Published in final edited form as:

Stat Med. 2013 January 30; 32(2): 196–205. doi:10.1002/sim.5497.

Null but Not Void: Considerations for Hypothesis Testing

Running Head: Null but Not Void

Pamela A. Shaw, Ph.D.¹ and Michael A. Proschan, Ph.D.¹

¹Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, Bethesda, Maryland, USA

Abstract

Standard statistical theory teaches us that once the null and alternative hypotheses have been defined for a parameter, the choice of the statistical test is clear. Standard theory does not teach us how to choose the null or alternative hypothesis appropriate to the scientific question of interest. Neither does it tell us that in some cases, depending on which alternatives are realistic, we may want to define our null hypothesis differently. Problems in statistical practice are frequently not as pristinely summarized as the classic theory in our textbooks. In this article, we present examples in statistical hypothesis testing in which seemingly simple choices are in fact rich with nuance that, when given full consideration, make the choice of the right hypothesis test much less straightforward.

Keywords

Binomial proportion; hypothesis testing; Lachenbruch test; mixed models; repeated measures; strong null hypothesis

1 Introduction

Much of what we learned in statistics courses began with, “Let X_i be independent and identically distributed (*iid*) random variables...” How often have we heard instead “Let X_i be *iid* if the null hypothesis is true, but dependent if the alternative hypothesis is true?” That is exactly what happens in many cases, as illustrated by our first example. In this work, we consider examples where a mismatch can easily occur between the scientific question of interest and the chosen statistical test. Part of the problem is a tendency to reduce scientific questions to a single test of a one-dimensional quantity like a mean or proportion. Physical phenomena, such as patient response to treatment, are often more complicated than that. The variance of an outcome is frequently considered a nuisance parameter, but increased variance might capture a meaningful aspect of the impact of the experimental condition under study. For instance, increased variance in the treatment arm of a trial could be due to an important sub-population responding, or not responding, to treatment.

An important principle demonstrated by examples presented here is that our beliefs about what is realistic under the alternative hypothesis should dictate the appropriate tests. For example, consider a therapy aimed at limiting the progression of atherosclerosis. Should the parameter capturing the treatment effect summarize the change in blockage in arteries that were partially blocked at baseline, prevention of new blockages, or number of blockages over a certain size? Depending on the mechanism of action of the drug, and what is most clinically meaningful in terms of important patient outcomes, any one of these could be of interest.

General principles of hypothesis testing can be found in any introductory statistics textbook and include the basics of defining the null and alternative hypotheses for parameters of interest and the associated type I and type II errors [1, 2]. There are also many well written textbooks dedicated to the elegant statistical theory behind hypothesis testing [3, 4]. Generally the focus of these texts is on understanding the mathematical properties of the statistical test, with less attention on the interaction of the scientific problem at hand and the construction of the hypothesis test. Books that focus on the translation of science into statistics are frequently primers written for a non-statistical audience and are not focused on more subtle nuances in the statistical issues involved in hypothesis testing.

In this paper we present a series of examples where choice of the null hypothesis is not so clear, and how nuances to the scientific question could fundamentally change how or what we test. In the first example we see how even with simple binary data, there can be a multitude of choices when translating the scientific question into a statistical test. We then consider a classic problem in continuous repeated measures data to see how inadvertent choices about the hypothesis we are testing can result from applying statistical software without carefully examining the underlying assumptions being made. Finally, we examine a series of examples for which the uncertain nature of the treatment effect leads to no single choice of parameter to test. We conclude with a discussion of some overarching themes which will be useful to consider in practice.

2 Example 1: Binary Data: the Simple Case?

Nine-year-old Emily Rosa designed and conducted a fourth grade science project that eventually netted her a television appearance and a paper in the Journal of the American Medical Association [5]. The project concerned therapeutic touch (TT), a practice by which the practitioner attempts to treat patients for a variety of medical conditions by placing and moving their hands just above the patient and “manipulating the human energy field” of the patient. Rosa *et al.* [5] examined whether TT practitioners could perceive a “human energy field” by measuring the following surrogate endpoint. Each TT practitioner sat behind a blind containing two holes through which they outstretched their hands. Rosa placed one of her own hands above one of the TT practitioner’s, chosen at random, and asked the practitioner to guess which hand her hand was placed over. She repeated this 10 times for each TT practitioner. The rationale was that a TT practitioner could not have the ability to manipulate the human energy field if they could not even detect such a field in Rosa’s hand (and thereby determine which of the practitioner’s hands Rosa’s hand was over). Rosa did two similar experiments over a period of a year.

