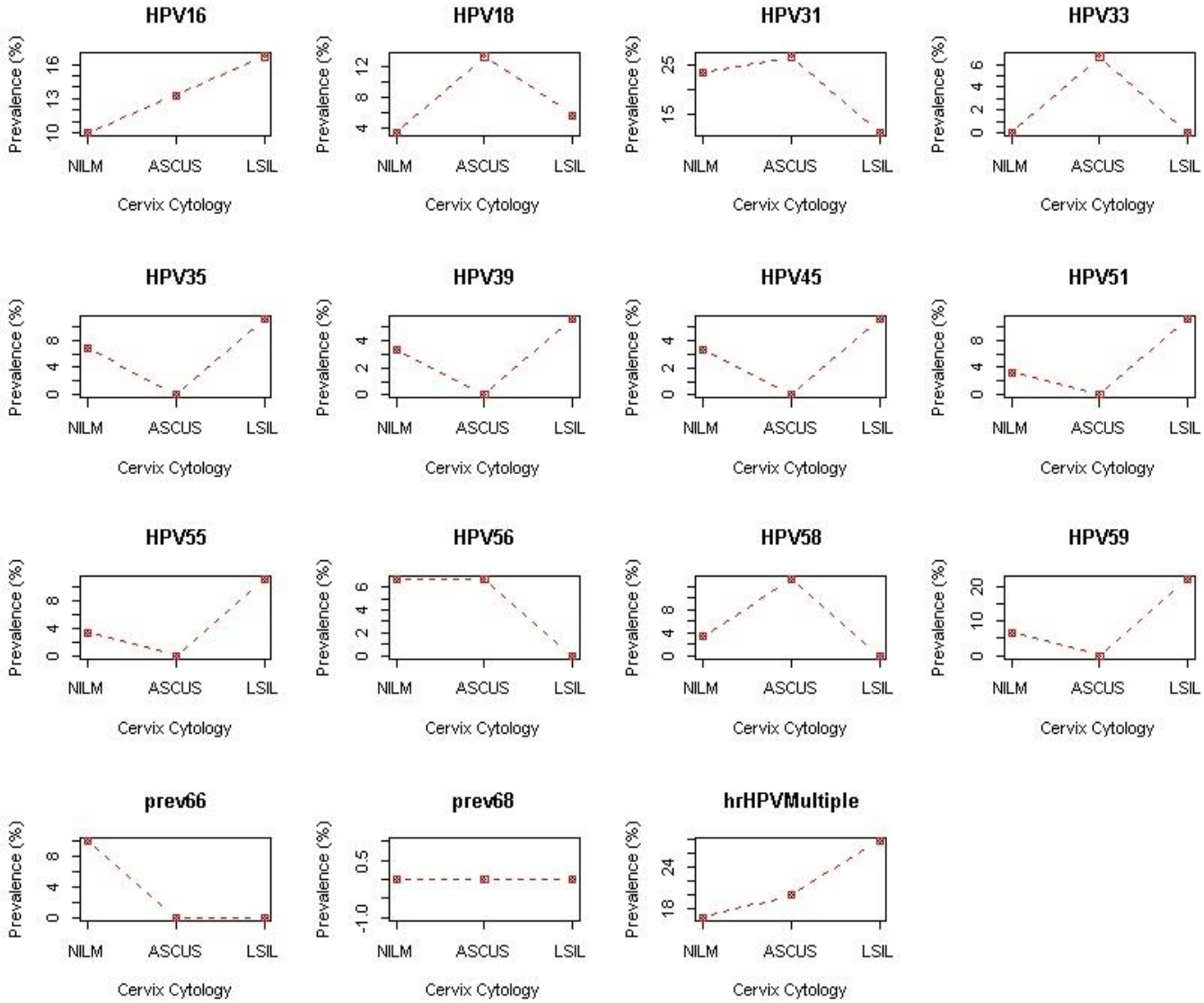
	Large database (N=1471)		**Small database (N=90)**	
	Number	**Prevalence (%)**	**Number**	**Prevalence (%)**
		Overall		
NILM	1060	72	72	80
ASCUS	183	13	9	10
AGC	3	0	-	-
LSIL	188	13	9	10
ASC-H	6	0	-	-
HSIL	29	2	-	-
Unknown	1	0		
		HR-HPV		
NILM	222/1060	21	15/72	21
ASCUS	145/183	79	7/9	78
AGC	3/3	100	-	-
LSIL	161/188	86	9/9	100
ASC-H	6/6	100	-	-
HSIL	29/29	100	-	-
Unknown	1	0		
		HPV 16/18		
NILM	62/1060	6	3/72	4
ASCUS	46/183	25	4/9	44
AGC	2/3	67	-	-
LSIL	70/188	37	4/9	44
ASC-H	3/6	50	-	-
HSIL	20/29	69	-	-
Unknown	1	0		

4.4.1 Small database

The prevalence for each individual high risk HPV genotype as a function of increasing abnormal cervical cytology was studied Figure 4-1. Sparseness of the data was an issue here because there were sampling zeroes in the cross-tabulation table cells. For HPV-16 and multiple HR-HPV genotypes, a trend test never indicated an increasing prevalence with increasing cervical cytology abnormality. The highest prevalence was observed in the LSIL group. For low risk and intermediate risk HPV genotypes, there was a decreasing prevalence trend with increasing abnormal cervical cytology (graph not shown). For HPV genotypes 16, 35, 39, 45, 51, 52, 59 and multiple HR-HPV infections, the highest prevalence was observed in the LSIL group. ASCUS diagnosis was observed

in FSW infected with HPV 18, 31, 33 and 58. FSW infected with low and intermediate risk HPV genotypes had normal smear.

Figure 4-1: Prevalence of each HPV genotype according to Cervix Cytological Diagnosis for the Small Database

