Clinical studies to prevent the development of food allergy have recently helped reshape public policy recommendations on the early introduction of allergenic foods. These trials are also prompting new research, and it is therefore important to address the unique design and analysis challenges of prevention trials. We highlight statistical concepts and give recommendations that clinical researchers may wish to adopt when designing future study protocols and analysis plans for prevention studies. Topics include selecting a study sample, addressing internal and external validity, improving statistical power, choosing alpha and beta, analysis innovations to address dilution effects, and analysis methods to deal with poor compliance, dropout, and missing data.

The prevalence of food allergy has been on the rise over the last 30 years with 6% to 8% of children being affected worldwide.1,2 Currently, there is no cure for IgE-mediated food allergy and the main treatment remains avoidance; thus, understanding the cause and developing strategies for the prevention of allergy has been at the forefront of current allergy research. The past decade has seen an increase in trials aimed at the prevention of food allergy through early life nutritional interventions.3,4 These prevention trials, in contrast to therapeutic trials, apply to subjects at risk of developing a future food allergy and therefore tend to be drawn from an at-risk pediatric population.

Although prevention trials can lead to valuable public health recommendations (eg, childhood vaccination or early consumption of peanuts3), their design, implementation, and interpretation pose unique and significant challenges. Prevention trials often take longer to complete, show smaller treatment effects, and require larger numbers of participants than do studies designed to test a therapy on a preexisting illness. Because participants are ostensibly healthy, the risk-benefit ratio of aggressive intervention is often shifted toward safer, more conservative strategies. Conservative interventions can lead to smaller treatment effects and therefore require larger sample sizes to achieve adequate power. Moreover, a drug’s side effects are experienced by only the small number of people treated with the drug. Conversely, harmful effects resulting from a broadly applied public policy recommendation can eliminate the public utility of the intervention because adverse events will be experienced over a large portion of the population.

The data analysis of prevention studies can also present unique challenges. Because a large proportion of the study sample typically does not develop the disease of interest, these participants can dilute or add variability to the metrics used to evaluate safety and efficacy. Prevention trials are often longer in duration to coincide with the incidence of disease. Unfortunately, participants enrolled in lengthy studies tend to have higher rates of dropout and lower rates of compliance, especially if they perceive little or no immediate benefit. Nevertheless, the many challenges that exist with conducting and analyzing prevention trials can be addressed with appropriate study design features and statistical methodologies.

This article focuses on the statistical challenges of food allergy prevention studies. However, the concepts apply to the design and analysis of nearly all prevention trials and particularly to diseases with low prevalence. Examples are drawn from 2 recently published randomized controlled prevention trials: Learning Early About Peanuts (LEAP) and Enquiring About Tolerance (EAT). Briefly, the LEAP and EAT studies enrolled 640 and
1303 infants, and took 7 and 7.5 years to complete, respectively. The LEAP study participants were recruited from an at-risk population (severe eczema and/or egg allergy) and the EAT study participants were recruited from a general population of exclusively breast-fed infants. At completion of the trials, the peanut allergy prevalence in the control group was 17.2% in the LEAP study and 2.5% in the EAT study, and compliance with the intervention was 92% and 54%, respectively. Using these trials as the main examples, topics in the following areas of food allergy prevention studies will be addressed:

1. Study Design: enrollment criteria, external validity
2. Statistical Power: choice of alpha and beta and 1- or 2-sided hypothesis testing
3. Analysis innovations to address dilution effects
4. Analysis methods to deal with poor compliance, dropout, and missing data

**STUDY DESIGN**

**Whom to enroll?**

Determining whom to enroll for a prevention study involves additional challenges not typically present for a therapeutic trial. When testing a new drug or therapy, participants with the disease of interest need to be identified and enrolled. Conversely, in prevention trials, participants must be enrolled before the illness presents. If disease prevalence is low (e.g., peanut allergy at ~2%), a random sample from the overall population needs to be very large to provide sufficient power. Moreover, the large proportion of participants unaffected by an illness often perceives less immediate study benefit. Thus, they may be unmotivated to comply with an intervention and continue study participation. Poor compliance and dropout impair analysis interpretation by decreasing statistical power and producing results that lack internal or external validity. Therefore, selecting a high-risk population can offer key advantages.

