American Journal of Therapeutics 10, 112–121 (2003)

Perils of the Random Experiment

Lemuel A. Moyé[1]* and Anita Deswal[2]

Most medical research is executed on samples selected from large populations. Nevertheless, health care researchers often blur the difference between interpreting sample-based research and evaluating research that included the entire population of interest. This is an implication-critical distinction; in population research, every result applies to the population (because the entire population was included in the analysis), although only a few results from sample-based research can be extended to the population at large. Treating every result from sample-based research as if that result applies to the population is misleading. Using nonmathematic terminology, this article develops the reason for the differences in the implications of these two research perspectives. In sample-based research, the best indicators of which results should be extended from the sample to the population are the presence of (1) a prospective plan for that experiment; and (2) the execution of the experiment according to that plan (concordant execution). The absence of these two features produces execution and analysis decisions based on the incoming data stream—the hallmark of the random experiment. In this latter paradigm, allowing the data to influence the execution and analysis decisions renders the usual estimates of effect size, standard errors, confidence intervals, and *P* values untrustworthy. Readers of clinical trial results must be vigilant for nonprotocol-driven research and understand that the results from these programs are at best exploratory and cannot be used to answer scientific questions.

INTRODUCTION

Many guides are written for clinical scientists concerning the correct design, execution, and analysis of research programs.[1,2] An important part of that information focuses on settling on an analysis plan prospectively.[3] Nevertheless, wise investigators recognize that the experience of a clinical trial can be unpredictable, producing surprising and unexpected results. In some circumstances, the anticipated finding never materializes.[4] In others, a modest finding that was expected is overshadowed by a stupendous finding from another analysis.[5,6] Sometimes, the analysis that produced the stupendous finding was planned; at other times, it was not. The advice from methodologists is that these ancillary findings do not carry persuasive weight, primarily because they were not planned prospectively.[7]

To many researchers, this concern seems much ado about nothing. The data are, in the end, the data. Allowing the data to decide the result of the experiment might seem to be the fairest and least prejudicial evaluation of the data. This policy also relieves the investigator from the responsibility of choosing arbitrary rules during the planning stage of the experiment, which may subsequently be demonstrated by the data to be the wrong choices. From the investigator's perspective, it may seem far better to preserve some flexibility in the experiment's interpretation by saying little during the design of the experiment about the end point selection or analysis procedures and letting the data choose the best analysis and end point selection as long as these selections are consistent with the goals of the experiment.

This point of view may be bolstered by the observation that obtaining the research sample correctly for

[1]*University of Texas School of Public Health, and* [2]*Winters Center for Heart Failure Research and Houston Center for Quality of Care and Utilization Studies, Veterans Administration Medical Center, and Baylor College of Medicine, Houston, Texas.*
A.D. is a recipient of the Veterans Administration Cooperative Studies Program Clinical Research Career Development Award (CRCD 712B).
**Address for correspondence: University of Texas School of Public Health, RAS Building E815, 1200 Herman Pressler, Houston, TX 77030, U.S.A.; e-mail: lmoye@sph.uth.tmc.edu*

clinical research can be an onerous, time-consuming, and expensive process. Intelligent and well-developed methodologies are required to choose the optimum sample size.[8–12] Well-tested randomization mechanisms are put in place to avoid systematic biases, which can produce destabilizing imbalances and confound the research interpretation. In fact, the fundamental principle of simple random sampling is to produce a sample that is representative of the population.[13]

Many investigators believe that after carrying out these procedures, they have provided the greatest assurance that the results from the sample are applicable to the population at large. If investigators have patiently carried out this preparatory work, why are they prohibited from believing the answers their sample provides to the questions the investigators ask, regardless of whether the questions were prospectively posed? Why cannot the researcher be allowed to take advantage of that representation by generalizing results from this "representative" sample to the population it represents? To investigators, withholding belief in a surprise finding's validity can seem like denying credit to Columbus for discovering the New World because his discovery was, after all, "not part of his protocol."

DIFFERENCE BETWEEN THE SAMPLE AND THE POPULATION

It is understandable that a scientist chooses to rely on the findings from his sample, which were obtained at great logistic and financial cost. Much is invested in the sample; therefore, much is demanded from it. However, the presence of a representative sample does not infer that every factoid in the sample portrays a reliable reflection of a population finding. To understand this requires only a brief review of what a sample is and how it is obtained. We must first observe that in health care research, samples, although obtained at great expense, are often miniscule when compared in size with populations (Fig. 1).

