

HPV detection patterns in young women from the PAPCLEAR longitudinal study: implications for HPV screening policies.

Thomas Beneteau^{1,*}, Soraya Groc², Carmen Lia Murall³, Vanina Boué¹, Baptiste Elie¹,
Nicolas Tessandier⁴, Claire Bernat^{1,5}, Marine Bonneau⁶, Vincent Foulongne², Christelle Graf⁶,
Sophie Grasset¹, Massilva Rahmoun¹, Michel Segondy², Vincent Tribout⁷, Jacques Reynes⁸,
Christian Selinger^{1,9}, Nathalie Boulle⁶, Ignacio G. Bravo¹, Mircea T. Sofonea^{2,10}, Samuel Alizon^{1,4,*}

¹ MIVEGEC, Univ Montpellier, CNRS, IRD, France

² PCCEI, Univ Montpellier, Inserm, EFS, Montpellier, France

³ Department of Biological Sciences, Université de Montréal, Montréal, Canada

⁴ Center for Interdisciplinary Research in Biology (CIRB), Collège de France, CNRS, INSERM, Université PSL,
Paris, France

⁵ Institut de Génomique Fonctionnelle, Université de Montpellier, CNRS, INSERM, Montpellier, France

⁶ Department of Obstetrics and Gynaecology, Centre Hospitalier Universitaire de Montpellier, Montpellier, France

⁷ Center for Free Information, Screening and Diagnosis (CeGIDD), Centre Hospitalier Universitaire de
Montpellier, Montpellier, France

⁸ Department of Infectious and Tropical Diseases, Centre Hospitalier Universitaire de Montpellier, Montpellier,
France

⁹ Swiss Tropical and Public Health Institute, Basel, Switzerland

¹⁰ CHU de Nîmes, Nîmes, France.

* Corresponding authors: thomas.beneteau@ird.fr and samuel.alizon@cnrs.fr

Abstract

Objectives. HPV infections are ubiquitous. For most infections, we lose track of the presence of the virus in host in less than three years after the start of infection. The mechanisms regulating the persistence of HPV infection are still partially understood. In this work, we focus on incident HPV detection in young women and we characterise the dynamics of these infections and evaluate the effect of genotype and host socio-economic factors on the duration of HPV detection and time between detection.

Methods. We investigated human papillomavirus (HPV) genotype detection patterns in 182 young women in Montpellier, France. We relied on SPF₁₀-LiPA25 screening assay for the simultaneous detection of 25 HPV genotypes. We used survival analysis tools with frailty effects to investigate the contribution of viral and host factors to variations in the time of HPV detectability, time of first incident detection, and time before re-detection.

Results. Women of the PAPCLEAR cohort experienced numerous positive HPV events, including frequent redetection of the same genotype. We retrieve classical results that HR-genotypes are detected for longer duration than LR-genotypes. HR-genotypes were also more likely to be detected than LR-genotypes during the follow-up. The number of lifetime sexual partner was strongly associated with increased risk of new positive detection while vaccination was related to a lower risk of displaying incident infections. Covariates related to socio-economic difficulties were associated with longer duration of HPV positivity.

Conclusions. Young women display numerous event of HPV detection, with frequent codetections of multiple genotypes at the same time and redetection of the same type after periods of no detection. These new detection are almost certainly the result of new acquisition from sexual partners, with little evidence of re-emergence of latent infections. A better characterisation of transient infections might help unveil doubts and misconception on HPV physiopathology, favouring adherence to preventive policies.

Keywords

Papillomaviruses ; Vaccination ; Epidemiology ; Public Health

28 Human Papillomaviruses (HPV) are the most oncogenic viruses known to infect humans, accounting
29 for more than 600,000 deaths worldwide each year [1]. They are also one of the most common sexually
30 transmitted infections, with estimates suggesting that by 45yo, more than 80% of the people are or
31 have already been infected by an HPV [2]. It is generally accepted that the initial HPV infection is
32 acquired during the first sexual exposures, with the prevalence peaking after sexual debut, and that the
33 risk of contracting a new HPV infection increases with the number of sexual partners [3]. HPV presence
34 generally goes undetected within the first three years, an event generally referred as HPV clearance [4].
35 This clearance, however, may not necessarily imply true immune clearance. The interpretation of re-
36 detection events is delicate as they might originate from various sources: true new infection, transient
37 deposition from a sexual partner, or detection of latent infection [5].

38 Most certainly, HPV detection is a combination of these different pathways. Deciphering the cause
39 of new HPV detection is still a major challenge. Answering this question is crucial in the optimisation
40 of future public health policies, to evaluate the effectiveness of catch-up vaccination or organise the age
41 stratification of vaccination policies.

42 Longitudinal studies are valuable data both in terms of density and length of follow-up. Extracting
43 full potential of such raw material is challenging and require the use of rigorous statistical tools. In this
44 work, we used the PAPCLEAR cohort which of samples collected every 8 weeks in 132 young women,
45 aged 18-25, for which we test the presence of 25 HPV genotypes using the SPF10-LiPA25 technique [6].
46 In particular, we evaluated viral genotypes and host factors involved with attention to frailty effects at
47 the patient level and accounting for the censoring in the data to maximize the quality of the analysis.

48 **Materials and methods**

49 **Study design and participants**

50 The PAPCLEAR cohort has been detailed elsewhere [7]. In short, this monocentric longitudinal study
51 included 189 women, who were between 18 and 25 years old at inclusion, lived in the area of Montpellier
52 (France) and reported having at least one new sexual partner over the last 12 months. Women with
53 a history of HPV-associated pathology were excluded from the study. Pregnant women or women who
54 were planning a pregnancy within the first year of inclusion were also excluded from the study. A graphic
55 summarising the inclusion protocol can be found in the Supplementary materials S4. A total of 150 women
56 were followed longitudinally for up to 2 years between 2016 and 2020. The additional 39 participants
57 were part of a cross-sectional analysis, that was prematurely suspended due to the pandemic. On-site
58 visits of infected participants took place every 8 weeks with a gynaecologist or a midwife, who performed a
59 cervical smear. Except for inclusion, participants were told to avoid sexual contact the day before the visit
60 took place to avoid unwanted transient sexual deposition from the partner. At inclusion, the participants
61 self-completed an extensive questionnaire related to demographic, socioeconomic, and behavioural risk
62 factors. For the next visits, the participants also filled in follow-up questionnaires to notify changes in
63 their habits. All participants provided written informed consent.

64 **Genotyping**

65 We first tested for the presence of alpha papillomaviruses in the cervical smears using the DEIA assay
66 [8]. DEIA-positive samples were genotyped using the LiPA25 assay, which was chosen for its sensitivity
67 and can detect up to 25 different HPV genotypes [6]. Among these, we refer to high-risk (HR) genotypes
68 for HPV16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 66, 68 [9] and to low-risk (LR) for the remaining
69 12 (HPV6, 11, 34, 40, 42, 43, 44, 53, 54, 70, 74). If the LiPA25 test was negative, the genotype was
70 determined using the PGMY PCR amplification [10] and Sanger sequencing of the PCR product. If the
71 sequencing did not yield a clear result, samples were labelled as 'non-typable'.

72 **Statistical analyses**

73 All statistical analyses were conducted using R 4.2.2 with additional packages listed in Supplementary
74 materials A6.

75 We excluded all the women with less than three visits, meaning all the cross-sectional group and
76 18 participants from the longitudinal group (1 participant quit the study, 4 were seen once, and 13
77 were seen only twice). All analyses were genotype-specific, with the unit of observation being the HPV
78 genotype. Therefore, each participant could contribute to multiple observations. Following earlier studies,
79 we assumed the dynamics of each genotype to be independent at the participant level and between
80 participants [11]. If not specified, the results were pooled across all genotypes.

