ESTIMATION OF VACCINE EFFICACY USING THE CASE-EXPOSURE METHOD

Donna Brogan and Alvin Rampey, Jr., Emory University
Mac Otten and Michael Deming, Centers for Disease Control
Donna Brogan, Statistics Department, Emory University, Atlanta GA 30322

I. INTRODUCTION

A recent article by Orenstein et al (1985) discusses the estimation of vaccine efficacy (VE) under different strategies for data collection. The case-exposure strategy (Greenland et al. (1986), Hogue et al. (1983), Hogue et al. (1986), Orenstein (1985)) involves the identification of cases from a surveillance system and the assessment of vaccination coverage from a cluster survey which may also be stratified. However, Orenstein et al (1985) do not discuss the standard error of the estimated vaccine efficacy for the case-exposure method. This paper presents a procedure for estimating the standard error of the estimated vaccine efficacy. The procedure is generalized to cluster surveys involving stratification and to analytical domains (subgroups) of interest. Simplified formulas, amenable to hand or calculator calculation, are available from the authors for a specific cluster survey design which is typical of EPI (Expanded Programme on Immunization) surveys.

II. DATA COLLECTION STRATEGY

A geographic area of interest is defined for study, followed by a vaccination coverage survey to estimate the proportion of children who are vaccinated. For the purpose of generality, this paper assumes a multiple dose vaccine with D doses recommended for full protection. Adequately vaccinated is defined as at least A doses where $0<A\leq D$. The unvaccinated population is defined as receiving no (zero) doses of the vaccine. Survey data include the occurrence and dates of all administered vaccines, allowing the estimation of P_a, the proportion of children who are adequately vaccinated, and P_o, the proportion of children who are unvaccinated. For multiple dose vaccines $P_o \leq 1-P_a$, whereas for single dose vaccines $P_o = 1-P_a$.

The total number of cases of the disease is known for children in the area in the same age range as in the survey. For each case age, sex, residence and occurrence and dates of all administered vaccines are known. The coverage survey should be fielded at the midpoint of the time interval during which cases are counted. Furthermore, the vaccination status of the population is assumed to remain constant during the time period when cases are counted.

The coverage survey can be designed as a complex sample survey which may include stratification, unequal probability of selection, several stages of sampling, clustering, nonresponse adjustments and poststratification. In this case the "ultimate cluster" approximation discussed by Kish (1965) is used wherein the estimation of the standard error recognizes the primary sampling unit (PSU) to which each element (child) belongs but does not take into account the subsequent sampling stages within the PSU to arrive at the sample elements.

The EPI Surveys (Henderson & Sundareson (1982); Lemeshow & Robinson (1985)) are examples of complex sample surveys which include clustering. In these surveys at least 30 clusters (the PSU's) are chosen with probability proportional to size (usually estimated population) from a list of all clusters (villages/towns, cities or parts of cities) which cover the geographic area. Once a sample cluster is chosen, a fixed number of households, women or children are selected within that cluster, where all eligible women and children per household are included in the survey. Generally a starting household is randomly selected from the households in the chosen cluster, and "next-nearest" households are visited by interviewers until the predetermined number of households, women or children are obtained. These procedures should produce an equal probability sample of children in a defined age range and, thus, the statistical analysis for EPI vaccine coverage surveys often is done unweighted.

III. DEFINITION OF ATTACK RATE AND VACCINE EFFICACY

Attack rate (AR) is defined as the number of cases (C) over some time interval divided by the number of children at risk (N) at the midpoint of the time interval, i.e.

$$AR = C/N \qquad (1)$$

The attack rates specific to the adequately vaccinated and unvaccinated populations are denoted, respectively, by AR_a and AR_o and are given as

$$AR_a = C_a/N_a \qquad (2)$$

$$AR_o = C_o/N_o \qquad (3)$$

C_a and C_o are the total number of cases who are adequately vaccinated and unvaccinated, respectively, and N_a and N_o are the total number of chidren in the population who are adequately vaccinated and unvaccinated, respectively.

Vaccine efficacy (VE) is defined in terms of attack rates as follows:

$$VE = \frac{AR_o - AR_a}{AR_o} = 1 - AR_a/AR_o \qquad (4)$$

Note that VE=100% if $AR_a=0$ and that VE=0 if $AR_a=AR_o$.

