

## Identifying Predictors of Anal HPV Status in HPV-Vaccinated MSM: A Machine Learning Approach

Honorina Ocagli, Daniele Bottigliengo, Giulia Lorenzoni, Francesco Fontana, Camilla Negri, Gian Michele Moise, Dario Gregori & Libera Clemente

To cite this article: Honorina Ocagli, Daniele Bottigliengo, Giulia Lorenzoni, Francesco Fontana, Camilla Negri, Gian Michele Moise, Dario Gregori & Libera Clemente (2022): Identifying Predictors of Anal HPV Status in HPV-Vaccinated MSM: A Machine Learning Approach, Journal of Homosexuality, DOI: [10.1080/00918369.2022.2132574](https://doi.org/10.1080/00918369.2022.2132574)

To link to this article: <https://doi.org/10.1080/00918369.2022.2132574>



© 2022 The Author(s). Published with license by Taylor & Francis Group, LLC.



[View supplementary material](#)



Published online: 04 Nov 2022.



[Submit your article to this journal](#)



Article views: 336





[View related articles](#)



[View Crossmark data](#)

## Identifying Predictors of Anal HPV Status in HPV-Vaccinated MSM: A Machine Learning Approach

Honorina Ocagli, MA <sup>a</sup>, Daniele Bottigliengo, PhD<sup>a</sup>, Giulia Lorenzoni, PhD<sup>a</sup>, Francesco Fontana, MD<sup>b</sup>, Camilla Negri, MD<sup>c</sup>, Gian Michele Moise, MD<sup>c</sup>, Dario Gregori, PhD <sup>a</sup>, and Libera Clemente, MD<sup>b</sup>

<sup>a</sup>Unit of Biostatistics, Epidemiology and Public Health, Department of Cardiac, Thoracic, Vascular Sciences, and Public Health, University of Padova, Padova, Italy; <sup>b</sup>Division of Laboratory Medicine, University Hospital Giuliano Isontina (ASU GI), Trieste, Italy; <sup>c</sup>STI-AIDS Unit, University Hospital Giuliano Isontina (ASU GI), Trieste, Italy

### ABSTRACT

Anal human papillomavirus (HPV) infection has a high prevalence in men who have sex with men (MSM), resulting in an increased risk for anal cancer. The present work aimed to identify factors associated with HPV in a prospective cohort of HPV-vaccinated MSM using a random forest (RF) approach. This observational study enrolled MSM patients admitted to an Italian (sexually transmitted infection) STI-AIDS Unit. For each patient, rectal swabs for 28 different HPV genotype detection were collected. Two RF algorithms were applied to evaluate predictors that were most associated with HPV. The cohort included 135 MSM, 49% of whom were HIV-positive with a median age of 39 years. In model 1 (baseline information), age, age sexual debut, HIV, number of lifetime sex partners, STIs, were most associated with the HPV. In model 2 (follow-up information), age, age sexual debut, HIV, STI class, and follow-up. The RF algorithm exhibited good performances with 61% and 83% accuracy for models 1 and 2, respectively. Traditional risk factors for anal HPV infection, such as drug use, receptive anal intercourse, and multiple sexual partner, were found to have low importance in predicting HPV status. The present results suggest the need to focus on HPV prevention campaigns.


### KEYWORDS

MSM (men who have sex with men); epidemiology; quantitative methods; sexual behavior; health; sexual health; education

## Introduction

Human papillomavirus (HPV) infection is strongly associated with the risk of anogenital cancer. Globally, 35,000 cases of anal cancer and 13,000 cases of penile cancer were attributable to HPV infection in 2012 (De Martel et al., 2017). In Italy, the trend in the incidence of anal squamous cell carcinoma (ASCC) increased during the period 1988–2007 for both sexes (Islami et al., 2017). The incidence of anal cancer, despite being low in the general population, is increasing in developed countries (De Martel et al., 2012).

**CONTACT** Dario Gregori  [dario.gregori@unipd.it](mailto:dario.gregori@unipd.it)  Unit of Biostatistics, Epidemiology and Public Health, University of Padova, Via Loredan 18, Padova, Italy.

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/00918369.2022.2132574>

© 2022 The Author(s). Published with license by Taylor & Francis Group, LLC.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

The incidence of anal cancer is higher in women, men who have sex with men (MSM), immunosuppressed people (i.e., HIV-positive and organ transplant recipients), and women with gynecological HPV cancer or precancerous lesions (Bruni et al., 2019). A recent meta-analysis reported an incidence rate (IR) per 100,000 person-years of anal cancer in the MSM living with HIV of 85 (95% CI 82–89; Clifford et al., 2020). Anal HPV overall pooled prevalence among MSM reaches 78.4% in a recent systematic review (Farahmand et al., 2021).

In the case of HPV-related cervical cancer, HPV vaccination, before the start of sexual activity, has been shown to protect against infection with the strains covered by vaccine, although the results need further confirmation in the coming years. In Italy, the coverage for HPV vaccination in girls varies from 72% to 87%, depending on the characteristics of the vaccination program (Carrozzi et al., 2015). The same HPV vaccination strategy is used for both anal and cervical cancers, given their similarities of mucosa and development of similar precancerous changes. In an observational study, it was shown that HPV vaccination with bivalent HPV (bHPV), quadrivalent HPV (qHPV), and nonavalent HPV (nHPV) vaccines potentially prevented anal cancers by 79%, 90%, and 96%, respectively (Hillman et al., 2014). A recent systematic review showed that HPV vaccination reduced anal intraepithelial lesion grades 2 and 3 by 61.9% and 46.8%, respectively, in naïve males (Harder et al., 2018). Despite this interest, in Italy vaccination campaigns have mainly focused on young females, and they were extended to young males only in 2017. To reduce the incidence of HPV infection and consequently anal cancer, together with the implementation of vaccination campaigns, it is extremely important to identify factors related to HPV infection. In the literature, multiple sexual partners, receptive anal intercourse, and smoking habits are considered the factors that most affect HPV incidence (Machalek et al., 2012). These factors are usually retrieved from descriptive studies, which are typically assessed using traditional statistical methods. In recent years, machine learning techniques (MLTs) have been increasingly used for building classification models that predict HPV status, for example, in oropharyngeal squamous cell carcinoma (Ren et al., 2020). MLTs have also been used for evaluating sentiment analysis on HPV vaccine-related tweets (Du et al., 2017).