**FIGURE 1.** Three sampling strategies (different shades of green) are shown from an overall population with a prevalence of peanut allergy of 2% (depicted by the 2 red squares out of a 100 green squares). The outer band represents a population-based study where all squares are randomly sampled to produce a representative sample with a 2% prevalence of peanut allergy. As the bands move inward, the selection criteria is more restrictive and the proportion with peanut allergy increases from 2% to 4% to 12.5%. The annotated P values and power levels (white blocks) are from Fisher exact tests between the simulated control and treatment groups using a 2-sided test of significance at alpha = 0.05. Table I and Figure 1 demonstrate a dramatic increase in statistical power with more focused selection criteria, despite the intervention effect (80%) and sample size (n = 1000) being held constant in each simulation.
effect from the study sample onto the overall population. Koplin et al. provide a recent example of this exercise using the LEAP and HealthNuts studies. However, this recommendation assumes that the intervention is effective in both high- and low-risk populations. If the pathophysiology is such that, for example, the high-risk group has progressed too far toward the development of the disease of interest, the intervention may be “too late” to prevent the onset of disease. In this scenario, the selection of a high-risk population may be detrimental to the intervention’s effectiveness and subgroup analyses may be specified to address this shortcoming. As described by Permutt et al., stratified randomization is not necessary when performing a stratified analysis; however, a priori subgroup specification is strongly recommended to control type I error. In summary, selecting a high-risk sample using inclusion and exclusion criteria with known population prevalence estimates can allow results to be extrapolated to a general population, thus improving the study’s external validity. However, this approach has limitations if the intervention works differently for high- and low-risk groups.

### STATISTICAL METHODS

#### Statistical powering — choice of alpha and beta

The current paradigm in both prevention and therapeutic trials is to set the false-positive rate ($\alpha$) 2 to 4 times lower than the false-negative rate ($\beta$). For example, most studies set an alpha of 0.05 and a beta of 0.1 or 0.2, which represents 90% or 80% power to reject a null hypothesis with a $P$ value of less than 0.05. National Institutes of Health studies have a long history of setting power at

*Table I. Simulation study parameters and results*

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Treatment effect</th>
<th>Prevalence in population</th>
<th>Risk group</th>
<th>Prevalence in study sample</th>
<th>Number allergic in the control group</th>
<th>Number allergic in the treated group</th>
<th>Risk difference</th>
<th>Power</th>
<th>Fisher exact test $P$ value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>80%</td>
<td>2%</td>
<td>High</td>
<td>12.5%</td>
<td>63</td>
<td>13</td>
<td>10.0%</td>
<td>&gt;99%</td>
<td>1.3 x 10^-19</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Moderate</td>
<td>4.1%</td>
<td>20</td>
<td>4</td>
<td>3.3%</td>
<td>91.0%</td>
<td>.0014</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Low</td>
<td>2.0%</td>
<td>10</td>
<td>2</td>
<td>1.6%</td>
<td>57.0%</td>
<td>.04</td>
</tr>
</tbody>
</table>

*FIGURE 2.* Medians (black lines) are not statistically different for peanut-specific Ara h2 in the box and violin plot (left) ($P > .05$ by Wilcoxon test). The means (black diamonds) show more of a difference in the opposite direction as the medians do because they are influenced by the very high titer levels. However, clear differences between the randomized groups exist in the upper ends of the distributions and this region corresponds to a high risk of allergy (filled red diamonds represent allergic participants). The proportion density plots (right) examine how the randomized groups are represented proportionally across the range of titer values. The black diamonds mark the overall mean distribution, which corresponds to a very low risk of peanut allergy. The proportion of the distribution encompassed by the avoidance group grows when moving to higher titer levels and significantly diverges ($P < .05$) at the dashed reference line.
90% for definitive studies. The principal precept of bioethics, "primum non nocere," or "first, do no harm," motivates the relatively lower type I (false-positive) versus type II (false-negative) error rates for treatment studies. For example, drug approval often operates under a conservative, precautionary principle when evaluating compounds as safe and effective. This tradeoff seems appropriate when the consequences of a spurious finding are harmful, for example, the introduction of a drug with known side effects. However, an unbalanced acceptance of type II error over type I error can have different consequences for prevention trials.