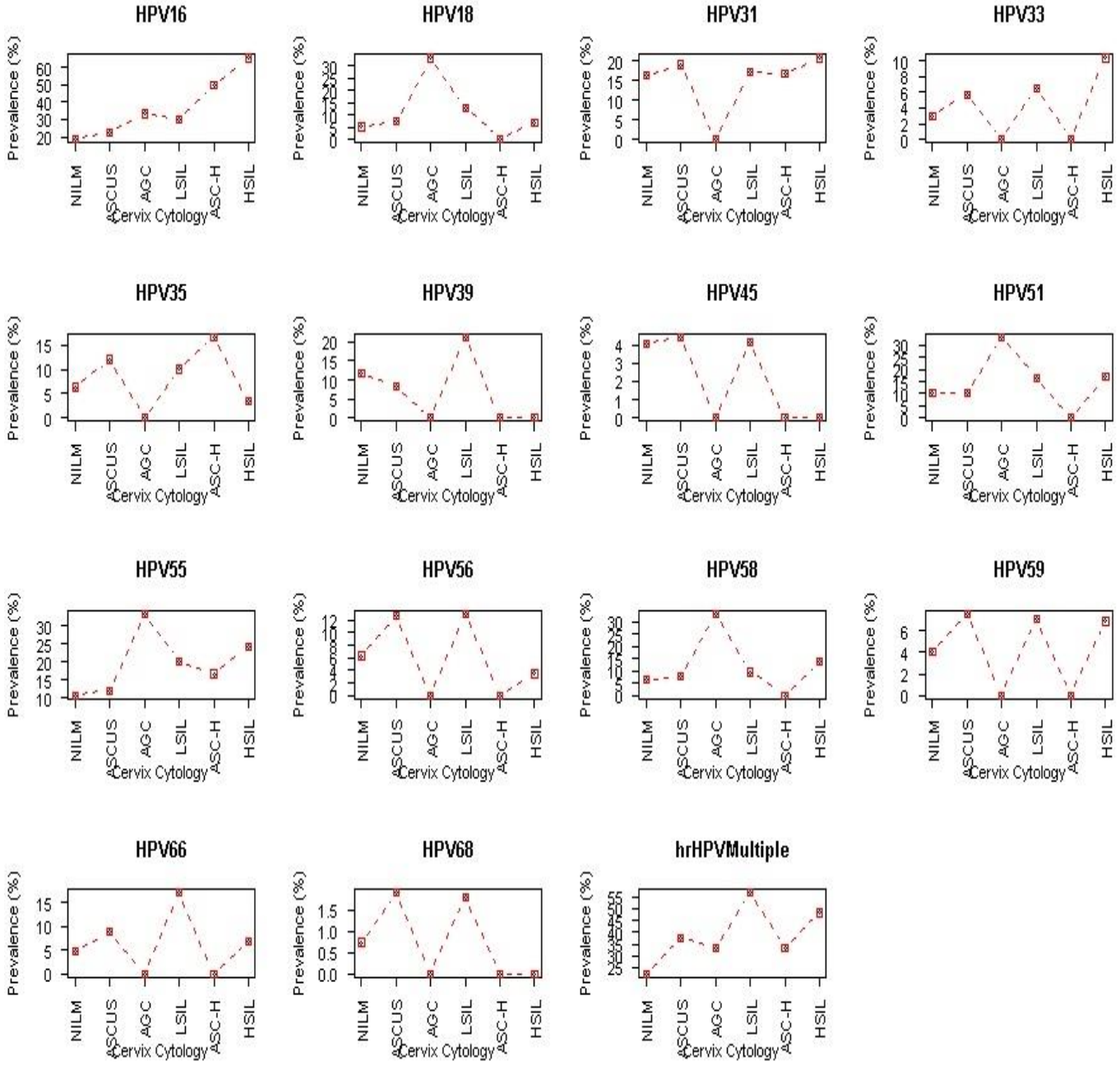


4.4.2 Large database

To study the prevalence of dysplasia in the large database, an overview of individual HR-HPV genotype as a function of increasing abnormal cervical cytology was studied, Figure 4-2. We observe that infection with HR-HPV genotypes 16, 31, 33, 55 and multiple HR-HPV genotypes depict an increasing prevalence with increasing abnormal cervical cytology, with the highest prevalence observed in FSW diagnosed with LSIL, ASC-H and HSIL, though to ascertain this, a trend test was performed. There was a significant trend test for the following HR-HPV genotypes: HPV-16 (χ_1^2=28.2385, p<0.0001), HPV-18 (χ_1^2=4.9081, p =0.0267), HPV-33 (χ_1^2=3.9127, p=0.0474), HPV-52 (χ_1^2=10.0733, p=0.0015), HPV-66 (χ_1^2=10.1049, p=0.0015) and multiple HR-HPV infection

(χ_1^2=44.4095, p<0.0001) while infection with HR-HPV genotypes 31, 35, 39, 45, 51, 56, 58 and 68 had insignificant trend test (results not shown) (Figure 4-2).

Figure 4-2: Prevalence of each HPV Genotype According to Cervix Cytological Diagnosis for the Large Database

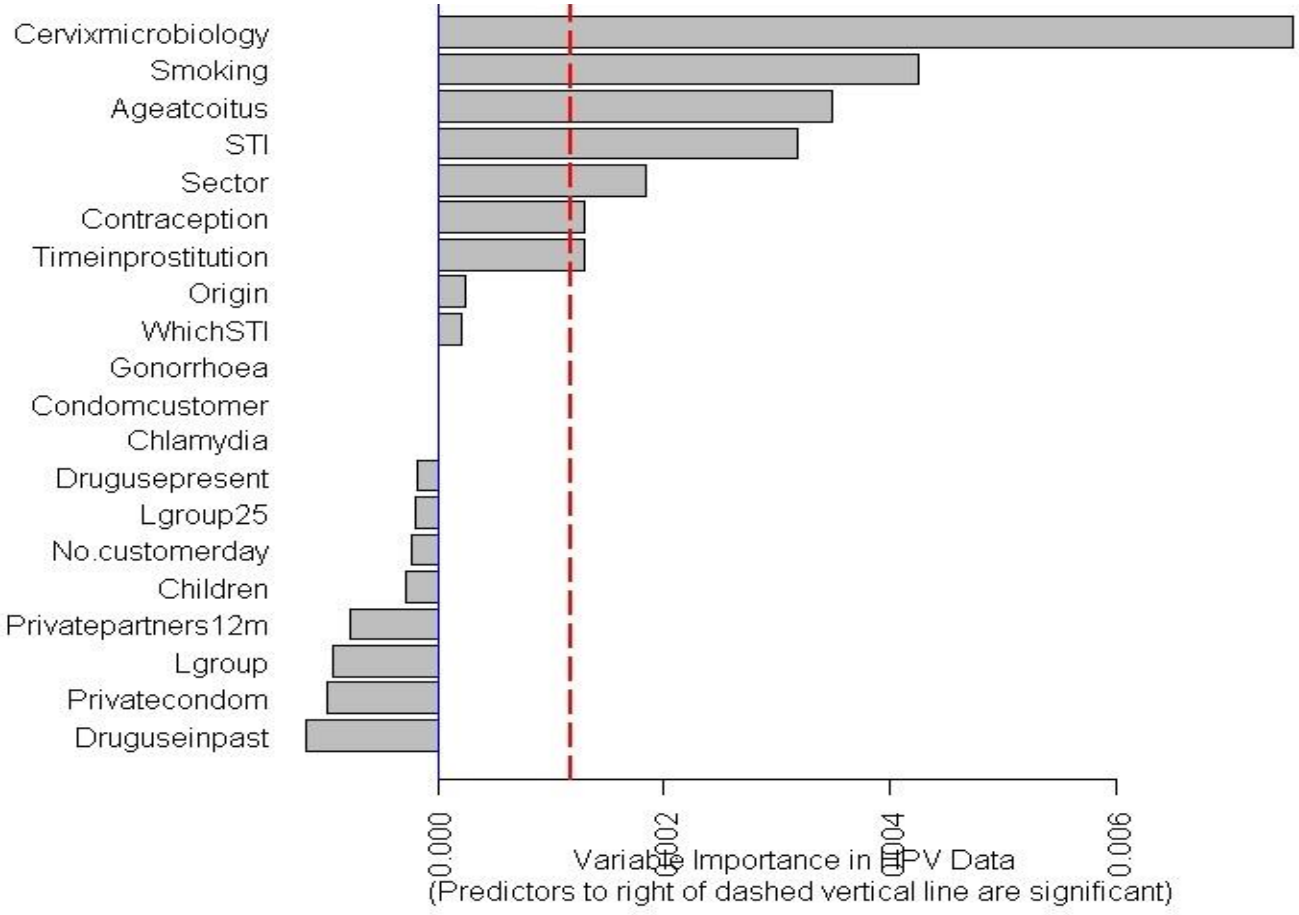


4.5 Risk Factors

Data mining applications were applied to all input predictor variables (possible risk factors) which are seldom equally relevant. The aim was to obtain a few of the input predictor variables which have substantial influence on the response. Therefore, random forests and Lasso logistic regression were considered in learning the relative importance or contribution of each input predictor variable in predicting the response. The results obtained from these two approaches were used in the model

building process. Figure 4-3 present results from RF and identified time in prostitution (years), age at first coitus (years), cervix microbiology, smoking status, current sector of work, contraception use and ever had STI at survey point as the important variables (predictors on the right of the red dotted line).

