Clinical Trials: Vaccine Development

Ivan S. F. Chan, William W. B. Wang, and Joseph Heyse
Merck Research Laboratories, West Point, Pennsylvania, U.S.A.

Abstract
This entry reviews some of the special considerations in vaccine development, including an increased interest in surrogate markers of activity, a special need for assessing immune responses, and a need for more specificity in endpoint definitions of clinical efficacy based on exposure.

INTRODUCTION
Clinical trials designed to demonstrate the safety and efficacy of new vaccines have a rich history. In 1954, the largest medical experiment in history tested a vaccine to prevent poliomyelitis, one of the most feared childhood diseases. This was the Salk vaccine, named after Dr. Jonas Salk. Close to 2 million children across the United States and Canada participated in this field trial. The vaccine became one of a group of required vaccinations for all children in the United States. Meier[1,2] discussed the key statistical considerations of that landmark trial. In many ways, the development of clinical trial methods for vaccines has paralleled rather than coincided with the development of trial methods for new drugs and other interventions. Even in the general area of primary prevention research, vaccine trials occupy a distinct place.

Vaccines are biological products that work primarily by introducing antigen or attenuated live virus into the body. Vaccination triggers an immune response in the form of antibodies and T-cell immunity that are specific for the target infectious agent. The presence of antibodies and T-cell immunity plays an important role in limiting the spread of most pathogens and preventing disease infection. Immunologic memory of immune responses enables the body to rapidly produce a large amount of antigen-specific antibodies and T-cells when the infectious agent is detected. The objective of vaccination is to educate the immune system by inducing the immune memory in the absence of the unpleasant clinical features associated with the target infection. Booster doses of vaccines are sometimes administered after several years to ensure that the immune memory is maintained.

Vaccines are used for the prevention of disease and play a critical role in public health policy. Widespread vaccination in a population not only protects vaccinees who are exposed to the infectious agent, but can change the exposure to infection for people who are not vaccinated. Because vaccines are developed for administration to millions of healthy subjects, often children, there is a premium on safety. A comprehensive evaluation is made to assure that the benefits of vaccination outweigh the risks. Vaccine review at the United States Food and Drug Administration (FDA) is undertaken by the Center for Biologics Evaluation and Research (CBER). The Centers for Disease Control and Prevention (CDC) are also integrally involved in establishing vaccination policy.

OVERVIEW
Clinical trials in vaccine development generally have four phases.[3,4] Phase I trials are early, small-scale human studies that explore the safety and immunogenicity of multiple dosage levels of a new vaccine. Phase II trials assess the safety, immunogenicity, and, sometimes, efficacy of selected doses of the vaccine, and generate hypotheses for later testing. Phase III trials, usually large in scale, seek to confirm the efficacy of the vaccine in the target population or prove the consistency of manufacturing processes. This is the last stage of clinical studies before licensure of the vaccine is requested. Phase IV trials are often conducted after licensure to collect additional information on the safety, immunogenicity, or efficacy of the vaccine to meet regulatory commitments or postmarketing objectives. Some sponsors further categorize phase IV trials and refer to postmarketing studies as phase V.

This article reviews some of the special considerations in vaccine development, including an increased interest in surrogate markers of activity (often termed as correlates of protection), a special need for assessing immune responses, a need for more specificity in endpoint definitions of clinical efficacy based on exposure, a greater need for confidence in safety, the unique regulatory process, and special health economic considerations of the public health impact of vaccination programs.
REGULATORY PROCESS FOR VACCINE LICENSURE

Vaccines have been successfully developed by using weakened live viruses, inactivated viruses, purified bacterial proteins and glycoproteins, and recombinant, pathogen-derived proteins. In the United States, vaccine development and licensing applications are regulated by the Center for Biologics Evaluation and Research within the Food and Drug Administration. The Center for Biologics Evaluation and Research’s mission is to “protect and enhance the public health through regulation of biological and related products including blood, vaccines, and biological therapeutics...”[5] The manufacture of biological products such as vaccines presents unique regulatory concerns in relation to quality control, consistency, stability, microbial contamination, product administration, and source of the material.[6] The Center for Biologics Evaluation and Research is responsible for ensuring the safety, purity, and efficacy of biological products intended for use in the treatment, prevention, or cure of diseases or conditions in humans.[5]

Vaccines are the most common class of biologic products regulated by CBER. The regulatory process starts with a filing of an Investigational New Drug application (IND), when preliminary testing of a new vaccine shows promising results in animals. All clinical trials under an IND are reviewed by CBER. The licensing application for a new vaccine is called the Biologic License Application (BLA), which is similar to the New Drug Application for a drug. When phase I/II studies are near completion and the planning of phase III (pivotal) trials is underway, the sponsor will request a pre-phase III meeting with CBER during the period of late phase II/pre-phase III to discuss a variety of issues, which may include designs of phase III trials, statistical analysis strategies, proposed indications of the vaccine, product profiles, manufacturing process, and facilities. Once phase III trials have been conducted and the sponsor is preparing for the submission of a licensing application, a pre-BLA meeting is usually scheduled with CBER to discuss a variety of topics such as the format and timing of submission, structure of clinical data, and any potential issues that may lead to refusal to file. The last step of applying for vaccine licensure is to file a BLA with CBER. Under the Prescription Drug User Fee Act of 1992, a standard review cycle (12 mo) is assigned for all submissions requiring the review of clinical data that cannot be classified as priority.[5] The Center for Biologics Evaluation and Research’s review of submissions that do not contain clinical data (such as filing of manufacturing changes) are to be completed within 6 mo. During the review of a BLA, CBER may request appropriate FDA advisory committee review of the new vaccine in terms of its efficacy, safety, and public health implications.

The manufacturing process is a fundamental characteristic of the vaccine product. As a result, the description of chemistry, manufacturing, and controls of the vaccine is particularly important in the licensing application. Because of the inherent variability in the manufacturing process, the concept of a generic drug has not been applied to vaccines. In fact, vaccines made under different manufacturing processes could be considered distinctly different products, although they are indicated for the same disease protection.[6]

ASSESSING VACCINE EFFICACY

One of the most critical steps of evaluating a new vaccine is to assess the protective efficacy of the vaccine against the target disease. If early phases (phase I/II) of clinical trials have demonstrated that a new vaccine is safe and able to induce immune responses, an efficacy trial (usually phase III, large scale) will often be conducted to evaluate whether the vaccine, as compared with a control, can completely prevent the disease of interest or at least reduce the incidence and/or the severity of the disease in the target population. When the control is a placebo injection or a vaccine that has no effect on the disease, the trial can estimate absolute vaccine efficacy. When the control is an already licensed vaccine for the same disease indication, the trial can estimate relative vaccine efficacy.[8]

Absolute vaccine efficacy is best assessed in a prospective, randomized, double-blind, placebo-controlled trial in which participants are randomly assigned to receive either a new vaccine (T) or a placebo control (C). Let \( P_T \) and \( P_C \) represent the true disease incidence rates among \( N_T \) vaccinees and \( N_C \) controls randomly assigned in the trial, respectively. The vaccine efficacy, denoted by \( \pi \), measures the relative reduction of the disease incidence in the vaccine group compared with the placebo group. A widely used measure of vaccine efficacy (\( \pi \)) is given by

\[
\pi = 1 - R, \quad \text{where} \quad R = \frac{P_T}{P_C}
\]  

(1)

Here \( R \) represents the relative risk of contracting the disease between the vaccine and placebo groups. In general, it is assumed that a new vaccine candidate will not be worse than placebo (\( P_T \leq P_C \)). A vaccine is 100% efficacious (\( \pi = 1 \)) if it prevents the disease completely (\( P_T = 0 \)), and it has no efficacy (\( \pi = 0 \)) if \( P_T = P_C \). Technically, however, the efficacy \( \pi \) can be negative if the vaccine group has a higher disease incidence rate than the placebo group.