For example, a clinical research program may go to great effort and expense to identify 300 patients with type II diabetes, randomly allocating them to either a new pharmacologic intervention or control therapy. Despite the great effort involved in this enterprise, it must be acknowledged that because there are 15 million patients with diabetes in the United States,[14] the research sample contains only 0.002% of the total number of patients with diabetes. Put another way, 99.998% of patients with diabetes are specifically not included in the sample. Also, it is easily computed that

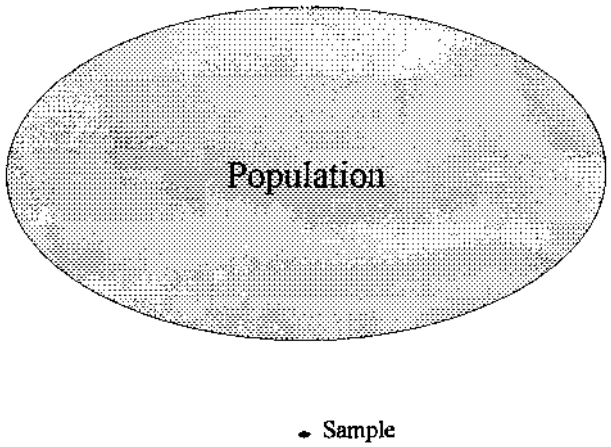

Fig. 1. The difficulty with inferring the sample results to the population.

15,000,000 / 300 or 50,000 different samples of the same size can be obtained from this same population. Each of these samples contains a kernel of truth about the population but not the complete truth. In addition, the "version" of the truth varies across samples. This sample-to-sample variability is termed *sampling error*. Because these samples contain different patients with different life experiences and different data, discerning the "truth" about a population by merely evaluating a single sample is problematic. Although it may be reasonable to conclude that this sample contains a nugget of truth that can be extended to the population, it is unreasonable to believe that every finding of information the one sample contains represents the truth about the population.

Therefore, extending results from a single sample to the population is a fragile process. Sampling error cannot be removed from this process. However, the extent to which sampling error is responsible for the results from the sample can be estimated. Recognizing this, experienced research designers focus on one subcomponent of the information in the sample. That one component is the study question, representing the single issue that the sample addresses. Nevertheless, because the presence of random sample variability can still present an obstacle to the extension of the sample's results to the population, investigators turn to the quantitative procedures of epidemiology and biostatistics. These fields have provided the computations that convert the data's information to the best unbiased estimates of the intervention's effect size (e.g., mean effect, odds ratio, or relative risk) and effect size variability. It is important to note that these estimators do not remove sampling error; instead, they incorporate it. If the researcher is also interested

in inference (i.e., hypothesis testing), statistical procedures can also channel sampling error into *P* values (the likelihood that there is no effect in the population but that the population has misled us by producing a sample in which there is an effect) and power (the likelihood that there is an important effect of therapy in the population but that the population has misled us by producing a sample in which no effect of therapy is observed). Thus, when used correctly, epidemiologic and statistical methodologies appropriately recognize and channel sampling error to familiar quantities that researchers have useful experience in interpreting.

For this process to be informative to the researcher and, ultimately, the medical community, the estimators used in interpreting the research must be reliable, that is, they must accurately measure what they were designed to measure. This accuracy comes automatically if the experiment is executed according to a well-designed and detailed protocol; however, this accuracy is lost when protocol violations occur. As a case in point, consider the following example.

THE EXAMPLE

An enthusiastic young researcher, Dr. B, interested in demonstrating the progressive increase in stroke incidence in patients with untreated atherosclerotic disease, designs a simple research program to examine the increase in the incidence of ischemic stroke over time in patients who are at risk for having a stroke. Dr. B recruits a sample of patients and follows them for 2 years, measuring the occurrence of stroke at the end of the 2-year period. Although Dr. B is focused on strokes, he measures other morbidity (i.e., fatal and nonfatal myocardial infarction, use of revascularization procedures, hospitalization for cardiovascular disease) as his cohort ages. All measures of morbidity are measured with the same precision. At the study's conclusion, Dr. B discovers, to his surprise and horror, that there has not been an important increase in the incidence of stroke but there has been an unmistakable increase in the incidence of fatal and nonfatal myocardial infarction. He therefore decides to change the end point for the program from an increase in stroke to an increase in fatal and nonfatal myocardial infarction and reports the findings of the latter in the peer-reviewed literature as the result of the experiment.