Figure 4-3: Box plot of Random Forest for Permutation Importance Variable Selection



Lasso logistic regression procedure penalizes unimportant predictors and shrinks them to zero. Cross-validation using log-likelihood obtained a lambda (λ) value of 3.5. The input predictor variables obtained which are strongly related to the response using LLR model were smoking status, contraception use, ever had STI, time in prostitution (years) and cervix microbiology. Smoking status and contraception use were highly related to HR-HPV infection even for $\lambda > 4.5$ Figure 4-4-2. The results from *glmnet* and *penalized* R packages were different; however, I could not find explanations from literature.

The predictors found using the two approaches were then preliminarily assessed graphically to see how they relate (no relation, linear or quadratic) with the predicted probability of the response. Predictors smoking status, ever had STI, and cervix microbiology have a linear relationship with HR-HPV, while 'age at first coitus' (years), time in prostitution (years), contraception use and current sector of work by FSW have a quadratic relationship with the response, HR-HPV (Figure 4-5).

Figure 4-4: Lasso Coefficient estimates versus Lambda (λ) values for LLR model

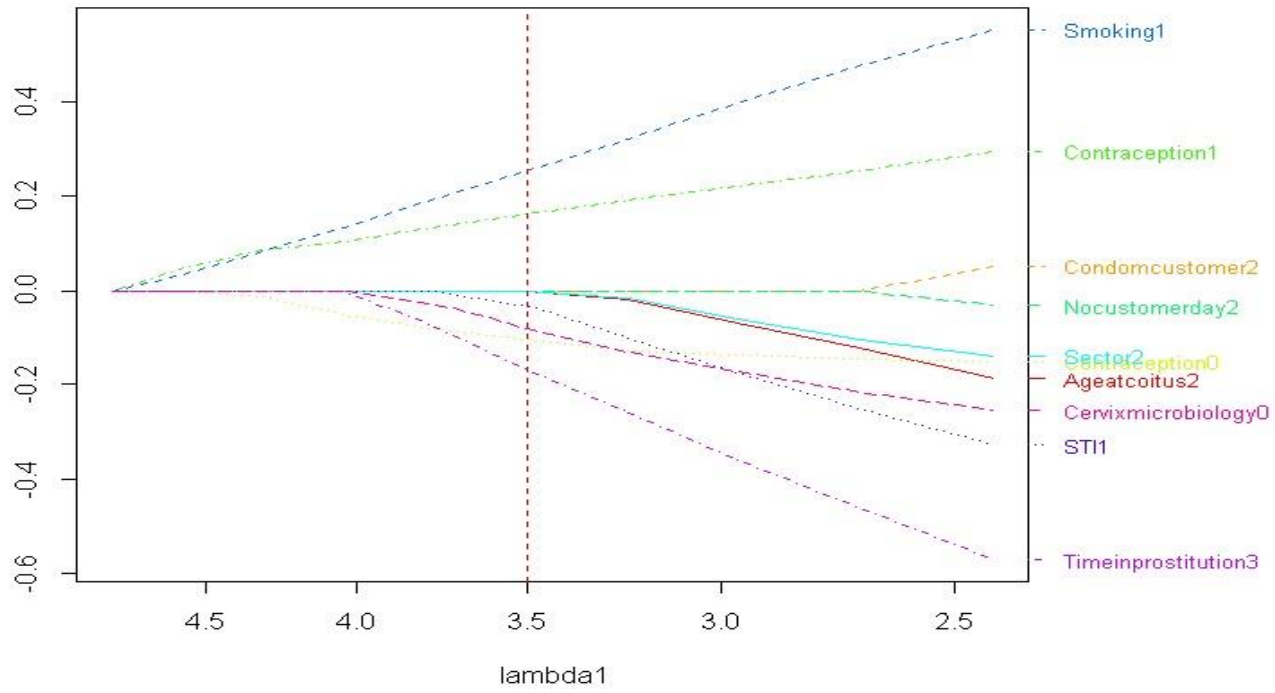
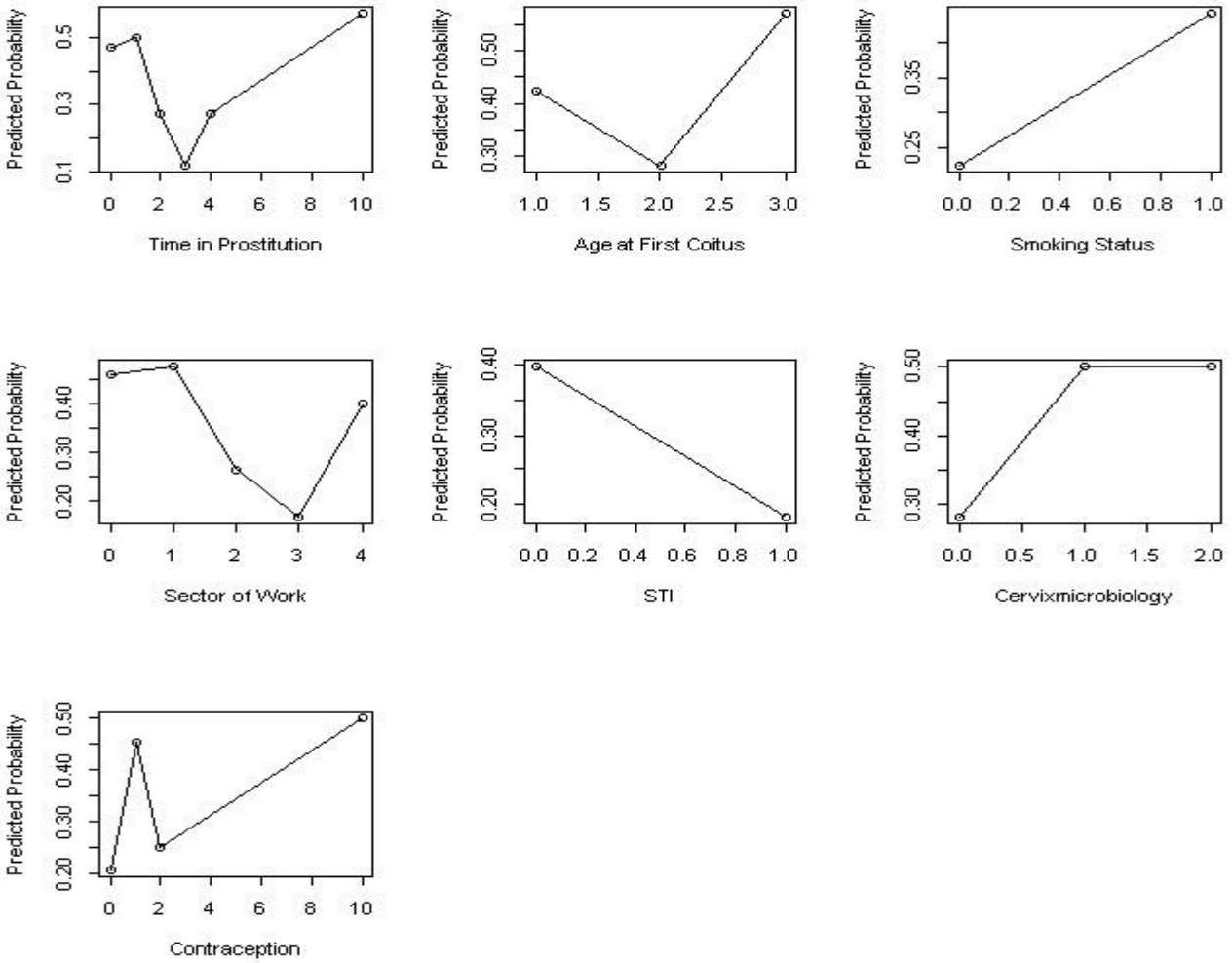