The practice of preventative medicine generally does not have harmful consequences. For example, the early introduction of peanut has been shown to be safe and nutritious in a high-risk population of infants.\textsuperscript{11} In these scenarios, the "do no harm" principle could instead be applied to type II error (false negative) because failing to overturn a harmful recommendation (eg, early avoidance of peanut during an immunological window of opportunity to induce tolerance) would be considered equally, if not more harmful than a spurious (false-positive) finding. A beta of 0.2 (80% power) leaves a 20% chance of not finding an effect, if one exists. The consequence for a false-negative outcome in a prevention study setting can have harmful public policy implications because failing to reject the null hypothesis results in the continuing implementation of harmful recommendations (eg, avoidance of peanut in early childhood).

According to Lilford and Johnson,\textsuperscript{12} "if the ‘costs’ (including all the side effects) of each treatment are the same, then alpha should equal beta, since either a false-positive or false-negative result would be equally undesirable.\textsuperscript{3,5} …if the specified improvement in outcome (delta) is sufficient to outweigh the known disadvantages of the more costly treatment, then a false-negative trial result is no longer preferable to a false-positive." Although the context is study dependent, generally, acceptance of a higher type II error rate in a prevention

\textbf{Instrumental Variable Analysis}

\textbf{Causal Diagram}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{causal_diagram.png}
\caption{Instrumental variable analysis causal diagram.}
\end{figure}
trial can have worse consequences than in a therapeutic trial. Similarly, spurious type I errors in prevention trials are generally more desirable than in therapeutic trials. Scientists planning studies to overturn potentially harmful health practices should therefore consider designs where the false-negative (type II) error rate is just as important as the false-positive (type I) error rate.

1-sided or 2-sided hypothesis testing

It is most common to use a 2-sided hypothesis test for both prevention and therapeutic trials. A 2-sided test allot half of the alpha (typically 0.05) to testing statistical significance in one direction and half of the alpha to testing statistical significance in the other direction. This equates to testing simultaneously whether a treatment or intervention is better or worse than a control. Because treatments can unexpectedly cause harm or be significantly less efficacious than a control, 2-sided tests are often justified. However, there are circumstances in which 1-sided tests are appropriate. Greenland et al provide a guide to common statistical misinterpretations, one of which is that researchers should always use 2-sided $P$ values; they should not. The construction of a hypothesis test should match the hypothesis. For example, a 1-sided test is appropriate if researchers are only interested in a difference that goes in one direction, and, hypothetically, if a large difference was observed in the unexpected direction, they would take the same action as if no difference at all was observed. Many authors have given justification and examples where 1-sided tests are appropriate including toxicity studies, safety data monitoring, testing whether a new drug or intervention is at least as good as one currently in use, and laboratory studies where biological constraints limit a difference in one direction. For example, if researchers were interested in testing whether a new intervention was "at least as good" as the LEAP intervention (eg, smaller amounts of recommended daily consumption of peanut), this could appropriately be designed using a 1-sided test. Importantly, Koch and Gillings have shown that under typical conditions of hypothesis testing, the required sample size for 80% and 90% power is increased by 27% and 23%, respectively, for a 2-sided test. As discussed earlier, prevention studies typically require larger sample sizes and longer durations. Thus, unnecessarily increasing the sample size to use alpha in a direction of no clinical importance comes with large cost and value implications and should not be practiced simply because 2-sided tests are conventionally used.