This paper assumes that the counts of cases (C, C_a, C_o) are known constants; they have no sampling variability since a complete count or enumeration of cases is done. The coverage survey is used to estimate P_a and P_o. Using the known value of N and the estimates of P_a and P_o, the numbers of vaccinated and unvaccinated children at risk are estimated. It is shown later that knowledge of N is necessary for the estimation of attack rates but is not needed for the estimation of vaccine efficacy.

IV. ESTIMATION OF ATTACK RATE AND STANDARD ERROR USING RATIOEST -NO STRATIFICATION

For the attack rates AR_a and AR_o in equations (2) and (3), the numerators C_a and C_o are known constants and the vaccination coverage survey is used to estimate the two denominators N_a and N_o. A coverage survey with no stratification is assumed for now.

Let p_a be the survey estimate of P_a, the proportion of children who are adequately vaccinated. Let p_o be the estimate of P_o, the proportion of children who are unvaccinated. Thus N_a is estimated by Np_a and N_o is estimated by Np_o. The two attack rates then are estimated by:

$$\hat{AR}_a = C_a/Np_a \text{ and } \hat{AR}_o = C_o/Np_o \; . \qquad (5)$$

Obtaining the standard error of the two estimated attack rates in equation (5) is not straightforward since the random variables (the estimated proportions) are in the denominator. Another complicating feature is that the variability of the proportions p_a and p_o cannot be estimated by standard statistical methods for simple random samples since the vaccination coverage surveys are cluster samples or complex sample surveys.

For these reasons it is helpful to write the attack rates in equation (5) as ratio estimators and then use PROC RATIOEST (Shah, 1981) to obtain both the point estimate and its standard error. RATIOEST is a SAS procedure (PROC) which analyzes data from complex sample surveys; it runs in conjunction with SAS. For its recommended use here only the PSU identification for each element is retained, one stratum is assumed (i.e. no stratification), and the finite population correction (fpc) factor is ignored. PROC RATIOEST estimates the standard error of ratio estimators in complex sample surveys by (1) expanding the estimator in an infinite Taylor Series and then ignoring higher order terms and (2) estimating within stratum variability by quantifying the variability between the weighted PSU totals within a stratum.

The estimator p_a is calculated as

$$p_a = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} w_{ij}\, y_{ija}}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} w_{ij}\, x_{ij}} \qquad (6)$$

where

y_{ija} = 1 if child j in PSU i is vaccinated adequately
= 0 otherwise

x_{ij} = 1 for all i, j

w_{ij} = the sampling weight for child j in PSU i

n_i = the number of children in PSU i

k = the number of PSU's in the survey

The weight w_{ij} is defined as the inverse of the probability of selection of child j in PSU i. An interpretation of w_{ij} is that survey child j in PSU i represents w_{ij} children from the inference population. If an equal probability sample is selected and it is not necessary to adjust the survey for nonresponse, a common occurrence in EPI surveys, then w_{ij} is a constant since all children have the same statistical weight.

Using the same notation, p_o is estimated as

$$p_o = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} w_{ij}\, y_{ijo}}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} w_{ij}\, x_{ij}} \qquad (7)$$

where

y_{ijo} = 1 if child j in PSU i has no doses of the vaccine
= 0 otherwise

Substituting (6) and (7) into (5) yields

$$\hat{AR}_a = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} C_a\, w_{ij} x_{ij}/N}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} w_{ij}\, y_{ija}} \qquad (8)$$

and

$$\hat{AR}_o = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} C_o\, w_{ij} x_{ij}/N}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} w_{ij} y_{ijo}} \qquad (9)$$

The point estimate $\hat{AR}_a$ and its standard error can be obtained from PROC RATIOEST by defining $C_a x_{ij}/N$ as the numerator variable, y_{ija} as the denominator variable, and w_{ij} as the sampling weight. The point estimate $\hat{AR}_o$ and its standard error can be obtained from PROC RATIOEST with $C_o x_{ij}/N$ as the numerator variable, y_{ijo} as the denominator variable and w_{ij} as the weight.