Figure 4-5: Scatter Plot for the Important Variables by Response Variable Predicted Probabilities



4.6 Modelling

Multiple logistic regression analysis was used to identify risk factors associated with HR-HPV infection. The predictor variables obtained using RF and LLR approaches were all used as main effects. However, due to small sample size and sparseness, no interactions were considered during model building process (Agresti, 2002). The predictor variable with highest insignificant p-value was dropped from the model. The parameter values of the remaining main effects were assessed for any changes due to possible confounding. It was noted that all the predictors confounded each other. Multiple logistic regression analysis accounting and not accounting for survey design showed that 'smoking status' and 'ever had STI' before the study point were statistically significantly associated with HR-HPV infection. The final LR model was

$$logit\ [\Pr[Y = 1|\ \boldsymbol{X} = \boldsymbol{x}]] = -0.930 + 1.114 * Smoking - 1.451 * Ever\ had\ STI$$

Current sector of work, age at first coitus (years), time in prostitution (years), contraceptive use and cervix microbiology were not associated with HR-HPV infection. While accounting for survey design, the risk of infection with HR-HPV genotypes was three times higher in smokers (OR 3.045 95% CI: 1.129 to 8.213) compared to non-smokers keeping all the other predictors constant. FSW who ever had history of an STI before the survey were at a lower risk of HR-HPV infection as compared to those never had STI (OR 0.234, 95% CI: 0.067 to 0.815) keeping all the other predictors constant (Table 4-4). We note that there are some differences in the parameter estimates, standard errors and p-values between ordinary LR and survey LR. This is an indication that accounting for survey design via the age group strata had an effect. In model diagnostics, logistic regression accounting and not accounting for survey design, Hosmer-Lemeshow goodness-of-fit test was not statistically significant an indication that the model fits the data well. The model inference did not change when the cases of categories causing sparseness (marked * in Table 4-1) were deleted from the data set.

Table 4-4: Estimates of Adjusted Odds Ratios and 95% CI

		Ordinary LR			Survey LR		
Predictor[a]	**Category**	**Estimate (se)**	**OR(95% CI)**	**P-Value**	**Estimate (se)**	**OR(95% CI)**	**P-Value**
Intercept	Constant	-1.071(0.391)	0.343(0.159;0.737)	0.0061	-0.930(0.405)	0.395(0.178;0.873)	0.0217
Smoking	Non-smoker		1	0.0169		1	0.0278
	Smoker	1.194(0.500)	3.300(1.274;9.167)		1.114(0.506)	3.045(1.129;8.213)	
Ever had STI	No		1	0.0353		1	0.0225
	Yes	-1.341(0.637)	0.262(0.066;0.840)		-1.451(0.636)	0.234(0.067;0.815)	

Note: n=90; OR: Odds Ratio; CI: Confidence Interval; LR: Logistic Regression; se: Standard error
***a** : reference categories for categorical predictors are Smoking: Non-smokers and Ever had STI: No*

4.7 Model Prediction

Model prediction of the binary response using test data with the final model was done. Boosting, logistic regression accounting and not accounting for survey design methods were used. To accomplish this, the survey data was split into training (70% =62 cases) and test (30% =28 cases) data sets. Main effects were used during boosting because of its strength, that is, the boosting algorithm is robust and that interactions and nonlinearities need not be explicitly specified. Research on comparison of classification of predicted binary response results using boosting and logistic regression has been done as stated by Schapire (2003). From Table 4-5 below, ordinary LR (misclassification rate of 28.57%) gave a better prediction (that is, explaining the variability in question) compared to boosting with a misclassification rate of 32.14%. LR accounting and not accounting for survey design gave the same misclassification rate and area under the curve an indication that they give the same prediction.

Table 4-5: Classification Results of Boosting, Ordinary logistic and Survey Logistic Regression

	Boosting			**Ordinary LR**			**Survey LR**		
	Observed			**Observed**			**Observed**		
Predicted	**Negative**	**Positive**	**Total**	**Negative**	**Positive**	**Total**	**Negative**	**Positive**	**Total**
Negative	16	4	20	13	4	17	13	4	17
Positive	5	3	8	4	7	11	4	7	11
Total	21	7	28	17	11	28	17	11	28
Error	0.3214			0.2857			0.2857		
Area Under Curve				0.7487			0.7487		

5. Discussion and Conclusion

Human papillomavirus is a public health concern for women. This study found a high prevalence of HR-HPV infection (34%) in FSW followed up at Ghapro, Antwerp, Belgium (Del Amos et al., 2004; Juarez-Figueroa et al., 2001; Thomas et al., 2001 & Wang et al., 2013). Prevalence of HR-HPV infection among FSW showed a decreasing trend test with increasing age in both the small ($p=0.066$, borderline) and the large ($p<0.001$) database which is consistent with other studies. This inverse relationship between increasing age and HR-HPV prevalence decrease has been attributed to the development of acquired immunity over time after multiple exposures and clearance of HPV infections (del Amos et al., 2004, Kjaer et al., 2000; Juarez-Figueroa et al., 2001 & Burk et al., 1996). The highest HR-HPV prevalence was observed in FSW aged 21 to 25 years in the small and large database, this is in agreement with the above discussion. 'Time in prostitution' and 'ages at first coitus' in years were not associated with HR-HPV types infection in both LR accounting and not accounting for survey design. These data are nevertheless sensitive; and are often missing information due to possible susceptibility to recall bias. Contraceptive use was not associated with HR-HPV genotype infection in both multiple logistic regression analysis. However, in the univariate analysis, FSW using hormonal contraceptive had a three-fold (OR 3.214, 95% CI: 1.157 to 8.926) increased risk of HR-HPV types infection compared to those FSW using only condoms. In this cross-sectional study no data was collected on the duration of hormonal contraceptive use and also there is no documentation as to whether the FSW using hormonal contraceptive also used condoms, it is likely that these FSW might be also at higher risk of multiple HPV exposures from multiple sexual partners.

Smoker FSW had a three-fold increased risk of HR-HPV infection compared to non-smokers in multiple logistic regression analysis accounting and not accounting for study design. However, past studies have shown mixed signals. That is, smoking is a risk factor for HPV persistence and malignant transformations although others have found a lower HPV infection rate in smokers. Smoking is said to influence the incidence and persistence of HPV infections by suppressing local immune function, increased cellular proliferation, unregulated pro-inflammatory factors, or induced host DNA damage resulting in increased susceptibility to HPV infection. Gunnell et al (2006) suggest that HPV infection and smoking behavior may create a biochemical synergy that propels cervical cancer in women if they are HPV-positive smokers. Past history of STI before the survey study was associated with HR-HPV types infection. FSW who ever had STI before the survey study

had a protective effect compared to FSW who never had past history of STI while accounting and not accounting for survey design.

The survey sample size was small. This more often results into sparse tables with sampling and structural zeroes. Sampling zeros are part of the data set and contributes to the likelihood function and model fitting. Structural zero is not an observation and is not part of the data (Agresti, 2002). Sampling zeros can affect the existence of finite ML estimates of loglinear and logit model parameters with the supremum of the likelihood function being finite (Agresti, 2002). During model building process, some main effects and interaction terms had $\pm\infty$ ML parameter estimates implying that ML fitted values equal zero in some cells, and some odds ratio estimates equal ∞ or zero an indication that the iterative fitting process does not converge. The resultant standard errors are extremely large and numerically unstable. Thus, data sparseness realize infinite estimated effects and hence report estimated effects and results of statistical inferences that are invalid and highly unstable (Agresti, 2002). When the pattern of empty cells forces certain fitted values for a model to equal zero, this affects the degrees of freedom for testing model fit (Haslett 1990). Empty cells and sparse tables can cause sampling distributions of goodness-of-fit statistics (X^2, G^2) to be far from chi-squared since the adequacy of chi-squared approximation depends both on sample size and number of cells. From the foregoing discussion, the model building process might have been influenced by the sparseness in the data and thus some inferences might be misleading.

