Web-based Supplementary Materials for Estimating strain-specific and overall efficacy of polyvalent vaccines against recurrent pathogens from a cross-sectional study by Kari Auranen, Hanna Rinta-Kokko and M. Elizabeth Halloran

WEB APPENDIX A

Here we provide the proofs of Results 1 and 2. The necessary notation was introduced in Sections 2 and 3. Recall that V and $\bar{V}$ denote the sets of vaccine and non-vaccine strains, respectively, and $\bar{V}_0 = \bar{V} \cup \{0\}$. Given the set $W \subseteq V$ of the target strains, sets $\bar{\mathcal{V}}_0$, $\mathcal{W}$ and $\mathcal{R}$ define a partition of the colonisation states $\mathcal{S}$ (Figure 2).

The proof of results 1 and 2 is based on the reversibility of the underlying processes of colonisation in the vaccinees (T) and controls (C) and is given here in full detail for model A. The applicability of these results under model B is then discussed. Recall that conditions (A1) and (A2) are true also in the vaccinated group under the assumed vaccine model (A3)–(A4). In the proof of the following lemma the superscripts T and C are omitted for simplicity.

Lemma. The Markov process satisfying conditions (A1) and (A2) is reversible, i.e., the condition of detailed balance, $p_{[h]}q_{[h][k]} = p_{[k]}q_{[k][h]}$, holds for all $[h], [k] \in \mathcal{S}$. Here p is the stationary distribution of the respective process.

Proof. Consider the set of equations: $p_0 q_{0,j} = p_j \mu$, $j = 1, \ldots, n$, and $p_i q_{i,ij} = p_{ij} q_{ij,i}$, $i = 1, \ldots, n-1$; $j = i+1, \ldots, n$. Together with the normalising condition $\sum_{i=0}^n p_i + \sum_{i<j}^n p_{ij} = 1$, these $n(n+1)/2 + 1$ equations have a unique solution vector p. The detailed balance with respect to distribution p thus applies between state 0 and any of the states $j \in S$ as well as between any state i and (i,j) when $i < j$. It follows from assumptions (A1) and (A2) that the detailed balance also applies between state i and (i,j) when $i > j$, since $p_i q_{i,ij} = p_i(q_{0,j}/q_{0,i})q_{j,ij} = p_j q_{j,ij} = p_{ij}\mu = p_{ij}q_{ij,i}$.

It follows from reversibility and the fact that the hazards of colonisation ($q_{0,i}$) are the same in vaccinees and controls for any of the non-vaccine strains ($i \in \bar{V}$) that the following equalities hold for all $i \in \bar{V}$ (non-vaccine strains), for all $j \in S$ (any strain), and for $k = T, C$ (for both vaccinees and controls):

$$
\begin{aligned}
(a) \quad & p_i^k / p_0^k = q_{0,i}^k / \mu = q_{0,i}^C / \mu, \\
(b) \quad & p_{ij}^k / p_0^k = (p_{ij}^k / p_j^k)(p_j^k / p_0^k) = (q_{j,ij}^k / \mu)(q_{0,j}^k / \mu) \\
& = (q_{j,ij}^C / \mu)(q_{0,j}^k / \mu), \\
(c) \quad & (p_i^k / p_0^k)(q_{i,ij}^k / q_{0,j}^k) = (p_{ij}^k / p_0^k)(p_i^k q_{i,ij}^k / p_{ij}^k)(1 / q_{0,j}^k) \\
& = (p_{ij}^k / p_0^k)(\mu / q_{0,j}^k) = q_{j,ij}^C / \mu.
\end{aligned}
$$

Starting from the definition of $\mathrm{VE}_{W|\bar{V}_0}$ in equation (2), and applying equalities (a) and (c) it follows that

$$
\begin{aligned}
\mathrm{VE}_{W|\bar{V}_0} &= 1 - \frac{p_0^T \sum_{j \in W} q_{0,j}^T + \sum_{i \in \bar{V}, j \in W} p_i^T q_{i,ij}^T}{p_0^T + \sum_{i \in \bar{V}} p_i^T} \times \frac{p_0^C + \sum_{i \in \bar{V}} p_i^C}{p_0^C \sum_{j \in W} q_{0,j}^C + \sum_{i \in \bar{V}, j \in W} p_i^C q_{i,ij}^C} \\
&= 1 - \frac{p_0^T \sum_{j \in W} q_{0,j}^T \left(1 + \sum_{i \in \bar{V}} (p_i^T / p_0^T)(q_{i,ij}^T / q_{0,j}^T)\right)}{p_0^C \sum_{j \in W} q_{0,j}^C \left(1 + \sum_{i \in \bar{V}} (p_i^C / p_0^C)(q_{i,ij}^C / q_{0,j}^C)\right)} \times \frac{p_0^C \left(1 + \sum_{i \in \bar{V}} (p_i^C / p_0^C)\right)}{p_0^T \left(1 + \sum_{i \in \bar{V}} (p_i^T / p_0^T)\right)} \\
&= 1 - \frac{\sum_{j \in W} q_{0,j}^T \left(1 + \sum_{i \in \bar{V}} q_{j,ij}^C / \mu\right)}{\sum_{j \in W} q_{0,j}^C \left(1 + \sum_{i \in \bar{V}} q_{j,ij}^C / \mu\right)}. \qquad \text{(E1)}
\end{aligned}
$$

This can be written as a weighted average of strain-specific efficacy values, with weights $q_{0,j}^C (1 + \sum_{i \in \bar{V}} q_{j,ij}^C / \mu),\ j \in W$.

