Supplemental Digital Content of

Estimating the Human Papillomavirus Genotype Attribution in Screen-detected High-grade Cervical Lesions

by Birgit I. Lissenberg-Witte, Johannes A. Bogaards, Wim Quint & Johannes Berkhof

The reader is referred to the main article for all terms and symbols not explicitly defined in this supplemental material.

eAppendix 1. Derivation and maximization of the log-likelihood

For each woman, the observed data is the vector (D, S). Conditional on $S = s$, the probability that $D = 0$ is equal to

$$\mathrm{P}(D = 0 \mid S = s) = \prod_{k\in s}(1-\pi_k),$$

and the probability that $D = 1$ is equal to

$$\mathrm{P}(D = 1 \mid S = s) = 1 - \prod_{k\in s}(1-\pi_k).$$

Therefore, the density of (D, S) under $\boldsymbol{\pi}$ can be written as

$$h_{\boldsymbol{\pi}}(d, s) = \left\{1 - \prod_{k\in s}(1-\pi_k)\right\}^{d} \left\{\prod_{k\in s}(1-\pi_k)\right\}^{1-d} \mathrm{P}(S = s).$$

The log-likelihood based on n observations $(d_1, s_1), \ldots, (d_n, s_n)$ of (D, S) is proportional to

$$\ell(\boldsymbol{\pi}) = \ell\big(\boldsymbol{\pi} \mid (d_1, s_1), \ldots, (d_n, s_n)\big) \propto \sum_{i=1}^{n} \left\{ d_i \log\Big(1 - \prod_{k\in s_i}(1-\pi_k)\Big) + (1-d_i)\log\Big(\prod_{k\in s_i}(1-\pi_k)\Big) \right\}.$$

$\mathrm{P}(S = s)$ does not depend on the genotype-specific risks π_k, as a consequence of assumption (A2), hence it can be left out of the log-likelihood.

The maximum likelihood estimator (MLE) maximizes the log-likelihood over all $\boldsymbol{\pi}$, i.e.

$$\hat{\boldsymbol{\pi}} = \underset{\boldsymbol{\pi}\in[0,1]^K}{\operatorname{argmax}}\, \ell(\boldsymbol{\pi}).$$

Here K is the total number of possible HPV genotypes in the set S, i.e. $K = 25$ when estimating the genotype-specific CIN2+ risk for all 25 HPV genotypes detectable by the SPF$_{10}$ and $K = 15$ when estimating the risks for only the 15 high-risk and probable high-risk HPV genotypes.

Newton's method is used to maximize the log-likelihood. The first and second order derivatives of $\ell(\boldsymbol{\pi})$ with respect to $\boldsymbol{\pi}$ are given below, first for the model where the genotype-specific risks π_k do not depend on age, then for the model where π_k depends on the woman's age x via a logit-link function

$$\operatorname{logit}\big(\pi_k(x)\big) = \alpha_k + \beta x. \tag{1}$$

Since for any k

$$\frac{\partial}{\partial \pi_k} \prod_{j\in s_i}(1-\pi_j) = -\prod_{j\in s_i\setminus\{k\}}(1-\pi_j) = -\frac{1}{1-\pi_k}\prod_{j\in s_i}(1-\pi_j),$$

it holds that

$$\begin{aligned}\frac{\partial}{\partial \pi_k} \log \Big(1 - \prod_{j \in s_i} (1-\pi_j)\Big) &= \frac{1}{1-\prod_{j\in s_i}(1-\pi_j)} \cdot \frac{\partial}{\partial \pi_k}\Big(1 - \prod_{j\in s_i}(1-\pi_j)\Big) \\ &= \frac{\prod_{j\in s_i}(1-\pi_j)}{(1-\pi_k)\Big(1-\prod_{j\in s_i}(1-\pi_j)\Big)},\end{aligned}$$

and

$$\begin{aligned}\frac{\partial}{\partial \pi_k} \log \prod_{j\in s_i}(1-\pi_j) &= \frac{1}{\prod_{j\in s_i}(1-\pi_j)} \cdot \frac{\partial}{\partial \pi_k} \prod_{j\in s_i}(1-\pi_j) \\ &= -\frac{1}{(1-\pi_k)}.\end{aligned}$$

Hence, the first order derivative of $\ell(\boldsymbol{\pi})$ is equal to

$$\frac{\partial \ell}{\partial \pi_k} = \sum_{i:k\in s_i} \left\{ d_i \frac{\prod_{j\in s_i}(1-\pi_j)}{(1-\pi_k)\Big(1-\prod_{j\in s_i}(1-\pi_j)\Big)} - (1-d_i)\frac{1}{1-\pi_k} \right\}.$$

Straightforward calculations for the second order derivative give

$$\begin{aligned}\frac{\partial^2 \ell}{\partial \pi_k^2} &= -\sum_{i:k\in s_i} \left\{ d_i \frac{\prod_{j\in s_i}(1-\pi_j)^2}{(1-\pi_k)^2\Big(1-\prod_{j\in s_i}(1-\pi_j)\Big)^2} + (1-d_i)\frac{1}{(1-\pi_k)^2} \right\} \\ \frac{\partial^2 \ell}{\partial \pi_k \partial \pi_l} &= -\sum_{i:k,l\in s_i} d_i \frac{\prod_{j\in s_i}(1-\pi_j)^2}{(1-\pi_k)(1-\pi_l)\Big(1-\prod_{j\in s_i}(1-\pi_j)\Big)^2}\end{aligned}$$