The present work aimed to identify factors associated with HPV in a prospective cohort of HPV-vaccinated MSM in the STI-AIDS Unit of Gorizia's Hospital in northern Italy using a random forest (RF) approach. The results of the RF algorithm are then used to profile patients in terms of HPV risk.

## Materials and methods

### Study design

This observational study enrolled MSM patients admitted to the (sexually transmitted infection) STI-AIDS Unit of Gorizia's Hospital from

January 2017 and is still ongoing. This study is part of a larger collaboration among Fondazione Cassa di Risparmio di Gorizia (Carigo) and the STI-AIDS Unit of the University Hospital Giuliano Isontina and the University of Padova. The primary objective of this collaboration is to evaluate HPV vaccination in the MSM population admitted to the center.

### **Setting**

The study was conducted in the STI-AIDS Unit of University Hospital Giuliano Isontina, which is part of the national surveillance network for sexually transmitted infections (STIs), coordinated by the *Istituto Superiore di Sanità*. In Friuli Venezia Giulia, the National HPV Immunization program was launched in 2008 for girls between 12 and 15 years of age as the primary target (D.G.R, 856, 2008); it was then extended to HIV-positive people, boys born since 2004, and MSM people since 2015 (DGR, 2014).

### **Study population**

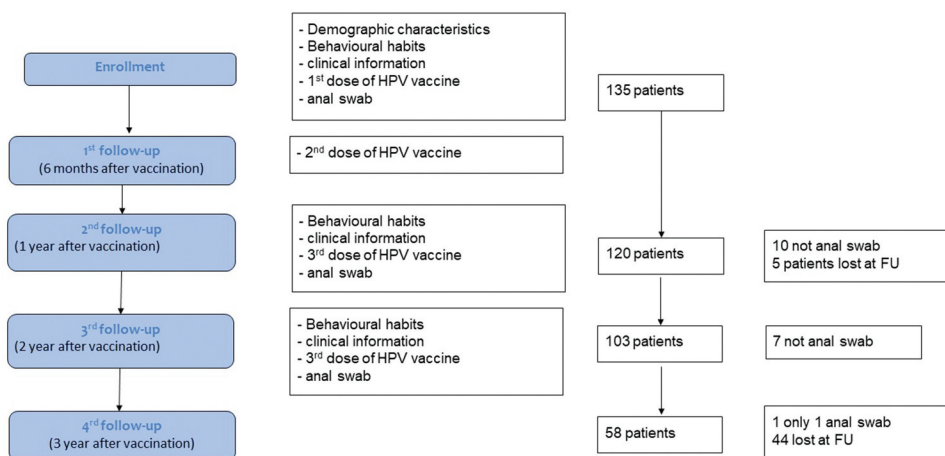
Participants were eligible if they *i*) provided written informed consent, *ii*) were 18 years old or older, *iii*) were admitted consecutively to the center, and *iv*) did not receive previous HPV vaccination at baseline assessment.

### **Study procedure**

Demographic information (nationality, level of education) was collected for each participant at baseline. Behavioral habits (age first intercourse, circumcision, risk factors, sexual behavior, use of sexual protection), and clinical information (HIV status, comorbidities, drug use) were collected at baseline assessment and each follow-up (Figure 1). HPV vaccination was administered at baseline (first dose), first follow-up (second dose), and the second follow-up (third dose). In the beginning, only the qHPV vaccine was used, followed by the introduction of the nHPV formulation in our patients.

### **Samples collection**

Anal samples were collected by inserting a swab 3 cm into the anal canal by an expert physician at baseline and each follow-up. The anal sample was tested for 28 HPV genotypes (6, 11, 16, 18, 26, 31, 33, 35, 39, 40, 42, 43, 44, 45, 51, 52, 53, 54, 56, 58, 59, 61, 66, 68, 69, 70, 73, 82) with *Anyplex<sup>tm</sup> II HPV28 detection* (Seegene, Seoul, Korea) according to the classification proposed by the International Agency for Research on Cancer of the World Health Organizations and for sexually transmitted infections (*Chlamydia trachomatis*, *Neisseria gonorrhoea*, *Mycoplasma hominis*, *Mycoplasma genitalium*,



**Figure 1.** Flowchart of the study.

*Trichomonas vaginalis*, *Ureaplasma urealyticum*, and *Ureaplasma parvum*) with Anyplex™ II STI-7 Detection (Seegene, Seoul, Korea).

### **Ethical aspects**

The study was approved by the ethics committee of the Friuli Venezia Giulia region with protocol number 33,298. Personal data were used according to both Italian and European regulations on this matter.

### **Random forest models**

RF algorithm, one of the most popular methods among MLTs, was developed by Leo Breiman for prediction tasks (Breiman, 2001). It belongs to the family of ensemble-of-trees methods, a class of algorithms that models the relationships between a set of predictors and an outcome by aggregating the results of classification and regression trees (CARTs). CARTs work by dividing the space of predictors into different regions using recursive binary splitting, and they predict the expected outcome in each region. RF is typically implemented by fitting a CART on several bootstrap replicates of the original dataset. By building each CART using only a random subset of the predictors, the algorithm allows each tree to be very different from the other. The random subset of predictors, which is usually called *mtry* in the RF literature, is the main parameter of the algorithm. Despite default values have been proposed for different situations, the parameter value is typically determined using the out-of-bag (OOB) prediction error, i.e., the value of *mtry* that minimizes the prediction error of the algorithm on the set of observations discarded by each bootstrap samples. Results returned by each CART are then averaged to obtain the final prediction. The strength of the RF algorithm is to

exploit the information of several trees that singularly might have “weak” performances but can become very powerful if pooled together (Hastie et al., 2009).