Alternatively, starting from expression (6) and applying equalities (a) and (b) it follows that

$$1-\frac{\displaystyle\sum_{[j]\in\mathcal{W}} p_j^T \Big/ \sum_{[j]\in\bar{\mathcal{V}}_0} p_j^T}{\displaystyle\sum_{[j]\in\mathcal{W}} p_j^C \Big/ \sum_{[j]\in\bar{\mathcal{V}}_0} p_j^C}$$

$$=1-\frac{p_0^T\left(\displaystyle\sum_{j\in W}(p_j^T/p_0^T)+\sum_{i\in\bar{V},j\in W}(p_{ij}^T/p_0^T)\right)}{p_0^C\left(\displaystyle\sum_{j\in W}(p_j^C/p_0^C)+\sum_{i\in\bar{V},j\in W}(p_{ij}^C/p_0^T)\right)}\times\frac{p_0^C\left(1+\displaystyle\sum_{i\in\bar{V}}\left((p_i^C/p_0^C)+\sum_{\substack{j\in\bar{V}\\ j\neq i}}(p_{ij}^C/p_0^C)\right)\right)}{p_0^T\left(1+\displaystyle\sum_{i\in\bar{V}}\left((p_i^T/p_0^T)+\sum_{\substack{j\in\bar{V}\\ j\neq i}}(p_{ij}^T/p_0^T)\right)\right)}$$

$$=1-\frac{\displaystyle\sum_{j\in W} q_{0,j}^T\left(1+\sum_{i\in\bar{V}} q_{j,ij}^C/\mu\right)}{\displaystyle\sum_{j\in W} q_{0,j}^C\left(1+\sum_{i\in\bar{V}} q_{j,ij}^C/\mu\right)}. \tag{E2}$$

This equals the expression of $\mathrm{VE}_{W|\bar{V}_0}$ as derived above, completing the proof of Result 2. Result 1 for $\mathrm{VE}_{W|0}$ follows by omitting the terms $\sum_{i\in\bar{V},j\in W}(p_{ij}/p_0)$ in the numerator and denominator of equation (E2), which then reads as

$$1-\frac{\displaystyle\sum_{[j]\in W} p_j^T \Big/ \sum_{[j]\in\bar{\mathcal{V}}_0} p_j^T}{\displaystyle\sum_{[j]\in W} p_j^C \Big/ \sum_{[j]\in\bar{\mathcal{V}}_0} p_j^C}=\mathrm{VE}_{W|0}.$$

Efficacy against the non-vaccine strains. It follows from equality (a) that the expression (5) of the vaccine efficacy against the non-vaccine strains is 0:

$$\mathrm{VE}_{\bar{V}|0}=1-\frac{\displaystyle\sum_{[i]\in\bar{V}} p_i^T\Big/p_0^T}{\displaystyle\sum_{[i]\in\bar{V}} p_i^C\Big/p_0^C}=1-\frac{\displaystyle\sum_{i\in\bar{V}} q_{0,i}^C/\mu}{\displaystyle\sum_{i\in\bar{V}} q_{0,i}^C/\mu}=0.$$

Cross-sectional estimation under model B. Above, cross-sectional estimators were considered under model A (Figure 1A). Result 1 holds also under model B (Figure 1B) which fulfils conditions (B1)–(B4). The proof follows from equation (E2) by omitting all terms that involve doubly-colonised states. Of note, estimator (6) is not defined under model B so that cross-sectional estimation of $\mathrm{VE}_{W|\bar{V}_0}$ is not possible under model B. In this paper, model B was applied to three aggregated states $\mathcal{V}_0$, $\mathcal{W}$, and $\mathcal{R}$ for estimation of vaccine efficacy against a subset of strains W (see Section 5).

