2

Search Versus Research

2.1 Introduction

As we saw in Chapter One, one of the difficulties of collecting a simple random sample was that it was not enough for the sample to be selected completely at random. In addition, the sample must be collected with a particular goal, or research question in mind. Just as one rarely has a successful vacation by simply packing up the car and leaving without knowing the destination, random-sampling without a research plan leaves most every desirable goal out of reach. And, just as leaving for one destination precludes seeing others, sampling that is designed to answer one question often makes it impossible to shed important light on other questions, even though they contain scientific interest of their own accord.

However, there is a second problem with carrying out an unplanned analysis suggested by the data. Like an iceberg, it is this second, submerged component that is commonly the most damaging when not recognized.

2.2 Catalina's Dilemma

Consider the following hypothetical example that is an illustration of a commonly occurring issue

2.2.1. Can Catalina Generalize?

An enthusiastic young researcher, Catalina, is interested in demonstrating the progressive deterioration in heart function observed in patients with mild CHF whose medical management is appropriate. She has designed a research program to examine the changes in left ventricular function over time, with specific attention to examining changes in end diastolic volume (EDV). During a six-month time period, Catalina recruits her sample randomly from the population of people with CHF seen at her hospital, and follows each patient for two years. For each patient, she measures heart function at baseline (i.e., when the patient agrees to enter the study) and again in 24 months. Every patient returns to have their heart function measured at the two-year time point. Although Catalina is focused on the change in EDV, the technology of the measuring tool permits her to obtain estimates of left ventricular

ejection fraction (LVEF), end systolic volume (ESV), stroke volume (SV), and cardiac output (CO) as well.

Her colleagues who have contributed their own patients to Catalina's investigation are anxious to learn of the conclusions of her study. At the anticipated time, Catalina examines her data with great anticipation.* Comparing the baseline to 24 month change in EDV for each of her patients, she discovered to her surprise that the anticipated increase in EDV did not materialize. However although EDV has not increased, there has been a substantial increase in ESV. She therefore chooses to change the endpoint for the program from EDV to ESV. She presents her results in terms of the change ESV, saying little about the absence of a change in EDV.

2.2.2 Do Logistical Issues Block Generalization?

Admittedly, many researchers would have no problem with Catalina's last minute endpoint change. They would argue that since ESV and EDV reflect measurements of the same underlying pathophysiology and jointly measure the progress of CHF, they should be interchangeable as endpoints. In fact, they would applaud Catalina's doggedness in looking beyond the disappointing EDV finding, ferreting out the hidden change in ESV.

Others who have read Chapter One would argue that the analysis of ESV was inappropriate because the study was not designed to detect changes in ESV (even though that is exactly what the study did). These critics, drawing on the arguments demonstrated in Figures 1.1 and 1.2 would point out that a sample that was optimum for assessing the change in EDV might not be optimum in detecting the change in ESV. For example, a comparison of the standard deviation of the two measurements might lead to a different minimum sample size for the two analyses.

However Catalina, anticipating this argument, examines the data that led her to choose the sample size that she ultimately selected for the study. She finds that the sample size necessary for detecting the change in EDV would be sufficient for detecting changes in ESV. In fact, after a thorough examination of the design issues that would preface the analysis for each variable, she concludes that she would have made no change in design of the study if she had set out to detect changes in ESV rather than EDV. As it turns out, the sample was optimally selected to analyze either of these two variables. Does removal of the logistics argument now permit Catalina to generalize the ESV results?

2.2.3 Disturbed Estimators

Even though the logistical impediment to generalization has been removed, problems with the generalizability argument persists. There remains a critical difficulty with the statistical estimators that Catalina uses to measure the effect in ESV in her study — a difficulty that is induced by the change in endpoint from EDV to ESV.

---

* Since no intervention is being provided and she is not required to carry out any interim monitoring of her patients, Catalina can wait until the end of the two-year follow-up period to examine her data.

The decision to change the endpoint from EDV to ESV was based solely on the findings in the sample. Specifically it was not Catalina's foresight, but the data, that suggested ESV should be analyzed (because its results were positive) and not EDV. This data-based change introduces a new effect of sampling error, and the statistical estimators that we use (means, standard deviations, odds ratios, relative risks, confidence intervals, and *p*-values) are not designed to incorporate this effect. Specifically, her statistical estimators do a fine job of estimating the mean change when the variable is fixed and the data are random. However, when the variable itself is random (i.e., randomly chosen) these estimators no longer fulfill their functions well.

Did Catalina choose the ESV variable randomly? From her point of view, no. However, the sample chose it for her, and the sample contains random sample-to-sample variability.

To examine this issue further, let's say that another investigator (Susi) sampled from the same population as Catalina. Like Catalina, Susi chooses EDV for her endpoint. At the conclusion of Susi's study, her data reveal that, as was the case for Catalina, EDV did not change over time. However neither did ESV. For Susi's sample, it was the variable SV that changed over time. She therefore chooses to report the change in SV as the major endpoint in her study.

Finally, Al, a third investigator, sampling from the same population and like his two colleagues, focused on EDV as the variable of interest, finds that neither EDV, ESV, nor SV changed. For him, it was the change in CO that was positive. Thus, the three different researchers report their three different findings, (Catalina reports ESV, Susi reports SV, and Al reports CO). Nobody reports EDV, which was the prospectively identified endpoint chosen by each of these researchers. How can these results be interpreted?

Each investigator acts as though they can have complete confidence in their statistical estimators. However, for each there are now two sources of variability where there was only supposed to be one. The first source of variability is the variability of the measurements from subject to subject — easily anticipated and easily handled by the statistical estimators. These estimators incorporate this component well. The sample mean and standard deviation are accurate, a test statistic is computed, and the *p*-value nicely incorporates this subject-to-subject variability.

However, there is another source of variation that was never anticipated that is present in these research efforts — the variability of the endpoint selection. Each investigator selected, independent of the data, EDV as their endpoint. However, each investigator allowed the data to select another endpoint for them. Yet the selection mechanism was a random one, since each data set exhibits sample-to-sample variability.

Essentially, in the case of each of these investigators, the data have provided an enticing answer to a question that the researcher didn't think to ask. When the data determine the analysis, as in this case, our commonly used statistical estimators (i.e., means, standard deviations, confidence intervals, and *p*-values) do not function reliably. They were never designed to apply to this scenario, and the familiar formula for these quantities are no longer accurate. What has dismembered the

formula is that there are now two sources of error, when they were designed to handle only one. This is the hallmark of *exploratory analyses*, or *random research*.

2.3 Exploratory Analysis and Random Research

Exploratory analysis is the process by which the investigator allows the data to answer specific questions that the investigator did not plan to use the data to address. There are two problems with exploratory or *hypothesis-generating research*. The first is that commonly, the sample is not an optimal one, since the investigator can only design a sample to answer questions that they knew to ask.

The second difficulty is a more pernicious one, requiring additional elaboration. We pointed out earlier that the careful selection of a sample to address the scientific question of interest does not prevent random-sampling error from generating the sample's answers. In order to measure the role of sampling error accurately, the investigator turns to the mathematical procedures supplied by statistics. From statistics, we find the computations that convert the sample's information (the data) into the best estimates of effect size (e.g., means or other measures of effect size, standard deviations, confidence intervals, and *p*-values). Researchers rely on the accuracy of these estimators to inform them about the population from which the sample was drawn

It is important to note that these estimators do not remove sampling error. Instead, they channel this sampling error into both the effect size estimates (e.g., means) and the variability of these estimates (e.g., standard deviations and confidence intervals). If the researcher is also interested in inference (i.e., statistical hypothesis testing), then statistical procedures will channel sampling error into *p*-values. Thus, when used correctly, statistical methodology will appropriately recognize and transmit sampling error into familiar quantities that researchers can interpret (Figure 2.1).

Unfortunately, these familiar estimators are corrupted when there is a source of random variability beyond that produced by sampling error. In the case of our investigator Catalina, the second source of variability is produced by random analysis selection.