Two RF algorithms were created to identify patients positive for HPV based on a selected set of predictors. The first model (Model 1) was used to detect subjects positive for HPV using data collected at baseline. The second model (Model 2) aimed to profile individuals who were vaccinated in terms of the probability of being positive for HPV using both the information collected at baseline and at each follow-up.

Model 1 was tuned by choosing the parameter value that minimized the OOB prediction error.

The parameter value of Model 2 was chosen such that it minimized the cross-validation (CV) prediction error. Model 2 was tuned using repeated k-fold CV, using fivefolds, and repeating the process 10 times (Hastie et al., 2009). To address dependency among the observations of the same subjects 2, Model 2 was trained using a stratified sampling scheme considering the identifiers of patients as strata.

#### ***Variable importance in predicting the outcome***

Predictors were ranked in terms of the strength of their association with the outcome using the decrease of the prediction error (DPE) obtained with the permutation approach (Ishwaran, 2007). Briefly, the approach works as follows: first, an RF algorithm is built using a copy of the variable of interest obtained by permuting the observations, thus removing the eventual association of the variable with the outcome; then, a second RF algorithm with the original predictor is fitted; finally, the DPE of the second model to the first model is computed: the higher the DPE, the higher the strength of the association of the predictor. The relationships between the most associated predictors and the probability of being positive for HPV were explored using partial dependence plots (PDPs; Friedman, 2001). Both RF algorithms were created using 10,000 trees.

#### ***Imbalance control and missing data imputation***

To handle the high unbalanced outcome in model 1 (ratio of positive cases to negative cases much larger than one), the random forests quantile-classifier (RFQ) was employed (O’Brien & Ishwaran, 2019). In model 2, the imbalance in the outcome was handled using the SMOTE method, which oversamples synthetic cases of the minority class (Chawla et al., 2002).

Missing data were handled using an RF algorithm that imputes missingness with a multivariate unsupervised split approach (percentages of missing data; Tang & Ishwaran, 2017).

## **Variables**

The variables included in the model are the following: age, age sexual debut, HIV status (negative/positive), STI class (bacterial, viral), number of lifetime sex partners, multiple sexual partner (yes/no), smoker (yes/no), occasional partner (yes/no), use of medication (yes/no), use of drugs (yes/no), oro-genital sexual contact (yes/no), circumcision (yes/no), swallowing sperm (yes/no), risk factors (yes/no if at least one among tattoo, piercing, blood transfusion, incarceration is reported), comorbidities (yes/no), last risk intercourse (days), sexual positioning practices (exclusively insertive, exclusively receptive, or both), condom use (yes/no), other vaccination (yes/no), educational level (high school, degree, secondary school), number sex partner since last follow-up ( $\leq 1$ , 2–5,  $\geq 6$ ), anogenital contact, orogenital contact. A patient was defined HPV positive if at least positive to one of the 28 genotypes tested at each follow-up.

## **Patient profiling**

To show how the RF algorithm can be used to characterize patients in terms of HPV risk, HPV risk profiles were identified for three hypothetical patients based on the strongest predictors according to the algorithm and their values observed in the sample of the study. For each profile, the algorithm was used to compute the expected risk of HPV given the values of the predictors. This approach can be potentially extended to any type of patient to aid physicians in targeting the best treatment strategies according to the characteristics of individuals.

## **Statistical analysis**

Descriptive statistics were reported as absolute numbers and percentages (%) for categorical variables and as median values (I and III quartiles) for continuous variables.

Predictive performances were assessed using the following metrics for both models: area under the ROC curve (AUC), accuracy, sensitivity, specificity, positive predictive values (PPV), and negative predictive values (NPV). The Wald test was used to assess significant differences between the performance measures of the two models.

Statistical analysis was performed using R software for statistical computing (version 4.0.2; R Core Team, 2020). The baseline model was fitted using the *randomForestSRC* R package (version 2.9.3; Ishwaran et al., 2020), and the follow-up model was trained using the *caret* (version 6.0–86) and *randomForest* (version 4.16–14) R packages (Cutler & Wiener, 2018; Kuhn et al., 2020).

## Results

### Study population

The presented analysis is based on data from the first 135 patients enrolled in the STI-AIDS Unit of University Hospital Giuliano Isontina. Baseline characteristics are reported in [Table 1](#) according to HIV status.

Patients had a median age of 43 years old (1<sup>o</sup>Q 36, 3<sup>o</sup>Q 50). Only 10% of the population (14) was below 26 years old, and the category 36–45 years was the most represented (53, 39%). The sample is primarily composed of homosexuals (121, 90%). The median age of sexual debut was 18 years old (16, 23). Forty-seven percent of participants were HIV positive (N = 63). Among the HIV-positive and HIV-negative groups at baseline, there were no statistically significant differences in other variables.

### HPV prevalence

The overall prevalence of HPV in our sample ranged from 88% at the second follow-up to 93% at the fourth follow-up. HPV was most

**Table 1.** Descriptive characteristics of the sample at baseline assessment. Continuous variables were reported as I, II (median), and III quartiles, categorical variables were presented as absolute numbers and percentages.

Variable	Variable levels	N	Statistics
Age		135	36/43/50
Age categories	≤ 26	135	10% (14)
	26–35		12% (16)
	36–45		39% (53)
	46–55		24% (33)
	> 55		14% (19)
Nationality	European	135	11% (7)
	Italian		86% (54)
	Other		3% (2)
Educational level	High school	135	58% (78)
	Degree		29% (39)
	Secondary school		13% (18)
Self reported sexual preference	Straight/heterosexual	135	10% (13)
	Gay/homosexual		90% (121)
	Transsexual		1% (1)
			1% (2)
Received money for sex		134	1% (2)
Age sexual debut		132	16/18/23
Circumcision		135	7% (9)
Risk factors: tattoo		135	18% (24)
Risk factors: piercing		135	15% (20)
Risk factors: blood transfusion		135	1% (2)
Risk factors: incarceration		135	1% (1)
Time HIV (years)		62	3.2/7.5/12.0
Age diagnosis HIV		62	29/36/44
Immunodepression CD4 cells	< 350	62	26% (16)
	>500		50% (31)
	350–500		24% (15)
			47% (63)
HIV status	Positive	135	47% (63)
	Negative	135	53% (72)



**Table 2.** Distribution of HPV genotypes at each follow-up according to HIV status. For each category, percentages and absolute value % (N) are reported.

