Published in final edited form as:
J Clin Epidemiol. 2007 November ; 60(11): 1127–1131.

Methodological Issues in Design and Analysis of a Matched Case-Control Study of a Vaccine's Effectiveness

Linda M. Niccolai, Ph.D.[*], **Lorraine G. Ogden, Ph.D.**[$], **Catherine E. Muehlenbein, M.P.H.**[#], **James D. Dziura, Ph.D.**[#,+], **Marietta Vázquez, M.D.**[#], and **Eugene D. Shapiro, M.D.**[*,#,+]

[*]*Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT*

[#]*Department of Pediatrics, Yale University School of Medicine, New Haven, CT*

[+]*Department of General Clinical Research Center, Yale University School of Medicine, New Haven, CT*

[$]*Department of Preventive Medicine and Biometrics, University of Colorado School of Medicine and Health Sciences Center, Denver, CO*

Abstract

Objective—Case-control studies of the effectiveness of a vaccine are useful to answer important questions, such as the effectiveness of a vaccine over time, that usually are not addressed by pre-licensure clinical trials of the vaccine's efficacy. This report describes methodological issues related to design and analysis that were used to determine the effects of time since vaccination and age at the time of vaccination.

Study Design and Setting—A matched case-control study of the effectiveness of varicella vaccine.

Results—Sampling procedures and conditional logistic regression models including interaction terms are described.

Conclusion—Use of these methods will allow investigators to assess the effects of a wide range of variables, such as time since vaccination and age at the time of vaccination, on the effectiveness of a vaccine.

Keywords

case-control study; vaccines; statistical methods; sampling

What is new?

- Matched case-control studies are a statistically powerful method of assessing a vaccine's effectiveness as it is used in the field, particularly to address questions about effects of time since vaccination and of age at the time of vaccination.
- We describe the sampling plan (risk set sampling) and the statistical methods (conditional logistic regression models with interaction terms) that can be used by investigators to answer multiple questions about a vaccine's effectiveness.

Address for correspondence: Eugene D. Shapiro MD, Yale Department of Pediatrics, 333 Cedar Street, P.O. Box 208064, New Haven, CT 06520-8064, Eugene.Shapiro@yale.edu, Phone: 203-688-4555, Fax: 203-785-3932.

- Future studies of a vaccine's effectiveness should incorporate these methodological considerations to maximize the relevance of findings to address clinical issues and to inform future research.

For a new vaccine to be approved for use by the Food and Drug Administration, it is virtually always necessary to conduct a randomized clinical trial to document its protective efficacy. Although randomized trials are the gold standard for scientific validity, they do have a number of limitations. [1] To study the efficacy of a vaccine over time, a randomized trial would require extensive longitudinal follow-up of a large number of subjects, which would be both logistically difficult to conduct and very expensive. Consequently, pre-licensure trials generally are conducted for a relatively short period of time and often do not answer questions such as whether the vaccine's efficacy wanes over time or whether the vaccine's efficacy differs in specific subgroups of patients. Case-control studies are a statistically powerful alternative method to answer such questions once a vaccine is approved and is in general use. [1-4] We conducted a matched case-control study of the effectiveness of varicella vaccine to address two major questions: 1) Does the vaccine's effectiveness wane over time? and 2) Does age at time of vaccination affect the vaccine's effectiveness? [5-6]

The methods and results of this study have been described previously. [6] Briefly, case subjects, identified by active surveillance of 20 pediatric practices in southern Connecticut, consisted of 339 eligible children clinically diagnosed with chickenpox who also had a positive test result by polymerase chain reaction assay. Two controls matched by both date of birth (±1 month) and pediatric practice were selected for each case subject. Medical records of all subjects were reviewed to determine when vaccinations were administered and to obtain information on covariates. Results indicated that overall the vaccine's effectiveness was 87% (95% confidence interval [CI]: 81%-91%). There was a significant effect of time since vaccination; effectiveness was 97% in the first year after vaccination compared with 84% in subsequent years ($P<0.01$). The effectiveness one year after vaccination was significantly lower among children vaccinated at <15 months of age compared with children vaccinated at ≥15 months of age (73% vs. 99%, $P=0.01$).