Relation of the two efficacy estimands. The two efficacy estimands, $\mathrm{VE}_{W|0}$ and $\mathrm{VE}_{W|\bar{V}_0}$, are both weighted averages of strain specific efficacy estimands. According to equations (1) and (E1), the weights are $q^C_{0,j}$ and $w_j = q^C_{0,j}(1 + a_j)$, respectively, where $a_j = \sum_{i \in \bar{V}} q^C_{j,ij}/\mu$. The following crude approximations hold:

$$\frac{1 + a_{min}}{1 + a_{max}} \mathrm{VE}_{W|0} = \frac{1 + a_{min}}{1 + a_{max}} \frac{\sum\limits_j q^C_{0,j} \cdot \mathrm{VE}_j}{\sum\limits_j q^C_{0,j}} \leq \frac{\sum\limits_j w_j \cdot \mathrm{VE}_j}{\sum\limits_j w_j}$$

$$\leq \frac{1 + a_{max}}{1 + a_{min}} \frac{\sum\limits_j q^C_{0,j} \cdot \mathrm{VE}_j}{\sum\limits_j q^C_{0,j}} = \frac{1 + a_{max}}{1 + a_{min}} \mathrm{VE}_{W|0}.$$

It follows that

$$\frac{1 + a_{min}}{1 + a_{max}} \leq \frac{\mathrm{VE}_{W|\bar{V}_0}}{\mathrm{VE}_{W|0}} \leq \frac{1 + a_{max}}{1 + a_{min}}.$$

The ratio of the two estimands is close to one if there is little variation in the weights a_j or they are all small. According to equation (c), $a_j = \sum_{i \in \bar{V}} (p^C_i/p^C_0)(q^C_{i,ij}/q^C_{0,j})$. The first possibility thus occurs if between-strain competition is homogeneous across all strains, i.e., the relative reduction $(q^C_{i,ij}/q^C_{0,j})$ in the acquisition hazard due to current colonisation has the same value for all strain pairs. The second possibility occurs in a setting in which non-vaccine strain and double colonisation are not very common. For example, if $p_0 \geq 1/3$ and $q^C_{i,ij}/q^C_{0,j} < 0.1$ for all i, j, $a_j < 0.3 p_{\bar{V}}$ (for all j) where $p_{\bar{V}}$ is the prevalence of non-vaccine strains.

The aggregated model.

Here we prove that estimand (8) of vaccine efficacy under the aggregate model is equivalent to estimand (2) for $\mathrm{VE}_{W|\bar{V}_0}$. Starting from (8) and using the fact that $q^T_{[h],[k]} = q^C_{[h],[k]} = 0$ for all $[h] \in \bar{\mathcal{V}} \backslash \bar{V}_0$ (doubly-colonised states with non-vaccine strains), we find that

$$
\begin{aligned}
& 1 - \frac{q^T_{\bar{\mathcal{V}}_0,\mathcal{W}}}{q^C_{\bar{\mathcal{V}}_0,\mathcal{W}}} \\
&= 1 - \frac{\displaystyle\sum_{[h]\in\bar{\mathcal{V}}_0} \bar{p}^T_{[h]|\bar{\mathcal{V}}_0} \left(\sum_{[k]\in\mathcal{W}} q^T_{[h][k]} \right)}{\displaystyle\sum_{[h]\in\bar{\mathcal{V}}_0} \bar{p}^C_{[h]|\bar{\mathcal{V}}_0} \left(\sum_{[k]\in\mathcal{W}} q^C_{[h][k]} \right)} = 1 - \frac{\displaystyle\sum_{[h]\in\bar{V}_0} p^T_{[h]|\bar{\mathcal{V}}_0} \left(\sum_{[k]\in\mathcal{W}} q^T_{[h][k]} \right)}{\displaystyle\sum_{[h]\in\bar{V}_0} p^C_{[h]|\bar{\mathcal{V}}_0} \left(\sum_{[k]\in\mathcal{W}} q^C_{[h][k]} \right)} \\
&= 1 - \frac{\displaystyle\sum_{[h]\in\bar{V}_0} p^T_{[h]} \left(\sum_{[k]\in\mathcal{W}} q^T_{[h][k]} \right)}{\displaystyle\sum_{[h]\in\bar{V}_0} p^C_{[h]} \left(\sum_{[k]\in\mathcal{W}} q^C_{[h][k]} \right)} \frac{\displaystyle\sum_{[h]\in\bar{\mathcal{V}}_0} p^C_{[h]}}{\displaystyle\sum_{[h]\in\bar{\mathcal{V}}_0} p^T_{[h]}} \\
&= 1 - \frac{\displaystyle\sum_{[h]\in\bar{V}_0} p^T_{[h]} \left(\sum_{[k]\in\mathcal{W}} q^T_{[h][k]} \right)}{\displaystyle\sum_{[h]\in\bar{V}_0} p^C_{[h]} \left(\sum_{[k]\in\mathcal{W}} q^C_{[h][k]} \right)} \frac{\displaystyle\sum_{[h]\in\bar{V}_0} p^C_{[h]}}{\displaystyle\sum_{[h]\in\bar{V}_0} p^T_{[h]}} \\
&= 1 - \frac{\displaystyle\sum_{[h]\in\bar{V}_0} \bar{p}^T_{[h]|\bar{V}_0} \left(\sum_{[k]\in\mathcal{W}} q^T_{[h][k]} \right)}{\displaystyle\sum_{[h]\in\bar{V}_0} \bar{p}^C_{[h]|\bar{V}_0} \left(\sum_{[k]\in\mathcal{W}} q^C_{[h][k]} \right)} = \mathrm{VE}_{W|\bar{V}_0}.
\end{aligned}
$$