Consider the simple example of the sample mean. If we select observations $x_1, x_2, x_3, \ldots x_n$, from a population, we commonly use

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

as an estimate of the population mean. We have come to accept its precision and reliability in many applications of research, and its ubiquity represents the confidence we have in this formula. It has taken statisticians a long time to figure out how to estimate from a sample [1].*

* The idea of repeating and combining observations made on the same quantity appears to have been introduced as a scientific method by Tycho Brae toward the end of the sixteenth century. He used the arithmetic mean to replace individual observations as a summary measurement. The demonstration that the sample mean was a more precise value than a single

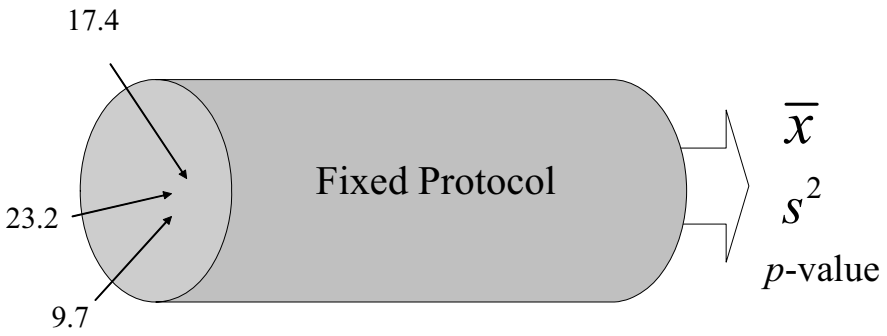

Fig. 2.1. A fixed, prospectively anchored protocol accepts random data and channels it into reliable, statistical estimators.

However, this formula was designed to work in a setting when there is only one source of error, sampling error. Specifically, it was created to work in the setting where the variable x is fixed, and then the sample of x's is selected randomly. While this is true in many circumstances, it is not the case with Catalina's EDV-ESV analysis. She did not choose the ESV analysis, but instead, allowed the ESV analysis to be selected for her by the data.* Thus, her original question has been supplanted by her observation that the ESV did change. This replacement was suggested to the investigator by the data. She would not have selected the ESV evaluation if the data didn't suggest that it deteriorated over time.

The sample data produced an answer to a question that the researcher had not asked, inducing the researcher to ask the question in a *post hoc* fashion. The effect of this data-driven choice is that the data contain sampling error. Therefore, as demonstrated with the experiences of Susi and Al, other datasets would produce

---

measurement did not appear until the end of the seventeenth century, and was based on the work by the astronomer Flamsten. At this point, the simultaneous organization of the formulation of discrete probability laws and the development of the differential calculus permitted an analytical examination of the probability distribution of the mean. In 1710, Simpson's work proved that the arithmetic mean has a smaller variance than a single observation.

* We can assert this claim because the investigator would never have highlighted the ESV analysis if its results did not stand out. The fact that she was drawn to the evaluation because of its magnitude and not because of any prospective plan to look at it is the mechanism of a data-driven analysis.

other intriguing results due to sampling error. Thus, different samples obtained from the same population would provide not just a different answer to the EDV deterioration over time, but, would also supply other "answers" to questions that had not been asked. Since the data are random and the data are proposing the research analyses, the analyses themselves are random.

In the random research environment, statistical estimators, e.g., the sample mean, lose their best properties of precision and reliability. They are constructed to operate best when 1) the analysis is chosen and fixed prospectively and 2) the resulting data are random. Their performance degrades when the analysis plan itself is random, distorting all measures of the magnitude and significance of the relationship. Operating like blind guides, they mislead us in the random research environment about what we would see in the population based on our observations in the sample (Figure 2.2). Therefore, since we do not have good estimators of the effect of interest in the population, the best that we can say is that these *post hoc* findings are exploratory, and require confirmation.

2.4 Gender–Salary Problem Revisited

Returning to the initial salary example, recall that this evaluation carried out by our researcher in Chapter One found that the salaries of female physicians were larger than the salaries of physicians who were men. We already understand that the impact of missing gender data can skew her analysis. The presence of this missing data means that we cannot even be sure what has happened in our sample, much less try to extend the sample result to the population.

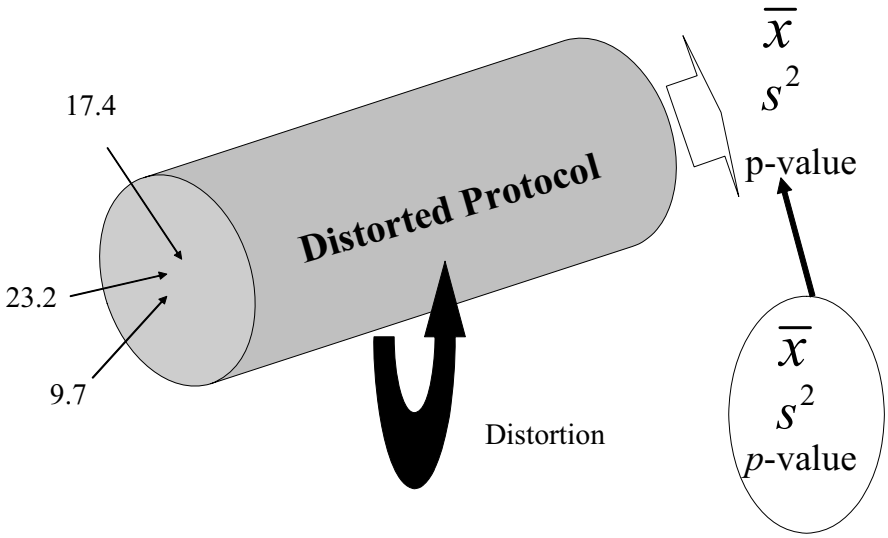

Fig. 2.2. A protocol, perturbed by a data-based, as opposed to a prospectively chosen-analysis plan, distorts the statistical estimators.

However, now assume that there is no missing gender data in the sample. Thus, the gender–salary analysis includes data from everyone in the sample. As we might anticipate from the previous discussion in this chapter, even though the sample database is complete, this gender–salary analysis is still likely to be misleading i.e., not at all representative of the population.

The unreliability of this sample-based result is rooted in the way in which the scientist's attention was drawn to the gender–salary relationship. Unlike with the overall salary evaluation that was designed a priori, the gender–salary analysis was unplanned. The researcher did not design the study to find this relationship. Instead she designed the study to obtain another measure, and was drawn to the gender salary relationship by its magnitude. In fact, the investigator was drawn to this finding by carrying out several non-prespecified evaluations, thereby discovering what the sample revealed about gender and salary. Since the source of this influence is sampling error, its presence generates misleading estimators (Figure 2.3).

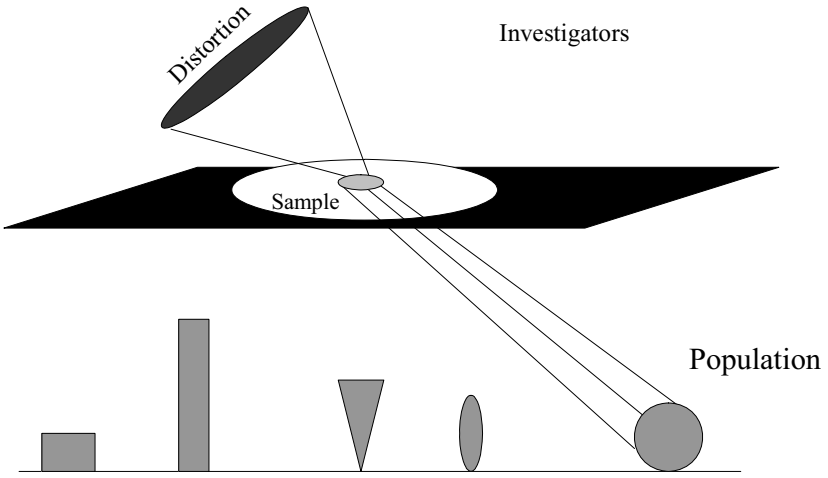

Fig 2.3. In the random research paradigm, a misdirected sampling scheme, in concert with untrustworthy estimators, misleads the investigators about the characteristics of the relationship. This is the hallmark of exploratory analyses.

As was the case with Catalina, this investigator did not choose the analysis. The data chose the gender–salary evaluation for her. Thus, her original question has been replaced by the question of the equality of salaries between men and women physicians. This replacement was suggested to the investigator by the data. Specifically, this means that the researcher would not have selected this question if the data did not suggest that the salaries of men and women were different.

