TOO MANY PROBABILITIES: STATISTICAL EVIDENCE OF TORT CAUSATION

DAVID W. BARNES*

I

INTRODUCTION

Judges and lawyers first encountering statistical evidence want to believe that scientific standards are tougher than legal standards. A court will reject an assumption that there is no causal connection between an act and an injury if the evidence makes causation "more likely than not." A scientist will reject an assumption that there is no relationship between two variables only if there is less than a five percent probability that the statistical evidence showing a relationship is due to chance. The law appears willing to accept no more than a forty-nine percent chance of error while science appears willing to accept no more than a five percent chance of error. This perception is incorrect, but hard to change. It is a matter of such serious concern to statisticians and scientists that they often raise the issue, but lay people seldom understand it. This article offers those uninitiated into the statistical guild several reasons to look behind the probabilities when evaluating scientific evidence.

The lack of congruity between legal and scientific standards is most egregious when testimony based on statistical/scientific methods is used to prove causation in tort law. Medical scientific testimony is often expressed in terms of two different probabilities. The first of these estimates the increased probability of harm if a person is exposed, for example, to a toxin. The second is the probability that the observed relationship is an artifact of the experimental method, rather than an actual causal relationship between the toxin and the injury. These two probabilities measure different phenomena, and neither measures whether causation is more likely than not. This article demonstrates that neither probability, taken alone or together, measures whether the "preponderance of the evidence" test is met.

In many cases involving statistical/scientific evidence, probabilistic observations may be conveniently captured under three headings: the "belief probabil-

This article is also available at http://www.law.duke.edu/journals/64LCPBarnes.

* Distinguished Research Professor of Law, Seton Hall University.

The author would like to express his gratitude to David Peterson, whose assistance was, as always, invaluable.

ity," the "fact probability,"[1] and the "sampling error probability." The belief probability relates to evidentiary requirements imposed by the law, and the fact probability relates to the facts relevant to legal cases. These two probabilities are directly related to the civil law evidentiary requirement that the proponent of a claim must prove that the other's act is more likely than not a cause of harm. By contrast, the sampling error probability is a characteristic of statistical science. Appreciating the distinctions among these probabilities facilitates an understanding of the relationship between the preponderance of the evidence standard and the probabilities reported by statisticians.

II

THE THREE PROBABILITIES

The belief probability refers to the credibility—the believability—of the evidence in support of a party's factual claims. In tort causation, the belief probability describes the factfinder's confidence in a party's evidence about cause. In civil cases, the law requires that the proponent of a fact convince the factfinder by a preponderance of the evidence. The factfinder is instructed that the plaintiff's claim of causation must be more likely than not true. Thus, the belief probability for the proponent's factual assertion must exceed fifty percent in civil cases for the proponent's assertion to be accepted by the factfinder as true.[2]

The fact probability describes a separate feature of a party's evidence related to cause. The fact probability is the likelihood that the defendant's actions led to the adverse outcome. A fact probability may be based on, or stated in terms of, percentages. For example, when a physician delays his diagnosis of a patient's disease, the fact probability measures the percentage point reduction, due to the defendant physician's delay, in the plaintiff's chances of survival or recovery (or, in general, of obtaining a better outcome than she obtained). If the physician's delayed diagnosis reduced the patient's chance of recovery from thirty percent to ten percent, the relevant probabilities are those percentages. How confident we are that the physician caused a twenty percentage point reduction in the probability of survival depends on the strength of the evidence, which is the belief probability. We may, for instance, be only forty percent sure (belief probability) that the reduction is twenty percentage points (fact probability). If that were the case, the factfinder would be constrained to conclude that the preponderance of the evidence test has not been satisfied.

---

1. *See* Steve Gold, *Causation in Toxic Torts: Burdens of Proof, Standards of Persuasion, and Statistical Evidence*, 96 YALE L.J. 376, 382-84 (1986) (describing the difference between the fact probability and the belief probability); *see also* Philip Cole, *Causality in Epidemiology, Health Policy, and Law*, 27 ENVTL. L. REP. 10,279, 10,280, 10,284 (1997) (discussing the credibility of scientific evidence in relation to civil and criminal burdens of proof).

2. Philip Cole describes the "preponderance of the evidence" standard as occupying a position on the "spectrum of credibility." Thus, what the law describes as a party's burden of persuasion could be equated to a requirement that the belief in the party's assertion occupy at least some minimum position on that spectrum. *See* Cole, *supra* note 1, at 10,280.

Risk ratios are frequently relied upon to establish fact probabilities. Risk ratios measure the percentage change in the incidence of a specified harm, such as a disease. A risk ratio compares a background rate, where the stimulus in question is not present, to the rate that obtains when the stimulus is present. For example, in a routine tort case alleging that a negligent failure to light a stairway caused a fall, a risk ratio might compare the incidence of falling down stairs when the stairs are well-lit to the incidence of falling when the stairs are unlit. In a case based on medical/scientific evidence of causation, a risk ratio might compare the incidence of birth defects when the drug in question is not taken by the mother to the incidence when the drug is taken. A risk ratio greater than one indicates that risks are increased. For instance, risk ratios of 1.5 and 3 indicate that the stimulus (for example, lack of lighting) increases the risk of falling by 50% and 200%, respectively.

A correlation or regression coefficient may also be the basis for a fact probability. A correlation coefficient measures the mathematical correspondence or degree of mathematical association between the stimulus and the harm. Squaring this coefficient allows us to estimate, in percentage terms, the amount of variation in one variable mathematically accounted for by variation in another. A regression coefficient also measures the association between those variables but adds to the information provided by the correlation coefficient. It estimates how much the harm varies when the amount of stimulus is changed. A regression coefficient may indicate the percentage increase in the incidence of injuries associated with a specified dosage of stimulus.

The numerical value of any of these fact probabilities is supposed to depend only on the actual (biological, physical, chemical, or "natural") connection between the events in the population being studied. The fact probability from a poorly conducted study will reflect that population poorly and will lack credibility. The confidence that the estimate given by a study or a statistical analysis is correct, known as the belief probability, is a characteristic of the scientific method as applied to the sample of the population.

The sampling error probability refers to a statistical property of data underlying evidence offered to prove a relevant fact, such as the connection between the defendant's act and the plaintiff's harm. Statistics used to prove causation are typically derived from a study of the relationship between acts like the defendant's and harms like the plaintiff's. Such a study is typically based on a sample, because of the literal and practical impossibility of measuring the effect of that act on all living human beings. Studying only a sample inevitably gives rise to the possibility that the sample chosen is atypical of a larger group, the population represented by the sample.

Even when a sample is composed of randomly chosen subjects, those subjects may not represent accurately the population. That possibility means that any observed statistical relationship between acts like the defendant's and harms like the plaintiff's revealed by a study of a sample may be due to the happenstance of having drawn randomly an atypical sample. Thus, the sam-

pling error probability measures the likelihood that an observed statistical relationship was due to the random selection of subjects to include in the study. If statistical evidence based on a sample is used to establish a fact probability, the statistical evidence concerning the fact probability always has an associated sampling error probability. An example of how we evaluate the sampling error in ordinary discourse may facilitate an understanding of the concept.