To derive the first and second order derivative of ℓ with respect to α_k and β in (1), we write the log-likelihood as

$$\begin{aligned}\ell(\boldsymbol{\pi}) &\propto \sum_{i=1}^n d_i \left\{ \log\Big(1-\prod_{k\in s_i}(1-\pi_k)\Big) - \log\Big(\prod_{k\in s_i}(1-\pi_k)\Big)\right\} + \sum_{i=1}^n \log\Big(\prod_{k\in s_i}(1-\pi_k)\Big) \\ &= \sum_{i=1}^n d_i \log \frac{1-\prod_{k\in s_i}(1-\pi_k)}{\prod_{k\in s_i}(1-\pi_k)} + \sum_{i=1}^n \log\Big(\prod_{k\in s_i}(1-\pi_k)\Big) \\ &= \sum_{i=1}^n -d_i \text{logit}\Big(\prod_{k\in s_i}(1-\pi_k)\Big) + \sum_{i=1}^n \log\Big(\prod_{k\in s_i}(1-\pi_k)\Big),\end{aligned}$$

and $\pi_k(x_i)$ as

$$\pi_k(x_i) = \frac{1}{1+\exp\big(-(\alpha_k+\beta x)\big)}.$$

Since for any k

$$\begin{aligned}1-\pi_k(x_i) &= \frac{\exp\big(-(\alpha_k+\beta x)\big)}{1+\exp\big(-(\alpha_k+\beta x)\big)}, \\ \frac{\partial}{\partial \alpha_k}\pi_k(x_i) &= \frac{1}{\big\{1+\exp\big(-(\alpha_k+\beta x)\big)\big\}^2} \cdot -1 \cdot \frac{\partial}{\partial \alpha_k}\exp\big(-(\alpha_k+\beta x)\big) \\ &= \frac{\exp\big(-(\alpha_k+\beta x)\big)}{\big\{1+\exp\big(-(\alpha_k+\beta x)\big)\big\}^2} \\ &= \pi_k(x_i)\big(1-\pi_k(x_i)\big), \\ \frac{\partial}{\partial \beta}\pi_k(x_i) &= x_i\pi_k(x_i)\big(1-\pi_k(x_i)\big),\end{aligned}$$

it holds that

$$
\begin{aligned}
\frac{\partial}{\partial\alpha_k}\prod_{j\in s_i}\big(1-\pi_j(x_i)\big) &= \prod_{j\in s_i\setminus\{k\}}\big(1-\pi_j(x_i)\big)\frac{\partial}{\partial\alpha_k}\big(1-\pi_k(x_i)\big)\\
&= \Big(\prod_{j\in s_i\setminus\{k\}}\big(1-\pi_j(x_i)\big)\Big)\cdot -\pi_k(x_i)\big(1-\pi_k(x_i)\big)\\
&= -\pi_k(x_i)\prod_{j\in s_i}\big(1-\pi_j(x_i)\big),\\
\frac{\partial}{\partial\beta}\prod_{j\in s_i}\big(1-\pi_j(x_i)\big) &= \sum_{l\in s_i}\frac{\partial}{\partial\beta}\big(1-\pi_l(x_i)\big)\cdot\prod_{j\in s_i\setminus\{l\}}\big(1-\pi_j(x_i)\big)\\
&= x_i\left(\sum_{l\in s_i}\pi_l(x_i)\right)\prod_{j\in s_i}\big(1-\pi_j(x_i)\big).
\end{aligned}
$$

Consequently,

$$
\begin{aligned}
\frac{\partial}{\partial\alpha_k}\log\Big(\prod_{j\in s_i}\big(1-\pi_j(x_i)\big)\Big) &= \frac{1}{\prod_{j\in s_i}\big(1-\pi_j(x_i)\big)}\cdot -\pi_k(x_i)\prod_{j\in s_i}\big(1-\pi_j(x_i)\big)\\
&= -\pi_k(x_i),\\
\frac{\partial}{\partial\beta}\log\Big(\prod_{j\in s_i}\big(1-\pi_j(x_i)\big)\Big) &= -x_i\sum_{j\in s_i}\pi_k(x_i),
\end{aligned}
$$

so that

$$
\begin{aligned}
\frac{\partial}{\partial\alpha_k}\frac{\prod_{j\in s_i}\big(1-\pi_j(x_i)\big)}{1-\prod_{j\in s_i}\big(1-\pi_j(x_i)\big)} &= \frac{-\pi_k(x_i)\prod_{j\in s_i}\big(1-\pi_j(x_i)\big)}{\Big(1-\prod_{j\in s_i}\big(1-\pi_j(x_i)\big)\Big)^2},\\
\frac{\partial}{\partial\alpha_k}\mathrm{logit}\Big(\prod_{j\in s_i}\big(1-\pi_j(x_i)\big)\Big) &= \frac{1-\prod_{j\in s_i}\big(1-\pi_j(x_i)\big)}{\prod_{j\in s_i}\big(1-\pi_j(x_i)\big)}\cdot\frac{-\pi_k(x_i)\prod_{j\in s_i}\big(1-\pi_j(x_i)\big)}{\Big(1-\prod_{j\in s_i}\big(1-\pi_j(x_i)\big)\Big)^2}\\
&= \frac{-\pi_k(x_i)}{1-\prod_{j\in s_i}\big(1-\pi_j(x_i)\big)},\\
\frac{\partial}{\partial\beta}\frac{\prod_{j\in s_i}\big(1-\pi_j(x_i)\big)}{1-\prod_{j\in s_i}\big(1-\pi_j(x_i)\big)} &= \frac{-x_i\sum_{j\in s_i}\pi_j(x_i)\prod_{j\in s_i}\big(1-\pi_j(x_i)\big)}{\Big(1-\prod_{j\in s_i}\big(1-\pi_j(x_i)\big)\Big)^2},\\
\frac{\partial}{\partial\beta}\mathrm{logit}\Big(\prod_{j\in s_i}\big(1-\pi_j(x_i)\big)\Big) &= \frac{-x_i\sum_{j\in s_i}\pi_j(x_i)}{1-\prod_{j\in s_i}\big(1-\pi_j(x_i)\big)}.
\end{aligned}
$$