81 As used in previous works [12], we defined an HPV genotype as ‘prevalent’ if detected at inclusion.
82 We also defined a genotype as ‘incident’ if detected at posterior visits but not at inclusion. Patterns of
83 positive detection separated by only one negative visit, sometimes referred to as ‘intermittent detection’
84 [12], are handled differently between studies and there does not appear to be consensus on the way to deal
85 with such data. In the main analysis, we considered the two episodes as separated but we also conducted
86 analysis with merged intermittent patterns (Supplementary materials A5).

87 Model selection was performed using the corrected Akaike Information Criterion (AICc) as a metric
88 for the penalised goodness of fit [13] We evaluated the goodness of fit for all sub-models of the maximum
89 model (i.e. with all the covariates) and estimated the hazard ratios of the Cox regression using a full
90 averaging procedure on the best models. Thorough details are available in the Supplementary materials
91 A4. To test for differences between the two populations, we used Fisher’s exact test for qualitative
92 variables and Wilcoxon-Mann-Whitney’s test for quantitative variables. We displayed the raw p -values
93 in the results Tables and Figures. In the following, a p -value < 0.05 is considered significant.

94 Survival analyses

95 For each genotype and each participant, we analysed the duration of HPV detectability, the duration
96 between HPV-positive episodes, and the time to incident HPV infection. Survival functions for these
97 quantities were computed using the non-parametric Nelson-Aalen estimator [14, 15]. For the time until the
98 first incident detection, we fitted a Weibull distribution to the survival function to predict the cumulative
99 risk of incident detection at 5 years since inclusion (see the Supplementary materials A7 for more details).

100 For a given episode, we defined the time of HPV detectability as the duration between the midpoint at
101 the start of the episode and the midpoint at the end of the episode. We included all incident episodes, even
102 the shortest episodes that were only detected during one visit, elsewhere referred to ‘transient’ infections,
103 but hereafter called ‘singletons’, and the right-censored observations. The latter corresponds to patients
104 who tested positive for HPV at their last scheduled visit. We also included prevalent episodes whose
105 start is unknown and for which the duration of HPV detectability is right-censored. Similarly, the time
106 between consecutive positive episodes was computed as the duration between the midpoint at the end of
107 the expired episodes and the midpoint at the start of the new episodes. The time until the first incident
108 infection here corresponds to the time from inclusion to the midpoint at the start of the first incident
109 detection. Extensive information can be found in Supplementary Methods A3.

110 We checked for differences in HPV detectability and time to first incident detection between HR-
111 genotypes and LR-genotypes using log-rank tests [16, 17]. To evaluate the effect of non-viral variables,
112 we used Cox proportional hazards models [18]. We stratified the Cox regression with different baseline
113 hazard functions for genotypes not detected, first detection, first redetection, and second redetection. We
114 assumed no interaction between the strata. For all Cox regression analyses, we checked the validity of the
115 proportional hazards (PH) assumption using Schoenfeld’s residuals [19]. The covariates included in the
116 analysis are the number of lifetime sexual partners (LTSP), the BMI at baseline, the self-declared ethnic
117 origin (Caucasian vs. non-caucasian or mixed-origin), the HPV vaccination status, the sexual affinity,
118 the use of condom or contraceptive pills, an indicator of financial difficulties (participant had to decline

119 medical care because of financial reasons), the number of years between inclusion and first menstruation,
120 the number of years between inclusion and first sexual intercourse and the smoking stats (past, current
121 or never). The numbers for each category can be found in Table 1.

122 We considered here that the unit of observation was the HPV genotype at the patient level. Thus as
123 one participant would experience codetection of multiple genotypes, this could induce some correlations
124 between the observations. To account for correlations between observations of the same cluster (i.e. a
125 participant), we add shared frailty effects at the patient level in the Cox regression [20]. We tested for the
126 relevance of adding the frailty at the patient level using a likelihood ratio test with one degree of freedom
127 between the two models (with and without the frailty effect).

128 Results

129 Descriptive analysis

130 The 150 participants of the PAPCLEAR cohort came for 6 visits on average (Poisson 95% CI: 5.63 - 6.43).
131 For the 132 women included in the analysis, the follow-up duration was on average 311 days (IQR: 182 -
132 431), this accounted for a total of 1543 months of follow-up. The PAPCLEAR participants included in
133 the analysis were on average aged 21.3yo (IQR: 20 - 23) at inclusion and around half of them (62/132)
134 were vaccinated against HPV (56 with Gardasil and 6 with Cervarix). Baseline characteristics for the 132
135 women included can be found in Table 1.

136 In about 74% of the women included in the analysis (98/132), we detected at least one episode during
137 their follow-up and 47% (62/132) experienced codetections (i.e. the simultaneous detections by more than
138 one genotype). Overall, we detected 342 distinct episodes, 186 incident (including 137 first detections, 44
139 redetections and 5 second redetections) and 156 prevalent, including 211 from HR types and 114 from LR
140 types. For 17 episodes, we could not determine the genotype detected. The three most frequent detected
141 types in descending order were HPV51, HPV53, and HPV66, in agreement with previous results on the
142 PAPCLEAR cohort [7]. A total of 83 (62.9%) participants were positively detected for at least one HPV

143 genotype.

144 An average of 2.6 (Poisson 95% CI: 2.32-2.88) episodes per woman were detected during the whole
145 follow-up, which yielded an average attack rate of 2.99 HPV episodes detected per person-year.

146 **Detection of first incident HPV episode**

147 Overall, we detected 137 first incident detection for all detectable genotypes and all participants. At
148 one year, we estimated a 4.80% (log-log 95% CI: 3.99 - 5.77) cumulative proportion of first incident
149 HPV detection pooled across all genotypes. After two years of follow-up, the proportion of first incident
150 infection detection increased to 7.26% (log-log 95% CI: 5.84 - 9.02). To assess the variation after 5 years,
151 we used the Weibull fit and predicted the proportion to reach 16.56% (95% CI: 7.37 - 29.73). Information
152 regarding the Weibull fit can be found in the Supplementary materials A7.

153 We assessed the rate of incident detection by oncogenic risk. We found that HR-HPV were more
154 likely to be detected than LR-HPV over the whole follow-up (log-rank p-value < 0.01). We, however,
155 lacked statistical power to verify that the difference in survival functions between HR genotypes and LR
156 genotypes was consistent for redetection or second redetection.

157 **Risk factors for the time between consecutive detection**

158 In addition to the 137 observed first incident detection (3271 right-censored), we detected 156 preva-
159 lent episodes, 44 observed redetection (213 right-censored) and 5 second redetection (36 right-censored).
160 Among the redetection, 33 consecutive episodes were only separated by one negative visit, pattern else-
161 where referred to as ‘intermittent’ detection[12]. Compared to participants reporting 1 or 2 LTSP, women
162 who reported 3-10 LTSP had increased risk of experiencing new detection(hazard ratio: 2.40 ; 95% CI:1.07
163 - 5.39), first incident or redetection. We observed a similar trend for participants reporting more than 10
164 LTSP, compared to the reference of 1 or 2 LTSP. However, we lacked statistical power to report significant
165 association (hazard ratio: 2.15 ; 95% CI:0.94 - 4.93). Merging intermittent patterns (Figure S2) yielded
166 similar results, this time with a significant association for the group reporting more than 10 LTSP. We

Table 1: **Baseline description of the PAPCLEAR cohort for participants included in the analysis (> 2 visits).** Except for the vaccine used not included in the analysis, the first level display for all categorical variables is the reference level used in the Cox analysis. Missing observations were removed from the analysis.