In our experience, many knowledgeable scientists would have no problem with Dr. B's switch from stroke to myocardial infarction, arguing that each of these end points is a measure of the same underlying physiologic and pathophysiologic criteria. They would claim that because these end points jointly measure the progress of the same disease process, there is no harm in interchanging them as study end points. They might contend that Dr. B. should not be held to the impossible standard of having to guess right about the best measure of atherosclerotic disease to choose as his end point. Because he had the insight to measure several different facets of atherosclerotic morbidity, perhaps he should be commended for his foresight in measuring the incidence of myocardial infarction and his decision to raise the significant result to a prominent place in his report. Others among us would be uncomfortable with the end point change, but we may be unclear as to exactly what the problem is. These critics might say that the decision to change the end point was "data driven." Well, what is so wrong with that? Are not the results of any study "data driven"?

RANDOM DATA VERSUS RANDOM EXPERIMENTS

As pointed out in the previous section, sampling error must be controlled and channeled in an interpretable sample-based research effort. Estimators used to accomplish this are effective but rely on a critical assumption—only the data can be subject to random influence; the research protocol must remain fixed. Estimators that we commonly use (e.g., mean changes, odds ratios, relative risks, confidence intervals, *P* values) were designed to work well in this environment. Nevertheless, when these estimators are computed in a research environment, where the protocol and analysis rules are influenced by the data (i.e., a random environment), the estimators no longer have the capability to assess sampling error in the data in which sampling error has influenced the research protocol. Our usual estimators can handle one source of variability, the fact that the data in one sample differ from those in another in measuring a predetermined end point, but they cannot handle the fact that a different sample with different data would lead to a different end point choice. The estimators therefore function irregularly, returning aberrant estimates of what they were designed to measure. Thus, to ensure the research effort's interpretability, the sampling error contained in the data must be segregated from the research procedures and analysis plans. This separation ensures that the statistical procedures applied to the data provide accurate estimates of effect size, standard errors, confidence intervals, and *P* values. No longer

anchored to its protocol, the research has become discordant and random, and the statistical estimates have been corrupted.* Drawing conclusions from these estimators is like combing one's hair using a distorted and blurred mirror. Because the reflection is not a true one, the result is unsatisfactory, and, ultimately, the exercise must be repeated. Only the data can be subject to random variation to comply with the assumptions on which the statistical tools rely. All else must be fixed.

This last point is worthy of elaboration. The difficulty with the computations is not that the random end point selection produces sloppy calculations. To the contrary, great effort is expended on these computations, using modern computing facilities and procedures. Nevertheless, because the research is no longer anchored to its protocol but has become random, the computations for effect sizes, confidence intervals, and type I/II errors can no longer accurately estimate sampling error. These computations were designed for only one source of variability—the data. The end point selection randomness has destroyed this paradigm. We know that the computations are wrong, and we do not know how to make them right. Therefore, estimators developed for the first (arbitrary, prospective) paradigm are now useless when the paradigm has shifted to the second (random) one. Because they are incorrect and uncorrectable, we cannot integrate them in our fund of knowledge; unfortunately, they must be discarded. The only protection against this dilemma is the prospective specification of the end points. This is the central motivation for the research tenet, "First, say what you will do; then, do what you said."

Consider the exercise a statistician goes through when asked to consider the development of a test statistic for the incidence of stroke. The statistician commonly starts with a statement "Let x_i, where i = 1,…, n be a sample of observations from a population in which $x_i = 1$ if the 'ith' patient had a stroke, and $x_i = 0$ if the 'ith' patient did not have a stroke. Then,

$$\sum_{i=1}^{n} x_i$$

has the following probability distribution…" From these statements, estimators of effect size, standard errors, confidence intervals, and *P* values are available to effectively convey the strength of evidence the data contain. The statistician can proceed by saying, "If we wish to estimate the population incidence rate, it is a straightforward result from the maximum likelihood theory, least squares theory, or optimality theory that the sample mean