As was the case with Catalina's cardiology research, the sample data produced an answer to a question that the researcher had not asked, persuading the researcher to ask the question in a *post hoc* fashion. She is not aware that other datasets will, due to the influence of sample-to-sample variability, generate other enticing results. For example a second data set might find no disparity at all between the salaries of male and female physicians but instead "find" a racial disparity, attracting the investigator's attention. A third sample may find neither of the previous two, but find a disparity in salary by zodiac sign. Thus, different samples obtained from the same population would provide not just a different answer to the gender–salary relationship question, but in addition, would supply other "answers" to questions that had not been asked. Since the data are random, and the data are proposing the research analyses, the analyses themselves are random. The estimators are derived to have the data selected randomly from sample to sample, not for the analysis plan itself to exhibit sample-to-sample variability.

A follow-up analysis designed to look at the gender–salary disparity would (1) choose the optimum sample (i.e., a sample with enough males and females), and (2) have the gender–salary question determined *a priori* (Figure 2.4).

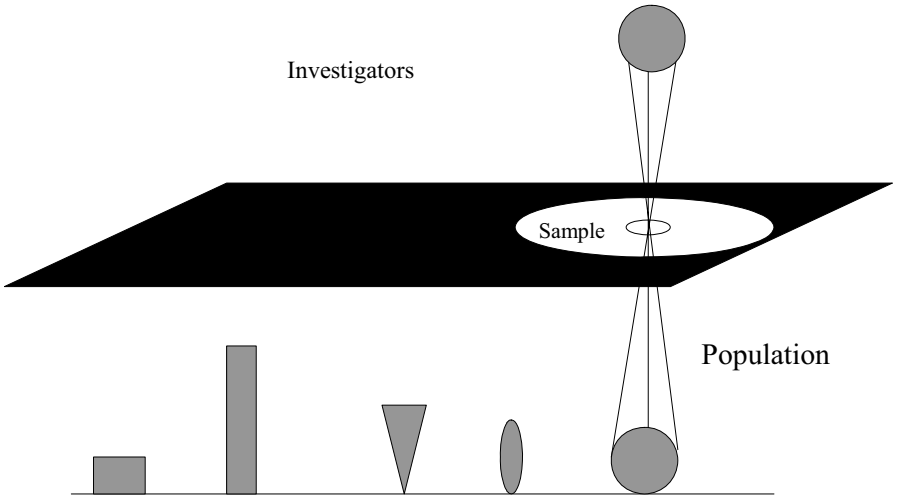

Fig 2.4. In the confirming analysis, the sample is specifically designed to focus on the aspect of the population of interest. The fixed analysis plan allows the statistical estimators to perform optimally. This is the hallmark of a confirmatory analysis.

2.5 Exploratory Versus Confirmatory

Both the hypothetical example from cardiology and the salary survey demonstrate the hallmark of exploratory analyses. By exploring, the investigator identifies relationships that were not anticipated but that the sample suggests are present. However, the presence of a relationship in the sample does not announce the presence of

the relationship in the population. Additionally the statistical investigators were not designed to function when the analysis is random. They therefore add another level of distortion to the effect "identified" by the exploratory work.

The confirmatory evaluation provides the clearest measure of the magnitude of the effect of interest. Having its location determined by the exploratory analysis, the sample is optimally configured for the relationship that was suggested in the *post hoc* evaluation. Also, with the analysis fixed (i.e., the variable in which interest lies has been chosen prospectively and plans for its evaluation are already in place) the statistical estimators perform well, providing a reliable sense of effect magnitude and the degree to which that magnitude may vary from sample to sample. Generalization from the sample to the population is strongest when it rests on a confirmatory finding.

However, generalization should not be attempted when its basis is exploratory analysis. In the setting, the usual sample estimators are undermined because the assumption on which their accuracy is based is false. The two sources of variability wreck the capacity of our commonly used estimators to provide reliable estimates of the true population measures. This is the trait of exploratory estimators. Unfortunately, once the random research paradigm is in place, it can be very difficult to repair the exploratory analysis. While they can occasionally provide some light on the answer to a question the researcher did not ask, they require confirmation before we can accept the results.

2.5.1 Cloak of Confirmation

We will soon see how exploratory or hypothesis-generating research can be most useful. However, one area in which it is harmful is when it is represented as confirmatory research. This is what the salary researcher of Chapter One and Catalina of this chapter intended to do. Each believed that their research results were accurate and worthy of generalization to the population at large. However, further evaluations revealed that only a portion of the results (the mean salary for the entire sample for the salary researcher, and the EDV results for Catalina) can be generalized; the residual was exploratory or hypothesis-generating.

Exploratory analyses are dangerous when they are covered with the cloak of confirmation. This misrepresentation of exploratory analyses as confirmatory can be dangerous, misleading, and must be identified at once. The need for the research result can be overwhelming; yet, its correct identification (exploratory or confirmatory) requires vigilance, patience, and discipline. As Miles [2] points out "If the fishing expedition catches a boot, the fishermen should throw it back, not claim that they were fishing for boots."

As an example, consider the plight of a young parent who discovers that his child is sick. An emergency visit to the pediatrician reveals that the child suffers from an acute illness, but one that can be easily treated with the prompt use of a prescription medicine. Minutes later, the pharmacist reviews the prescription, telling the parent that the required medication is a combination of three compounds that can be quickly mixed and administered. The pharmacist returns with the preparation, and, anxious to follow the doctor's orders, the parent prepares to give the child the first teaspoon of medication right there in the pharmacy. However, just

when the teaspoon is placed on the child's lips, the pharmacist rushes out, telling the parent that, although the medicine contains the right constituents, the proportions of those compounds are wrong because the device he used to mix them is defective.

The exploratory estimator, like the defective medication mixture, is a distorted and unusable concoction that should be avoided. Most parents would steel themselves, withdrawing the teaspoon containing the defective medication. Just as the bad medicine cannot be given, the researcher must exert discipline and avoid hasty interpretation of the exploratory estimator. It is the fortunately rare (and perhaps, rarely fortunate) parent who would insist on giving the defective compound to their child in the face of this news.

Now that the parent recognizes that the compound is faulty, what steps can be taken to correctly adjust the preparation at hand. Should it be diluted? If so, by how much? Should additional compounds be added? If so, in what quantities should they be added? Since the precise defect in the compound's formulation cannot be identified, the parent only knows that the medication is defective and that he cannot correct it. All he can do is ask the pharmacist to dispose of what he has in hand and then start the process again, this time using the required compounds in the right proportions.

Similarly, an exploratory estimator cannot be repaired. We only know that a critical assumption in the estimator's construction has been violated. Since we cannot rehabilitate the estimator, we can only ask that a study be carried out that does not violate the estimator's assumptions. Specifically, this means that the study is (1) designed to answer the prospectively asked question, and (2) the study is executed and its data analyzed as described in the protocol (concordant execution). In this paradigm, the estimators are trustworthy measures of population effects.

2.6 Exploration and MRFIT

A prime example of the harm that comes from exploratory analyses that are represented as confirmatory is one of the results from the Multiple Risk Factor Intervention Trial (MRFIT) [3] study. Published in 1982, it was designed to demonstrate that reductions in the risk factors associated with atherosclerotic cardiovascular disease would be translated into reduction in clinical events, e.g., myocardial infarction and stroke. Patients in the intervention group received treatment for elevated blood pressure, joined cigarette smoking cessation programs, reduced their weight, and lowered their serum lipid levels; patients in the control group followed their usual accepted standard of living. At the conclusion of the study, the investigators found and reported that there was no difference in clinical outcome between those patients who received risk factor intervention and those who did not.

These null findings were a disappointment to the risk factor interventionists, who then poured over the data to identify if any effect could be found in a fraction of the patients that might explain the null overall effect. They found one, and it was a bombshell. When the researches ignored the results in the entire randomized cohort (i.e., all randomized patients), and concentrated on men who were hypertensive and had resting electrocardiograph (ECG) abnormalities at baseline, they dis-

covered that these patients had a worse outcome when randomized to antihypertensive therapy than those who received no such therapy.