The Taylor Series approximation utilized in PROC RATIOEST to estimate the standard error is valid only when the denominator of the ratio estimate has reasonable stability. A common guideline for stability is that the coefficient of variation (CV) of the denominator be small, usually regarded as being less than .10, .15 or .20. Unfortunately, PROC RATIOEST does not include as part of its output the CV of the denominator. The CV can be estimated, however, by using PROC SESUDAAN (Shah, 1981), in the same software package as PROC RATIOEST, to obtain the point estimate and standard error of the denominator variable, either as an estimated mean or estimated total. The estimated CV then is calculated as the ratio of the standard error to

the point estimate. Furthermore, PROC RATIOEST should be used only when the denominator variable is always positive or zero (or always negative or zero). This condition is satisfied for the situations considered in this paper.

V. ESTIMATION OF VACCINE EFFICACY AND STANDARD ERROR USING RATIOEST - NO STRATIFICATION

The vaccine efficacy in equation (4) can be written as

$$VE = 1 - \frac{C_a}{N P_a} \cdot \frac{NP_o}{C_o} = 1 - C_a P_o / C_o P_a \qquad (10)$$

Thus, the estimated vaccine efficacy is obtained by substituting p_a for P_a and p_o for P_o, i.e.

$$\hat{VE} = 1 - \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} w_{ij} C_a y_{ijo} / C_o}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} w_{ij} y_{ija}} \qquad (11)$$

Hence, the point estimate $(1-\hat{VE})$ and its standard error can be obtained from PROC RATIOEST by defining the numerator variable as $C_a y_{ijo}/C_o$, the denominator variable as y_{ija}, and the weight as w_{ij}. Be sure to subtract the PROC RATIOEST point estimate from 1 to obtain $\hat{VE}$. Note also that the standard errors of $(1-\hat{VE})$ and $\hat{VE}$ are equal.

The CV of the denominator should be checked for stability to be sure that the estimated standard errors for vaccine efficacy are valid.

VI. ESTIMATION OF ATTACK RATES AND VACCINE EFFICACY FOR STRATIFIED DESIGNS

A. Notation for stratified sampling

In an extension of sections III. through V. to a stratified design, let the subscript hij denote child j within PSU i within stratum h. The symbols and definitions introduced earlier have the obvious extensions. w_{hij} is the sampling weight for child hij, and x_{hij} is defined to equal 1. Further,

y_{hija} = 1 if child j in PSU i in stratum h is vaccinated adequately
= 0 otherwise.

and

y_{hijo} = 1 if child j in PSU i in stratum h has no doses of vaccine
= 0 otherwise, for $0 \le d \le D$.

Let there be H strata, k_h PSU's within stratum h, and n_{hi} children within PSU i within stratum h. Furthermore, let the total number of cases C be comprised of C_h cases from stratum h, i.e.

$$C = \sum_{h=1}^{H} C_h . \qquad (12)$$

Likewise, let the number of adequately vaccinated and unvaccinated cases in stratum h be denoted by C_{ha} and C_{ho}, respectively. Then

$$C_a = \sum_{h=1}^{H} C_{ha} \quad \text{and} \quad C_o = \sum_{h=1}^{H} C_{ho} . \qquad (13)$$

B. Estimates of Attack Rates for Combined Strata

It is assumed that all H strata are added over, although the formulas presented here also are valid for adding over only S of the strata. The attack rate for the adequately vaccinated, comparable to equation (8) for only one stratum, is given as

$$\hat{AR}_{ast} = \frac{\frac{C_a}{N} \sum_{h=1}^{H} \sum_{i=1}^{k_h} \sum_{j=1}^{n_{hi}} w_{hij} x_{hij}}{\sum_{h=1}^{H} \sum_{i=1}^{k_h} \sum_{j=1}^{n_{hi}} w_{hij} y_{hija}} \qquad (14)$$

Note that the stratum-specific values of C_{ha} and N_h are not used in the estimation of AR_a (except, of course, for the role that N_h plays in determining w_{hij}). Equation (14) can be written as

$$\hat{AR}_{ast} = C_a / N p_{ast} \qquad (15)$$

where p_{ast} is the stratified estimator of P_a, the proportion of children in the entire population of size N who are adequately vaccinated.

PROC RATIOEST can be used to calculate $\hat{AR}_{ast}$ in equation (14) and its estimated standard error by defining $C_a x_{hij}/N$ as the numerator variable, y_{hija} as the denominator variable, and w_{hij} as the weight. Furthermore, every subject must be identified for RATIOEST by stratum number and by PSU number within a given stratum. As always, the CV of the denominator variable needs to be checked for stability.