$$\frac{\sum_{i=1}^{n} x_i}{n} = \bar{X}$$

is the best estimator of the population incidence." This paradigm falls apart when x is not prospectively and arbitrarily chosen but is instead selected by the data, which contain sampling error. In this second (random) paradigm, there is now a new probability distribution that governs the selection of the end point variable itself. This change in assumption leads to a more complicated second situation in which our commonly used familiar estimators are no longer optimal. In this new paradigm, our statistician cannot begin with, "Let x represent the incidence of a stroke." He/she must consider something along the following lines, starting with, "The data will determine the end point that will be used for this study, so by what random mechanism will the end point be chosen?" Consider the circumstance that there are five possible end points: V, W, X, Y, and Z. Let v_i be the realization of V, w_i be the realization of W, x_i be the realization of X, y_i be the realization of Y, and z_i be the realization of Z for the "ith" patient, i = 1,…,n. Because we do not know which of V, W, X, Y, or Z will be the final end point (the data determine that), we have to explicitly consider that each may be the end point for the study. For this, we need a probability distribution for the end point selection. Let the probability that the random variable V is chosen be P_v, the probability that the end point W is chosen be P_w, with analogous assignments for P_x, P_y, and P_z such that $P_v + P_w + P_x + P_y + P_z = 1$. In this paradigm, the best estimator for the sample mean is not $\bar{X}$. Instead, it is the more complicated estimate:

$$e = p_v \bar{V} + p_2 \bar{W} + p_x \bar{X} + p_y \bar{Y} + p_z \bar{Z}$$

Applying this weighted estimate to the experiment of Dr. B, it is clear that Dr. B should have used a weighted incidence measure for the possible end points in this study. This estimator, although best for the experimental execution, is problematic for two reasons. First, there is no known procedure for choosing the values of the probabilities P_v, P_w, P_x, P_y, and P_z prospectively. Second, this weighted end point is exceedingly difficult to interpret. If Dr. B's candidate end

*There are additional problems with changing the end point from stroke to myocardial infarction. The incidence rates may be quite different between the two end point measures, leading to a different sample size requirement. Although logistically important, this specific implication for the end point choice is not the focus of this article.

points were continuous, the interpretability problem gets much worse. For example, if V reflected change in low-density lipoprotein cholesterol, W reflected change in very low-density lipoprotein cholesterol, X reflected change in high-density lipoprotein cholesterol, Y reflected a change in triglycerides, and Z was a change in apolipoprotein B cholesterol, the appropriate estimate would be a weighted average of these changes, a measure that defies interpretation.

The key point here is that it is the choice of the estimator that is linked to the research execution. The random research paradigm of Dr. B requires difficult and obtuse estimators for the correct analysis. The correct analysis is difficult to interpret, but this is the analysis he should use. The fact that the estimators are difficult to use and interpret is merely an expression that the random research environment is difficult. The problem is not the estimators but the paradigm. Using the customary estimators of the fixed paradigm in this random paradigm is also incorrect, because these estimators are not interpretable in this random environment. Even though Dr. B (now) knows that choosing $\bar{X}$ was wrong, he does not know precisely how to correct it, because such a correction requires him to estimate values of the probabilities P_v, P_w, P_x, P_y, and P_z. Thus, his estimate of the effect size for his experiment is wrong, and his estimates of variability, confidence intervals, and any statistical inference (*P* values and power) that he carries out are also wrong. The experiment as designed was interpretable; however, the random end point selection process executed during the research execution has destroyed that interpretability.

The action that led to the corruption of the research was allowing the data stream with its sampling error to affect the analysis procedure. The root of this difficulty is how to handle sampling error. This dilemma is resolved if the researcher can study every subject in the population. In this circumstance, there is no sampling error. In fact, there is no estimation, because the population parameters are directly measured. For example, a laboratory researcher is interested in characterizing the measure of abnormal glucose metabolism in diabetic patients admitted to a community hospital. There are two possible candidates for the research end point: glycosylated hemoglobin or fasting blood glucose levels. In this circumstance, there is no requirement for choosing only one prospectively. The decision to measure one or the other or both can be made at any time, because there is no sampling error. Thus, choosing one of these prospectively is not necessary if the investigator has no interest in generalizing the results of the study to another population (e.g., to the city hospital or community hospital diabetic patients to be admitted in the future). There is freedom in choosing the end point here. The end point selection liberty gained by studying the entire population is counterbalanced by the generalizability restriction, however.

FINAL PERSPECTIVE ON RANDOM RESEARCH

In the more common paradigm of sample-based research, random data containing sampling error cannot be allowed to change a fixed research protocol. Like the fruit from a poisonous tree, the results from the altered research protocol cannot be understood and absorbed and must be shunned. Such data-driven protocol deviations are a klaxon for poor estimators, inaccurate standard errors, and untrustworthy type I and type II error calculations. The only protection against this dilemma is the prospective specification of the end points in complete detail, leaving nothing to chance. This is the central motivation for the research tenet, "First, say what you will do; then, do what you said." The following are areas where random research arises and some techniques to avoid them.

