Journal of Data Science **4**(2006), 189-205

Estimating Vaccine Efficacy from Household Data Using Surrogate Outcome and a Validation Sample

Xiaohong M. Davis[1] and Michael Haber[2,*]
[1] *Centers for Disease Control Prevention and* [2] *Emory University*

Abstract: Household data are frequently used in estimating vaccine efficacy because it provides information about every individual's exposure to vaccinated and unvaccinated infected household members. This information is essential for reliable estimation of vaccine efficacy for infectiousness (VE_I), in addition to estimating vaccine efficacy for susceptibility (VE_S). However, accurate infection outcome data is not always available on each person due to high cost or lack of feasible methods to collect this information. Lack of reliable data on true infection status may result in biased or inefficient estimates of vaccine efficacy. In this paper, a semiparametric method that uses surrogate outcome data and a validation sample is introduced for estimation of VE_S and VE_I from a sample of households. The surrogate outcome data is usually based on illness symptoms. We report the results of simulations conducted to examine the performance of the estimates, compare the proposed semiparametric method with maximum likelihood methods that either use the validation data only or use the surrogate data only and address study design issues. The new method shows improved precision as compared to a method based on the validation sample only and smaller bias as compared to a method using surrogate outcome data only. In addition, the use of household data is shown to greatly improve the attenuation in the estimate of VE_S due to misclassification of the outcome, as compared to the use of a random sample of unrelated individuals.

Key words: Mismeasured observations, semiparametric model surrogate outcome, vaccine efficacy for infectiousness, vaccine efficacy for susceptibility, validation sample.

1. Introduction

Estimation of vaccine efficacy has traditionally focused on the vaccine-induced reduction in susceptibility to infection, or vaccine efficacy for susceptibility (VE_S). However a vaccine, such as a prophylactic HIV vaccine, may also lower the infectiousness of a vaccinated person who became infected (Longini *et al.*, 1996). The relative reduction in infectiousness due to a vaccine is the vaccine efficacy for

infectiousness or VE_I. Both VE_S and VE_I are measures of the true biological effects of a vaccine.

In general, VE is expressed as 1-RR, where RR is a measure of relative risk in vaccinated individuals compared to unvaccinated individuals, under the assumption of equal exposure to the infectious agent. Different levels of information are required to estimate VE_S depending on what parameterization is used (Halloran *et al.*, 1997). Haber *et al.* (1991) defined VE_S in terms of the transmission probability to a susceptible individual who makes a contact with an infectious person. VE_S is defined as one minus the ratio of the transmission probabilities to a vaccinated and an unvaccinated susceptible person when both are exposed to the same source of infection. VE_I measures the effect of a vaccine on infectiousness of a vaccinated infected person. It is defined as one minus the ratio of the transmission probabilities from a vaccinated and an unvaccinated infected individual when they make contacts with a susceptible person (Koopman and Little, 1995). Estimation of VE_I is challenging because it requires information on exposure to infection, and gathering this type of information is often expensive, difficult or even impossible. Therefore, VE_I cannot be estimated from a sample of unrelated individuals. Data based on a sample of households provide information on everyone's exposure to both vaccinated and unvaccinated infected individuals. The information on infections contracted from vaccinated persons who became infected is essential for reliable estimation of VE_I. Davis and Haber (2001) developed a maximum likelihood method for the estimation of VE_S and VE_I from household data.

The problem of estimating VE_S and VE_I is further complicated by the fact that reliable infection outcome data is often expensive or difficult to collect from each individual in a vaccine study. For example, in an influenza vaccine study, a culture or a quick test of a sputum or a blood sample would be required to confirm infection (Halloran and Longini, 2001). Confirming all individuals in the study by cultures or samples can be very expensive and time consuming. Often, a closely related outcome may be used as a surrogate for the infection outcome. For example, an illness outcome defined as 'any respiratory illness,' can be used as a surrogate for the infection outcome in an influenza vaccine study.

The use of surrogate outcome variables is common in medical research, especially in clinical settings (Prentice, 1989; Wittes *et al.*, 1989; Fleming *et al.*, 1994). In identifying 'valid' surrogates, Prentice (1989) suggested the criteria that a test of the null hypothesis using a surrogate w provides valid inference regarding the true outcome x. He also provided general guidelines for choosing variables to satisfy this definition of surrogacy. According to his definition, a key property of a potential surrogate is that $P(x|w, m) = P(x|w)$ almost surely, where m is a covariates or a treatment indicator. This implies that the effect of

treatment on the true outcome should act solely through the surrogate w. This is the foundation for making inference about the true outcome based solely on the surrogate. However, this assumption may not be satisfied in many applications. For example, in the case of an infectious disease it is possible that the vaccine affects the probability that an ill person is indeed infected. To relax this assumption, Pepe (1992) proposed a semi-parametric method that uses a validation sample to relate the true and surrogate outcomes. She showed that this semiparametric method allows direct inference regarding the association between the true outcome and the covariates.

Golm *et al.* (1998, 1999) explored the use of semiparametric methods with validation samples for exposure-to-infection information to estimate VE_I in trials of human immunodeficiency virus vaccines. Their methods assume that exposure-to-infection, which is a covariate, may be mismeasured while the outcome (infection) is always correctly assessed. Halloran and Longini (2001) illustrated the use of validation sets to correct the attenuated estimate of VE_S for mismeasured outcome data. They used an example of influenza vaccine efficacy and effectiveness trials under the assumption that the group of influenza-like cases includes true and misspecified influenza infection cases. Halloran and Longini multiplied an estimated probability (which is assumed constant over time) of an influenza-like case being true influenza infection in each vaccination stratum (i.e., vaccinated or nonvaccinated) when estimating VE_S alone from final attack rates. Currently, there is no method available for estimating VE_S and VE_I from data with mismeasured outcome information.