In designing a vaccine efficacy trial, one needs to ensure that the study has sufficiently high power to test the hypothesis

\[
H_0 : \pi = \pi_0 \quad \text{versus} \quad H_1 : \pi > \pi_0
\]  

(2)

where \( \pi_0 \) denotes the minimal level of efficacy considered to be acceptable for the new vaccine. This is equivalent to requiring a high level of confidence that the vaccine
efficacy is greater than the prespecified lower bound ($\pi_0$). While $\pi_0$ could be 0, as is typically the case in establishing therapeutic efficacy of a new drug, considering a lower bound of 0 in vaccine trials involving healthy participants is usually not sufficient. In fact, it is a regulatory requirement to demonstrate that the protective efficacy of a new vaccine is significantly greater than some nonzero lower bound ($\pi_0 > 0$); this requirement was established so that researchers define more precisely the benefit of the vaccine and justify the risk of vaccinating potentially millions of healthy individuals. This nonzero lower bound is often called a “super efficacy” requirement. Orenstein, Bernier, and Hinman\cite{11} gave an excellent review of the design considerations for vaccine efficacy evaluation in terms of case definition, case finding, vaccination status ascertainment, and assuring comparability of vaccinated and unvaccinated groups. Lachenbruch\cite{20} discussed how the specificity and sensitivity of the case definition affects the vaccine efficacy estimate and concluded that a less-specific case definition will lead to a severe underestimation of the vaccine efficacy. Ellenberg and Dixon\cite{10} and Rida and Lawrence\cite{15} discussed many important statistical issues related to HIV vaccine trials.

**Vaccine Efficacy Based on the Ratio of Two Binomial Proportions**

In a randomized vaccine efficacy trial, the number of disease cases in the vaccine and control groups can be assumed to follow independent binomial distributions with parameters ($N_v$, $P_v$) and ($N_c$, $P_c$), respectively. This assumption is practical if the trial has a relatively short duration or all participants are completely followed through the end of the trial. If so, then a natural estimate of vaccine efficacy is

$$\hat{\pi} = 1 - \hat{R}, \quad \text{where} \hat{R} = \hat{P}_v / \hat{P}_c$$

(3)

Here $\hat{P}_v$ and $\hat{P}_c$ are the observed proportions of participants in the vaccine and control groups, respectively, who develop disease during the trial.

In this case, asymptotic methods for analyzing vaccine efficacy (hypothesis test and confidence interval) are direct applications of methods for estimating the relative risk of two binomial proportions; these methods have been extensively discussed in the literature.\cite{12-18} Of common use is the Z-type method proposed by Miettinen and Nurminen:\cite{7,13,16}

$$Z_x = \frac{\hat{P}_v - (1 - \pi_0) \hat{P}_c}{\hat{\sigma}_{\pi} \sqrt{N_v}}$$

(4)

where

$$\hat{\sigma}_{\pi} = \sqrt{\frac{\hat{P}_v(1 - \hat{P}_v) + \frac{1}{u}(1 - \pi_0)^2 \hat{P}_c(1 - \hat{P}_c)}{u N_v/N_v}}$$

(5)

and ($\hat{P}_v$, $\hat{P}_c$) are the constrained maximum likelihood estimates of ($P_v$, $P_c$) under the null hypothesis given in Eq. (2) based on the observed responses ($\hat{P}_v$, $\hat{P}_c$). Detailed expressions of ($T$, $C$) are given in Refs. [7,13]. The vaccine efficacy is demonstrated (rejecting the $H_0$) at the one-sided $\alpha$ (usually 2.5%) significance level if $Z_x < -Z_{\alpha}$, where $Z_{\alpha}$ is the $100(1 - \alpha)$ percentile of the standard normal distribution. This $Z_x$ test is equivalent to the likelihood score test\cite{16} and performs very well in many practical applications.\cite{7,13,16}

In addition to hypothesis testing, this method provides a test-based confidence interval that ensures consistent inference with the corresponding $p$-value. Asymptotic formulas for sample size and power calculations for vaccine efficacy (or relative risk) studies can be found in reports by O’Neill,\cite{13} Farrington and Manning,\cite{15} Blackwelder,\cite{16} and Nam.\cite{14} Exact inference of vaccine efficacy has been proposed by Chan\cite{20} and discussed by Chan and Bohidar\cite{7} in terms of power and sample size calculation. An example of sample size and power calculation based on the $Z_x$ test is given in “Sample Size and Power Calculation for Vaccine Efficacy Trials.”

**Vaccine Efficacy Based on the Ratio of Two Rates**

When the disease incidence rate is very low, a large-scale study is usually required to demonstrate vaccine efficacy. For sufficiently large sample sizes and small incidence of disease, the numbers of cases in the vaccine and placebo groups may be approximated by independent Poisson distributions with rate parameters $\lambda_v$ ($=N_v P_v$) and $\lambda_c$ ($=N_c P_c$), respectively. In this case, the number of cases in the vaccine group, given the total number of cases ($S$), is distributed as binomial ($S$, $\theta$) where

$$\theta = \frac{\lambda_v}{\lambda_v + \lambda_c} = \frac{N_v P_v}{N_v P_v + N_c P_c} = \frac{R}{R + u} = \frac{1 - \pi}{1 - \pi + u}$$

(6)

where $u = N_c/N_v$.

Because $\theta$ is decreasing in $\pi$, the efficacy hypothesis in Eq. (2) is equivalent to

$$H_0 : \theta \geq \theta_0 \text{ versus } H_1 : \theta < \theta_0$$

(7)

where $\theta_0 = (1 - \pi_0)/(1 - \pi_0 + u)$. Blackwelder\cite{16} discussed many asymptotic methods for constructing hypothesis tests and interval estimation regarding vaccine efficacy. For testing the classical null hypothesis with $\theta_0 = 0$, an exact conditional inference based on the Clopper and Pearson method\cite{21} has been proposed.\cite{22,23} Chan and Bohidar\cite{7} discussed a generalization of this exact conditional method for testing efficacy hypotheses with nonzero lower bounds. They found that the exact conditional test works extremely well for low incidence rates. This exact conditional method has been used to analyze the efficacy of vaccines for the prevention of hepatitis A, herpes zoster, and cervical cancer.
This exact conditional method[7] can also be used to design a study whose goal is to accrue a fixed number of events instead of running for a fixed duration. Once the total number of events is fixed, the power of the study depends on incidence rates only through the true efficacy, not the disease incidence rate. Thus one can avoid a situation sometimes encountered in a fixed-duration trial in which the anticipated power was not achieved at the end of the trial because the unexpectedly low incidence rates resulted in too few events.[7] In addition, the Poisson assumption also allows the exact conditional method to handle person–time data easily. This flexibility is important in long-term studies, in which follow-up often differs between the treatment groups. Sample size and power calculation based on this exact conditional method are discussed in Chan and Bohidar,[7] and an example is given in “Sample Size and Power Calculation for Vaccine Efficacy Trials.”

In longitudinal vaccine efficacy trials, time-to-event type analyses such as the Cox proportional hazards model[25] may also be used to compare the disease outcomes between the treatment groups. In this case, the vaccine efficacy may be estimated as the ratio of the hazard rates between vaccine and placebo groups.

### Trials to Estimate Relative Vaccine Efficacy

When the control is an already licensed vaccine, a randomized efficacy trial can estimate the relative efficacy ($\pi$) of a new vaccine via the relative risk ($R = P_T/P_C$) based on the relationship $\pi = 1 - R$. If the absolute efficacy of the control ($\pi_C$) has been established, one can estimate the absolute efficacy of the new vaccine ($\pi_T$) through the indirect argument[26,27]

$$\pi_T = 1 - R(1 - \pi_C)$$  \hspace{1cm} (8)

A trial to compare the relative efficacy focuses on the relative risk ($R$). It is usually designed as a noninferiority trial and tests the hypothesis

$$H_0 : R \geq R_0 \text{ versus } H_1 : R < R_0$$  \hspace{1cm} (9)

where $R_0$ ($>1$) is a prespecified noninferiority margin or relative risk threshold. The trial is powered to show that the relative risk $R$ is less than $R_0$ such that the lower bound of the confidence interval for the estimated absolute vaccine efficacy ($\pi_T$) is greater than $(1 - R_0(1 - \pi_C))$ with a high likelihood.[26] The choice of the noninferiority margin should be based on clinical, regulatory, and statistical judgments so that ruling out a relative risk of $R_0$ larger between the new and control vaccines will imply that the new vaccine preserves a large proportion of the efficacy of the control vaccine. Although a noninferiority margin that preserves 50% of the treatment effect has been proposed for the evaluation of drug treatments,[28–30] there is a general perception that a narrower margin should be used in preventive vaccine trials because the vaccine will be given to healthy individuals for prophylaxis.