The equivalence on the second line above is based on the fact that the ratio of the normalising constants pertaining to sets $\bar{\mathcal{V}}_0$ and $\bar{V}_0$ are equal by (a) and (b):

$$\frac{\sum\limits_{i\in\bar{V}_0} p_i^T + \sum\limits_{i,j\in\bar{V};i\neq j} p_{ij}^T}{\sum\limits_{i\in\bar{V}_0} p_i^C + \sum\limits_{i,j\in\bar{V};i\neq j} p_{ij}^C} = \frac{(\sum\limits_{i\in\bar{V}_0} q_{0,i}^C + \sum\limits_{i,j\in\bar{V};i\neq j} q_{i,ij}^C q_{0,j}^C/\mu)p_0^T}{(\sum\limits_{i\in\bar{V}_0} q_{0,i}^C + \sum\limits_{i,j\in\bar{V};i\neq j} q_{i,ij}^C q_{0,j}^C/\mu)p_0^C}$$

$$= \frac{(\sum\limits_{i\in\bar{V}_0} q_{0,i}^C)p_0^T}{(\sum\limits_{i\in\bar{V}_0} q_{0,i}^C)p_0^C} = \frac{\sum\limits_{i\in\bar{V}_0} p_i^T}{\sum\limits_{i\in\bar{V}_0} p_i^C}.$$

Convergence to stationarity. The speed of convergence towards stationarity depends on the second largest eigenvalue of the transition intensity matrix of the underlying Markov chain (Levin, Peres and Wilmer, 2009). Assume that conditions (A1)-(A4) hold and there is strong between-strain competition in the sense that $q_{j,ij}/q_{0,i} = 0$ for all i, j. For $n = 2$ strains, the transition probability matrix of the approximating Markov chain is

$$D = (d_{ij}) = \begin{pmatrix} d_{11} & p_{0,1} & p_{0,2} \\ p_{\cdot,0} & d_{22} & 0 \\ p_{\cdot,0} & 0 & d_{33} \end{pmatrix},$$

where $p_{i,j}$ is the transition probability from state i to state j over a short (infinitesimal) time interval Δt, and $d_{ii} = 1 - \sum_{j;j\neq i} d_{ij}$. The eigenvalues of D are 1, $1-p_{\cdot,0}$, and $1-p_{\cdot,0}-p_{0,1}-p_{0,2}$ so that the second largest eigenvalue is $1-p_{\cdot,0} = 1-\mu\Delta t$. Likewise, in a model with n strains, the second largest eigenvalue can be shown to be $1 - \mu\Delta t$ (with multiplicity $n-1$). This implies that under strong competition ($q_{j,ij}/q_{0,i} \simeq 0$) the parameter with the most influence on convergence is the clearance rate of colonisation.

Additional reference:

Levin D. A., Peres Y., Wilmer E. L. Markov chains and mixing times (2009). American Mathematical Society, Providence, Rhode Island.