This result was published, and had a major impact on the momentum to treat essential hypertension. At this time in clinical medicine, the importance of identification and treatment of hypertension galvanized physicians. Screening programs were well underway. New therapies (e.g., hydrochlorothiazide, alphamethyldopa, and clonidine) for the hypertensive patient became available. All of the necessary forces for a war on undiagnosed and untreated hypertension were maneuvering into position when the MRFIT analyses were released. This finding slowed the momentum for the treatment of hypertension by raising disturbing, and ultimately unhelpful, questions e.g., "Maybe not all hypertension was bad after all?", or "Maybe hypertensive disease itself was bad, but the treatment was worse?"

The real question however, was, "Is it just a distorted treatment effect?" For years after this finding, clinical trials in hypertension were forced to address this unusual result. None of the major studies ever found that hypertensive men with resting ECG abnormalities were better off when their hypertension remained unchecked. Nevertheless, an exploratory analysis, dressed as a confirmatory one, produced an important interruption in the treatment of a deadly cardiovascular disease.

2.7 Exploration in the ELITE Trials

Another, more recent example of the misdirection produced by exploratory analyses are the ELITE trials. There are many medications used to treat heart failure, one of which is angiotensin-converting enzyme inhibitor (ACE-i). This effective CHF medication's use dramatically increased in the 1980s. Unfortunately, many ACE-i treated patients experience undesirable side effects of this therapy; among the worst of these is renal insufficiency.

As a response to this undesirable side effect profile, angiotensin II type I receptor blockers were developed. In order to compare the relative safety of angiotensin type II type receptor blocker to ACE-i therapy, the Evaluation of Losartan in the Elderly Study (ELITE) was undertaken [4]. The primary analysis of ELITE was the comparison of the two drug's abilities to preserve renal function.

ELITE recruited 722 patients and followed them in for 48 weeks. At its conclusion, ELITE investigators determined that kidney function was equally preserved by the two medications. However, the investigators discovered that 17 deaths occurred in the losartan group and 32 deaths in the captopril group (p = 0.035). This finding received the principle emphasis in the discussion section of the manuscript. Although the need to repeat the trial was mentioned in the abstract, the balance of the discussion focused on the reduced mortality rate of losartan. According to the authors, "This study demonstrated that losartan reduced mortality compared with captopril; whether the apparent mortality advantage for losartan over captopril holds true for other ACE inhibitors requires further study." Others even went so far as to attempt to explain the mechanism for the reduction in sudden death observed in ELITE 1 [5, 6].

To the investigators' credit, ELITE II [7] was executed to confirm the superiority of losartan over captopril in improving survival in patients with heart fail-

ure. The primary endpoint in ELITE II was the effect of therapy on total mortality. This study required 3,152 patients (almost five times the number of patients recruited for the ELITE I) and also had to follow patients for 18 months (almost twice as long as the duration of follow-up in ELITE I). At the conclusion of ELITE II, the cumulative all-cause mortality rate was not significantly different between the losartan and captopril groups. The investigators conceded "More likely, the superiority of losartan to captopril in reducing mortality, mainly due to decreasing sudden cardiac death, seen in ELITE should be taken as a chance finding."

Although the finding in ELITE I may have been due to chance alone, the principle difficulty presented by the first study was that the statistical estimators commonly used to measure the mortality effect were inaccurate when applied to this surprise finding. However, since the sample was random (selected as one of millions of possible samples from patients with CHF), the selection mechanism for the analysis is random (since other samples would have produced other unanticipated findings). Specifically, by allowing their focus to be shifted to surprise mortality effect, the research paradigm became a random one (Figure 2.2). In this random analysis setting, the usual statistical estimators provide misleading information about the population effect from the observed findings in the sample. This is the hallmark of the random protocol. By letting the data decide the analysis, the analysis and the experiment becomes random, and the resulting statistical estimators become untrustworthy.

2.8 Necessity of Exploratory Analyses

Recognizing the importance of a guiding hypothesis is critical to the scientific thought process. This central hypothesis generates the finely tuned research design that, when executed per plan, permits a clear, defensible test of the core scientific hypothesis. This deliberative procedure stands in stark contrast to discovery or "exploration" whose use in sample-based research has limits as pointed out here and elsewhere [8,9].

Nevertheless, it is quite undeniable that discovery is important to science in general and in healthcare research in particular. Such findings can have important implications. The arrival of Christopher Columbus at the island of San Salvador led to the unanticipated "discovery" of the New World. Madam Curie "discovered" radiation. These researchers did not anticipate and were not looking for their discoveries. They stumbled upon them precisely because these discoveries were not in their view. However, as these prominent illustrations demonstrate, despite the weaknesses of the exploratory process, discovery has and will play an important role in healthcare research. This undeniable importance of this style of research is aptly demonstrated by the example of compound 2254RP.

2.8.1 Product 2254RP

During the height of World War II in France, Janbon and colleagues at the Infectious Disease Clinic of the Medical School in Montpellier quietly studied the effects of a compound that held out promise as an antibiotic.* An offshoot of the new class of sulfonamides, compound 2254RP was quite possibly a new treatment for typhoid fever. However the researchers were unable to maintain focus on the antimicrobial abilities of this agent because of its production of seizures. Even patients with no known medical history of epilepsy would commonly experience profound convulsions after the institution of 2254RP.

Further evaluation of these patients revealed that seizures were more likely to occur in patients who were malnourished. This unanticipated findings generated further queries, revealing that exposed patients were also rendered hypoglycemic by the compound.

Puzzled, Jabon transmitted these observation to a colleague, Auguste Loubatières who himself was engaged in research on the characteristics of seizure disorders in patients exposed to high concentrations of insulin. Loubatières hypothesized that both insulin and 2254RP produced hypoglycemia. After demonstrating the sequence of hypoglycemia followed by seizures in dogs treated with 2254RP, he verified the induction of hypoglycemia in three female patients by the compound. This collection of research efforts generated the development of the sulfonylureas as oral hypoglycemic agents in diabetes mellitus.

2.8.2 The Role of Discovery Versus Confirmation

Certainly, the production of hypoglycemia and seizure disorder was not part of the research protocol for Dr. Janbon and colleagues, and their unanticipated findings fell into the category of exploratory research. Janbon and colleagues would not be criticized for pursuing the surprise findings of their research efforts. The discovery that the antihypertensive agent minoxidil could unexpectedly reduce hair loss, and the finding that the antihypertensive, anti-anginal compound sidenafil can temporarily reverse erectile dysfunction are contemporary examples of the fruits of discovery.

Discovery must certainly play a major role in an environment where compounds have unanticipated and sometimes, even unimagined effects. The difficulty in sample-based research arises in separating true discovery from the misdirection of sampling error. What distinguished the "discovery" in MRFIT (that some hypertensive men should not be treated for their hypertension) from the discovery that the sulfonylureas had effects above and beyond antimicrobial abilities was confirmation. The MRFIT finding could not be confirmed. The 2254RP findings were. Therefore, discovery must be confirmed before it can be accepted as a trustworthy. Eccentricities of the discovery process e.g., faulty instrumentation, sample-to-sample variability, and outright mistakes in measurements and observations can each mislead honest researchers and their audience.

---

* This is taken from Chapter 6, Sulfonylurea Receptors, ATP-Sensitive Potassium Channels, and Insulin Secretion from LeRoith D, Taylor SI, Olefsky JM (2000). *Diabetes Mellitus: A Fundamental and Clinical Text. Second Edition.* Philadelphia: Lippincott Williams, and Wilkins.

This integration can be achieved by setting the discovery as the central *a priori* hypothesis for a new experiment that seeks to verify the discovery. For example, while it is true that Columbus discovered the new world, it is also true that he had to return three additional times before he was given credit for his discovery.[*]

Additionally, given that claims of discovery can commonly misdirect us, it is important to distinguish between an evaluation that uses a research effort to confirm a prospectively stated hypothesis, a process that we will define as *confirmatory analysis* (truly "*re*-searching") versus the identification of a finding for which the research effort was not specifically designed to reliably detect ("searching").

