LOST IN TRANSLATION: Statistical Inference in Court

Erica Beecher-Monas[*]

Scientists and jurists may appear to speak the same language, but they often mean very different things. The use of statistics is basic to scientific endeavors. But judges frequently misunderstand the terminology and reasoning of the statistics used in scientific testimony. The way scientists understand causal inference in their writings and practice, for example, differs radically from the testimony jurists require to prove causation in court. The result is a disconnect between science as it is practiced and understood by scientists, and its legal use in the courtroom. Nowhere is this more evident than in the language of statistical reasoning.

Unacknowledged difficulties in reasoning from group data to the individual case (in civil cases) and the absence of group data in making assertions about the individual (in criminal cases) beset the courts. Although nominally speaking the same language, scientists and jurists often appear to be in dire need of translators. Since expert testimony has become a mainstay of both civil and criminal litigation, this failure to communicate creates a conundrum in which jurists insist on testimony that experts are not capable of giving, and scientists attempt to conform their testimony to what the courts demand, often well beyond the limits of their expertise.

This garbled communication has severe consequences in both civil and criminal litigation. Particularly in medical causation and criminal identification cases, courts routinely exclude testimony that is scientifically sound, and admit expert testimony that is wholly lacking in scientific basis. Not only do jurists misunderstand the meaning of common terms like statistical significance, confidence intervals, and relative risk, but they pervasively misunderstand the limits of statistical inference drawing.

Statistics are crucial to the scientific enterprise. Statistics can be very helpful in determining the size of a scientific study, accounting for randomness, and comparing risks, among other things.[1] All scientific fields make use of statistics. But what statistics cannot do—nor can the fields

[*] Professor of Law, Wayne State University Law School. Thanks to Steve Winter and Terence Anderson for their comments, and to the excellent editors of this Law Journal for their editorial assistance.

[1] *See, e.g.*, MICHAEL B. BRACKEN, RISK, CHANCE, AND CAUSATION: INVESTIGATING THE ORIGINS AND TREATMENT OF DISEASE 127 (2013) (discussing the role of statistics in designing studies).

employing statistics, like epidemiology and toxicology, and DNA identification, to name a few—is to ascribe individual causation.

Statistics is the law of large numbers. It can tell us much about populations. It can tell us, for example, that so-and-so is a member of a group that has a particular chance of developing cancer. It can tell us that exposure to a chemical or drug increases the risk to that group by a certain percentage. What statistics cannot do is tell which exposed person with cancer developed it because of exposure. This creates a conundrum for the courts, because nearly always the legal question is about the individual rather than the group to which the individual belongs.

Not that experts are unwilling to ascribe individual causation. On the contrary; such testimony is a mainstay in both civil and criminal litigation. The problem is that such testimony exceeds the capacity of science, and the experts are therefore testifying beyond the limits of their scientific expertise.

For example, in toxic torts, where courts demand testimony on both general and specific causation, while general causation in the form of a population statement is similar to what these experts (epidemiologists and others) do in their respective fields, the specific causation statement is not. Specific causation experts—generally medical doctors—attempt to determine individual causation through what the courts call "differential etiology."[2] Differential etiology, however, "is an exercise designed for the courtroom."[3] Medical schools do not teach it. It is not a part of doctors' normal practice.

Diagnosis—which doctors are trained in—involves assessing symptoms and running tests, but outside of infectious diseases, does not ordinarily involve determining causation. Figuring out what the illness is and treating it are the normal tasks for doctors. The doctor who diagnoses "cancer" does not (and cannot scientifically) determine the cause. There may be many causes, some of them interacting. Yet courts require testimony that goes beyond what medicine can do.

While the chasm in medical causation cases is the inability of science to reason from the general data to the individual, conversely, in criminal cases—outside of DNA testimony—there is a striking absence of general data. Instead, the reasoning in criminal identification focuses strictly on the

[2] David L. Faigman et al., *Group to Individual (G2I) Inference in Scientific Expert Testimony*, 81 U. CHI. L. REV. 417, 435 n.79 (2014).

[3] A. Philip Dawid et al., *Fitting Science into Legal Contexts: Assessing Effects of Causes or Causes of Effects?*, 43 SOC. METHODS & RES. 359, 369 (2014); *see also* CAUSALITY: STATISTICAL PERSPECTIVES AND APPLICATIONS, at xxiii (Carlo Berzuini, Philip Dawid & Luisa Bernardinelli eds., 2012) (noting that "it may simply be impossible, even with the best data in the world, to estimate causes of effects at the individual level without making arbitrary and empirically untestable additional assumptions.").

individual, claiming that each individual is unique, without general data to support that claim. Forensic science experts are willing to testify to the uniqueness of particular patterns (in fingerprints, for example) without any general population data. The assumptions on which such testimony is based exceed the current bounds of science, and reflect a profound misunderstanding of statistical inference drawing. Nonetheless, this testimony is routinely admitted in our criminal courts.

The lack of understanding about what statistics can and cannot do—and therefore what scientists who rely on statistics can legitimately say about the issues before the court—has severe repercussions in a legal system that depends as heavily as ours on expert testimony. The goal of this article is to provide translation of complex statistical concepts currently confounding judicial admissibility decisions, enabling courts to better determine when expert testimony is genuinely helpful to the jury. This article proceeds in three parts. Part II discusses judicial gatekeeping requirements, relevance and reliability standards, as they relate to expert testimony. Part III focuses on statistical misunderstandings in civil and criminal courts. Part IV explores possible solutions to the conundrum faced by experts and the courts. Part V concludes that correct translation of statistical concepts and their inherent limitations is key to achieving justice in our courts.

[redacted] GATEKEEPING: THE MEANING OF RELEVANCE

The commitment to a rational system of evidence entails the exclusion of irrelevant information.[4] Even scholars arguing for "free proof" acknowledge the importance of screening information to ensure that it has some tendency to make a disputed issue in the case more or less probable.[5] Only facts having rational probative value should be admissible in the search for truth.[6] If something is not logically probative, no rational system of evidence should

[4] *See* William Twining, *The Rationalist Tradition of Evidence Scholarship*, *in* RETHINKING EVIDENCE: EXPLORATORY ESSAYS 32 (1990) (discussing the rationalist tradition).

[5] *See, e.g.*, Michael S. Pardo, *On Misshapen Stones and Criminal Law's Epistemology*, 86 TEX. L. REV. 347, 354–63 (2007) (book review).

[6] The doctrines of relevance and probativity are expressed as follows under the Federal Rules of Evidence: "Evidence is relevant if: (a) it has any tendency to make a fact more or less probable than it would be without the evidence; and (b) the fact is of consequence in determining the action." FED. R. EVID. 401; and "[t]he court may exclude relevant evidence if its probative value is substantially outweighed by a danger of one or more of the following: unfair prejudice, confusing the issues, misleading the jury, undue delay, wasting time, or needlessly presenting cumulative evidence." FED. R. EVID. 403.

consider it. Something is relevant or not, in relation to a disputed legal issue (the facet of relevance that the *Daubert* court referred to as "fit"[7]).

Admissibility of expert testimony in federal courts is governed by Federal Rule of Evidence 702, which places the threshold of admissibility at helpfulness to the jury.[8] In its transformative *Daubert* opinion, the Supreme Court replaced the nearly universal general consensus standard for the admissibility of scientific expert testimony[9] with a requirement that judges must evaluate the scientific validity of expert testimony.[10] As the *Daubert* Court explained, the requirement that expert testimony assist the trier of fact "goes primarily to relevance."[11] *Daubert*, currently the predominant rule on the admissibility of expert testimony,[12] emphasized "appropriate validation" and "good grounds" as the cornerstones of admissibility.[13] Gatekeeping for judges who simply do not understand the statistical inferences that they are required to evaluate becomes a muddle. In the civil cases, courts tend to rely on rules of thumb and bright line cut-offs (like requiring relative risks of two or more and rejecting confidence intervals that include relative risks of one) and asking for medical testimony that doctors are not capable of giving. In the criminal cases (with the exception of DNA evidence) the statistical misunderstandings are nearly the reverse of those in the civil toxic tort cases. Rather than insisting on general population testimony first (as courts do with general causation testimony in toxic torts), criminal courts start with the individual (what in toxic torts would be called specific causation) and never get to the general. This misunderstanding of statistics is extremely troubling because it affects the search for truth on which our legal system is based.

The *Daubert* Court also noted that "evidentiary reliability will be based on scientific validity."[14] So when it comes to expert testimony, relevance must be considered in tandem with reliability.[15] *Daubert* and amended Rule

[7] Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 579, 591 (1993).

[8] Federal Rule of Evidence 702 begins with the proviso that expert testimony is admissible only if "the expert's scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue." FED. R. EVID. 702.

[9] The general consensus standard originated with Frye v. United States, 293 F. 1013, 1014 (D.C. Cir. 1923) (scientific testimony must "be sufficiently established to have gained general acceptance in the particular field in which it belongs.").

[10] *Daubert*, 509 U.S. at 591.

[11] *Id.*

[12] *See generally* David L. Faigman & John Monahan, *Psychological Evidence at the Dawn of the Law's Scientific Age,* 56 ANN. REV. PSYCHOL. 631, 632 (2005) (observing that the *Daubert* test applies in all federal cases, and a majority of states have adopted the *Daubert* framework).

[13] *Daubert*, 509 U.S. at 590.

[14] *Id.* at 509 n.9.

[15] Justice Blackmun explained that reliability for admissibility purposes is different from what scientists call reliability (which he defined as getting "consistent results") in that for legal

702 both stress reliability of expert testimony as a facet of relevance, and therefore of admissibility.[16] This is a particular problem in criminal identification cases, because reliability requires sufficient data, a requirement that—in the absence of general data about the prevalence of particular patterns in the population—criminal individuation testimony cannot meet. In toxic torts, the reliability problem appears in specific causation testimony, which is not reliable because the inference leap from general data to individual causation is unsupported by science.

The courts' muddle over statistically based testimony goes in both directions: individuation testimony may appear to have legal fit but, because it lacks empirical support, fail to be relevant. General causation testimony based on statistical significance, relative risk and confidence intervals, on the other hand, has both legal fit and scientific basis, and ought not to be excluded for failing to meet judicially imposed standards that do not affect their validity.

Admitting only relevant evidence is key to preventing the danger that irrelevancies may be mistaken as bearing on the question at hand. Admitting irrelevant information may make the ultimate decision unfounded and inaccurate (or, if accurate, only by chance). Such evidence is affirmatively misleading. If the input is wrong, no reasoning process can be expected to make correct inferences.[17]

Although inaccuracy is a possible factor in any evidence, not just expert testimony, baseless expert testimony is particularly pernicious because the entire reason it is being admitted is that the jury lacks the background knowledge necessary to evaluate it.[18] (So do judges, but judges at least have the benefit of training in critical thinking, guidelines for the evaluation of scientific testimony, repeat exposure and a measure of accountability.)[19]

purposes, reliability means scientific validity (which he defined as "the principle supports what it purports to show" and "trustworthiness"). *Id.*

[16] FED. R. EVID. 702 (to be admissible, expert testimony must be based on sufficient data, and reliable methods); *Daubert*, 509 U.S. at 589 (courts should screen expert evidence for relevance and reliability).

[17] *See* Alvin I. Goldman, *Simple Heuristics and Legal Evidence*, 2 L. PROBABILITY & RISK 215, 219 (2003) (explaining that even deductive reasoning requires true premises in order to reach true conclusions).

[18] *See* Mark P. Denbeaux & D. Michael Risinger, *Kumho Tire and Expert Reliability: How the Question You Ask Gives the Answer You Get*, 34 SETON HALL L. REV. 15, 30 (2003) (*Daubert* implies a view that misleading expert evidence is worse—and less amenable to correction through cross-examination—than misleading lay testimony).

[19] *See* ERICA BEECHER-MONAS, EVALUATING SCIENTIFIC EVIDENCE: AN INTERDISCIPLINARY FRAMEWORK FOR INTELLECTUAL DUE PROCESS 33–35 (2007) (discussing why judicial gatekeeping has more potential for reaching accurate conclusions about expert testimony than simply admitting the evidence subject to cross-examination).