The first order derivatives are then equal to

$$
\begin{aligned}
\frac{\partial\ell}{\partial\alpha_k} &= \sum_{i:k\in s_i}\left\{d_i\frac{\pi_k(x_i)}{1-\prod_{j\in s_i}\big(1-\pi_j(x_i)\big)}-\pi_k(x_i)\right\},\\
\frac{\partial\ell}{\partial\beta} &= \sum_{i=1}^{n}\left\{d_i\frac{x_i\sum_{j\in s_i}\pi_j(x_i)}{1-\prod_{j\in s_i}\big(1-\pi_j(x_i)\big)}-x_i\sum_{j\in s_i}\pi_j(x_i)\right\}.
\end{aligned}
$$

The second order derivatives then follow from straightforward calculations

$$
\begin{aligned}
\frac{\partial^2 \ell}{\partial \alpha_k^2} &= \sum_{i:k\in s_i} \left\{ d_i \frac{\pi_k(x_i)\Big(1-\pi_k(x_i)-\prod_{j\in s_i}\big(1-\pi_j(x_i)\big)\Big)}{\Big(1-\prod_{j\in s_i}\big(1-\pi_j(x_i)\big)\Big)^2} - \pi_k(x_i)\big(1-\pi_k(x_i)\big) \right\}, \\
\frac{\partial^2 \ell}{\partial \alpha_k \partial \alpha_l} &= \sum_{i:k,l\in s_i} d_i \frac{\pi_k(x_i)\pi_l(x_i)\prod_{j\in s_i}\big(1-\pi_j(x_i)\big)}{\Big(1-\prod_{j\in s_i}\big(1-\pi_j(x_i)\big)\Big)^2}, \\
\frac{\partial^2 \ell}{\partial \alpha_k \partial \beta} &= \sum_{i:k\in s_i} \left\{ d_i \frac{x_i\pi_k(x_i)\Big\{\big(1-\pi_k(x_i)\big)\Big(1-\prod_{j\in s_i}\big(1-\pi_j(x_i)\big)\Big) - \prod_{j\in s_i}\big(1-\pi_j(x_i)\big)\sum_{j\in s_i}\pi_j(x_i)\Big\}}{\Big(1-\prod_{j\in s_i}\big(1-\pi_j(x_i)\big)\Big)^2} \right. \\
&\qquad\qquad \left. - x_i\pi_k(x_i)\big(1-\pi_k(x_i)\big) \right\}, \\
\frac{\partial^2 \ell}{\partial \beta^2} &= \sum_{i=1}^{n} \left\{ d_i \frac{x_i^2\Big\{\Big(1-\prod_{j\in s_i}\big(1-\pi_j(x_i)\big)\Big)\sum_{j\in s_i}\pi_j(x_i)\big(1-\pi_j(x_i)\big) - \prod_{j\in s_i}\big(1-\pi_j(x_i)\big)\Big(\sum_{j\in s_i}\pi_j(x_i)\Big)^2\Big\}}{\Big(1-\prod_{j\in s_i}\big(1-\pi_j(x_i)\big)\Big)^2} \right. \\
&\qquad\qquad \left. - x_i^2 \sum_{j\in s_i}\pi_j(x_i)\big(1-\pi_j(x_i)\big) \right\}.
\end{aligned}
$$

eAppendix 2. R-code for maximum likelihood estimation

We provide the R-code to maximize the log-likelihood via Newton's method (`mle`) as well as an example of how to use the function for an example dataset of size $n = 400$. This dataset, in `.csv`-format, can be downloaded from the journal website. The first 5 women of the example dataset are displayed below to illustrate the structure of the input for our R-code.

1. Woman 1 is 29 years old, has a CIN2+ and tested positive only for HPV16.
2. Women 2 is 43 years old, does not have a CIN2+ and tested positive for HPV51, -52, and -54.
3. Women 3 is 28 years old, has a CIN2+ and tested positive for HPV35, -66, and -43.
4. Women 4 is 38 years old, has a CIN2+ and tested positive for HPV31, and -59.
5. Women 5 is 39 years old, does not have a CIN2+ and tested positive for HPV6, HPV44 and HPV74.

```
> data.full <- read.csv("example.csv",sep=",")
> head(data.full)
  age CIN2 HPV16 HPV18 HPV31 HPV33 HPV35 HPV39 HPV45 HPV51 HPV52 HPV53 HPV56 HPV58
1  29    1     1     0     0     0     0     0     0     0     0     0     0     0
2  43    0     0     0     0     0     0     0     0     1     1     0     0     0
3  28    1     0     0     0     0     1     0     0     0     0     0     0     0
4  38    1     0     0     1     0     0     0     0     0     0     0     0     0
5  39    0     0     0     0     0     0     0     0     0     0     0     0     0
  HPV59 HPV66 HPV68 HPV6 HPV11 HPV34 HPV40 HPV42 HPV43 HPV44 HPV54 HPV70 HPV74
1     0     0     0    0     0     0     0     0     0     0     0     0     0
2     0     0     0    0     0     0     0     0     0     0     1     0     0
3     0     1     0    0     0     0     0     0     1     0     0     0     0
4     1     0     0    0     0     0     0     0     0     0     0     0     0
5     0     0     0    1     0     0     0     0     0     1     0     0     1
> age <- data.full[,1]
> data <- data.full[,-1]
```