That such a simple experiment has hidden complexities becomes clear when we try to formulate appropriate hypotheses. We must think about the experiment from the point of view of both a skeptic and a believer. To a skeptic, the practitioner is purely guessing; outcomes of guesses are Bernoulli random variables X_i with success probability $p = 1/2$. There is no distinction between two outcomes for a given practitioner and one outcome from each of two different practitioners. On the other hand, a believer would be confident that the success probability exceeds $1/2$, but might also speculate that different practitioners have different p s. A very good practitioner might have success probability 0.9 , whereas someone with less skill might have $p = 0.6$. Conditioned on the practitioner-specific p , the 10 observations are independent, but when the totality of outcomes of different practitioners are amassed, they are not. The set of all Bernoulli random variables from different practitioners and different repetitions is independent under a skeptic's null hypothesis, but dependent under a believer's alternative hypothesis. The believer might treat the practitioner as the unit of analysis and use as outcome Y , the number of correct guesses out of 10. The null hypotheses expressed in terms of the means p and μ of X and Y are

- H_{0X} : The X s are *iid* Bernoulli with $p = 1/2$
- H_{0Y} : The Y s are *iid* with mean $\mu = 5$.

The first null hypothesis implies the second and is therefore stronger. It is theoretically possible to have results very inconsistent with the skeptic's view of *iid* Bernoulli $(1/2)$ random variables, but consistent with H_{0Y} . For instance, suppose that half of the practitioners got 0 correct and half got 10 correct. This would offer no evidence against H_{0X} , but strong evidence against *iid* Bernoulli $(1/2)$ random variables. We will return to this point later. For now we note that natural test statistics for H_{0X} and H_{0Y} , respectively, are

$$Z = \frac{\hat{p} - .5}{\sqrt{\frac{(.5)(1-.5)}{10n}}} \text{ and } T = \frac{\bar{Y} - 5}{\sqrt{S_Y^2/n}}, \quad (1)$$

where n is the number of TT practitioners and s_y^2 is the sample variance for Y . The Z statistic is approximately standard normal (though an exact binomial test can also be used), while T has an approximate t -distribution with $n - 1$ degrees of freedom if the Y_i are not too skewed. Rosa *et al.* [5] used T and the one-tailed alternative $\mu > 5$.

The first experiment with 15 practitioners yielded $T = -0.7174$, one-tailed $p = 0.76$. The follow-up experiment with 13 TT practitioners, 7 of whom were also in the first experiment, resulted in $T = -2.222$, one-tailed $p = 0.98$. Rosa *et al.* declared that there was no evidence supporting the ability of TT practitioners to discern the correct hand.

In an author's response to letters to the editor, Rosa *et al.* [6] reiterated their deliberate choice of a t -test over a binomial test. A t -test has some appeal on robustness grounds in case of unforeseen variability resulting from subtle flaws in the experiment. For instance, there might be a cooling effect of air movement of the experimenter's hand being placed suddenly over the subject's hand. This could lead to too many or too few correct guesses depending on how it is interpreted (e.g., a practitioner who expects to sense a warm hand, but instead senses cool air, might assume there is no hand). While the authors mentioned an

attempt to check for evidence that such cues existed prior to the start of their main experiment, this check was done informally. Another possibility is that some TT practitioners misunderstood the instructions and thought they were supposed to guess which of the experimenter's hands was above the practitioner's. Depending on the prevalence of the different problems, there could be either systematically too many or too few correct guesses, or a clumping effect whereby a higher than expected proportion of participants get extreme results, some getting too many and others getting too few correct. Flaws are causing such consequences, one must question the entire premise of the experiment. A null result caused by flaws in the design or execution of the experiment would not be convincing evidence that the critics are correct. Validation of the critic's viewpoint requires demonstration of data consistent with H_{0X} . The robustness of the t -test is actually a disadvantage because person-to-person variability that should shake the foundation on which the experiment is based actually makes it easier to corroborate the critic's conclusion.

One can formulate specific hypothesis tests to determine whether the data may support the concerns of the preceding paragraph. The two-tailed t -test for H_{0Y} , with p -value = 0.04 in the second experiment, reveals that the TT practitioners did substantially worse than would be expected by chance. A two-tailed binomial test for the second experiment also yields

$Z = \frac{(53 - 65)}{\sqrt{130/4}} = -2.105$, $p = 0.04$. The statistically significant result for the binomial test suggests that the outcomes of guesses are either *iid* $\text{Ber}(p)$ with $p < 0.5$ (systematically poor guessing), or are not all *iid* with the same p (there is a clumping effect). The fact that the two-tailed t -test was also statistically significant corroborates the systematically poor guessing explanation. After all, the t -test uses an empirical variance estimate that is valid whether or not there is clumping. Had only the z -statistic been statistically significant, one would then have suspected the clumping explanation.