The problem for individuation testimony (specific causation and criminal identification) is that because it is without scientific basis, it cannot assist the jury. Unlike testimony based on studies that fail statistical significance, relative risk or confidence interval limits set by the courts—which may make the testimony shaky but still admissible—individuation testimony should be excluded. While cross-examination and the presentation of contradictory expert testimony are the traditional cures for "attacking shaky but admissible evidence,"[20] and thus correct testimony based on studies with statistical significance levels less than 95%, relative risks less than two, and confidence intervals that include the relative risk of one, expert testimony that lacks any empirical basis is resistant to this kind of correction. In the absence of data, the assumptions made by an expert sound perfectly plausible.[21] That is a problem for both specific causation and criminal identification evidence.[22]

[redacted]CONUNDRA IN THE COURTS

A. Civil Toxic Torts

Courts in toxic tort cases require proof of causation at two levels: general causation and specific causation. To prove general causation, courts tend to require testimony based on epidemiology that a particular chemical to which the plaintiff was exposed is capable of causing injuries like that suffered by the plaintiff. Sometimes this testimony is supplemented by toxicology, physiology and chemical structure testimony. Specific causation is proved

[20] *Daubert*, 509 U.S. at 596 ("Vigorous cross-examination, presentation of contrary evidence, and careful instruction on the burden of proof are the traditional and appropriate means of attacking shaky but admissible evidence.").

[21] Justice Learned Hand (well over a century ago) expressed the jury's dilemma with respect to expert testimony, "how can the jury judge between two statements each founded upon an experience confessedly foreign in kind to their own?" Learned Hand, *Historical and Practical Considerations Regarding Expert Testimony*, 15 HARV. L. REV. 40, 54 (1901).

[22] In mock jury studies about the effectiveness of cross-examination in criminal cases, it apparently made little difference whether the defense challenged the expert testimony; whether the defense pointed out in cross examination that the expert's conclusions were inconsistent with prior research and that the expert had not followed standard methodology; and whether the defense not only cross-examined the prosecution expert, but also put on its own expert. *See* Joseph Sanders, *The Merits of Paternalistic Justifications for Restrictions on the Admissibility of Expert Evidence*, 33 SETON HALL L. REV. 881, 936 (2003) (discussing the experimental work of Shari Diamond et al., and concluding that "rulings excluding unreliable evidence promote jury accuracy even if we assume jurors are as good as judges in assessing reliability."). Although jurors in these studies discussed the expert evidence in their deliberations, and although there was a strong correlation between the prosecution expert's testimony and the jury's verdict preferences, the results did not vary among the first three conditions. *Id.* at 934.

through testimony of medical doctors that exposure to the defendant's chemical was the cause of the plaintiff's injury.

This sounds pretty straight-forward, but it is not so simple. Causation in the biological sciences is complex and probabilistic. It is not like Newtonian causation (an example of which would be throwing a stone through a window, causing it to shatter). Rather, biologic causation is probabilistic, with many factors converging to cause disease. Both genetics and environment are undoubtedly involved in nearly every case of illness.[23]

Moreover, quite a few diseases appear in the general population without known cause (a phenomenon referred to in medical practice as "idiopathic" disease). The same kind of cancers, for example, may appear in the general population without any known toxic exposure as may appear after toxic exposure. And determining which identically manifested disease was caused by which factor is beyond the capacity of medical science.[24] Indeed, epidemiologists speak in terms of causal pies rather than a single cause. It is simply not possible to infer logically whether a specific factor caused a particular illness.[25]

Causation in cancer (and probably other diseases as well) tends to have multiple pathways.[26] Long latency periods between exposure to a carcinogen and disease manifestation add to the uncertainty. Epidemiologists are also increasingly considering the role of individual genetic susceptibility exacerbated by environmental exposure.

Probabilistic reasoning relies on statistical concepts of randomness. Judges (and most people) struggle with these statistical concepts. Although statistical concepts are basic to understanding causation in biological systems, this kind of probabilistic reasoning does not mesh well with tort law, and some judges handle the uncertainty better than others.[27] Adding to the

[23] Kenneth J. Rothman, a preeminent epidemiologist, illustrates the concept that "every causal mechanism involves the joint action of a multitude of component causes" as a causal pie, in which "some component causes play a more important role than other factors in the causation of disease." KENNETH J. ROTHMAN, EPIDEMIOLOGY: AN INTRODUCTION 24–25 (2d ed. 2012).

[24] For example, phenylketonuria, which results in mental retardation, is usually considered a genetic disease, but the mental retardation that results from it can be prevented by diet. *See id.* at 24.

[25] As Rothman explained the problem, "it is not possible to infer logically whether a specific factor was the cause of an observed event." *Id.* at 250.

[26] *See, e.g.*, Mel Greaves, *Cancer Causation*: *The Darwinian Downside of Past Success?*, 3 LANCET ONCOLOGY 244 (2002) (proposing a causal network explanation of cancer causation).

[27] *See, e.g.*, Jed S. Rakoff, *Science and the Law: Uncomfortable Bedfellows*, 38 SETON HALL L. REV. 1379, 1391 (2008) (explaining how, as a federal district court judge, he resolved the uncertainty issue raised by the absence of an epidemiology study in the ephedra litigation by permitting plaintiffs' experts to testify "that there is a reliable basis to believe that ephedra may be a contributing cause of cardiac injury and strokes in people with high blood pressure, certain serious heart conditions, or a genetic sensitivity to ephedra—provided that such experts qualify

confusion is the judicial bifurcation of proof into general and specific causation, neither of which is a concept used by scientists in their practices. All this uncertainty tends to make causation a highly problematic area for toxic torts.

[redacted] General Causation: Cursory Checklists and Bottom Lines

Terms such as statistical significance, relative risk (or its close cousin, the odds ratio), and confidence interval are ubiquitous in epidemiological testimony.[28] Rather than attempting to understand these as inter-related concepts, judges far too often treat them as separate thresholds that each study relied upon must cross. Using these concepts as exclusionary rules is a drastic misinterpretation of their meaning.[29]

[redacted] *Statistical Significance*

The courts frequently treat statistical significance as either being present or not. These judges exclude testimony based on studies that fail to meet statistical significance at an observed significance level of 95% (or P-value of 0.05), reducing statistical significance to a crude dichotomy. That interpretation, however, garbles the meaning of statistical significance.[30]

A better way to understand statistical significance is as a description of the role of chance.[31] It measures the consistency between data and the hypothesis being tested if the model used to compute the P-value is correct.[32] The P-value is the probability, assuming the null hypothesis (of no effect) is true

their testimony with the acknowledgment that none of this has been the subject of definitive study and may yet be disproved" and noting that the case settled shortly after his ruling).

[28] *See* ROTHMAN, *supra* note 23, at 148 (observing that statistics in epidemiology has two central roles, to "measure variability in the data in an effort to assess the role of chance, and . . . to estimate effects after correcting for biases such as confounding").

[29] *See* Sander Greenland & Charles Poole, *Problems in Common Interpretations of Statistics in Scientific Articles, Expert Reports, and Testimony*, 51 JURIMETRICS J. 113, 116 (2011) ("Much misinterpretation arises from attempts to simplify complex concepts or statistical conclusions.").

[30] *Id.* at 120 (using example of diet drug testimony). The P-value is a complex concept, hypothesizing randomly selected subjects in all conceivable study repetitions, defined as the "probability of getting data that conflict with the tested hypothesis as much as or more than the observed data conflict with the tested hypothesis," providing that the tested hypothesis is correct, all other assumptions used in computing the P-value are correct, and conflict with the tested hypothesis is gauged by a particular measure called the test statistic. *See id.* at 117–18. Different models (of which there are many "conflicting . . . candidates for the correct model") or test statistics may yield different conclusions. *Id.* at 118.

[31] *See* ROTHMAN, *supra* note 23, at 247 (discussing statistical significance).

[32] *See* Greenland & Poole, *supra* note 29, at 116 (explaining the concept of statistical significance).

(and the study is free of bias) of observing as strong an association as was observed.[33] While this may sound impenetrable to lawyers and judges, what they need to understand is that statistical significance is a measure of the relative consistency of the null hypothesis and the data. For example, a P-value of 0.01 means that the data are not very consistent with the null hypothesis, whereas a P-value of 0.5 means the data are reasonably consistent with the null hypothesis.[34]

Statistical significance level is a choice, not a mandate.[35] Judicial rejection of studies that fail to meet "statistical significance" (meaning a P-value of 0.05 or significance level of 95%) reflects a misunderstanding of why a particular P-value (or significance level) is chosen. The level of statistical significance chosen is a tradeoff between false positives and false negatives.[36] By setting a high significance level (low P-value), a scientist may avoid claiming an association where there is none, but at the risk of missing an association that is there. If the significance level is set too low (i.e., the P-value is set too high), a scientist may include associations that do exist, but at the risk of claiming an association where there is none. Scientists conventionally attempt to minimize the probability of failing to reject a false hypothesis by setting the significance level at 95% (P-value at 0.05).[37] But there may be good reasons for choice of a different significance level.

The size of the study is one reason for choosing a significance level other than 95%; the size of the expected effect is another. Statistical significance depends on both the size of the study and on the size of the observed effect. Larger studies may achieve statistical significance even where there is no effect, while small studies may not show statistical significance even though there is an effect. [38]

Power, the probability that the study in which the hypothesis is being tested will reject the alterative hypothesis when it is false, increases with the size of the study.[39] Power also increases with the degree of difference from

[33] ROTHMAN, *supra* note 23, at 150.

[34] *Id.*

[35] *See* THEODORE COLTON, STATISTICS IN MEDICINE 128 (1974) (noting that the P-value of 0.05 is chosen to minimize false positive errors).

[36] *See* BEECHER-MONAS, *supra* note 19, at 60–68 (discussing the inter-relationships between power, study size and biological context).

[37] *See* COLTON, *supra* note 35.

[38] Statistical tests, such as significance and relative risk, depend on the size of the study; in large studies, even small effects may be significant, while in a small study even a large effect may not be statistically differentiated from chance. *See* ROTHMAN, *supra* note 23, at 247 ("[F]or a given strength of association, more data results in a smaller *P* value.").

[39] *See, e.g.*, David H. Kaye & David A. Freedman, *Reference Guide on Statistics*, *in* FED. JUDICIAL CTR., REFERENCE MANUAL ON SCIENTIFIC EVIDENCE 211, 253–54 (3d ed. 2011) (discussing the statistical concept of power).

the null hypothesis (the effect size). The investigator will thus choose the significance level based on the size of the study, the size of the effect, and the trade-off between Type I (incorrect rejection of the null hypothesis) and Type II (incorrect failure to reject the null hypothesis) errors.[40]

Defense experts in toxic tort cases frequently assert that failure to disprove the null hypothesis means that the chemical in question has no effect.[41] The significance level (or P-value) cannot tell you whether the null hypothesis is correct, however. Failure to reject the null hypothesis only means that the data could as well be explained by chance. In order to test the hypothesis that the chemical in question has no harmful effects, one would have to study that hypothesis directly, and the data would have to demonstrate that the alternative hypothesis (that there is an effect) can be rejected. In other words, ambiguous evidence is not negative evidence.[42]

The unthinking use of statistical significance as a screening device leads to both over- and under-inclusiveness.[43] Judges who categorically exclude any testimony based on studies with statistical significance levels of less than 95% fail to recognize that the numbers are chosen because they reveal something about the study. Using statistical significance as a screening device is thus mistaken on many levels.[44]

[redacted]*Relative Risk*

Judges similarly misunderstand the concept of relative risk, often requiring a relative risk of two, or a doubling of the risk, before admitting epidemiology testimony.[45] Relative risk is an epidemiologic term referring to

[40] *See* Beecher-Monas, *supra* note 19, at 63 (2007) (discussing trade-offs between Type I and Type II errors).

[41] *See, e.g.*, Transcript of Record at 18, Baxter Healthcare Corp. v. Denton, No. 99CS00868, 2002 WL 31600035 (Cal. App. Dep't Super. Ct. Oct. 3, 2002) (defense expert testifying that "human data demonstrates that DEHP does not pose any risk of cancer" when the data only failed to reject the null hypothesis).