### Other Considerations for Vaccine Efficacy Estimates

The vaccine efficacy estimates discussed in previous sections measure the direct effect of the vaccine on the prevention of the disease infection (reduction of disease incidence). Sometimes a vaccine may reduce both the incidence and the severity of the target disease. For example, the chickenpox vaccine has been shown to be more than 90% efficacious in preventing chickenpox in children 1 to 12 yr of age.[31–33] Furthermore, in vaccinated children who developed chickenpox, the severity (number of lesions and fever) was generally much milder than in unvaccinated children who contracted chickenpox.[33] Chang, Guess, and Heyse[34] proposed a combined measure of efficacy, which considers both incidence and severity in evaluating the total direct effect of a vaccine. Recent clinical trials have used this composite efficacy measure, called burden-of-illness, to evaluate the efficacy of new vaccines for the prevention of rotavirus diarrhea in children and herpes zoster in elderly persons. The burden-of-illness approach has also been adapted to compare injection-site symptoms of two hepatitis A vaccines.[35]

Besides the direct benefits, vaccination also often produces indirect benefits to unvaccinated persons by reducing person-to-person transmission of the disease in the population. Examples of vaccines conferring indirect benefits include the oral polio vaccine, the *Haemophilus influenzae* type B vaccine, and the measles–mumps–rubella vaccine. This indirect effect of the vaccine is called herd immunity.[36] Extensive work has focused on design and analysis in vaccine trials to estimate the indirect effect and the vaccine efficacy for the reduction of disease transmission rate or secondary attack rate.[4,37–40] This vaccine efficacy on disease transmission rate is an important measure for evaluating new vaccines intended to prevent diseases with cluster exposures, such as HIV infection.

Investigation of waning (or long-term) vaccine efficacy is also an important issue in vaccine trials. If long-term follow-up data on persons who received vaccines and those who received placebo are available, one can divide the study duration into different time intervals, estimate vaccine efficacy for each interval, and then examine the estimates for trends.[4] Durham et al.[41] used a nonparametric survival method to estimate the long-term efficacy of a cholera vaccine in the presence of waning protection. In many instances, long-term follow-up data on placebo controls are not available because these participants would have been offered vaccination once the vaccine has been shown to be efficacious. In such cases, long-term disease breakthrough rates among vaccinees can be compared with age-specific rates from the unvaccinated susceptible population to evaluate the long-term vaccine efficacy.[33]
Time-to-event analyses can also be performed to examine whether breakthrough rates among vaccinees change over time. However, assessing waning vaccine efficacy without concurrent controls may be subject to bias because the disease epidemiology may change over time, particularly with the increased vaccination coverage after licensure of the vaccine.

**Sample Size and Power Calculation for Vaccine Efficacy Trials**

As discussed in “Vaccine Efficacy Based on the Ratio of Two Binomial Proportions” and “Vaccine Efficacy Based on the Ratio of Two Rates,” there are many methods of calculating sample size and power for vaccine efficacy trials. Here we briefly describe the methods based on the commonly used asymptotic $Z_{01}$ test and the exact conditional test in a placebo control trial setting. For testing the null hypothesis $H_0: \pi = \pi_0$ against a specific alternative $H_1: \pi = \pi_1 (\pi_1 > \pi_0)$, the asymptotic power of an $\alpha$ level $Z_{01}$ test is given by

$$1 - \beta = \Phi \left( \frac{1}{\sigma_1} \left[ -Z_{01} \bar{\sigma}_n + \sqrt{N_1 P_c (\pi_1 - \pi_0)} \right] \right)$$

where $\sigma_1$ is the value of the expression in Eq. (5) when $P_1$ and $P_c$ are replaced with $P_0$ and $P_c$, respectively, $\bar{\sigma}_n$ is the limiting value of $\sigma_n$ in Eq. (5) obtained by calculating $\hat{P}_1, \hat{P}_c$ at the point $(\hat{P}_1, \hat{P}_c) = (P_1, P_c)$, and $\Phi$ is the standard normal distribution function. Similarly, the asymptotic sample size for an $\alpha$ level $Z_{01}$ test with $(1 - \beta)$ power is given by

$$N_1 = \left( Z_{01} \bar{\sigma}_n + Z_{\beta} \sigma_1 \right)^2 / [P_c (\pi_1 - \pi_0)]^2$$

and $N_c = uN_1$.

For example, suppose we design a study to test the efficacy hypothesis with a lower bound of $\pi_0$ of 0.2 at the one-sided 2.5% level, and we would like to achieve 95% power if the true vaccine efficacy is 0.8. Assuming equal sample sizes and a disease incidence rate of 0.6% in the placebo control group, the total number of subjects needed for the study will be approximately 10,838 based on Eq. (11).

To calculate the power using the exact conditional method, we first compute the critical value $Y_c$ corresponding to the one-sided $\alpha$ level test. Given a total number of cases ($T$) for the study and $\theta_0 = (1 - \pi_0)/(1 - \pi_0 + u)$ under the null hypothesis, $Y_c$ can be determined as the maximum number of cases in the vaccine group such that the null hypothesis is rejected: $Pr(Y \leq Y_c | Y \sim \text{Binomial}(T, \theta_0), H_0) \leq \alpha$. Then the power is given by Chan and Bohidar as

$$1 - \beta = \Pr(Y \leq Y_c | Y \sim \text{Binomial}(T, \theta_0), H_0) \leq \alpha \text{ as } \sum_{k=0}^{Y_c} \binom{T}{k} (1 - \theta_0)^{T-k} \left( \frac{\beta \Phi^{\pi_1 - \pi_0}}{\pi_1 - \pi_0} \right)^{k}$$

where $\theta_0 = (1 - \pi)/(1 - \pi + u)$. One can also use Eq. (12) to iteratively determine the total number of cases ($T$) required for the study to achieve the desired power. Because the unconditional expected value of $T$ is $(N_1P_1 + N_cP_c)$, we can estimate the number of participants needed for the study as

$$N_1 = T / [(1 - \pi_1 + u)P_c] \text{ and } N_c = uN_1$$

For the above example, a total number of cases of 37 will ensure the study to have at least 95% power to show the vaccine efficacy. Based on the disease incidence rate of 0.6% in the placebo control group, the total number of subjects required for enrollment is estimated to be 10,278 using Eq. (13), which compares favorably with the sample size of 10,838 obtained based on the asymptotic formula.

**ASSESSING VACCINE IMMUNOGENICITY**

Vaccine immunogenicity clinical trials study the immune response to vaccination, which are usually measured by serum antibody titers or T-cell responses. Many immunogenicity trials are conducted in the early phases of vaccine development to assess whether the vaccine can induce immunity before large-scale efficacy trials are performed. In addition, immunogenicity of a vaccine is typically assessed in an efficacy trial to determine whether an immune marker to the vaccine can be used as a surrogate or correlate of disease protection (see the section “Immune Surrogate and Correlate of Protection for Vaccines” for more details).