2.8.3 Tools of Exploration

While confirmatory analyses are prospectively designed, focused, and disciplined, hypothesis-generating analyses have other characteristics. Untethered by any early planning, they employ tools of elementary pattern recognition, requiring an open mind, and sometimes, a spark of imagination. Such analyses include, but are not limited to (1) changing a study's endpoint (i.e., exploring a new endpoint), (2) changing the analysis of an endpoint,[†] (3) subgroup analyses,[‡] and (4) data-based model building.[§] This type of investigational perspective is prevalent in healthcare research, is exciting to carry out, and is almost always interesting.

However, whenever the execution of a research effort is altered due to an unanticipated finding in the data as is the case with 1–4 above, the protocol becomes random, the research becomes discordant (i.e., its execution is no longer governed by the prospectively written protocol) and the analyses are hypothesis-generating. In these cases we must tightly bind the exploratory conclusions with the lock of caution until a subsequent confirmatory analysis can unlock and thereby generalize the result. To enforce this point, these exploratory, or hypothesis-generating results should be reported without *p*-values. *Z* scores would suffice very nicely here, since they provide a normed effect size, without mixing in the sampling error issue.

2.9 Prospective Plans and "Calling Your Shot"

It comes as no surprise that the advice from research methodologists is that exploratory or surprise findings do not carry persuasive weight primarily because they were not planned prospectively [10]. However, to many researchers, this requirement of "calling your shot," i.e., of identifying prospectively what analyses will have persuasive influence, seems much ado about nothing. After all, the data are, in the end, the data. To these critics, allowing the data to decide the result of the experiment can appear to be the fairest, least prejudicial evaluation of the message

---

[*] Columbus was essentially forced by Queen Isabelle of Spain to prove that he could find the New World when he was actually looking for it. Only when he did this three times (driving himself into poverty during the process) and other ship captains confirmed its location, was the New World finally accepted.

[†] For example, changing the evaluation of a 0-1 or dichotomous event to take into account the time until the event occurred (e.g., life table analysis).

[‡] Subgroup analyses is the subject of a later chapter.

[§] Regression analysis is discussed in Chapter Eleven.

they contain. When these debates involve patient well-being, the discussions can be explosive. Consider the case of carvedilol.

2.9.1 The US Carvedilol Program

In the late 1980s interest in the heart failure research community focused on the use of beta blockers. Considered anathema for the treatment of chronic CHF, re-investigation of the issue suggested that beta blockade could be useful in relieving the symptoms of heart failure. The US Carvedilol program [11] evaluated the medication carvedilol (previously approved for the treatment of essential hypertension) for the treatment of CHF.

In these research efforts 1,094 patients were selected for entry into one of four protocols, then randomized to either standard therapy plus placebo, or standard therapy plus carvedilol. There were 398 total patients randomized to placebo and 696 to carvedilol. At the conclusion of approximately one year of follow-up, 31 deaths had occurred in the placebo group and 22 in the active group (relative risk = 0.65, p-value = 0.001). The program's oversight committee recommended that the program be terminated in the face of this overwhelming mortality benefit. Both the investigators and the sponsor believed that since mortality effects were important, the beneficial effect of carvedilol on the total mortality rate should compel the federal Food and Drug Administration (FDA) to approve the compound as effective in reducing the incidence of total mortality in patients with heart failure.

However, FDA review of the program revealed some aspects of the carvedilol research effort that were not clearly elucidated in the *New England Journal of Medicine* manuscript. Although that paper correctly stated that patients were stratified into one of four treatment protocols [1, page 1350], the manuscript did not state the fact that each of these protocols had its own prospectively identified primary endpoint, nor did it acknowledge that the total mortality rate was neither a primary nor a secondary endpoint for any of the trials. Three of the trials had exercise tolerance as a primary endpoint and measures of morbidity from CHF as secondary endpoints. Furthermore, a protocol-by-protocol delineation of the primary endpoint results was not included in the manuscript.

Additional interrogation revealed that the finding for the prospectively defined primary endpoint in three of the studies were statistically insignificant ($p > 0.05$). The fourth study had as its primary endpoint hospitalization for heart failure (the statistical analysis for this primary endpoint was $p < 0.05$). Each of these four studies had secondary endpoints that assessed CHF morbidity, some of which were nominally significant ($p < 0.05$), others not. However, as pointed out by Fisher [12], total mortality was not an endpoint of any of the four studies in the program. Thus a research effort that at first glance appeared to be a single, cohesive trial with total mortality as its primary endpoint, was upon close inspection a *post hoc* combined analysis for a non-prospectively defined endpoint.

This observation, and subsequent arguments at the FDA advisory committee meeting produced a host of problems for the experiment's interpretation. The verbal debates that ensued at the Maser Auditorium in May 1996 between the Sponsor's representatives and the Advisory Committee were intense. The sponsor was taken aback by the unanticipated strong criticisms of its program with its strong

mortality findings. Viewing the criticisms as small arguments over Lilliputian technicalities, the sponsor responded quickly and vehemently. The contentions spilled over from verbal debate into the literature, first as letters to the editor [13,14] and then as full manuscripts in their own right [12,15,16,17].

The proponents of carvedilol assembled a collection of compelling arguments both during and after the disputatious FDA debate. There was no question about the strength of the US Carvedilol's programs findings. Lives were prolonged for patients who had a disease that had no known definitive therapy and whose prevalence was rapidly rising. The Data Safety and Monitoring Committee* of the program had stated that, in the face of the large mortality benefit, it was unethical for the trial to continue merely to demonstrate an effect on its weaker endpoints. The FDA was reminded that the most compelling of all endpoints in clinical trials is total mortality, and the most laudable goal was prolonging life. There was no doubt that carvedilol had produced this precise result in the US Carvedilol Program.

In addition, the research community was asked to recall that at the outset of a research effort, investigators cannot be expected to identify the totality of the analyses that they will want to carry out at the study's conclusion. While scientists will certainly want to have a prospectively asked question that motivate the generation of the sample, the logistical efficiency of the research effort requires that they collect data extending above and beyond the prospectively declared analyses. When the manuscripts are submitted for publication, the authors must affirmatively reply to persistent journal reviewers and editors who may like to see additional analyses carried out so that the reader will have all relevant information required to make an independent judgment about the intervention's risk-benefit balance. To do anything else would open the investigators to the criticism of concealing critical information.

Finally, the sponsors were obligated to collect safety data. It was the FDA, the Advisory Committee was told, that required the US Carvedilol Program to collect mortality data in order to provide assurances that carvedilol did not shorten lives. Was the FDA, they asked, now to ignore the data they themselves demanded when the data demonstrated an unsuspected benefit of the drug? †

The US Carvedilol Program had complied with the requirements of the FDA in the design of the program and the studies' interim monitoring. From its advocates' point of view, scientists who criticized the US Carvedilol Program for not finding a statistically significant reduction in weaker endpoints e.g., exercise tolerance when carvedilol was found to prolong lives was tantamount criticizing a baseball player for not stealing any bases when all he did was hit home runs.

---

* The Data Safety and Monitoring Committee is a collection of distinquished scientists responsible for evaluating the interim results of the study to ensure the ethical treatment of its participants. Unblinded to therapy assignment, and privy to all of the data, this committee can recommend that the study be prematurely terminated if unanticipated harm or benefit has occurred.

† Certainly, if carvedilol had been shown to increase mortality, a clear hazard to patients, the sponsor's request for approval for a CHF indication would have been denied, regardless of the findings for the programs official panoply of primary morbidity endpoints. The asymmetry of safety findings is discussed in Chapter Eight.

However, critics of the program were equally vehement. There was no question about the findings of the US Carvedilol program. Lives had been saved. But, what did this collection of studies of just over 1,000 patients imply about the population of millions of patients with heart failure? Sample extrapolation is a dangerous process, and discipline, not hope should govern what sample results should be extended from the sample to the population. Samples are replete with "facts'; most of these facts apply only to the sample and not to the entire population. Healthcare has seen these kinds of failures of generalization before. Experience (e.g., that of MRFIT) suggested that the primary outcome analyses most likely represented the effects of carvedilol in the larger population.

Additionally, pre-specification of the anticipated analysis in the protocol of a trial has been an accepted standard among clinical trial workers [18] and certainly must be included in a manuscript describing that trial's results. In addition, the non-reporting of non-significant endpoints in clinical trials has been criticized [19]. Each of these principles was clearly violated in the manuscript published in *The New England Journal of Medicin*e. The scientific community expects, and clinical trial workers require that analyses be provided for all prospectively stated endpoints. The fact that the results of a program claiming major benefit did not specifically define and report the analysis of the primary endpoint is a serious deficiency in the manuscript that purports to describe the effects of therapy.