RANDOM RESEARCH AND MULTIPLE ANALYSES

By an analysis, we mean any statistical procedure that leads to a hypothesis test. Thus, the analysis of the effect of therapy among different treatment groups in a clinical trial with more than two arms, the evaluation of multiple end points in a single clinical trial, the study of the effect of the intervention in a collection of subgroups of the research cohort, or combinations of these evaluations are examples of multiple analyses. The consequences of the preceding discussion for the reliability of the conclusions from these multiple analyses are direct and immediate. Investigators who ask the question about the intervention-disease relation first (prospective design) and then execute the experiment according to that design are assured that the measures of (1) the magnitude of the intervention-disease relation; (2) its standard error; and (3) the type I and type II errors[†] are trustworthy.

Even in this circumstance, there can be difficulties. If too many questions are asked of the research sample

[†]Type I and type II errors are the estimates of sampling error that measure whether sampling error produced the findings the investigator observed in the population.

Table 1. Consequences of alternative strategies for selection of multiple analyses.

Strategy	Difficulty
Analysis plan is based on the incoming data	Untrustworthy estimates of effect size, standard errors, confidence intervals, and *P* value
Prospective choice of analyses, no a priori alpha allocation	Trustworthy estimates available, but type I error inflation occurs
Prospective choice of analyses with a priori alpha allocation	Trustworthy estimates are available with good type I error control

(even if these questions are asked prospectively), the likelihood that the answers to these questions are the result of sampling error grows. This is reflected in inflated type I and type II errors. Just like the probability of at least one "tails" in the successive flips of a coin increases as the coin is repeatedly tossed, the probability of at least one false answer increases if the sample is queried repeatedly for information that is generalized to the population. In a concordantly executed[‡] experiment, the analysis for one end point may result in a *P* value of 0.05; when two questions are asked, the probability that at least one of the answers is a result solely of sampling error is $1 - (1–0.05)(1–0.05) = 0.0975$. What we observe is that the likelihood of at least one of those conclusions being wrong (referred to as the family-wise type I error rate) becomes too large too quickly. We do not know which of the end point findings is wrong; we can only say that the likelihood that at least one of them is wrong has become too large.

The challenges and difficulties of using multiple analyses in clinical trials and the consequences of these decisions are easily described (Table 1). The planning of the analyses has a profound effect on the interpretation of the analyses. If the analyses are post hoc evaluations with no a priori planning, the estimators of effect from these analyses are untrustworthy. Any hypothesis testing based on these estimators is also unreliable, and the results of analyses are best viewed as exploratory. In these circumstances, the results of hypothesis tests might best be reported as *z* scores, which are normalized effect sizes rather than *P* values. Because exploratory analyses cannot be generalized to the population, type I error for these exploratory analyses is irrelevant, and the salient information about the scaled effect size is communicated more clearly by reporting solely the test statistic.

If the analysis is planned prospectively and the experiment is executed concordantly (i.e., executed according to the protocol), the estimators derived from the analysis are trustworthy. The inflation of type I error from the multiple analyses decreases the persuasive force of the findings from these evaluations, however. The most direct interpretable evaluation from multiple analyses is produced from well-designed, prospectively planned, and concordantly executed analyses with adequate type I error allocation to each of the confirmatory analyses.

[‡]Concordant execution simply means that the experiment was executed according to its protocol.

APPLICATION OF THE PRINCIPLE

The following are areas of multiple analyses in clinical trials with which the previous discussion has a direct connection.

End point selection

Two commonly occurring problems in health care research directly involve the choice of the trial's end points. The first is that the end points may not have been determined before the experiment began. This identification of end points post hoc makes the analysis uninterpretable, because the end point was determined based on the data, which, of course, contain an important random component. The difficulties in interpretation of programs with this deficiency have been elaborated in the previous discussion. Second, even when chosen a priori, if too many end points are evaluated, type I error accumulates rapidly as the number of conclusions drawn from a sample increases. Therefore, multiple end point selection must be prospective, and there must be adequate type I error protection in a clinical trial. It may be useful to think of type I error occurrence as an issue of population protection. Every intervention has side effects. The justification for their use is that the intervention also has a beneficial effect. The occurrence of a type I error denotes that the efficacy of the intervention in the population is like a placebo. Therefore, commission of a type I error is tantamount to exposing the community to an intervention that pro-

duces no efficacy but does produce real adverse effects. Thus, controlling type I error rate is a way to provide community protection from a potentially harmful intervention that has no efficacy.