	Baseline	2nd Follow up	3rd Follow up	4th Follow up
	(N = 134)	(N = 120)	(N = 103)	(N = 58)
HPV positive	91% (122)	88% (105)	92% (95)	93% (54)
Genotype qHPV vaccine <sup>1</sup>	57% (76)	48% (58)	54% (56)	53% (31)
Genotype nHPV vaccine <sup>2</sup>	73% (98)	64% (77)	74% (76)	71% (41)
High risk genotype <sup>3</sup>	81% (108)	79% (81)	79% (81)	83% (48)

<sup>1</sup>qHPV: genotypes 6, 11, 16, and 18, <sup>2</sup>nHPV genotype vaccine: genotypes 6, 11, 16, 18, 31, 33, 45, 52, and 58; <sup>3</sup>high-risk genotype: genotypes 16, 18, 31, 33, 45, 52, 58, 35, 39, 51, 56, 59, 66, and 68

prevalent in HIV-positive patients at each follow-up, ranging from 94% to 97%. The same was true for qHPV (genotypes 6, 11, 16, and 18), nHPV genotype vaccine (genotypes 6, 11, 16, 18, 31, 33, 45, 52, and 58), and high-risk genotype (genotypes 16, 18, 31, 33, 45, 52, 58, 35, 39, 51, 56, 59, 66, and 68; [Table 2](#)).

### Model performances

Model 1 detects subjects positive for HPV considering data collected at baseline, including demographic characteristics, clinical characteristics, and behavioral characteristics. A total of 134 patients were included in the first model, 122 HPV positive and 12 HPV negative. The percentage of missingness in the data used by Model 1 was 3.1%. Model 2 considered all patients at each follow-up for a total of 415 observations, 376 positives, and 39 negatives for HPV. The percentage of missing data was 3.8% for Model 2. In [Table 3](#), the performances of the RF models are reported.

The accuracy in model 1, i.e., 0.61, was significantly lower than model 2, i.e., 0.81 (p-value <0.001). Model 2 showed a higher sensitivity value (0.83) than model 1 (0.57) (p-value <0.001), whereas similar specificity values were observed for both models (0.58 for model 1 and 0.55 for model 2, p-value = 1). Model 1 exhibited low discrimination (AUC of 0.63), whereas an AUC of 0.77 was observed for model 2, suggesting a good discrimination ability when the information at different follow-ups was used.

**Table 3.** Performances of random both forest model. Sensitivity, specificity, predictive positive value (PPV), accuracy, and negative predictive value (NPV) were calculated using a confusion matrix.

	Sensitivity	Specificity	PPV	NPV	Accuracy	AUC
<b>Model 1</b>	0.57	0.58	0.93	0.12	0.61	0.63
<b>Model 2</b>	0.83	0.55	0.95	0.27	0.81	0.77

**Importance of predictors**

Figure 2 shows the importance of the selected variables in predicting the outcome of HPV. The importance is presented as percentages in the DPE associated with each variable.

In model 1, age (DPE 9.94%), age of sexual debut (DPE 6.64%), HIV (DPE 5.84%), and STIs (DPE 1.25%) were most associated with the outcome. In model 2, age (DPE 19.09%), age of sexual debut (DPE 14.27%), HIV (DPE 12.33%), STI class (DPE 12.33%), and follow-up (DPE 4.69%) were included (Figure 3).

People aged 40 to 50 at baseline appeared to be more likely to be HPV positive. In model 1, the risk of HPV positive was approximately 90% (Figure 1). In model 2, a maximum of 80% was achieved at the third follow-