Once an immune surrogate or correlate of protection has been established, immunogenicity trials can be used as efficient alternatives (for time and economical considerations) to efficacy trials in evaluating the effectiveness of future vaccines. After the successful demonstration of vaccine efficacy, for example, immunogenicity trials are typically used to demonstrate the consistency of vaccine manufacturing process; to prove the vaccine effectiveness when manufacturing processes, storage conditions, or vaccination schedules are modified; to justify concomitant use with other vaccines; and to develop a new combination vaccine with multiple components. The primary objective of such immunogenicity studies is to demonstrate that a new (or modified) vaccine is noninferior (or equivalent) to the current vaccine by ruling out a prespecified clinically relevant difference in the immune response.

**General Approaches to Immunogenicity Testing**

**Immunogenicity Endpoints**

Two types of immunologic endpoints are commonly used to assess vaccine immunogenicity: (1) an immune...
response rate, defined as the percentage of participants who achieve a certain level of immune response after vaccination (yes/no); and (2) a geometric mean titer (GMT) or geometric mean concentration of immune response after vaccination. When a particular level of immune response has been shown to be correlated with disease protection, the percentage of participants achieving the "protective level" after vaccination is usually considered the primary endpoint for immunogenicity analyses. In studying participants with preexisting immunity, additional endpoints may include the percentage of participants who develop a certain (e.g., fourfold) fold-rise and the geometric mean fold-rise in immune response from before to after vaccination.

Approaches for Analyzing Immune Response Rates

For the analyses of immune response rates, we consider the general setting comparing a new vaccine (T) to a control (C) in a randomized study. If the control is a placebo, the study is designed as a superiority trial. If the control is a licensed vaccine, the study is often designed as a noninferiority trial. Using the notation of “Assessing Vaccine Efficacy,” let \( P_T \) and \( P_C \) represent the true immune response rates of the control (with \( N_C \) subjects) and the new vaccine group (with \( N_T \) subjects), respectively. Then the number of successes (i.e., participants who achieve a certain level of immune response) in the control and new vaccine groups is independently distributed as binomial (\( N_C, P_C \)) and binomial (\( N_T, P_T \)), respectively. In general, the study can be designed to test the statistical hypothesis (superiority or noninferiority)

\[
H_0: P_T - P_C \leq -\delta \quad \text{versus} \quad H_1: P_T - P_C > -\delta 
\]

where \( \delta \) is a prespecified small quantity defining the noninferiority margin. The nominal significance level for the one-sided test is generally set at half of the conventional significance level for a two-sided test for the difference in proportions. This approach has been adopted in regulatory environments, as suggested in the International Conference on Harmonization E9 (ICH E9) guidelines.\(^{42}\)

When \( \delta = 0 \), the hypothesis (Eq. [14]) reduces to the classical one-sided hypothesis aimed to show that the new vaccine is superior to the control. Standard asymptotic methods for testing a null hypothesis of no difference in two proportions can be used.\(^{35}\) In addition, exact methods are available for testing this hypothesis\(^{44,45}\) and for constructing confidence intervals.\(^{46,47}\) Finally, some discussions and debates of one-sided versus two-sided hypothesis for superiority testing can be found in Peace\(^{48}\) and Fisher.\(^{49}\)

When \( \delta \) is a small positive quantity, the hypothesis above (Eq. [14]) aims to show that the new vaccine is noninferior (or equivalent) to the control vaccine. Considerations for the choice of noninferiority margin \( \delta \) are discussed below at the end of “General Approaches to Immunogenicity Analysis.” Asymptotic statistical tests of the noninferiority hypothesis (Eq. [14]) for two independent treatment groups with a dichotomous endpoint have been extensively discussed in the literature (e.g., Refs. [13,15,50–52]). Many authors have proposed Z-type test statistic with different standard error estimates.\(^{13,51,52}\) Of common use and better performance is the Z-type method proposed by Miettinen and Nurminen:\(^{13,15}\)

\[
Z_\alpha = \frac{\hat{P}_T - \hat{P}_C + \delta}{\sigma_{\alpha} / \sqrt{N_T}} 
\]

where

\[
\sigma_{\alpha} = \left[ \frac{\hat{P}_T(1 - \hat{P}_T)}{u} + \frac{\hat{P}_C(1 - \hat{P}_C)}{u} \right]^{1/2}, \quad u = N_T / N_T
\]

and \( (\hat{P}_T, \hat{P}_C) \) are the constrained maximum likelihood estimates of \( (P_T, P_C) \) under the null hypothesis given in Eq. (14) based on the observed responses \( (\hat{P}_T, \hat{P}_C) \). Detailed expressions of \( (\hat{P}_T, \hat{P}_C) \) are given in Refs.\(^{13,15}\). The noninferiority is established \( (H_0 \text{ rejected}) \) at the one-sided significance level if \( Z_\alpha \geq Z_\alpha \) where \( Z_\alpha \) is the 100(1 - \( \alpha \)) percentile of the standard normal distribution. For immunogenicity trials with small sample sizes, as are often conducted in the early phases of vaccine development, exact tests and confidence intervals\(^{42,46,53}\) can be used to compare treatments and test the hypothesis in Eq. (14).

Some immunogenicity trials also involve a hypothesis regarding whether the new vaccine induces an adequate immune response compared with a historical control.\(^{154}\) The statistical hypothesis of interest is that \( H_{0_1}: P_T \leq p_0 \) versus \( H_1: P_T > p_0 \) where \( p_0 \) is the prespecified lower limit determined from the historical control data. The statistical criterion for this hypothesis is equivalent to requiring that the lower bound of the confidence interval on the single group immune response is larger than the prespecified lower limit. A confidence interval for immune response rate can be computed by using asymptotic or exact methods for a single binomial proportion.

Approaches for Analyzing Geometric Mean Titers

For the comparison of GMTs, an immunogenicity study is usually designed to test the following general statistical hypothesis:

\[
H_0: \frac{\text{GMT}_T}{\text{GMT}_C} \leq K \quad \text{versus} \quad H_1: \frac{\text{GMT}_T}{\text{GMT}_C} > K
\]

where \( \text{GMT}_T \) and \( \text{GMT}_C \) represent the GMTs of the control and new vaccine groups, respectively, and \( K \) is a prespecified positive quantity relevant to the comparison of fold-difference between the two GMTs. Titers are usually log-transformed in the statistical analysis. The confidence interval for individual GMTs is often calculated on the basis of the asymptotic t-distribution.
When $K = 1$, the hypothesis (Eq. [17]) reduces to the classic one-sided hypothesis of superiority of the new vaccine compared with the control. When $0 < K < 1$, the rejection of the null hypothesis $H_0$ in 17 at a prespecified $\alpha$ level (usually one-sided $\alpha = 0.025$) leads to the conclusion that the new vaccine is noninferior to the control with respect to the GMTs, with an interpretation that a difference in GMTs (control/new) is $< 1/K$-fold. This statistical criterion corresponds to the lower bound of the corresponding two-sided $(1 - 2\alpha)$ 100% confidence interval on the ratio of GMTs being greater than $K$. Some considerations for the choice of noninferiority margin $K$ are given next. The statistical testing and confidence interval estimation regarding GMTs can be performed by using an analysis of variance (ANOVA), an analysis of covariance, or a linear mixed-effects model that includes natural log of the antibody titer as the dependent variable and uses treatment group, baseline, and stratification variables as explanatory variables.

In addition to these parametric analyses on GMTs, graphical displays of reverse cumulative distribution curves of $\Pr(X \geq x)$, which give percentages of participants with titers greater than or equal to varying levels of titers ($x$), are commonly used for exploratory evaluation of immune response to vaccination. Nonparametric methods have also been proposed to estimate the overlap or the proportion of similar response in distributions, which can be used to measure the similarity between two distributions of immune response.