[42] *See* Austin B. Hill, *The Environment and Disease: Association or Causation?*, 58 Proc. Royal Soc'y Med. 295, 300 (1965) ("[T]oo often we deduce 'no difference' from 'no significant difference.'").

[43] *See* Kenneth J. Rothman, *Significance Questing*, 105 Annals Internal Med. 445, 446 (1986) ("An algorithm for inference cannot substitute for thinking about the problem.").

[44] *See, e.g.*, *In re* Breast Implant Litig., 11 F. Supp. 2d 1217, 1226–27 (D. Colo. 1998); Haggerty v. Upjohn Co., 950 F. Supp. 1160, 1164 (S.D. Fla. 1996), *aff'd*, 158 F.3d 588 (11th Cir. 1998) ("[S]cientifically valid cause and effect determinations depend on controlled clinical trials and epidemiological studies.").

[45] *See, e.g.*, Allison v. McGhee Med. Corp., 184 F.3d 1300, 1315 n.16 (11th Cir. 1999) (upholding exclusion of expert testimony based on epidemiological study with relative risk of 1.24); Cotroneo v. Shaw Envt'l & Infrastructure, Inc., No. H-05-1250, 2007 WL 3145791 (S.D. Tex. Oct. 25, 2007) (excluding testimony based on a relative risk of less than 2); Hall v. Baxter Healthcare, 947 F. Supp. 1387, 1403 (D. Or. 1996) (excluding testimony unless based on relative

the increase of risk in exposed versus unexposed populations. Relative risk statistically describes the measured strength of association between a disease and a risk factor.[46] A relative risk of one indicates that there was no increase in effect. Any increase above a relative risk of one indicates that there is some effect. The larger the relative risk, the stronger the effect. As epidemiologists have tried to explain to the courts, any increase in group risk from exposure to a chemical (that is, any relative risk greater than one) may be attributable to causation of the effect experienced by individuals within the group.[47]

Using a relative risk of two as a cutoff for admissibility misconstrues these principles. Some courts mistakenly reasoned that a relative risk of 2 (indicating a doubling of the risk) was required to meet the more probable than not standard for civil proof.[48] That reasoning mixes apples and oranges.[49] Equating legal and scientific standards is logically unsound.[50] As one prominent epidemiologist remarked, "It is possible that relative risks below 2 meet the criteria for causality, and it is commonplace for relative risks well above 2 to fail to do so."[51]

Relative risk is a statistical test that (like statistical significance) depends on the size of the population being tested. It is defined as the percentage of risk in the exposed population attributable to the agent under investigation. Increasingly, courts are beginning to acknowledge that any relative risk greater than 1.0 shows some increase of risk in the exposed population.[52] As long as there is a relative risk greater than 1.0, there is some association, and

risk of 2); Sanderson v. Int'l Flavors & Fragrances, 950 F. Supp. 981, 999–1000 (C.D. Cal. 1996) (excluding testimony based on studies with relative risk less than 2); Merrill Dow Pharm., Inc. v. Havner, 953 S.W.2d 706 (Tex. 1997) (conflating a doubling of the risk with the burden of proof necessary to establish causation and excluding as irrelevant testimony based on studies that did not meet this standard). For a discussion of this problematic view of relative risk, see Sander Greenland, *The Need for Critical Appraisal of Expert Witnesses in Epidemiology and Statistics*, 39 WAKE FOREST L. REV. 291, 294 (2004).

[46] *See* DAVID E. LILIENFELD & PAUL D. STOLLEY, FOUNDATIONS OF EPIDEMIOLOGY 200–02 (3d ed. 1994) (defining relative risk as the percent of risk in the exposed population attributable to the agent under investigation).

[47] *See, e.g.*, Greenland, *supra* note 46, at 294 (discussing judicial misinterpretation of relative risk).

[48] *See, e.g.*, Wells v. SmithKline Beecham Corp., No. A-06-CA-126-LY, 2009 WL 564303, at *6 (W.D. Tex. 2009) (requiring testimony based on studies with a relative risk > 2 in order to meet the "more likely than not" legal standard).

[49] *See* BRACKEN, *supra* note 1, at 250 (discussing the differences between legal and scientific goals).

[50] *See* BEECHER-MONAS, *supra* note 19, at 65–67 (explaining the logical fallacy of confounding legal and scientific standards).

[51] BRACKEN, *supra* note 1, at 250.

[52] *See* Greenland, *supra* note 46, at 294.

experts should be permitted to base their causal explanations on such studies.[53]

[redacted]*Confidence Intervals*

The judicial exclusionary approach to relative risk seems to have shifted to confidence intervals, but entailing even more confusion. Confidence intervals, like statistical significance and relative risk, tend to be used by courts as thresholds for admissibility. With some regularity, courts exclude expert epidemiology testimony if the confidence interval includes the relative risk of one.[54] (Recall that relative risk of one means that the null hypothesis of no effect cannot be rejected). This is a mistaken interpretation of confidence intervals.

A confidence interval is defined as a range of possible (relative risk) values at a given significance level (P-value).[55] A 95% confidence interval means that, over a vast number of repetitions, 95% of the intervals generated would contain the true association if the model were correct.[56] If the model used to compute the confidence interval is correct, the data and the model provide more support for data points inside the limits of the interval than outside.[57] A relative risk of one within the confidence interval does not mean there is no association, because confidence intervals include a range of values. If, for example, we have a 95% CI [1-10], the interval includes the relative risk of one, but it also includes the relative risk of ten.[58] Thus the relative risk is as likely to be ten as it is one.

Rather, the confidence interval limits indicate the values within which a certain percentage of all data are likely to fall.[59] The whole point of a confidence interval is to give a range of values that if a study were replicated many times, would include the correct value 95% (or whatever other

[53] *See* King v. Burlington N. Santa Fe Ry. Co., 762 N.W.2d 24, 46 (Neb. 2009) (declining to set a minimum threshold above 1.0 for relative risk because the studies "need not draw definitive conclusions on causation before experts can conclude that an agent can cause a disease"); s*ee, e.g.*, Lofton v. McNeil Consumer & Specialty Pharm., No. 05-CV-1531-L(BH), 2008 WL 4878066, at *3 (N.D. Tex. 2008) (finding admissible testimony based on several epidemiology studies, although none of them had found a relative risk of 2).

[54] *See, e.g.*, *In re* Viagra Prod. Liab. Litig., 572 F. Supp. 2d 1071, 1078–79 (D. Minn. 2008) (excluding testimony based on studies with confidence intervals that included 1).

[55] *See* Michael D. Green et al., *Reference Guide on Epidemiology*, *in* FED. JUDICIAL CTR., REFERENCE MANUAL ON SCIENTIFIC EVIDENCE 580 (3d ed. 2011).

[56] *See* Greenland & Poole, *supra* note 29, at 124 (explaining confidence intervals).

[57] *Id.* at 125.

[58] *See id.* at 123 (discussing common misinterpretations of confidence intervals).

[59] BEECHER-MONAS, *supra* note 19, at 61 (discussing statistical fallacies in the courts).

arbitrarily set level is chosen) of the time.[60] The confidence interval is a "general guide to the amount of random error in the data."[61]

The rationale courts often give for the categorical exclusion of studies with confidence intervals including the relative risk of one is that such studies lack statistical significance.[62] Well, yes and no. The problem here is the courts' use of a dichotomous meaning for statistical significance (significant or not).[63] This is not a correct understanding of statistical significance.

The higher the significance level (the lower the P-value), the more stringent the exclusion of possible random error, and the wider the confidence interval.[64] In other words, confidence intervals are supposed to inform the decisionmaker about relative risk through the width and location of the interval. In using confidence intervals as a surrogate for statistical significance, courts "ignore the potentially useful quantitative information

[60] *See* ROTHMAN, *supra* note 23, at 150. Rothman explains the concept:

> A given confidence interval is tied to an arbitrarily set level of confidence. Commonly, the level of confidence is set at 95% or 90%, although any level in the interval 0% to 100% is possible . . . [i]f the level of confidence is set to 95%, it means that if the data collection and analysis could be replicated many times, and the study were free of bias, the confidence interval would include within it the correct value of the measure 95% of the time . . . . It is better not to consider a confidence interval to be a literal measure of statistical variability, but rather a general guide to the amount of random error in the data.

Id.

[61] *Id.*

[62] *See, e.g.*, Pritchard v. Dow Agro Scis., 430 Fed. Appx. 102, 104 (3d Cir. 2011) (upholding exclusion of expert testimony where expert tried to explain that if a lower than 95% confidence interval—say, a 90% confidence interval—would not contain the relative risk of 1.0); Scharff v. Wyeth, No. 2:10-CV-220-WKW, 2011 WL 4361634, at *18 (M.D. Ala. Sept. 19, 2011) (finding that study with 95% confidence interval (limits 0.9-22.4) was insufficient to establish pharmaceutical defendant's knowledge of breast cancer risk because it included the relative risk 1.0); Tumlinson v. Advanced Micro Devices, Inc., No. 08C-07-106, 2012 WL 1415777, at *3 (Del. Super. Ct. Jan. 6, 2012) (excluding testimony based on studies with 95% confidence intervals with limits including relative risk 1.0); Faust v. BSNF Ry. Co., 337 S.W.3d 325, 337 (Tex. Ct. App. 2011) ("To be considered reliable scientific evidence of general causation, an epidemiology study must (1) have a relative risk of 2.0 and (2) be statistically significant at the 95% confidence level.").

[63] *See, e.g.*, ROTHMAN, *supra* note 23, at 164 ("For those who inappropriately place emphasis on whether a confidence interval contains the null value (thereby converting the confidence interval into a statistical test) . . . placing emphasis on the exact location of a confidence interval, equivalent to placing emphasis on statistical significance, is an inappropriate and potentially misleading way to interpret data.").

[64] *Id.* at 150.

that the confidence interval provides."[65] Failure to understand that the confidence interval provides information through the width and location of the interval rather than as a cut-off for scientific validity is both inappropriate and misleading.[66]

A striking illustration of the utility of confidence intervals is provided to judges in the Federal Judicial Center's Reference Manual on Scientific Evidence.[67] There, the authors provide a graph demonstrating how the confidence interval changes with P-value, so that while a confidence interval with P <0.05 (that is, a 95% confidence interval, corresponding to a 95% significance level) may include the relative risk of one, choosing a P-value <0.1 (that is, a 90% confidence interval) would include the values 1.1-2.2 and thus demonstrate some effect with confidence interval set at 90%.

In other words, there is nothing magic about a confidence interval including (or excluding) a relative risk of one—it depends on the P-value (or statistical significance level) chosen.[68] So, the courts' frequent quotation that "where a confidence interval contains a relative risk of 1.0, the results of the study are not statistically significant"[69] means only that at a P-value of 0.05 (or 95% significance level, or confidence interval), no effect was one of the values included. Other values will be included, depending on the width of the confidence interval. At a lower (say 90%) confidence interval, the interval may include only relative risks greater than one. Interpreting the confidence interval is not a dichotomous choice, any more than is statistical significance. It is a description of the study, not a dichotomous signal.

The complexity of causation assessments makes the urge to simplify the evaluation process through cursory checklists and rules of thumb hard to resist. Yet insisting on using statistical significance, relative risk and confidence intervals as screening devices simply makes no sense. It is a complete mistranslation of the science behind expert testimony. Rather than

[65] *See id.* at 147 (using as an example the erroneous interpretation of a flutamide study in which the researchers claimed there was no effect on prostate cancer, contradicting the previous 10 studies that had shown a modest benefit).

[66] *Id.* at 164 (noting that such an interpretation of confidence intervals is "an inappropriate and potentially misleading way to interpret data.").

[67] *See* Green et al., *supra* note 56, at 580, fig.4 (discussing the use of confidence intervals).

[68] For example, in a case control study of the risks of congenital heart disease when mothers took chlordiazepoxide in early pregnancy, the researchers concluded that the results, showing 95%CI(1-10.5), meant that there was no effect. In re-analyzing the data, however, Rothman found that their conclusion unwarranted. *See* ROTHMAN, *supra* note 23, at 154. That is because the interval, which includes the relative risk of one, also includes the relative risk of 10.5 with the same compatibility, and there is no reason to prefer rr=1 over rr=10.5. Moreover, the same data re-analyzed at 90%CI(>1<10), so that at 90% significance level, it is clear that there is some effect on congenital heart disease. *Id.*

[69] Green et al., *supra* note 56, at 621.

exclude studies that fail to meet these court-imposed thresholds, courts should admit and allow the adversary process to inform the jury about the strengths and weaknesses of these studies.