In the small database, there was association between contraceptive use and cervical neoplasia lesions accounting for survey design (Thomas et al., Bangok I,II,III, 2001 & IARC, 1999). Research has shown that current multiple and persistent HR-HPV infections have been associated with increased cervical dysplasia. For example, detection of HPV-16/18 in women carries a five-fold greater risk of developing the precancerous CIN 2/3 than detection of other HR-HPV types (Castle et al., 2005). Research has shown that cervical cancer key driver is persistent HR-HPV infection. Therapeutic HPV vaccination has been proposed and implemented to different women groups based on age and sexual activity in different countries. Therefore, in conclusion smoking status and ever had STI risk factors were found to be associated with HR-HPV in this survey. However, the high prevalence of HR-HPV genotypes and the limited indications in literature of additional benefit of HPV vaccination to this population at higher risk of HR-HPV infection do certainly support the need for a HPV vaccine trial with either Gardasil or multivalent HPV (6/11/16/18/31/33/45/52/58) vaccines. The therapeutic vaccination may offer cross-protection and protection offered against HPV subtypes included in the vaccine but with which the population might not be infected. This will also offer opportunities to protect FSW against a job related health risk.

6. Limitations and Recommendations

HPV infection is known to naturally resolve as a result of a cell-mediated immune response (Stanley 2008). Failure to induce an effective immune response is related to inefficient activation of innate immunity and ineffective priming of the adaptive immune response; this defective immune response facilitates viral persistence, a key feature of HR-HPV infection (Stanley 2008). This was considered as a limitation as the test for HPV genotypes could not capture all the HPV infection a subject ever had except only at the survey study point. The population variability in terms of the number of hidden nationalities and lack/erratic history of prostitution in another country were possible sources of bias in survey study.

Additionally, the answers to the questions in the survey might be biased because the survey was carried out by HCW of the Ghapro. Participants are often answering in line with what HCW expect them to do, withholding information that would not be in line with recommendations on, for example, safe sex techniques etc. The same counts for taboo, like the use of drugs. People intend to answer what is socially accepted as good behaviour. There was language bias as only four languages were used for this survey study.

To address the problem of sparseness in variable selection, we recommend use of penalized likelihood using concave penalty functions. Also use of random effects models and Bayesian methods to address sparseness. Model prediction comparison using boosting and logistic regression have been studied, however, we were not able to compare model prediction results from boosting and logistic regression taking into account survey design since no research has been done to this end and thus we recommend further research.

References

About human papillomavirus. Available at: http://www.genticel.com/web/en/35-hpv-and-cervix-carcinoma.php, accessed on 20th August, 2013.

Agresti, A. (2002). *An Introduction to Categorical Data Analysis*, 2nd ed., New York: John Wiley & Sons.

American College of Obstetricians and Gynecologists. (2010). *Human papillomavirus Vaccination. Committee Opinion No. 467*. (Accessed on 03rd September, 2013.)

American College of Obstetricians and Gynecologists. (2005). *Human papillomavirus. ACOG Practice Bulletin No. 61*. Obstet Gynecol; 105:905–18.

Arbyn, M., Autier, P., and Ferlay, J. (2007). *Burden of cervical cancer in the 27 member states of the European Union: estimates for 2004*. Ann Oncol 18:1423-1425.

Arbyn, M., Ronco, G., Anttila, A., Meijer, C.J., Poljak, M., Ogilvie, G., Koliopoulos, G., Naucler, P., Sankaranarayanan, R., and Peto, J. (2012). *Evidence regarding human papillomavirus testing in secondary prevention of cervical cancer*. Vaccine 30 Suppl. 5: F88-99.

Aslam, J.A.; Popa, R.A.; and Rivest, R.L. (2007, Aug). *On Estimating the Size and Confidence of a Statistical Audit, Proceedings of the Electronic Voting Technology Workshop* (EVT '07), Boston, MA.

Baay, M., Verhoeven, V., Wouters, K., Lardon, F., Van Damme, P., Avonts, D., Van Marck, E., Van Royen, P., Vermorken, J.B. (2004). *The Prevalence of the Human Papillomavirus in Cervix and Vagina in Low-risk and High-risk Populations*. Scand J Infect Dis 36: 456-459.

Binder, D. A. (1983), *"On the Variances of Asymptotically Normal Estimators from Complex Surveys,"* International Statistical Review, 51, 279–292.

Bosch, F.X., Burchell, A.N., Schiffman, M., Giuliano, A.R., de Sanjose, S., Bruni, L., Tortolero-Luna, G., Kjaer, S.K., and Munoz, N. (2008). *Epidemiology and natural history of human papillomavirus infections and type-specific implications in cervical neoplasia*. Vaccine 26 Suppl 10: K1-16.

Breiman, L. (2001). *Random forests*. *Machine Learning*, 45: 5-32.

Breiman, L. (1996). *"Bagging predictors"*. *Machine Learning*. Vol. 24 (2): 123–140. doi:10.1007/BF00058655. CiteSeerX: 10.1.1.121.7654.

Burk, R.D., Kelly, P., Feldman, J., Bromberg, J., Vermund, S.H., DeHovitz, J.A. and Landesman, S.H. (1996). *Declining prevalence of cervicovaginal human papillomavirus infection with age is independent of other risk factors*. Sex Transm Dis; 23(4):333–41.

Brown, B., Blas, M., Cabral, A., Byraiah, G., Guerra-Giraldez, C., Sarabia-Vega, V., Carcamo, C., Gravitt, P.E., Halsey, N.A. (2011). *HPV Prevalence, Cervical Abnormalities, and Risk Factors among Female Sex Workers in Lima, Peru*. International Journal of STDs and AIDS; 23 242-247

Brown, B., Carcamo, C., Blas, M., Valderrama, M., and Halsey, N. (2010). *Peruvian Female Sex Workers: Understanding HPV and barriers to vaccination.* Vaccine 28, 7743-7747.

Caster O. (2007). *Mining the WHO Drug Safety Database Using Lasso Logistic Regression*,

UUDM Project Report, http://www.diva-portal.org/smash/get/diva2:304279/FULLTEXT01.pdf

Castellsague, X., de Sanjose, S., Aguado, T., Louie, K.S., Bruni, L., Munoz, J., Diaz, M., Irwin, K., Gacic, M., Beauvais, O., Albero, G., Ferrer, E., Byrne, S., Bosch, F.X. (2007, Nov). *HPV and Cervical Cancer in the World: 2007 Report (Edited).* Vaccine Vol 25(Suppl. 3), pg c1-230. Accessed on 30.07.2013 http://www.sciencedirect.com/science/journal/0264410X/25/supp/S3

Castellsague, X., Diaz, M., de Sanjose, S., Munoz, N., Herrero, R., Franceschi, S., Peeling, R.W., Ashley, R., Smith, J.S., Snijders, P.J., et al. (2006). *Worldwide human papillomavirus etiology of cervical adenocarcinoma and its cofactors: implications for screening and prevention*. J Natl Cancer Inst 98:303-315.

Castle, P.E., Solomon, D., Schiffman, M., Wheeler, C.M. (2005). *Human papillomavirus type 16 infections and 2-year absolute risk of cervical precancer in women with equivocal or mild cytologic abnormalities.* J Natl Cancer Inst Vol 97(14):1066–1071. [PubMed: 16030304]

Chambers, R.L. and Skinner, C.J. (2003). *Analysis of Survey Data*. John Wiley & Sons, Ltd.