Considerations for Noninferiority/Equivalence Margins

In general, the choice of $\delta$ and $\gamma$, called noninferiority or equivalence margins, depends on the level of correlation between immune responses and the vaccine efficacy, the variability of immunogenicity measures, and the importance of these endpoints in the trial. The choice should be based on statistical reasoning as well as on clinical and regulatory judgments so that ruling out a difference of $\delta$ in immune response rates (or a fold-difference of $1/K$ in GMTs) between the new vaccine and the control will demonstrate that the effectiveness of the new vaccine is comparable to that of the control. For example, in studies of the hepatitis A vaccine VAQTA®, a $\Lambda$ of 10 percentage points is typically used as the noninferiority margin for comparing immune response rates, which are usually greater than 90%. A $K$ of 0.67, 0.5, or 0.25 (corresponding to 1.5-, 2-, or 4-fold difference) are often used in comparing GMTs. These noninferiority margins should be discussed proactively between the sponsor and regulatory agencies. Regulatory agencies usually apply the same requirements of equivalence margins for the same endpoints when reviewing similar products from different companies. More general discussions on the choice of noninferiority or equivalence margins can be found in Temple, Jones et al., Ebbutt and Frith, the ICH E9 guidelines, the ICH E10 guidelines, Siegel, and Wiens.

Different Types of Vaccine Immunogenicity Studies

Superiority Immunogenicity Study

Immunogenicity studies are often conducted in the early phases of vaccine development to assess vaccine immunogenicity compared with a placebo. In addition, immunogenicity trials can also be performed to claim superiority of one vaccine over another vaccine from a different manufacturer with respect to immune responses. Such immunogenicity trials should be designed as a superiority (difference-detection) trial rather than an equivalence or noninferiority trial. In this case, the conventional null hypothesis of no difference should be tested against a one-sided alternative or two-sided alternative. The analysis strategy follows the methods for $\Lambda = 0$ or $K = 1$ described in the section “General Approaches to Immunogenicity Analysis.”

Dose Response Immunogenicity Study

During vaccine development, a dose–response study may be conducted to assess the immunologic response across different dosage levels of the vaccine and thus determine the minimum effective and safe dose. For many live-virus vaccines, the potency of the vaccine may decay over time. A dose–response study will help investigators understand the kinetic of potency decay, determine the release and end-expiry dosage level, and specify the shelf life of the vaccine. Immune response rates and GMTs are often co-primary endpoints in this type of study. The existence of a dose–response trend in the immune response rate can be tested by using the Cochran–Armitage (C–A) trend test. A dose–response trend with respect to GMTs can be evaluated by using an ANOVA model. Multiple testing procedures can be used to control the overall type I error rate if multiple trend tests or comparisons are made.

Consistency Lots Study

Vaccines are biological products, and their manufacturing process generally varies much more widely than does that of chemical drug products. Before a vaccine can be licensed, the sponsor must demonstrate consistency of the vaccine manufacturing process through analytical and clinical testing. A clinical consistency lot study typically uses three lots of vaccines made from the same manufacturing process (called consistency lots) and a control vaccine (see an example in Ref. [57]), and is intended to (1) rule out a clinically significant difference in either direction between any two pairs of the three consistency lots; and (2) rule out a clinically significant difference
between the combined immune response of three consistency lots and the response of the control. A consistency lot study is typically designed as a two-sided equivalence study with respect to both an immune response rate and a GMT.

The hypothesis of consistency among three lots of vaccine with respect to immune response rate can be written as

\[ H_0 : |P_i - P_j| \leq \Lambda \text{ for all } (i, j) \]

versus \[ H_1 : |P_i - P_j| > \Lambda \text{ for some pair } (i, j) \]

where \( P_i \) and \( P_j \) are the immune response rates for lots \( i \) and \( j \), \( 1 \leq i < j \leq 3 \). Testing this hypothesis is equivalent to testing three pairs of noninferiority hypothesis tests, all of which must be rejected for consistency to be declared among the three lots. The nominal level of significance for the consistency lots hypothesis is usually set at the \( \alpha = 0.05 \) level, at which all six hypotheses are tested. The statistical criterion is equivalent to requiring that all two-sided 90% confidence intervals for the three pairwise differences in immune response rates fall within the interval of \( (\pm \delta, \delta) \). The analysis approach for each noninferiority hypothesis can follow the methods described in “General Approaches to Immunogenicity Analysis.” Wiens and Iglewicz proposed a slightly improved statistic to test the hypothesis (Eq. [18]). Once the consistency among the three vaccine lots has been demonstrated, the immune response for the three lots can be pooled and compared with the response of the control in a noninferiority hypothesis setting. The hypothesis testing strategy for clinical lot consistency with respect to GMTs can be structured in a fashion similar to that for the immune response rate. The sample size and power for consistency lot studies can be calculated by using simulation. More discussions on sample size and power calculations can be found in the section “Sample Size and Power Considerations for Vaccine Immunogenicity Studies.”

**Bridging Study**

After vaccine licensure, manufacturing processes, storage conditions, routes of administration, or dosing schedules may be changed to improve production yields, potency stability, or convenience of vaccination schedule. It is a regulatory requirement to conduct a bridging study (comparing a modified vaccine with the current vaccine) to demonstrate that such changes do not have adverse effects on vaccine effectiveness. If an immune marker correlates with disease protection (see “Immune Surrogate and Correlate of Protection for Vaccines” for more details), immunogenicity bridging trials can be conducted instead of efficacy trials to show clinical equivalence or noninferiority of the modified vaccine to the current vaccine.

An immunogenicity bridging trial is commonly designed as a noninferiority trial aimed at excluding a clinically significant difference in the immune response between the modified vaccine and the current vaccine. The analysis strategy usually follows the methods described in “General Approaches to Immunogenicity Analysis.” The noninferiority hypothesis is usually tested at \( \alpha = 0.025 \). The key for determining the study hypothesis is the choice of the noninferiority margin, namely, a clinically tolerable difference in proportions (\( \Lambda \)) or fold-difference in GMTs (\( K \)) between the modified vaccine and the current vaccine. As discussed in “General Approaches to Immunogenicity Analysis,” the choice of \( \Lambda \) and \( K \) depends on the level of correlation between immune responses and the vaccine efficacy, the variability of immunogenicity measures, and the importance of these endpoints.

**Combination or Multivalent Vaccine Study**

A combination or multivalent vaccine consists of two or more live organisms, inactivated organisms or purified antigens combined by the manufacturer or mixed immediately before administration. The vaccine is intended to prevent multiple diseases, prevent one disease caused by different strains or serotypes of the same organism, or reduce the number of injections and health care visits. The immunogenicity of all the vaccine components in the combination or multivalent vaccine should be studied to rule out the clinically significant differences in immune response rates or GMTs between the combined vaccine and the separate but simultaneously administered antigens. In addition, acceptable levels of immunogenicity should be demonstrated for each serotype or component.

In most cases, the endpoints for each component are considered co-primary, and the success requires the demonstration of similarity for all components (see case 2 in Ref. [70]). In some cases, the success requires the demonstration of similarity for any \( M \) of the \( K \) components (see case 3 in Ref. [70]). For this kind of problems, Capizzi and Zhang proposed a slightly improved statistic to test the hypothesis (Eq. [18]). Once the consistency among the three vaccine lots has been demonstrated, the immune response for the three lots can be pooled and compared with the response of the control in a noninferiority hypothesis setting. The hypothesis testing strategy for clinical lot consistency with respect to GMTs can be structured in a fashion similar to that for the immune response rate. The sample size and power for consistency lot studies can be calculated by using simulation. More discussions on sample size and power calculations can be found in the section “Sample Size and Power Considerations for Vaccine Immunogenicity Studies.”

**Immunologic Persistence Study**

The clinical development phase of a vaccine often takes several years before reaching its licensure. However, the duration of vaccine-induced immunity is expected to be considerably longer than the duration of the clinical studies. Thus there are both conceptual and methodological challenges in assessing whether immunity is waning and in estimating long-term immunologic persistence (defined as the presence of vaccine-induced antibody or cell-mediated immunity over time). To this end, immunologic studies (often open-labeled) are conducted, even after vaccine licensure, to annually measure the immune responses over...
multiple years. Life-table or time-to-event analyses can be used to analyze the cumulative immunologic persistence rate. In addition, several modeling strategies have been proposed to predict the duration of vaccine-induced immunity based on the extrapolation of observed antibody data from the clinical studies. Although prediction from these extrapolation models relies on many assumptions, they may be useful for developing vaccine booster recommendation and epidemiologic surveillance.