[redacted] Statistics in Context: The Problem of Atomistic Admissibility

Although statistics is fundamental to epidemiology, epidemiology is more than just applied statistics. Causal assessments do not exist in a vacuum. Statistical tests may tell us less about the data than other types of data analysis.[70] They tell us something about how to design studies, and what those studies mean. Statistical results rarely imply high certainty for a hypothesis.[71] Multiple studies showing small but consistent effect may increase the certainty.[72] Knowing the mechanism of action also increases the certainty.[73] But where the mechanism of action is unknown, or the observed risk level small, a wide range of hypotheses, including the null and alternative hypotheses, are possible.[74]

This makes biological plausibility assessments vitally important, and such assessments require many studies, and many kinds of studies. Despite judges' familiarity with the concept of separate pieces of circumstantial evidence building on each other to create a whole picture (a single brick does not build a wall), they appear to have difficulty with this concept when it comes to evaluating scientific evidence.

Courts have been remarkably reluctant to view the studies on which experts base their testimony as parts of a coherent whole. In assessing causation, epidemiologists look well beyond a single study, to see how its results compare to the results in similar studies, and to "the entire body of research on the topic."[75] That would include toxicology studies (involving animal studies and in vitro studies) as well as chemical function tests.

But many judges have reflexively required epidemiology studies to demonstrate causation; animal studies, cell studies and chemical structure

[70] *See* W. Douglas Thompson, *Statistical Criteria in the Interpretation of Epidemiologic Data*, 77 AM. J. PUB. HEALTH 191, 191–94 (1987) (discussing data analysis).

[71] *See* Greenland & Poole, *supra* note 29, at 128–29 (explaining that even where there exists a large and consistent body of epidemiologic evidence there is a large degree of uncertainty).

[72] *See* BEECHER-MONAS, *supra* note 19, at 68 (discussing the importance of multiple studies in causation assessments).

[73] *See id.* (giving examples of infections diseases and medical device failures).

[74] *Id.*

[75] BRACKEN, *supra* note 1, at 128.

studies are often rejected as irrelevant, or treated as beside the point.[76] This has made the more scientifically accurate kind of causation argument, based on multiple lines of study, frequently impossible to advance in court. Although many courts expressly stated that epidemiology evidence is not essential to proving causation, they have then excluded testimony based on studies other than epidemiology.[77]

Instead, using the template set out by the Supreme Court in *General Electric Co. v. Joiner*,[78] courts have exhibited a misguided tendency to separate the studies relied on by plaintiffs' experts, requiring that each study relied upon wholly support the expert's causation opinion, rather than examining whether the studies as a whole provide support for the general causation opinion.[79] The tendency of lower courts to follow the analytic

[76] *See, e.g.*, Hollander v. Sandoz Pharm. Corp., 95 F. Supp. 2d 1230, 1238 (W.D. Okla. 2000), *aff'd*, 289 F.3d 1193 (10th Cir. 2002) (rejecting the relevance of animal studies to causation arguments).

[77] *See* Rider v. Sandoz Pharm. Corp., 295 F.3d 1194, 1198–99, 1203 (11th Cir. 2002) (citing Eighth, Tenth and Eleventh Circuit opinions holding that epidemiology was not required as a basis for causation testimony, but then excluding causation testimony that was not based on epidemiological evidence); *In re* Meridia Products Liab. Litig., 328 F. Supp. 2d 791, 800–01 (N.D. Ohio 2004).

[78] 522 U.S. 136 (1997).

[79] While each study should indeed be examined, it is the evidence as a whole that must be examined for its cumulative force. *See, e.g.*, BEECHER-MONAS, *supra* note 19, at 47–49 (discussing the importance of assessing the cumulative impact of studies used to support a biological theory of causation); *see also* CARL F. CRANOR, TOXIC TORTS: SCIENCE, LAW, AND THE POSSIBILITY OF JUSTICE 138–40 (2006) (arguing that the Supreme Court made a mistake in *Joiner* by affirming the trial court's piecemeal analysis rather than undertaking a cumulative analysis).

template of the Supreme Court[80] is understandable, but it is not good analysis, and the result has been a very unscientific approach to causation evidence.[81]

Because individual studies almost never sufficiently support a complex determination like causation, courts' use of this kind of atomistic approach means that proving causation in many toxic tort cases is nearly impossible. First, it is highly unlikely that an epidemiology study with the precise parameters of a tort plaintiff's exposure has ever been done, so it is unlikely that the kind of evidence judges prefer even exists. Second, multiple lines of inquiry in different disciplines reinforce the biological explanation of the links along the causal pathway between exposure and effect.

[80] The Supreme Court in *Joiner* upheld the trial court's exclusion of expert testimony proffered by an electrician who claimed that his lung cancer had been caused by the polychlorinated biphenyls (PCBs) contaminating the dielectric fluid used as a coolant in the transformers that the plaintiff often had to repair. 522 U.S. at 139–40. The plaintiff argued that his exposure to PCBs promoted the lung cancer to which he was prone to by virtue of genetics and through his own history of smoking. *Id.* Although PCBs were known to cause cancer, and Congress had banned their production or sale since 1978, the district court excluded the plaintiff's expert causation testimony because it "did not rise above [the level of] 'subjective belief or unsupported speculation.'" *Id.* (quoting Joiner v. Elec. Co., 864 F. Supp. 1310, 1326 (N.D. Ga. 1994)). The Court of Appeals reversed, finding that the district court had substituted its conclusions for those of the experts rather than determining the legal reliability of the testimony. *Id.* at 140. (citation omitted). The Supreme Court disagreed. *Id.* at 141. Examining each of the studies the experts relied on, the Court deemed each of the studies insufficient, standing alone, to support causation. *Id.* at 144–46. First, it upheld the exclusion of the animal studies, because they had used infant mice, exposed at high doses, and had developed a different form of cancer. *Id.* at 144. Then, the Court turned to the epidemiology studies, and found each of them flawed as support. *Id.* at 145. The first study involved Italian workers who developed cancer from exposure to PCBs in a capacitor plant, but it concluded only that there was an association between exposure and the disease. *Id.* The second epidemiology study found an increase in lung cancer deaths of workers manufacturing PCBs, but not a statistically significant one. *Id.* The third study did show a statistically significant increase of lung cancer deaths, but the workers were exposed to mineral oil, rather than the transformer oil that Joiner was exposed to. *Id.* at 145–46. The fourth study also found a statistically significant association between exposure to PCBs and lung cancer, but the subjects of the study had also been exposed to toxic rice oil. *Id.* at 146. The Court held that "it was within the District Court's discretion to conclude that the studies upon which the experts relied were not sufficient, whether individually or in combination, to support their conclusions," and so reversed and remanded. *Id.* at 146–47. Thus, while the Court appears to be saying that it would have found no problem had the trial court evaluated the studies as a whole, neither the trial court nor the Supreme Court did so. Indeed, the Supreme Court went out of its way to analyze the studies one by one to find each insufficient to support the expert's entire argument rather than examining each as supporting a small part of the overall argument. This has resulted in the overwhelming majority of courts evaluating the expert evidence seriatim rather than as a whole.

[81] *See* Bernard D. Goldstein, *Toxic Torts: The Devil Is in the Dose*, 16 J.L. & POL'Y 551, 573–74 (2008) (noting the courts' failure to understand the importance of animal toxicity studies, the courts' over reliance on "rules of thumb" like relative risk of two, and statistical significance at the 95% level, as well as on epidemiology studies).

Excluding information from any source that contributes to this explanation is myopic. This approach results in courts scrutinizing the evidence for imperfections while simultaneously ignoring basic reasoning.[82] Rather than adopting the interdisciplinary systems approach of the scientific community when evaluating evidence, courts using a piecemeal form of analysis became overly reductionist.[83]

This piecemeal approach to scientific validity is rather surprising because judges are used to assessing circumstantial evidence as parts of a whole in areas outside of scientific evidence. As Susan Haack explains, "this epistemological question is really quite general, arising in virtually every area of inquiry."[84] Why courts think that biological causation is any different in this regard remains a mystery.

Courts need to be aware that the evidence they consider most relevant—human epidemiology studies—may simply not exist, or may be flawed in various ways, making them less than perfectly analogous to the plaintiff's situation.[85] But even more important is the recognition that no single piece of evidence about toxicity is ever likely to support a complex causation opinion.[86] The best scientific reasoning is based on multiple lines of evidence, integrated into a theoretical whole.

[82] *See, e.g.*, Greenland, *supra* note 46, at 297–300 (giving examples of courts' failures to recognize experts' logical and statistical flaws).

[83] *See* Goldstein, *supra* note 82, at 569 (castigating courts for their "simple uni-dimensional solutions for toxic tort issues which increasingly exclude modern scientific reasoning" built on interdisciplinarity).

[84] Susan Haack, *Proving Causation: The Holism of Warrant and the Atomism of* Daubert, 4 J. HEALTH & BIOMEDICAL L. 253, 263 (2008) ("[R]eliance on a whole mesh of evidence is ubiquitous—the rule, not the exception.").

[85] *See, e.g.*, McCarrell v. Hoffman-La Roche, Inc., No. A-3280-07T1, 2009 N.J. Super Unpub. LEXIS 558, at *9, *41 (N.J. Super. Ct. App. Div. Mar. 12, 2009) (holding expert testimony based in part on animal studies admissible).

[86] *See, e.g.*, Hyman & Armstrong, P.S.C. v. Gunderson, 279 S.W.3d 93, 104 (Ky. 2009) (quoting Globetti v. Sandoz Pharm. Corp., 111 F. Supp. 2d 1174, 1180 (N.D. Ala. 2000)) (upholding admissibility in a parlodel case of causation testimony based on animal studies, case reports, chemical structure and activity, because "[s]cience, like many other human endeavors, draws conclusions from circumstantial evidence when other, better forms of evidence is not available"). For example, in In re *Welding Fume Products Liability Litigation*, No. 1:03-CV-17000, MDL 1535, 2005 WL 1868046, at *33 (N.D. Ohio Aug. 8, 2005), when faced with assessing the reliability of plaintiffs' general causation testimony, the court, acknowledging that "none of the 'epi-studies' presented by either side is perfect," explained that these studies had to be examined in conjunction with other evidence of biological plausibility, including animal studies, case reports and case series. Similarly, the Kentucky Supreme Court's approach in *Gunderson* was a holistic one. *See* 279 S.W.3d at 102. Arguing that case reports, animal studies and chemical analogies were "merely anecdotal" and that the only reliable method of proving causation was through an epidemiology study, defendants appealed a jury verdict. *Id.* There were two epidemiology studies of post-partum women taking Parlodel, both commissioned by the defendant pharmaceutical company Sandoz, but both had significant flaws. In one study, the

[redacted]Specific Causation

Specific causation—that the plaintiff's injury was caused by exposure to the defendant's chemical—is almost universally required in medical causation cases.[87] Moreover, heeding the caution in *Kumho Tire* about expert *ipse dixit*,[88] the overwhelming majority of courts require medical experts to engage in a process the courts call "differential etiology."[89] The courts describe this process as "ruling in" all the possible causes of the injury, and then "ruling out" all but the chemical or drug to which the plaintiff was exposed.[90] Notably, this is not the same as differential diagnosis, in which doctors identify the illness from which the patient is suffering in order to provide treatment.[91]

While differential etiology may sound logical to lawyers and judges, this exercise simply is not part of medical (as opposed to courtroom) practice.[92] First, "ruling in" all potential causes cannot be done. Nearly every disease is caused by both genetics and environment. We simply do not know enough about either to decide what should be ruled in. Second, ruling out all but one cause is not feasible either. Epidemiologists now speak in terms of causal

patients were not trackable, and there were apparent misclassifications within the study. In the other, commissioned by the manufacturer to allay FDA concerns about adverse reaction reports, the data had been manipulated to lower it from a statistically significant risk rate of 2.86 to a statistically insignificant risk rate of 1.61 by excluding women with a history of seizures and women who had also taken a related drug, regardless of whether that other drug was in the woman's system at the time of the adverse event. *Id.* A damaging memo in Sandoz's file requested the researchers in the study to '"recut the data on late-onset seizures"' in the final report. *Id.* The study had been rejected as misleading by three peer-reviewed publications. *Id.* No other epidemiology study had been done. *Id.* That left the plaintiffs with case reports, including challenge-dechallenge-rechallenge case reports, animal studies using dogs and rats, and chemical analogy studies with ergot alkaloids (the class of drug of which Parlodel is a member). *Id.* at 103. Although the court acknowledged that no piece of this evidence standing alone could prove causation, "together they tend to show that Parlodel can cause postpartum seizures in women taking the drug for [postpartum lactation suppression]." *Id.* at 106.