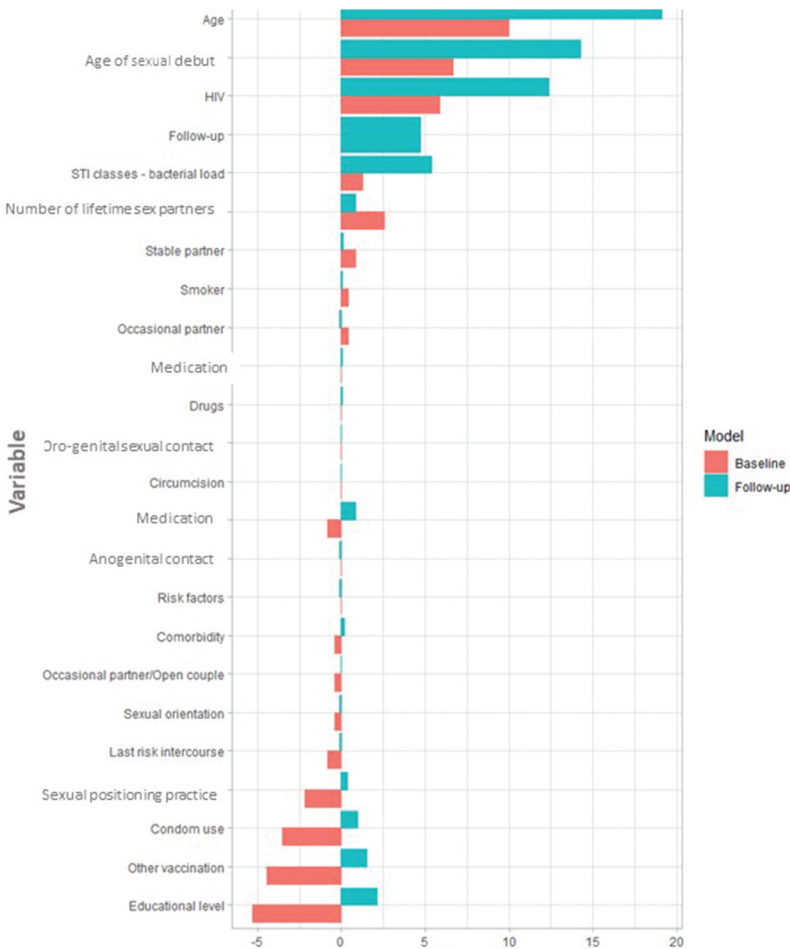
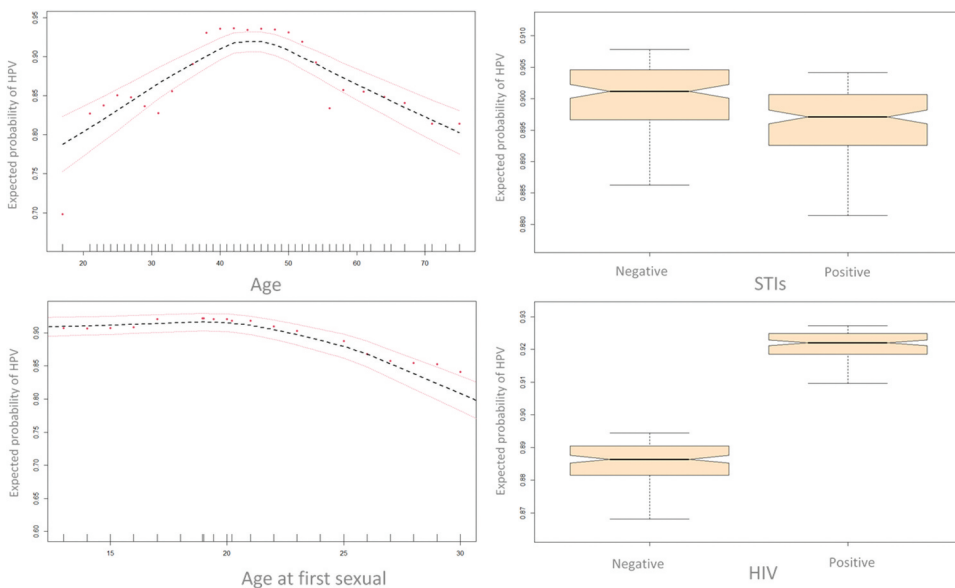


Figure 2. Variable importance in the models. For each model, the variable and percentage of the ability to decrease the performance model were reported.



**Figure 3.** Association between the first four most important variables and predicted risk of HPV in model 1.

up. In both models, the probability of being positive for HPV decreased with increasing age (Figure 4).

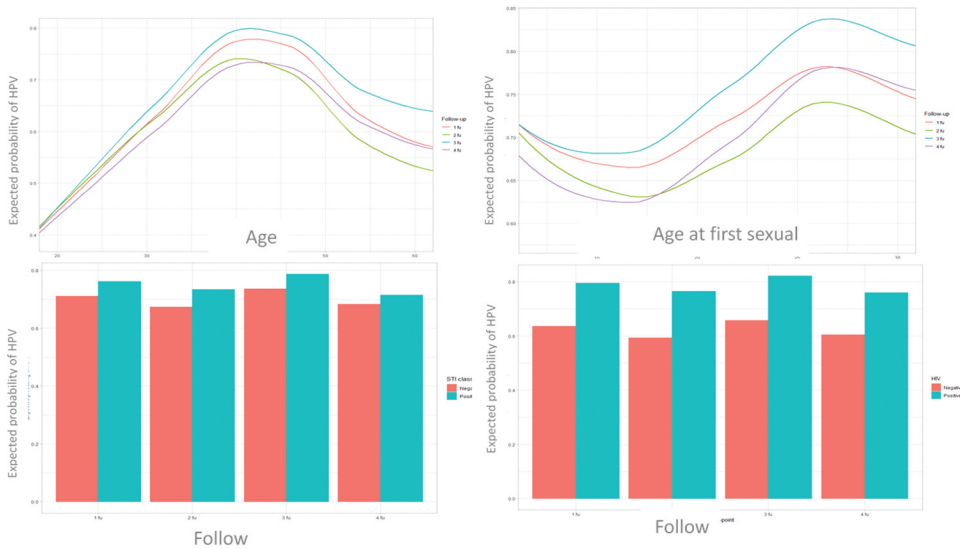
Regarding the age of sexual debut, the risk of having HPV was greater for age below 20 years in model 1, with a probability of nearly 90%. In model 2, the risks of having HPV were between 70% and 81% if the age of sexual debut was between 18 and 28 years. Higher risks were observed for individuals aged less than 15 years old.

In model 2, patients with concomitant bacterial STIs exhibited increased probabilities of having HPV compared to those negative for any STIs.

HIV-positive individuals had a greater probability of developing HPV in both models. In model 1, HIV positivity was associated with a 92% risk of developing HPV, whereas, in model 2, the risk of HIV positivity being associated with HPV positivity was lower to a maximum of nearly 80% at the third FU.

### **Patient profiling**

Results of the variable importance in Model 1 and the partial dependency plots allowed us to identify three different levels of risk for HPV: low, medium, and high. Supplementary Figure S1 reports the primary characteristics for each level of risk. Therefore, in a clinical setting, clinicians can define the level of risk of already having HPV according to baseline characteristics. For example, a patient with a high risk of developing HPV, according to the results of our



**Figure 4.** Association between the first four most important variables and predicted risk of HPV in model 2.

model, is 45 years of age, had his sexual debut at 15 y of age, is positive for HIV and STI, has more than six partners, and has behavioral risks, such as occasional partner and smoking habits. In contrast, patients aged 20 years with their sexual debut at 19 and negative for both HIV and STI with no risky behavior had a lower risk of HPV infection. The probability of developing HPV according to Model 1 was 95%, 89%, and 75% for patients at high, medium, and low risk, respectively.