III

"MALE ANSWER SYNDROME" EVIDENCE: SAMPLING AND OTHER SOURCES OF ERROR

An analogy to an imaginary psychological disorder, "Male Answer Syndrome," may promote an intuitive appreciation of the differences among these three probabilities. "Male Answer Syndrome" ("M.A.S.") describes a subject's exaggerated willingness to answer (or inability to resist the impulse to answer) factual questions regardless of the subject's lack of knowledge of the relevant phenomena or circumstances. Thought to be behaviorally related to the exaggerated unwillingness to ask for directions when lost, M.A.S. is most commonly observed in the behavior of males of the human species, lawyers, and pedagogues in all disciplines. A single response by an individual suffering from M.A.S. illustrates the differences among fact, belief, and sampling error probabilities.

The sampling error probability helps us to evaluate the M.A.S. victim's response to a question. Consider the following question demanded (perhaps even rhetorically) of one afflicted with M.A.S.: Why is the train so late? The respondent may have some data that he thinks are relevant to the question. The sampling error probability assumes without question that the data are relevant. To calculate this probability, a statistician does not ask where the data originated, but rather how much data there are and how internally consistent they are. The sampling error probability depends on the sample size ("how much data") and the variation within the data on the relevant quantified variable of interest (the "internal consistency").

A prudent person who questions the M.A.S. victim will want to consider the range of experiences the respondent brings to bear on his analysis. Returning to the train question, if the speaker's experiences with late trains have been few and at odds with one another (such as one late train due to a fire the first time and due to rush hour congestion the only other time), the prudent questioner will conclude that the factual response should be evaluated skeptically, because the sample is small and the observations are quite different from one another. Thus, whatever fact is asserted, the sampling error makes a prudent factfinder question the assertion. The sampling error affects the credibility of the assertion, which is the belief probability.

Someone listening to the M.A.S. victim will, however, want to know more than only the sample size and the variation before deciding whether to believe the speaker. It is a characteristic of the syndrome that, despite the lack of sub-

stantial evidence, the responding M.A.S. sufferer has an answer: "Rush hour crowds are causing the delay"; or, if he is a pedant, "There is an eighty-five percent chance that rush hour crowds are the cause." The fact probability in this example is eighty-five percent. Is it more likely than not that rush hour crowds are the cause of the delay? Yes, if we believe the respondent. Do we? The sample error probability tells us something—but not everything—we need to know about whether we should believe the respondent. The fact probability alone tells us nothing at all about the likelihood that the answer is correct.

The preponderance of the evidence test for causation in torts has substantive and procedural components. The substantive component comes from the law of causation, which requires that the stimulus be a necessary event in the chain of causation that resulted in harm. A stimulus is a "but for" cause in torts if the harm would not have occurred absent the stimulus. In cases involving harmful drugs and other toxic agents, the type of harm suffered by the complainant may occur even if the stimulus is not present. The substantive issue in the complainant's case is whether the stimulus was present and was necessary to produce the harm.

By contrast, the procedural component of the preponderance of the evidence test describes how convincing the proponent's evidence of "necessity" must be. It only requires that the proponent's claim that the stimulus was a necessary antecedent event is more likely than not true. If the evidence is not persuasive enough to make the fact more likely than not true, the procedural effect is that the proponent has failed to prove his claim. The procedural component describes the weight or credibility of the evidence—its believability—not the substance of what the evidence demonstrates.[3] Both components must

---

3. Thus, in loss-of-a-chance cases, where there is no dispute regarding the cause of death, a defendant may not claim that there was a 70% chance that the patient would have died anyway and argue that recovery should be denied because "more likely than not" she would have died anyway. The 70% figure is a fact probability. The "more likely than not" test refers to the belief probability, which is the credibility of the 70% figure. Thus, the belief probability is not necessarily the 70% claimed by the defendant.

The basic confusion of fact probabilities and belief probabilities in loss-of-a-chance cases is illustrated by *Fennell v. Southern Maryland Hospital Center*. In *Fennell*, Maryland's highest court stated:

> [T]raditional tort law is based on probabilities. If a patient had a 49% chance of dying from an injury or disease and if the patient was negligently treated and dies, full recovery will be permitted because, absent the negligence, it was more likely than not that the patient would have survived. Based on the 51% probability of surviving the injury or disease, we exclude the injury or disease as the cause of death. Damages are not reduced by the fact that there was a strong possibility that the patient would have died absent the negligence. Conversely, if the patient had a 51% chance of dying from an injury or disease, and was negligently treated and died, it was probably the pre-existing medical condition, not the negligence, that killed the patient, and there is no recovery. Damages must be proven by a preponderance of the evidence. Damages are not proven when it is more likely than not that death was caused by the antecedent disease or injury rather than the negligence of the physician.

580 A.2d 206, 214 (Md. 1990). The *Fennell* court also stated:

> We are unwilling to relax traditional rules of causation and create a new tort allowing full recovery for causing death by causing a loss of less than 50% chance of survival. In order to demonstrate proximate cause, the burden is on the plaintiff to prove by a preponderance of the evidence that "it is more probable than not that defendant's act caused his injury."

Id. at 211 (citation omitted).

be evaluated in any analysis of scientific evidence of causation. Thus, fact probabilities alone—whether risk ratios, correlations, regression results, or other statistical estimates—may not be considered without an accompanying measure of the belief probability. The sampling error probability does not measure the belief probability because there are so many sources of potential error other than the unrepresentativeness of the sample.

For our M.A.S. victim, the belief probability is, by definition, low. That does not mean he is wrong. It only means that we are highly uncertain that he is right. He responds that the train will be late due to rush hour crowds despite the small number of times he has ridden the trains and the conflicting experiences he has had on the trains, problems the sampling error probability is designed to address in scientific studies.

He also responds despite his inability to do more than guess at the causes in those cases where the train was late, his inability to know if there were other factors (such as weather or labor shortages affecting the schedule), his lack of aptitude for reasoned analysis, and his inability to extrapolate from his experience on subways to a conclusion about commuter rails generally or about this particular train. Each of these sources of potential error should lessen our confidence in the eighty-five percent figure he asserts. When evaluating statistical evidence, as when evaluating the responses of a M.A.S. victim, we have no scientific way to adjust this eighty-five percent fact probability to reflect these additional uncertainties. We cannot know whether the fact probability is believable enough to meet the preponderance of the evidence standard. All we know is that our belief probability is below 100%.