Caster, O. (2007). *Mining the WHO Drug Safety Database Using Lasso Logistic Regression*. UUDM Project Report, http://www.diva-portal.org/smash/get/diva2:304279/FULLTEXT01.pdf

Chawla, N., Moore Jr., E. T., Bowyer, K.W., Hall, L. O., Springer, C., and Kegelmeyer, P. (2002). *Bagging Is A Small-Data-Set Phenomenon*. In International Conference on Computer Vision and Pattern Recognition (CVPR).

Clogg, C. C. and Eliason, S. R. (1987). *Some common problems in log-linear analysis*. Sociol. Methods Res. 15: 4 44.

Cogliano, V., Baan, R., Straif, K., Grosse, Y., Secretan, B., and El Ghissassi, F. (2005). *Carcinogenicity of human papillomaviruses*. Lancet Oncol 6:204.

de Villiers, E.M., Fauquet, C., Broker, T.R., Bernard, H.U., and zur Hausen, H. (2004). *Classification of papillomaviruses*. Virology 324:17-27.

Del Amos, J., Gonzalez, C., Losana, J., Clavo, P., Munoz, L., Ballesteros, J., Garcia-Saiz, A., Belza, M., Ortiz, M., Menendez, B., del Romero, J., and Bolumar, F. (2004). *Influence of age and geographical origin in the prevalence of high risk human papillomavirus in migrant female sex workers in Spain*. Sex Transm Infect. 81(1): 79–84. doi: 10.1136/sti.2003.008060

Depuydt, C.E., Criel, A.M., Benoy, I.H., Arbyn, M., Vereecken, A.J., Bogers, J.J. (2012). *Changes in type-specific human papillomavirus load predict progression to cervical cancer.* J Cell Mol Med.; 16(12):3096-104. doi: 10.1111/j.1582-4934.2012.01631.x

Depuydt, C.E., Benoy, I.H., Bailleul, E.J., Vandepitte, J., Vereecken, A.J., Bogers, J.J. (2006, Dec). *Improved endocervical sampling and HPV viral load detection by Cervex-Brush® Combi.* Cytopathology. 17(6):374-81. doi/10.1111/j.1365-2303.2006.00386.x

Depuydt, C. E., Vereecken, A. J., Salembier, G. M., Vanbrabant, A. S., Boels, L. A., van Herck, E., Arbyn, M., Segers, K., and Bogers, J. J. (2003). *Thin-layer liquid-based cervical cytology and PCR for detecting and typing human papillomavirus DNA in Flemish women.* Br J Cancer.; 88(4): 560–566. doi: 10.1038/sj.bjc.6600756.

FUTURE II Study Group. (2007). *Quadrivalent vaccine against human papillomavirus to prevent high-grade cervical lesions*. N Engl J Med;356:1915-27.

Franco, E. L., Duarte, F. E., Ferenczy, A. (2001). *Cervical cancer: epidemiology, prevention and the role of human papillomavirus infection*. Can Med Assoc J; 164: 1017- 25.

Friedman, J., Hastie, T. and Tibshirani, B. (2010). *Regularization Paths for Generalized Linear Models via Coordinate Descent.* Journal of Statistical Software (http://www.jstatsoft.org/) Vol. 33(1).

Freund, Y. and Schapire, R. E. (1997). *A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting*. Journal of Computer and System Sciences, 55(1):119-139

Genital HPV Infection — CDC Fact Sheet. http://www.cdc.gov/std/hpv/stdfact-hpv.htm. Centers for Disease Control and Prevention (CDC). July 25, 2013. Retrieved 03 September, 2013.

Genkin, A., Lewis, D. D. and Madigan, D. (2007). *Large-scale bayesian logistic regression for text categorization*. Technometrics, 49, 291–304.

Goeman, J. J. (2009): *L1 Penalized Estimation in the Cox Proportional Hazards Model*. Biometrical Journal, Vol 52(1); 70-84

Gunnell, A. S., Tran, T. N., Torra˚ng, A., Dickman,P. W., Spare´n, P., Palmgren, J., and Ylitalo, N. (2006). *Synergy between Cigarette Smoking and Human Papillomavirus Type 16 in Cervical Cancer In situ Development*. Cancer Epidemiol Biomarkers Prev; 15(11): 2141-7

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*, Second Edition. New York: Springer.

Hens, N. (2013). *Data Mining*. University of Hasselt, unpublished course notes

Hosmer, D. W., Lemeshow, S. (2000). *Applied Logistic Regression*, New York: Wiley, ISBN 0-471-61553-6.

Hothorn, T., Hornik, K. and Zeileis, A. (2006). *Unbiased recursive partitioning: A conditional inference framework*. Journal of Computational and Graphical Statistics, 15(3):651–674.

International Agency for Research on Cancer. (1999). *Hormonal contraception and post-menopausal hormonal therapy*. IARC Monogr Eval Carcinog Risks Hum; 72:1–660.

Jemal, A., Siegel, R., Ward, E., Hao, Y., Xu, J., and Thun, M.J. (2009). *Cancer statistics, 2009*. CA Cancer J Clin 59:225-249.

Juarez-Figueroa, L. A., Wheeler CM, Uribe-Salas FJ, Conde-Glez, C. J., Zampilpa-Mejía, L. G., García-Cisneros, S., Hernández-Avila, M. (2001). *Human papillomavirus: a highly prevalent sexually transmitted disease agent among female sex workers from Mexico City*. Sex Transm Dis; 28(3): 125–30.

Kjaer, S. K., Svare, E. I., Worm, A.M., Walboomers, J. M., Meijer, C.J. and van den Brule, A. J. (2000). *Human papillomavirus infection in Danish female sex workers. Decreasing prevalence with age despite continuously high sexual activity*. Sex Transm Dis; 27(8): 438–45.

Liaw, A. and Wiener, M. (2002). *Classification and Regression by randomForest*. R News 2(3), 18-22.

Medical need provoked by HPV infection. Available at: http://www.genticel.com/web/en/36-medical-need-provoked-by-hpv-infection.php, accessed on 20th August, 2013.

Meier, L. D. (2008). *High-Dimensional Regression Problems with Special Structure*. A PhD dissertation submitted to Eth Zurich.

Meier, L., van de Geer, S. and Bühlmann, P. (2008). *The group lasso for logistic regression*. J. R. Statist. Soc. 70, Part 1, pp. 53–71

Micalessi, I. M., Boulet G.A.V., Bogers, J.J., Benoy, I.H., Depuydt, C.E. (2012). *High-throughput detection, genotyping and quantification of the human papillomavirus using real-time PCR.* Clin Chem Lab Med.; 50: 655–61.

Moscicki, A. (2008, Oct). *HPV vaccines: Today and in the future*. J Adolesc Health. Vol 43(4 Suppl): S26–S40. doi:10.1016/j.jadohealth.2008.07.010.

Munoz, N., Kjaer, S.K., Sigurdsson, K., Iversen, O.E., Hernandez–Avila, M., Wheeler, C.M., Perez, G., Brown, D.R., Koutsky, L.A., Tay, E.H., Garcia, P.J., Ault, K.A., Garland, S.M., Leodolter, S., Olsson, S., Tang G.W.K., Ferris, D.G., Paavonen, J., Bosch, M.S.X., Dillner, J., Huh, W.K., Joura, E.A., Kurman, R.J., Majewski, S., Myers, E.R., Villa, L.L., Taddeo, F.J., Roberts, C., Tadesse, A., Bryan, J.T., Lupinacci, L.C., Giacoletti, K.E.D., Sings, H.L., James, M.K., Hesley, T.M., Barr, E., Haupt. R.M. (2010, Feb). *Impact of human papillomavirus (HPV)-6/11/16/18 vaccine on all HPV-associated genital diseases in young women.* JNCI J Natl Cancer Inst 102 (5): 325-339. doi: 10.1093/jnci/djp534

Munoz, N. N., Castellsague, X., de Gonzalez, A. B., & Gissmann, L. L. (2006). *Chapter 1: HPV in the etiology of human cancer*. Vaccine, 24(Supplement 3), S1-S10.