[87] *See, e.g.*, Green et al., *supra* note 56, at 609 (noting that "the specific causation issue is a necessary legal element in a toxic substance case").

[88] Kumho Tire v. Carmichael, 526 U.S. 137 (1999).

[89] *See, e.g.*, Hendrix *ex rel.* G.P. v. Evenflo Co., 609 F.3d 1183, 1195 (11th Cir. 2010) (citation omitted) ("[D]ifferential etiology is a medical process of elimination whereby the possible causes of a condition are considered and ruled out one-by-one, leaving only one cause remaining."). Notably, this is not the same as differential diagnosis, in which doctors identify the illness from which the patient is suffering in order to provide treatment. *Id.* at 1194 n.4.

[90] *See, e.g.*, *id.* at 1198 n.5.

[91] *See, e.g.*, *id.* at 1194 n.4.

[92] *See* Faigman et al., *supra* note 2, at 420 (observing that making group to individual inferences is "rarely a focus of the basic scientific enterprise").

pies, where a number of causes may act in concert.[93] Every sufficient cause involves the joint action of a multitude of component causes.[94] Theorizing about the causal connection and testing the theories with data is the most that science can achieve.

Moreover, applying a causal inference from epidemiologic data to a specific person is circular reasoning that "defeats the validity of the epidemiologic process."[95] Disease should be defined on the basis of certain criteria (signs and symptoms) rather than exposure.[96] Running tests, observing signs and symptoms, are the basis of the diagnostic process, which distinguishes those individuals with a specific disease from those without it.

Judicial failure to understand the inability of medical scientists to pinpoint individual causation is, in part, another translation problem. When we say that a patient's symptoms of fever, chills, general weakness, or whatever are caused by kidney disease, what we mean is that these symptoms are the effects of the kidney disease. Certainly doctors are trained and experienced in linking symptoms with a particular disease. What they are not trained to do is determine the cause of the kidney disease. The most they can say is that certain risk factors have been identified, and perhaps inquire whether the individual patient was exposed to any of these risks.

In other words, what medical doctors do is decide what illness the patient has in order to treat it. Even this process is often irretrievably subjective. It involves intuition, guesswork and judgment. The advent of evidence-based medicine in recent years, with its emphasis on population statistics, has increased the objectivity of diagnostic decision making.[97] But this is a process of deciding what disease the individual is suffering from and how best to treat it. The idea is that integrating clinical expertise with the best available data will provide a "helpful framework for providers navigating the uncertainty inherent in patient care."[98] Nothing in this process requires (or permits) the

[93] *See* ROTHMAN, *supra* note 23, at 24 ("[E]very causal mechanism [sometimes called a sufficient cause] involves the joint action of a multitude of component causes.").

[94] Even infectious diseases require at least two component causes: exposure to a pathogen and lack of immunity. *Id.* at 250.

[95] *Id.*

[96] *See id.* at 250 (remarking that although disease criteria should have nothing to do with exposure, there is a long list of diseases where this is common, and citing the example of analgesic nephropathy, which is defined as kidney failure induced by analgesic drugs, making it impossible to evaluate the relationship of analgisics to kidney failure).

[97] *See* David L. Sackett et al., *Evidence Based Medicine: What It Is and What It Isn't: It's About Integrating Individual Clinical Expertise and the Best External Evidence*, 312 BRIT. MED. J. 71, 71 (1996) (characterizing evidence based medicine as the "conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients").

[98] MARK B. MCCLELLAN ET AL., EVIDENCE-BASED MEDICINE AND THE CHANGING NATURE OF HEALTH CARE: 2007 IOM ANNUAL MEETING SUMMARY, at v (2008).

medical doctor to ascribe individual causation. Treatment in many cases is applied using the physician's experience, but it can also be trial-and-error. For example, depression is treated with anti-depressants, some of which are not effective in some patients, and many patients require multiple drugs, but only in certain circumstances.

Statistical inference from the general to the specific is simply not something science can do.[99] Statistical inference is a useful tool for describing average effects on populations (of people exposed to chemicals, for example). But it cannot tell us whether a particular individual within that population suffered the effect in question from exposure or for some other reason (genetic or environmental).[100]

This is no denigration of the medical profession. Doctors are healers, and modern medicine has made incredible advances in saving human lives and keeping humans healthy. Many once-fatal diseases are now merely an inconvenience, thanks to treatments the medical profession has devised. But, like most experts, testifying doctors have a tendency to expound on matters in which they have neither training nor experience, and specific causation is one of those matters.

Courts have long been familiar with medical experts testifying to a "reasonable degree of medical certainty," and in the early years of *Daubert*-imposed gatekeeping, courts rarely excluded testimony based on differential diagnosis. This, however, created opportunities for abuse, and in a few mass torts, such as asbestos, silicosis, and fen-phen, highly suspect mass medical screenings by a few doctors for litigation rather than treatment purposes were admitted as specific causation testimony, often without challenge.[101] In a stinging, though advisory, opinion, Judge Janis Jack brought this problem to the fore in a multi-district litigation involving 10,000 silicosis claims that

[99] *See* A.P. Dawid, *Causal Inference Without Counterfactuals*, 95 J. AM. STAT. ASS'N 407, 408 (2000) (arguing the importance of making the distinction between the causes of effects and the effects of causes); A.P. Dawid, *The Role of Statistical and Scientific Evidence in Assessing Causality*, *in* PERSPECTIVES ON CAUSATION 133–47 (Richard Goldberg ed., 2011) (noting the importance of making this distinction).

[100] *See* Dawid et al., *supra* note 3, at 369 (2014) ("[T]he imperative to admit testimony about whether the specific effect is an instance of some general effects is more the product of the demands of the law than the power of the science.").

[101] *See* Lester Brickman, *The Use of Litigation Screenings in Mass Torts: A Formula for Fraud?*, 61 SMU L. REV. 1221, 1226, 1229 (2008) (discussing problems posed to the justice system by "doctors who are willing to mass produce mostly unreliable and arguably fraudulent diagnoses" and "criminal and civil justice systems [which] appear largely incapable of detecting or deterring, let alone sanctioning, [their] actions.").

were generated through mass screenings.[102] This opinion appears to have raised judicial consciousness regarding the problem of mass screenings, and to the subjective nature of differential diagnosis in general.

Nonetheless, courts continue to require medical testimony of specific causation.[103] When medical experts are honest enough to acknowledge that many diseases—like cancer—are mostly idiopathic (that is, of unknown cause and occurring with some baseline frequency in the general population), courts tend to exclude their testimony even where there is valid epidemiologic testimony establishing a causal link to exposure in humans.[104] This is yet another problem of statistical mistranslation.

As discussed above, causation in cancer (and probably many other diseases as well), tends to have multiple causes.[105] Some of these factors may be unknown (or idiopathic, in medical terminology)—unidentified genes or environmental exposures. Lung cancer, for example, may appear in the general population without any known toxic exposure. Nonsmokers get lung cancer.

The lung cancer risk of nonsmokers not exposed to asbestos is one in one hundred thousand.[106] By definition that is the relative risk of one. That is the background rate, the rate due to other, mostly unidentified, causes. The incidence of lung cancer in the population is greatly increased by smoking (rr=10).[107] So one can say that smoking is a strong risk factor for lung cancer.[108]

For any individual case of lung cancer, however, smoking is no more important than any of the other component causes, some of which may be

[102] *In re* Silica Products Liab. Litig., 398 F. Supp. 2d 563, 634 (S.D. Tex. 2005) ("[A]ssembly line diagnosing . . . is an ingenious method of grossly inflating the number of positive diagnoses.").

[103] *See, e.g.*, Dawid et al., *supra* note 3, at 366 ("[C]ourts' expectations regarding the specificity of the proof are dictated by legal standards, rather than the ability of scientists to provide the necessary proof.").

[104] *See, e.g.*, Milward v. Acuity Specialty Prods. Grp., Inc., 969 F. Supp. 2d 101, 109 (D. Mass. 2013) (excluding expert testimony on specific causation because of the "high probability that a cause cannot be identified."). A Westlaw search of the all states database for the terms "specific causation" and "idiopathic" revealed 53 cases in which the courts rejected expert medical testimony of specific causation because the expert could not rule out idiopathic causes. In these cases, the courts have misunderstood the concept of biological causation, which very often has multiple causes, some of which are unknown. *See infra* notes 106–10 and accompanying text.

[105] *See* ROTHMAN, *supra* note 23 at 25 (noting that nearly every causal mechanism involves some environmental and some genetic factors).

[106] *Id.* at 200 (discussing lung cancer risk).

[107] *Id.*

[108] *Id.(*explaining that smoking is a strong cause of lung cancer because it causes a large proportion of the cases).

unknown. As Rothman explains, "[w]ith respect to an individual case of disease . . . every component cause that played a role was necessary to the occurrence of that case."[109] From the individual perspective, there is no distinction between strong and weak causes; either something is a factor in the resulting disease or it isn't.[110] And determining which identically manifested disease was caused by which factor is beyond the capacity of medical science.[111] It simply is not possible to infer logically whether a specific factor in a causal pie was the cause of an observed event.[112]

The reason epidemiologists engage in relative risk analysis is to determine whether exposure to a chemical or other pathogen had any increase in effect over the baseline. When we know that several factors may be involved, we can measure risk among people exposed to the known factors, and then measure risk among people exposed to one but not the others.[113] If there are unknown causes as well as known causes, the best we can do is determine the increased risk from exposure to the known causes in the population.[114] Looking backward from an individual case of lung cancer, in a person exposed to both asbestos and smoking, to try to determine the cause, we cannot separate which factor was primarily responsible. If all factors are

[109] *Id.* at 25.

[110] *Id.*

[111] *Id.* at 204 (noting that "there is no way to tell, by direct observation alone, which clas of causal mechanisms is responsible for an individual case").

[112] *Id.* at 203 (discussing the concept of causal pies). Rothman explains that both fair skin and exposure to ultraviolet light are risk factors for melanoma. *Id.* Both are causes, parts of the same causal pie. The parts of a causal pie may cause disease without any direct interaction, if they act through different mechanisms, or they may interact with each other and with background unidentified causes. *Id.* Each may interact with the unidentified causes. Alternatively, it could all be the interaction of both with unidentified causes. *Id.* There simply is no way to tell, by direct observation.

[113] *Id.* at 204 (discussing risk differences in smokers , people exposed to asbestos, and both to demonstrate that most of the risk among people with joint exposure to both smoking and asbestos is due to biological interactions between smoking and asbestos).

[114] For example, the court in *Soldo v. Sandoz Pharm. Corp.*, 244 F. Sup. 2d 434 (W.D. Pa. 2003), was particularly confused about causation, despite having three court-appointed experts to guide it (all of whom acknowledged that qualified experts could legitimately disagree on the issue). First the court rejected the plaintiffs' experts' general causation experts for failing to show that Parlodel could cause intracerebral hemorrhage. *Id.* There were no epidemiology studies on the chemical. Then the court conflated the idea of general causation with specific causation, finding that because a large proportion of intracerebral hemorrhage have no known cause, plaintiffs' specific causation experts would not be able to rule out unknown causes. *Id.* at 479–80. For an excellent discussion of *Soldo*, see Joe B. Cecil, *Construing Science in the Quest for "Ipse Dixit": a Comment on Sanders and Cohen,* 33 SETON HALL L. REV. 967, 971–86 (2003). Taking their cue from *Soldo*, a series of recent cases have excluded specific causation testimony for failure to consider idiopathic causes. *See* William E. Padgett, *Etiology Unknown: Using the Idiopathic Cause in Your Specific Causation Defense*, FOR DEFENSE, Jan. 2013, at 56 (discussing the idiopathic defense and citing cases where it has been successful).

present, all were the cause. We cannot determine which case of lung cancer in an individual exposed to smoking, asbestos and unknown factors arose from which factor.