Whether the uncertainties inherent in the statement of a fact reduce our belief probability below that required by the preponderance of the evidence standard cannot be objectively measured. We can say, however, that the belief probability is not measured by either the fact probability or the sampling error probability. The fact probability is a statement about the event, not about the credibility of the person or study reporting the event. The sampling error probability is a statement about the sample size and variation within the data. The number and variety of observations are two sources of uncertainty about the conclusion that affect the belief probability, but the sampling error probability in no way accounts for non-random selection of data, nonquantifiability of ob-

---

There are two probabilities initially involved in this case. The belief probability is reflected in the preponderance of the evidence rule requiring that the greater weight of the evidence regarding causation support the plaintiff's claim. The plaintiff's claim is based on the fact probability that the improper intubation, rather than the disease, caused the death. The preponderance of the evidence rule requires that the plaintiff's version of the facts are "more likely than not" true, whatever those facts are. The *Fennell* court assumed (that is, treated the belief probability as 100%) that the defendant's act caused the death; the fact is treated as known. The court denied recovery, however, because the plaintiff's likelihood of surviving meningitis was reduced from 40% to 0%. Whatever policy reasons there may be for denying recovery for that loss, it cannot be based on either too low a belief probability (it was assumed to be 100%) or the fact probability (without the physician's act, the patient would not have died when she did). The denial cannot be based on the preponderance of the evidence rule as it relates to causation. If some other policy—related perhaps to the fact that the patient did not have long to live anyway—should limit the amount of damages the plaintiff should recover, then that is another issue.

servations, improper exclusion of relevant variables, or the difficulties in generalizing from samples or particularizing from generalities.

Notwithstanding the fervent wishes of courts, advocates, and expert witnesses, none of these three probabilities is equivalent to the probability that the defendant's act caused the plaintiff's harm. Nor do other measures commonly employed, such as risk ratios, produce numbers equivalent to the probability of causation. The following section describes how statistical evidence compares to the information we would ideally possess when determining causation.

IV

WHAT STATISTICS CONCEAL

Statisticians study the mathematical relationships among variables, while tort lawyers are interested in the causal relationships among variables. Both might be concerned, for example, with the relationship between the way some people act and the harm other people suffer. A particular kind of conduct may be associated (mathematically or by physical laws) with a particular kind of harm. For example, ingestion by mothers of the miscarriage preventative diethylstilbestrol may be associated with cancerous vaginal and cervical growths in their daughters. It is well recognized, however, that the fact of mathematical association between variables does not mean that one is physically related to or caused by the other.

The statistical conclusion about an association between the two variables refers to the extent to which a mathematical relationship between stimulus and harm is observed in the data. A statistical coefficient provides an estimate that summarizes the nature of the relationship, including characteristics such as the consistency or strength of the relationship. The tort law conclusion refers to whether the harm would have arisen had the act not occurred. This conclusion is concerned with the "physics" of the relationship—with what came first, the chronological connection between an act and a result—rather than just whether there is a connection or the strength of the connection. Thus, fundamentally different tasks confront the statistician and the legal factfinder.

What do statistics reveal about causation? Tort lawyers would like a summary statistic—a coefficient—that reveals whether, and the extent to which, a particular act caused a particular plaintiff's harm. Other factors might be relevant as well, but the tort lawyer wants to know whether this act or stimulus is one of the necessary antecedent events. This is the "but for" cause question in torts. A coefficient summarizes a relationship between two variables by mathematically summarizing the individual observations in the sample. It purports to summarize how the positions of people like the plaintiff are affected by acts like the defendant's, but what turns that bare number into sufficiently convincing proof of causation?

A. Statistical Sampling Error

Even with a perfectly designed study of the sample of a population, it is possible that the size of the coefficient is due to the happenstance that the individual subjects of the study, even if randomly chosen, are not representative of the population. The control group in a sample may have contained, by chance, a larger proportion of individuals who are naturally immune to the disease than the population from which they were randomly drawn. Alternatively, the sample may be too small to reflect the diversity within the population from which it was drawn.

Note that the sampling error is not an error in the design of the sample. Indeed, it is not an error attributable to any person. It is an unavoidable property of inferential statistics, the process of estimating attributes of a population by examining a sample. Statisticians use the "p-value" to measure the sampling error probability, which is the probability that the observed relationship is due to the unrepresentative nature of the randomly selected subjects studied, rather than characteristic of the population from which they were drawn.

Varying between 0 and 1.00, the p-value measures the likelihood that it is the happenstance of selection of the particular individuals included in the study, rather than any relationship in the underlying population, that accounts for the observed relationship. The calculation depends completely on the size of the sample and the variation within the sample. For a given sample size, more variation usually results in a higher p-value, closer to 1.00. For a given amount of variation, a smaller sample size usually results in a higher p-value. This makes sense; it is hard to generalize from just a few observations if everything observed is different. The higher the p-value, the harder it is to conclude that a stimulus caused a result. A p-value closer to zero indicates a smaller probability that the error results from the sampling, leaving open the possibility that it was one of the many other sources of error identified below that caused any error in the estimate.

Statisticians tend to conclude that, if a scientific study is otherwise reliable, a five percent (a p-value of .05) sampling error probability is sufficiently small to reject the assumption that there is no relationship between the variables of interest. There is no magic to the five percent value.[4] All we can say for sure is that if everything else remains constant, as the p-value gets smaller, evidence of

4. The statistical significance of the study's findings only measures the probability that the random selection of subjects explains the observed relationship. There is no reason why a court should find a statistically insignificant coefficient at least relevant to the question of cause, except that scientific experts (for whatever historical reasons) do not find it sufficient for their purposes. It does say something about causation, even if it is not as persuasive as other evidence. A p-value of .15, for instance, tells a scientist that there is a 15% probability that it is the selection of the sample's members rather than "some other reason" for the observed relationship. This may be valuable information. Two reasons for ignoring studies with p-values greater than .05 are that judges do not want juries confused by the subtleties of p-values, or that they are willing to defer to scientists in determining evidentiary standards.

a relationship gets stronger.[5] To that extent, the p-value affects the belief probability. A higher p-value (closer to 1.0) reflects an increased probability that the coefficient's magnitude is due to the selection of subjects rather than any underlying relationship. Thus, the strength of our belief in the factual probability reflected in the coefficient should decline.[6]

Part V explains that the sampling error probability is not conceptually the same as the belief probability. Before explaining that fundamental difference,

---

5. Professor David H. Kaye recommends requiring that statisticians provide confidence intervals, rather than just testifying as to whether the coefficient is statistically significant as a way of emphasizing that it is the court's job, and not the statistician's, to decide what level of proof is required and what the statistical evidence proves or disproves. *See* David H. Kaye, *Is Proof of Statistical Significance Relevant?*, 61 WASH. L. REV. 1333, 1363-64 (1986). According to Professor Kaye, the source of the .05 significance level convention is Sir R.A. Fisher, who wrote:

> [I]t is convenient to draw the line at about the level at which we can say, "Either there is something in the treatment, or a coincidence has occurred such as does not occur more than once in twenty trials." . . . If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 per cent point), or one in a hundred (the 1 per cent point). Personally, the writer prefers to set a low standard of significance at the 5 per cent point, and ignore entirely all results which fail to reach that level. A scientific fact should be regarded as experimentally established only if a properly designed experiment *rarely fails* to give this level of significance.