WEB APPENDIX B

This Appendix provides the R code that can be used to analyse the 9-strain example as presented in Table 3 in the main text.

```
# This is the main R file for the estimation of vaccine efficacy
# against colonisation with multiple strains. The program with its
# example data set is implemented for a setting with 9 strains, analogous
# to the setting presented in Table 3 of the article (Auranen K,
# Rinta-Kokko H, Halloran ME. "Estimating strain-specific and overall efficacy of
# polyvalent vaccines against pathogens with recurrent dynamics from a
# cross-sectional study").
#
# HR-K & KA, May 31, 2012
#
# For the vaccine strains, three different estimands can be considered,
# either (a) for VE_{W|0} (estimand (1) of the article), or
#        (b) VE_{W|V0}    (estimand (2) of the article).
#        (c) "HRK" (a simple odds ratio, see the article for details).
#
# For the non-vaccine strains, the estimand is always given by expression (3)
# in the article. See the article for more details.
#
# The following scripts are used:
source("VE_grouping.R") # To group the states of colonisation
                        # and to call function VE_publ.R
source("VE_estimate.R") # To estimate the vaccine efficacy

# Simulated example data sets for vaccinees and controls, based on one
# cross-section of 1000 individuals in both groups, are provided below
# in the data file data_vaccinees_controls.csv.
#
# There are nine strains in this example, strains 1-4 being the vaccine strains
# and strains 5-9 the non-vaccine strains. In the simulation of these data,
# the vaccine efficacies against strains 1-4 and the overall vaccine
# efficacy against types 1-4 combined are 0.7, 0.4, 0.7, 0.4, 0.61, respectively.
#
# N.B. This is a simulated example data. The results in Table 3 are based on
#      1000 repetitions from the model.

# Read in the data sets in vector format (states of colonisation coded as explained below)
datav = data.frame(read.table("data_vaccinees_controls.csv",header=TRUE,sep=";"))[,1] # vaccinees
datac = data.frame(read.table("data_vaccinees_controls.csv",header=TRUE,sep=";"))[,2] # controls

# The number of strains in the example data
n_s = 9

# The n_s*(n_s-1)/2 = 45 states of colonisation (in the 9-strain model)
states = c(1,2,3,4,5,6,7,8,9,12,13,14,15,16,17,18,19,23,24,25,26,27,28,29,34,35,36,37,38,39,45,46,47,48,49,
56,57,58,59,67,68,69,78,79,89)

# Sub-sets of colonisation states (defined for the 9-strain model below)
#   (a) single colonisation with a vaccine strain;
#   (b) single colonisation with a non-vaccine strain;
#   (c) double colonisation with two vaccine strains;
#   (d) double colonisation with a vaccine strain and a non-vaccine strain
#   (e) double colonisation with two non-vaccine strains
```

```
vt     = c(1,2,3,4)                        # (a)
nvt    = c(5,6,7,8,9)                      # (b)
vtvt   = c(12,13,14,23,24,34)              # (c)
vtnvt  = c(15,16,17,18,19,25,26,27,28,29,35,36,37,38,39,45,46,47,48,49) #(d)
nvtnvt = c(56,57,58,59,67,68,69,78,79,89)  # (e)


# The target set of strains referst to those strains against which
# the vaccine efficacy is to be estimated. In the current implementation,
# this set of strains must be chosen so that all strain belong to either
# to the vaccine of non-vaccine strains.
#
# In the following, vaccine efficacy is estimated against each of the individual
# strains (1,2,...,9) as well as 'overall' against all vaccine-strains (1,2,3,4)
# and against all non-vaccine strains (5,6,7,8,9).
#
# In general, the estimand is defined by choosing the input parameter 'estimand'
# as either "W|V" or "W|0" in the routine that calculates the efficacy (see below).
# (a)  "W|V"  (this yields an estimate for estimand (2), i.e. VE_W\vert\bar V_0)
# (b)  "W|0"   (this yields an estimate for estimad (1), i.e. VE_W\vert 0)
#
# In the following example, the estimate for the vaccine strains is calcualated for "W|V".
#
# The output is a matrix with 11 rows (vaccine efficacy (VE) for each of
# 9 strains individuals, overall VE against the vaccine strains,
# and overall VE against the non-vaccine strains). The columns are:
# VE and the lower and upper bounds of a 90\% confidence interval.

# Initialise the output matrix
resmat    = matrix(0, ncol=3, nrow=n_s+2, dimnames=list(c(seq(1:n_s), "VT", "NVT"),
                   c("VE", "CI_VE_low", "CI_VE_high")))

# Choose the esimator (the other option estimator = "W|0")
estimand   = "W|V"

resmat[1,]  = VE_grouping(datav, datac, estimand, target = 1, vt, nvt, nvtnvt)   # strain 1
resmat[2,]  = VE_grouping(datav, datac, estimand, target = 2, vt, nvt, nvtnvt)   # strain 2
resmat[3,]  = VE_grouping(datav, datac, estimand, target = 3, vt, nvt, nvtnvt)   # strain 3
resmat[4,]  = VE_grouping(datav, datac, estimand, target = 4, vt, nvt, nvtnvt)   # strain 4
resmat[5,]  = VE_grouping(datav, datac, estimand, target = 5, vt, nvt, nvtnvt)   # strain 5
resmat[6,]  = VE_grouping(datav, datac, estimand, target = 6, vt, nvt, nvtnvt)   # strain 6
resmat[7,]  = VE_grouping(datav, datac, estimand, target = 7, vt, nvt, nvtnvt)   # strain 7
resmat[8,]  = VE_grouping(datav, datac, estimand, target = 8, vt, nvt, nvtnvt)   # strain 8
resmat[9,]  = VE_grouping(datav, datac, estimand, target = 9, vt, nvt, nvtnvt)   # strain 9
resmat[10,] = VE_grouping(datav, datac, estimand, target = vt, vt, nvt, nvtnvt)  # all vaccine strains
resmat[11,] = VE_grouping(datav, datac, estimand, target = nvt, vt, nvt, nvtnvt) # all non-vaccine strains
resmat



########################
# Subroutine VE_grouping
########################

VE_grouping = function(datavacc, datacontr, estimand, target, vt, nvt, nvtnvt){
#
# HR-K & KA, May 31, 2012
#
# This function calculates the vaccine efficacy against a select set of
# strains, i.e., the target set of strains. The strains of the target
# set must all belong to either the vaccine or non-vaccine sets of strains.
```

```
#
# If estimator "W|V" is used, the function first appropriately classifies
# the states of colonisation.
#
# This function call function VE_publ.R to actually estimate the vaccine efficacy.
#
# INPUT
#  target:    a vector of strain numbers for which the vaccine efficacy is to be estimated;
#             eg., target = c(2) for strain 2; e.g., target = vt for all vaccine strains
#             listed in the input list 'vt'
#  datavacc:  the observations from the vaccinees, in a vector
#  datacontr: the observations from the controls, in a vector
#  estimand:  which vaccine efficacy estimand is considered: "W|0" or "W|V"
#  vt:        single vaccine strains in a list
#  nvt:       single non-vaccine strains in a list
#  nvtnvt:    states of double colonisation with two non-vaccine strains in a list
#
# OUTPUT = a vector of vaccine efficacy and the lower and upper bounds of the 90\% confidence interval.
#
# This funtion calls function "VE_estimate.R".

# Initialse the output matrix
res = rep(0,4)

###########################################