Further testing specifically designed to detect clumping supports the conclusion that it is not present. Specifically, if we concede that $p = 1/2$, we can test whether different people have different p s. The likelihood ratio test of a common p versus participant-specific p s rejects for large values of

$$W = \prod_{i=1}^n (y_i/10)^{y_i} (1 - y_i/10)^{10-y_i}.$$

The null distribution of W , i.e., its distribution under the hypothesis that the individual Bernoullis are *iid* p (not necessarily $1/2$), depends on p . To circumvent this problem, we conditioned on the total number $M = \sum_{i=1}^n Y_i$ of correct responses across all participants. The distribution of $Y_1, \dots, Y_n$, conditioned on M , is that of a multinomial with M total balls and 28 cells, each with probability $1/28$, conditioned on no cell having more than 10 successes. We computed a p -value by repeatedly simulating from this multinomial distribution, discarding simulated multinomials with more than 10 successes in any cell, and determining whether the simulated W value exceeded that of the actual data. The p -value associated with this test of clumping was $p = 0.39$. The lack of evidence for clumping

provides support that it was systematic bias rather than clumping driving the rejection of H_{0X} .

One aspect of the experiment that we have implicitly glossed over is the fact that 7 TT practitioners were in both experiments. One could argue that this should not cause a problem, in terms of an inflated type I error rate, for any of the tests if the strong null hypothesis H_{0X} is true. Under H_{0X} , all of the Y_i are independent, whether or not two or more come from the same participant. This is not true under a weaker null hypothesis that includes person-person variability, but we maintain that H_{0X} is the real null hypothesis of interest. Under an alternative hypothesis, the data may be dependent, in which case using some of the same practitioners in both experiments may be problematic. For instance, suppose that participants who did poorly more often chose to repeat the experiment in order to “redeem themselves.” If those who did poorly in the first experiment also did poorly in the second experiment, that would suggest that they were detecting some cues and interpreting them the opposite of the way they should have. The fact that the results of the second experiment were statistically significantly worse than chance may be an indication that this is what happened.

In summary, although robustness is often viewed as an advantage in statistical analyses, it is a disadvantage in this experiment. In choosing the t -statistic over the z -statistic, one essentially admits that there may be some flaws in the experiment against which one wants to hedge his/her bet. Such flaws could lead to either clumping or systematic effects. The one-tailed t -test does not allow the ability to test for systematically getting too few correct, and would likely be nonsignificant if there were substantial clumping. In actuality, there appears to be no clumping, but systematic under-achievement that may suggest a problem with the experiment or an unexpected finding related to TT.

3 Example 2: Repeated Measures

This example is adapted from an investigation of the proportion of naive CD4 T-cells expressing the cell surface marker CD31 in blood samples from HIV patients [7]. In practice, blood samples are often frozen and stored to be measured later. The purpose of this study was to see whether this process of freezing and thawing blood might actually change the results. Two potential problems are:

- P1: There is a systematic effect of freezing on the proportion of naive CD4 T-cells expressing CD31 (e.g., consistent underestimation).
- P2: Freezing causes underestimation in some patients and overestimation in others, leading to increased variability.

P1 and P2 can occur together. P1 is usually the more serious problem because it leads to an incorrect answer even on a mean level. That is, it leads to an incorrect answer for the mean of a group of people irrespective of group size. P2 in the absence of P1 is less serious because it adds noise rather than systematic bias. We focus on tests of P1, contrasting those whose validity does or does not depend on P2 being absent.

Let Y_0 be the proportions of naive CD4 T-cells expressing CD31 when the sample is analyzed right away (time 0), and Y_1 be the proportion when blood for the same patient is frozen and then thawed and analyzed one day later (time 1). Two possible null hypotheses are:

- H_{01} : $E(Y_0) = E(Y_1)$ (P1 absent).
- H_{02} : Y_0 and Y_1 have identical distributions (P1 absent, P2 absent).

Of course H_{02} implies H_{01} . Under the simplifying assumption of normality made throughout this example, H_{02} is equivalent to the means and variances being the same at times 0 and 1.

3.1 One Measurement Per Time Point

Imagine first that there is one observation per time point for each of n people. Let $D = Y_1 - Y_0$ be a paired difference between time 1 and time 0 for a given patient.

A natural test of problem P1 (hypothesis H_{01}) uses the paired t -statistic. This is equivalent to using a mixed model with time as a fixed effect and patient as a random effect, and testing whether the time effect is 0. This procedure has type I error rate α under H_{01} , whether or not the stronger hypothesis H_{02} holds. If the variance is increased at time 1 relative to time 0, that will be reflected in the sample variance of differences used in the paired t -statistic. The applicability of the paired t -test under either null hypothesis is clear, but the selection of the equivalent mixed model is much less obvious. The paired t -statistic is easy to understand and ideal for assessing whether P1 is present without requiring the additional assumption that P2 is absent.