Covariates	States/mean [IQR]	# women	% women
Age at inclusion	21.27 [20 - 24]	132	100
Age of first menstruation	12.85 [12 - 14]	131	99.24
Years between 1st menstruation and inclusion	8.42 [7 - 10]	131	99.24
Age of first sexual intercourse	16.39 [15 - 17]	132	100
Years between 1st intercourse and inclusion	4.89 [3 - 7]	132	100
Vaccination	Unvaccinated	66	50.00
	Vaccinated	66	50.00
Vaccine used	Cervarix	6	9.09
	Gardasil	56	84.18
	Non-specified	4	6.06
Self-declared ethnicity	Mixed or other	27	20.45
	Caucasian	105	79.55
Self-declared sexual affinity	Bi-/Homosexual	12	9.09
	Heterosexual	120	90.90
Smoking	Never	61	46.21
	Past	19	14.39
	Current	52	39.39
Contraceptive pills †	Not using	64	48.48
	User	68	51.51
Male condoms	Not using	56	42.42
	User	76	57.57
Number of lifetime sexual partner	1;2	20	15.15
	3;10	63	47.72
	11+	48	36.36
	missing	1	0.76
Financial difficulties *	No experience	117	88.64
	Experienced in the last 12mo	15	11.36

† emergency pills not included

* defined as a participant who declined medical care because of financial reasons

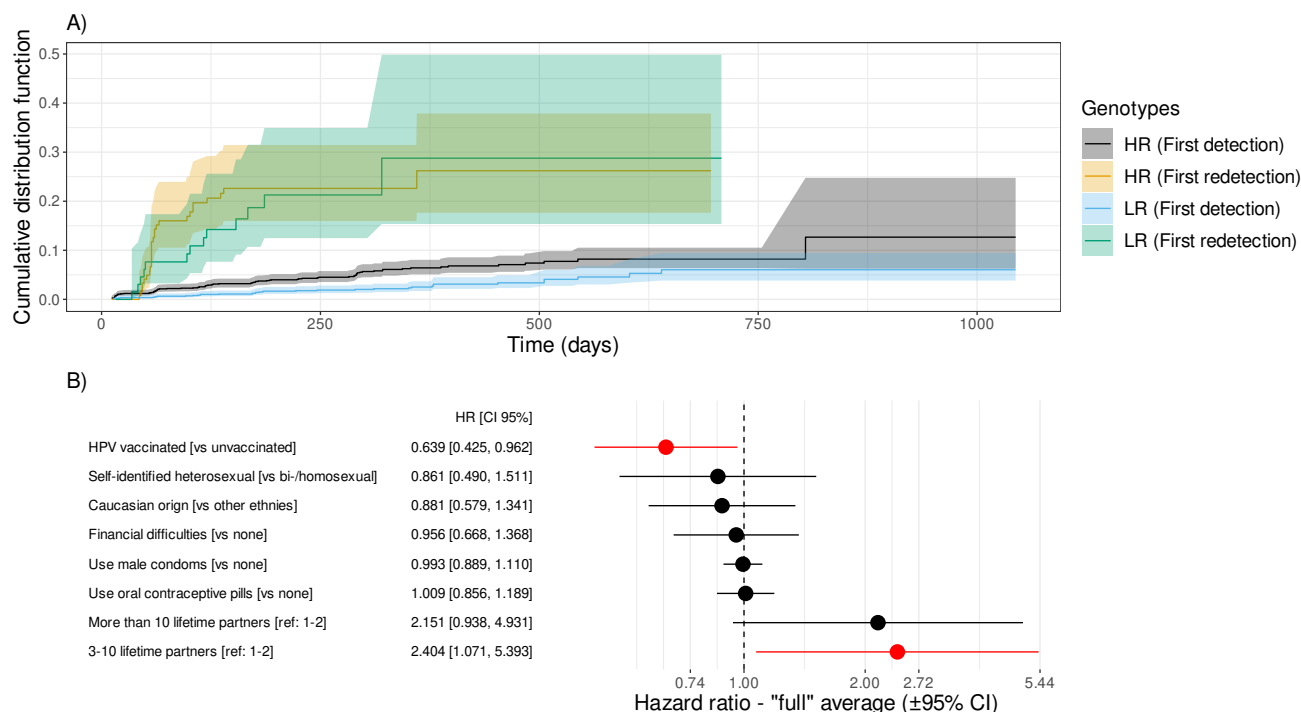


Figure 1: **Cumulative distribution functions for the time to first incident detection and time to first genotype redetection stratified by HR/LR genotypes and effects of host covariates on these estimates.** A) Cumulative distribution function (CDF) of the time to first incident HPV detection since inclusion and the time to the first redetection, stratified by HPV genotype status (HR and LR). B) Hazard ratio for the best models selected by Cox regression with frailty at the patient level. Significant covariates are in red and hazard ratios greater than 1 indicate the covariate is associated with an increased risk of detection, hence lower duration between consecutive episodes.

167 also found that vaccinated participants were less likely to display new incident detection or redetection
 168 compared to unvaccinated participants (hazard ratio: 0.64 ; 95% CI: 0.43-0.96). Thorough results of the
 169 Cox analysis for the time between consecutive episodes are displayed in Figure 1.

170 Loss of HPV detection

171 On the total of 342 detected episodes, 156 were prevalent HPV detection and for 40 episodes we did not
 172 observe the loss of detectability. For 17 episodes, the participants entered positive to a genotype and left
 173 the follow-up still positive for that same genotype, without any negative visit in-between. The majority
 174 of the episodes were positive for only one visit (198/342 ; 57.9%), but a significant proportion of them are
 175 censored observations (86/198 ; 43.4%) and thus potentially just a glimpse of a longer event. We estimated
 176 a median period of HPV detectability of 113 days (log-log 95% CI: 92.5 - 124). Our results suggest that

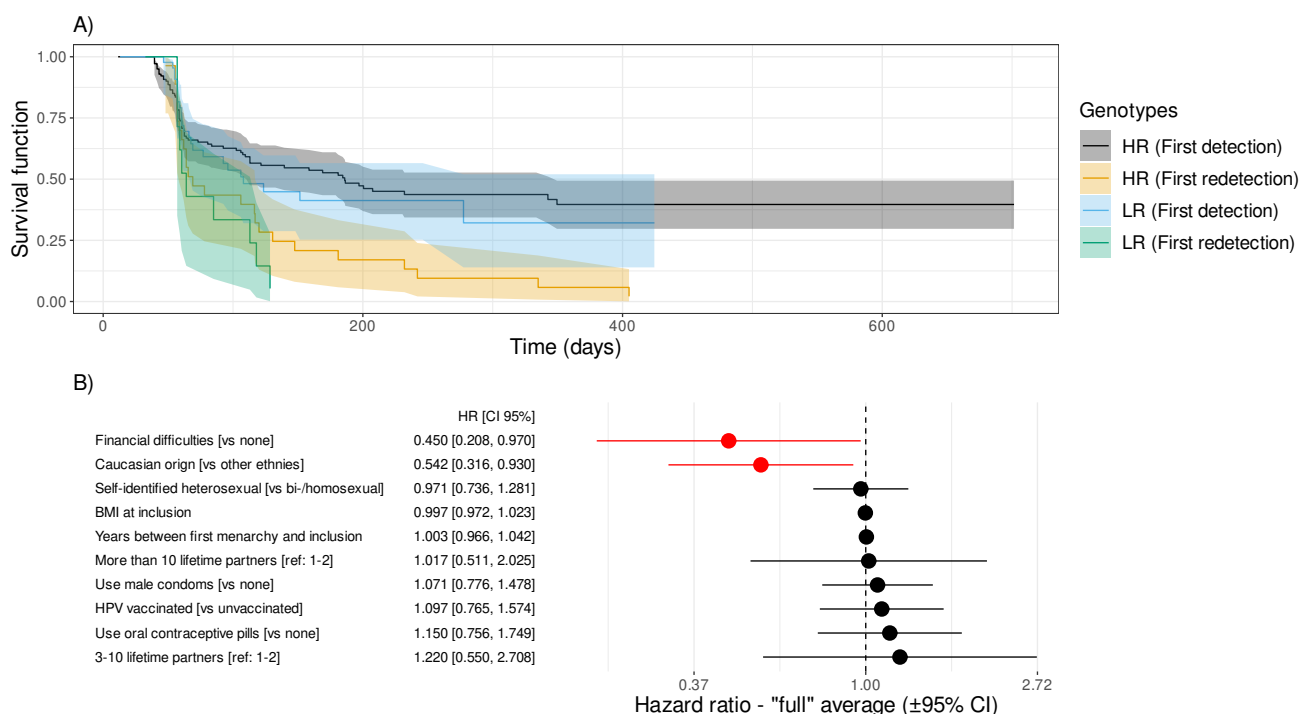


Figure 2: **Survival function for the time of HPV detectability stratified by HR/LR genotypes and effects of host covariates on this estimate.** A) Survival functions stratified by the genotype (HR/LR) for the time to loss of HPV detection. B) Hazard ratios for the host factors. Significant factors are in red and a hazard ratio lower than one indicates that the trait is associated with a decreased rate of loss of HPV DNA detection, hence longer survival functions. The reference level is indicated in the bracket for the qualitative variables (see Methods for details).