Munoz, N., Bosch, F.X., de Sanjose, S., Herrero, R., Castellsague, X., Shah, K.V., Snijders, P.J., and Meijer, C.J. (2003). *Epidemiologic classification of human papillomavirus types associated with cervical cancer*. N Engl J Med 348:518-527.

Munoz, N. and F.X. Bosch, (1997). *Cervical cancer and human papillomavirus: epidemiological evidence and perspectives for prevention*. Salud Publica Mex,. 39 (4): p. 274-82.

Nelder, J. A. and Wedderburn, R. W. M. (1972), "*Generalized Linear Models*," Journal of the Royal Statistical Society, Series A, 135, 370–384.

Other multivalent vaccine candidates. Available at: http://www.genticel.com/products/multivalent-vaccines/, accessed on 20th August, 2013.

Paavonen, J., Naud, P., Salmeron, J., Wheeler, C.M., Chow, S.N., Apter, D., et al. (2009, Jul). *Efficacy of human papillomavirus (HPV)-16/18 AS04-adjuvanted vaccine against cervical infectionand precancer caused by oncogenic HPV types (PATRICIA): final analysis of a double-blind, randomised study in young women.* Lancet; 374(9686):301-14. doi: 10.1016/S0140-6736(09)61248-4.

Petter, A., Heim, K., Guger, M., Ciresa-Konig, A., Christensen, N., Sarcletti, M., Wieland, U., Pfister, H., Zangerle, R., and Hopfl, R. (2000). *Specific serum IgG, IgM and IgA antibodies to human papillomavirus types 6, 11, 16, 18 and 31 virus-like particles in human immunodeficiency virus-seropositive women.* Journal of General Virology, 81, 701–708

Rao, J. N. K., Wu, C. F. J., and Yue, K. (1992), *Some Recent Work on Resampling Methods for Complex Surveys*. Survey Methodology, 18, 209–217.

Särndal, C. E., Swensson, B., and Wretman, J. (1992), *Model Assisted Survey Sampling*, New York: Springer-Verlag.

Sawant, A.A. and Chawan, P. M. (2013). *Study of Data Mining Techniques used for Financial Data Analysis.* International Journal of Engineering Science and Innovative Technology (IJESIT); Vol 2(3).

Schapireat, R. E. (2003). *Nonlinear Estimation and Classification*, Springer. The Boosting Approach to Machine Learning An Overview (2001). T Labs.

Schmitt, M., Depuydt, C., Benoy, I., Bogers, J., Antoine, J., Arbyn, M., Pawlita, M.; VALGENT Study Group. (2012). *Prevalence and viral load of 51 genital human papillomavirus types and three subtypes.* Int J Cancer; 132(10):2395-403. doi: 10.1002/ijc.27891.

Shevade, S. and Keerthi, S. (2003). *A simple and efficient algorithm for gene selection using sparse logistic regression*. Bioinformatics, 19, 2246–2253.

Smith, J.S., Lindsay, L., Hoots, B., Keys, J., Franceschi, S., Winer, R., and Clifford, G.M. (2007). *Human papillomavirus type distribution in invasive cervical cancer and high-grade cervical lesions: a meta-analysis update*. Int J Cancer 121:621-632.

Stanley, M. (2008). *Immunobiology of HPV and HPV vaccines*. Gynecologic Oncology. Vol. 109 (2), Pages S15–21

Strobl, C., Hothorn, T., and Zeileis, A. (2009a). *Party on! A New, Conditional Variable Importance Measure for Random Forests Available in the party Package*. BMC Bioinformatics, 8:25.

Strobl, C., Hothorn, T., and Zeileis, A. (2009b). *An Introduction to Recursive Partitioning: Rational, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests*. Psychological Methods. 14(4): 323-348.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. (2008). *Conditional variable importance for random forests*. BMC Bioinformatics, 9:307.

Strobl, C., Zeileis, A. (2008). Danger: *High power! – Exploring the statistical properties of a test for random forest variable importance*. In Proceedings of the 18th International Conference on Computational Statistics, Porto, Portugal, 2008.

Thomas, D. B., Ray, R.M., Kuypers, J., Kiviat, N., Koetsawang, A., Ashley, R. L., Qin, Q. and Koetsawang, S. (2001). *Human Papillomavirus and cervical cancer in Bangkok. III. The role of husbands and commercial sex workers*. Am J Epidemiol; 153(8): 740-8.

Thomas, D. B., Ray, R.M., Kuypers, J., Kiviat, N., Koetsawang, A., Ashley, R. L., Qin, Q. and Koetsawang, S. (2001). *Human Papillomavirus and cervical cancer in Bangkok. II. Risk factors for in situ and invasive squamous cell cervical carcinomas*. Am J Epidemiol; 153(8): 732-9.

Thomas, D. B., Ray, R.M., Kuypers, J., Kiviat, N., Koetsawang, A., Ashley, R. L., Qin, Q. and Koetsawang, S. (2001). *Human papillomaviruses and cervical cancer in Bangkok. I. Risk factors for invasive cervical carcinomas with human papillomavirus types 16 and 18 DNA*. Am J Epidemiol; 153(8):723–31.

Tibshirani, R. (1996). *Regression shrinkage and selection via the lasso*. J. R. Statist. Soc. B, 58, 267–288.

Walboomers, J.M., Jacobs, M.V., Manos, M.M., Bosch, F.X., Kummer, J.A., Shah, K.V., Snijders, P.J., Peto, J., Meijer, C.J., Muñoz, N. (1999) *Human papillomavirus is a necessary cause of invasive cervical cancer worldwide*. J Pathol. 189:12–19

Wang, X., Gu, D., Lou, B., Xu, B., Qian, F., and Chen, Y. (2013). *Hospital-based prevalence of high-risk cervical HPV types infecting the general population and female sex workers in Huzhou, China.* Int J Gynaecol Obstet., 120(1): 37-41. doi:10.1016/j.ijgo.2012.07.019

Wolter, K. M. (2007). *Introduction to Variance Estimation*, Second Edition, New York: Springer.

Yuan, M. and Lin, Y. (2006). *Model selection and estimation in regression with grouped variables*. Journal of the Royal Statistical Society Series B, 68: 49-67.