We first run the code without age as covariate, then with age as covariate. In the latter case, the risks and corresponding 95% confidence intervals (CIs), estimated number of lesions and the attributable fractions (AFs) are computed at the mean age.

```
> mle(data)
$mle
     HPV16      HPV18      HPV31      HPV33      HPV35      HPV39      HPV45
0.63180646 0.30478390 0.49427357 0.48147720 0.51897816 0.00001000 0.34159582
     HPV51      HPV52      HPV53      HPV56      HPV58      HPV59      HPV66
0.00001000 0.32962417 0.00001000 0.06744069 0.49307657 0.44134865 0.14090195
     HPV68       HPV6      HPV11      HPV34      HPV40      HPV42      HPV43
0.10565148 0.00001000 0.00001000 0.00001000 0.00001000 0.00001000 0.00001000
     HPV44      HPV54      HPV70      HPV74
0.00001000 0.00001000 0.00001000 0.00001000

$results
      #HPV+   MLE 95% CI       #lesions    AF
HPV16   137 0.632  0.538 0.726     86.6 0.430
HPV18    36 0.305  0.094 0.515     11.0 0.055
HPV31    56 0.494  0.322 0.666     27.7 0.138
HPV33    27 0.481  0.243 0.720     13.0 0.065
HPV35    27 0.519  0.278 0.760     14.0 0.070
HPV39    30 0.000  0.000 0.373      0.0 0.000
HPV45    14 0.342  0.000 0.684      4.8 0.024
HPV51    44 0.000  0.000 0.338      0.0 0.000
HPV52    60 0.330  0.164 0.495     19.8 0.098
HPV53    36 0.000  0.000 0.358      0.0 0.000
HPV56    24 0.067  0.000 0.193      1.6 0.008
HPV58    22 0.493  0.253 0.733     10.8 0.054
HPV59     6 0.441  0.000 1.000      2.6 0.013
HPV66    48 0.141  0.000 0.307      6.8 0.034
HPV68    25 0.106  0.000 0.292      2.6 0.013
```

```
HPV6     23 0.000  0.000 0.413      0.0 0.000
HPV11    19 0.000  0.000 0.462      0.0 0.000
HPV34     3 0.000  0.000 1.000      0.0 0.000
HPV40     3 0.000  0.000 1.000      0.0 0.000
HPV42     5 0.000  0.000 0.979      0.0 0.000
HPV43     2 0.000  0.000 1.000      0.0 0.000
HPV44    11 0.000  0.000 0.692      0.0 0.000
HPV54    12 0.000  0.000 0.600      0.0 0.000
HPV70     9 0.000  0.000 0.718      0.0 0.000
HPV74    11 0.000  0.000 0.641      0.0 0.000

> mle(data,logit=TRUE,cov=age)
$mle
      HPV16       HPV18       HPV31       HPV33       HPV35       HPV39       HPV45
  0.9695925  -0.1504835   0.3117804   0.3997860   0.6567388  -2.7831184  -0.1177573
      HPV51       HPV52       HPV53       HPV56       HPV58       HPV59       HPV66
 -2.9556953  -0.6593780  -3.2790702  -1.8581688   0.2267747 -14.3345966  -1.1551660
      HPV68        HPV6       HPV11       HPV34       HPV40       HPV42       HPV43
 -1.3739698 -14.3345966  -2.2988732  -5.8288991  -9.7604366  -5.5810660  -9.9640151
      HPV44       HPV54       HPV70       HPV74
 -3.2924816  -4.5718076  -2.2198799  -4.8961975  -0.0159972

$results
      #HPV+   MLE 95% CI       #lesions    AF
HPV16   137 0.575  0.473 0.672     78.8 0.439
HPV18    36 0.307  0.143 0.541     11.0 0.061
HPV31    56 0.413  0.255 0.591     23.1 0.129
HPV33    27 0.434  0.223 0.672     11.7 0.065
HPV35    27 0.498  0.274 0.723     13.4 0.075
HPV39    30 0.031  0.001 0.456      0.9 0.005
HPV45    14 0.314  0.086 0.689      4.4 0.024
HPV51    44 0.026  0.001 0.335      1.1 0.006
HPV52    60 0.210  0.098 0.394     12.6 0.070
HPV53    36 0.019  0.001 0.280      0.7 0.004
HPV56    24 0.074  0.011 0.368      1.8 0.010
HPV58    22 0.392  0.191 0.638      8.6 0.048
HPV59     6 0.000  0.000 1.000      0.0 0.000
HPV66    48 0.139  0.038 0.401      6.7 0.037
HPV68    25 0.115  0.017 0.499      2.9 0.016
HPV6     23 0.000  0.000 1.000      0.0 0.000
HPV11    19 0.049  0.002 0.548      0.9 0.005
HPV34     3 0.002  0.000 1.000      0.0 0.000
HPV40     3 0.000  0.000 1.000      0.0 0.000
HPV42     5 0.002  0.000 1.000      0.0 0.000
HPV43     2 0.000  0.000 1.000      0.0 0.000
HPV44    11 0.019  0.000 0.897      0.2 0.001
HPV54    12 0.005  0.000 0.978      0.1 0.000
HPV70     9 0.053  0.001 0.823      0.5 0.003
HPV74    11 0.004  0.000 0.994      0.0 0.000

$cov
0.145
```