The purpose of this work is to develop and evaluate a semiparametric method for simultaneous estimation of VE_S and VE_I from household data when the true infection status is observed on everybody in a validation sample of households and a surrogate illness outcome is observed on every study participant. We extend the method of Pepe (1992) to the case where the units of analysis are households of various sizes, the true outcome is the array of the (correlated) infection statuses of all household members, the surrogate outcome is the corresponding array of illness statuses and the treatment indicator is the corresponding array of vaccination statuses. As we mentioned earlier, household data is used because it contains information on the vaccination and infection or illness status of each household member. In other words, it provides information about every individual's exposure to vaccinated and unvaccinated infected or ill household members, which is necessary for reliable estimation of the vaccine effect on infectiousness. One should note that for a study participant in a household where the true infection statuses may be misclassified, we have incomplete information on both the outcome variable (her/his own infection status) and the exposure variables (the infection statuses of all other household members).

2. Estimation Methods

2.1 Study design

We consider an outbreak of an infectious disease which is transmitted from person to person in a closed community. Once a susceptible person becomes infected, she or he is infectious to others for a relatively short time and then becomes immune at least until the outbreak is over. The community consists of many small transmission units, which will be referred to as households. (Sexual partnerships can be viewed as households of size two). For simplicity, we assume that everybody, except for a small number of initial infectives, is susceptible at the beginning of the study. (Individuals who are initially immune can be excluded from the study without loss of any relevant information). A susceptible person can become infected from an infectious household member or from 'the community', i.e. from an infectious person in another household. Prior to the outbreak, individuals may be vaccinated with a 'leaky' vaccine, i.e., a vaccine that reduces their susceptibility by lowering their probability of becoming infected. The vaccine may also reduce an individual's infectiousness by lowering her/his probability of infecting others in the case she/he becomes infected (a vaccine breakthrough). The main purpose of the study is to evaluate the vaccine's effects on the susceptibility and infectiousness of a vaccinee as compared to an unvaccinated person.

For the purpose of the study, we assume that two samples of households are selected from the community. In the first sample, which will be referred to as the validation sample, both the true infection outcome and a related surrogate outcome, which is usually based on illness symptoms, are available for all the members of every household. We denote the number of households in the validation sample by N_v. In the second sample, which will be called the surrogate sample, only the surrogate illness outcome is known for everyone. There are $N - N_v$ households in the surrogate sample, where N is the total number of households in the study.

Consider a household with $s = s_0 + s_1$ initial susceptibles, where s_0 and s_1 are the number of unvaccinated and vaccinated susceptible household members, respectively. Let m_i denote the vaccination status of person i, with $m_i = 1$ for vaccinated and $m_i = 0$ for unvaccinated. The array $\mathbf{m} = (m_1, ..., m_s)$ denotes the vaccination statuses of all the susceptible household members. Let x_i be the infection status of person i at the end of the outbreak, with $x_i = 1$ for infected and $x_i = 0$ for uninfected. Finally, let w_i be the surrogate outcome (i.e., illness) of person i with $w_i = 1$ for ill and $w_i = 0$ for not ill. For households in the validation sample, the true infection outcome array $\mathbf{x} = (x_1, ..., x_s)$ and the surrogate outcome $\mathbf{w} = (w_1, ..., w_s)$ are known. For households in the surrogate

sample, only the surrogate outcome array **w** is known. Table 1 describes the data structure.

Table 1: Sample data structure of a household study with a validation sample of N_v households and a surrogate sample of $N - N_v$ households; s is the size (number of initial susceptibles) of the household; **m** is the array of vaccination statuses of all household members; **x** is the true infection outcome; **w** is the surrogate illness outcome. $\triangle$ is an indicator with 1 representing a household in the validation sample and 0 representing a household in the surrogate sample.

Household	s	**m**	**x**	**w**	$\triangle$
1	3	(0,1,0)	(0,0,1)	(0,1,1)	1
2	2	(1,1)	(0,0)	(0,1)	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
N_v	4	(0,1,1,0)	(1,0,1,0)	(1,1,0,1)	1
$N_v + 1$	2	(1,0)	–	(0,1)	0
$N_v + 2$	4	(0,1,0,1)	–	(1,1,0,1)	0
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
N	3	(1,0,0)	–	(1,0,0)	0

2.2 Calculation of $P(\mathbf{x}|\mathbf{m})$

To write an expression for the probability $P(\mathbf{x}|\mathbf{m})$ of infection outcome **x** in a household with vaccination pattern **m**, we first need to define the transmission probabilities and the effects of the vaccine. Let β denote the probability that an unvaccinated susceptible becomes infected from the community during the course of the epidemic, and let γ denote the probability that the same person is infected from an unvaccinated household member while the latter is infectious. The vaccine efficacy for susceptibility, VE_S, is the relative reduction due to vaccination in the transmission probability to a vaccinated susceptible. The vaccine efficacy for infectiousness, VE_I, is the relative reduction due to vaccination in the transmission probability from a vaccinated infectious person. Define $\theta = 1 - VE_S$ and $\varphi = 1 - VE_I$. Then the transmission probability from the community to a vaccinated susceptible is $\beta \cdot \theta$. The transmission probability from an infected person to a susceptible household member is $\gamma \cdot \theta$ when the susceptible person is vaccinated and the infected is unvaccinated; it is $\gamma \cdot \varphi$ when the susceptible

is unvaccinated and the infected is vaccinated; and it is $\gamma \cdot \theta \cdot \varphi$ when both are vaccinated.