**Sample Size and Power Considerations for Vaccine Immunogenicity Studies**

Sample size and power calculation for vaccine immunogenicity trials usually follow the same principles and formula as in the drug trials. Appropriate sample size and power estimations for immunogenicity studies depend on (1) reasonable assessments of source and magnitude of immunologic variability; (2) meaningful determinations of clinically significant difference/threshold in immune response rates or geometric mean titers; (3) clearly stated study hypotheses and endpoints; and (4) well-planned statistical analysis strategies.

**Bridging Study**

For bridging studies that are designed to test the noninferiority hypothesis (Eq. [14]) with respect to immune response rates, there are many methods of calculating the sample size and power. Of common use is the sample size formula proposed by Farrington and Manning, which is based on the \( Z \)-type test statistic (Eq. [15]). To have a \( 1 - \beta \) power to claim noninferiority at one-sided \( \alpha \) level, the approximate sample size for testing the hypothesis (Eq. [14]) should be

\[
N_t = \left[ Z_{\alpha} + Z_{\beta} \right]^2/(P_t - P_c + \delta)^2
\]

and

\[
N_c = uN_t
\]

where \( u \) is the prespecified sample size ratio between the control and the test vaccine groups, \( P_c \) and \( P_t \) are the expected immune response rates for control and test vaccine groups in the planned study, respectively, \( Z_{\alpha} \) is the value of the expression in 16 when the constrained maximum likelihood estimates \( \hat{P}_t \) and \( \hat{P}_c \) are replaced with \( P_t \) and \( P_c \), respectively, and \( \hat{\sigma}_{\log} \) is the limiting value of \( \hat{\sigma}_{\log} \) in Eq. (16) obtained when \( \hat{(P_t, \hat{P}_c)} = (\hat{P}_t, \hat{P}_c) \). Similarly, the asymptotic power of a one-sided \( \alpha \) level test based on the normal approximation is given by

\[
1 - \beta = \Phi \left[ \frac{1}{\sigma_{\log}} \left( -Z_{\alpha} \hat{\sigma}_{\log} + \sqrt{N_t} (P_t - P_c + \delta) \right) \right]
\]

where \( \Phi \) is the standard normal distribution function.

To illustrate the sample size calculation under this setting, let us assume that one needs to design a bridging study to rule out a 10 percentage-point difference in immune response rate between a new test vaccine and a control vaccine. The noninferiority margin in the study hypothesis (Eq. [14]) is 10% (i.e., \( \Delta = 0.1 \)). Assuming the hypothesis is tested at one-sided \( \alpha = 0.025 \) level and the expected immune response rates for both groups are 90% in the planned study (i.e., \( P_c = P_t = 0.9 \)), then in order to have 95% power to rule out a 10 percentage-point difference between the equally sized test and control groups, ~250 subjects per group are required to be evaluable for the primary analysis. Assuming a 10% nonevaluable rate because of protocol violation, loss to follow-up, invalid assay results, etc., ~280 subjects should be enrolled in each group.

For bridging studies that are designed to test the noninferiority hypothesis in Eq. (17) with respect to the GMTs, the sample size calculation is similar to that for testing the conventional hypothesis of no difference under the lognormal assumption. Let \( I \) be the standard deviation of the log-transformed antibody titer. In order to achieve \( 1 - \beta \) power in testing the hypothesis (Eq. [17]) at the one-sided \( \alpha \) level, the sample size for the test vaccine group is:

\[
N_t = (1 + I/(Z_{\alpha} + Z_{\beta})^2 I^2/\log(R_{GMT} - \log(\gamma))^2
\]

and

\[
N_c = uN_t
\]

where \( u \) is the prespecified sample size ratio between the control and the test vaccine groups and \( R_{GMT} \) represents the ratio of GMTs between the new and control vaccine groups under the alternative hypothesis. Similarly, the asymptotic power of a one-sided \( \alpha \) level test based on the normal approximation is given by

\[
1 - \beta = \Phi \left[ \frac{1}{\sigma_{I}} \left( -Z_{\alpha} \hat{\sigma}_{I} + \sqrt{N_t} (I(P_t - P_c + \delta)) \right) \right]
\]

**Consistency Lot Study**

When planning a consistency lot study to compare three clinical lots, no closed-form formula exists for sample size calculation, and simulation study can be used to adequately plan the sample size. For power estimation, Wiens, Heyse, and Matthews recommended performing simulations with the three population parameters under the alternative hypothesis of being not all equal. For example, for hypothesis testing with respect to the immune response rate, they recommended performing simulations under the setup that the true immune response rates for the three groups are \( P_t \), \( P_t' \), and \( P_t + 0.5A \), instead of being all equal. This gives a conservative estimate of power when the alternative hypothesis is true.
Combination or Multivalent Vaccine Study

In a combination or multivalent vaccine study, the hypotheses regarding each component or each serotype are often considered as co-primary. In planning such a study, sample size for each primary hypothesis can be calculated using formulas 19, 21, or other related sample size formulas for that hypothesis. To address the concerns of multiplicity, one approach is to plan the power for each primary hypothesis large enough such that the overall power is acceptable under the assumption of independence among the multiple hypotheses. This approach is generally conservative (estimating lower bound of power) because the correlation between components is usually positive. For example, for a multivalent vaccine with \( k \) components, one can plan the power for hypothesis testing on each component to be at least \( 1 - \beta/k \) so that the overall power is guaranteed to be no less than \( 1 - \beta \).

**ASSESSING VACCINE SAFETY**

A complete evaluation of safety is an important component in all vaccine development programs. The safety considerations of vaccines differ from those of the pharmaceutical products because of the fundamental role of vaccines in public health initiatives. Vaccines are administered to millions of healthy individuals around the world with the objective of preventing infectious diseases. Many vaccines are mandated for entry into public schools, day care and preschool programs, and even universities. Safety evaluations begin with the earliest preclinical experiments and continue throughout the clinical development program and marketing life span of the vaccine. Phase IV postlicensure regulatory commitments typically include a formal safety evaluation in a large managed health care setting. In addition to postmarketing surveillance systems monitored by individual corporate sponsors, the FDA and the CDC have developed a national adverse event reporting system for U.S. licensed vaccines.

The ICH E9 guidelines include specific recommendations on safety evaluation in clinical trials of products that pertain to both drugs and vaccines. Safety is always an important consideration in vaccine trials, although these trials are mostly designed with a sample size related to a primary hypothesis of efficacy or immunogenicity. It is important to monitor all clinical trials for adverse events that become apparent. Independent data monitoring committees are often established in large multicenter trials to assess the safety and efficacy data at predefined study intervals. The committee may recommend to the sponsor to modify or terminate a trial on the basis of an evaluation of the risks and benefits to the study participants.

The methods and measurements chosen to establish the safety of a vaccine depend on many factors, including the type of vaccine (e.g., live attenuated viruses or bacterial vaccines) and its specific mechanism for eliciting immune responses. Vaccination can cause allergic or anaphylactic reactions because it induces the immune system. The reactions are typically local to the site of the infection, such as swelling, tenderness, or redness. Systemic reactions, such as fever or muscle ache, can also occur. The vaccine may also produce an illness similar to that produced by the wild-type infectious agent. Coplan et al. developed and validated a standardized vaccine report card for use in vaccine trials. The pediatric and adult instruments were developed to ensure that the wording was understandable, to promote ease of use, and to enhance accuracy and completeness of reporting. The validation phase showed that the use of a standardized vaccine report card provides a practical and complete method of eliciting complaints after vaccination.

With the increasingly aggressive vaccination schedule in children, an interest in the development of combination vaccines has been expressed. Ellenberg discussed the statistical considerations in evaluating the safety of combination vaccines. Although combination vaccines generally include individual components that have been previously studied and used, new or more severe adverse reactions might occur in the combination product. Clinical trials usually compare the combination vaccine to the individual components. Although the safety database may not need to be as large as that for the novel vaccines with new immunogens, a reliable evaluation of safety should still be performed. The CBER developed guidelines for the evaluation of combination vaccines.