Because type I error rate control is an important community protection device, it must be kept to an acceptably low level in health care research that studies an intervention. This responsibility requires both a clear understanding of and a tight rein on type I error rates. The Bonferroni approach (i.e., dividing the total type I error by the number of hypothesis tests to be executed)[15] and its adaptations,[16] commonly used for secondary end points, does not work well as generally applied (although successfully used in some research), because the type I error threshold for each test can decrease to an unusable low level quickly for each additional test. In addition, the current use of type I error requires that research efforts with null findings for the primary end point but with positive findings on secondary end points be considered null trials, which is a matter of great frustration to investigators who resist being compelled to place all their "alpha eggs" in one primary end point "basket."

This difficulty has elicited discussion recently.[17–23] The major recommendations from this body of work are to require that (1) each primary and secondary endpoint be prospectively chosen and (2) each of these end points have type I error attached in a prospective and reasoned fashion. This collection of procedures increases the rigor for the prospective statements concerning secondary end points, although permitting the straightforward interpretation of a research effort that is positive for secondary end points but in which the primary end point is not statistically significant. Most recently, the Clopidogrel in Unstable Angina to Prevent Recurrent Events Trial demonstrated the advantage of not just choosing end points prospectively but of making an a priori alpha assignment to each.[24] This study prospectively considered two coprimary end points. The first was the composite of death from cardiovascular causes, nonfatal myocardial infarction, or stroke. The second primary outcome was the composite of the first primary outcome or the occurrence of refractory ischemia. The investigators allocated 0.045 to the primary outcome and 0.01 for the second coprimary end point.[§]

Random influences in modeling

A commonly used analysis tool in clinical research is regression analysis. Its use has allowed health care researchers to carefully examine the relation between an explanatory variable of interest and the end point measure while simultaneously identifying, isolating, and removing the effect of intermediary and related variables. Regression analysis is of particular interest to researchers who are not in the position of being able to assign the risk factor (e.g., as intervention) randomly. This is a common issue in postmarketing studies in which the patients and not the investigators choose the medication they are taking. Consider, for example, the difficulty posed by the investigation of the relation between antihistamine use and the occurrence of sudden death. In such a study, one follows patients who have been exposed to antihistamine 1, antihistamine 2, or antihistamine 3 over time, [‖] collecting information on the occurrence of sudden death subsequent to exposure.

Because the investigator did not choose the antihistamine for these patients, the factors that led the patient or the patient's physician to choose the antihistamine (e.g., comorbidity or concomitant medications) may be more closely related to the true risk factor for sudden death. Regression analysis (e.g., logistic regression analysis or Cox proportional hazard analysis) is a commonly used tool to address this problem. Although not as effective as the random allocation of exposure, these procedures permit the removal of the influence of these additional variables. How should the investigator choose the adjusting variables? A commonly followed procedure is to collect information on the comorbid events and other patient characteristics that may or may not be related to both antihistamine use and sudden death and to let the regression model choose which adjusting variables are important. This broad database search for significant confounders represents a thorough use of the data set; in fact, this procedure maximizes the data set's ability to explain important variability in the occurrence of sudden death. Nevertheless, this maximum use of database information comes at the price of confusing the view of the relation between antihistamine use and sudden death in the population.

This confusion has at its heart the realization that the random aggregation of data in the sample produces variable interrelations that are the result of

[§]Dependence between these two end points led to a family-wide type I error level of 0.05, less than the 1 – (1–0.045) (1–0.01) = 0.055 from the standard Bonferroni computation.

[‖]We ignore in this discussion the difficulty caused by the sequential (or concomitant) use of different antihistamines.

chance alone, reflecting nothing about the relations between these variables in the population. Thus, variables may seem to be noteworthy adjusters because they have a relation with either antihistamine use or sudden death in the sample, when no such relation exists in the population. For example, consider two different researchers drawing different samples from the same population to answer this question. Each of them uses the data to determine the variables for which they wish to adjust the antihistamine–sudden death relation. Because the data sets are random, however, and the data sets choose the covariates, the covariate selection procedure is also random. These investigators are likely to produce two different sets of adjusting covariates. Which is correct? The difficulty becomes all the more pernicious if the investigators produce two different conclusions about the risk that antihistamines pose for the occurrence of sudden death based on the covariates for which the analysis was adjusted.