A matched case-control study can be used to control for the effects of potential confounding variables in the study design. The major advantages of matching are that it may be both cost-effective and efficient under the right set of circumstances. Another advantage of matching on potential confounding variables, rather than adjusting for potential confounders during analysis, is that the functional form for the relationship between the confounding variable and the outcome need not be specified, and thus matching provides a superior method to control confounding. However, there are several potential disadvantages of a matched case-control design that must also be considered. First, it may be difficult to obtain matched controls for all cases. Second, the effect of the matching variables on disease cannot be estimated. Finally, special methods (described below) are required to analyze data from matched designs.

The purpose of this report is to describe the sampling plan - risk set sampling - and the statistical methods - conditional logistic regression models with interaction terms - to answer research questions that have important clinical implications in a matched case-control study of the vaccine's effectiveness.

Design Issue: Risk Set Sampling

Appropriate selection of controls in case-control studies is essential to produce valid, unbiased estimates of risk. Ideally, controls will be a sample of individuals from the source population that gave rise to the cases. In other words, controls are individuals who, had they developed the disease of interest, would have been a case in the study. When controls are selected from a list at a single point in time, that is, without regard for the amount of time these individuals

were at risk for developing the disease, then parameter estimation of the odds ratio approximates relative risk only for rare diseases. However, if risk set sampling is used, then the odds ratio approximates relative risk for outcomes of any frequency. Risk set sampling is the process by which controls are chosen from the set of individuals in the source population who are at risk for the disease (i.e., in the risk set) at the time the case was diagnosed. [7] In this way, the element of time at risk is accounted for because cases are matched to controls with respect to sampling time and the odds ratio is a valid estimate of relative risk. Risk set sampling was employed in this study because controls were selected from lists of active patients at the pediatric practice at the time the case was diagnosed. Using risk set sampling means that a control can later become a case, and controls may be included more than once because if these individuals remain in the risk set, then they should be eligible subsequently for selection. Indeed, in this study, on several occasions controls later did become cases. Though this may seem a subtle procedural matter in recruitment and eligibility of subjects, it is critical for valid estimation of risk in case-control studies.

Statistical Analysis: Conditional Logistic Regression Using Interaction Terms

Overall Effectiveness of the Vaccine

In matched case-control studies, the log odds of disease for individual *j* in matched set *i* can be modeled using a linear logistic model as shown in Equation (1), which includes a stratum-specific constant term, α_i, for each of the matched sets (i.e., strata) and a beta parameter coefficient for each of the *k* variables of primary interest:

$$\log \text{odds}\left(\text{varicella}_{ij}\right) = \alpha_i + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \ldots + \beta_k x_{kij}. \quad (1)$$

To determine the overall effectiveness of a vaccine in a matched case-control study, the model of interest will estimate the log odds ratio (OR) of being a case (having varicella) for those who were vaccinated compared with those who were not vaccinated. This can be achieved by creating a variable vaccine that is equal to 1 if the subject is vaccinated anytime (or more than a specified time—e.g., 2 weeks) before focal time—the time of diagnosis for the case subject, and 0 if the subject is not vaccinated before this time. The model shown in Equation (2) can be used to estimate the overall effectiveness of the vaccine.

$$\log \text{odds}\,(\text{varicella}) = \alpha_i + \beta_1\,(\text{vaccine}) \quad (2)$$

where β_1 is the log OR for children who were vaccinated compared with children who were not vaccinated. Using Equation (3), one may obtain the OR comparing these two groups.

$$\text{OR} = \exp\left(\beta_1\right) \quad (3)$$

The vaccine's effectiveness (VE) expressed as a percentage may be estimated by Equation (4) where the OR is used as an estimate of the relative risk (RR).