All safety variables encountered in a clinical trial require attention, and the protocol should specify the analytical approach. All adverse events should be reported, regardless of whether they are considered related to vaccination. When studies include hypotheses about specific adverse events, the usual statistical framework is appropriate. However, the main difficulty with the statistical analysis of adverse events is the possibility of new and unanticipated reactions. Treating these safety variables as confirmatory information is not appropriate. For this type of safety evaluation, it is best to apply descriptive statistics supplemented by confidence intervals. Measures of risk differences or relative risk are useful for this purpose. Reporting individual \( p \)-values can sometimes be used to help evaluate a specific difference of interest or as a “flagging” device applied to many safety variables to highlight possible vaccine effects that are worth further attention.

The ICH E9 guidelines recognize the importance of considering multiplicity in the interpretation of safety data. They noted the need to quantify the type I error by using statistical adjustments for multiplicity while recognizing the concerns of type II errors in safety assessment. Mehrotra and Heyse proposed a two-step use of adjusted \( p \)-values based on the false discovery rate method of Benjamini and Hochberg as a flagging method. They used data from three vaccine clinical trials to illustrate...
the proposed double false discovery rate approach and to reinforce the potential effect of failure to account for multiplicity.

Because vaccines will typically be administered to millions of otherwise healthy individuals, evaluating vaccines for rare but serious adverse events is important. Common reactions to vaccines are readily identified in the investigational phase of clinical development programs. However, rare but serious events may be missed in the clinical trials. To exemplify the problem, we refer to the “rule of three,”[80] which states that \(3/n\) is an upper 95% confidence bound for the binomial probability \(P\) when no events occur among \(n\) independent trials. Applying this rule to a vaccine program of \(n = 5000\) participants gives an upper bound estimate of \(P = 0.0006\) for adverse events not observed in the studies. Administration of the vaccine to 1 million persons could yield 600 severe reactions.

Several activities have been put in place to protect against this situation. Clinical development programs include prelicensure safety studies to build up the safety database. In addition, sponsors undertake Phase IV studies to address important safety concerns. For example, Black et al.[81] reported the results of a prospective postmarketing evaluation of the safety and effectiveness of varicella vaccination in 89,753 adults and children in the preventive care program of the Northern California Kaiser Permanente Medical Care Program. These types of studies are possible in large managed care organizations because of the large patient populations and the sophisticated use of automated clinical databases of hospitalizations, emergency department visits, and clinic visits. Other types of epidemiologic investigations of vaccine safety are also undertaken. Ascherio et al.[82] reported the results of a nested case-control study of a possible association between hepatitis B vaccination and the risk of multiple sclerosis. The study was prompted by spontaneous reports of multiple sclerosis development after vaccination. Ascherio et al.[82] reported no such association in the context of this well-conducted evaluation.

The FDA and the CDC have created the Vaccine Adverse Event Reporting System (VAERS) for postmarketing safety surveillance.[83] This system accepts reports of adverse events that may be associated with U.S. licensed vaccines from health care providers, manufacturers, and the public. The reports are continually monitored for any unexpected patterns or changes in rates of adverse events. Niu, Eriwn, and Braun[84] used empirical Bayes data-mining techniques of DuMouchel[85] in VAERS to investigate the detection of intussusception after rotavirus vaccination.

Postmarketing evaluations are complex and have their drawbacks.[86,87] In addition, many persons may be at risk during the early phases of the vaccine integration to the marketplace. Ellenberg[87] argued that the most reliable way to assess causality in safety is a controlled study and that the size of clinical trials of new vaccines must be increased to be able to detect serious rare effects. She proposed sample sizes on the order of 60,000 to 80,000 participants using the methods appropriate under simple binomial sampling. This information would increase the confidence in the safety of vaccines and would be a valuable resource for assessing spontaneous reports of adverse events after licensure. Overall, a comprehensive and continual assessment of the safety of a vaccine reduces the risk that a vaccine may cause severe reaction to a small proportion of vaccinees.

The association between rotavirus vaccination and intussusception provides an excellent example of the issues relating to evaluating vaccine safety. Intussusception is a rare but serious bowel condition that occurs in young children 3 mo to 2 yr of age. It occurs when a portion of the bowel folds in on itself and causes a constriction that interrupts blood flow. The incidence of intussusception is approximately 1 case per 2000 person-yr. In prelicensure studies of a rhesus-human reassortant rotavirus tetravalent vaccine (RRV-TV), 5 cases of intussusception were identified among 10,054 vaccinated infants and 1 case was identified among 4633 placebo recipients. This result was not statistically significant, and Rennels et al.[88] concluded that the findings did not support an apparent association between intussusception and RRV-TV. The RRV-TV was licensed on August 31, 1998, and was recommended by the Advisory Committee on Immunization Practices as a three-dose series given at 2, 4, and 6 mo of age.

Murphy et al.[89] reported on an investigation that was initiated after 9 cases of intussusception were reported to VAERS. They undertook a thorough data collection exercise and analyzed data for 429 infants with intussusception and 1763 matched controls in a case-control analysis, as well as for 432 infants with intussusception in a case series analysis. A case series analysis[90,91] estimates the relative incidence of clinical events in defined time intervals after vaccination compared with control periods using only cases. Murphy et al.[89] concluded a strong causal association between intussusception and RRV-TV vaccination, especially within 1 week after the first dose. Niu, Eriwn, and Braun[84] used empirical Bayes data-mining methods on the VAERS intussusception data and showed that even earlier detection of rare serious adverse events was possible. The Advisory Committee on Immunization Practices withdrew its recommendation of the vaccine, and the vaccine manufacturer voluntarily withdrew its product from the market.[92]

Sadoff et al.[93] described the study design considerations necessary to detect an increased risk of intussusception in a randomized clinical safety study. By using extensive monitoring for intussusception cases through multiple stopping boundaries, they noted that efficiencies in the design are possible. Still, studies to detect rare safety outcomes require very large sample sizes. The complex design described by Sadoff et al.[93] used Monte Carlo simulation methods for a study that involves at least 60,000 infants. This sample size
is consistent with the sample sizes generally proposed by Ellenberg[77] for vaccine safety studies.

As a postscript, it is interesting to note that 4 of the 5 cases of intussusception among RRV-TV recipients occurred within 2 weeks of vaccination. The single case in placebo recipients occurred several weeks after vaccination. This observation supports O’Neill’s[94] recommendation to consider the timing of events in the analysis of safety data.

**ADDITIONAL STATISTICAL ISSUES FOR ANALYZING VACCINE TRIALS**

**Intention-to-Treat or Per-Protocol Analysis**

A typical intention-to-treat population for vaccine trials includes all randomly assigned participants, regardless of whether they met study entry criteria, the vaccine they actually received, and subsequent withdrawal or deviation from the protocol. In comparison, a typical per-protocol population includes a subset of randomly assigned participants who met the study entry criteria, received study vaccines as prescribed by the protocol, completed the study without major protocol violations, and had endpoint measurements at specific time points. Some vaccine efficacy studies also use a modified intention-to-treat population, which includes only participants who complete all doses of the assigned vaccine and disease cases occurring after some specific time following the full series of vaccination to capture the full effect of the vaccine. More general discussions of these analysis populations in clinical trials can be found in Gillings and Koch,[95] the ICH E9 guidelines,[42] Lachin,[96] and Lange.[97]

For the analysis of vaccine efficacy, an intention-to-treat analysis generally provides unbiased assessment of vaccination strategy and controls false-positivity rates in placebo-controlled efficacy trials. However, this analysis may artificially inflate the false-positivity rate in trials aimed to show equivalent/noninferior efficacy or immunogenicity of two vaccines. A per-protocol or modified intention-to-treat analysis of vaccine efficacy generally aims to assess the biological effect of the vaccine. Historically, the per-protocol or modified intention-to-treat analysis of vaccine efficacy has often been considered the primary approach, and the intention-to-treat analysis has been considered the secondary approach.[98] This is probably because bias is less an issue in vaccine trials involving healthy participants; high adherence is usually achievable and missing data/dropouts as a result of adverse experience are usually very low.[99] In addition, the intention-to-treat and modified intention-to-treat/per-protocol analyses usually yield the same conclusion about the vaccine efficacy. However, if bias is a major concern of the trial (such as in open-label trials), or the goal is to compare the effectiveness of vaccination strategy (such as comparing a three-dose regimen at 1, 2, and 6 mo versus 2, 6, and 12 mo), then intention-to-treat analysis should be used.[99]

A per-protocol analysis is usually the primary approach for immunogenicity, which is designed to assess the biological or immunologic effects of vaccination. In addition, an intention-to-treat or modified intention-to-treat analysis is typically planned to support or confirm the per-protocol analysis. Safety analysis of vaccine trials is typically based on a modified intention-to-treat population that includes all participants vaccinated and for whom safety follow-up data are available.