177 around 23.5% (log-log 95% CI: 16.5 - 31.4) of the HPV episodes were still detectable after 700 day of
 178 follow-up. We found that HR-HPV types were detected significantly longer than LR-HPV infections
 179 (log-rank p-value < 0.05), the survival functions are displayed in Figure 2. The median duration of
 180 detectability was 130 days (log -log 95% CI 106 - 186) for HR genotypes and 96 days (log-log 95% CI 67.5
 181 - 113) for LR genotypes.

182 Finally, we investigated the effect of key host factors (listed in Table 1) on the duration of HPV
 183 detectability using LiPA assays. Our analysis showed that HPV was detected for a significantly longer
 184 duration in participants who experienced financial difficulties (defined as a participant who declined
 185 medical care because of financial reasons) in the last 12mo before inclusion compared to participants who
 186 did not experience it (Hazard ratio: 0.45 ; 95% CI: 0.21 - 0.97). Infections were also detected longer in
 187 participants who identified themselves solely as 'Caucasians' (Hazard ratio: 0.54 ; 95% CI: 0.32 - 0.96)

188 compared to participants who identified themselves as non-Caucasian or mixed-origin (Figure 2). These
189 results were similar when merging intermittent patterns (Figure S3).

190 Discussion

191 In this work, we analysed the HPV detection patterns of 132 young women from the PAPCLEAR cohort.
192 We estimated a cumulative probability of first incident HPV detection since baseline, pooled across all
193 genotypes, of 4.80% (log-log 95% CI: 3.99 - 5.77) at one year and predicted it to reach 16.56% (95% CI:
194 7.37 - 29.73) at five year. While we used convenient denominations such as ‘first incident detection’ or
195 ‘first redetection’ following previous works, these expressions surely not bare the same relevance in terms
196 of natural history. It is very unlikely for most of the investigated participants that a first detection during
197 the follow-up corresponds to a true first exposure, only 11.36% participants declared having had their first
198 sexual partner in the last 12 months. Besides, these definitions are dependent of the sampling frequency,
199 as the probability of documenting a short period between redetections increases with the sampling rate.
200 Thus, it is not straightforward to compare our results with other studies. Our results for first detection
201 are very similar to results for first redetection but quite discordant with estimates for first incident detection
202 from other studies [12]. Those differences might first come from the difference in the cohort design, as
203 the sampling rate for the Ludwig-McGill cohort was about every 6 months. Besides, the two populations
204 were quite different. Participants here are all between 18-25 years old at inclusion. In the latter, there
205 is a wider age diversity among the women, with around 80% of the women being older than 25 years
206 old. Sexual activity is negatively correlated with age after sexual debut, thus our population might be
207 more exposed to HPV due to increased sexual activity [21]. Additionally, the Ludwig-McGill cohort was
208 sampled from low-income families from Brazil, while we included women without income criteria.

209 We found that the participants experienced numerous detected episodes (2.99 per women-year), most
210 of them being positively detected for only one visit (198/374 singletons, 86/198 censored). In a quarter
211 of the participants included in the analysis (34/132), no alphapapillomaviruses was detected, while for 18
212 women we detected more than 5 episodes during the follow-up. Redetection, were not uncommon as in
213 about a third of women displaying at least one HPV positive visit we detected redetections of the same
214 genotype. The frequency of codetections is similarly high (47%) and is consistent with previous studies

215 [22].

216 To date, except for some specific populations (e.g. abstinent women [23]), it is quite difficult to settle
217 on the origin of new HPV detection in humans. In our cohort, the number of lifetime partners was
218 negatively associated with the duration between episodes. We found that in participants reporting a
219 number of lifetime partners higher than 3, the time between episodes was significantly shorter than for
220 women with 1 or 2 reported LTSP at inclusion, with little difference between participants reporting 3 to 10
221 LTSP and those reporting more than 11 LTSP. The number of LTSP was not associated with the number
222 of years between first intercourse and inclusion. Besides, women reporting a higher number of LTSP
223 at inclusion were also more likely to report intercourse with new partner during the follow-up. Taken
224 together, our results suggest that new detection and redetection observed here are more likely to be new
225 acquisition than re-emergence of latent infection. Additionally, we recall that only women who declared
226 having sex with a new partner in the last 12 months were included in the cohort [7]. It was noted elsewhere
227 that in the setting of new sexual partners, true incident infection was the preferential explanation [24, 25].
228 Compared to unvaccinated participants were less at risk of displaying incident infections. These results
229 are consistent with observations in a dozen country [26].

230 We found that HR-HPV types were more likely to be detected for a longer period than LR-HPV types,
231 which corroborates earlier studies [27]. Conversely, we found that the time between HPV positive events
232 was shorter for HR-HPV types compared to LR-HPV types, consistent with other work [22].

233 The mean duration of HPV detectability was globally lower than previous studies [12, 22, 23, 27]. This
234 can be partially explained by the difference in sampling rates with compared to studies (8 weeks compared
235 to 6 months) and the inclusion of all positive episodes, including the singletons, sometimes excluded from
236 analysis [23]. Participants that experienced financial difficulties (defined as a participant who declined
237 medical care because of financial reasons) in past 12 months prior to inclusion displayed longer periods
238 of HPV detectability compared to participants that did not declared suffering from financial struggles.
239 People in situations of poverty generally tend to live in areas with low medical coverage, thus also limiting

240 their access to medical care [28]. While there was no significant difference in vaccination uptake between
241 participants who faced financial difficulties and those who did not, taken together, our results suggest that
242 people with financial difficulties might be less prone to seek medical guidance, especially since specialists
243 are not fully reimbursed in France, thus putting them more at risk of genital infections and complications.
244 Our results also suggest that self-declared Caucasian participants experienced longer periods of HPV
245 detection. However, we lacked information to assess if that trend originated from genetic origin or socio-
246 demographic/behavioural differences between the two groups. Besides, our population is relatively limited
247 in number (132 women, 27 mixed origin or non-Caucasian origin, 105 Caucasian origin) and restricted to
248 a specific region in France. It does not reflect the French population diversity, and thus might just be the
249 results of selection bias.

250 Clarifying the dynamic of HPV infection, especially regarding the distinction between re-detection
251 and new acquisition is decisive to inform public health policies. Efficient screening policies and prevention
252 have been implemented to limit progression towards cancer with good compliance to these measures
253 [29]. While most HPV infections are generally benign, testing positive during HPV screening can cause
254 psychological stress and anxiety [30], especially if self-sampling becomes widespread [31].

255 We believe a better characterisation of HPV infections, especially regarding the link between infection
256 status and detection data, will help unveil doubts and misconception on HPV physiopathology, favouring
257 adherence to preventive measures [32].

258 **Acknowledgements**

259 The authors thank all the participants of the PAPCLEAR study and the clinical staff and nurses for their
260 help.

261 **Financial support**

262 TB is funded by la *Ligue contre le Cancer* (grant No TAKX21133). This work was supported by the
263 European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation
264 programme (grant agreement No 648963 to SA). The sponsor had no role in the study design; in the
265 collection, analysis, and interpretation of data; in the writing of the report; and in the decision to submit
266 the article for publication.