3.2 Two Measurements Per Time Point

Now imagine a more thorough design placing blood from each individual into four test tubes, two analyzed immediately (time 0) and the other two frozen and analyzed 1 day later (time 1). The two tubes at each time point allow the assessment of variability of the measurement process. Ironically, the better design made the testing problem harder because inclusion of more than one observation per time point can lead to an analytical trap, as seen below.

A new potential problem becomes observable with a repeat measurement at each time point:

- P3: Freezing increases the inherent variability in the assay.

H_{02} (P1 absent, P2 absent) is now a weak null and P3 becomes part of a new strong null hypothesis:

- H_{03} : The four observations on a given patient are exchangeable (P1 absent, P2 absent, P3 absent).

With so many repeated measures, a natural tendency is to consider a mixed model. With one observation per time point, the paired t -test is equivalent to using a mixed model with time as a fixed effect and patient as a random effect, and testing whether the effect of time is 0. This model remains attractive because of its parsimony, focusing on the parameter of interest, a shift in means between time points, and incorporating a random intercept to

account for the within-person correlation. Then Y_{ijk} , the k th observation during time j ($j = 0$ is day 0 $j = 1$ is day 1; $k = 1; 2$) follows:

$$Y_{ijk} = \mu + \tau j + b_i + \varepsilon_{ijk}, \quad (2)$$

where μ is the overall mean, τ is the difference in means between time 1 and time 0, and the ε_{ijk} are *iid* random errors independent of the person-specific random effect b_i . But notice that $\tau = 0$ implies the strong null hypothesis H_{03} . Therefore, although the type I error rate for rejecting $\tau = 0$ is guaranteed to be controlled under the strong null hypothesis H_{03} , it is not necessarily controlled under weaker null hypotheses like H_{01} or even H_{02} .

A simple approach is as follows. Once again, we compute a paired t -statistic, with the effect of freezing/thawing for a given patient calculated as the difference D between the average of the two day-1 measurements and the average of the two day-0 measurements. We compute $\bar{D}$, the average paired difference over patients and compare

$$t = \frac{\bar{D}}{\sqrt{s_D^2/n}}, \quad (3)$$

where s_D^2 is the sample variance of D , to a t -distribution with $n - 1$ degrees of freedom. As with one observation per time point, the paired t -test controls the type I error rate under the weaker null hypothesis H_{01} , and is therefore ideal for testing whether P1 is present without requiring the assumption that P2 or P3 be absent.

Now suppose that the the strong null hypothesis H_{03} is false but the weaker null hypothesis H_{02} is true. Freezing/thawing might increase the number of CD4 T-cells expressing CD31 in some patients, and decrease it in others. In that case a more appropriate model than (2) might be

$$Y_{ijk} = \mu + \tau j + b_{0i} + b_{1i}j + \varepsilon_{ijk}, \quad (4)$$

where b_{0i} and b_{1i} are the time-specific random effects. Under this model, differences between time 0 and 1 would vary from patient to patient even if there were no measurement variability (i.e., even if $\sigma_\varepsilon^2 = 0$); the variance of $D_i = Y_{i1} - Y_{i0}$ is now $\sigma_\varepsilon^2 + \sigma_{b_1}^2 > \sigma_\varepsilon^2$. Unless $\sigma_{b_1}^2 = 0$ model (4) rules out H_{03} , although a fourth weak null $H_{04} = (\text{P1 absent, P3 absent})$ may still hold. What this means is, if model (2) is fit and $\tau = 0$ is rejected, then one cannot without further testing conclude a shift in the mean is the underlying cause of statistical significance (see appendix for details). The conscientious statistician would have likely examined residual plots or otherwise checked for heterogeneity, and hence considered the validity of model (2). However, a failure to detect heterogeneity does not guarantee that model (2) is correct. The paired t -test, on the other hand, is a valid test under both models (2) and (4).

One interesting conclusion is that the simplest analyses actually correspond to the most complicated models. The fully robust paired t -test, which is valid without stronger assumptions like H_{02} or H_{03} , is associated with the least parsimonious mixed model. Again

the paired t perspective allows us to see that it is robust against unequal variances at the two time points, a conclusion that is more opaque when we consider the equivalent mixed model. There is a trend toward preferring sophisticated “black box” methods, such as mixed models, over simpler approaches like t -tests. The underlying test-statistic is less evident for a mixed model, which in turn makes it more difficult to appreciate the exact null hypothesis being tested. At face value model (2) could appear to be a more obvious choice for someone interested in a change in the mean and familiar with Occam’s razor.