$$\text{VE} = [1 - \text{RR}] * 100 \quad (4)$$

In the study, the model in Equation (2) produced a parameter estimate of $ß_1 = -2.0$. Therefore, we estimated the OR for being vaccinated, comparing cases with controls, to be 0.13, corresponding to an overall effectiveness of: [1 - 0.13] * 100 = 87%. This means that the odds of being a case are 87% lower for vaccinated children compared with children of the same age and pediatric practice who were not vaccinated, in other words, an 87% reduction in the risk of infection. These and all subsequent analyses may be adjusted for other potential confounding variables by including the necessary parameters in the model. In the study, terms for sex, race, attendance group day care, asthma, use of a corticosteroid, and receipt of varicella vaccine within 28 days after receiving measles-mumps-rubella vaccine were included in all models for this purpose.

Effect of Time since Vaccination on the Vaccine's Effectiveness

In addition to knowing its overall effectiveness, investigators may also want to know the effect of time since vaccination to assess whether effectiveness wanes over time. Because Equation (2) does not include a parameter for the effect of time since vaccination, it cannot be used to assess this potential effect. The variable vaccine includes both the effect of vaccination and the effect of time since vaccination in a single parameter, so these effects cannot be separated. However, this may be an important effect to estimate because if effectiveness declines at a certain rate or falls below a specified level after a certain period of time, then a booster dose of the vaccine may be recommended.

To estimate the effect of time since vaccination on the effectiveness of a vaccine, additional variables need to be created for fitting another model. A term for each possible year during which vaccinated children could have developed disease was created. These terms equal 1 for subjects whose status was vaccinated in a given year prior to focal time (the time at which the case subject developed varicella) and 0 for everybody else, including all unvaccinated study participants. For this model, 7 new terms were created because there were up to 7 years of follow-up after a child could have been vaccinated. Y1 was equal to 1 for subjects who had been vaccinated <1 year before focal time and it was equal to 0 for everyone else (subjects who were vaccinated ≥1 year prior to focal time and all unvaccinated subjects); Y2 was equal to 1 for subjects who were vaccinated between 1 and 2 years before focal time and was equal to 0 for everyone else; Y3 = 1 for subjects who were vaccinated between 2 and 3 years before focal time and was equal to 0 for everyone else, etc. up through Y7. The model in Equation (5) that includes all terms Y1 through Y7 produces estimates for the log OR of being a case for children who were vaccinated in different years prior to focal time compared with the unvaccinated.

$$\log \text{odds}(\text{varicella}) = \alpha_i + \beta_1(Y1) + \beta_2(Y2) + \beta_3(Y3) + \beta_4(Y4) + \beta_5(Y5) + \beta_6(Y6) + \beta_7(Y7) \quad (5)$$

In this model, the referent group is unvaccinated children because this group is not parameterized. Each parameter estimate can be exponentiated using Equation (3) to produce estimates of the vaccine's effectiveness in each year after vaccination. In this study, the parameter estimates were ß$_1$= -3.6, ß$_2$= -2.0, ß$_3$= -1.8, ß$_4$= -1.7, ß$_5$= -1.8, ß$_6$= -1.6, and ß$_7$= -1.7, corresponding odds ratios of 0.03, 0.14, 0.17, 0.19, 0.16, 0.19, and 0.19, respectively. These translate to an effectiveness of 97% 1 year after vaccination, 86% 2 years after vaccination, 83% 3 years after vaccination, 81% 4 years after vaccination, 84% 5 years after vaccination, 81% 6 years after vaccination, and 81% 7 years after vaccination.

In addition to determining the vaccine's effectiveness in each year after vaccination, investigators may also want to determine whether there is a significant change in the vaccine's effectiveness over time. To do this, another model was created, shown in Equation (6), that includes the vaccine variable and the Yi variables with the term for year one excluded to provide the referent category.

$$\log \text{odds}(\text{varicella}) = \alpha_i + \beta_1(\text{vaccine}) + \beta_2(Y2) + \beta_3(Y3) + \beta_4(Y4) + \beta_5(Y5) + \beta_6(Y6) + \beta_7(Y7) \quad (6)$$