267 **Declaration of Competing Interest**

268 J. R. reports personal fees from Gilead (consulting and payment or honoraria for lectures, presentations,
269 speaker’s bureaus, manuscript writing, or educational events), Janssen (payment or honoraria for lec-
270 tures, presentations, speaker’s bureaus, manuscript writing, or educational events), Merck (payment or
271 honoraria for lectures, presentations, speaker’s bureaus, manuscript writing, or educational events), Ther-
272 atechnologies (payment or honoraria for lectures, presentations, speaker’s bureaus, manuscript writing,
273 or educational events), and ViiV Healthcare (consulting and payment or honoraria for lectures, presenta-
274 tions, speaker’s bureaus, manuscript writing, or educational events) and support for attending meetings
275 and/or travel from Gilead and Pfizer, outside of the submitted work. All the other authors do not report
276 any conflict of interest or personal relationships that could have appeared to influence the work reported
277 in this paper.

278 **References**

- 279 [1] Plummer M, Vaccarella S, Franceschi S. Multiple human papillomavirus infections: the exception or
280 the rule? *J Infect Dis.* 2011;203(7):891–3.
- 281 [2] Chesson HW, Dunne EF, Hariri S, Markowitz LE. The estimated lifetime probability of acquiring
282 human papillomavirus in the United States. *Sexually Transmitted Diseases.* 2014;41(11):660–664.

- 283 [3] Bruni L, Diaz M, Castellsagué M, Ferrer E, Bosch FX, de Sanjosé S. Cervical Human Papillomavirus
284 Prevalence in 5 Continents: Meta-Analysis of 1 Million Women with Normal Cytological Findings.
285 The Journal of Infectious Diseases. 2010 Dec;202(12):1789–1799. Available from: [https://doi.org/
286 10.1086/657321](https://doi.org/10.1086/657321).
- 287 [4] Schiffman M, Wentzensen N, Wacholder S, Kinney W, Gage JC, Castle PE. Human Papillomavirus
288 Testing in the Prevention of Cervical Cancer. J Natl Cancer Inst. 2011 Mar;103(5):368–383. Publisher:
289 Oxford Academic. Available from: <https://academic.oup.com/jnci/article/103/5/368/905734>.
- 290 [5] Gravitt P, Winer R. Natural History of HPV Infection across the Lifespan: Role of Viral Latency.
291 Viruses. 2017 Sep;9(10):267. Available from: <http://www.mdpi.com/1999-4915/9/10/267>.
- 292 [6] Geraets DT, Struijk L, Kleter B, Moliijn A, van Doorn LJ, Quint WGV, et al. The original SPF10
293 LiPA25 algorithm is more sensitive and suitable for epidemiologic HPV research than the SPF10
294 INNO-LiPA Extra. J Virol Meth. 2015;215-216:22–29.
- 295 [7] Murall CL, Rahmoun M, Selinger C, Baldellou M, Bernat C, Bonneau M, et al. Natural history,
296 dynamics, and ecology of human papillomaviruses in genital infections of young women: protocol of
297 the PAPCLEAR cohort study. BMJ Open. 2019 Jun;9(6):e025129. Available from: [http://bmjopen.
298 bmj.com/content/9/6/e025129.abstract](http://bmjopen.bmj.com/content/9/6/e025129.abstract).
- 299 [8] Kleter B, van Doorn LJ, ter Schegget J, Schrauwen L, van Krimpen K, Burger M, et al. Novel
300 Short-Fragment PCR Assay for Highly Sensitive Broad-Spectrum Detection of Anogenital Human
301 Papillomaviruses. Am J Pathol. 1998;153(6):1731–1739.
- 302 [9] IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. Biological Agents.
303 vol. 100B of IARC monographs on the evaluation of carcinogenic risks to humans. Lyon, France:
304 International Agency for Research on Cancer; 2012.
- 305 [10] Coutlée F, Gravitt P, Kornegay J, Hankins C, Richardson H, Lapointe N, et al. Use of PGMY

- 306 Primers in L1 Consensus PCR Improves Detection of Human Papillomavirus DNA in Genital Sam-
307 ples. *Journal of Clinical Microbiology*. 2002 Mar;40(3):902–907.
- 308 [11] Dillner J, Arbyn M, Unger E, Dillner L. Monitoring of human papillomavirus vaccination. *Clin Exp*
309 *Immunol*. 2011;163(1):17–25.
- 310 [12] Malagón T, Trottier H, El-Zein M, Villa L, Franco E. Human papillomavirus intermittence and
311 risk factors associated with first detections and redetections in the Ludwig-McGill cohort study of
312 adult women. *The Journal of infectious diseases*. 2023 Feb;Publisher: J Infect Dis. Available from:
313 <https://pubmed.ncbi.nlm.nih.gov/36790831/>.
- 314 [13] Hurvich CM, Tsai CL. A Corrected Akaike Information Criterion for Vector Autoregressive Model
315 Selection. *Journal of Time Series Analysis*. 1993;14(3):271–279.
- 316 [14] Nelson W. Theory and Applications of Hazard Plotting for Censored Failure Data. *Technometrics*.
317 1972;14(4):945–966.
- 318 [15] Aalen O. Nonparametric Inference for a Family of Counting Processes. *The Annals of Statistics*.
319 1978;6(4):701–726.
- 320 [16] Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration.
321 *Cancer Chemotherapy Reports*. 1966 Mar;50(3):163–170.
- 322 [17] Peto R, Peto J. Asymptotically Efficient Rank Invariant Test Procedures. *Journal of the Royal*
323 *Statistical Society Series A (General)*. 1972;135(2):185–207.
- 324 [18] Cox DR. Regression Models and Life-Tables. In: Kotz S, Johnson NL, editors. Breakthroughs in
325 *Statistics: Methodology and Distribution*. Springer Series in Statistics. New York, NY: Springer;
326 1992. p. 527–541.
- 327 [19] Schoenfeld D. Partial Residuals for The Proportional Hazards Regression Model. *Biometrika*.
328 1982;69(1):239.

- 329 [20] Wienke A. Frailty Models in Survival Analysis. New York: Chapman and Hall/CRC; 2010.
- 330 [21] Herbenick D, Reece M, Schick V, Sanders SA, Dodge B, Fortenberry JD. Sexual behavior in the
331 United States: results from a national probability sample of men and women ages 14-94. The Journal
332 of Sexual Medicine. 2010 Oct;7 Suppl 5:255–265.
- 333 [22] Ramanakumar AV, Naud P, Roteli-Martins CM, de Carvalho NS, de Borba PC, Teixeira JC, et al. In-
334 cidence and duration of type-specific human papillomavirus infection in high-risk HPV-naïve women:
335 results from the control arm of a phase II HPV-16/18 vaccine trial. BMJ Open. 2016;6(8):e011371.
- 336 [23] Paul P, Hammer A, Rositch AF, Burke AE, Viscidi RP, Silver MI, et al. Rates of New Human
337 Papillomavirus Detection and Loss of Detection in Middle-aged Women by Recent and Past Sexual
338 Behavior. The Journal of Infectious Diseases. 2021 Apr;223(8):1423–1432. Available from: <https://doi.org/10.1093/infdis/jiaa557>.
- 339 <https://doi.org/10.1093/infdis/jiaa557>.
- 340 [24] Trottier H, Ferreira S, Thomann P, Costa MC, Sobrinho JS, Prado JCM, et al. Human papillomavirus
341 infection and reinfection in adult women: the role of sexual activity and natural immunity. Cancer
342 Research. 2010 Nov;70(21):8569–8577.
- 343 [25] Moscicki AB, Ma Y, Farhat S, Darragh TM, Pawlita M, Galloway DA, et al. Redetection of cervical
344 human papillomavirus type 16 (HPV16) in women with a history of HPV16. The Journal of Infectious
345 Diseases. 2013 Aug;208(3):403–412.
- 346 [26] Brotherton JML. Impact of HPV vaccination: Achievements and future challenges. Papillomavirus
347 Research (Amsterdam, Netherlands). 2019 Jun;7:138–140.
- 348 [27] Goodman MT, Shvetsov YB, McDuffie K, Wilkens LR, Zhu X, Thompson PJ, et al. Prevalence,
349 Acquisition, and Clearance of Cervical Human Papillomavirus Infection among Women with Normal
350 Cytology: Hawaii Human Papillomavirus Cohort Study. Cancer Research. 2008;68(21):8813–8824.
- 351 [28] Vallée J, Cadot E, Grillo F, Parizot I, Chauvin P. The combined effects of activity space and