Id. at 1344 (citing R.A. Fisher, *The Arrangement of Field Experiments*, 33 J. MINISTRY AGRIC. GR. BRIT. 504 (1926)), *quoted in* Leonard J. Savage, *On Rereading R.A. Fisher*, 4 ANNALS STAT. 441, 471 (1976).

6. The legal admissibility of evidence that would not be accepted by scientists raises difficulties for courts. The dissent in *Hodges v. Secretary of Department of Health and Human Services*, 9 F.3d 958, 965 (Fed. Cir. 1993) (Newman, J., dissenting), observed that

> although the data may not establish a causal relationship to a medical certainty, they may nonetheless meet the more-likely-than-not standard of the law. It was error for the special master to refuse to evaluate Kara Hodges' death on its particular facts and the available evidence, applying the requisite standard of proof.

This recognizes that statistical evidence is revealing and may shed light on a legal or factual issue even though it does not meet the standards of science. Nevertheless, courts are hesitant to admit such evidence coming from experts, because it appears that the experts are speaking outside their area of competence. Thus, the court in *Allen v. Pennsylvania Engineering Corp*., 102 F.3d 194, 197 (5th Cir. 1996), stated:

> While appellants' experts acknowledge the lack of statistically significant epidemiological evidence, they rely on certain studies as "suggestive" of a link between EtO exposure and brain cancer. "Suggestiveness" is not by the experts' own admission statistical significance, nor did the appellants' experts show why and how mere "suggestiveness" scientifically supports a causal connection; this basis for their scientific opinion must be rejected.

In an associated footnote, the *Allen* court quotes *Braun v. Lorillard Inc.,* 84 F.3d 230, 235 n.4 (7th Cir. 1996), *cert. denied*, 519 U.S. 992 (1996), saying:

> Courts should particularly pay close attention when expert witnesses depart from generally accepted scientific methodologies. As the Seventh Circuit noted in *Braun v. Lorillard Inc*., "[a] judge or jury is not equipped to evaluate scientific innovations. If, therefore, an expert proposes to depart from the generally accepted methodology of his field and embark upon a sea of scientific uncertainty, the court may appropriately insist that he ground his departure in demonstrable and scrupulous adherence to the scientist's creed of meticulous and objective inquiry."

Allen, 102 F.3d at 197.

This suggests that the court's rejection of evidence that is not statistically significant is due to the rules of evidence restricting experts to testimony based on the usual standards of their own professions or the court's own lack of expertise, rather than the perceived uselessness of studies with p-values greater than .05.

this article describes other sources of error that may make the sampling error probability and the fact probability themselves unreliable. Every study has other potential sources of error discussed below but will yield a p-value that ignores them. More important, the calculation of the p-value relies on assumptions that are inappropriate when estimating the belief probability. The mathematics of calculating sampling error probabilities proceeds unaware of any other sources of error and as if all underlying assumptions were true. The sources of error that arise when those assumptions are not true are described below. Each source of error affects the belief probability.

B. Experimental Design and Measurement Error

Before we can conclude anything about the causal connection between the particular defendant's act and the particular plaintiff's harm, we must know whence the coefficient came. What was measured? How was it measured? The well-known gold standard for experimental design is the randomized, controlled, double-masked study. As Peterson and Conley's paper discussing the Polio Trials demonstrates,[7] a study that fails to exhibit any of these three characteristics is not necessarily doomed but contains the potential for creating useless results. The significance of each characteristic depends on the type of study, so an independent evaluation must be made for each type of research. Courts are aware that samples must be well designed. The Court of Appeals for the Fifth Circuit has held that relying on non-random samples to answer troubling issues of causation when handling aggregated claims in mass torts cases violates the United States Constitution's requirement of due process.[8] The court emphasized the need for a proper inferential statistical technique as a means of ensuring "a sufficient level of confidence" in the results.[9] A sampling error probability may as easily be calculated from a poorly designed study as from a randomized, controlled, double-masked study. The credibility of that probability and any fact probability derived from that study, however, depends on the quality of the study design.

The poorer the experimental design, the more irrelevant and unreliable are both fact and sampling error probabilities. The effect of design error cannot generally be quantified, though qualitative judgments about the "closeness" of the study to the ideal may be made and evaluated when deciding how much credibility to give to the coefficient. The lack of quantifiability, however, means the effect of poor design on belief probability, fact probability, and sampling error probability cannot be measured. All we can know is that the probability estimates are not accurate. For example, we know that a sampling error probability of .05 from a study of subjects not randomly chosen does not even accurately measure the probability that an observed association is not due to

---

7. *See* David W. Peterson & John M. Conley, *Of Cherries, Fudge, and Onions: Science and Its Courtroom Perversion*, 64 LAW & CONTEMP. PROBS. 213, 217 (Autumn 2001).

8. *See In re* Chevron U.S.A., Inc., 109 F.3d. 1016, 1019-20 (5th Cir. 1997).

9. *See id*. at 1020.

chance. We know that the fact probability is unreliable. The credibility of those estimates depends on the quality of the study.

Similarly, the credibility of the estimate, the belief probability, will suffer if any of the following are true: the measurements were imprecise because of sloppy work or careless research assistants; inaccurate tools were used (using a yardstick to measure the diameter of a microscopic cell); the underlying question is not readily testable (tests of toxins cannot be performed on people, so they are performed on rats, which may not respond like people at all);[10] the variable of interest cannot easily be quantified (how much pain did the injured party suffer, or how bored is a juror listening to an expert witness); or unobserved phenomena account for the relationship (there may be an apparent relationship between the number of storks in the sky in September and the number of births that month, but both may have been caused by an extremely cold January). It may be that we can recognize that a study is biased, in the sense that it systematically errs in one direction, but the study itself does not provide a measure of how far off the estimate is.[11] When calculating the sampling error probability, none of these uncertainties that affect credibility are taken into account. The sampling error probability, then, does not measure the belief probability.

C. Statistical Modeling Error

Even if the experiment is well designed, the measurements are precise, and the p-value is low, both the coefficient and the probability of sampling error may be meaningless. The values of numbers reflected in the summary statistic and the p-value are only as accurate as the match between the statistical model and the underlying phenomenon. Statistical modeling error affects our confidence in the magnitude of the coefficient and the sample error probability, thereby affecting both the fact probability (we cannot trust that the coefficient accurately describes the relationship) and the belief probability (we cannot trust the p-value to measure the likelihood that the selection of aberrant individuals for the study, rather than some other explanation, accounts for the apparent relationship).

Two sources of statistical modeling error involve decisions regarding what variables interact with the two variables of interest and how all of the variables relate to one another. A statistical model makes assumptions about what influences an outcome and how different variables affect an outcome. Since our ultimate question is exactly that—that is, does the defendant's act (one of the variables) affect the plaintiff's condition—those assumptions are critical.[12] Pe-

10. *See* Allen v. Pennsylvania Eng'g Corp., 102 F.3d 194, 197 (5th Cir. 1996) (concluding that it would be improper to generalize about an effect on humans from tests on rats showing that EtO produced brain cancer because, among other reasons, EtO did not produce brain cancer in mice).