For the infection outcome $\mathbf{x}$, let j_0 and j_1 be the number of infected persons among the unvaccinated and vaccinated household members, respectively. Then $j = j_0 + j_1 = \sum x_i$. Let $\mathbf{J}$ denote the subset of the j household members who became infected. Then for $j = 0, 1, 2, ..., s-1$ (i.e., not everybody in the household became infected):

$$\begin{aligned} P(\mathbf{x}|\mathbf{m}) &= P(\mathbf{1}|\mathbf{J})(1-\beta)^{s_0-j_0}(1-\theta\beta)^{s_1-j_1}(1-\gamma)^{j_0(s_0-j_0)}(1-\theta\gamma)^{j_0(s_1-j_1)} \\ &\quad (1-\varphi\gamma)^{j_1(s_0-j_0)}(1-\theta\varphi\gamma)^{j_1(s_1-j_1)}. \end{aligned} \tag{2.1}$$

The first term in (2.1) denotes the probability that everybody in subset $\mathbf{J}$ became infected if there were no other members in the household. The second and third terms are the probabilities that all non-infected unvaccinated and vaccinated household members, respectively, escaped infection from the community. The next two terms are the probabilities that all non-infected unvaccinated and vaccinated household members escaped infection from the j_0 unvaccinated infected members. The last two terms are the corresponding escape probabilities from the j_1 vaccinated infected members. For a proof of (2.1) see Longini *et al.* (1988).

The probability that everybody in the household became infected, i.e., $P(\mathbf{x} = \mathbf{1}|\mathbf{m})$, is obtained as one minus the sum of all the expressions (2.1) over $j = 0, 1, 2, ..., s-1$. Thus, a recursive computation is involved in calculating the probabilities of the infection outcomes, $\mathbf{x}$. For a household of size s, one needs to first calculate the probabilities of all possible outcomes for all the households of sizes $s' = 1, 2, ..., s-1$.

If the true infection outcome is available for all the study participants, then the likelihood function is obtained as the product of all the terms $P(\mathbf{x}|\mathbf{m})$ over all the households in the study. Maximization of the likelihood will then provide estimates of the parameters β, γ, θ and φ (Davis and Haber, 2001).

2.3 The semiparametric method

We propose a semiparametric method to estimate θ and φ (i.e., VE_S and VE_I) using the surrogate and validation samples. The validation sample is used to relate the true and the surrogate outcomes ($\mathbf{x}$ and $\mathbf{w}$) and thus to reduce the bias of the parameter estimates. The surrogate sample is used to improve the efficiency of the estimates. A semiparametric method is used to avoid specification or misspecification of the relationship between the true outcome and the surrogate outcome while still making valid inference on the parameters of interest (Pepe, 1992). A semiparametric method that places no structure on the conditional

probability function $P(\mathbf{w}|\mathbf{x},\mathbf{m})$ is desirable since the relationship between the true outcome $\mathbf{x}$ and the surrogate outcome $\mathbf{w}$ is not of primary interest.

Given that no structure is specified for $P(\mathbf{w}|\mathbf{x},\mathbf{m})$, we assume that $P(\mathbf{w}|\mathbf{x},\mathbf{m})$ is independent of $\mathbf{\Theta}$, where $\mathbf{\Theta} = (\beta,\gamma,\theta,\varphi)$. In other words, the *parameters* related to transmission and vaccine effects do not affect that probability that an infected person develops illness symptoms. On the other hand, we allow the probability of illness given infection to depend on the *actual* vaccination status. Then, an empirical estimator of $P(\mathbf{w}|\mathbf{x},\mathbf{m})$ is found using the validation sample:

$$\hat{P}(\mathbf{w}|\mathbf{x},\mathbf{m}) = \hat{P}(\mathbf{w},\mathbf{x},\mathbf{m})/\hat{P}(\mathbf{x},\mathbf{m}),$$

where

$$\hat{P}(\mathbf{w},\mathbf{x},\mathbf{m}) = \frac{1}{N_v}\sum_{i\in V} I[\mathbf{w_i}=\mathbf{w},\mathbf{x_i}=\mathbf{x},\mathbf{m_i}=\mathbf{m}],$$

$$\hat{P}(\mathbf{x},\mathbf{m}) = \frac{1}{N_v}\sum_{i\in V} I[\mathbf{x_i}=\mathbf{x},\mathbf{m_i}=\mathbf{m}].$$

$I[.]$ is the indicator function, V denotes the validation sample and N_v is the number of households in the validation sample.

Define $\hat{P}_{\mathbf{\Theta}}(\mathbf{w}|\mathbf{m}) = \sum_{\mathbf{x}} P_{\mathbf{\Theta}}(\mathbf{x}|\mathbf{m})\hat{P}(\mathbf{w}|\mathbf{x},\mathbf{m})$.