The parameter β_1 represents the log OR for children vaccinated within one year of focal time relative to unvaccinated children and is equivalent to the parameter β_1 from Equation 5. The parameters ß $_2$ through ß $_7$ now represent the effect of vaccination given in years two through seven before focal time relative to those given vaccine within one year of focal time. In this model, statistically significant coefficients for ß $_2$ through ß $_7$ indicate that the vaccine's effectiveness in the first year after vaccination is significantly different from its effectiveness in each of the subsequent years. For example, in the study, the coefficient ß $_2$= 1.6 and P=0.01. The odds ratio calculated using Equation 3 was 5.1. Thus we concluded that there was a 5.1 fold increase in the odds of varicella among children two years after vaccination compared

with the odds of varicella among children one year after vaccination, and the difference in the effectiveness of the vaccine in years 1 and 2 after vaccination was statistically significant. One can see that β_2 from Equation (6) could have also been calculated from the parameters in Equation (5) as the difference between the beta coefficients for Y1 and Y2, that is, $\beta_2(5) = \beta_2(4) - \beta_1(4) = -2.0 - (-3.6) = 1.6$. Calculation of its standard error and P-value are slightly more complicated, and thus fitting the model from Equation (6) is a convenient way of obtaining this estimate directly. It is important to note here that even though the question of interest is to compare relative effectiveness of vaccination at different points in time, it is not possible to do this by excluding unvaccinated children from the analysis. Because this is a matched study, dropping an unvaccinated case would result in the loss of the controls from that stratum, and dropping an unvaccinated control would result in the loss of the other two individuals (a case and a control) from that matched stratum.

Similar models can be designed to compare this effect in any year with the effects in other years by excluding terms for different years. For example, the model in Equation 7 uses 2 years after vaccination as the referent group and evaluates whether the vaccine's effectiveness in year 1 and in years 3 through 7 were significantly different from its effectiveness in year 2.

$$\text{log odds}(\text{varicella}) = \alpha_i + \beta_1(\text{Vaccine}) + \beta_2(\text{Y1}) + \beta_3(\text{Y3}) + \beta_4(\text{Y4}) + \beta_5(\text{Y5}) + \beta_6(\text{Y6}) + \beta_7(\text{Y7}) \quad (7)$$

In this instance, if the coefficients β_3 through β_7 are not statistically significant, then the vaccine effectiveness in years 3 through 7 is not significantly different from its effectiveness in year 2.

Such analyses allow investigators to identify whether and at what point a vaccine's effectiveness begins to wane and/or whether waning effectiveness subsequently stabilizes. The choice to collapse levels of a categorical variable is often desirable for multiple reasons. Creating dichotomous measures increases the statistical power of the analysis because the data are not spread sparsely over many strata, and this also may make it easier to interpret the results. Furthermore, it simplifies the analysis when interaction terms need to be created. The choice of which levels to combine (e.g., what cutpoint to use for dichotomizing) may not always be straightforward. This choice may be done *a priori* in which a theoretical or practical rationale is provided, or this may be done post-hoc in which the data themselves suggest a meaningful cutpoint. This study provides examples of both as described below.

To more efficiently model time since vaccination and to improve interpretability of our final model, we used a data-driven approach for determining the cutpoint to dichotomize this variable. We observed that the vaccine's effectiveness decreased significantly between years 1 and 2 after vaccination, and that this was the only statistically significant decline in effectiveness through year 7 after vaccination. When Y2 was excluded from the model as in Equation (7), P-values for Y3 through Y7 were all substantially greater than 0.05, indicating no statistically significant difference in the vaccine's effectiveness between year 1 and years 2 through 7. Consequently, years 2 through 7 were grouped together for subsequent analyses to estimate the effects of time since vaccination. The variable Y2toY7 was created, which was equal to 1 if the subject was vaccinated 2 through 7 years before focal time and 0 if the subject either was vaccinated 1 year before focal time or was unvaccinated. Therefore, the following model was developed to produce the summary estimates:

$$\text{log odds}(\text{varicella}) = \alpha_i + \beta_1(\text{vaccine}) + \beta_2(\text{Y2toY7}) \quad (8)$$

β_1 is interpreted as the effect of vaccine during the first year after vaccination relative to unvaccinated children (the effect of the vaccine when Y2toY7= 0) and β_2 is interpreted as the change in the effect of vaccine after 1 year (when Y2toY7= 1) compared with the effect of vaccine within the first year.