Important to note is that the data matrix that serves as input for the R-function `mle` can only contain HPV genotypes for which at least one woman tested positive.

```
mle <- function(data,start=NULL,eps=10ˆ(-9),logit=FALSE,cov=NULL){
  # INPUT:
  #  data: data matrix with
  #        1) observed outcome D (0: no; 1: yes)
  #        2) observed types at risk for (0: no; 1: yes)
  #  start: starting vector for the algorithm
  #  eps: accuracy parameter to stop the algorithm
  #  logit: logical to indicate whether the logit model is used
  #         (i.e. correct for a covariate)
  #  cov: covariate vector to correct for (only if logit=TRUE)
  #
  # OUTPUT:
  #  mle: vector with estimated parameter values
  #  results: table with results (nr HPV+, MLE, 95% CI, nr lesions, AF)
  #  cov: p-value for the slope of the logit-model (only if logit=TRUE)

  K <- length(data[1,-1]); n <- length(data[,1])

  if (is.null(start)){
    start <- colSums(data[,-1])/n; start[start==0] <- 0.01; start[start==1] <- 0.99
    if (logit){
      start <- c(log(start/(1-start)),0)
    }
  }
  est <- as.matrix(start)

  ll <- numeric(0)
  m <- 1; diff1 <- diff2 <- 1

  while (max(abs(diff1),abs(diff2))>eps){
    ll.m <- log_lik(est[,m],data,logit,cov)
    if (((m>1)&&(ll.m$ll<ll[m-1]))||(ll.m$ll=="NaN")){
      x0 <- est[,m-1]
      ll.0 <- log_lik(x0,data,logit,cov)
      ll0 <-ll.0$ll
      dir <- ll.0$H.inv%*%ll.0$g
      est.n <- line_search(x0,ll0,dir,data,logit,cov)
      ll.m <- log_lik(est.n,data,logit,cov)
      est <- cbind(est[,-m],est.n)
      m <- m-1
    } else {
      ll <- c(ll,ll.m$ll)
      est <- cbind(est,est[,m]-ll.m$H.inv%*%ll.m$g)
    }

    if (logit){
      est[est[,m+1]<(-15),m+1] <- -15; est[est[,m+1]>10,m+1] <- 10
      diff1 <- 1/(1+exp(-est[,m+1]))-1/(1+exp(-est[,m]));
    } else {
      est[est[,m+1]<=eps*10ˆ(-3),m+1] <- 10ˆ(-5)
      est[est[,m+1]>=1-eps*10ˆ(-3),m+1] <- 1-10ˆ(-5)
      diff1 <- est[,m+1]-est[,m]
    }
    if (m>1){ diff2 <- ll[m]-ll[m-1] }
    m <- m+1
  }

  ll.m <- log_lik(est[,m-1],data,logit,cov)
  se <- sqrt(-diag(ll.m$H.inv))

  if (logit) { mle <- c(est[1:K,m-1]-mean(cov)*est[K+1,m-1],est[K+1,m-1])
  } else { mle <- est[,m-1] }

  if (logit){
    a <- mle[1:K]; b <- mle[K+1]; x <- mean(cov)
    mle.C <- a+b*x
    y1 <- mle.C-qnorm(0.975)*se[-(K+1)]; y2 <- mle.C+qnorm(0.975)*se[-(K+1)]
    CI <- cbind(1/(1+exp(-y1)),1/(1+exp(-y2)))
    mle.C <- 1/(1+exp(-mle.C))
  } else {
    CI <- cbind(mle-qnorm(0.975)*se,mle+qnorm(0.975)*se)
  }
  CI[CI[,1]<0,1] <- 10^(-5); CI[CI[,2]>1,2] <- 1-10^(-5)

  Wald.score <- abs(mle-rep(0,K+logit))/se
  p.val <- 2*(1-pnorm(Wald.score))

  if (logit) { nr <- colSums(data[,-1])*mle.C
  } else { nr <- colSums(data[,-1])*mle }
  AF <- nr/sum(nr)

  if (logit){
    out <- cbind(colSums(data[,-1]),round(mle.C,3),round(CI,3),round(nr,1),round(AF,3))
    dimnames(out)[[2]] <- c("#HPV+","MLE","95% CI","","#lesions","AF")
    list(mle=mle,results=out,cov=round(p.val[K+1],3))
  } else {
    out <- cbind(colSums(data[,-1]),round(mle,3),round(CI,3),round(nr,1),round(AF,3))
    dimnames(out)[[2]] <- c("#HPV+","MLE","95% CI","","#lesions","AF")
    list(mle=mle,results=out)
  }
}
```