# Estimand (1) of the article (VE_{W|0})
###########################################
if(estimand=="W|0"){
target_set   = target # the target set of colonisation states is equal to
                           # the singly colonised states with a vaccine strain
res          = VE_estimate(target_set, datavacc, datacontr, nvt,  nvtnvt, estimand)
}

##################################################
# Estimand (2) of the article (VE_{W|V_0})
##################################################
if(estimand=="W|V"){

if(target[1] \%in\% vt){

# Form the appropriate set of colonisation states from
# single colonisation with a target strain and double
# colonisation of a target srain with any of the non-vaccine strains


       # Initialise
target_set = rep(0,length(target)*(length(nvt)+1))
h = 1
       for(j in 1:length(target)){
         target_set[h] = target[j] # single colonisation
           h    = h+1
           k    = 1

           # Concatenate the target strain with the non-vaccine strain, one at a time
     while(k <= length(nvt)){
             d             = as.numeric(paste(target[j],nvt [k], sep=""))
  target_set[h] = d
  h             = h+1
  k             = k+1
      }
  }
}
if(target[1] \%in\% nvt){
target_set = target
}

    res = VE_estimate(target_set, datavacc, datacontr, nvt, nvtnvt,  estimand)
}



###########################################
# Estimand "HRK"
###########################################
if(estimand=="HRK"){
target_set   = target # the target set of colonisation states is equal to
                           # the singly colonised states with a vaccine strain
res          = VE_estimate(target_set, datavacc, datacontr, nvt,  nvtnvt, estimand)
}



  # Output: VE, the lower and upper bounds of the 90\% confidence interval
  VE_grouping = c(res[1][[1]], res[2][[1]], res[3][[1]])

}



########################
# Subroutine VE_estimate
########################

VE_estimate = function(target_set, datavacc, datacontr, nvt, nvtnvt, estimand){
#
# HR-K & KA, May 31, 2012
#
# This function calculates the estimate of vaccine efficacy as an odds ratio.
# Estimator (1) or (2) of the article is used. This is defined by grouping of
# states of coloniation realised already in the calling subroutine (VE_groupind.R).
#
# INPUT
#  target_set : the target *set* of colonisation states (realised by the calling subroutine;
#               see more details below)
#  datavacc   : the observations from the vaccinees
#  datacontr  : the observations from the controls
#  nvt        : the non-vaccine strains as a list
#  nvtnvt     : the states of double colonisation with two non-vaccine strains as a list
#
# OUTPUT      : the output is a table of vaccine efficacy and the lower and upper bounds of the 90\% CI



 # Vaccinees: determine the total numbers of samples in the target
 # and appropriate reference states of colonisation
 nvacc_target = length(datavacc[datavacc \%in\% target_set]) # the total number of those colonised with a
                                                           # target strain (singly or with a non-vaccine strain)
 nvaccNVT     = length(datavacc[datavacc \%in\% nvt])        # the total number of those colonised with one
                                                           # non-vaccine strain only
 nvaccNVTNVT  = length(datavacc[datavacc \%in\% nvtnvt])     # the total number of those colonised with two
                                                           # non-vaccine strains
 nvacc0       = length(datavacc[datavacc==0])              # the total number of non-colonised
 nvaccTot   = length(datavacc)        # the total number of samples



 # Controls: determine the total numbers of samples in the target
 # and appropriate reference states of colonisation
 ncontr_target = length(datacontr[datacontr \%in\% target_set]) # the total number of those colonised with a
                                                              # target strain (singly or with a non-vaccine strain)
 ncontrNVT     = length(datacontr[datacontr \%in\% nvt])        # the total number of those colonised with
                                                              # one non-vaccine strain only
 ncontrNVTNVT  = length(datacontr[datacontr \%in\% nvtnvt])     # the total number of those colonised with
                                                              # two non-vaccine strains
 ncontr0       = length(datacontr[datacontr==0])              # the total number of non-colonised
 ncontrTot   = length(datacontr)           # the total number of samples


if(estimand == "W|V" | estimand == "W|0"){

# For any non-vaccine strain(s), the reference set is the non-colonised
if(target_set[1] \%in\% nvt){
target_vacc     = nvacc_target
reference_vacc  = nvacc0
target_contr    = ncontr_target
reference_contr = ncontr0
}



  # For any of the vaccine strain(s), the reference set is the  non-colonised + those colonised
  # with one (only) or two non-vaccine strains
if(target_set[1] \%in\% vt){
target_vacc     = nvacc_target
reference_vacc  = nvacc0+nvaccNVT+nvaccNVTNVT
target_contr    = ncontr_target
reference_contr = ncontr0+ncontrNVT+ncontrNVTNVT
}
}

if(estimand == "HRK"){
target_vacc = nvacc_target
reference_vacc = nvaccTot-nvacc_target
target_contr = ncontr_target
reference_contr = ncontrTot-ncontr_target
}



# Next, the odds ratio is calculated using logistic regression.
# Point estimates of the odds and odds ratio could be calculated also as follows:
# odds_target = target_vacc/target_contr             # the odds of being vaccinated among those
                                                     # colonised with the target states
# odds_reference = reference_vacc/reference_contr    # the odds of being vaccinated among those
                                                     # colonised with the reference states
# OR              = odds_target/odds_reference



# Working matrix for logistic regression
tr = data.frame(matrix(0,ncol=2,nrow=target_vacc+target_contr+reference_vacc+reference_contr,
     dimnames=list(seq(1:(target_vacc+target_contr+reference_vacc+reference_contr)),
     c("Vacc","Targ"))))
if(target_vacc>0){
tr[(1:target_vacc),1]<-rep(1,target_vacc)
}
if(reference_vacc>0){
tr[(target_vacc+target_contr+1):(target_vacc+target_contr +reference_vacc),1]<-rep(1,reference_vacc)
}
if(target_vacc>0 | target_contr>0){
tr[1:(target_vacc+target_contr),2]<-rep(1,(target_vacc +target_contr))
}


# Logistic regression
glm.fit = glm(tr\$Vacc~tr\$Targ, family=binomial(link="logit"))


logOR = coef(glm.fit)[2]
OR    = exp(coef(glm.fit)[2])[[1]]


# Vaccine efficacy
VE = 1-OR


# Standard error, 90\% confidence interval
SE_logOR       = summary(glm.fit)\$coef[2,2]
LowerCI_logOR  = logOR+1.645*SE_logOR
HigherCI_logOR = logOR-1.645*SE_logOR


LowerCI_VE  = 1-exp(LowerCI_logOR)
HigherCI_VE = 1-exp(HigherCI_logOR)


# Output
VE_estimate = data.frame(VE = VE, LowerCI_VE = LowerCI_VE, HigherCI_VE = HigherCI_VE)


}
```