Then the estimated likelihood function is:

$$\hat{L}(\mathbf{\Theta}) = \prod_{i\in V} P_{\mathbf{\Theta}}(\mathbf{x}_i|\mathbf{m}_i)\prod_{j\in\bar{V}} \hat{P}_{\mathbf{\Theta}}(\mathbf{w}_j|\mathbf{m}_j). \tag{2.2}$$

2.4 Properties of the maximum estimated likelihood estimates

Under regularity conditions, the maximum estimated likelihood estimates $\hat{\mathbf{\Theta}}$ satisfies the score equation $\partial\widehat{\log L}(\mathbf{\Theta})/\partial\mathbf{\Theta} = 0$ and is consistent (Pepe, 1992). If derivatives are available, the Newton-Raphson iteration scheme can be used to find $\hat{\mathbf{\Theta}}$. The estimates of VE_S and VE_I are obtained as $1-\hat{\theta}$ and $1-\hat{\varphi}$, repectively. The properties of $\hat{\mathbf{\Theta}}$ (details of the proof can be found in Pepe, 1992) are:

a. If the validation sample fraction N_v/N has a nonzero limit ρ_v then $n^{\frac{1}{2}}(\hat{\mathbf{\Theta}}-\mathbf{\Theta})$ converges in distribution to a mean zero normal random variable with variance

$$J^{-1}(\mathbf{\Theta}) + \frac{(1-\rho_v)^2}{\rho_v} J^{-1}(\mathbf{\Theta})\kappa(\mathbf{\Theta})J^{-1}(\mathbf{\Theta}),$$

where

$$J(\mathbf{\Theta}) = \rho_v E\{-\frac{\partial^2 \log P_{\mathbf{\Theta}}(\mathbf{x}|\mathbf{m})}{\partial\mathbf{\Theta}^2}\} + (1-\rho_v)E\{-\frac{\partial^2 \log P_{\mathbf{\Theta}}(\mathbf{w}|\mathbf{m})}{\partial\mathbf{\Theta}^2}\},$$

$$\kappa(\boldsymbol{\Theta}) = \text{var}[E\{\frac{D_{\boldsymbol{\Theta}}(\mathbf{w}_{\bar{\mathbf{v}}}|\mathbf{m}_{\bar{\mathbf{v}}})}{P_{\boldsymbol{\Theta}}(\mathbf{w}_{\bar{\mathbf{v}}}|\mathbf{m}_{\bar{\mathbf{v}}})} - \frac{D_{\boldsymbol{\Theta}}(\mathbf{w}|\mathbf{m}_{\bar{\mathbf{v}}})}{P_{\boldsymbol{\Theta}}(\mathbf{w}|\mathbf{m}_{\bar{\mathbf{v}}})}|\mathbf{x}_{\bar{\mathbf{v}}} = \mathbf{x}, \mathbf{m}_{\bar{\mathbf{v}}} = \mathbf{m}, \mathbf{x}, \mathbf{w}\}]$$

$D_{\boldsymbol{\Theta}}(...|...) = \partial P_{\boldsymbol{\Theta}}(...|...)/\partial\boldsymbol{\Theta}$, $\bar{v}$ denotes an arbitrary household in the surrogate sample, and $\mathbf{m}_{\bar{\mathbf{v}}}$ is the vector of vaccination statuses of that household.

b. The estimate $\hat{J}(\boldsymbol{\Theta}) = n^{-1}\partial^2 \log\hat{L}(\boldsymbol{\Theta})/\partial\boldsymbol{\Theta}^2$ is consistent for $J(\boldsymbol{\Theta})$, and

$$\hat{\kappa}(\boldsymbol{\Theta}) = (N_v)^{-1}\sum_{i\in V}\{\hat{Q}_i^{\bar{v}}(\boldsymbol{\Theta})\}\{\hat{Q}_i^{\bar{v}}(\boldsymbol{\Theta})\}^T$$

is consistent for $\kappa(\boldsymbol{\Theta})$ where

$$\begin{aligned}\hat{Q}_i^{\bar{v}}(\boldsymbol{\Theta}) &= \frac{1}{N_v\hat{P}(\mathbf{x_i}, \mathbf{m_i})}\sum_{\mathbf{j}\in\bar{\mathbf{v}}}\{(I[\mathbf{w_j} = \mathbf{w_i}] - \hat{P}(\mathbf{w_j}|\mathbf{x_i}, \mathbf{m_i}))I[\mathbf{m_i} = \mathbf{m_j}]\} \\ &\times \{\frac{D_{\boldsymbol{\Theta}}(\mathbf{x_i}|\mathbf{m_j})}{\hat{P}_{\boldsymbol{\Theta}}(\mathbf{w_j}|\mathbf{m_j})} - \frac{\hat{D}_{\boldsymbol{\Theta}}(\mathbf{x_i}|\mathbf{m_j})}{\hat{P}^2_{\boldsymbol{\Theta}}(\mathbf{w_j}|\mathbf{m_j})}P_{\boldsymbol{\Theta}}(\mathbf{x_i}|\mathbf{m_j})\}.\end{aligned}$$

3. Simulation Results

We conducted a simulation study to investigate the empirical bias and precision of the estimates of θ and φ, and to compare the performance of the parameter estimates with different validation sample sizes and misclassification probabilities. Four estimation methods were used. (1) The full data method, i.e., the ML method that one would use if the true infection outcome could be measured on every study participant. (2) The validation method that uses only the true outcomes in the validation sample. (3) the surrogate method that uses the surrogate outcomes from *all* the N households. (4) The semiparametric method that uses the true and the surrogate data from the validation sample and the surrogate data from the surrogate sample. One expects the first method to produce the most accurate and precise estimates as it uses the true infection outcome for all the households in the study. Obviously, this method cannot be used when the the true outcome is only observed on a subset of households, but we included it in the simulation study for comparisons with the other methods. The second method completely ignores the surrogate outcomes. The third method ignores the true outcomes in the validation sample; this method was included in the simulation study as it is based on the data that would be available if it was impossible to obtain the true outcome on any study participant. The fourth method uses all the available data, hence it is expected to produce estimates that are more accurate than in method 3 and more precise than in method 2.