Misinterpretations of mixed models are especially likely when analyses are done by nonstatisticians, but even statisticians could benefit from trying to think about exactly what their statistical software is doing. For instance, we recall an incident in which, for an analysis of a multicenter clinical trial, a statistician argued that the model for the treatment effect must include a random effect of center, but he did not include a treatment by center interaction. He was surprised to find that it made virtually no difference for the test of treatment effect whether center was treated as fixed or random. But this is not surprising at all because the center effect drops out when we compute the treatment effect within each center. A careful examination of the test statistic would have revealed that the inclusion of a random intercept alone would not capture heterogeneity in the treatment effect across centers. Verbeke and Molenberghs (2000) identify other situations where the naive application of mixed models can lead to inappropriate statistical tests [8]. Two such examples are the inappropriate use of familiar maximum likelihood tests for comparing apparent nested models when REML (frequently the default in statistical software) is used to fit parameter estimates, and inappropriately using a standard z -statistic to test the null hypothesis that all random effects are 0. These authors highlight that the proper test for the latter scenario depends on whether one views the mixed model as arising from a marginal or nested hierarchical model, something probably not in the minds of many individuals implementing these models.

While much of this discussion can be seen as the classic tradeoff between robustness and efficiency, it also highlights a second, often unintentional, tradeoff, the one between transparency and sophistication. It is always advisable to do simple analyses to corroborate results of more complicated ones; when there is a conflict, trust the simpler analyses or be prepared to roll up your sleeves for a deeper investigation.

4 Example 3: When Treatment Affects Who or What is Measured

If the mechanism of action for the treatment at hand is multidimensional, as is likely with complex diseases like HIV or cardiovascular disease, one dimensional summaries may be insufficient. In fact, whether and how the treatment works could even change what should be measured, as the following examples show.

4.1 HIV Vaccines

In the setting of HIV and the search for an effective vaccine, an interesting testing problem arises. It is hypothesized that an efficacious vaccine could work in two ways: 1) it could lower the probability p of becoming infected and 2) it could lower the severity of disease for vaccinees who become infected, say captured by the viral set point Y , which is the viral load

following acute infection. In a randomized controlled trial of a vaccine candidate versus placebo, the test for efficacy naturally has the multidimensional null hypothesis:

$$H_0: p_{\text{vaccine}} = p_{\text{placebo}} \text{ and the distribution of } Y \text{ is identical for vaccine and placebo.}$$

One could consider formulating two one-dimensional (weak null) tests, one testing the difference in proportion infected between groups and one testing for a location shift in the distribution of Y in the two groups. A composite summary of the two tests could also be formed. One such summary is the Lachenbruch test, which compares the sum of the squared z -score from the two-group test of proportions and the standardized Wilcoxon rank-sum test for a shift in the median viral load to a chi-squared distribution with two degrees of freedom [9]. The test of proportions would be a valid test of the weak null that the vaccine had no effect on the proportion infected, but this test requires very large sample sizes, as vaccine trials tend to have low rates of infection. There would also be an efficiency loss using this one-dimensional test if there was an additional vaccine effect on the viral load amongst the infected. Constructing a valid test of a location shift in Y is more difficult, as this is a comparison between two non-randomized groups. This could lead to an erroneous conclusion of efficacy if the vaccine affected the viral load mean in the vaccine group by increasing infections in individuals who, with or without the vaccine, would have a low viral load set point. In fact, Lachenbruch's test is particularly advantageous when the vaccine effect on viral load is in the opposite direction than the infection proportion, say when a vaccine increases infection risk but lowers the viral set point. But having high power against alternatives showing that the vaccine may do more harm than good is not helpful. Even putting this concern aside, assuming that comparisons of the mean Y would be valid, constructing the best test of H_0 is not clear. Combining the two Z scores would have worse power than the univariate test for a location shift in the Y if the vaccine had no impact on risk of infection, but only affected the distribution of Y . Interpreting the results of such a combination test relies on understanding what the likely alternatives could be.

As was done in example 1, a series of supportive and complementary statistical tests could be done to better understand the possible mechanism of action and under what set of assumptions the proposed tests of efficacy would remain valid. Several authors have in fact proposed testing schemes for HIV vaccine efficacy [10, 11, 12, 13]. Mehrotra *et al.*[10] compare a number of composite statistics, which amount to different weighting schemes for the two univariate tests, and discuss their relative performance under different possible alternatives. Gilbert *et al.*[11] and others consider the framework of potential outcomes to construct sensitivity analyses to better understand the existing evidence for efficacy under different possible scenarios for differences in infected populations on the two arms [12]. Follmann *et al.*[13] propose a test designed to have good power for a location shift in Y but in a way that would not see power gains under the harmful vaccine scenario described above. These and many other instructive and creative papers in this area demonstrate how careful elucidation of alternatives and study of proposed tests under these alternatives can not only expand the scientific insights to be gained from the data but may be a necessary step to avoiding erroneous inference.