The function `mle` uses the functions `log_lik` and `line_search`.

```
log_lik <- function(par,data,logit=FALSE,cov=NULL){
  # INPUT:
  #  par: parameter vector to determine the log-likelihood at
  #  data / logit / cov: as in mle-function
  #
  # OUTPUT:
  #  ll: value of the log-likelihood  at par
  #  g: gradient vector with partial derivative of log-likelihood
  #  H: Hessian matrix with the second partial derivative of log-likelihood
  #  H.inv: inverse of H

  K <- length(data[1,-1]); N <- length(data[,1])
  d <- data[,1]; Y <- data[,-1]
  if (logit){ X <- cov-mean(cov) }

  if (logit){
    pred <- matrix(rep(par[1:K],N),byrow=TRUE,ncol=K)+par[K+1]*X
    pi <- 1/(1+exp(-pred))
  } else {
    pi <- matrix(rep(par,N),byrow=TRUE,ncol=K)
  }

  prod <- apply((1-pi)^Y,1,prod)
  if (logit){
    sum.1 <- apply(1-(1-pi)^Y,1,sum)
    sum.2 <- apply((pi^Y)*(1-pi^Y),1,sum)
  }

  ll <- sum(d*log(1-prod)+(1-d)*log(prod))
  g <- numeric(K+logit); H <- matrix(0,ncol=K+logit,nrow=K+logit)

  for (k in 1:K){
    ind.k <- which(Y[,k]==1)

    if (logit){
      g[k] <- sum((d*pi[,k]/(1-prod)-pi[,k])[ind.k])
      H[k,k] <- sum((d*pi[,k]*(1-pi[,k]-prod)/(1-prod)ˆ2-pi[,k]*(1-pi[,k]))[ind.k])
      H[k,K+1] <- H[K+1,k] <- sum((X*(d*pi[,k]*((1-pi[,k])*(1-prod)-prod*sum.1)/(1-prod)ˆ2
                                  -pi[,k]*(1-pi[,k])))[ind.k])
    } else {
      g[k] <- sum((d*prod/((1-pi[,k])*(1-prod))-(1-d)/(1-pi[,k]))[ind.k])
      H[k,k] <- -sum((d*prodˆ2/((1-prod)ˆ2*(1-pi[,k])ˆ2)+(1-d)/(1-pi[,k])ˆ2)[ind.k])
    }

    for (l in min((k+1),K):K){
      if (k!=l){
        ind.kl<- which((Y[,k]==1)&(Y[,l]==1))

        if (logit){
          H[k,l] <- H[l,k] <- sum((d*pi[,k]*pi[,l]*prod/(1-prod)ˆ2)[ind.kl])
        } else {
          H[k,l] <- H[l,k] <- -sum((d*prodˆ2/((1-prod)ˆ2*(1-pi[,k])*(1-pi[,l])))[ind.kl])
        }
      }
    }
  }

  if (logit){
    g[K+1] <- sum(X*(d*sum.1/(1-prod)-sum.1))
    H[K+1,K+1] <- sum(Xˆ2*(d*((1-prod)*sum.2-prod*sum.1ˆ2)/(1-prod)ˆ2-sum.2))
  }

  if (logit){
    # invert H via block-inversion for case of "singularity" due to precision
    A <- H[1:K,1:K]; B <- H[K+1,1:K]; C <- H[1:K,K+1]; D <- H[K+1,K+1]
    A.inv <- solve(A); A.inv_B <- A.inv%*%B; C_A.inv <- C%*%A.inv
    L1 <- A.inv + (A.inv_B%*%L4)%*%C_A.inv; L2 <- -A.inv_B%*%L4
    L3 <- -L4%*%C_A.inv; L4 <- 1/(D-C%*%A.inv%*%B)
    H.inv <- rbind(cbind(L1,L2),cbind(L3,L4))
  } else {
    H.inv <- solve(H)
  }

  list(ll=ll,g=g,H=H,H.inv=H.inv)
}

line_search <- function(x0,ll0,dir,data,logit=FALSE,cov=NULL){
  # INPUT:
  #  x0: current estimate to start line search algorithm from
  #  ll0: log-likelihood value at x0
  #  dir: direction for line search (Hˆ(-1)*g)
  #  data / logit / cov: as in mle-function
  #
  # OUTPUT:
  #  x.n: new estimate based on line search

  lambda <- 1
  ll.n <- log_lik(x0,data,logit,cov)

  # determine the correction direction
  x.d <- x0-0.01*dir;
  if (logit) { x.d[x.d>10] <- 10; x.d[-x.d>15] <- -15
  } else { x.d[x.d<=0] <- 10ˆ(-5); x.d[x.d>=1] <- 1-10ˆ(-5) }
  ll.d <- log_lik(x.d,data,logit,cov)
  if (ll.d$ll<ll.n$ll) {dir <- -dir}
  ll.n$ll <- ll.n$ll-1

  while(ll.n$ll<ll0){
    lambda <- lambda/2
    x.n <- x0-lambda*dir;
    if (logit){ x.n[x.n>10] <- 10; x.n[-x.n>15] <- -15
    } else { x.n[x.n<=0] <- 10ˆ(-5); x.n[x.n>=1] <- 1-10ˆ(-5) }
    ll.n <- log_lik(x.n,data,logit,cov)
  }

  return(x.n)
}
```

eAppendix 3. Supplementary tables and figures

eTable 1: Number and proportion of women with a positive HPV genotype result in the cervical samples and in the lesions as detected by laser capture microdissection-PCR, stratified by country.

	the Netherlands						**Spain**					
	cervical sample		**CIN2+**		**CIN3+**		**cervical sample**		**CIN2+**		**CIN3+**	
HPV genotype	n	**(%)**	n	**(%)**[a]	n	**(%)**[a]	n	**(%)**	n	**(%)**[a]	n	**(%)**[a]
(probable) high-risk												
16	28	(35)	22	(79)	15	(54)	161	(37)	107	(66)	60	(37)
18	9	(11)	3	(33)	2	(22)	34	(7.9)	13	(38)	6	(18)
31	14	(17)	10	(71)	5	(36)	71	(16)	31	(44)	8	(11)
33	5	(6.2)	2	(40)	0	(0)	24	(5.6)	16	(67)	6	(25)
35	8	(9.9)	3	(38)	3	(38)	22	(5.1)	6	(27)	3	(14)
39	5	(6.2)	0	(0)	0	(0)	32	(7.4)	5	(16)	0	(0)
45	0	(0)	0	-	0	(0)	12	(2.8)	4	(33)	1	(8.3)
51	14	(17)	6	(43)	1	(7.1)	46	(11)	4	(8.7)	3	(6.5)
52	6	(7.4)	1	(17)	1	(17)	57	(13)	13	(23)	5	(8.8)
53	9	(11)	0	(0)	0	(0)	35	(8.1)	1	(2.9)	0	(0)
56	5	(6.2)	2	(40)	1	(20)	31	(7.2)	4	(13)	1	(3.2)
58	3	(3.7)	2	(67)	1	(33)	21	(4.9)	10	(48)	5	(24)
59	6	(7.4)	3	(50)	2	(33)	5	(1.2)	0	(0)	0	(0)
66	12	(15)	3	(25)	0	(0)	39	(9.0)	4	(10)	2	(5.1)
68	4	(4.9)	0	(0)	0	(0)	27	(6.3)	4	(15)	2	(7.4)
low risk												
6	2	(2.5)	0	(0)	0	(0)	17	(3.9)	0	(0)	0	(0)
11	5	(6.2)	0	(0)	0	(0)	10	(2.3)	0	(0)	0	(0)
34	1	(1.2)	0	(0)	0	(0)	6	(1.4)	0	(0)	0	(0)
40	0	(0)	0	-	0	-	2	(0.46)	0	(0)	0	(0)
42	0	(0)	0	-	0	-	5	(1.2)	0	(0)	0	(0)
43	0	(0)	0	-	0	-	2	(0.46)	0	(0)	0	(0)
44	2	(2.5)	0	(0)	0	(0)	7	(1.6)	0	(0)	0	(0)
54	3	(3.7)	0	(0)	0	(0)	13	(3.0)	1	(7.7)	0	(0)
70	3	(3.7)	0	(0)	0	(0)	7	(1.6)	0	(0)	0	(0)
74	3	(3.7)	0	(0)	0	(0)	10	(2.3)	0	(0)	0	(0)
Total	147		57		31		696		223		102	

[a] percentage of women positive for that HPV genotype.

eTable 2: Estimated genotype-specific CIN2+ and CIN3+ risk (π_k) and number of lesions (n) for all 25 HPV genotypes.