Unfortunately, by violating these fundamental methodology tenets, the Carvedilol investigators open themselves to the criticism that they selected the total mortality analysis because of its favorable results, thus biasing the conclusions[*] and tainting the research effort. The mortality findings for the US Carvedilol were a surprise finding. They were an intriguing result, but the cardiology community was reminded that surprise "good findings are not uncommonly followed by surprise bad findings as the vesnarinone experience demonstrated.[†] Unanticipated surprise findings on non-primary endpoints weaken rather than strengthen the results of a clinical trial.

---

[*] This is a classic illustration of a random analysis.

[†]The contemporary reversal of the vesnarinone findings added more fuel to this raging fire. Vesnarinone was a positive inotropic agent that increased the pumping ability of the heart, holding out initial promise for improving the treatment of CHF. The first study, designed to randomize 150 patients to each of the three treatment arms, and follow them for 6 months, revealed a 62% reduction in all-cause mortality (95% CI 28 to 80; $p = 0.002$) which was not the primary endpoint of the trial (Feldman AM, Bristow MR, Parmley, WW et al. (1993). Effects of vesnarinone on morbidity and mortality in patients with heart failure. *New England Journal of Medicine* **329**:149–55.). However a second clinical trial reversed these findings to the confusion of the cardiology community. The vesnarinone investigators stated, "Examination of the patient populations in the two trials reveals no differences that could reasonably account for the opposite response to the daily administration of 60-mg of vesnarinone." (Cohn J, Goldstein SC, Feenheed S et al. (1998). A dose-dependent increase in mortality seen with vesnarinone among patients with severe heart failure. *New England Journal of Medicine* **339**:1810–16.)

2.9.2 Let the Data Decide!

Much of the discussion involving carvedilol revolved around the policy of disassembling the notion of a hierarchy of prospectively planned analysis, in favor of "letting the data decide." This latter point of view has a seductive, egalitarian sound. At first glance "letting the data decide" appears to liberate the interpretation of a study's results from the biases and preconceived notions of the investigator. It also frees the investigator from the responsibility of choosing arbitrary endpoint decisions during the planning stage of the experiment, selections that subsequently may be demonstrated by the data to be the "wrong choices." In fact, it can appear to be far better for the investigators to preserve some flexibility in their experiment's interpretation by saying little during the design of the experiment about either the endpoint selection or the analysis procedures. This would allow the data collected to choose the best analysis and endpoint as long as these selections are consistent with the goals of the experiment.

This "let the data decide" point of view may appear to be bolstered by the observation that researchers by and large understand and appreciate the importance of choosing the research sample with great care. Intelligent, well developed methodologies are required to choose the optimum sample size [20 — 24]. The sampling mechanism, i.e., the process by which patients are selected from the population requires careful attention to detail. Well-tested mechanisms by which patients are randomized to receive the intervention or the control therapy are put into place in order to avoid systematic biases that can produce destabilizing imbalances. In fact, the fundamental motivation for the execution of the simple random-sampling mechanism is to produce a sample that is representative of the population [25]. This effort can be an onerous, time consuming, and expensive process, but investigators have learned that it can pay off handsomely by producing a sample that "looks like" the population at large. Therefore, having invested the time and energy into producing a sample that is "as large and as random as possible," they would like the right to generalize any result "the data decides" from the sample.

Unfortunately, the utility of this approach as a confirmation instrument is undone by the wide range of sample-to-sample variability. The results from Table 1.3 demonstrate the wide variability in estimates of the mortality effect of an intervention. It is impossible to determine which sample provides the most accurate estimate of the effect in the population. In fact, from Table 1.3, we see that "letting the data decide" leads to a cacophony of disparate results from different samples. The data in fact doesn't decide anything, since the data vary so widely from sample to sample.

There are two additional problems with allowing the data to decide. The first is that the sample is commonly not drawn correctly to provide the best viewpoint to provide the answer. The appropriate sample size, as well as inclusion and exclusion criteria, can only be implemented if the research question is asked prospectively. Additionally, as discussed earlier, the statistical estimators in this approach do not perform optimally when *post hoc* analyses are carried out. This dangerous combination of suboptimal sampling frames and statistical estimators can misinform investigators. The delineated experiences of the MRFIT, Vesnarinone, ELITE, and PRAISE investigators requires us to distance ourselves from the allur-

ing but ultimately misleading and disappointing approach of letting the data decide. While we must rely on data, the reliance pays its greatest dividend when the data are derived from a detailed, prospective plan.

The FDA refused the sponsor's application for the approval of Carvedilol for life prolongation in patients with CHF. However, six months later the compound was approved for CHF symptom amelioration. A more thorough evaluation of this second meeting is available [12,13], as well as an elaboration of the additional entanglements faced by clinical trials assessing the role of beta blockage in CHF [26]. Alternative points of view are available [12,14,17].

2.10 Tight Protocols

Well written protocols are the first good product of a well-designed research effort. The development of a tight protocol, immune to unplanned sampling error contamination, is a praiseworthy effort but takes a substantial amount of time and requires great patience. The investigators who are designing the study must have the patience to design the study appropriately. Its literature review is not just thorough and informative, but constructive, in that it guides the final assembly of a relevant research question.[*] They must take the time to understand the population from which they will sample, make a focused determination of the necessary endpoint measurements, and assess endpoint measures accurately and precisely. Occasionally, they should also have the patience to carry out a small pilot study, postponing the main trial until they have tested recruitment and data collection strategies. In the well written protocol, the methods section is elaborated in great detail. The source and numbers of research subjects or patients are provided. Data collection methods are elaborated. Statistical analysis tools are expounded. The specification of each endpoint's ascertainment, verification, and analysis is laid out in great detail.[†]

Those who make these efforts are rewarded with well-considered protocols that are executable, and have a clearly articulated analysis plan that is data independent. This is a heavy burden.[‡] This complete elaboration serves several useful purposes. A requirement of its development demands that the research effort itself be thoughtfully considered prospectively, a requisite for good research execution. In addition, the specifications in the protocol permit the conditions of the research to be fully illuminated, permitting other researchers to replicate the research design. Finally, a well-written protocol serves as an indispensable anchor for the study, keeping the analyses from being cast adrift in the eddies of the incoming data stream.

After the protocol is written and accepted, investigators must insist on nothing less than its rigorous execution. The protocol is the rule book of the trial,

---

[*] Arthur Young wrote of the necessity for a thorough review of the literature as a tool to create productive and useful research efforts. (Chapter One, page 12).

[†] In some cases, the protocol may discuss an analysis even though neither the endpoints of the analysis nor the details for the analysis are known during the design phase of the trial. An example is blood banking.

[‡] Investigators may profit from remembering the Quaker admonition "strength in this life, happiness in the next."

and all involved must strictly adhere to it. Information about a violation of the trial protocol must immediately raise concerns for the presence of study discordance with attendant alpha error corruption. As for midstream endpoint changes, consider the following true story.

> A 34 year-old automobile mechanic from Alamo, Michigan, was troubled by a noise coming from the truck he and his friend were driving. The mechanic insisted on finding the noise's source. Instructing his friend to continue driving so the noise would continue, the mechanic maneuvered outside of the truck and then, while the truck was still being driven, continued to squirm and maneuver until he was underneath the truck. While carrying out his investigation, his clothing caught on part of the undercarriage. Some miles later when his friend finally stopped the car and got out, he found his mechanic friend "wrapped in the drive shaft" – quite dead.

Avoid the temptation of changing a research program that is already underway, so you don't get "wrapped in drive shaft."

2.11 Design Manuscripts

Occasionally the investigators will choose to publish the protocol of their study as a "design manuscript." There are several advantages to this procedure. First, the appearance of the protocol in the peer-reviewed medical literature broadcasts to the research and the medical community that a research effort is being conducted to answer the posed scientific question. In addition, important assumptions underlying the sample size computation, aspects of the inclusion and exclusion criteria, and endpoint determinations are carefully elaborated, commonly to a greater degree than is permissible in the manuscript that will describe the study's final results.