11. *See* Cole, *supra* note 1, at 10,281 (discussing bias in experimental design).

12. In *Paxton v. Union National Bank*, 519 F. Supp. 136, 163 (E.D. Ark. 1981), *aff'd in part, rev'd in part*, 688 F.2d 552 (8th Cir. 1982), plaintiffs offered a coefficient showing that, on average, black bank employees made $213.35 less than whites. The associated p-value was .01. Taking other factors such as

terson and Conley's paper also describes how an analyst's model of behavior can influence the outcome.[13] As the Court of Appeals for the Seventh Circuit has explained,

> [a] statistical study is not inadmissible merely because it is unable to exclude all possible causal factors other than the one of interest. But a statistical study that fails to correct for salient explanatory variables, or even to make the most elementary comparisons, has no value as causal explanation and is therefore inadmissible in a federal court.[14]

Thus, a study with a low sampling error probability may be inadmissible because its credibility is so low. There are statistical tools to aid modeling, but the lesson is that a coefficient ought not to be taken at face value even if the experiment is well designed.

To the extent that the statistical model incorrectly describes the underlying relationships, both the coefficient (the basis for the fact probability) and the sampling error probability are less accurate. We should correspondingly have less faith in the magnitude of the coefficient or the probability that the estimate is a result of the selection of individuals for the sample. Without more information, we cannot say whether the estimates are too high or too low. The added degree of uncertainty cannot be measured, but our belief probability certainly should decline. Again, the mathematical calculations underlying the fact and sampling error probabilities are blind to the quality of the statistical modeling, and the p-value does not reflect errors in modeling. Most importantly, the sampling error probability cannot be used to measure whether the evidence embodied in the fact probability is more likely than not true.

D. Hypothesis Testing Error

Even if (1) there is no error in the design of the experiment or in measuring the variables of interest, (2) the probability of sampling error is small, and (3) the statistician fully understands and correctly models the underlying phenomena, the p-value is not a reliable measure of the credibility of the evidence or the belief probability. Two additional sources of error involve the choice of methods for calculating p-values and the failure of statistical methods to detect relationships that exist.

Hypothesis testing is the process of making assumptions about relationships and then comparing the assumptions to the data by means of a statistical analysis. One might, for instance, assume that ingesting a drug is not related to the incidence of an adverse consequence, then test that assumption by collecting and analyzing data. The calculation of a sampling error probability is one

---

experience and education into account, the defendant's coefficient was only -$4.19 and had an associated p-value greater than .50, which demonstrates the significance of model design.

13. *See* Peterson & Conley, *supra* note 7, at 213.

14. People Who Care v. Rockford Bd. of Educ., School Dist. No. 205, 111 F.3d 528, 537-38 (7th Cir. 1997) (rejecting study claiming that discrimination caused poor student achievement test performance without considering the effects of poverty, parental involvement, education level of parents, or other factors on performance).

means of considering the assumption in light of the evidence. A p-value may, for instance, be based on the assumption that there is no relationship. When combined with a factual probability suggesting a relationship, a small p-value for the coefficient between the drug and the adverse result calls the assumption into question.

For many problems where a statistician wishes to measure the probability of sampling error, there is a choice of methods. Whether or not statisticians generally agree about the correct way to calculate a p-value in a given case, the mere appearance of a p-value does not guarantee that the correct choice of measures was made.[15]

Furthermore, not all proponents of statistical evidence want to prove that a relationship exists; some want to prove that it does not. Should a high p-value (closer to 1.0) suggest that there is no relationship? No. Different statistical tests differ in their ability or power to detect relationships that exist, and the power of any statistical test is dramatically affected by the size of the sample from which the coefficient was calculated.[16] A court might accept evidence that there is no relationship between a stimulus and a harm—no relationship, for instance, between progestins and birth defects. The hypothesis must be tested with a methodology with sufficient statistical power to detect a causal relationship if one exists. In *Ambrosini v. Labarraque*, the court inferred from the plaintiff's expert testimony that "[c]onventionally, in order to be considered meaningful, negative studies, that is, those which allege the absence of a causal relationship, must have at least an 80 to 90 percent chance of detecting a causal link if such a link exists; otherwise, the studies cannot be considered conclusive."[17] This eighty to ninety percent figure measures the power of the test but bears no relationship to the sampling error probability. Both the power of the statistical test and the sampling error probability affect the credibility of scientific testimony regarding cause, the belief probability.

A low-powered test (one that is less able to discern relationships or one based on a small sample size) may indicate no relationship even where there is one. Low power will affect the credibility of the proponent's evidence; this is yet another reason why the belief probability is divorced from the sampling error probability. There is no neat mathematical adjustment to the p-value to reflect the inability of the test to capture a relationship.

The ideas of power and statistical significance are complementary. Lack of power is one of many reasons a study might fail to find a relationship even though there is one. High sampling error is one reason a study might have "revealed" a relationship even though none exists. While the p-value measures the probability of finding a relationship where there is none because of the random

15. *See infra* Part V (illustrating the results of applying alternative methods of calculating p-values).

16. *See* Kaye, *supra* note 5, at 1357-62 (discussing power functions).

17. 101 F.3d 129, 136 (D.C. Cir. 1996).

composition of the study, power refers to the probability of not finding a relationship when one exists because of the limitations of the statistical test.

E. Extrapolation Error

Even if no disconnect results from any of the considerations identified above, the sampling error probability would be divorced from the belief probability just where we need it—where we ask whether the particular defendant's act was a necessary event in producing the particular plaintiff's harm.[18] The p-values from scientific/medical studies are based on general observations, not observations of the plaintiff. It is based on acts like the defendant's, not the defendant's act. It remains to be proven that what is a possible, even a probable, cause in general actually was a cause in this case. This is the problem of inferring specific causation from general causation. The p-value does not measure the probability that extrapolation is logical or permissible.

The article in this issue by Professors Freedman and Stark illustrates the difficulties of extrapolating from scientific/medical evidence.[19] In particular, they discuss why general observations may not apply to specific individuals who are atypical of the general population studied. This extrapolation error is distinct from the sampling error, which results from studying a sample that is different from the general population. In a torts case, the plaintiff may be different from both the sample and the general population. Because the p-value describes only the relationship between the sample and the general population, it says nothing about the applicability of the study to the plaintiff.