4.2 Atherosclerosis

Coronary angiography is used to measure progression of heart disease. Clinical trials sometimes use as their outcome a measure of the change between angiograms performed at baseline and the end of study, say 3 years later. One popular technique is to divide coronary arteries into multiple segments and use the minimum lumen diameter as a measure of disease severity within each segment; smaller diameters are bad because they indicate more blockage that might eventually result in total occlusion and heart attack. The same segments are examined at baseline and the end of the study, and the change in minimum lumen diameter is computed for each segment. These changes are then averaged over segments. The question then becomes: which segments should be included in the analysis? One could use all segments, but many of them would have no occlusion at baseline or end of study, so including such segments would dampen any signal. It seems more natural to use only segments that are occluded at baseline because they can least accommodate disease progression. Therefore, one popular approach is to count segments that are at least, say, 50% occluded at baseline. But what if the treatment benefit is through limiting the size of new lesions rather than slowing progression in segments that are already occluded? Then inclusion of only the segments that were occluded at baseline might obscure the treatment effect.

We can formulate this discussion mathematically as follows. Imagine selecting a segment at random from a patient, and let (X, D, I) denote the baseline percentage occlusion, change in minimum diameter from baseline to end of study, and indicator of detectable lesion at end of study, respectively for the selected segment. The strong null hypothesis is

$$H_0: \text{The joint distribution of } (X, D, I) \text{ is identical for treatment and placebo.} \quad (5)$$

Weaker null hypotheses include between-arm equivalence of 1) the mean of D or 2) the conditional distribution of D given that $X \geq 50$ or 3) the conditional mean of D given that $X \geq 50$, or 4) the probability that $I = 1$. The strong null implies all of these weaker nulls.

The Women's Angiographic Vitamin and Estrogen (WAVE) steering committee considered all of the above issues in deciding which segments to include in its factorial clinical trial of the effect of estrogen replacement therapy and antioxidant vitamins on progression of coronary artery disease in postmenopausal women [14]. They decided to include segments that were occluded on either the baseline or end of study angiogram. Whether a segment that is clear at baseline is included depends on what happens after randomization, namely whether it becomes occluded by the end of the study. Therefore, whether it is included depends in part on how well the treatment works. A patient in the placebo arm might develop a new lesion that is counted in the primary analysis, whereas had that same patient been assigned to treatment, she may have been prevented from developing a new lesion. Therefore, the corresponding segment would not have been included. This does not cause a problem under the strong null hypothesis in (5) above because in that case there would not be any between-arm difference in which segments are included. As in the vaccine example, one could imagine scenarios under an alternative hypothesis in which the treatment effect estimate could be misleading. For instance, if treatment actually caused very small new

lesions, then these small progressions would be included in the treatment, but not the placebo average. The average treatment progression would be artificially lowered by a harmful treatment effect, namely proliferation of small new lesions. The WAVE Steering committee discussed these issues and decided that the latter scenario was extremely unlikely. As in the vaccine example, careful thought is required about the likely treatment mechanisms. This is true for any trial, but especially so in the potentially perilous setting where who or what is included depends on how well the treatment works. Supportive analyses can be done after the trial is completed to explore other dimensions of the treatment effect and whether there was any indication of undesirable effects not captured by the primary analysis.

5 Discussion

In this paper we examined hypothesis testing in several settings where nuisance parameters had important implications for the scientific question at hand, namely whether the experimental condition had an effect on the target of interest. On autopilot, we as statisticians can be lulled into a sense of security with the term “nuisance parameter” and feel comfortable ignoring any parameter that carries this moniker. We are further assured by statistical attributes, such as the ancillary nature of a parameter that we are ignoring, or parsimony, which pleases our statistical sensibilities when choosing a model. Nevertheless, a distinction must be made between nuisance parameters that are truly unimportant to the mechanisms we wish to understand and parameters that may complicate the testing problem but contain relevant scientific information.

In the first two examples, aspects of the random variation contained evidence that the experimental condition affected the scientific quantity under study. When choosing a statistic, we as statisticians frequently consider mean/variance tradeoff, but we do this perhaps without fully considering or explaining to our collaborators the scientific tradeoffs in the types of conclusions we can make from the selected test(s). Full discussion of what is scientifically interesting could lead to a better test choice. For example, if such discussions highlight that aspects of both mean and variation could be interesting scientifically, the statistician might choose a test that is sensitive to this type of heterogeneity, such as the likelihood ratio test in the therapeutic touch example, rather than the usual t -test of means. In problems with repeated measures, such as the frozen samples example, it is easy to choose a test that is not completely consistent with the null hypothesis of interest, particularly when parsimony is used as the guiding principle for specifying a model in standard statistical software and the corresponding test is buried deep in the documentation for the chosen software. Ironically, the simplest and best statistic can actually correspond to the least parsimonious model. Better communication between the statistician and scientific collaborators allows for a fully informed choice, not one driven by ease or convention.