## Discussion

The primary results of this study are as follows: *i*) the prevalence of HPV ranged from 88% at baseline assessment to 93% at the fourth follow-up in a cohort of 134 patients; *ii*) age and age of sexual debut were the most important variables for predicting the risk of HPV, with 9.95% and 6.63% MDP in model 1 and 19.09% and 14.27% MDP in model 2, respectively; *iii*) STIs and HIV were relevant for HPV infection; *iv*) traditional risk factors for HPV infection were considered less relevant by our models.

In our sample, the prevalence of any HPV tested was higher than the estimates reported in the literature. The prevalence in anal site in the meta-analysis of Farajmand et al. (Farahmand et al., 2021) were 78.4% (95% CI: 75.6%–81.0%). In China Zhou et al. in their meta-analysis reported an estimated prevalence of anal HPV of 85.1% in HIV-positive and 53.6% in HIV-negative. In the Italian MSM population, the prevalence of HPV at anal sites was estimated as 56% (95% CI 41.3–70.0; Sammarco et al., 2016) and 93%

(95% CI 88.1–96.3; Donà et al., 2015) in HIV-positive populations. Similar levels have been reported in HIV-negative patients: 72.1% (95% CI 67.6–76.2; Donà et al., 2015) to 88.9% (95% CI 51.8–99.7; Pierangeli et al., 2008). The higher prevalence reported in our study may be related to the fact that we have considered a patient positive for HPV if he was positive for any genotype for he was tested for.

Our results showed that age and age of sexual debut were among the most relevant factors that affected HPV onset both before and after the vaccination. People between 40 and 50 years of age in our sample experienced less benefit from recent vaccination and HPV prevention campaigns. This might be related to the long period of sexual activity, and, therefore, they have been exposed to the risk of infections longer. Another reason may be the recent awareness of the problem. Indeed, the inclusion of males in existing HPV vaccination programs, along with females, has recently begun, given the similarity between anal and cervix cancers (Harder et al., 2018; De Martel et al., 2017; Schmeler & Sturgis, 2016). HPV vaccination for males starting from the 2006 cohort is freely available in all Italian territories with the National Vaccine Prevention Plan 2017–2019. In the Friuli Venezia Giulia region, the HPV vaccination coverage for males was extended to males born in 2004. The coverage was 52.68% in 2017, slightly lower than the female coverage in the same cohort in the same year, 64.57%.

Young age at the time of sexual debut was confirmed to be a relevant risk factor for HPV acquisition, as shown in a previous study (Frisch et al., 1997; De Martel et al., 2017). This might be related to the adolescents being unaware of HPV and, more generally, of STIs (Ciccarese et al., 2020; Samkange-Zeeb et al., 2013). This lack of knowledge may be related to the decrease in the age of sexual debut and, eventually, to educators and media focusing most of their attention on HIV prevention. On the other hand, the presence of a previous sexually transmitted infection is relevant for both models. STIs are responsible for an estimated 376 million new infections (World Health Organization, 2018), and the most frequent is the bacterial form (i.e., chlamydia, gonorrhea, syphilis, and trichomoniasis; World Health Organization, 2019).

HIV status in model 2, after vaccination, seems to have a relevant impact on HPV status, i.e., HIV-positive individuals have a nearly 0.10 higher probability of being positive for HPV than HIV-negative subjects. HIV-positive status is considered one of the most important factors affecting anal cancer, especially in MSM, as confirmed by a recent meta-analysis (Clifford et al., 2020).

Traditional risk factors for HPV infection, such as drug use, receptive anal intercourse, and multiple sexual partner, were found to have lower importance in predicting HPV status in our classification algorithms, in contrast with other studies (Giuliano et al., 2010; Kang et al., 2018). These results may be related to the imbalance of these variables in our dataset, a feature of the

sample that might potentially affect the direction of the association of these variables in predicting HPV. For example, in the present study, although individuals older than other studies were enrolled, these subjects were poorly represented in the sample, leading to more variable estimates at the extremes of the distributions.

### **Limitations**

The sample considered in this study has a large proportion of HPV infection, such that the low differences in risk estimates still corresponded to clinically significant differences in the probability of infection. Larger samples should be considered to recalibrate our estimates and to verify the direction of the association of some factors, especially those related to behaviors. Moreover, it would be interesting to evaluate different models for HIV+ and HIV- people with larger sample sizes, since HIV induces huge immunological changes that could influence HPV predictors. The low sample size poses some limitations to the generalizability of the present study. Future research with higher sample sizes is needed to guarantee the replicability of the present findings.

Our results showed that there is a great prevalence of HPV positivity in the MSM population. This might be related to the fact that the vaccination campaign is recent, even for girls. Our MSM population was unaffected by the campaign's prevention, as they had already started their sex life when the importance of HPV prevention begins to spread. Furthermore, the MSM population was not affected by herd immunity, as might be the case for girls. Consequently, considering that the MSM population is at high risk for HPV, it is important to implement an awareness campaign for sexual behaviors that can minimize the risk of HPV infection. In particular, prevention should start early, when boys are not yet actively involved in sexual intercourse, as stated by other authors (Brotherton et al., 2016; Sundaram et al., 2020).

Considering the peculiarity of this population, it would be interesting to evaluate the effects of vaccination in patients who have not already engaged in sexual intercourse. In these situations, it is clear that vaccination loses its prophylactic scope. Nevertheless, evaluation of an eventual curative effect should be considered. Of course, this assessment will be performed in subsequent years, since vaccination of this population has only recently begun.

### **Conclusions**

Identification of the variables that most influence HPV onset can help to define the patient profile and consequently adapt prevention campaigns for characteristics of the patients that are most at risk. In our case, for example, knowing that HPV prevalence is higher in older subjects suggests that

developing additional strategies to lower the impact of HPV consequences and evaluate relapses is beneficial. Further developments of our work may focus on profiling patients at risk for infection with high-risk genotypes.