Statistical analysis of well-designed scientific/medical studies focuses on the probability that "chance," the possible unrepresentativeness of the sample, accounts for an observed relationship between treatment and outcome. The "but for" cause in torts is concerned with the likelihood that the outcome can be explained by "not chance," the probability that the treatment accounts for the outcome. The belief probability also relates to "not chance" rather than the probability of "chance." Since the treatment is only one of the "not chance" explanations for the harm, the probability of "chance" has nothing to do with the probability that the treatment is the correct "not chance" explanation. Perfect or ideal experimental design eliminates the possibility of "not chance" explanations other than the variable of interest influencing the outcome. The p-value does not measure the probability that the design was perfect; rather, it assumes the design was perfect. A factfinder must evaluate the characteristics of design and testing described earlier in this Part. Doing so, however, does not eliminate the possibility that the plaintiff in a particular case is different from

18. In *Cipolone v. Liggett Group, Inc.*, 893 F.2d 541, 561 (3d Cir. 1990), *aff'd in part, rev'd in part*, 505 U.S. 504 (1992), the court held that the but for cause test did not necessarily mean that there was a 50% chance that the defendant's conduct caused the harm. Rather, the test was whether, more likely than not, the act was a necessary part of the chain of events.

19. *See* David A. Freedman & Philip B. Stark, *The Swine Flu Vaccine and Guillain-Barré Syndrome: A Case Study in Relative Risk and Specific Causation*, 64 LAW & CONTEMP. PROBS. 49 (Autumn 2001).

the subjects sampled. The ideal of testing only subjects identical in all ways to the plaintiff is difficult, if not impossible, in practice.[20]

F. Risk Ratios and the Preponderance of the Evidence Rule

Risk ratios, a measure used by epidemiologists in their study of causation, suffer from the same problems as other coefficients. A risk ratio varies between zero and infinity, higher ratios reflecting increased probability of a risk materializing as a result of exposure to acts like the defendant's. A risk ratio of two, for instance, indicates that the probability of an adverse outcome doubles when people are exposed to the stimulus in question. A risk ratio of two is treated as a magical threshold because it is interpreted to mean that the exposure causes as many occurrences of the adverse outcome as background conditions cause. If the risk ratio is greater than two, some conclude, the preponderance of the evidence test is met—the risk is more likely than not caused by the exposure. This reasoning erroneously either ignores belief probabilities or conflates the fact and belief probabilities.

The risk ratio should be treated as related to the fact probability rather than the belief probability. It describes a fact, a numerical summary of the relationship between exposure and risk akin to the ideal coefficient the lawyer seeks. But there is nothing in that number that describes its credibility. Risk ratios are, however, likely to be based on a coefficient for which the sampling error probability has been or can be calculated. One might suspect that a combination of a p-value and a risk ratio would provide something akin to the belief probability. For all of the reasons described above, neither the sampling error probability, nor the fact probability, nor the two combined, measure the belief probability.

The risk ratio and the sampling error are only the start of an inquiry into the reliability of scientific evidence. Only if the fact probability is sufficiently high so that the risk ratio is greater than two, the fact probability is based on a study with a p-value that is sufficiently low, and the fact probability is based on a study meeting all of the additional challenges discussed above, can one begin to infer that there is a relationship between treatment and outcome. One would still not have a measure of the belief probability.

Courts look for studies with low sampling error probabilities and high factual probabilities (high risk ratios), citing those two criteria as the basis for establishing a credible case of causation.[21] Neither bears any necessary relation-

20. Freedman & Stark, *supra* note 19, illustrate the difficulty of concluding that treatment with flu vaccine caused the plaintiff in *Manko v. United States*, 636 F. Supp. 1419 (W.D. Mo. 1986), *aff'd in part, remanded in part*, 830 F.2d 831 (8th Cir. 1987), to suffer from Guillain-Barré syndrome. Their discussion identifies several factors leading to different rates of adverse outcomes for different people, including the age and health of the vaccinated individual.

21. *Cf.* Lucinda M. Finley, *Guarding the Gate to the Courthouse: How Trial Judges are Using Their Evidentiary Screening Role to Remake Tort Causation Rules*, 49 DEPAUL L. REV. 335, 347 n.49 (1999) ("Some judges have improperly blurred the two very distinct concepts of relative risk and statistical significance by labeling the results of epidemiological studies which derived a relative risk of less than 2.0 as 'statistically insignificant.'"). As an example, Finley cites *In re Joint Eastern and Southern District*

ship to credibility. In *Allison v. McGhan Medical Corp.*,[22] the court reviewed the district court's treatment of a study that showed a statistically significant correlation between silicone and increased antinuclear antibodies and a relative risk of 1.24. The district court concluded that the risk ratio of 1.24 was "so significantly close to 1.0" that the study was not worth serious consideration for proving causation.[23]

It is strange but true that a perfectly credible study may not be worth serious consideration. Imagine that one can be perfectly confident that 1.24 is a precise and accurate description of the increased risk of an adverse outcome resulting from treatment for the specific plaintiff. This evidence provides a sound basis for inferring that there is a causal connection but is not strong enough evidence. We are "confident," so the 1.24 figure is credible, and the belief probability is high. It is a precise estimate, so the sampling error is small. The problem is that the fact probability, while credible, is too small. To prove but-for causation by a preponderance of the evidence, both the belief and fact probabilities must be taken together.

The court in *Allison* recognized that a relative risk of 2.0 permits an inference that the plaintiff's disease was as likely as not caused by the agent, because it implies a fifty percent likelihood that the agent caused the disease.[24] Based on the probabilities represented in a relative risk of 2.0, *if they are true*, the treatment is as likely to have caused the outcome as all other causes combined. A risk ratio greater than two makes the treatment more likely than not a but for cause. Under a standard adopted by the Court of Appeals for the Ninth Circuit, on remand from the Supreme Court's decision in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, "plaintiffs' experts would have had to testify either that Bendectin actually caused plaintiffs' injuries (which they could not say) or that Bendectin more than doubled the likelihood of limb reduction birth defects (which they did not say)."[25] Plaintiffs must offer "either specific evidence that the drug actually caused their injuries or epidemiological proof that the drug 'more than doubles' the risk of birth defects because epidemiological evidence presents statistical likelihoods rather than direct information about the cause of

Asbestos Litigation, 827 F. Supp. 1014, 1041-42 (S.D.N.Y. 1993), *rev'd*, 52 F.3d 1124 (2d Cir. 1995), which describes epidemiological studies that yielded relative risks between 1.0 and 1.5 as "statistically insignificant." *See also* Charles Tomljanovic et al., *Anthropogenic Electromagnetic Fields and Cancer: A Perspective*, 8 RISK: HEALTH, SAFETY & ENV'T 287, 290 (1997) (stating that "[a] relative risk ratio of 2.0 or more may be considered a strong, statistically significant association between exposure and disease, and supports a causal relationship").

22. 184 F.3d 1300, 1315 (11th Cir. 1999).

23. *Id.* Causation can logically be established even if the risk ratio is less than two if there is particularized evidence, such as tissue samples, standard tests, and patient examination. Risk ratios, like other statistical tests, are only as believable as the underlying design, which includes the testability of the causal link. It may be the inherent difficulty of epidemiology, rather than the lack of a causal relationship, that is responsible for the lack of epidemiological evidence.

24. *See id.* at 1315 n.16 (citing REFERENCE MANUAL ON SCIENTIFIC EVIDENCE 168-69 (Federal Judicial Ctr. ed., 1994)).