One setting that is fraught with potential danger is when one of the proposed treatment mechanisms cannot be measured without breaking the randomization. One example is a vaccine that might affect not only the proportion of people who become infected, but also the severity of disease among those who become infected. Infected patients may differ systematically across arms because treatment might affect who gets infected. Another

example involved the choice of segments to include in an angiographic trial of progression of coronary artery disease. Investigators, postulating that one possible treatment mechanism of action is the limiting of the size of new lesions, include in the analysis segments that were occluded either at baseline or end of study. Which segments are included is affected by how well treatment works. Any time who or what is measured depends in part on whether treatment has any effect, there is the potential for misleading conclusions, such as when a vaccine causes mild infections or a drug causes mild new lesions to appear on the angiogram. Additional statistical tests may be required to ensure that this is not occurring.

It is important to consider the scientific subject matter first when deciding what test best fits the testing problem at hand. We must more fully involve our biological collaborators in the process of vetting what should be considered a nuisance parameter versus what is informative scientifically. With a more complete examination of our models and their associated null and alternative hypotheses, we will be in a better position to select a test statistic that best addresses the scientific questions of the investigator.

Appendix

We now examine more carefully what is happening with the mixed model (2), focusing first on patient i . Assume throughout that the underlying variables are normally distributed. The random effect b_i drops out when we form patient i 's paired difference D_i , so $\text{var}(D_i) = \text{var}\{\tau + \bar{\epsilon}_{i1} - \bar{\epsilon}_{i0}\}$, where the dot denotes an average over the 2 observations at the given time point. Therefore,

$$\text{var}(D_i) = \sigma_\epsilon^2/2 + \sigma_\epsilon^2/2 = \sigma_\epsilon^2.$$

The fact that any estimator of σ_ϵ^2 is also unbiased for $\text{var}(D_i)$ leads to a within-patient estimate of σ_D^2 as an alternative to the between-patient estimate s_D^2 . Consider the sample variance s_{i0}^2 of the two observations at day 0 from patient i . Again the random effect b_i drops out, and s_{i0}^2 is simply the sample variance of e_{i01}, e_{i02} . It follows that $s_{i0}^2/\sigma_\epsilon^2$ has a chi-squared distribution with 1 degree of freedom and, because the mean and variance of *iid* normal observations are independent, s_{i0}^2 is independent of $\bar{\epsilon}_{i0}$. Similarly, the sample variance s_{i1}^2 of the two observations at day 1 is independent of $\bar{\epsilon}_{i1}$. Also, the fact that $(\bar{\epsilon}_{i0}, s_{i0}^2)$ is a function of (e_{i01}, e_{i02}) and $(\bar{\epsilon}_{i1}, s_{i1}^2)$ is a function of (e_{i11}, e_{i12}) means that $(\bar{\epsilon}_{i0}, s_{i0}^2)$ and $(\bar{\epsilon}_{i1}, s_{i1}^2)$ are independent. It follows that the within-patient paired difference $D = \tau + \bar{\epsilon}_{i1} - \bar{\epsilon}_{i0}$ is independent of the within-patient pooled variance estimate. See Table 1 for a summary.

Therefore, another estimator of $\text{var}(D)$ is the within-patient estimator

$$\hat{\sigma}_w^2 = \sum_{i=1}^n s_i^2/n.$$

We can combine the two statistically independent estimators s_D^2 and $\hat{\sigma}_W^2$. Each of them, when multiplied by its degrees of freedom and divided by σ_ε^2 , is a chi-squared random variable. The hybrid estimator

$$\hat{\sigma}_H^2 = \frac{(2n)\hat{\sigma}_W^2 + (n-1)s_D^2}{2n+n-1} \approx (2/3)\hat{\sigma}_W^2 + (1/3)s_D^2, \quad (6)$$

which has a chi-squared distribution with $3n - 1$ degrees of freedom, is actually the error term for mixed model (2) (see Table 2).

The problem is that under model (4), $\text{var}(D_i) = \sigma_\varepsilon^2 + \sigma_{b1}^2$, as noted earlier. Although s_D^2 estimates this variance, $\hat{\sigma}_W^2$ estimates $\sigma_\varepsilon^2 < \text{var}(D_i)$. Therefore, $\hat{\sigma}_H^2$ also underestimates $\text{var}(D_i)$ and leads to inflation of the type I error rate.