## Acknowledgments

D.G. conceived and supervised the study. D.B. and H.O. completed the analysis. H. O. synthesized analyses and led the writing. L.C. and G.L. assisted with the study and the writing. C.N., F.F., and G.M.M were responsible for data collection. All authors discussed the results and contributed to the final manuscript.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

This work was supported by the Fondazione cassa di risparmio di Gorizia (carigo) [U1401.2018/AI.952.MF 8.11.2018].

## ORCID

Honoria Ocagli  <http://orcid.org/0000-0002-5823-1659>

Dario Gregori  <http://orcid.org/0000-0001-7906-0580>

## Ethical approval

The study was approved by the ethics committee of the Friuli Venezia Giulia region with protocol number 33,298.

## Patient consent

Each respondent provided written informed consent before starting the survey.

## References

- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Brotherton, J. M. L., Giuliano, A. R., Markowitz, L. E., Dunne, E. F., & Ogilvie, G. S. (2016). Monitoring the impact of HPV vaccine in males—Considerations and challenges. *Papillomavirus Research*, 2, 106–111. <https://doi.org/10.1016/j.pvr.2016.05.001>
- Bruni, L., Albero, G., Serrano, B., Mena, M., Collado, J.J., Gómez, D., and de Sanjosé S. (2019). *ICO/IARC Information Centre on HPV and Cancer (HPV Information Centre). Human Papillomavirus and Related Diseases in the World*. Summary Report 22 October 2021.



- Carrozzi, G., Sampaolo, L., Bolognesi, L., Sardonini, L., Bertozzi, N., Rossi, P. G., Ferrante, G., Campostrini, S., Ferrante, G., Masocco, M., Minardi, V., D'Argenzio, A., Moghadam, P. F., Quarchioni, E., Ramigni, M., Trinito, M. O., Salmaso, S., & Zappa, M. (2015). Cancer screening uptake: Association with individual characteristics, geographic distribution, and time trends in Italy. *Epidemiol Prev*, 39(Suppl 1), 9–18. 26405772. [https://epiprevit.serversiucuro.it/materiali/2015/EP2015\\_I3S1\\_009.pdf](https://epiprevit.serversiucuro.it/materiali/2015/EP2015_I3S1_009.pdf)
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321–357. <https://doi.org/10.1613/jair.953>
- Ciccarese, G., Drago, F., Herzum, A., Reborà, A., Cogorno, L., Zangrillo, F., & Parodi, A. (2020). Knowledge of sexually transmitted infections and risky behaviors among undergraduate students in Tirana, Albania: Comparison with Italian students. *Journal of Preventive Medicine and Hygiene*, 61(1), E3–E5. <https://doi.org/10.15167/2421-4248/jpmh2020.61.1.1413>
- Clifford, G. M., Georges, D., Shiels, M. S., Engels, E. A., Albuquerque, A., Poynten, I. M., . . . Stier, E. A. (2020). A meta-analysis of anal cancer incidence by risk group: Toward a unified anal cancer risk scale. *International Journal of Cancer*, 1, 11. <https://doi.org/10.1002/ijc.33185>
- Cutler, F., original by L. B. and A., & Wiener, R., port by A. L. and M. (2018). *randomForest: Breiman and cutler's random forests for classification and regression* (Version 4.6-14). <https://CRAN.R-project.org/package=randomForest>
- de Martel, C., Ferlay, J., Franceschi, S., Vignat, J., Bray, F., Forman, D., & Plummer, M. (2012). Global burden of cancers attributable to infections in 2008: A review and synthetic analysis. *The Lancet. Oncology*, 13(6), 607–615. [https://doi.org/10.1016/S1470-2045\(12](https://doi.org/10.1016/S1470-2045(12)
- de Martel, C., Plummer, M., Vignat, J., & Franceschi, S. (2017). Worldwide burden of cancer attributable to HPV by site, country and HPV type. *International Journal of Cancer*, 141(4), 664–670. <https://doi.org/10.1002/ijc.30716>
- DGR. (2014, December 18). *DGR 2535 Aggiornamento ed estensione dell'offerta vaccinale nella regione FVG*. [http://mtom.regione.fvg.it/storage//2014\\_2535/Testo%20integrale%20della%20Delibera%20n%202535-2014.pdf](http://mtom.regione.fvg.it/storage//2014_2535/Testo%20integrale%20della%20Delibera%20n%202535-2014.pdf)
- D.G.R., 856. (2008). *Programma regionale di vaccinazione antipapilloma virus per la prevenzione dello sviluppo del tumore del collo dell'utero*. <https://bur.regione.fvg.it/newbur/visionaBUR?bnum=2008/05/28/22>
- Donà, M. G., Gheit, T., Latini, A., Benevolo, M., Torres, M., Smelov, V., McKay-Chopin, S., Giglio, A., Cristaudo, A., Zaccarelli, M., Tommasino, M., & Giuliani, M. (2015). Alpha, beta and gamma human papillomaviruses in the anal canal of HIV-infected and uninfected men who have sex with men. *The Journal of Infection*, 71(1), 74–84. <https://doi.org/10.1016/j.jinf.2015.02.001>
- Du, J., Xu, J., Song, H., Liu, X., & Tao, C. (2017). Optimization on machine learning based approaches for sentiment analysis on HPV vaccines related tweets. *Journal of Biomedical Semantics*, 8(1), 9. <https://doi.org/10.1186/s13326-017-0120-6>
- Farahmand, M., Monavari, S. H., & Tavakoli, A. (2021). Prevalence and genotype distribution of human papillomavirus infection in different anatomical sites among men who have sex with men: A systematic review and meta-analysis. *Reviews in Medical Virology*, 31(6). <https://doi.org/10.1002/rmv.2219>
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189–1232. <https://doi.org/10.1214/aos/1013203451>
- Frisch, M., Glimelius, B., van den Brule, A. J., Wohlfahrt, J., Meijer, C. J., Walboomers, J. M., Goldman, S., Svensson, C., Adami, H.-O., & Melbye, M. (1997). Sexually transmitted