### ANALYSIS INNOVATIONS TO ADDRESS DILUTION EFFECTS

The analysis shown in Figure 1 demonstrates the decrease in statistical power that occurs when a study sample has a high proportion of healthy participants—this is known as a dilution effect. During the design phase of a trial, dilution effects are addressed with sample size calculations. However, sample sizes are typically based on a comparison of the primary end point (eg, the difference in the proportion with allergy between randomized groups), and more variable secondary end points (eg, specific IgE titer levels) can often be underpowered to detect clinically meaningful effects. Nevertheless, a gain in statistical efficiency can be obtained by selecting more powerful analysis methodologies. For example, Aban et al have demonstrated increased efficiency using the Poisson or negative binomial distribution instead of commonly used methods such as the Wilcoxon test or the $t$ test. Moreover, statistical innovations in vaccine trials (another type of prevention study) have recently been developed to provide more powerful methods to address dilution effects. These testing procedures compare the randomized groups by essentially removing an equal number of zeros (or patients with undetectable titer levels) from each group before performing a permutation test on the remaining, less-diluted sample.

Last, an approach used in the LEAP study secondary analyses was to examine the upper end of the peanut-specific IgE distribution, rather than the means or medians. The LEAP study evaluated a high-risk cohort and after 5 years of follow-up, 17% developed peanut allergy in the Avoidance (control) group compared with 3% in the Consumption (intervention) group. Therefore, if peanut consumption was responsible for modulating IgE production, it might have done so in only 14% (17% minus 3%) of those in the Consumption group who would have developed allergy, had they not consumed peanut. In other words, the testable assumption was that if immunological differences exist in IgE between the randomized groups, these differences could be diluted by the 86% who were "destined" not to be allergic. Moreover, the rationale was that these differences would not be apparent when examining the means or medians of the overall distribution. Therefore, the upper end of the IgE distribution was considered more relevant because this region corresponds to levels more highly predictive of peanut allergy. Bootstrap sampling was used to determine if and where the upper percentiles of the IgE distribution significantly differed from the null distribution.

### TABLE II. Complier average causal effect (CACE) model

<table>
<thead>
<tr>
<th>Status</th>
<th>Compliance rate</th>
<th>Treatment</th>
<th></th>
<th></th>
<th></th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Symbol</td>
<td>No. allergic</td>
<td>n</td>
<td>Proportion</td>
<td>Symbol</td>
<td>No. allergic</td>
<td>n</td>
</tr>
<tr>
<td>Compliant</td>
<td>54%</td>
<td>$C_{1}$</td>
<td>0</td>
<td>310</td>
<td>0.00%</td>
<td>$C_{c}$</td>
</tr>
<tr>
<td>Not compliant</td>
<td>46%</td>
<td>$N_{1}$</td>
<td>7</td>
<td>261</td>
<td>2.68%</td>
<td>$N_{c}$</td>
</tr>
<tr>
<td>Overall outcome</td>
<td>(ITT row)</td>
<td>T</td>
<td>7</td>
<td>571</td>
<td>1.23%</td>
<td>C</td>
</tr>
</tbody>
</table>

The CACE is a measure of the causal effect of the intervention on the people who received it as intended by the original group allocation. The model assumes that the difference between the 2 bolded allergy rates above (2.68% and 2.51%) are due to confounding factors “U” (Figure 3) that affect compliance (X) and allergy (Y). These confounding effects are “pulled out” to create the adjusted allergy rate (2.47%) in the theoretical compliant stratum of the control arm ($C_{c}$).

Assumptions: Members of the control group have the same probability of being a “noncomplier” as do members of the intervention group. Being offered early introduction of peanut (as opposed to actually eating) has no effect on the outcome.

*Calculated rather than observed values using the CACE approach.

### TABLE III. Comparison of treatment effects (relative risks and risk differences) by type of analysis

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Calculation</th>
<th>Relative risk</th>
<th>Risk difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITT</td>
<td>$T/C$</td>
<td>48.8%</td>
<td>−1.29%</td>
</tr>
<tr>
<td>PP</td>
<td>$C_{i}/C$</td>
<td>0.0%</td>
<td>−2.51%</td>
</tr>
<tr>
<td>CACE&quot;</td>
<td>$C/C_{c}$/&quot;</td>
<td>0.0%&quot;</td>
<td>−2.47%&quot;</td>
</tr>
</tbody>
</table>

*Calculated rather than observed values using the CACE approach.
between the randomized groups. A progressive divergence in very high peanut-specific IgE titers was detected and displayed using proportion density plots. Reference lines were drawn to indicate where the distributions began to differ at the alpha \( \alpha = 0.05 \) level of significance (Figure 2). These types of analyses underscore the utility of implementing unique statistical methods for prevention studies, as they are useful for identifying the consequences of an intervention when dilution effects are present.