25. 43 F.3d 1311, 1322 (9th Cir. 1995).

birth defects for the particular plaintiff."[26] The relative risk permits such an inference only if it is credible, which means that the belief probability is greater than fifty percent. The preponderance of the evidence standard is not met without a belief probability greater than fifty percent *and* a risk ratio greater than 2.0.

Risk ratios suffer from the same extrapolation problems as other coefficients and do not help courts avoid the difficulties presented by statistical science. These difficulties are not unknown to epidemiologists or medical science. Establishing a statistical relationship between the stimulus and the response—the act and the harm—is the first step in an epidemiologist's search for causation. An epidemiologist then looks to see whether there is "consistency in the findings of multiple scientific studies [to eliminate some of the sources of error mentioned above], biologic [sic] plausibility, the time sequence of the prospective cause and effect, the quantitative strength of the association, and the specificity with which the two phenomenon [sic] correlate."[27] When looking for the cause of a particular individual's harm, it would be most desirable to have an individualized risk ratio, because risk ratios are different for different groups of people. The lack of data for different groups of people enhances the difficulties inherent in extrapolation, reducing the belief probability associated with the fact probability based on a risk ratio.[28]

V

A FUNDAMENTAL DISCONTINUITY BETWEEN THE BURDEN OF PROOF AND THE SAMPLING ERROR

It is easy to be misled by many discussions of the role of statistical evidence. Consider an excerpt from one discussion of the scientific method. In this discussion, the author repeats the notion that statistical significance is the only test for reliability of scientific evidence:

> The legal "preponderance of the evidence" standard has nothing to do with assessing whether data is scientifically reliable. Since *Daubert* seeks to exclude scientifically unreliable evidence, the scientific evidence must conform to the accepted convention of 95 percent probability to be admissible. Once all scientifically reliable evidence (that is, that meeting the 95 percent threshold) is introduced, the fact finder will determine by the preponderance of all admissible scientifically reliable evidence, whether the plaintiff has met the burden of proof.[29]

---

26. Ambrosini v. Labarraque, 101 F.3d 129, 135 n.8 (D.C. Cir. 1996) (describing the standard adopted by the Ninth Circuit in *Daubert*).

27. *Id*. at 136.

28. *See Daubert*, 43 F.3d at 1321 n.16 (discussing how a risk ratio less than two could be combined with other evidence to show causation); *see also* Mark Parascandola, *What is Wrong with the Probability of Causation?*, 39 JURIMETRICS J. 29, 35-36 (1998) (explaining why risk ratios should be based on plaintiffs' particularized probability of causation and how there is no necessary relationship between the size of the risk ratio (fact probability) and the strength of the evidence (belief probability)).

29. Bruce R. Parker, *Effective Strategies for Closing the Door on Junk Science Experts*, 65 DEF. COUNS. J. 338, 347 (1998).

A reader of this excerpt from and article by Bruce Parker might believe that all statistically significant evidence is scientifically reliable, though the previous section has demonstrated that this is not true. It would be easy to cite Parker's article as authority for the proposition that scientific evidence is reliable if it is statistically significant, even though elsewhere Parker reveals his awareness that the fact of statistical significance does not mean that a study's results are "good scientific evidence."[30] Parker observed that the study from which the p-value was calculated might have been poorly designed or controlled.[31] He was also aware that even statistically significant results from a well-controlled study may have no biological significance.[32] The previous sections have illustrated not only that there is no magical significance to the .05 p-value (translated in the excerpted text into a ninety-five percent threshold), but that reliability depends on much more than statistical significance.[33]

There is no convenient way to translate the .05 p-value into a ninety-five-percent confidence that the fact probability is correct, credible, believable, or true. There is a fundamental disconnect between what factfinders want to know and the information a statistical study provides. In theoretical discussions of the use of statistical evidence in legal proof, it is well recognized that statistical significance is not the same as the civil burden of proof.

Consider again the but for cause test and the sampling error probability. The but for cause test in civil cases requires the factfinder to determine the probability that the treatment caused the outcome. In scientific/medical cases, that is the plaintiff's assertion or hypothesis. In mathematical terms, the plaintiff bears the burden of showing that the probability *P* of the hypothesis *H* being true, given the evidence *E*, is greater than fifty percent, that is, plaintiff must demonstrate that $p(H \mid E) > .50$.

While the plaintiff must prove that the hypothesis is (more likely than not) true, the p-value assumes that the hypothesis is true. The p-value measures the probability of finding such evidence *E* if *H* is true. For a statistical result to be statistically significant by conventional standards, a scientist must show that the sampling error probability is less than .05, that is, the scientist must demonstrate that $p(E \mid H) < .05$. Since the statistical calculation assumes that the hypothesis

30. *Id*.

31. *See id*.

32. *See id*.

33. It is not difficult to find quotations that misrepresent the test for reliability of scientific evidence. This is not necessarily because commentators are unaware of the limits of statistical science. Rather, scholarly commentators and courts are often considering only a piece of the statistical puzzle. As the previous section illustrated, this is quite common in discussions of the relationship between risk ratios and statistical significance. *See*, *e.g.*, M. Elizabeth Karns, *Statistical Misperceptions*, FED. LAW., June 2000, at 19, 21 (focusing on the importance of statistical significance when evaluating risk ratios and discussing how statistically significant risk ratios less than two may still be worth a court's consideration); *see also* Finley, *supra* note 21. The rest of the picture is the credibility of both of those numbers. *See* Erica Beecher-Monas, *A Ray of Light for Judges Blinded by Science: Triers of Science and Intellectual Due Process*, 33 GA. L. REV. 1047, 1102 n.327 (1999) (discussing how more goes into the validity calculus when evaluating the legal admissibility of expert testimony than either relative risk or statistical significance).

is true, it does not measure whether it is true. Significance testing begs the questions courts need ultimately to answer.

The closest any mathematical approach has come to translating p-values into belief probabilities is Bayes Theorem.[34] Bayes Theorem specifies the additional information needed to translate something like a p-value into the likelihood that a hypothesis is true given the evidence.[35] Bayes Theorem states that the likelihood that the hypothesis is true depends on the sampling error, $p(E|H)$, the probability that the hypothesis is generally true without regard to the particular evidence, $p(H)$, and the probability of occurrence of evidence such as that observed given the probability that the hypothesis is true:[36] $p(H|E) = [p(H) \times p(E|H)] \div p(E)$. To translate the p-value into a measure of the likelihood that the treatment caused the outcome, the factfinder needs to know, *a priori*, the probability that the hypothesis is true, $p(H)$. Then the factfinder needs to calculate the likelihood of finding such evidence given that probability. This information is missing from medical/scientific studies of tort causation. Statistical science has, at this point, provided all the information it can.

Where statistical science ends its inquiry, logic must take over. It is the proper design and analysis of scientific studies that leads scientists from p-values to inferences about causation. Having designed and carefully analyzed the study, and having discovered that the sampling error is unlikely to explain the outcome, the logical alternative choice is the treatment.