The input parameters for the simulations are δ, θ, φ, ε_0, and ε_1. δ is the daily transmission probability from an unvaccinated infected person to an unvaccinated susceptible household member. ε_0 and ε_1 are the daily transmission probabilities from the community to an unvaccinated and a vaccinated person, respectively. Note that the simulation program uses the transmission probabilities in one day, and hence they differ from β and γ defined in Section 2.2. The probability of an unvaccinated person becoming infected in one day is $1-(1-\delta)^{x_0} * (1-\delta * \varphi)^{x_1} * (1-\varepsilon_0)$. Here x_0, and x_1 are the numbers of infected unvaccinated and vaccinated persons in the household, respectively, on the previous day. The probability of a vaccinated person becoming infected in one day is $1-(1-\delta * \theta)^{x_0} * (1-\delta * \theta * \varphi)^{x_1} * (1-\varepsilon_1)$. In all the simulations, the length of infectious period was set to one day. Prior to the beginning of the 'outbreak', each individual was 'vaccinated' with a probability of 0.5, independently of all other individuals. Based on the results from our earlier paper (Davis and Haber, 2001), this random vaccination design produces the most precise parameter estimates. For each scenario, we generated 200 simulations and reported the mean parameter estimate and the mean estimated standard error over the 200 simulations.

The true infection outcome was obtained for each study participants in each simulation. We now describe the generation of the surrogate outcomes. For a given individual of vaccination status m, define $P(w|x,m)$ as the probability of surrogate outcome w given infection outcome x. Four probabilities were used to generate the surrogate outcome given one's infectious outcome and vaccination status: $P1 = P(w=1|x=1,m=0)$, $P2 = P(w=1|x=1,m=1)$, $P3 = P(w=1|x=0,m=0)$, and $P4 = P(w=1|x=0,m=1)$. To choose the values for these four probabilities, we first followed the assumption made by Halloran and Longini (2001) for an hypothetical influenza vaccine study. They assumed that every infected person becomes ill, and that an uninfected person may also develop illness symptoms. This implies that $P1 = P2 = 1$, $P3 > 0$, and $P4 > 0$. We then varied the values of $P3$ and $P4$ to explore the effect of the probability that an uninfected person becomes ill on the properties of the estimated parameters. Later we relaxed the assumption $P1 = P2 = 1$ and chose values less than 1.0 for these probabilities.

Fortran programs were used to generate the data and obtain the parameter estimates along with their standard errors. Since the likelihood is very complicated and there is no closed form for the derivatives, we followed conventional ways of obtaining the standard errors from Fortran IMSL routines. The subroutine DB2ONF was used in maximizing the likelihood using a quasi-Newton method and a finite-difference gradient. The Hessian matrix is obtained from this subroutine and then the routine DLINRG was used to compute the information

matrix.

Table 2: Mean $\hat{\theta}$ with mean standard errors for household sizes of 3 and 4. Each set has 200 simulations. Input values: $\delta = 0.6, \varepsilon_0 = 0.008, \varepsilon_1 = 0.002$. Validation sample size=100 households; surrogate sample size=400 households.

		Full		Validation		Surrogate		Semiparametric	
HH size	$P3 = P4^\star$	$\hat{\theta}$	$se(\hat{\theta})$	$\hat{\theta}$	$se(\hat{\theta})$	$\hat{\theta}$	$se(\hat{\theta})$	$\hat{\theta}$	$se(\hat{\theta})$
					$\theta = 0.4$				
3	0.2	0.400	0.061	0.389	0.139	0.342	0.085	0.472	0.075
	0.4	0.400	0.061	0.389	0.139	0.305	0.133	0.493	0.068
	0.6	0.400	0.061	0.389	0.139	0.298	0.246	0.499	0.081
4	0.2	0.400	0.043	0.409	0.103	0.372	0.060	0.424	0.045
	0.4	0.400	0.043	0.409	0.103	0.353	0.085	0.461	0.048
	0.6	0.400	0.043	0.409	0.103	0.332	0.157	0.498	0.053
					$\theta = 0.6$				
3	0.2	0.607	0.070	0.625	0.172	0.528	0.094	0.646	0.076
	0.4	0.607	0.070	0.625	0.172	0.486	0.148	0.665	0.081
	0.6	0.607	0.070	0.625	0.172	0.431	0.230	0.689	0.086
4	0.2	0.597	0.056	0.605	0.131	0.570	0.068	0.592	0.056
	0.4	0.597	0.056	0.605	0.131	0.552	0.096	0.623	0.059
	0.6	0.597	0.056	0.605	0.131	0.539	0.158	0.657	0.068
					$\theta = 0.8$				
3	0.2	0.809	0.085	0.825	0.191	0.729	0.117	0.806	0.091
	0.4	0.809	0.085	0.825	0.191	0.665	0.148	0.822	0.087
	0.6	0.809	0.085	0.825	0.191	0.640	0.241	0.826	0.091
4	0.2	0.794	0.064	0.816	0.158	0.753	0.079	0.779	0.070
	0.4	0.794	0.064	0.816	0.158	0.734	0.105	0.796	0.077
	0.6	0.794	0.064	0.816	0.158	0.728	0.177	0.823	0.079

$\star$ $P3$ and $P4$ are the probabilities that an unvaccinated and a vaccinated person, respectively, develop illness symptoms when they are infected.

3.1 Reduced model — estimating VE_S when $VE_I = 0$

A reduced version of our model for estimating vaccine efficacy can be obtained by assuming that the vaccine affects only susceptibility, i.e., $VE_I = 0$ ($\varphi = 1$). We explored the performance of $\hat{\theta}$ under different scenarios. One would expect the bias of the methods that use surrogate outcomes to depend on the misclassification probabilities $P3$ and $P4$.