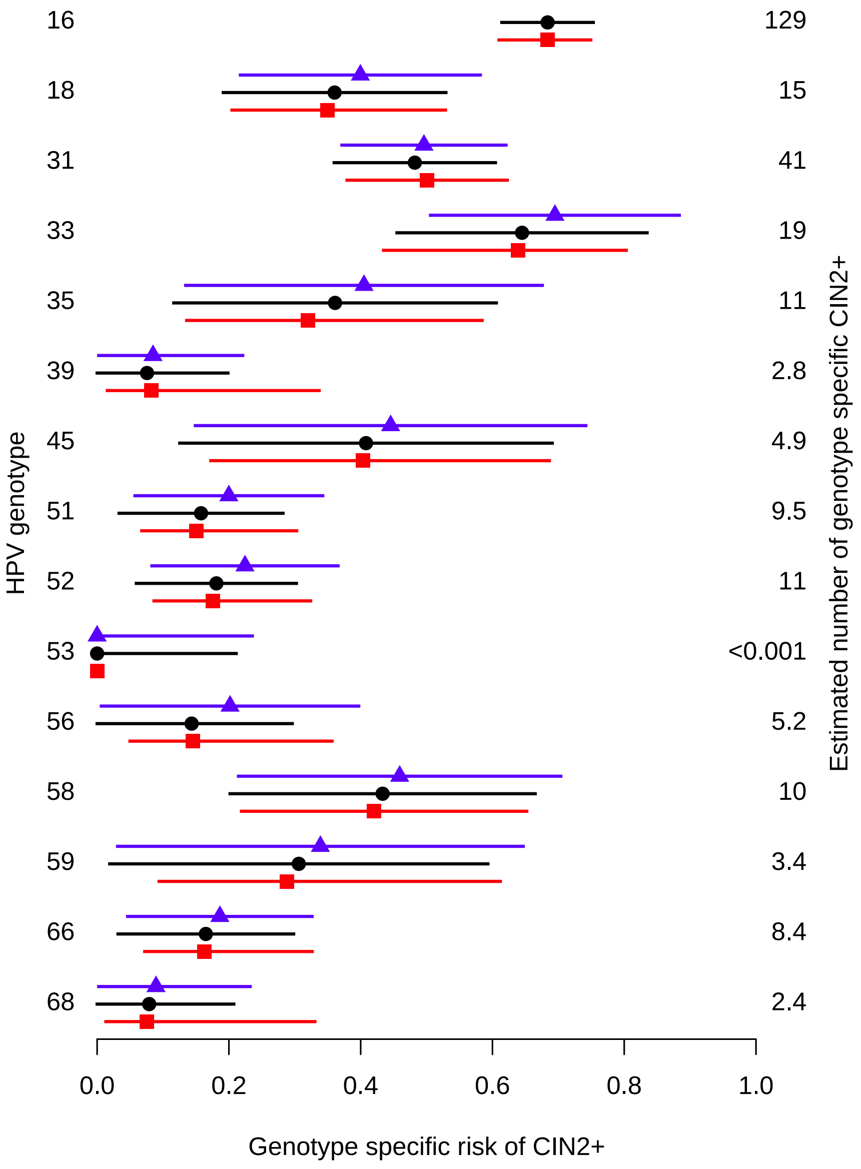
	CIN2+			**CIN3+**		
HPV genotype	π_k	**95% CI**	n	π_k	**95% CI**	n
(probable) high-risk						
16	0.68	0.61, 0.75	129	0.41	0.34, 0.49	78
18	0.36	0.19, 0.53	15	0.14	0.027, 0.26	6.2
31	0.48	0.36, 0.60	41	0.11	0.031, 0.18	9.0
33	0.64	0.46, 0.83	19	0.22	0.064, 0.37	6.3
35	0.36	0.12, 0.60	11	0.20	0.028, 0.38	6.1
39	0.076	<0.001, 0.20	2.8	<0.001	<0.001, 0.32	<0.001
45	0.41	0.13, 0.69	4.9	0.14	<0.001, 0.32	1.7
51	0.16	0.033, 0.28	9.5	0.058	<0.001, 0.13	3.5
52	0.18	0.059, 0.30	11	0.061	<0.001, 0.14	3.8
53	<0.001	<0.001, 0.21	<0.001	<0.001	<0.001, 0.29	<0.001
56	0.14	<0.001, 0.30	5.2	<0.001	<0.001, 0.30	<0.001
58	0.43	0.20, 0.66	10	0.23	0.039, 0.43	5.6
59	0.31	0.019, 0.59	3.4	0.24	<0.001, 0.51	2.6
66	0.16	0.032, 0.30	8.4	0.043	<0.001, 0.10	2.2
68	0.079	<0.001, 0.21	2.4	<0.001	<0.001, 0.13	0.004
low risk						
6	<0.001	<0.001, 0.35	<0.001	<0.001	<0.001, 0.44	<0.001
11	<0.001	<0.001, 0.51	<0.001	<0.001	<0.001, 0.51	<0.001
34	<0.001	<0.001, 0.85	<0.001	<0.001	<0.001, 0.75	<0.001
40	<0.001	n.r.	<0.001	<0.001	n.r.	<0.001
42	<0.001	<0.001, 0.88	<0.001	<0.001	<0.001, 0.88	<0.001
43	<0.001	n.r.	<0.001	<0.001	n.r.	<0.001
44	<0.001	<0.001, 0.35	<0.001	<0.001	<0.001, 0.65	<0.001
54	<0.001	<0.001, 0.53	<0.001	<0.001	<0.001, 0.48	<0.001
70	0.007	<0.001, 0.58	0.066	<0.001	<0.001, 0.59	<0.001
74	<0.001	<0.001, 0.63	<0.001	<0.001	<0.001, 0.52	<0.001

CI=confidence interval; n.r.=95% CI not reported because estimate was indeterminate.