WEB APPENDIX C

Here we provide more detailed results to some simulation analyses in Sections 5 and 6.

The slightly negative bias in the point estimates in Table 3 is due to bias in odds-ratio estimates from finite samples. A general reference supporting this claim is Nelson (1972). We further investigated this by simulating two scenarios: (i) a sample of the same size (N=1000 per group, as in Table 3) but taking the sample at day 730, i.e., very long after vaccination to ensure that the samples would be drawn from the equilibrium distribution; (ii) a very large sample (N=10,000 per group), taking the sample of colonisation from each individual at the same time (183 days) after vaccination as in the example of Table 3. Under (i), we found that there was larger bias for the rare strains. By contrast, under (ii), when the sample size was very large, the bias disappeared also for the rare strains. The results from scenario (ii) are presented in Table C1.

[Table 1 about here.]

[Table 2 about here.]

[Table 3 about here.]

The theoretical results of Section 5 mean that the only assumption required for cross-sectional estimation of overall and strain-specific efficacies is (B5). Table C2 shows results from a simulation study in which vaccination enhances or decelerates clearance of the vaccine strains, thus violating condition (B5). Under enhanced clearance, the combined efficacy against acquisition and clearance can be estimated from a cross-sectional sample. The combined efficacy cannot be estimated equally well under decelerated clearance.

The theoretical results of Section 5 also mean that cross-sectional estimators for overall and strain-specific efficacy are applicable with any difference in the clearance rates of the target strain and those in the class “rest”, at least under symmetric competition. We verified this by a simulation study in which the ratio of the clearance rates for the target and the rest was as large (small) as 2 (0.5) (Table C3).

Additional reference:

Nelson W. Statistical methods for the ratio of two multinomial proportions. The American Statistician 1972;26(3):22-27.