Table 3: Mean $\hat{\theta}$ with mean standard errors for a fixed sum of $P3$ and $P4$. Household size=4. Each set has 200 simulations. Input values: $\delta = 0.6, \varepsilon_0 = 0.008, \varepsilon_1 = 0.002$. Validation sample size=100 households; surrogate sample size=400 households.

$P3+P4^{\star}$	$P3/P4$	Full		Validation		Surrogate		Semiparametric	
		$\hat{\theta}$	$se(\hat{\theta})$	$\hat{\theta}$	$se(\hat{\theta})$	$\hat{\theta}$	$se(\hat{\theta})$	$\hat{\theta}$	$se(\hat{\theta})$
					$\theta = 0.4$				
0.2	1	0.400	0.043	0.409	0.103	0.383	0.050	0.409	0.043
	2	0.400	0.043	0.409	0.103	0.396	0.049	0.398	0.042
	4	0.400	0.043	0.409	0.103	0.407	0.049	0.392	0.042
0.4	1	0.400	0.043	0.409	0.103	0.372	0.060	0.424	0.045
	2	0.400	0.043	0.409	0.103	0.394	0.057	0.406	0.045
	4	0.400	0.043	0.409	0.103	0.414	0.057	0.389	0.042
					$\theta = 0.6$				
0.2	1	0.597	0.056	0.605	0.131	0.580	0.060	0.577	0.055
	2	0.597	0.056	0.605	0.131	0.594	0.063	0.569	0.054
	4	0.597	0.056	0.605	0.131	0.605	0.061	0.563	0.051
0.4	1	0.597	0.056	0.605	0.131	0.570	0.068	0.591	0.056
	2	0.597	0.056	0.605	0.131	0.593	0.069	0.572	0.053
	4	0.597	0.056	0.605	0.131	0.612	0.070	0.555	0.052
					$\theta = 0.8$				
0.2	1	0.794	0.064	0.816	0.158	0.770	0.071	0.772	0.070
	2	0.794	0.064	0.816	0.158	0.787	0.071	0.767	0.069
	4	0.794	0.064	0.816	0.158	0.798	0.073	0.762	0.067
0.4	1	0.794	0.064	0.816	0.158	0.753	0.079	0.779	0.070
	2	0.794	0.064	0.816	0.158	0.779	0.082	0.759	0.064
	4	0.794	0.064	0.816	0.158	0.801	0.082	0.746	0.065

$\star$ $P3$ and $P4$ are the probabilities that an unvaccinated and a vaccinated person, respectively, develop illness symptoms when they are infected.

The case $P3 = P4$

Table 2 presents the mean of $\hat{\theta}$ and of its standard error for various input parameter values for household sizes 3 and 4 when $P3 = P4$. The semiparametric method is more robust than the surrogate method and more precise than the validation method. This is more evident for larger values of θ and larger values of $P3 = P4$. We also see that for larger household sizes all four methods perform better (smaller bias and smaller standard error). In order to reduce the impact of the simulation-induced variability, we chose the same seed in all the simulation using a fixed value of θ. Therefore, the results for the full data and the validation

methods, which do not depend on $P3$ and $P4$, are the same for the same value of θ.

Table 4: Mean $\hat{\theta}$ with mean standard errors for various sampling fraction P_v for the validation sample. Household size=4, $P3 = P4^{\star} = 0.4$. Each set has 200 simulations. Input values: $\delta = 0.6, \varepsilon_0 = 0.008, \varepsilon_1 = 0.002$. Total number of households in study=500.

	Full		Validation		Surrogate		Semiparametric	
P_v	$\hat{\theta}$	$se(\hat{\theta})$	$\hat{\theta}$	$se(\hat{\theta})$	$\hat{\theta}$	$se(\hat{\theta})$	$\hat{\theta}$	$se(\hat{\theta})$
				$\theta = 0.4$				
0.2	0.400	0.043	0.409	0.103	0.353	0.085	0.461	0.048
0.4	0.400	0.043	0.405	0.070	0.353	0.085	0.468	0.047
0.6	0.400	0.043	0.402	0.057	0.353	0.085	0.454	0.046
				$\theta = 0.6$				
0.2	0.597	0.056	0.605	0.131	0.552	0.096	0.623	0.059
0.4	0.597	0.056	0.593	0.087	0.552	0.096	0.635	0.057
0.6	0.597	0.056	0.596	0.072	0.552	0.096	0.631	0.056
				$\theta = 0.8$				
0.2	0.794	0.064	0.816	0.158	0.734	0.105	0.796	0.077
0.4	0.794	0.064	0.802	0.105	0.734	0.105	0.813	0.071
0.6	0.794	0.064	0.796	0.085	0.734	0.105	0.810	0.067

$\star$ $P3$ and $P4$ are the probabilities that an unvaccinated and a vaccinated person, respectively, develop illness symptoms when they are infected.

Unequal $P3$ and $P4$

We now consider situations when $P3$ and $P4$ are not equal. One can expect the performance of the methods that use surrogate data to depend on both the magnitude and the ratio of the misclassification probabilities. The ratio is important because if the misclassification probabilities for vaccinated and unvaccinated persons are very different then the ratio of the frequencies of ill persons between vaccinees and nonvaccinees will be a biased estimate of the vaccine effect. To investigate the effect of the ratio on the performance of the estimates we conducted simulations with a fixed value $P3 + P4$ while varying the ratio $P3/P4$. Table 3 presents the results of these simulation for $P3 + P4$=0.2, 0.4, and $P3/P4 = 1, 2, 4$. We can see that the semiparametric method is quite robust even when $P3/P4 = 4$. It is intersting to note that as $P3/P4$ increases, the standard error for the semiparametric method decreases while the standard error from the surrogate method increases.