**Missing/Coarse Data and Loss to Follow-Up**

Missing data and loss to follow-up may occur in vaccine studies. These issues present fewer problems in estimating the efficacy, immunogenicity, or safety of vaccines than in drug trials. Because vaccination is usually administered once or as a few doses over time, treatment adherence is typically very high. In addition, as mentioned earlier, few dropouts are a result of vaccine-related adverse events. In practice, missing data and loss to follow-up from vaccine trials are generally a result of the causes unrelated to the study vaccine. Therefore missing data arising from vaccine trials can reasonably be assumed to be missing at random or even missing completely at random. In an efficacy trial, one can estimate the loss of information (disease cases) because of loss to follow-up and perform sensitivity analyses. Multiple imputation techniques[100] are useful in these analyses.

Serial-dilution assays are commonly used to measure immune responses. Without a standard curve, the reported immune responses from a serial-dilution assay are often interval-censored, which can be considered as a special form of coarse data.[101] The lower limit of the interval is routinely reported and treated as the exact titer value in immunogenicity analyses. However, ignoring the coarseness of the immune responses may result in severe bias in parameter estimation. Different approaches have been discussed for addressing this issue.[102]

**Analyses with Stratification Variables**

Many vaccine trials are conducted in multiple study centers or are stratified by variables such as age, sex, or other prognostic factors. In these cases, stratified analysis can be used to increase clinical trial (or assay) sensitivity and to reduce estimation bias.[103,104] An overall vaccine effect can be estimated as a weighted average across strata, and the weighting scheme may depend on whether the stratification factor is prognostic or nonprognostic, or whether treatment-by-stratification interaction exists.[105–108]

**Interim Analyses**

In large-scale vaccine efficacy trials, formal interim analyses for potential early stopping of the trial are often
planned. Interim monitoring of efficacy and safety is typically performed by an independent data monitoring committee. The FDA has recently developed a draft guidance on the establishment and the operation of these committees. Determination of early stopping because of overwhelming efficacy is usually guided by statistical stopping criteria such as those of O’Brien and Fleming and Lan and DeMets. Trial termination because of safety concerns occurs far less frequently in vaccine trials than in drug trials because participants receive only one or a limited number of vaccinations over a short period. Of course, futility analysis (for early stopping of the trial because of a low likelihood of success at study end) can be performed to minimize resource utilization.

**IMMUNE SURROGATE AND CORRELATE OF PROTECTION FOR VACCINES**

A critical task in vaccine development is identifying immune response markers (antibody or T-cell response) that can be used as surrogates for clinical vaccine efficacy so that more efficient evaluation of new vaccines or manufacturing process changes can be performed using the immune surrogates. Immune surrogates are best evaluated in a prospectively planned efficacy trial. The classic methods for validating surrogate endpoints may be used to assess the validity of an immune surrogate. In particular, these methods are applicable for vaccines aimed at increasing the immune responses of persons who may have preexisting immunity; in these situations, the validation criteria are similar to the validation of CD4 counts and viral RNA levels as surrogates for AIDS. However, these classic methods may break down in evaluating pediatric vaccines, which are targeted toward persons who are immune-naive (without preexisting immunity before vaccination). If a vaccine is highly immunogenic, the immune responses of vaccinated persons will clearly be separated from that of unvaccinated persons (who have no immunity). As a result, the immune response status will be completely confounded with the vaccination status, and it will be very difficult, if not impossible, to apply Prentice’s criteria to show that given an immune response, the probability of developing the disease is the same regardless of whether a person is vaccinated or not.

Because of the above-mentioned reasons, vaccine researchers tend to assess the relationship between immune responses and protective efficacy via the concept of “correlate of protection.” By examining vaccine failures, researchers aim to establish a “protective level” of an immune response that can be used to determine whether a person is “completely protected” or “not protected and still susceptible to the disease.” On a population basis, this “protected level” means that 100% of the persons who have immune responses at this level or higher will be completely protected from the disease.

In some instances, it is difficult to identify a clear-cut value of the immune response as the “protective level” because protective efficacy tends to increase with higher immune responses. For example, clinical trials with a live attenuated varicella (Oka/Merck) vaccine have shown a high protective efficacy against chickenpox and an inverse relationship between the frequency of chickenpox breakthroughs among vaccinees and the post-vaccination antibody titer measured by the glycoprotein enzyme-linked immunosorbent assay. However, no particular level of antibody can be considered a clear-cut “protective level.” Chan et al. have proposed the use of statistical models (such as accelerated failure time models) to link the whole distribution of antibody titers and long-term disease protection conferred by vaccination, and Li et al. have proposed the use of the concept of “approximate protective level” at which a high proportion (e.g., >95%) of individuals are expected to be protected from the disease. If feasible, the use of placebo group and different vaccine dosages in the same study will also enhance the assessment of the immune correlate. Once an immune marker is validated as a “correlate of protection,” vaccine efficacy can be subsequently evaluated through this immune marker. This will facilitate efficient evaluation of vaccine manufacturing process changes, vaccine consistency lots, concomitant vaccination, or new combination vaccines. Chan et al. have also proposed the use of statistical models to predict vaccine efficacy based on the immune responses collected in immunogenicity trials.

**ECONOMIC EVALUATION OF VACCINE**

There is increasing interest in conducting economic evaluations of medical interventions. Several textbooks review the study types and methods used to evaluate the cost-effectiveness of health care programs. Cost-effectiveness analysis identifies all relevant costs involved and evaluates the expected health gains derived in particular programs. The goal of economic evaluation is to maximize the health benefits per dollar spent. The usual measure is the cost-effectiveness ratio, defined as the ratio of the incremental costs of undertaking the health care program to the incremental health effects. Cost-effectiveness analysis is often used as an aid to national public health decision making.

Edmunds, Medley, and Nokes argued that while most cost-effectiveness analyses are based on the benefits to individual persons for vaccine programs, the population is the appropriate target for the mass immunization program. Therefore, the indirect effects (herd immunity) are important and should be included in the cost-effectiveness analysis. This approach measures the health effect of the vaccination program in terms of the proportion of the total population that will be protected after mass immunization. The effects are a result of both the efficacy of the vaccine
and a reduction in the number of infectious individuals in the population. Edmunds, Medley, and Nokes\textsuperscript{129} called this a dynamic model of an infectious disease and showed that the cost-effectiveness ratio seen with the use of this approach tends to decline over time.

Using the dynamic approach involves developing a population-based mathematical model of the spread of the infection. Halloran et al.\textsuperscript{130} developed such a model for routine varicella immunization of preschool children in the United States. Using newly available data, Brisson et al.\textsuperscript{131} updated Halloran et al.’s estimates. The Halloran model has been used as the basis for conducting cost-effectiveness analysis of varicella vaccinations in several countries.\textsuperscript{115}

**REFERENCES**

71. Ruger, B. The maximal significance level of the test which consists of rejecting H_0 when K of a given test leads to rejection (German). Metrika 1978, 25, 171–178.