The best solution to this problem is to remove the random covariate selection procedure. Investigators should deliberately choose the adjusting variables based on information collected before the inception of the experiment. This requires taking the time to identify previous regression models built on other data sets that have identified adjusters. Although requiring additional work during the design feature of the research, this has the advantage of providing a model that is interpretable and directly extendable to the population at large.

Random subgroup analyses

Subgroup analysis is the examination of a treatment effect within a fraction of the entire cohort. In clinical trials, subgroup analysis is produced when the focus on the effect of therapy in the entire cohort is narrowed further to the effect of therapy in only a fraction of the cohort. The number of subgroup analyses executed in a clinical trial can quickly mushroom, because there are so many subgroups available to be analyzed. In its worst form, subgroup analysis can degenerate to data dredging, in which the data set is examined for statistical significance in every possible subgroup.[25] This can lead to untrustworthy estimators as well as to profligate type I errors.

Subgroup analyses have been covered in the clinical research literature. Yusuf et al[26] have clearly presented the specific methodologic difficulties with subgroup analyses as executed in health care research. One of the common problems with subgroups is the absence of a prospectively stated analysis plan, making the estimators of subgroup effect untrustworthy. Another difficulty with subgroup analysis is that the criteria for determining subgroup membership may be based not on baseline information but on information that only becomes available after randomization. The problem here is that factors that influence the effectiveness of therapy may have influenced subgroup membership criteria. In this setting, the effect of therapy cannot be disentangled from the subgroup entry criteria, confusing the interpretation of the result.

Finally, there should be appropriate protection for type I and type II error. Multiplicity of type I error in concert with small sample sizes leads to type I error inflation. The commonly used tool to control this difficulty (i.e., increasing the sample size of the subgroup) can only be applied rarely. When so much work is invested in obtaining a research cohort large enough to answer the scientific question with some statistical certainty, it is almost too much to ask investigators also to ensure that the subgroup is large enough to answer the same question. Nevertheless, some recent work has identified conditions under which prospectively designed subgroup analyses with adequate power may be designed with resultant confirmatory analyses.[27,28]

Data dredging is random research in extremis. This search for significant findings in the research may be well motivated, and the dredgers are driven by the notion that if they look hard enough and long enough and dig deep enough, they will turn up something "significant." Although it is possible to discover a jewel in this strip-mining operation, it is also more likely that for every rare jewel identified, there will be many false alarms, fakes, and shams. It takes tremendous effort to sort out all these findings. In his book, *Experimental Agriculture* (1849), James Johnson[29] states that a badly conceived experiment is not only wasted time and money but also leads to the adoption of incorrect results in standard books, the loss of money in practice, and the neglect of further research along more appropriate lines. This is the legacy of random research. It is not enough to design the research well. Believing that well-designed research can overcome protocol violations is like believing that a well-stocked kitchen can turn out a good meal despite the skills of the chef. Both a well-prepared kitchen and a skilled chef are necessary.

CONCLUSIONS

The fact that a sample was obtained with great care using state-of-the-art sampling procedures does not ensure that every finding the sample produces can be applied to the population from which the sample was

derived. Extension of the sample results to the population is a delicate process, quickly disrupted by allowing the incoming data to determine the analysis. Random research, generated from allowing the random data stream to change the analysis plans of a clinical investigation, has important implications. The commonly used estimators of effect size, standard errors, confidence intervals, and *P* values are no longer trustworthy and therefore not worthy of further consideration in the random research paradigm.

Even though they are not confirmatory, exploratory research efforts can shed first light on new directions for future research. An exploratory analysis can be of great value if it is announced before it is carried out, exerts discipline through the early choice of an analysis plan, and limits itself to examinations that are plausible based on the understanding of the mechanism of the disease. *P* values should be avoided, and the reported effect sizes should be used to ask the question rather than to answer one.

Finally, the readership must develop a new skill of discrimination. Keeping in mind that the role of the investigator is not as a "searcher" who stumbles on an unexpected finding but as a "*re*searcher" who confirms an a priori hypothesis with scientific rigor, readers of the peer-reviewed medical research literature must separate confirmatory from exploratory analyses. Confirmatory analyses are those for which there is a prospective specification of an analysis plan in complete detail, including type I error allocations, leaving nothing in the analysis plan to be determined later by the data. This is the best way to ensure that the estimators the investigators have provided are trustworthy. Data-driven protocol deviations, which are the hallmarks of random research, are alarm bells for type I and type II error aberrations and can serve only to produce preliminary exploratory evaluations.