Different sampling fractions for the validation sample

Let P_v denote the fraction of the validation sample size out of the total number of households in the study. Table 4 lists the simulation results for various sampling fractions with $P3 = P4 = 0.4$ and a total of 500 households of size 4 in the study. We can see that the sampling fraction of the validation sample does not have a significant effect on the bias of the semiparametric estimate of θ. For the standard errors, a slight decrease is observed with increasing the sampling fraction of the validation sample. Thus, it seems that the semiparametric method works as well for a smaller sampling fraction ($P_v = 0.2$) as for a larger sampling fraction.

Table 5: Mean $\hat{\theta}$ and $\hat{\varphi}$ with mean standard errors for $P3 = P4^{\star} = 0.2$. Household size=4; each set has 200 simulations. Input $\delta = 0.6, \varepsilon_0 = 0.008, \varepsilon_1 = 0.002$. Validation sample size=100 households. Surrogate sample size=400 households.

	Full		Validation		Surrogate		Semiparametric	
	mean	se	mean	se	mean	se	mean	se
				$\theta = 0.4$ and $\varphi = 0.4$				
$\hat{\theta}$	0.397	0.053	0.408	0.123	0.412	0.070	0.412	0.055
$\hat{\varphi}$	0.392	0.112	0.367	0.273	0.292	0.115	0.508	0.130
				$\theta = 0.4$ and $\varphi = 0.6$				
$\hat{\theta}$	0.397	0.052	0.410	0.121	0.402	0.071	0.415	0.052
$\hat{\varphi}$	0.589	0.113	0.582	0.262	0.455	0.114	0.674	0.125
				$\theta = 0.6$ and $\varphi = 0.4$				
$\hat{\theta}$	0.599	0.061	0.622	0.148	0.610	0.079	0.617	0.066
$\hat{\varphi}$	0.404	0.083	0.389	0.189	0.318	0.088	0.453	0.099
				$\theta = 0.6$ and $\varphi = 0.6$				
$\hat{\theta}$	0.604	0.062	0.616	0.153	0.617	0.081	0.611	0.065
$\hat{\varphi}$	0.595	0.091	0.603	0.205	0.474	0.091	0.670	0.108

$\star$ $P3$ and $P4$ are the probabilities that an unvaccinated and a vaccinated person, respectively, develop illness symptoms when they are infected.

3.2 The full model — estimating VE_S and VE_I

In this section we drop the assumption $VE_I = 0$ and compare the performance of the four methods with respect to the simultaneous estimation of θ and φ. Table 5 presents the results for the case $P3 = P4 = 0.2$ when 100 out of a total of 500 households of size 4 are included in the validation sample. The estimates of θ produced by the semiparametric method have small bias and standard error.

The semiparametric estimates of φ have a positive bias, but this bias is usually smaller than the (negative) bias of the surrogate method. On the other hand, the standard errors of the semiparametric estimates are only slightly larger than those produced by the surrogate method. Hence, the use of the true outcomes from the validation sample improves the estimation of φ.

So far we have always assumed that $P1 = P2 = 1$, i.e., every person who is infected indeed develops the illness symptoms. We now consider situations where some of the infected persons remained symptom-free (silent infections). Here we report the results for the case $P1 = P2 = 0.9$, $P3 = P4 = 0.2$, $\theta = \varphi = 0.4$, and all the remaining quantities are set to the same values as in Table 5. For the estimation of θ, the surrogate method produced a severely biased estimate of 0.69 while the the bias from each of the other three methods was very small. For the estimation of φ, the estimates produced by the full, validation, surrogate and semiparametric methods were 0.39, 0.37, 0.19 and 0.53, respectively. Thus, the last two methods produces biased estimates. Of these three methods, the semiparametric estimate has the smallest standard error (0.13), compared to 0.18 for the surrogate method and 0.27 for the validation method.

4. Discussion

Estimation of VE_S and VE_I is often complicated by lack of reliable information on exposure to infection and on the true infection outcome. This paper proposes a semiparametric method that uses data from two sample of households: (i) a surrogate sample, where only a surrogate outcome variable (such as illness symptoms) is observed, and (ii) a validation sample where both the true infection outcome and the surrogate outcome are observed. In estimating VE_S when $VE_I = 0$, this semiparametric method performs better than maximum likelihood methods that use the surrogate outcome data only or the true outcome data only. The semiparametric estimates have smaller standard errors than those based on the validation data only and smaller biases than those based on the surrogate data only. This suggests that the proposed method gains efficiency by including the surrogate data and corrects the misclassification bias associated with the surrogate data by including the true outcome data from the validation sample. In estimating VE_S and VE_I simultaneously, the semiparametric method estimates VE_S with very small bias and standard error, but it tends to underestimate VE_I, even though this underestimation is not severe when the true VE_I is small. The bias in estimating VE_I is always larger than in estimating VE_S, even when the true outcome is observed for every study participant (Davis and Haber, 2001). While we fixed the household size in each set of simulations, the estimation methods can be used when households of different sizes are included in the study.

Several studies found estimates of vaccine efficacy (VE_S) to be severely attenuated when surrogate illness outcomes are used instead of the true infection outcomes (Belshe *et al.*, 1998, 2000; Nichol *et al.*, 1999; Longini et al., 2000). In this work we found that the use of household data from a study consisting of a surrogate and a validation sample reduces the bias resulting from the inaccuracy of the surrogate data. For example, Halloran and Longini (2001) used data from a random sample of unrelated individuals and obtained an estimated VE_S of 0.25 when the true VE_S was 0.89. Using the semiparametric method and the study design described in this paper we found that the bias in the estimate of VE_S was usually less than 0.1. In addition, using household data allows simultaneous estimation of both VE_S and VE_I while data on unrelated individuals are not suitable for the estimation of VE_I (Davis and Haber, 2001).