- 352 neighbourhood of residence on participation in preventive health-care activities: The case of cervical
353 screening in the Paris metropolitan area (France). *Health & Place*. 2010 Sep;16(5):838–852. Available
354 from: <https://www.sciencedirect.com/science/article/pii/S1353829210000456>.
- 355 [29] Jansen EEL, Zielonke N, Gini A, Anttila A, Segnan N, Vokó Z, et al. Effect of organised cervical
356 cancer screening on cervical cancer mortality in Europe: a systematic review. *European Journal of*
357 *Cancer*. 2020 Mar;127:207–223. Available from: [https://www.sciencedirect.com/science/article/pii/](https://www.sciencedirect.com/science/article/pii/S0959804919308688)
358 [S0959804919308688](https://www.sciencedirect.com/science/article/pii/S0959804919308688).
- 359 [30] McBride E, Tatar O, Rosberger Z, Rockliffe L, Marlow LAV, Moss-Morris R, et al. Emotional response
360 to testing positive for human papillomavirus at cervical cancer screening: a mixed method systematic
361 review with meta-analysis. *Health Psychology Review*. 2021 Jul;15(3):395–429. Publisher: Routledge
362 _eprint: <https://doi.org/10.1080/17437199.2020.1762106>. Available from: [https://doi.org/10.1080/](https://doi.org/10.1080/17437199.2020.1762106)
363 [17437199.2020.1762106](https://doi.org/10.1080/17437199.2020.1762106).
- 364 [31] Nishimura H, Yeh PT, Oguntade H, Kennedy CE, Narasimhan M. HPV self-sampling for cervi-
365 cal cancer screening: a systematic review of values and preferences. *BMJ Global Health*. 2021
366 May;6(5):e003743. Publisher: BMJ Specialist Journals Section: Original research. Available from:
367 <https://gh.bmj.com/content/6/5/e003743>.
- 368 [32] Thompson EL, Vamos CA, Sappenfield WM, Straub DM, Daley EM. Relationship status im-
369 pacts primary reasons for interest in the HPV vaccine among young adult women. *Vaccine*.
370 2016 Jun;34(27):3119–3124. Available from: [https://www.sciencedirect.com/science/article/pii/](https://www.sciencedirect.com/science/article/pii/S0264410X16302286)
371 [S0264410X16302286](https://www.sciencedirect.com/science/article/pii/S0264410X16302286).
- 372 [33] Bogaerts K, Komárek A, Lesaffre E. *Survival Analysis with Interval-Censored Data*. Chapman and
373 Hall/CRC; 2017.
- 374 [34] Law CG, Brookmeyer R. Effects of mid-point imputation on the analysis of doubly censored data.
375 *Statistics in Medicine*. 1992;11(12).

- 376 [35] Nelson W. Hazard Plotting for Incomplete Failure Data. *Journal of Quality Technology*. 1969;1(1):27–
377 52.
- 378 [36] Burnham KP, Anderson DR. Multimodel Inference: Understanding AIC and BIC in Model Selection.
379 *Sociological Methods & Research*. 2004;33(2):261–304.
- 380 [37] Symonds MRE, Moussalli A. A brief guide to model selection, multimodel inference and model
381 averaging in behavioural ecology using Akaike’s information criterion. *Behavioral Ecology and So-*
382 *ciobiology*. 2011;65(1):13–21.
- 383 [38] Therneau TM. A Package for Survival Analysis in R; 2022. R package version 3.4-0. Available from:
384 <https://CRAN.R-project.org/package=survival>.
- 385 [39] Bartoń K. MuMIn: Multi-Model Inference; 2022. R package version 1.46.0. Available from: [https:](https://CRAN.R-project.org/package=MuMIn)
386 [//CRAN.R-project.org/package=MuMIn](https://CRAN.R-project.org/package=MuMIn).
- 387 [40] Therneau TM. coxme: Mixed Effects Cox Models; 2022. R package version 2.2-17. Available from:
388 <https://CRAN.R-project.org/package=coxme>.
- 389 [41] Dilley SE, Peral S, Straughn JM, Scarinci IC. The challenge of HPV vaccination uptake and oppor-
390 tunities for solutions: Lessons learned from Alabama. *Preventive Medicine*. 2018;113:124–131.

391 Supplementary Materials

392 A1 Ethics

393 The PAPCLEAR trial was promoted by the Centre Hospitalier Universitaire de Montpellier and ap-
 394 proved by the *Comité de Protection des Personnes (CPP) Sud Méditerranée I* on 11 May 2016 (CPP
 395 number 16 42, reference number ID RCB 2016A00712-49); by the *Comité Consultatif sur le Traite-
 396 ment de l'Information en matière de Recherche dans le domaine de la Santé* on 12 July 2016 (reference
 397 number 16.504); and by the *Commission Nationale Informatique et Libertés* on 16 December 2016 (ref-
 398 erence number MMS/ABD/AR1612278, decision number DR-2016488). This trial was authorised by
 399 the *Agence Nationale de Sécurité du Médicament et des Produits de Santé* on 20 July 2016 (reference
 400 20160072000007). The ClinicalTrials.gov identifier is NCT02946346. All participants provided written
 401 informed consent.

402 A2 Protocol of PAPCLEAR study

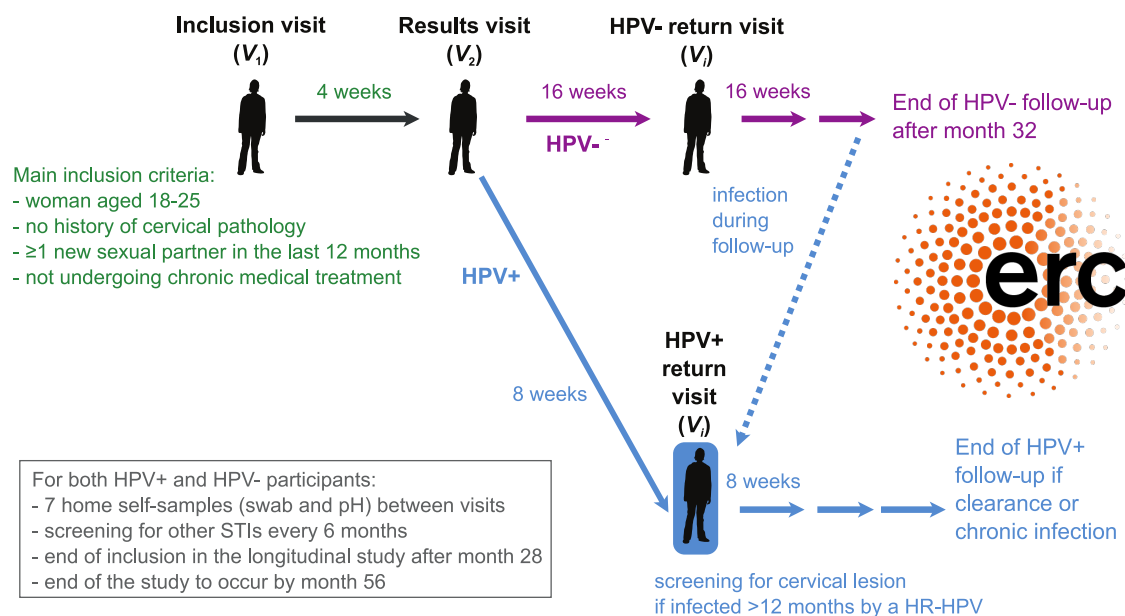


Figure S1: General structure of the PAPCLEAR study. [7]

403 **A3 Defining events and durations**

404 Results from the DEIA and LiPA25 assays yielded dated binary vectors. For each infectious event, we only
405 know the intervals during which the infection started and ended, which means the data is ‘doubly interval
406 censored’ and usually cumbersome to analyse [33]. To simplify the problem, we computed duration using
407 the conventional midpoint methodology. For this, we defined the start of an infection as the midpoint
408 between the last negative test before and the first positive test of the infection. Likewise, we defined the
409 end of an infection as the midpoint between the last positive test of the infection and the first negative
410 test after the infection). For incomplete data, we assume the start to be at inclusion for left-censored
411 observation and we assumed the end to be at last visit for right-censored observation. The bias associated
412 with this simpler method is expected to be limited since our sampling scheme is regular and short-spaced
413 [34].