Finally, a design manuscript is a message to the research community from the investigators that says "Here is the research question we wish to address. This is how we have decided to address it. Here are the rules of our trial. Be sure to hold us to them."[*] Examples of design manuscripts are available in several healthcare related fields [27 —32].

In addition, design manuscripts can be particularly useful for research efforts that examined disputed, sometimes litigious research questions in which strong, vocal, and influential forces have forcefully articulated their points of view before the research effort was conceived. At its conclusion, a well-designed, well executed research effort will be criticized because its results are at variance with the expectations of some, a level of criticism that is directly proportional to the controversial nature of the research question. Such criticism is inevitable and unavoidable.

---

[*] Design manuscripts have the additional advantages of (1) engaging the clinical trial investigators in the publishing process, an activity that can help to improve morale in a long trial, and (2) conserving space in the final manuscript that is published when the clinical trial has been completed by describing the trial's methodology in complete detail in the earlier appearing design manuscript.

However, one particularly sharp barb can be that the investigators sacrificed their objectivity by changing the research methodology to produce the desired answer. Like the unfortunate archeologist who was found with a hammer and chisel altering the length of one of the pyramids because it did not conform to his expectation, the researchers are commonly accused of warping their research (and their data) to get the desired answer. The appearance of a well-written design manuscript followed by the concordant execution of the protocol, like the apple of gold in the setting of silver, naturally fit together to blunt this especially visceral criticism.

2.12 Concordant Versus Discordant Research

As we have seen, the accuracy of statistical estimators in a sample-based research effort depend on the sources of variability. Define study *concordance* [33] as the execution of a research effort in accordance with a prespecified protocol and analysis plan. With study concordance, the analysis plan is prospectively fixed, sampling error only affects the values of the observations, and the estimators perform well. On the other hand, if the research plan is not chosen prospectively but is selected by the data, the influence of sampling error on the analysis selection perturbs the estimators. This circumstance is defined as study *discordance*. Since the random data have been allowed to transmit randomness to the analysis plan, and the research program's results, effect sizes, standard errors, confidence intervals, and *p*-values are unreliable since it is the result not just of random data, but of a random analysis plan.

However the notion of concordance or discordance is not one of absolutes. The unexpected affects every research effort, and some discordance is present in every research program.

2.12.1 Severe Discordance: Mortality Corruption

Severe study discordance describes the situation where the research execution has been so different from that prescribed in the prospectively written protocol that the sample provides a hopelessly blurred and distorted view of the population from which it was chosen. Estimators from these discordant studies are rendered meaningless and irrelevant to the medical community since the view of the population through the sample is smeared and distorted. This unfortunate state of affairs can be produced by the flawed execution of a well-written protocol. The situation can be so complicated that, in the case of a positive study, the estimators themselves cannot be incontrovertibly computed. Experiments that lead to this type of estimator corruption are essentially useless to the research community.

Consider the case of a clinical trial that will assign therapy to patients with advanced HIV infections. The study has two treatment arms, and patients are to be followed for four years. The endpoint of the study is the total mortality rate.

At the conclusion of the study, 15% of patients are lost to follow-up. What is the best interpretation of these results? The implications of this discordance are substantial, because the follow-up losses blur the investigator's view of the population. In essence, they obtained the sample to learn what would happen in the popu-

lation, but now, with substantial follow-up losses, they never learned what really happened in their sample.

There are additional problems. How should the analysis be carried out? The problem here is not that statistical estimators are data-based (every statistical estimator uses data). The problem introduced by severe study discordance is that the very method of computing the estimators is data based. Should the computation assume that all patients who are lost to follow-up are dead? Should it assume that patients who are lost to follow-up in the active group are dead, and those who were lost to follow-up and in the placebo group are alive? Should it assume that an equal fraction of patients who are lost to follow-up are alive, regardless of the therapy group assignment? The best choice from among these possibilities is not dictated by the protocol. Instead, the choice of computation is based on belief and the data, which itself is full of sampling error.* Variability has wrenched control of the statistical estimator's computation from the protocol and complicates this study's interpretation.

The degree of discordance here depends on the magnitude of the statistical estimators. If the effect size and *p*-value remained below the threshold of significance in the most stringent of circumstances i.e., assuming that all lost patients assigned to the placebo group were alive at the trial's conclusion, but that all lost patients in the active group were dead, we must conclude the discordance is mild because the worst implications of the follow-up losses do not vitiate the results. However, if the *p*-value fluctuates wildly in this sensitivity analysis, the discordance is severe.

2.12.2 Severe Discordance: Medication Changes

Another example of severe study discordance is produced from a clinical trial designed to assess the effect of an intervention on patients who have established heart failure at the time of randomization. Patients are randomized to either control therapy or control therapy plus an active intervention. The prospectively specified primary analysis for the study is a change in the background medication for heart failure (e.g., increase in digoxin or diuretic use during the trial, or the addition of ACE-i during the trial), a change that would be triggered by deterioration of the patient's left ventricular function over time. The trial requires a sample size of 482 patients.

The trial protocol assumes that all patients will have an endpoint assessment. However, during the trial's execution, 40% of the patients have missing medical records precluding the determination of endpoint status, allowing endpoint computation on only the remaining 60%. Is this experiment interpretable? The investigators must carry out some post-hoc computation if the research effort is to convey useful information about the population to the research and regulatory community.

Clearly, the view is distorted if 40% of the sample has missing endpoint information. As before, there is disagreement on the computation of the trial's statistical estimators since different assumptions about the medication records for the

* Some of these computations will provide statistical estimators that support the investigators' ideas, while other computations provide antithetical measures of effect size, standard errors, confidence intervals, and *p*-values.

40% of patients with missing information would lead to different values of the effect size and confidence interval widths. For example should the *p*-values be computed assuming that the missing patients had no medication changes? Assuming only active patients had medication changes? Each assumption leads to a different effect size computation. Experimental discordance is potentially extreme, and if a sensitivity analysis reveals wide variation in effect size estimator based on the assumption about the 40% of patients with lost records, then the study will be uninformative about the effect of therapy on the CHF population.

2.12.3 Discordance and NSABP

As a final example of discordance, consider the findings of the National Surgical Adjuvant Breast Project (NSABP) [34] that examined the effect of different therapies for breast cancer reduction. After the study's results were analyzed and published, it was discovered that 99 ineligible patients were deliberately randomized with falsified data. Is this experiment hopelessly corrupted?

One way to resolve this dilemma was to examine whether the estimates of effect size, standard error, confidence intervals, or *p*-values were substantially altered in a sensitivity analysis by excluding the fraudulently entered 99 patients. This sensitivity analysis revealed no important change in the statistical estimators. However, a second relevant question of the interpretation of the trial had to be addressed since the presence of fraudulent data admitted the possibility of dishonest behavior elsewhere in the trial apparatus. To address this discordance issue, a full audit of the study data was carried out by Christian et. al. [35]. Since the protocol discrepancies identified were small in number and magnitude, the degree of discordance was assessed to be mild, and the conclusions of NSABP were allowed to stand.*

2.13 Conclusions

The supremacy of hypothesis-driven research does not eliminate surprises. Quite the contrary, we learn by surprises in science. The key features of our principle are (1) we cannot blindly take every finding in a sample and extend that finding to a much larger population and (2) the magnitude of the statistical estimators does not determine which finding should be extended to the population, and which should be left behind in the sample. The findings for which the experiment is designed are most likely to be the generalizable findings — those we take to the population with us. This applies to any sample-based research, from small observational studies to large, expensive state-of-the-art clinical trials.

During the design of a research effort, anticipate that there will be many more exploratory analyses than primary analyses. Investigators should be encouraged to triage their analysis plans [26], identifying a small number of primary analyses, followed by a greater number of secondary analysis, and even a larger number of exploratory analyses. This permits them to select a sample that focuses

---

* The controversy that swirled around this study involved several layers of investigation, including congressional inquiries. See Moyé L (2004) *Finding Your Way in Science: How You Can Combine Character, Compassion, and Productivity in Your Research Career.* Vancouver: Trafford.

on a small number of research questions for which they can provide confirmatory solutions. While the exploratory questions are interesting, the answers provided are only hypothesis-generating.