- infection as a cause of anal cancer. *The New England Journal of Medicine*, 337(19), 1350–1358. <https://doi.org/10.1056/NEJM199711063371904>
- Giuliano, A. R., Anic, G., & Nyitray, A. G. (2010). Epidemiology and pathology of HPV disease in males. *Gynecologic Oncology*, 117(2), S15–19. <https://doi.org/10.1016/j.ygyno.2010.01.026>
- Harder, T., Wichmann, O., Klug, S. J., van der Sande, M. A. B., & Wiese-Posselt, M. (2018). Efficacy, effectiveness and safety of vaccination against human papillomavirus in males: A systematic review. *BMC Medicine*, 16(1), 110. <https://doi.org/10.1186/s12916-018-1098-3>
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction, second edition* (2nd ed.). Springer-Verlag. <https://www.springer.com/gp/book/9780387848570>
- Hillman, R. J., Garland, S. M., Gunathilake, M. P. W., Stevens, M., Kumaradevan, N., Lemech, C., Ward, R. L., Meagher, A., McHugh, L., Jin, F., Carroll, S., Goldstein, D., Grulich, A. E., & Tabrizi, S. N. (2014). Human papillomavirus (HPV) genotypes in an Australian sample of anal cancers. *International Journal of Cancer*, 135(4), 996–1001. <https://doi.org/10.1002/ijc.28730>
- Ishwaran, H. (2007). Variable importance in binary regression trees and forests. *Electronic Journal of Statistics*, 1, 519–537. <https://doi.org/10.1214/07-EJS039>
- Ishwaran, H., Kogalur, U. B., & Kogalur, M. U. B. (2020). Package ‘randomForestSRC’ (Version 2.9.3).
- Islami, F., Ferlay, J., Lortet-Tieulent, J., Bray, F., & Jemal, A. (2017). International trends in anal cancer incidence rates. *International Journal of Epidemiology*, 46(3), 924–938. <https://doi.org/10.1093/ije/dyw276>
- Kang, Y.-J., Smith, M., & Canfell, K. (2018). Anal cancer in high-income countries: Increasing burden of disease. *PLOS ONE*, 13(10), e0205105. <https://doi.org/10.1371/journal.pone.0205105>
- Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., . . . Hunt, T. (2020). *Caret: classification and regression training* (Version 6.0-86). <https://CRAN.R-project.org/package=caret>
- Machalek, D. A., Poynten, M., Jin, F., Fairley, C. K., Farnsworth, A., Garland, S. M., Hillman, R. J., Petoumenos, K., Roberts, J., Tabrizi, S. N., Templeton, D. J., & Grulich, A. E. (2012). Anal human papillomavirus infection and associated neoplastic lesions in men who have sex with men: A systematic review and meta-analysis. *The Lancet. Oncology*, 13(5), 487–500. [https://doi.org/10.1016/S1470-2045\(12\)70080-3](https://doi.org/10.1016/S1470-2045(12)70080-3)
- O’Brien, R., & Ishwaran, H. (2019). A random forests quantile classifier for class imbalanced data. *Pattern Recognition*, 90, 232–249. <https://doi.org/10.1016/j.patcog.2019.01.036>
- Pierangeli, A., Scagnolari, C., Degener, A. M., Bucci, M., Ciardi, A., Riva, E., Vullo, V., D’Ettore, G., Vullo, V., Antonelli, G., & Indinnimeo, M. (2008). Type-specific human papillomavirus-DNA load in anal infection in HIV-positive men. *Aids*, 22(15), 1929–1935. <https://doi.org/10.1097/QAD.0b013e32830fbd7a>
- R Core Team. (2020). *R: A language and environment for statistical computing* (Version 4.0.2). R Foundation for Statistical Computing. <https://www.R-project.org/>
- Ren, J., Yuan, Y., Qi, M., & Tao, X. (2020). Machine learning-based CT texture analysis to predict HPV status in oropharyngeal squamous cell carcinoma: Comparison of 2D and 3D segmentation. *European Radiology*, 30(12), 6858–6866. <https://doi.org/10.1007/s00330-020-07011-4>
- Samkange-Zeeb, F., Mikolajczyk, R. T., & Zeeb, H. (2013). Awareness and knowledge of sexually transmitted diseases among secondary school students in two German cities. *Journal of Community Health*, 38(2), 293–300. <https://doi.org/10.1007/s10900-012-9614-4>
- Sammarco, M. L., Ucciferri, C., Tamburro, M., Falasca, K., Ripabelli, G., & Vecchiet, J. (2016). High prevalence of human papillomavirus type 58 in HIV infected men who have sex with

- men: A preliminary report in Central Italy. *Journal of Medical Virology*, 88(5), 911–914. <https://doi.org/10.1002/jmv.24406>
- Schmeler, K. M., & Sturgis, E. M. (2016). Expanding the benefits of HPV vaccination to boys and men. *The Lancet*, 387(10030), 1798–1799. [https://doi.org/10.1016/S0140-6736\(16\)30314-2](https://doi.org/10.1016/S0140-6736(16)30314-2)
- Sundaram, N., Voo, T. C., & Tam, C. C. (2020). Adolescent HPV vaccination: Empowerment, equity and ethics. *Human Vaccines & Immunotherapeutics*, 16(8), 1835–1840. <https://doi.org/10.1080/21645515.2019.1697596>
- Tang, F., & Ishwaran, H. (2017). Random forest missing data algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 10(6), 363–377. <https://doi.org/10.1002/sam.11348>
- World Health Organization. (2018). *Report on global sexually transmitted infection surveillance 2018*.
- World Health Organization. (2019, June 14). *Sexually transmitted infections (STIs)*. Retrieved 14 October 2020, from [https://www.who.int/news-room/fact-sheets/detail/sexually-transmitted-infections-\(stis\)](https://www.who.int/news-room/fact-sheets/detail/sexually-transmitted-infections-(stis))