Nothing in relative risk analysis, in statistical analysis, nor anything in medical training, permits an inference of specific causation in the individual case. No expert can tell whether a particular exposed individual's cancer was caused by unknown factors (was idiopathic), linked to a particular gene, or caused by the individual's chemical exposure. If all three are present, and general causation has been established for the chemical exposure, one can only infer that they all caused the disease.[115] Courts demanding that experts make a contrary inference, that one of the factors was the primary cause, are asking to be misled. Experts who have tried to point that out, however, have had a difficult time getting their testimony admitted.[116]

B. Statistical Misunderstandings in Criminal Cases

Judges in criminal cases also struggle to understand statistical inference drawing. There are two primary areas in which this is apparent, diagnostic cases involving medical diagnoses of intentionally inflicted trauma and pattern identification cases.

[redacted] Diagnostic Issues in Criminal Cases: the Saga of Shaken-Baby Syndrome

All the problems discussed above relating to specific causation testimony are inherent in criminal cases involving medical diagnoses.[117] Nowhere is this more apparent than in shaken baby cases, where reasoning from individual instance (of a baby with a set of symptoms) to group data about the prevalence of these symptoms in the general population has confounded the

[115] ROTHMAN, *supra* note 23, at 25 (explaining that "every component cause that played a role was necessary to the occurrence of that case").

[116] *See, e.g.*, Collins v. Ashland Inc., No. 06C-03-339 BEN, 2011 WL 5042330 (Del. Super. Ct. Oct. 21, 2011) (excluding expert specific causation testimony for failing to rule out idiopathic causes); Milward v. Acuity Specialty Prods. Grp., Inc., 969 F. Supp. 2d 101, 115 (D. Mass. 2013) (excluding expert specific causation testimony because "differential etiology is not possible here given the large percentage of idiopathic cases of AML"); Pritchard v. Dow Agro Scis., 705 F. Supp. 2d 471, 496 (W.D. Pa. 2010) (excluding expert specific causation testimony for failing to rule out idiopathic causes).

[117] *See, e.g.*, United States v. Grigsby, 712 F.3d 964, 968 (6th Cir. 2013) (involving testimony to a reasonable medical certainty that defendant had a mental illness at the time of the offense).

courts.[118] Part of the problem is that the group data were flawed and incomplete.[119]

Notwithstanding decades of medical testimony (usually from pediatricians specializing in child abuse) that a triad of symptoms (subdural hemorrhage, retinal hemorrhage and brain damage) was diagnostic of violent shaking, current research has demonstrated that many other factors also can cause these symptoms.[120] In addition, research into the biomechanics of shaking has cast doubt on the ability of even severe shaking to result in such symptoms.[121]

The claim that only inflicted trauma can cause the triad of symptoms is simply incorrect.[122] A myriad of events other than trauma have been shown to be associated with the triad.[123] Subdural hemorrhages are found in normal infants.[124] Retinal hemorrhages are found in roughly 30% of newborns.[125] Brain damage in most shaken baby/abusive head trauma cases has been shown to be hypoxic-ischemic (lack of oxygen) rather than traumatic

[118] For a description of courts' widespread failure to respond to evolving evidence in shaken baby cases, see generally Deborah Tuerkheimer, *Science-Dependent Prosecution and the Problem of Epistemic Contingency: A Study of Shaken Baby Syndrome*, 62 ALA. L. REV. 513 (2011).

[119] *See generally* Keith A. Findley et al., *Shaken Baby Syndrome, Abusive Head Trauma, and Actual Innocence: Getting it Right*, 12 HOUS. J. HEALTH L. & POL'Y 209 (2012) (discussing the validity of the studies supporting and undermining shaken baby syndrome).

[120] A.N. Guthkelch, *Problems of Infant Retino-Dural Hemorrhage with Minimal External Injury*, 12 HOUS. J. HEALTH L. & POL'Y 201, 202 (2012) ("Since subdural and retinal hemorrages (with or without cerebral edema) may also be observed in accidental or natural settings, I suggest that elements of the classic triad of retinal hemorrhage, subdural hematoma and cerebral edema would be better defined in terms of their medical features."); *see also* Findley et al., *supra* note 120, at 214 (the triad has been found in many conditions, such as falls, birth trauma, congenital malformations, genetic conditions, metabolic disorders, infections, and toxins, to name a few).

[121] Findley et al., *supra* note 120, at 214 (noting that biomechanical studies have consistently shown that shaking is not sufficient). Because the biomechanical studies cast doubt on shaking as being capable of producing the triad, child abuse pediatricians have switched the name from shaken baby syndrome to abusive head trauma. *Id.* at 220. The central tenets of the testimony have not changed, however, despite the name change, and experts continue to testify that the triad is caused by inflicted trauma. *Id.* at 220–21.

[122] *See, e.g.*, Waney Squier & Julie Mack, *The Neuropathology of Infant Subdural Haemorrhage*, 187 FORENSIC SCI. INT'L 6, 7 (2009) (emphasizing the complex neuropathology of the brain and its coverings).

[123] Guthkelch, *supra* note 121, at 204 (one cannot assume the triad is caused by trauma, rather than by natural causes).

[124] V.J. Rooks et al., *Prevalence and Evolution of Intracranial Hemorrhage in Asymptomatic Term Infants*, 29 AM. J. NEURORADIOLOGY 1082, 1085 (2008) (finding that 46% of infants with normal births have subdural hemorrhages).

[125] M. Vaughn Emerson et al., *Incidence and Rate of Disappearance of Retinal Hemorrhage in Newborns*, 108 OPHTHALMOLOGY 36, 38 (2001).

injury.[126] In other words, the brain damage is not—as claimed by countless experts—a symptom of trauma, but a result of oxygen deprivation.

That the triad is associated with a host of factors other than inflicted trauma does not mean, of course, that an abused baby would never show signs of the triad. What it means is that the symptoms relied on as dispositive of abuse could be symptomatic of many conditions, only one of which is abuse. Experts, however, claim that if there is no other reasonable explanation given for the infant's condition by its caregivers, the presence of the triad is diagnostic of child abuse.[127] This is faulty logic, however. Just because the triad of symptoms is found and no one can think of another explanation does not justify the inference of criminal intent.[128]

The fundamental problem here is the flawed statistical reasoning of the child abuse experts, and the widespread failure of courts to examine this reasoning.[129] Even if such experts had data on the prevalence of the triad of symptoms in the infant/toddler population (if, hypothetically, we knew what the base rate of the triad was in the infant population, and what percentage of those cases were attributable to risk factors other than trauma) an expert still could not validly—that is with any more scientific basis than a rational juror—draw an inference from that data to the cause of an individual baby's

[126] J.F. Geddes et al., *Neuropathology of Inflicted Head Injury in Children II. Microscopic Brain Brain Injury in Infants*, 124 BRAIN 1299, 1304 (2001) (concluding that axonal damage in children with inflicted head injury is "diffuse vascular or hypoxic-ischaemic injury, attributable to brain swelling and raised intracranial pressure."); *see* Mark S. Dias, *The Case for Shaking*, *in* CHILD ABUSE AND NEGLECT: DIAGNOSIS, TREATMENT AND EVIDENCE 362, 368 (Carole Jenny ed., 2011) (noting that neuroimaging studies and post-mortem analyses make it increasingly clear that the cerebral and axonal damage in abusive head trauma cases is hypoxic-ischemic rather than traumatic in origin); Neil Stoodley, *Non-accidental Head Injury in Children: Gathering the Evidence*, 360 LANCET 271, 272 (2002) (in the pathophysiology of non-accidental head injury, hypoxic-ischemic damage is more common than traumatic axonal or shearing injury).

[127] *See, e.g.*, Maze v. State, No. M2008-01837-CCA-R3-PC, 2010 WL 4324377, at *6 (Tenn. Crim. App. Nov. 2, 2010) (expert testimony that no explanation other than shaken baby syndrome accounted for infant's injuries).

[128] *See* Guthkelch, *supra* note 121, at 203–04 ("[I]nstances in which both medical science and the law have gone too far in hypothesizing and criminalizing alleged acts of violence in which the only evidence has been the presence of the classic triad or even just one or two of its elements.").

[129] The Supreme Court has not performed this reasoning well either, when it comes to shaken baby syndrome. *See* Cavazos v. Smith, 132 S. Ct. 2 (2011) (reversing the 9th Circuit's grant of habeas in a shaken baby case involving a grandmother convicted of murdering her 7-week-old grandson). The 9th Circuit had granted habeas on the basis of newly discovered medical evidence (that other factors than intentional trauma could cause the triad) and the Supreme Court reversed, holding that because there was some evidence of trauma to the infant's brain the 9th Circuit was "plainly wrong" that there was no evidence supporting the prosecution's theory. *Id.* at 6–8. Justices Ginsburg, Breyer and Sotomayor dissented, based on changes in the scientific understanding of shaken baby syndrome and the paucity of nonmedical evidence. *Id.* at 11 (Ginsburg, Breyer & Sotomayor, J., dissenting).

death. The best the expert could offer the factfinder is to explain the general data and leave it to the jury to draw the inference.

That is not at all what is happening in our courts. Instead, in hundreds of cases, the jury hears that the triad of symptoms does not appear absent abuse.[130] The only exceptions such experts will acknowledge are falls from buildings and car crashes.[131] If the symptoms are otherwise unexplained, the experts testify that the triad is diagnostic of abuse.

Although such testimony is wholly lacking in empirical support, it is rarely questioned by the courts. Not only do the many published studies on shaken baby syndrome (or abusive head trauma as it is now called) fail to provide evidence in support of the triad/abuse hypothesis, numerous studies show that abuse is far from the only factor that can cause it.[132] These errors would never have been perpetuated, however, had courts recognized the statistical flaw of asserting causation in the individual case.

[redacted]Absence of General Population Data

In most criminal identification evidence cases—apart from those involving DNA testimony—the statistical reasoning issues arise from an absence of data.[133] The forensic science disciplines uniformly depend on pattern recognition, and the typical testimony is that each pattern is unique to an individual, so when patterns are found to match, that excludes all other

[130] *See* Deborah Turkheimer, *The Next Innocence Project: Shaken Baby Syndrome and the Criminal Courts*, 87 WASH. U. L. REV. 1, 10 (2009) (estimating about 200 convictions per year).

[131] *See, e.g.*, State v. Mesa, No. 2 CA-CR 2011-0171-PR, 2011 WL 4379428, at *1 ¶ 3 (Ariz. Ct. App. Sept. 16, 2011) (expert testimony that the only explanations other than shaken baby syndrome for the infant's triad of symptoms were falls from a second or third story, or down a flight of stairs); Maze v. State, No. M2008-01837-CCA-R3-PC, 2010 WL 4324377, at *3–6 (Tenn. Crim. App. Nov. 2, 2010) (expert testimony that no explanation other than shaken baby syndrome could account for infant's injuries); State v. Fortener, No. E2008-01775-CCA-R3-CD, 2010 WL 1241629, at *4 (Tenn. Crim. App. Mar. 31, 2010) (expert testimony that nothing other than car crashes or violent shaking could account for the triad of injuries).

[132] *See* Findley et al., *supra* note 120, at 224–26, 264–96 (discussing the scientific literature on studies pro and contra the abusive head trauma hypothesis).

[133] *See* NAT'L RESEARCH COUNCIL OF THE NAT'L ACAD. OF SCI., STRENGTHENING FORENSIC SCIENCE IN THE UNITED STATES: A PATH FORWARD (2009) [hereinafter NAS REPORT] (finding that other than DNA testimony, criminal identification testimony is presented categorically, and without probability estimates).

individuals.[134] The assertion being made by these experts is that "nature never repeats."[135]

This assertion is based theoretically on the product rule. The product rule yields the joint probability of individual events by multiplying the separate probabilities of independent events.[136] Forensic scientists maintain that because there are many individual features of a particular pattern, the odds of repetition are vanishingly small. This assertion, however logical it may appear, is not actually based on statistics.