Our simulation study shows that the semiparametric method is quite robust even when the number of households in the validation sample size is quite small (e.g., 20 percent) compared to the total number of households included in the study. We also found that the performance of the semiparametric estimates remains quite stable when the misclassification probabilities for vaccinated and unvaccinated persons are very different.

The semiparametric method proposed in this study extends the method of Pepe (1992) to the case where both the true and the surrogate outcomes are arrays of infection or illness statuses of individuals in the same household. Our simulations show that despite the multivariate nature of the outcome variable, the semiparametric method is very robust when one is interested in estimating VE_S regardless of the value of VE_I. The bias in the estimation of VE_I is not more severe than the bias associated with estimating VE_I when the true infection outcome is known for each individual.

Future studies can look into better ways to correct the bias in estimating VE_I with household data and may add a component in the semiparametric method to correct this underestimation. It is also desirable to explore methods to find the optimal sampling fraction for the semiparametric method proposed for the household data. Finally, one may try to extend this method to cases where the true infection outcome and the surrogate illness outcome are observed for some of the household members while only the illness outcome is observed for other members of the same household. Data of this type was collected in an influenza vaccine trial described in Hurwitz et al (2000).

References

Belshe, R. B., Mendelman, P. M., Treanor, J., *et al.* (1998). The efficacy of live attenuated, cold-adapted, trivalent, intranasal influenza virus vaccine in children. *New England Journal of Medicine* **20**, 1405-1412.

Belshe, R. B., Gruber, W. C., Mendelman, P. M., *et al.* (2000). Efficacy of vaccination with live attenuated, cold-adapted, trivalent, intranasal influenza virus against a variant (A/Sydney) not contained in the vaccine. *Journal of Pediatrics* **136**, 168-175.

Davis, X. M. and Haber, M. (2001). Estimation of vaccine efficacy from household data. *Proceedings of the American Statistical Association, Statistics in Epidemiology Section*, CD-ROM.

Fleming, T. R., Prentice, R. L., Pepe, M.S., and Glidden, D. (1994). Surrogate and auxiliary end-points in clinical-trials, with potential applications in cancer and AIDS research. *Statistics in Medicine* **13**, (9) 955-968.

Golm, G. T., Halloran, M. E., and Longini, I. M. (1998). Semiparametric models for mismeasured exposure information in vaccine trials. *Statistics in Medicine* **17**, 2335-52.

Golm, G. T., Halloran, M. E., and Longini, I. M. (1999). Semiparametric methods for multiple exposure mismeasurement and a bivariate outcome in HIV vaccine trials. *Biometrics* **55**, 94-101.

Haber, M., Longini, I. M. and Halloran, M. E. (1991). Measures of the effects of vaccination in a randomly mixing population. *International Journal of Epidemiology* **20**, 300-310.

Halloran, M. E. and Longini, I. M. (2001). Using validation sets for outcomes and exposure to infection in vaccine field studies. *American Journal of Epidemiology* **154**, 391-398.

Halloran, M. E., Struchiner, C. J. and Longini, I. M. (1997). Study designs for different efficacy and effectiveness aspects of vaccination. *American Journal of Epidemiology* **10**, 789-803.

Hurwitz, E. S., Haber, M., Chang, A., *et al.* (2000). Effectiveness of influenza vaccination of day-care children in reducing influenza-related morbidity among household contacts. *Journal of American Medical Association* **284**, 1677-1682.

Koopman, J. S. and Little, R. J. (1995). Assessing HIV vaccine effects. *American Journal of Epidemiology* **142**, 1113-1120.

Longini, I. M., Datta, S. and Halloran, M. E. (1996). Measuring vaccine efficacy for both susceptibility to infection and reduction in infectiousness for prophylactic HIV-1 vaccines. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology* **13**, 440-447.

Longini, I. M., Halloran, M. E., Nizam, A., *et al.* (2000). Estimation of the efficacy of live, attenuated influenza vaccine from a two-year, multi-center vaccine trial: implications for influenza epidemic control. *Vaccine* **18**, 1902-1909.

Longini, I. M., Koopman, J. S., Haber, M., and Cotsonis, G. A. (1988). Statistical inference for infectious diseases: Risk-specific household and community transmission parameters. *American Journal of Epidemiology* **123**, 845-859.

Nichol, K. L., Mendelman, P. M., Mallon, K. P., *et al.* (1999). Effectiveness of live, attenuated intranasal influenza virus vaccine in healthy working adults: a randomized controlled trial. *Journal of American Medical Association* **282**, 137-144.

Pepe, M. S. (1992). Inference using surrogate outcome data and a validation sample. *Biometrika* **79**, 355-365.

Prentice, R. L. (1989). Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine* **8**, 431-440.

Wittes, J., Lakatos, E., and Probstfield, J. (1989). Surrogate endpoints in clinical trials: cardiovascular disease. *Statistics in Medicine* **8**, 415-425.

Received November 17, 2004; accepted January 24, 2005.

Xiaohong M. Davis
Influenza Branch
National Centers for Disease Control and Prevention
1600 Clifton Road, MS A-32
Atlanta, GA 30341, USA.

Michael Haber
Department of Biostatistics
Rollins School of Public Health
Emory University
1518 Clifton Road NE
Atlanta, GA 30322, USA
mhaber@sph.emory.edu