**ANALYSIS METHODS TO DEAL WITH POOR COMPLIANCE, DROPOUT, AND MISSING DATA**

**Adjusting for noncompliance using instrumental variable analysis**

The literature describing analysis methods to address poor compliance, dropout, and missing data is vast and context-specific. The objective is not to provide a review, but rather to present a type of instrumental variable analysis that can be used to estimate a nonbiased intervention effect for food allergy prevention trials, while controlling for noncompliance. In addition, relevant imputation methods and sensitivity analyses for binary outcomes will be recommended.

Most clinical trials specify the intention-to-treat (ITT) analysis as primary but also include a supporting, per-protocol (PP) analysis. Each analysis has its advantages and disadvantages. Generally, the ITT approach answers the question whether the offer and initial agreement of treatment to the intervention population is effective. To more directly address if receiving the treatment is effective, the PP analysis is often recommended. However, with respect to a randomized comparison, the PP approach can be biased unless the probability of taking the treatment is random with respect to all predictors of the study’s outcome. The ITT effect is thought to be an overly conservative measure of an intervention’s effect when compliance is poor, as participants who received little or no intervention are accounted for in the intervention group. The PP analysis, although biased, is desirable because it directly measures the effect of treatment among the participants who received the intervention. Instrumental variable analyses in general and the complier average causal effect (CACE) model in particular have been recently proposed as nonbiased methods to estimate the causal effect of treatment.22-28 Figure 3 depicts a causal diagram typically used to illustrate instrumental variable methods, and Tables II and III demonstrates how the CACE model estimates an unbiased intervention effect using the EAT study data.

This reanalysis of the EAT study data uses summary allergy and compliance rates from the New England Journal of Medicine figures and limits PP compliance to only the Early Introduction group (intervention arm). Although there were also noncompliers in the Standard Introduction group (control arm), the CACE analysis is focused on compliance in the intervention group only. Under the assumption that randomization balances all factors, the CACE method projects the noncompliant rate in the intervention arm onto the control arm. By comparing the rate of allergy in the noncompliant, intervention arm (2.68%) with that...
in the overall control arm (2.51%), an estimate of confounding (symbolized by U in Figure 3) can be obtained. This confounding effect is removed from the PP CACE intervention effect. Typically, the CACE effect is between the ITT and PP estimates. In the case of the EAT study, the unbiased CACE estimate (risk difference of 2.47%) is nearly equal to the PP effect (2.51%). The similarity of the PP and CACE effect is a result of the noncompliant allergy rate in the intervention arm being approximately equal to the allergy rate in the control arm (ie, confounding factors [U] are not highly associated with allergy [Y]). Last, Jo29 has shown using Monte-Carlo simulations that the CACE model is just as powerful, if not more powerful than the ITT analysis under many scenarios.

The CACE model is relevant to food allergy trials because compliance with the intervention is typically defined as a successful outcome (ie, tolerating the food allergen of interest). For example, the PP analysis in this setting often excludes allergic participants only from the intervention group who cannot comply with continued consumption of the food. Participants may be noncompliant for many reasons, but there always exists the belief or criticism that the PP analysis is not credible because of reverse causality (ie, the participants are noncompliant because they are allergic or are becoming allergic; hence, they are unable to eat the recommended food allergen). Because the PP analysis excludes only these hypothetical participants from the intervention group, it may produce a biased estimate of the intervention’s effect. The appeal of the CACE model is that it adjusts for this apparent bias by introducing the following counterfactual: if those randomized to the control group had been randomized to the intervention group, they would have had an equal rate of noncompliance compared with the observed noncompliance in the intervention group. Moreover, this hypothetical noncompliant group would have an allergy rate equal to that of the