David Peterson has a favorite story to explain the limits of the sampling error probability.[37] He describes a clown entering a room flipping a coin. The clown flips the coin three times, obtaining three tails, then leaves the room. A statistician can calculate the probability of obtaining three tails from a sample of three just by chance if the coin is fair. That is the $p(E|H)$ described above. Without more information, however—information that is outside the statistician's normal province—the statistician cannot say whether the clown's coin is fair. That is the $p(H|E)$ above. If it is fair, the probability of that outcome arising by chance can be determined. Whether it is fair is a mystery until we bring in a metallurgist, a physicist, or perhaps a numismatist. Unfortunately, we are particularly interested in whether the "unfairness" of the coin explains the outcome.

---

34. An exposition of Bayes Theorem appears in RONALD J. WONNACOTT & THOMAS H. WONNACOTT, ECONOMETRICS § 10.1, at 198 (1970).

35. In Bayes Theorem, the likelihood of observing such evidence if the hypothesis is true [$p(E|H)$] corresponds to the p-value if "observing such evidence" means "observing evidence as inconsistent with or even more inconsistent with the hypothesis as was observed in these data."

36. *See*, *e.g.*, Kingsley R. Browne, *The Strangely Persistent "Transposition Fallacy": Why "Statistically Significant" Evidence of Discrimination May Not Be Significant*, 14 LAB. LAW. 437, 444 n.21 (1998) (applying Bayes Theorem to employment discrimination context); Jonathan J. Koehler & Daniel N. Shaviro, *Veridical Verdicts: Increasing Verdict Accuracy Through the Use of Overtly Probabilistic Evidence and Methods*, 75 CORNELL L. REV. 247, 255 n.27 (1990) (discussing Bayes Theorem).

37. *See* David Peterson, Judging Science, Lecture at the Duke University School of Law Private Adjudication Center Symposium (May 24, 2000).

Whether the coin is fair is analogous to the "but for" cause question in tort law. We are interested in comparing alternative explanations for an outcome. One possibility is that chance, the happenstance of getting three out of three heads even if the coin is fair, accounts for the result. Others are that an imbalanced (unfairly weighted) coin explains the greater number of tails, or perhaps an unfair coin-flipping method accounted for the result. Statistics can tell us the probability that chance accounts for the result if the coin is fair. If the sampling error is small, it suggests that something else—perhaps a weighted coin or perhaps the unfair flipping—accounts for the result. We need other experts to establish what that "something else" is.

Since the "something else" is the relevant question in tort law, it might seem that statistical science is of little practical utility. Combined with scientific research design, however, statistical science is very powerful. The elements of research design eliminate as logical alternative explanations for the observed outcome all but two explanations: chance (the happenstance of a nonrepresentative sample) and the explanation of interest (in torts, exposure to the defendant's product). Eliminating confounding explanations in the coin-flipping example requires designing the experiment so that, if chance is an unlikely explanation because the sampling error is very small, then an unfair coin is the only other logical alternative. Scientific research design in the torts causation context requires designing product tests so that, if sampling error is unlikely to explain the harm associated with ingestion of a drug, then it must be the drug itself. The sampling error probability does not measure the likelihood that the experiment has been scientifically designed; it assumes that the design was flawless and calculates the likelihood that the random selection of subjects accounts for the outcome.

VI

CONCLUSION

When an experiment has been scientifically designed to eliminate explanations other than the influences of chance and the product in question, statistical analysis may eliminate chance as an explanation, leaving only the product in question. Eliminating chance as a plausible explanation for the outcome is a sine qua non of drawing a logical conclusion that the product caused the outcome. It is naturally tempting to judge reliability of a scientific study by its statistical significance. This article has illustrated why the statistical significance question is only the beginning of the inquiry. It has identified a number of other factors relevant to the question of reliability. If these elements of scientific design and testing are not met, the measure of statistical significance is itself meaningless.

The coin-flipping clown example illustrates the errors resulting from improper application of the scientific method. Experimental design error may explain the outcome if the clown has learned to flip coins in a way that affects the outcome. Even if the clown is not skilled enough to make it land on heads

every time, starting with the head faced down every time and flipping so that the coin rotates as few times as possible may affect the outcome, regardless of whether the coin is fair or not. The result of experimental error is that the sampling error estimate and the result of three tails are unreliable descriptions of the coin's characteristics. Similarly, if the clown obscures the result of each flip and merely reports it, his impaired vision, rather than any characteristic of the coin, may account for the reported result.

Statistical modeling error may account for the low p-value. It might be that whether the coin starts its voyage into the air face down or face up makes a systematic difference in the outcome. If we incorrectly assume the contrary, then we are implicitly assuming that the design or weight distribution of the coin (its "fairness") is the only relevant explanatory factor. The simple model that assumes that only one factor influences the outcome may give us unreliable evidence about the coin. Unless the model is changed, omission of a relevant explanatory factor yields a sampling error based on false assumptions. The belief probability depends on the accuracy of the model and on the assumption that the sampling error is correct. The further the statistical model departs from reality, the less reliable the sampling error will be.

Hypothesis testing error may account for the low p-value. The p-value may be calculated by various methods. The best test is one designed for the particular experiment. Selecting the wrong test might change the p-value. For example, use of the Z-test, which is appropriate for larger samples than this, yields a p-value of .0416 in this case, while the preferable binomial test yields a p-value of .125, quite different compared to the .05 significance level.[38] The appropriate test depends on the reason why we are asking the question, the size of the sample, the process the analyst attempted to model statistically, and the properties of the underlying population. The p-value depends on the choice of test. Using either the improper test or a low-powered test will yield a misleading p-value and affect the belief probability.

Extrapolating from the results of the clown's flipping to the likely results of other people flipping other coins may lead to inaccurate results if this coin was unfair (or if the flipping was biased), a fact about which the statistical test has revealed little. The more differently other coins are constructed, the less reliable are either the fact of eight tails or the resulting sampling error. The belief probability suffers accordingly.

Courts cannot avoid evaluating the underlying scientific and statistical methodology when evaluating scientific evidence of causation. In particular, they may not rely on the statistical significance of the study, as measured by the sampling error probability, to conclude that the evidence is scientifically reliable

38. Similarly, one must choose in many cases between a two-tailed test (which identifies the probability of getting such an extreme number of heads or tails), or a one-tailed test (which identifies the probability of this many tails). The issue of whether a one- or two-tailed test is appropriate arises often in employment discrimination cases. *See*, *e.g.*, Palmer v. Schultz, 815 F.2d 84, 92-97 (D.C. Cir. 1987) (discussing one-tailed and two-tailed tests of statistical significance).

or whether the preponderance of the evidence test is met. Nor may they rely on the sampling error probability in combination with the fact probability. It is the belief probability, based only in part on an estimate of the sampling error probability, that determines whether the fact probability is more likely than not true. The belief probability also depends, however, on the other indicia of reliability discussed.