Table C1

Performance of cross-sectional estimators of vaccine efficacy when sample size is very large. Perfect detection of the singly and doubly-colonised states is assumed. The results are based on 1000 simulated data sets, each with 10000 vaccinees and 10000 controls; otherwise the model is identical to that of the Table 3.

Strain	**True efficacy**	**No. of singly and doubly colonised samples† (% of all samples)**		**Estimates of vaccine efficacy (SE log(OR); 90% cov. prob.)††**	
		Vaccinees	**Controls**	$\mathrm{VE}_{W\|\bar{V}_0}$	$\mathrm{VE}_{W\|0}$
1	**0.7**	656 + 193(8.5)	1658 + 608(22.7)	0.70(0.04; 0.91)	0.70(0.05; 0.91)
2	**0.4**	525 + 161(6.9)	660 + 312(9.7)	0.40(0.06; 0.89)	0.40(0.06; 0.89)
3	**0.7**	175 + 59(2.3)	441 + 217(6.6)	0.70(0.08; 0.91)	0.70(0.09; 0.89)
4	**0.4**	175 + 58(2.3)	220 + 113(3.3)	0.40(0.09; 0.89)	0.40(0.10; 0.90)
5	0	1466 + 337(18.0)	1103 + 469(15.7)	0.00(0.05; 0.92)*	0.00(0.05; 0.92)*
6	0	875 + 244(11.2)	661 + 311(9.7)	0.00(0.06; 0.89)*	0.00(0.06; 0.89)*
7	0	262 + 88(3.5)	198 + 102(3.0)	−0.00(0.10; 0.89)*	−0.00(0.10; 0.89)*
8	0	146 + 48(1.9)	110 + 59(1.7)	−0.01(0.13; 0.90)*	−0.01(0.13; 0.90)*
9	0	87 + 30(1.2)	66 + 35(1.0)	0.00(0.17; 0.90)*	0.00(0.17; 0.90)*
Vaccine strains	**0.61**	1531 + 408(19.4)	2979 + 962(39.4)	0.61(0.03; 0.89)	0.61(0.04; 0.87)
Non-vaccine strains	0	2836 + 546(33.8)	2138 + 825(29.6)	0.00(0.04; 0.90)*	0.00(0.04; 0.90)*
0		5024(50.2)	3770(37.7)	–	–

∗ For efficacy against non-vaccine strains, the reference state is 0 and estimator (5) was used.

† The mean number of samples in the 1000 simulated data sets.

†† The average standard error of the log odds ratio, based on the 1000 simulated data sets. The coverage probability of a 90% confidence interval, based on the 1000 simulated data sets.

Table C2

Cross-sectional estimation of the combined vaccine efficacy when the vaccine affects duration of colonisation. The results are based on 1000 simulated data sets (see Table 3). For the vaccine strains, the hazard of clearance in the vaccinees was set α times that in the controls ($\alpha = 1, 0.5, 2$). The simulation was started from state 0 (uncolonised) for each individual and the sample for calculating the vaccine efficacy was taken at day 183. The estimates are based on estimator $VE_{W|\bar{V}_0}$.

The relative clearance rate in vaccinees vs. controls (α)	**Combined vaccine efficacy***	**Estimate of vaccine efficacy (SE log(OR); 90% cov. prob.)†**
1	0.61	0.61(0.10; 0.91)
0.5	0.22	0.32(0.10; 0.55)
2	0.81	0.80(0.12; 0.88)

$*$ A combined measure of the effect of being vaccinated on susceptibility to acquisition and on duration of colonisation; see text.

† The average standard error for the log odds ratio, based on the 1000 simulated data sets. The coverage probability of a 90% confidence interval, based on the 1000 simulated data sets.

Table C3

Cross-sectional estimation of vaccine efficacy when the clearance rates of the target vaccine strain and other strains are different. The results are based on 1000 simulated data sets (see Table 3). The hazard of clearance of the target strain was set β times that of all other strains ($\beta = 1, 0.5, 2$).The simulation was started from state 0 (uncolonised) for each individual and the sample for calculating the vaccine efficacy was taken at day 183. The estimates are based on estimator $VE_{W|\bar{V}_0}$.

The relative clearance rate		**Target vaccine strain**		**Other vaccine strains**
of the target strain vs.	**True**	**Estimate**	**True**	**Estimate**
other strains	**efficacy**	**(SE log(OR); 90% cov. prob.)†**	**efficacy**	**(SE log(OR); 90% cov. prob.)†**
1	0.70	0.70(0.14; 0.89)	0.50	0.50(0.14; 0.91)
0.5	0.70	0.71(0.12; 0.87)	0.50	0.50(0.14; 0.89)
2	0.70	0.70(0.19; 0.91)	0.50	0.50(0.13; 0.89)

† The average standard error for the log odds ratio, based on the 1000 simulated data sets. The coverage probability of a 90% confidence interval, based on the 1000 simulated data sets.