FIGURE 5. Diagnostic algorithm for determination of peanut allergy in the absence of peanut-challenge results using dietary and reaction history, SPT, and peanut-specific IgE. An oral food challenge was used for the assessment of peanut allergy for 96% (617 of 640) of the participants; this diagnostic algorithm was therefore required for 13 study participants whose outcomes were as follows: 7 were peanut allergic, 4 were peanut tolerant, and 2 were nonevaluable. SPT, Skin prick test.
Sensitivity analyses for missing data

Missing data occur for many different reasons and can have drastic consequences on the interpretation of a study’s primary analysis. For example, if allergic participants in the intervention arm discontinue a study for fear of having an oral food challenge, this will clearly bias results by overestimating the effectiveness of the intervention. Conversely, if participants are not allergic and drop out because they perceive no continued benefit, the intervention effect will be underestimated. It is a misconception that a conservative and appropriate method to overcome this bias is simply to impute an allergic result to all missing outcomes. In prevention studies, the rate of allergy can be less common than the rate of dropout; therefore, allergic imputation can produce more bias than simply analyzing the nonimputed, complete-case data set. To properly address missing data, sensitivity analyses should be undertaken to examine the reason for dropout and to develop imputation methods for replacing missing values with probable outcomes. Generally, for a prevention study, the most probable outcome is a nonallergic participant; however, for a treatment study, the most probable outcome is an allergic participant. Figure 4 uses simulated data to demonstrate how simply imputing an allergic result to all missing outcomes can introduce progressively more bias when allergy rates are low and dropout is high. This simulation experiment illustrates how complete case analysis (no imputation) is preferred because it produces less bias than “conservative” allergic imputation.

As an a priori method to address missing data, the LEAP study used a diagnostic algorithm to impute outcomes for participants who refused an oral food challenge (Figure 5). Worst-case imputation was used to show that the study’s findings were robust to missing data. However, unlike the LEAP study, many prevention trials have small treatment effects, a low prevalence of disease, and higher than average dropout rates, with the consequence that a worst-case imputation analysis will present an unrealistically low estimate of an intervention’s effectiveness. Although the specification of analysis methods is study specific, for prevention studies, scientists should consider developing algorithms (a priori) to impute missing data when possible. For missing data that cannot reasonably be imputed, sensitivity analyses should be used to support the primary, complete-case ITT analysis. These sensitivity analyses will assess the robustness of a study’s findings with respect to missing data without introducing assumptions and bias into the primary ITT analysis. Last, Liublinska and Rubin’s enhanced tipping-point displays are recommended for binary data because they help visualize sensitivity analyses via an easy to comprehend heat map. Figure 6 shows an example tipping point analysis derived from data presented in Figure 4 part B of the Perkin et al study. The heat maps show results from sensitivity analyses from all possible combinations of imputed missing data (panel B) along with the more probable imputation scenarios (panel A).

DISCUSSION

The recent rise in food allergy and current lack of therapies has prompted an increase in clinical research studies aimed at the prevention of food allergy. Ideally, interventions aimed at preventing food allergy in a general population would be trialed using adequately powered population-based samples. However,
logistical and financial constraints usually mandate that smaller, high-risk cohorts be enrolled. Some of the presented recommendations, for example, a lower beta, will result in even larger sample sizes. However, reducing the chances of a false-negative trial may be justified when the study hypothesis aims to overturn a harmful medical recommendation (eg, the early avoidance of allergenic foods). In addition, several suggestions are made to increase statistical power, minimize the introduction of bias, and better estimate intervention effects. Together these recommendations may offset the costs associated with lowering type II error. Finally, CACE models, imputation methods, and sensitivity analyses are suggested as supplements to a primary complete-case ITT analysis.

Acknowledgments

Research reported in this publication was performed as a project of the Immune Tolerance Network, an international clinical research consortium headquartered at the Benaroya Research Institute and supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under Award Number U54AI109565. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

We thank Gerald Nepom, Amber Franz, Helen Fisher, and Cate Speake for their input reviewing the manuscript. Last, we thank Agustin Calatroni for introducing us to the “Chumplump” R package and his suggestion to bootstrap sample the upper percentiles of a continuous distribution in LEAP study analyses.

REFERENCES