AR_o can be estimated in a similar manner. Letting $\sum_{hij}$ denote the triple summations indicated in equation (14), the attack rate for the unvaccinated, comparable to equation (9) for an unstratified design, is given as

$$\hat{AR}_{ost} = \frac{\frac{C_o}{N} \sum_{hij} w_{hij} x_{hij}}{\sum_{hij} w_{hij} y_{hijo}} \qquad (16)$$

C. Estimates of Vaccine Efficacy for Combined Strata

The estimated vaccine efficacy, comparable to equation (11) in the unstratified design, is given as

$$\hat{VE}_{st} = 1 - \frac{\frac{C_a}{C_o} \sum_{hij} w_{hij} y_{hijo}}{\sum_{hij} w_{hij} y_{hija}} \qquad (17)$$

As discussed earlier, the point estimate in equation (17), as well as its estimated standard error, can be obtained from PROC RATIOEST with

appropriate definition of the numerator and denominator variables and a check on the stability of the denominator variability.

VII. EXTENSION TO ANALYSIS OF DOMAINS OR SUBGROUPS

A. Subgroups

Many epidemiologists are interested in estimating vaccination coverage, attack rates and vaccine efficacy for subgroups of the population. Epidemiologists often use the term strata for these subgroups. In this paper the term stratum is used to designate part of the sample design; strata typically are defined on a geographical basis. A common term used by samplers for subgroups of interest in the analysis is domains, although this paper uses the term subgroup.

When the subgroups are geographically defined, they may correspond with strata which are based on geography. In this instance, the subgroup analyses are done by combining strata as discussed in section VI.

Most frequently, however, the subgroups are based on demographic variables such as age, sex and/or race and, thus, the subgroups or domains span across all strata in the survey. This section discusses analysis for subgroups which occur within a stratum (for an unstratified design) or across all or several strata (for a stratified design).

B. Unstratified Design

The analysis of subgroups is handled by the use of indicator variables. The variables used earlier, i.e. y_{ija}, x_{ij} and y_{ijo} are redefined as follows:

$y_{sija} = y_{ija}$ if child j in PSU i belongs to subgroup s
$= 0$ otherwise

$x_{sij} = x_{ij}$ if child j in PSU i belongs to subgroup s
$= 0$ otherwise (18)

$y_{sijo} = y_{ijo}$ if child j in PSU i belongs to subgroup s
$= 0$ otherwise (19)

All of the procedures in Sections IV. and V. for estimating attack rates and vaccine efficacy can be applied by substituting y_{sija}, x_{sij} and y_{sijo} for y_{ija}, x_{ij}, and y_{ijo}, respectively.

In addition, the terms C_a and C_o must be replaced by the terms C_{sa} and C_{so}, defined respectively as the number of adequately vaccinated and unvaccinated cases within subgroup s.

PROC RATIOEST can be used as earlier by defining the appropriate variables y_{sija}, y_{sijo} and x_{sij} and multiplying by the appropriate constants C_{sa} and C_{so}. Alternatively, PROC RATIOEST can be instructed to calculate estimates for mutually exclusive and exhaustive subgroups. If this latter approach is used, the specific constants for each subgroup need to be multiplied by the PROC RATIOEST output since PROC RATIOEST cannot be programmed to include different constants across the subgroups.

C. Stratified Designs

When analogous indicator variables are defined for stratified designs, the methodology of section VI. can be used. The new variables are:

$y_{shija} = y_{hija}$ if child j within PSU i within stratum h belongs to subgroup s
$= 0$ otherwise

$x_{shij} = x_{hij}$ if child j within PSU i within stratum h belongs to subgroup s
$= 0$ otherwise

$y_{shijo} = y_{hijo}$ if child j within PSU i within stratum h belongs to subgroup s
$= 0$ otherwise

y_{shija}, x_{shij} and y_{shijo} are substituted for y_{hija}, x_{hij} and y_{hijo}, respectively, in section VI. Furthermore, the quantities C_a and C_o in section VI are replaced by C_{sa} and C_{so}, respectively, where these latter quantities are the adequately vaccinated and unvaccinated cases within subgroup s.