Similarly, reporting research results should follow a hierarchy as well. It is best to first report the findings that were the basis of prospective statements, on which some prior type I error has been allocated.* Once this has been completed, the researcher may announce other unexpected finding, but with the preamble that these are exploratory findings. These exploratory findings are disseminated to raise new questions — not to answer them.

Thus, research efforts are combinations of confirmatory and exploratory analyses. The sequence of events is typically that an interesting exploratory analysis is followed by a confirmatory one, the latter being used to confirm the former. This confirmatory analysis should not represent an attempt to reproduce slavishly the findings of exploratory analysis, as in the following humorous example:

> The chef at a hotel in Switzerland lost a finger in a meat cutting machine and, after a little hopping around, submitted a claim to his insurance company. The company, suspecting negligence, sent out one of its men to have a look for himself. He tried the machine out and lost a finger. The chef's claim was approved.

Exploratory analyses are commonly useful because they provide the first data-based view of the future. Thus, despite their limitations, exploratory analyses will continue to play an important role in research efforts. However, they should be reported at a level and tenor consistent with the inaccuracy of the estimators on which they are based.

On the other hand, the confirmatory work to reproduce the exploratory analysis should be designed to evoke and elaborate in detail the result of the exploratory analysis so that information about the mechanism that produced the exploratory relationship becomes clearer. The confirmatory analysis will, in all likelihood, require a different number of subjects, perhaps a more precise measure of the endpoint, and a different analysis than that presented in the exploratory study.

Ultimately, the value of the research depends on the generalizability of the research sample's results to larger populations. Well-designed research, well-chosen samples, and concordant execution are each required for this process to succeed.

References

1 . Lehmann EL (1983) *Theory of Point Estimation.* New York; John Wiley. p 3.
2. Miles JL (1993) Data torturing. *New England Journal of Medicine* **329**: 1196-1199.
3. MRFIT Investigators (1982) Multiple risk factor intervention trial. *Journal of the American Medical Association* **248**:1465-77.

* Some uncomplicated advice on this allocation is provided in Chapter Eight.

4 Pitt B, Segal R, Martinez FA et al. on behalf of the ELITE Study Investigators (1997) Randomized trial of losartan versus captopril in patients over 65 with heart failure. *Lancet* **349**:747–52.
5 Jensen BV, Nielsen, SL (1997) Correspondence: Losartan versus captopril in elderly patients with heart failure. *Lancet* **349**:1473.
6 Fournier A, Achard JM, Fernandez LA (1997) Correspondence: Losartan versus captopril in elderly patients with heart failure. *Lancet* **349:**1473**.**
7. Pitt B, Poole-Wilson PA., Segal R, et. al (2000) Effect of losartan compared with captopril on mortality in patients with symptomatic heart failure randomized trial–The losartan heart failure survival study. ELITE II. *Lancet.***355**:1582–87.
8. Meinert CL (1986) *Clinical Trials: Design, Conduct, and Analysis*. New York. Oxford University Press.
9. Friedman L, Furberg C, DeMets D (1996) *Fundamentals of Clinical Trials*. Third Edition. New York. Spinger.
10. Moyé LA (1998) *P*–value interpretation and alpha allocation in clinical trials. *Annals of Epidemiology.* **8**:351–357.
11. Packer M., Bristow MR Cohn JN et al (1996) The effect of carvedilol on morbidity and mortality in patients with chronic heart failure. *New England Journal of Medicine.* **334**:1349–55.
12 Fisher L (1999) Carvedilol and the FDA approval process: the FDA paradigm and reflections upon hypothses testing. *Controlled Clinical Trials* **20**:16–39.
13. Moyé LA, Abernethy D (1996) Carvedilol in Patients with Chronic Heart Failure (Letter) *New England Journal of Medicine.* **335**: 1318–1319.
14. Packer M, Cohn JN, Ccolucci WS (1996) Response to Moyé and Abernethy. *New England Journal of Medicine* 335:1318–1319.
15. Fisher LD, Moyé LA (1999) Carvedilol and the Food and Drug Administration Approval Process: An Introduction. *Controlled Clinical Trials*. **20**:1–15.
16. Moyé L.A (1999) *P* Value Interpretation in Clinical Trials. The Case for Discipline. *Controlled Clinical Trials* **20**:40–49.
17. Fisher LD (1999) Carvedilol and the Food and Drug Administration-Approval Process: A Brief Response to Professor Moyé's article. *Controlled Clinical Trials*. **20**:50–51.
18. Lewis JA (1995) Statistical issues in the regulation of medicines. *Statistics in Medicine*. **14:** 127–136.
19. Pocock SJ, Geller NL, Tsiatis AA (1987) The analysis of multiple endpoints in clinical trials. *Biometrics* **43**:487–498.
20. Lachim JM (1981) Introduction to sample size determinations and power analyses for clinical trials. *Controlled Clinical Trials* **2**:93–114.
21. Sahai H, Khurshid A (1996) Formulae and tables for determination of sample size and power in clinical trials for testing differences in proportions for the two sample design. *Statistics in Medicine* **15**:1–21.
22. Donner A (1984) Approach to sample size estimation in the design of clinical trials — a review. *Statistics in Medicine* **3**:199–214.

23. George SL, Desue, MM (1974) Planning the size and duration of a clinical trial studying the time to some critical event. *Journal of Chronic Disease* **27:**15–24.
24. Davy SJ and Graham OT (1991) Sample size estimation for comparing two or more treatment groups in clinical trials. *Statistics in Medicine* **10**:3–43.
25. Snedecor GW, Cochran WG (1980) *Statistical Methods, 7th Edition*. Iowa; Iowa State University Press.
26. Moyé LA (2003) *Multiple Analyses and Clinical Trials*. New York; Springer.
27. The SHEP Cooperative Research Group (1988) Rationale and design of a randomized clinical trial on prevention of stroke in isolated systolic hypertension. *Journal of Clinical Epidemiology* **41**:1197–1208.
28. Davis BR, Cutler JA, Gordon DJ, Furberg CD, Wright JT, Cushman WC, Grimm RH, LaRosa J, Whelton PK, Perry HM, Alderman MH, Ford CE, Oparil S, Francis C, Proschan M, Pressel S, Black HR, Hawkins CM for the ALLHAT Research Group (1996) Rationale and design for the antihypertensive and lipid lowering treatment to prevent heart attack trial (ALLHAT) *American Journal of Hypertension* **9**:342–360.
29. Moyé, LA for the SAVE Cooperative Group (1991) Rationale and design of a trial to assess patient survival and ventricular enlargement after myocardial infarction. *American Journal of Cardiology* **68**:70D–79D.
30. Pratt, CM, Mahmarian JJ, Morales-Ballejo H,Casareto R, Moyé, LA for the Transdermal Nitroglycerin Investigators Group (1998) The long–term effects of intermittent transdermal nitroglycerin on left ventircular remodeling after acute myocardial infaction: Design of a randomized, placebo controlled mulitcenter trial. *American Journal of Cardiology* **81**:719–724.
31. Moyé LA, Richardson MA, Post-White J, Justice, B (1995) Research Methodology in Psychoneuroimmunology: Rationale and design of the IMAGES–P (imagery and group emotional support study-pilot) clinical trial. *Alternative Therapy in Medicine* **1**:34–39.
32. Pfeffer MA, Sacks FM, Moyé LA et al. for the Cholesterol and Recurrent Events Clinical Trial Investigators (1995) Cholesterol and Recurrent Events (CARE) trial: A secondary prevention trial for normolipidemic patients. *American Journal of Cardiology* **76**: 98C–106C.
33. Moyé LA (1998) P value interpretation and alpha allocation in clinical trials. *Annals of Epidemiology*. **8**:351–357.
34. Fisher B, Bauer M, Margolese R. et al (1985) Eight year results of a randomized clinical trial comparing total mastectomy and segmental mastectomy with or without radiation in the treatment of breast cancer. *New England Journal of Medicine*.**312**:665–73.
35. Christian MC, McCabe MS, Korn EL, Abrams JS, Kaplan RS, Friedman MA (1995) The National Cancer Institute audit of the national surgical adjuvant breast and bowel project protocol B-06. *New England Journal of Medicine*. **333**:1469–1474.