For the product rule to be statistically sound, it must be based on empirical data. As the National Academy of Sciences pointed out in its report on forensic sciences,[137] that is precisely what is missing from forensic science testimony other than DNA testimony. One would need to know, for example, that each feature of a given pattern is independent of the other features. Also, we would need to know the probability of each characteristic, that is, the frequency of its occurrence in the general population. Without knowing the frequencies of each characteristic in the general population, what is there to multiply? Even if we knew these factors (independence and frequency), uniqueness could not be established. Any factor greater than zero would yield a nonzero probability of another individual matching the pattern. However small the chance of repetition, it is not zero.[138]

It is important, therefore, to know the size of the pool of possible matches. That is precisely what is missing from forensic science testimony other than DNA. Criminal identification techniques such as fingerprint, bite-mark, knife-mark, microscopic hair analysis, voice spectrography, shoe print identification, handwriting analysis and ballistics all suffer from an absence of general population data.[139] In each of the identification techniques, crime

[134] *See* KEITH INMAN & NORAH RUDIN, PRINCIPLES AND PRACTICE OF CRIMINALISTICS: THE PROFESSION OF FORENSIC SCIENCE 123 (CRC Press L.L.C. 2001) (noting that the "belief that uniqueness is both attainable and existent is central to our work as forensic scientists").

[135] For a discussion of this uniqueness fallacy and its origins, see Michael J. Saks & Jonathan J. Koehler, *The Individuation Fallacy in Forensic Science Evidence*, 61 VAND. L. REV. 199, 204 (2008) (explaining the fallacy of uniqueness assertions made by forensic science experts).

[136] *Id.* at 207.

[137] *See* NAS REPORT, *supra* note 134, at 149 (observing that forensic science experts present their conclusions categorically and without reference to population data).

[138] See Saks & Koehler, *supra* note 136, at 207 for an excellent explanation of the product rule.

[139] Recognizing the absence of general data behind most forensic science, Congress in 2006 charged the National Academy of Sciences to examine the scientific validity of forensic science across all disciplines. *See* H.R. REP. NO. 109-272, at 121 (2005) (Conf. Rep.) (charging the NAS "to conduct a study . . . as described in the Senate Report."); S. REP. NO. 109-88, at 46 (2005) (noting the "absence of data" underlying the forensic sciences and listing eight charges to the NAS).

scene evidence is examined for indicia that "match" indicia belonging to the suspect.[140]

Pattern matching is the foundation for most of the forensic science disciplines. But without knowing how frequently such matches occur in the general population, one cannot know how significant it is that the markings appear to be similar.[141] Even if such data existed for these forensic techniques, the testimony would have to be given in terms of population estimates, as it is, for example, in DNA testimony. That, however, is not what happens in pattern-matching forensic science disciplines. Instead, the usual testimony is that because patterns match, they must have come from the same individual.

Arson testimony suffers from the same lack of general data. Although fire causation is somewhat better studied in terms of general data than the criminal identification techniques, it still suffers from the problem of lack of general knowledge about fires.[142] Fire experts examine signs and symptoms left in the wake of a fire in order to decide whether the fire came from "natural" causes or was set by an arsonist.[143]

Many of the signs and symptoms experts describe as dispositive of arson, such as that accelerants burn hotter than natural fires, have been disproved.[144] The problem here is that not enough is known about the signs and symptoms of "natural" fires, so that ascribing criminal responsibility is a matter of speculation. At least in this area, however, the expert association has been engaging in research to attempt to obtain this general information. Nonetheless, permitting experts to testify beyond what is known about the characteristics of fires and ascribing causation is scientifically unjustified.

[140] *See* Saks & Koehler, s*upra* note 136, at 204 (explaining the fallacy of uniqueness assertions made by forensic science experts).

[141] *See id.* at 209 ("[T]he proper application of the [product] rule requires a set of reliable frequency estimates for the relevant set of forensic characteristics[, and even then] . . . the product rule necessarily falls short of establishing unique individualization.").

[142] *See, e.g.*, NAT'L FIRE PROTECTION ASS'N, NFPA 921: GUIDE FOR FIRE AND EXPLOSION INVESTIGATIONS ¶¶ 5.1–5.1.2 (2011) [hereinafter NFPA] (noting the many complex variables and unpredictability of fire behavior).

[143] One of the most controversial cases of arson testimony, *Texas v. Willingham*, involved the defendant's conviction for setting his home on fire, thereby killing his three children. *See* Paul C. Giannelli, *Junk Science and the Execution of an Innocent Man*, 7 N.Y.U. J.L. & LIBERTY 221, 222–24 (2013) (discussing the testimony in the case and the subsequent controversies over the testimony).

[144] *See* NFPA*, supra* note 143, ¶ 6.2.2.2 ("Wood and gasoline burn at essentially the same flame temperature.").

[redacted] POSSIBLE SOLUTIONS

Despite thousands of articles on *Daubert*, and evaluating scientific evidence, very few have focused on this problem of statistical mistranslation. Even fewer have attempted to solve the problem of inferring individual causation from general population data, and its converse assumption that one can infer something about the general population from the individual. Three recent notable articles however, have attempted to propose solutions to this conundrum. Although each proposed solution is, in my view, incomplete, each offers an important perspective.

A. Probability of Causation?

In their article *Fitting Science into Legal Contexts: Assessing Effects of Causes or Causes of Effects*, Professors Fienberg, Faigman and Dawid seek to "begin to remedy" the disconnect between science as it is practiced and understood by scientists, and its legal use in the courtroom.[145] The authors frame the issue as one of perspective: science infers the effects of causes, while law infers the causes of effects.[146] Making this distinction, the authors assert, would help the courts to reason more logically about the statistically based testimony offered by experts.[147]

By this, the authors mean that courts have difficulty reasoning from group data to the individual case. As examples of where the courts go astray, the authors proffer "differential etiology" in medical causation cases,[148] pattern recognition in criminal identification cases,[149] the opinion rule, in which experts are permitted to testify from the individual effect to the general cause,[150] and class certification in employment discrimination cases.[151] In each of these instances, the authors posit that the law's "unrealistic

[145] Dawid et al., *supra* note 3, at 361.

[146] *Id.* at 382–83.

[147] *Id.* at 382 (arguing that "having proof of both the [effects of causes] and the [causes of effects]" is the "key challenge" of translation).

[148] *Id.* at 367 (contending that differential etiology "has little scientific grounding and, indeed, is as much art as science.").

[149] *Id.* at 369–70 (explaining that declaring a "match" from two patterns is invalid in the absence of group data—without knowing the frequency of the pattern in the general population, which—outside of DNA evidence—does not exist for most forensic sciences).

[150] *Id.* at 372 (noting that this is "contrary to basic precepts of the scientific method").

[151] *Id.* at 379–80 (discussing the plaintiffs' failed attempt to use general data about corporate culture to infer effects of discrimination in *Wal-Mart Stores, Inc. v. Dukes*, 131 S. Ct. 2541 (2011)).

expectations" about what scientific proof can offer makes for illogical decisions.[152]

The authors discuss two possible solutions: the odds ratio, and the probability of causation. Although the odds ratio at first blush appears to be a good solution, it turns out to be a blind alley. Despite being symmetrical between cause and effect, meaning it can be estimated from both prospective and retrospective studies (in contrast to relative risk, which measures only the effect of the cause),[153] the odds ratio is not a great solution, the authors acknowledge, because "as was clear in the work of Galton . . . the regression of *X* on *Y* is not obtainable from the regression of *Y* on *X*."[154] In other words, odds ratio does not solve the problem of making individual inferences from general data.

Instead, the authors' posit a "Probability of Causation"[155] that involves several pages of elegant equations and tables. In the end, however, the authors are forced to acknowledge that "even if we start with the best possible information (perfect experimental results) about the [effects of causes], and use all relevant auxiliary information, we need to apply subtle logic to make inferences about the [causes of effects] (which will still, necessarily, remain imprecisely determined)."[156] The problem with this solution is that while those factors might exist in concert in a thought experiment, they are hard to come by in our less than perfect world.

Each of the factors is flawed. Few legal cases have perfect experiments that the expert can rely on. How "all relevant auxiliary" information could be known or available is an unanswered question. Asking judges to engage in subtle logic in a topic for which they were not trained seems rather preposterous. And few judges will permit testimony that is "imprecisely determined" because it sounds far too much like speculation.

In scientific practice, "perfect experimental results" rarely are obtained and are almost never available for courtroom testimony. The real problem for scientific experts and the lawyers that employ them is how to make sense of the imperfect studies that do exist. Relevant auxiliary information may be unknown, or unavailable for courtroom presentation and possibly not even guessed-at.

[152] *Id.* at 381.

[153] *Id.* at 375 ("[The odds ratio] can be estimated both from prospective studies where we 'control' or condition on *X* [cause] and then examine *Y* [outcome], and from retrospective studies, where we control or condition on *Y* and then observe *X*.").

[154] *See id.* (noting that while relative risk is often used as a measure of the effect of a cause, "its application to assessing the cause of an effect is more problematic").

[155] *Id.* at 377.

[156] *Id.* at 377–78.

As for subtle logic, some jurists are better at this than others, but almost everyone could use some guidelines as to what this means and how to employ it. The reason that many judges use bright-line exclusionary rules (requiring a relative risk of two for admissibility or rejecting confidence intervals that include a relative risk of one, statistical significance at p-level 0.05) is not because they have thought through the meaning of the concept in the particular case, but because a bright-line rule can be applied reflexively (if mistakenly).[157]

The authors suggest that a statistician be permitted to testify about the factors that go into the probability of causation, without actually saying what that probability is—something like eyewitness experts, who identify factors that may affect the accuracy of eyewitness testimony, without actually opining on the particular eyewitness's credibility. Similarly, a statistician could set up one of the authors' tables, explain the factors going into each box, and the formula to assess the evidence, and let the jury have at it (with proper instructions about the "subtle reasoning" that is required to make the leap from general to individual causation).

Perhaps so. If what the best scientists—including statisticians—validly can offer the jury is testimony about general causation, and if indeed there is "little scientific evidence to bear" on the question of individual causation, the courts should not be demanding this kind of testimony from them.

B. Best Practice Guidelines?

Professors Faigman, Monahan and Slobogin offer a slightly different approach to solving this problem of inferring individual causation from general data in their article, *Group to Individual (G2i) Inference in Scientific Expert Testimony*.[158] Noting that inference from group data to the individual is "rarely a focus of the basic scientific enterprise,"[159] the authors seek to offer courts a set of guidelines to help them accomplish this. Acknowledging that neither *Frye* nor *Daubert* addresses this issue, they nonetheless look to the *Daubert* trilogy in crafting their guidelines.[160]

First, the authors maintain that the standards applied to what they call framework testimony (and I've referred to as general causation or population data) ought to be applied differently to diagnostic testimony (specific

[157] For further discussion of these imperfect factors, and how judges can decide their helpfulness to the jury, see generally BEECHER-MONAS, *supra* note 19.

[158] Faigman et al., *supra* note 2, at 417–18.

[159] *Id.* at 420.

[160] *Id.* at 439–40.

causation or individual data).[161] The authors explain that while diagnostic testimony inferring individual causation may have "legal fit" in that it is relevant to an issue in the case (causation), it lacks "empirical fit" (whether the basis for the opinion can generalize to disputed legal issues).[162] Rather than categorically admitting or excluding diagnostic testimony, the authors suggest a more nuanced approach.