Effect of Age at the Time of Vaccination on the Vaccine's Effectiveness

The above analyses describe the main effects of vaccination and of time since vaccination on the effectiveness of the vaccine. However, investigators may also be interested in the effects of other variables, and the effects of interactions also should be considered when appropriate. We were interested in the effect of age at the time of vaccination because we hypothesized that the vaccine may have diminished effectiveness if administered at <15 months of age because such effects have been observed with other live-attenuated vaccines, such as measles vaccine, which are administered parenterally. To assess this possibility, a new variable, age, was created. Age was equal to 1 if the subject had been vaccinated at ≥15 months of age, and 0 if the subject either had been vaccinated at less than 15 months of age or was unvaccinated. In this case, the choice of dichotomizing at age <15 months was made a priori and was based on both practical issues and the results of previous research. Children typically have 15-month well baby visits, and other vaccines are also recommended at this time. Therefore, this time point is inherently of interest because if important differences were found, clinically relevant recommendations could be made. Furthermore, recent reports of outbreaks of varicella in which young age at the time of vaccination was assessed as a potential risk factor for vaccine failure have not used a consistent cutpoint for younger age at vaccination (range <14 to <18 months [9-12]); our choice is consistent with and in the middle of the range of ages used by others.

The final model, shown in Equation (9), also evaluated interactions between time since vaccination and age at vaccination:

$$\log \text{odds}(\text{case}) = \alpha_i + \beta_1(\text{vaccine}) + \beta_2(\text{Y2toY7}) + \beta_3(\text{age}) + \beta_4\left(\text{Y2toY7}^{*}\text{age}\right) \quad (9)$$

In this model, if the β_4 parameter is statistically significant, it suggests that the waning effect after one year is different between the two different categories of age at vaccination, and the analyses should be stratified by age at vaccination (and/or by time since vaccination). In our study, the coefficient β_4 was significantly different from zero (P=.02), so it was concluded that there was a significant interaction between age at vaccination and time since vaccination. Therefore, the effect of age at vaccination should be interpreted with respect to time since vaccination and the effect of time since vaccination should be interpreted with respect to age at vaccination. This model allows for the calculation of the vaccine's effectiveness over time separately for the two age groups, including estimates of the parameters for the 4 groups determined by all possible covariate patterns (vaccinated at <15 months of age and within 1 year of focal time, vaccinated at <15 months of age and within 2-7 years of focal time, vaccinated at ≥15 months of age and within 1 year of focal time, and vaccinated ≥15 months of age and within 2-7 years of focal time) compared with unvaccinated subjects. For example, output for the model in (9) produced the following estimates: β_1= -1.3, β_2= -0.3, β_3= -3.3, β_4= 3.1. The β_1 parameter represents the vaccine's effectiveness for the first year after vaccination among those vaccinated at <15 months of age (the effect when time = 0 and age = 0). The β_2 parameter represents the effect of time among those vaccinated at <15 months of age and the β_3 parameter represents the effect of age at vaccination among those vaccinated less than 1 year earlier. Therefore, the vaccine's effectiveness for the first year after vaccination among those vaccinated at <15 months of age was 73%, calculated as 1-exp(-1.3)*100 according to the formulas presented in Table 1. The vaccine's effectiveness for the first year after vaccination among those vaccinated at ≥15 months of age was 99%, calculated as 1-exp (-1.3+-3.3)*100. Thus, the β_3 parameter represents the difference in the vaccine's effectiveness for the first year after vaccination for those vaccinated <15 months of age compared to those vaccinated at ≥ 15 months of age, and the P-value for this parameter can be used to assess this difference, 73% vs. 99%. Similar methodology can be used to compare the vaccine's effectiveness between any two of the four groups defined by age at vaccination and time since vaccination.