REFERENCES

1. Meinert CL: *Clinical Trials Design, Conduct, and Analysis*. Oxford University Press, New York, 1986.
2. Friedman L, Furberg C, DeMets D: *Fundamentals of Clinical Trials*, 3rd ed. Mosby, New York, 1996.
3. Moyé LA: *Statistical Reasoning in Medicine—The Intuitive P Value Primer*. Springler-Verlag, New York, 2000.
4. MRFIT Investigators: Multiple risk factor intervention trial. JAMA 1982;248:1465–1477.
5. Pitt B, Segal R, Martinez FA, et al, on behalf of the ELITE Study Investigators: Randomized trial of losartan versus captopril in patients over 65 with heart failure. Lancet 1997;349:747–752.
6. Packer M, O'Connor CM, Ghali JK, et al, for the Prospective Randomized Amlodipine Survival Evaluation Study Group: Effect of amlodipine on morbidity and mortality in severe chronic heart failure. N Engl J Med 1996;335:1107–1114.
7. Moyé LA: P-value interpretation and alpha allocation in clinical trials. Ann Epidemiol 1998;8:351–357.
8. Lachim JM: Introduction to sample size determinations and power analyses for clinical trials. Control Clin Trials 1981;2:93–114.
9. Sahai H, Khurshid A: Formulae and tables for determination of sample size and power in clinical trials for testing differences in proportions for the two sample design. Stat Med 1996;15:1–21. (CHECK HARD COPY FOR ARTICLE TITLE)
10. Davy SJ, Graham OT: Sample size estimation for comparing two or more treatment groups in clinical trials. Stat Med 1991;10:3–43.
11. Donner A: Approach to sample size estimation in the design of clinical trials—a review. Stat Med 1984;3:199–214.
12. George SL, Desue MM: Planning the size and duration of a clinical trial studying the time to some critical event. J Chronic Dis 1974;27:15–24.
13. Rosner B: *Fundamentals of Biostatistics*, 3rd ed. PWS-Kent Publishing Company, Boston, 1990.
14. Harris MI, Flegal KM, Cowie CC, et al: Prevalence of diabetes, impaired fasting glucose, and impaired glucose tolerance in U.S. adults. The Third National Health and Nutrition Examination Survey, 1988–1994. Diabetes Care 1998;21:518–524.
15. Snedecor GW, Cochgran WG: *Statistical Methods*, 7th ed. Iowa State University Press, Ames, 1980.
16. Simes RJ: An improved Bonferroni procedure for multiple tests of significance. Biometrika 1986;73:751–754.17.
17. Westfall PH, Young S: P value adjustments for multiple tests in multivariate binomial models. J Am Stat Assoc 1989;84:780–786.
18. Westfall PH, Krishnen A, Young SS: Using prior information to allocate significance levels for multiple endpoints. Stat Med 1998;17:2107–2119.
19. Moyé LA: Alpha calculus in clinical trials: considerations and commentary for the new millennium. Stat Med 2000;19:767–779.
20. Moyé LA: Alpha calculus in clinical trials: considerations and commentary for the new millennium. Rejoinder. Stat Med 2000;19:767–779.
21. D'Agostino RB: Controlling alpha in clinical trials; the case for secondary endpoints. Stat Med 2000;19:763–766.
22. Koch GG: Alpha calculus in clinical trials: considerations and commentary for the new millennium [discussion]. Stat Med 2000;19:781–784.
23. O'Neill RT: Alpha calculus in clinical trials: considerations and commentary for the new millennium [commentary]. Stat Med 2000;19:785–793.
24. The Clopidogrel in Unstable Angina to Prevent Recurrent Events Trial Investigators: Effects of clopidogrel in addition to aspirin in patients with acute coronary syn-

dromes without ST-segment elevation. N Engl J Med 2001;345:494–502.
25. Mills JL: Data torturing [see comments]. N Engl J Med 1993;329:1196–1199.
26. Yusuf S, Wittes J, Probstfield J, et al: Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. JAMA 1991;266: 93–98.
27. Moyé LA, Deswal A: Trials within trials: confirmatory subgroup analyses in clinical trials. Control Clin Trials 2001;22:605–619.
28. Moyé LA, Power JH: Evaluation of ethnic minorities and gender effects in clinical trials: opportunities lost and rediscovered. J Natl Med Assoc 2001;93(Suppl):1–6.
29. Owen DB: *On the History of Probability and Statistics*. Marcel Dekker, New York, 1976.