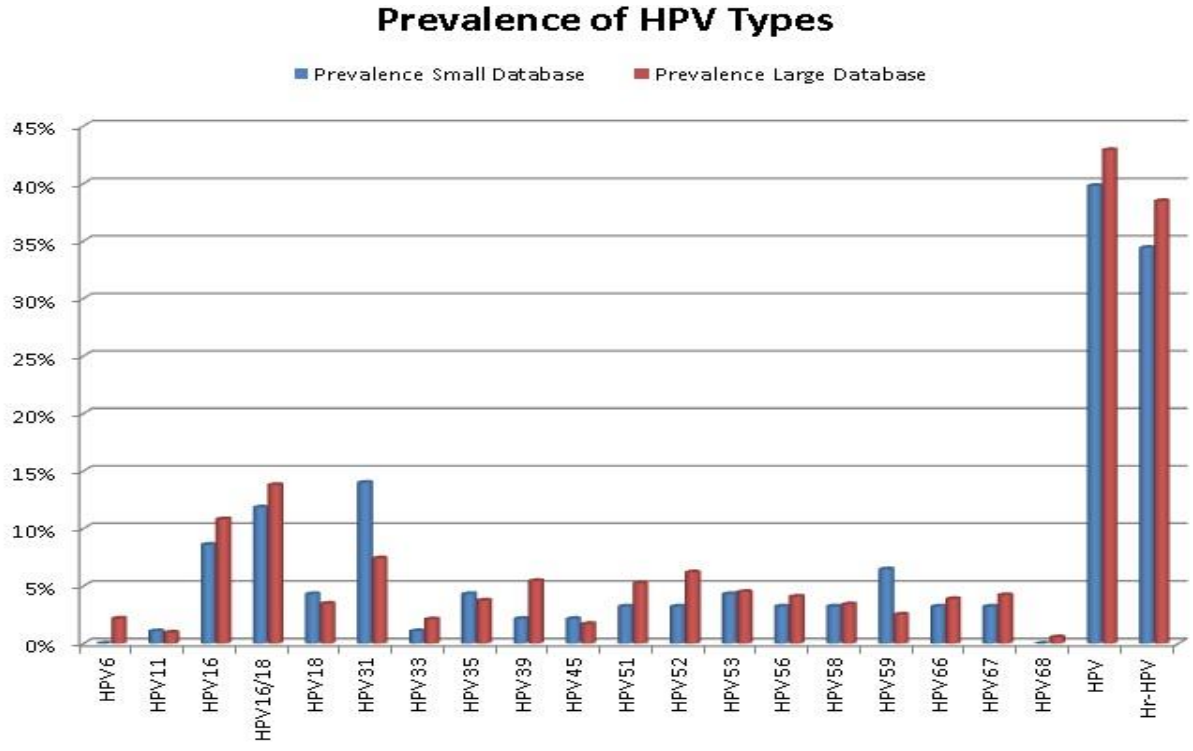
Appendix

Table A.1: Variables and Variable Coding used

Variables	Variable Type	Coding for the Variable
HR-HPV (Var: hrHPV)	Categorical	0: Negative 1: Positive
Age Group (Years) (Var: Lgroup)	Categorical	1: <21 2: 21-25 3: 26-30 4: 31-35 5: 36-40 6: > 40
Region of Origin (Var: Origin)	Categorical	0: Unknown 1: Africa (Sub-Sahara) 2: America 3: South-East Asia 4: Europe 5: Eastern Mediterranean 6: West Pacific Ocean
Current work sector (Var: Sector)	Categorical	0: Bar, Bar + Window 1: Private house, Massage parlours, Home, SM-studio 2: Window (red light district) 3: Escort 4 : Street/African women Cafe 10: Unprecedented
Number of Children (Parity) (Var: Children)	Categorical	0: 0 (None) 1: 1 2: >1 10: Unknown
Smoking	Categorical	0: No 1: Yes 10: Unknown
Contraceptives	Categorical	0: Only condom 1: Hormonal Contraceptive (HC) 2: Intrauterine device (IUD)/Tubal Ligation (TL) 10: Unknown
Condom use for each technique with customers (Var: condomcustomer)	Categorical	1: Always 2: Mostly, 3: Sometimes 4: Never 10: Unknown
Age of first intercourse (Years) (Var: Ageatcoitus)	Categorical	1: <16 2 : 16 - 20 3 : >20
Number of private partners (in the past 12 months) (Var: Privatepartners12m)	Categorical	0: 0 1: 1 2: 2 to 5 3: >5 10: Unknown
Time in prostitution (Years) (Var: timeinprostitution)	Categorical	0: <2 1: 2 2: 3-4 3: 5-9 4: > 9 10: Unknown
Condom use with private partners (Var: Condomprivate)	Categorical	0: No 1: Yes 10: Unknown
STI in the past (Var: STI)	Categorical	0: No 1: Yes

Which STI? (Var: whichSTI)	Categorical	0 : HIV 1: Syphilis (Treponema pallidum) 2: Hepatitis B 3: Neisseria Gonorrhoea 4: Chlamydia Trachomatis 5: Herpes Simplex 6: Crabs 7: Trichomonas Vaginalis 9: Not applicable 10: Unknown
Abnormal Pap smear in the past (Var: Abnormalpap)	Categorical	0: No 1: Yes 10: Unknown
Number of customers per day (Var: Nocustomer)	Categorical	0: 0-2 1: 3-5 2: 6-10 3: > 10 10: unknown
Drug use in the past (Var: Druginpast)	Categorical	0: No 1: Yes 10: Unknown
Drug use in the present (Var: Druginpresent)	Categorical	0: No 1: Yes 10: Unknown
Gonorrhoea at screening (Var: Gonorrhoea)	Categorical	0: Negative 1: Positive 10: Not available
Chlamydia at screening (var: Chlamydia)	Categorical	0: Negative 1: Positive 10: Not available
Cervix microbiology (Cervixmicrobiology)	Categorical	0: Negative 1: Gardnerella 2: Candida, Actinomyces, Trichomonas Vaginalis

Figure A-1: HPV Type Prevalence



Questionnaire A.2 : Survey Study Questionnaire

Questionnaire Gh@pro vzw Version 23/04/2007

PATIENT NUMMER:.....	
1. Age	……Years
2. Nationality	
3. Smoking behaviour	Yes / No
4. How many children (number)	
5. a) Oral contraception?	Yes / No
b) If Yes, for how long?	
c) If No, other contraception?	Yes / No
d) Specify which other contraception	Only condoom/ IUCD/ Injection/ Patch/ Implant at arm/ Nuvaring
6. How old at first sexual contact (years)	……Years
7. How many private sexual partners in last 12 months	0 1 2-5 6-10 >10
8. a) Private condom use	Yes / No
b) If Yes, against pregnancy or against STI?	STI / Pregnancy / Both
c) If Yes: always/mostly or sometimes	
9. Anal sex? Often/sometimes/never	
10. a) Ever had an STI?	Yes / No
b) If Yes which one : HIV/ Syphilis/ Hepatitis B/ Gonorrhea/ Chlamydia	HIV/ Syphilis/ Hepatitis B/ Gonorrhea/ Chlamydia
11. Ever had genital warts?	Yes / No
12. a) Ever had abnormal pap smear?	Yes / No
b) If Yes, what treatment??	
13. Ever had cancer of genital region?	Yes / No
14. a) At the moment anal problems? Fissurae, abcess…	Yes / No
b) If Yes for how long	
15. Ever had organ transplantation?	Yes / No

Questionnaire Gh@pro vzw Version 23/04/2007

Questions Specific for GHAPRO	
1. How long in sex work? (Years)	
2. What sector?	Bar/ Private/ Red light district/ Escort/ Street
3. a) Ever worked in other sector?	Yes / No
b) If Yes, which sector?	Bar/ Private/ Red light district/ Escort/ Street
4. How many days work a week	
5. How many clients a day	0-2 3-5 6-10 >10
6. What sexual techniques do you use?	
a) Vaginal?	Yes / No
b) Oral?	Yes / No
c) Anal?	Yes / No
7. Condom use per technique	
a) Vaginal	Always mostly sometimes never not applicable
b) Oral	Always mostly sometimes never not applicable
c) Anal	Always mostly sometimes never not applicable
8. Ever used drugs	Yes / No
a) If yes: iv or not iv?	IVD niet-IVD
9. You use currently drugs	Yes / No
If yes iv or non iv drug use	IVD niet-IVD

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Prevalence of high risk HPV in female sex workers in Antwerp, Belgium

Richting: **Master of Statistics-Biostatistics**
Jaar: **2013**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Musingila, Paul

Datum: **11/09/2013**