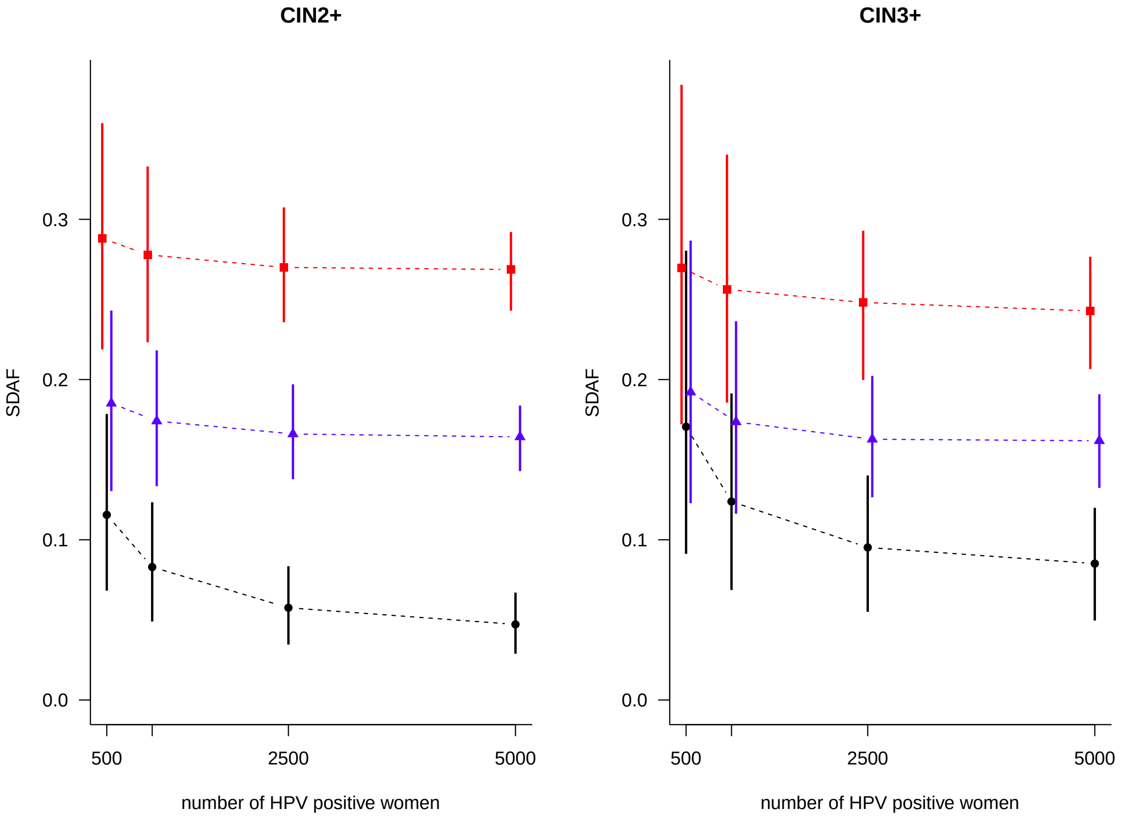
eTable 3: Estimated genotype-specific CIN2+ and CIN3+ risk (π_k) and number of lesions (n) for only the 15 high-risk or probable high-risk HPV genotypes.

	CIN2+			**CIN3+**		
HPV genotype	π_k	**95% CI**	n	π_k	**95% CI**	n
16	0.68	0.61, 0.75	128	0.40	0.33, 0.47	75
18	0.35	0.18, 0.52	15	0.14	0.026, 0.26	6.1
31	0.48	0.36, 0.60	41	0.11	0.032, 0.18	9.2
33	0.64	0.45, 0.83	19	0.22	0.065, 0.37	6.4
35	0.34	0.098, 0.57	10	0.21	0.029, 0.38	6.2
39	0.076	<0.001, 0.20	2.8	<0.001	<0.001, 0.31	<0.001
45	0.41	0.13, 0.69	4.9	0.14	<0.001, 0.34	1.7
51	0.15	0.039, 0.27	9.3	0.059	<0.001, 0.13	3.5
52	0.18	0.062, 0.30	12	0.064	<0.001, 0.14	4.0
53	<0.001	<0.001, 0.19	<0.001	<0.001	<0.001, 0.29	<0.001
56	0.14	<0.001, 0.29	5.1	<0.001	<0.001, 0.30	<0.001
58	0.43	0.20, 0.66	10	0.24	0.043, 0.44	5.8
59	0.31	0.019, 0.59	3.4	0.24	<0.001, 0.52	2.6
66	0.16	0.035, 0.28	8.1	0.043	<0.001, 0.10	2.2
68	0.079	<0.001, 0.21	2.4	0.001	<0.001, 0.14	0.024

CI=confidence interval



eFigure 1: Estimated genotype-specific risk of CIN2+ with corresponding 95% CIs, for all positive women not adjusted for age (black dots), for high-risk and probable high-risk positive women only adjusted for age (red squares), and HPV16 negative women not adjusted for age (blue triangles). The standard error of the age-adjusted estimate for HPV53 (red square) was indeterminate and the 95% CI could not be computed.



eFigure 2: Median (and 95% confidence interval) SDAF of our method (black dots), the proportional approach (blue triangle) and the hierarchical approach (red square) for different sample sizes of HPV-positive women in the simulated sample.