It should also be noted that another way to parameterize this model without using terms for main effects and their interactions would be to use an indicator variable for each of the four subgroups previously defined with the referent group being the unvaccinated subjects. This model would produce the same results, and it is straightforward to code the terms and interpret the results for dichotomous variables.

Discussion

Case-control studies are a statistically powerful method of assessing a vaccine's effectiveness as it is used in the field. As more new vaccines are approved for use, it is likely that such post-licensure studies will be of increasing importance to assess the performance of vaccines. Risk set sampling and use of conditional logistic regression models similar to those described above will allow investigators to assess the effects of a wide range of clinically important variables, such as time since vaccination and age at the time of vaccination, on the effectiveness of the vaccine.

Acknowledgements

Supported in part by grants # AI41608, AI01703, RR022477, M01-RR00125, and AI068280 from the National Institutes of Health and by a Robert Wood Johnson Minority Medical Faculty Development Award (Dr. Vazquez).

References

1. Clemens JD, Shapiro ED. The pneumococcal vaccine controversy: Are there alternatives to randomized clinical trials. Rev Infect Dis 1984;6:589–600. [PubMed: 6390636]
2. Shapiro ED. Case-control studies of the effectiveness of vaccines: Validity and assessment of bias. Pediatr Infect Dis J 2004;23:127–31. [PubMed: 14872178]
3. Breslow, NE.; Day, NE. Statistical methods in cancer research: Vol. 1. The analysis of case-control studies. World Health Organization; Lyon: 1980.
4. Schlesselman, JJ. Case-control studies: Design, conduct, analysis. Oxford University Press; New York: 1982.
5. Vázquez M, LaRussa PS, Gershon AA, Steinberg SP, Freudigman K, Shapiro ED. Effectiveness of Varicella Vaccine in Clinical Practice. N Engl J Med 2001;344:955–60. [PubMed: 11274621]
6. Vázquez M, LaRussa PS, Gershon AA, Niccolai LM, Muehlenbein CE, Steinberg SP, et al. Effectiveness of varicella vaccine over time. JAMA 2004;291:851–5. [PubMed: 14970064]
7. Rothman, KJ.; Greenland, S. Modern Epidemiology. second edition. Lippincott-Raven Publishers; Philadelphia PA: 1998.
8. Breslow, NE.; Day, NE. Statistical Methods in Cancer Research: The Analysis of Case-Control Studies. Oxford University Press; 1993.
9. Tugwell BD, Lee LE, Gilette H, et al. Chickpox outbreak in a highly vaccinated school population. Pediatrics 2004;113:455–459. [PubMed: 14993534]
10. Galil K, Lee B, Strine T, et al. Outbreak of varicella at a daycare center despite vaccination. NEJM 2002;347:1909–1915. [PubMed: 12477940]
11. Galil K, Fair E, Mountcastle N, et al. Younger age at vaccination may increase risk of varicella vaccine failure. JID 2002;186:102–105. [PubMed: 12089668]
12. Haddad MB, Hill MB, Pavia AT, et al. Vaccine effectiveness during a varicella outbreak among schoolchildren: Utah, 2002-2003. Pediatrics 2005;115:1488–1493. [PubMed: 15930208]

Table 1

Coding structure and covariate pattern, including interaction term, for Equation (9).

COVARIATE PATTERN	VACCINE TERM	TIME TERM	AGE TERM	TIME×AGE TERM	LOG OR BASED ON MODEL 8
Not vaccinated	0	0	0	0	(referent)
Younger age at vaccination and vaccinated one year prior to disease	1	0	0	0	$\exp(ß_1)$
Younger age at vaccination and vaccinated 2-7 years prior to disease	1	1	0	0	$\exp(ß_1 + ß_2)$
Older age at vaccination and vaccinated one year prior to disease	1	0	1	0	$\exp(ß_1 + ß_3)$
Older age at vaccination and vaccinated 2-7 years prior to disease	1	1	1	1	$\exp(ß_1 + ß_2 + ß_3 + ß_4)$