References

1. Casella, G.; Berger, R. Statistical Inference. 2nd edn. Duxbury Press: Florence, KY; 2001.
2. Rosner, B. Fundamentals of Biostatistics. 7th edn. Brooks/Cole Cengage Learning: Boston, MA; 2010.
3. Boes, D.; Graybill, F.; Mood, A. Introduction to the Theory of Statistics. 3rd edn.. McGraw-Hill: New York, NY; 1974.
4. Lehmann, E.; Romano, J. Testing Statistical Hypotheses. 3rd edn. Springer Verlag: New York, NY; 2005.
5. Rosa L, Rosa E, Sarner L, Barrett S. A close look at therapeutic touch. Journal of the American Medical Association. 1998; 279(13):1005–1010. [PubMed: 9533499]
6. Rosa L, Sarner L, Barrett S. An even closer look at therapeutic touch. Journal of the American Medical Association. 1998; 280(22):1908. [PubMed: 9851474]
7. Higgins J, Metcalf J, Stevens R, Baseler M, Proschan M, Lane H, Sereti I. Effects of delays in peripheral blood processing, including cryopreservation, on detection of CD31 expression on naive CD4 T cells. Clinical and Vaccine Immunology. 2008; 15(7):1141–1143. [PubMed: 18448620]
8. Verbeke, G.; Molenberghs, G. Linear mixed models for longitudinal data. Springer: New York, NY; 2000.
9. Lachenbruch P. Comparisons of two-part models with competitors. Statistics in Medicine. 2001; 20(8):1215–1234. [PubMed: 11304737]
10. Mehrotra D, Li X, Gilbert P. A comparison of eight methods for the dual-endpoint evaluation of efficacy in a proof-of-concept HIV vaccine trial. Biometrics. 2006; 62(3):893–900. [PubMed: 16984333]
11. Gilbert P, Bosch R, Hudgens M. Sensitivity analysis for the assessment of causal vaccine effects on viral load in HIV vaccine trials. Biometrics. 2003; 59(3):531–541. [PubMed: 14601754]
12. Shepherd B, Gilbert P, Jemai Y, Rotnitzky A. Sensitivity analyses comparing outcomes only existing in a subset selected post-randomization, conditional on covariates, with application to HIV vaccine trials. Biometrics. 2006; 62(2):332–342. [PubMed: 16918897]
13. Follmann D, Fay M, Proschan M. Chop-lump tests for vaccine trials. Biometrics. 2009; 65(3):885–893. [PubMed: 19210727]
14. Waters D, Alderman E, Hsia J, Howard B, Cobb F, Rogers W, Ouyang P, Thompson P, Tardif J, Higginson L, et al. Effects of hormone replacement therapy and antioxidant vitamin supplements on coronary atherosclerosis in postmenopausal women. Journal of the American Medical Association. 2002; 288(19):2432–2440. [PubMed: 12435256]

Table 1

Summary of mean and variance estimates under Model (2).

Time 0	Time 1
$Y_{i01} = \mu - (1/2)\tau + b_i + e_{i01}$ $Y_{i02} = \mu - (1/2)\tau + b_i + e_{i02}$	
$Y_{i11} = \mu + (1/2)\tau + b_i + e_{i11}$ $Y_{i12} = \mu + (1/2)\tau + b_i + e_{i12}$	
$s_{i0}^2 = s^2(Y_{i01}, Y_{i02}) = s^2(\varepsilon_{i01}, \varepsilon_{i02})$	$s_{i1}^2 = s^2(Y_{i11}, Y_{i12}) = s^2(\varepsilon_{i11}, \varepsilon_{i12})$
$D_i = \tau + e_{i1}, e_{i0}$	
$s_{i0}^2 = s^2(Y_{i01}, Y_{i02}) = s^2(\varepsilon_{i01}, \varepsilon_{i02})$	$s_{i1}^2 = s^2(Y_{i11}, Y_{i12}) = s^2(\varepsilon_{i11}, \varepsilon_{i12})$
$s_i^2 = (s_{i0}^2 + s_{i1}^2)/2 = [s^2(\varepsilon_{i01}, \varepsilon_{i02}) + s^2(\varepsilon_{i11}, \varepsilon_{i12})]/2$ $D_i = \tau + \bar{\varepsilon}_{i1} - \varepsilon_{i0}, D_i = \tau + e_{i1}, e_{i0}$	

Table 2

ANOVA table corresponding to Model (2).

Source	Degrees of freedom
Time	1
Patients	$n - 1$
Error	$3n - 1$
Total	$4n - 1$