414 To study the time to HPV infection clearance, we defined as an ‘event’ or ‘episode’ a series of at least
415 one positive LiPA25 detection for a given HPV type and a given participant. During the follow-up, we
416 often detected several events per participant (sometimes even by the same genotype). We assumed that
417 two consecutive episodes were independent even if only separated by one negative visit. Such patterns,
418 also called intermittent [12], are sometimes merged to form a longer episode instead of two separate
419 entities [23]. We evaluated the changes in the estimates using this methodology below XXXX.

420 To estimate the time of HPV detectability, i.e. the time of positive HPV detection, we computed the
421 duration between the midpoint at the start of an infection and the midpoint at the end of the infection. If
422 one or both of the endpoints were censored, we assumed the duration to be right-censored. We assumed
423 the events to be independent [11] and, therefore, defined the time between episodes to be independent
424 events. For the time to first incident infection, we excluded prevalent infection and computed the time
425 from inclusion to the midpoint at start of first incident detection for a genotype and a participant. If the
426 genotype is not detected during follow-up, we used a right-censored observation whose duration equals
427 the time of follow-up of the participant. When analysing the time between positive episodes, we included

428 all events and computed the time as the duration between the midpoint at start of expired episodes and
429 the midpoint at start of the new episodes. There is in general, a lower number of data of redetection than
430 expected because some participant were still positive for a genotype at end of follow-up, thus preventing
431 us from computing a time of redetection. In both cases, the cumulative distribution functions (CDF) or
432 survival functions were computed using the Nelson-Aalen estimator of the cumulative hazard rate function
433 [14, 15, 35].

434 **A4 Model comparison**

435 We compared the models using the corrected Akaike Information Criterion (AICc) as a metric for the
436 penalised goodness of fit [13]. Briefly, we first generated the maximum model with all the variables chosen
437 for the Cox regression and then performed the model selection by subsetting all possible combinations
438 from the maximum model and evaluating their respective AICc. We kept the models with an AICc smaller
439 than the minimum AICc+2, following standard practice [36]. We then averaged the coefficients of the
440 remaining models using a full averaging procedure to avoid artificial departure from 0. This was necessary
441 because we averaged on all the selected models, not just on the ones with the variable whose coefficient
442 was computed [37]. Finally, we computed the hazard ratio by taking the exponential of these averaged
443 coefficients.

444 **A5 Merging intermittent patterns**

445 Following previous notations, intermittent patterns corresponds to successive positive HPV detection
446 episodes separated only by one negative visit. Merging intermittent patterns modifies the data used for
447 analysis, diminishing the number of events, making them last longer in average. In total, we detect 33
448 intermittent patterns. Merging the patterns decreased the number of positive detectable events by the
449 same amount. However, it did not change much the results of the Cox regression. While the degree of
450 significance varies between the two datasets, the same trends were observed between the two cases.

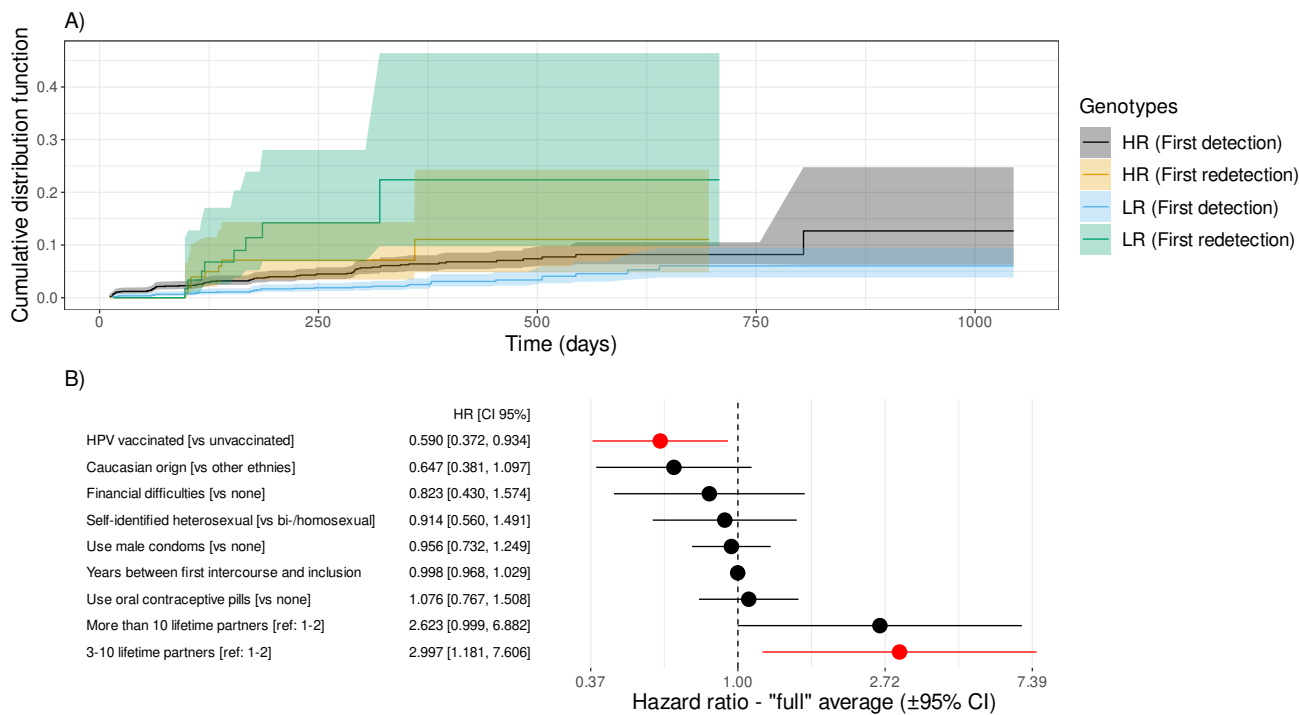


Figure S2: Cumulative distribution functions for the time to first incident detection and time to first genotype redetection stratified by HR/LR genotypes and effects of host covariates on these estimates for merged intermittent patterns. A) Cumulative distribution function (CDF) of the time to first incident HPV detection since inclusion and the time to the first redetection, stratified by HPV genotype status (HR and LR). B) Hazard ratio for the best models selected by Cox regression with frailty at the patient level. Significant covariates are in red and hazard ratios greater than 1 indicate the covariate is associated with a increased risk of detection, hence lower duration between consecutive episodes.