As discussed in section VII-B, PROC RATIOEST can be used in one of two ways to calculate attack rates and/or vaccine efficacies for subgroups. Recall that each child needs to be identified by stratum and by PSU. It is especially important to check the CV of the denominator variable since a small sample size in a subgroup may make the Taylor Series approximation invalid.

VIII. DISCUSSION

The sampling techniques and statistical analysis procedures for EPI vaccination coverage surveys are based on work by Serfling and Sherman (1965) in designing public health surveys in the U.S. The adaptation of this methodology to vaccination coverage surveys in Africa is detailed in a WHO manual "Evaluating Vaccination Coverage", although no indication is given of how to calculate standard errors for the estimated proportions. A review of the EPI survey methodology is presented in Henderson et al (1973), Henderson & Sundaresan (1982), Lemeshow et al. (1985) and Lemeshow & Robinson (1985), with only the latter article indicating a formula for the standard error of the estimated proportion (vaccination coverage) when the specific design is an equal number of children per sample cluster.

As epidemiologists have become more sophisticated about sample surveys, these EPI surveys have expanded in scope beyond the initial objective of estimating vaccination coverage. Some common newer objectives are estimation of child mortality, fertility, and health care practices and utilization. The statistical analysis, including point estimates and estimated standard errors, generally is more complicated in these recent surveys with expanded objectives. There does not appear to be a systematic treatment of estimation strategies for these newer objectives.

This paper has considered one particular expanded objective of EPI surveys, the estimation of attack rates and vaccine efficacy via the case-exposure method. The case data are assumed to constitute an enumeration of the total population of cases, and no variability is

assumed for the case data. Formulas for point estimates and estimated standard errors are given which should meet most analysis needs for these surveys.

One basic premise for the development of these formulas is the use of ratio estimators and the Taylor Series expansion of the estimator in order to obtain the estimated variance. This paper discusses variance of the estimator, rather than mean square error, because the sample size is assumed large enough so that the bias of the ratio estimator is negligible.

These formulas should be useful to epidemiologists who wish to estimate attack rates and vaccine efficacy from an EPI coverage survey and complete count data on cases of the disease.

REFERENCES

Greenland, Sander, D.C. Thomas, H. Morgenstern (1986). "The Rare-Disease Assumption Revisited: A Critique of Estimators of Relative Risk for Case-Control Studies", American Journal of Epidemiology, 124(6), 869-876.

Henderson, R.H., and T. Sundareson (1982). "Cluster Sampling to Assess Immunization Coverage: A Review of Experience with a Simplified Sampling Method," Bulletin of the World Health Organization, 60, 253-260.

Henderson, R.H., H. Davis, D.L. Eddins and W.H. Foege (1973). "Assessment of Vaccination Coverage, Vaccination Scar Rates and Smallpox Scarring in Five Areas of West Africa," Bulletin of the World Health Organization, 48, 183-194.

Hogue, Carol J.R., D.W. Gaylor, K.F. Schulz (1983). "Estimators of Relative Risk for Case-Control Studies," American Journal of Epidemiology, 118(3), 396-407.

Hogue, Carol J.R., D.W. Gaylor, K.F. Schulz (1986). "The Case-Exposure Study: A Further Explication and Response to a Critique," American Journal of Epidemiology.

Lemeshow, S., A.G. Tserkovnyi, J.L. Tulloch, J.E. Dowd, S.K. Lwanga and J. Keja (1985). "A Computer Simulation of the EPI Survey Strategy," International Journal of Epidemiology, 14(3), 473-481.

Lemeshow, S. and D. Robinson (1985). "Surveys to Measure Programme Coverage and Impact: A Review of the Methodology Used by the Expanded Programme on Immunization," World Health Statistical Quarterly, 38, 65-75.

Orenstein, Walter A., R.H. Bernier, T.J. Dondero,, A.R. Hinman, J.S. Marks, K.J. Bart, B. Sirotkin (1985). "Field Evaluation of Vaccine Efficacy," Bulletin of the World Health Organization, 63(6), 1055-1068.

Shah, B.V. (1981). SESUDAAN: Standard Errors Program for Computing of Standardized Rates From Sample Survey Data. Research Triangle Institute, Research Triangle Park, N.C.