Assessing Correlates of Protection in Vaccine Trials

Janet Wittes
Statistics Collaborative, Washington, DC

(See the major article by Gilbert et al on pages 1573–81.)

**Keywords.** correlate of protection; surrogate outcome; principal stratification; subgroups; vaccine efficacy; Zostavax; herpes zoster.

Readers interested in knowing how well fold rise in antibody titers serves as a correlate of protection (CoP) for the herpes zoster (HZ) vaccine Zostavax will find a clear answer in the article by Gilbert et al [1]: fold rise is an excellent CoP. The real strength of the article, however, is its answer to a more general question: how can one identify when a biomarker is a useful CoP? Embedded sotto voce in that question is another, not-yet-addressed important practical problem that faces vaccine development: is there a way to learn enough about the likelihood that a vaccine will be effective to obviate the need for certain field trials? In particular, can one exploit knowledge gained through the use of 1 or more CoPs to infer vaccine efficacy? By extension, the problem is relevant to surrogate outcomes outside of vaccine development, for a CoP is just a surrogate in the vaccine world. Gilbert et al use “CoP” to refer to any variable that is “reliably predictive of outcome,” whether or not it serves as a mechanism of protection. That nomenclature, if used more generally, will clarify the literature. They advocate, when useful, denoting a mechanistic and a nonmechanistic CoP as “mCoP” and “nCoP,” respectively.

The suggestion that a surrogate can possibly stand in for a clinical outcome may send a chill down the backs of clinical trialists. I can hear the reader exclaim, “Have we not learned enough from trials involving hormone replacement in women, high-density lipoprotein levels in men and women, and antioxidants in smokers to beware of the surrogate route?” But Gilbert et al are leading us on a way different from the usual surrogate path—they are walking us gingerly through the principal stratification method of causal inference [2]. This approach differs so much from the typical analysis used in clinical trials that it is worth pausing to describe its basic philosophy.

Conventional strata in clinical trials are those that are defined at the time of randomization, such as age category, race, prior exposure to an organism of interest, or baseline severity. Restricting attention to baseline variables allows unbiased assessment of the effect of an intervention by strata. For example, one can ask whether Zostavax is effective in males or whether the effect varies by sex. Other subgroups, called “improper” by Yusuf [3], are defined after randomization, such as people who adhere to the protocol regime and those who do not. Because such postrandomization subgroups are inextricably tied to outcome, they cannot form the basis of unbiased inference concerning the effect of treatment. Gilbert et al discuss neither of these types of strata. They address, instead, principal stratification of posttreatment variables, a method that cross-classifies participants in trials by the potential values of those variables under each of the treatments being studied. Here the variables are titers or fold rise and the treatments are Zostavax and placebo. Because treatment assignment does not affect these variables, they can be used like any baseline variable.

Principal strata address conditions contrary to fact: they posit what participants in a trial would have experienced had they been assigned the study group to which they had not been randomized. One can view populations in clinical trials comparing a treatment to control as comprising 4 types of participants: those who would be likely to fail, regardless of whether they received the experimental or control treatment; those who would be likely to succeed, regardless of whether they received the experimental or control treatment; those who would be more likely to succeed if they received the experimental or control treatment; and those who would be more likely to succeed if they received the control treatment. Depending on the study, “failure” might be a clinical outcome (eg, death from a gram-negative bacterial infection in an antibiotic trial or having a stroke in a trial of an antithrombotic drug) or it might be, as in the Gilbert et al analysis, achieving a specific value on a surrogate outcome or on a marker.
The analysis by Gilbert et al of the Zostavax Efficacy and Safety Trial (ZEST) [4] considers 2 potential markers that might serve as CoPs: antibody titer as measured by glycoprotein-based enzyme-linked immunosorbent assay (gpELISA) at week 6 and fold rise in antibody titer between baseline and week 6. Each person in the ZEST potentially received either vaccine or placebo, and treatment for each would be characterized as a success or failure depending on outcome. Logically, each person can experience one of 16 possible configurations so of the 2 potential CoPs and could fall into one of 6 possible strata (Table 1). In the ZEST, the authors only consider the cases in which the vaccine can produce increases, not decreases, in fold rise and titer. Further, they do not consider the case in which a person has the same response category under both treatment assignment for fold rise but a “high” response for titer under vaccine assignment and a “low” response for titer under placebo assignment. Their categorization leads to 3 strata: no response, partial response, and high response. Omitting some of the logically possible configurations may be relevant in the ZEST setting; however, one can imagine situations in other drug-marker settings in which all 16 configurations are possible.

In the ZEST, baseline and week 6 serum specimens were collected from all 22,000 participants. In a substudy, the titters were to be measured in specimens from a random sample of 10% of the participants, as well as from all 129 confirmed cases of HZ that occurred during the trial. This substudy gave the authors tools to identify the principal strata with respect to fold rise and titer and, importantly, to see how well these markers predicted vaccine efficacy. The trick was to use the substudy to build models in which baseline variables predicted week 6 fold rise and antibody titer. The authors then applied their model to placebo recipients. The high correlation between day 1 and week 6 titers allowed this sleight of hand. The results are striking: the analysis using principal stratification shows nonsignificant vaccine efficacy (VE) in nonresponders, 83% VE in partial responders, and 96% VE in high responders (Figure 4 in the article by Gilbert et al [1] and Table 1).

The article leaves us with at least 3 questions. First, for those interested
specifically in Zostavax it leads to the question of whether the results are so strong that they permit limiting further studies of the vaccine to measurement of immune response. The authors state, “[T]he fold rise in gpELISA antibody titers can be used as a correlate of protection against HZ for Zostavax in the trial population of the ZEST, suggesting that additional field trials of ZV in the ZEST population are not needed” [1]. It is not clear why one would want to replicate ZEST, even without these results. The more interesting and not answerable question is whether one could move to a different population and make conclusions about protection from fold rise alone. The Shingles Prevention Study (SPS) [5] had shown the benefit of Zostavax in men and women aged >60 years. If one were interested in use of the vaccine in people aged <50 years or in countries with higher rates of chicken pox in the population than experienced in the populations of ZEST and SPS, it is an open question whether an analysis of fold rise alone could substitute for a field trial. Use of principal stratification allows identification of predictors; it does not by itself allow use of a CoP to replace a trial of actual outcomes.

For those interested in vaccine development in general, the article raises questions about the conditions under which one can profitably use this approach and about how Gilbert et al’s methods should influence designs of further studies. The antibody substudy in the ZEST, the elegant analyses the authors describe, and the clear results speak to the importance of routinely designing future vaccine trials with this same type of substudy. Vaccine trials are often require tens of thousands of participants. Judicious subsampling of these study populations can mimic the analysis by Gilbert et al in other settings, perhaps producing important insights into vaccine efficacy. The method will only work, however, when the incidence of disease is low in the population studied. For conditions like malaria, in which a high percentage of trial participants will contract the disease of interest, the approach cannot yield accurate estimates. Also, none of the discussion deals with safety. Much of the reason for the need for very large sample sizes in vaccine trials is not to show efficacy but to be assured of safety.

The set of methods Gilbert et al describe is likely to be relevant to surrogates outside of vaccines. An interesting exercise would be to reanalyze data from large trials by using principal stratification to explore the conditions under which their approach could yield clearer insights into the relationship between surrogates and clinical outcomes.

Note

Potential conflict of interest. J. W. is president of Statistics Collaborative, which has served as a consultant to and has reported fees from Merck, Sanaria.

The author has submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

References