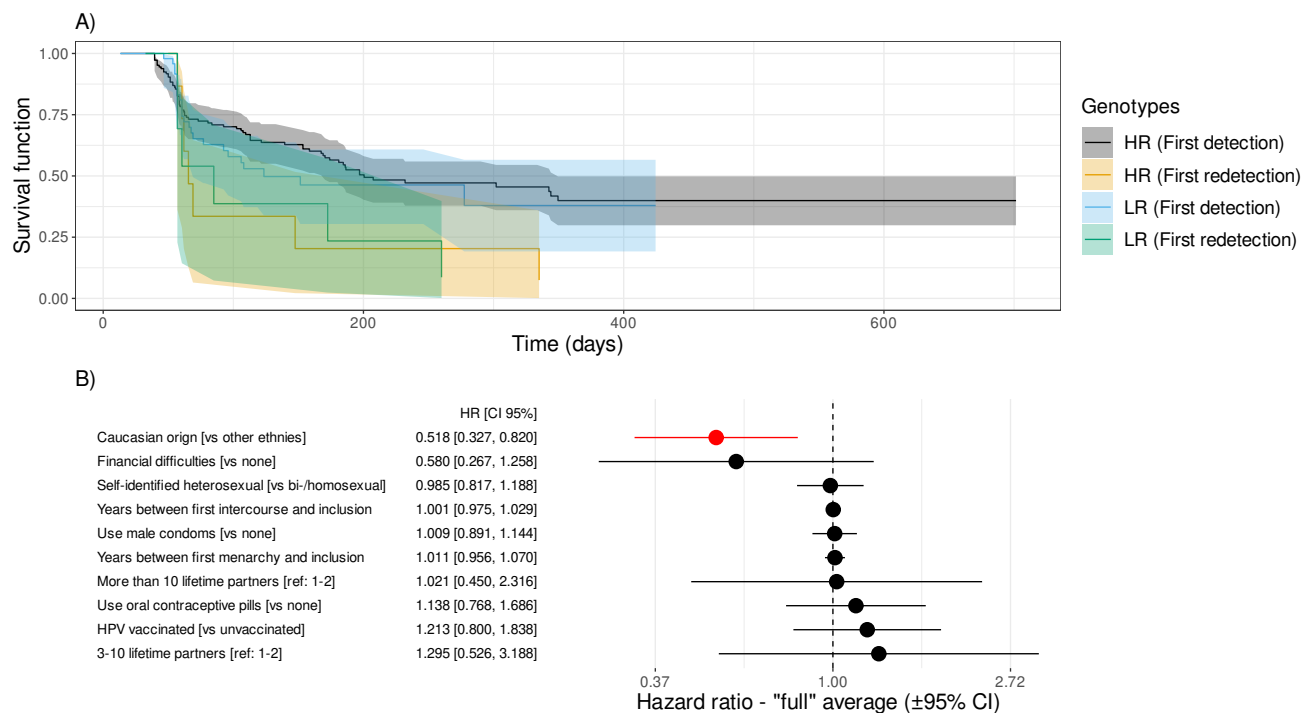


Figure S3: Survival function for the time of HPV detectability stratified by HR/LR genotypes and effects of host covariates on this estimate for merged intermittent patterns. A) Survival functions stratified by the genotype (HR/LR) for the time to loss of HPV detection. B) Hazard ratios for the host factors. Significant factors are in red and a hazard ratio lower than one indicates that the trait is associated with an decreased rate of loss of HPV DNA detection, hence longer survival functions. The reference level is indicated in the bracket for the qualitative variables (see Methods for details).

451 **A6 R packages**

- 452 – SURVIVAL: non-parametric and parametric estimators of the survival function and Cox regression
- 453 [38] ; version 3.5-3.
- 454 – MUMIN: model selection and model averaging [39] ; version 1.46.0.
- 455 – COXME: adding frailty effects to the hazard function as a centred Gaussian distribution [40] ; version
- 456 2.2-17.

457 **A7 Graphical Weibull fit**

Let $\lambda > 0$ be the scale parameter and $k > 0$ be the shape parameter, for all $t \in \mathcal{R}^+$, we can define the survival function of the Weibull distribution as:

$$S(t) = e^{-\left(\frac{t}{\lambda}\right)^k} \Leftrightarrow \log(-\log(S(t))) = k(\log(t) - \log(\lambda)) \quad (\text{S1})$$

458 Thus using the $\log(-\log(\cdot))$ transformation of the survival function, estimated using non-parametric
459 estimators like Nelson-Aalen or Kaplan-Meier, and plotting it versus the natural logarithm of the event
460 times, we can assess if a Weibull distribution is an appropriate model to describe the data by evaluating
461 the goodness of the fit as a linear model [41]. Clearly, for the time to first incident detection pooled across
462 all genotypes, and for both HR genotypes and LR genotypes grouping, the Weibull was relevant (panel
463 A in Figure S4). However, for the time to loss of HPV detection, we see a clear non-linear trend between
464 the $\log(-\log(\cdot))$ transformation of the survival function and the $\log(\text{time})$ (panel B in Figure S4), thus
465 discouraging us for trying to fit a Weibull distributions to this data. The parameter estimates of the
466 Weibull distribution for the time to first incident detection are displayed in Table S1.

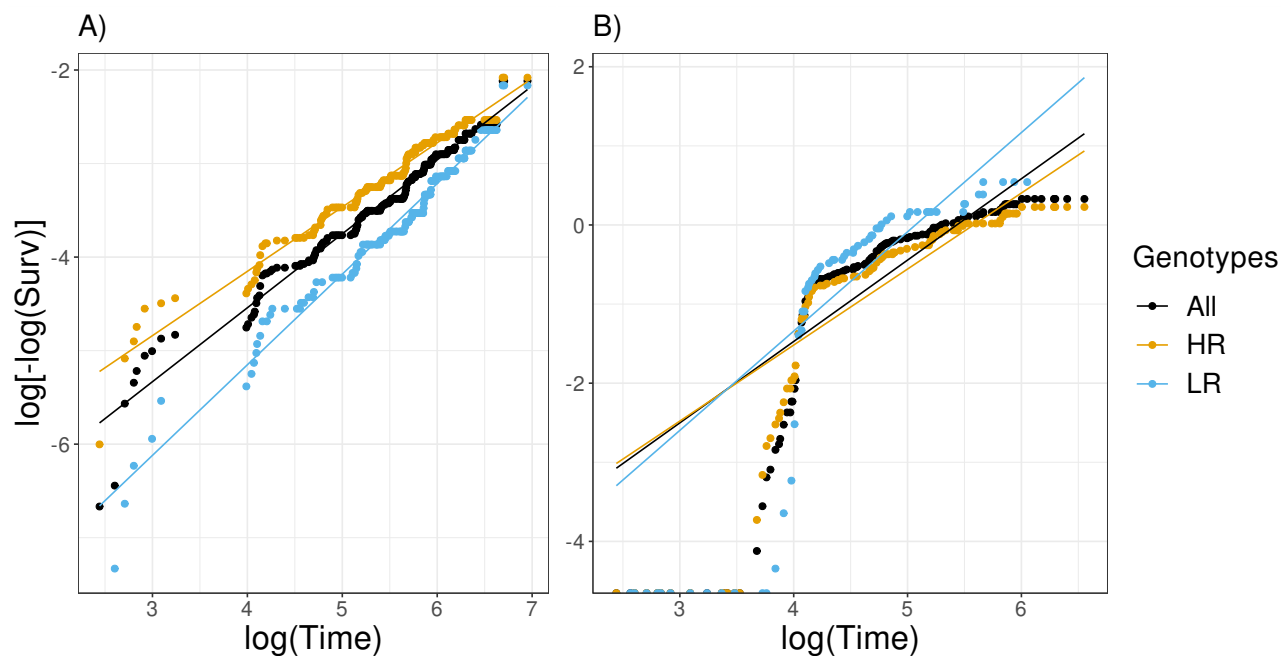


Figure S4: **Graphical assessment for the goodness of Weibull fit.** On panel A) we displayed the $\log(-\log(\cdot))$ transformation of the survival function for the time to first incident detection versus the $\log(\text{Time})$ and on panel B) the same transformation for the survival functions of the time to loss of HPV detection. For panel A) the linear fit is acceptable while for panel B) there is a clear non-linear trend.

Table S1: **Estimates for the Weibull parameters (shape and scale) for the time to first incident detection.**

	scale (λ)	shape (k)
All genotypes	$1.43 [0.79; 2.58] \times 10^4$	0.832 [0.713; 0.970]
HR genotypes	$1.12 [0.63; 2.01] \times 10^4$	0.832 [0.714; 0.971]
LR genotypes	$1.93 [1.27; 2.94] \times 10^4$	0.832 [0.714; 0.971]