Rather than jettison specific causation testimony entirely, the authors instead contend that judges should examine five factors: materiality (which they describe as legal and empirical fit); witness qualifications; internal validity (comprised of testing, standards, error rates, peer review, general acceptance, type of rigorous analysis expected in the field); helpfulness (whether the testimony provides useful information for the factfinder); and avoidance of testimony misleading or distracting the factfinder.[163]

Distinguishing between diagnostic testimony that identifies the illness from diagnostic testimony that identifies the cause of the illness, the authors maintain that the ability of diagnostic experts to gather information (by, for example, conducting various tests) is a benefit to the factfinder and therefore should not be categorically excluded, but run through the five factor test.[164] Applying their five factors to eyewitness expert testimony, , one would think that specific causation testimony would always fail the first factor. The authors hedge their bets, however, by maintaining that proficiency testing may substitute for empirical fit, and that courts need to look at how peer experts function.[165] In their eyewitness expert example, specific causation testimony would fail because (outside the courtroom) such experts do not assess whether a particular eye witness was misled.[166]

The second factor, qualifications, could not be met in the eyewitness expert example, because nothing in such experts' clinical training permits an individual assessment.[167] Subjecting eyewitness diagnostic testimony to

[161] *Id.* at 480 ("Differential application of the five factors . . . is crucial if courts are to succeed at balancing the numerous legal and scientific considerations that influence when general research may be heard in court and the extent to which experts may apply that research to help resolve specific cases.").

[162] *Id.* at 440–44.

[163] *Id.*

[164] *Id.* at 446.

[165] *Id.* at 451 (arguing that courts should use the *Frye* standard for determining the empirical fit of diagnostic testimony and *Daubert* for framework testimony).

[166] *Id.* at 466 ("[D]iagnostic testimony that reaches beyond the expert's customary practice may well be irrelevant.").

[167] *Id.* at 432. The authors (mistakenly, in my opinion) explicitly distinguish medical doctors, whom, they assert, have clinical training in making individual assessments. *Id.* at 447. But these individual assessments are made for treating the patient, not for determining causation.

stringent internal validity testing—the third factor—is problematic because there is no protocol for individual causation assessments (outside the courtroom), and no accuracy feedback loop. As for the helpfulness (value-added) prong, the authors contend that since the framework testimony (based on the general data for eyewitness identifications) is strong, the jury can draw its own conclusions, and therefore the diagnostic testimony is not helpful.[168] The fifth factor is whether the testimony avoids misleading or distracting the jury, and here the authors contend that diagnostic testimony "could distract the jury from the scientific uncertainty inherent in the field" and therefore "probably is" more prejudicial than probative.[169] The authors note that "diagnostic evidence is most likely to be prejudicial the further it departs from what framework scientists [general causation experts] can say with a high degree of certainty."[170] Here, however, the authors maintain that the adversary process will usually be able to limit the dangers.[171]

In applying these factors, the authors say that if the proposed testimony fails any of the five factors, it should be excluded.[172] Once the testimony passes the five-factor threshold test, then any weakness in one factor can be met by strength in another. And, the authors maintain, framework testimony [general causation] should be balanced differently from "particularized diagnostic evidence."[173]

I agree with the authors on many points. They have identified a major problem in expert testimony. What they refer to as the G2i problem is indeed endemic to our courts. This problem afflicts specific causation testimony in toxic torts, individual identification in criminal cases, as well as individual applications of general framework testimony such as eyewitness testimony. I also agree with their conclusion that unless this problem can be solved,

Doctors are trained to make diagnoses. They are experts in determining what illness the patient is suffering from; they are not trained to determine causation.

[168] *Id.* at 467. Helpfulness in diagnostic testimony depends on whether it helps jurors "reason from a valid empirical framework to a valid diagnostic judgment." *Id.* at 468.

[169] *Id.* at 480. Again, the authors contrast their conclusion for eyewitness diagnostic testimony with medical causation testimony, contending that the jury would be unable to assess the different strands of general causation testimony, such as toxicology, epidemiology, etc. They do not discuss the reasoning behind this assumption, nor provide us with any data as to its basis.

[170] *Id.* at 477.

[171] *Id.*

[172] The authors do not come right out and say so, but it appears that eyewitness diagnostic testimony fails all five factors. The authors do say that while the five factors point toward admissibility of eyewitness framework testimony, they point in the opposite direction for diagnostic testimony. *See generally id.* at 443, 446, 468, 470.

[173] *Id.* at 474.

"empirically valueless diagnostic speculation will undermine the adjudication process."[174]

Their proposed solution, however, does nothing to address the statistical root of this basic issue. The G2i problem is foundational. Legal tweaking of balancing factors cannot solve it. Using a general consensus standard (as the authors suggest) for individuation testimony cannot solve the problem because as even the authors recognize, just because a "field claims the ability to apply general research to a particular case does not make it so."[175] Just because medical experts claim to be able to determine individual causation (and forensic scientists claim to individuate patterns) does not make it so.

Rather, the only way to solve the G2i problem is to understand that statistics is the law of large numbers. Nothing in statistical science permits the individuation that courts demand. And although the authors do not say this directly, they imply as much by recognizing the lack of empirical fit of such testimony.

Moreover their analysis of eyewitness expert testimony reveals these flaws, and the authors' suggestion that such experts be limited to general framework testimony makes good sense. What is unclear, however, is why they would exempt medical testimony from this kind of analysis. Such testimony would fail for the same reasons that eyewitness individuation testimony fails.

Empirical fit is similarly lacking in both eyewitness diagnostic and medical specific causation testimony. Determining causation simply is not what doctors are trained to do, anymore than are eyewitness experts, so they similarly ought to fail the second test, of qualifications.[176] Medical doctors could not have received training in clinical causation assessments, so they are not qualified by training or experience (testifying in court does not qualify them).[177] Medical doctors, by training and in their ordinary practice, decide what the illness is, and then treat it. Apart from microbial diseases like malaria and cholera, and contrary to the authors' assertion,[178] doctors do not (and cannot) determine what caused the particular ailment. The fourth factor, "value-added" helpfulness, is where the authors explicitly distinguish medical from eyewitness individuation testimony. They feel that jurors could draw their own conclusions from eyewitness general testimony, but not from general causation testimony in toxic torts, assuming that jurors could not put

[174] *Id.* at 480.

[175] *Id.* at 440.

[176] *Id.* at 444.

[177] *Id.*

[178] *Id.* at 468 ("[D]octors help jurors decide whether ingestion of a drug known to be associated with cancer caused cancer in the case at hand.").

together information from different strands of expertise such epidemiology, toxicology, chemical structure studies and the like. The authors do not, however, explain why the jury would be disadvantaged here, nor do they offer empirical support for such an assumption. As for the fifth factor, medical diagnostic testimony would appear to similarly distract the jury from "the uncertainty inherent in the field," thus failing the fifth factor.

Curiously, the authors mention and then fail to develop what could be a solution to the G2i problem in the courts.[179] The Classification of Violence Risk ("COVR") is a computerized actuarial risk assessment instrument that is based on data regarding risk factors for violent behavior. Depending on the individual being evaluated (usually for civil commitments), the software makes a prediction about the risk of violence posed by the individual being tested.[180] Importantly, however, the instrument's predictions are about the group to which the individual belongs, not risk that the individual will commit a violent act. The sample result cited by the authors is:

> Based on the data used to construct the *Classification of Violence Risk*, one can say with 95% confidence, that between 20% and 32% of persons with the same score as [person's name] can be expected to commit a violent act toward another person in the next several months, with a best estimate of 26 percent.[181]

Note what the instrument does not say: that the tested individual is 26% likely to recidivate. He is merely part of a group with an average recidivism rate of 26%. Some will recidivate more, some less. The inference about the tested individual's likelihood of recidivism is left to the factfinder. This is similar to what Professors Faigman, Monahan and Slobogin propose as a solution to the eyewitness expert scenario: let the expert testify to the general data and let the factfinder draw the inference about the individual. The same could legitimately be done with medical causation testimony.

Professors Faigman, Monahan and Slobogin have identified an important problem. Deriving individual inferences from general data is beyond the scope of current scientific expertise. Although their solution does not solve the problem they have identified, this paper is an important contribution to the dialog about translational errors from science to law.

[179] *Id.* at 455 (discussing the Classification of Violence Risk instrument in the context of standards and error rates).

[180] *Id.*

[181] *Id.*

C. Ditch Daubert for Individuation Testimony?

Julie Seaman, in her article *A Tale of Two Dauberts*,[182] recognizes that judges are reluctant to apply rigorous standards for evaluating scientific evidence in criminal cases. She notes that the NAS Report castigating forensic science has had little influence on judges.[183] As a result, while courts nominally apply the *Daubert* standards to both civil and criminal evidence, there are systematic differences in application.[184]

Professor Seaman notes the work-around that some courts have employed, by permitting experts to testify to the general data in criminal cases (in her examples of handwriting analysis and arson testimony) while leaving inference-drawing to the factfinder.[185] She worries, however, that "it is questionable whether this limitation provides very much protection from the potentially unreliable conclusions against which it is directed."[186] She observes that when two exemplars are presented to the jury, one from the crime scene and one from the defendant, accompanied by expert testimony about similarities in the patterns, it is a foregone conclusion that the jury will find a match.[187]

The two examples Professor Seaman has chosen for her study, however, are very different with respect to whether general data exists. Arson (fire origin and causation) testimony is based on general data about fires,[188] burning points of various materials, wind directions, etc.; handwriting testimony has no such database, just the *ipse dixit* of the expert regarding similarities in the exemplars. So while limiting the testimony of arson experts to general data without permitting individuation (that is, without permitting the expert to opine that this case involved arson) makes good sense, while in the absence of general data, handwriting experts have no value added to help the jury resolve an issue in the case.

[182] Julie A. Seaman, *A Tale of Two Dauberts*, 47 GA. L. REV. 889, 903 (2013).

[183] *Id.* at 894 (noting that "many prosecutorial applications of the forensic 'sciences' that are routinely admitted, that have long been admitted, and that continue to be admitted despite the serious questions raised" in the NAS Report would not satisfy a *Daubert* inquiry as it is performed in civil toxic court cases).

[184] *Id.* at 892 ("[T]he *Daubert* standard indeed may be disparately applied to even very similar evidence when offered in criminal versus civil cases.").

[185] *Id.* at 902 (observing that a judge concerned about the reliability of the testimony might find this approach "a move in the right direction").

[186] *Id.* at 902–03 (noting that even when handwriting experts are limited to identifying similarities and differences in handwriting, prosecution experts will focus on the similarities and dismiss differences as "individual variation").

[187] *Id.* at 902.

[188] *Id.* at 908 (Professor Seaman notes that while much arson testimony was based on assumptions that later proved to be false, the field has made progress with "fire investigation training and methodology" relying to a greater extent on scientific methodology).

In the face of academic disapproval and the NAS Report, most courts continue to admit forensic expert testimony that lacks scientific foundation, insisting that the testimony satisfies the *Daubert* test, which is "dishonest and misleading."[189] Professor Seaman observes that judges have little incentive to exclude forensic science expertise offered against a criminal defendant, and compelling reasons for admitting it.[190] As a consequence, she suggests that rather than have trial judges pretending to apply *Daubert* even as they are busy circumventing it, a more honest approach would be just to ditch *Daubert* in criminal cases.[191]

Unfortunately, however, ditching *Daubert* does not make the courts more honest or transparent. Instead, it subverts the search for truth to which our legal system aspires.[192] The hypocrisy of having a higher decision standard in criminal cases, but permitting the jury to consider evidence that has no empirical basis, will gnaw at the foundation of our justice system. Moreover, if the rationale for ditching *Daubert* is that judges circumvent it, we would have to ditch *Daubert* for specific causation testimony in toxic torts as well. Evidence without empirical foundation should not be presented in either criminal or civil cases.

[redacted] EDUCATING BENCH AND BAR ABOUT STATISTICAL INFERENCE DRAWING

The most effective way to solve the problem of bogus testimony in our courts is through education. The issue of statistical inference drawing from general data to individual instance in scientific testimony has seldom been addressed. There are few academic articles on the problem. The Federal Judicial Center, although it has a chapter on statistics,[193] does not discuss this problem. The NAS Report, which does identify the problem, has been ignored by the judiciary.[194]

[189] *Id.* at 920.

[190] *Id.* at 918 (noting that in the face of prosecutors' claims that they will be seriously prejudiced without forensic science testimomy, and that trial judges' admissibility decisions are "exceedingly unlikely" to be overturned on appeal, judges are apt to cite longstanding legal precedent to permit expert testimony where the forensic sciences are concerned).

[191] *Id.* at 918–19.

[192] Professor Seaman also worries about this problem. *See id.* at 921, 921 n.129 (acknowledging that her solution may seem "perverse" as it "allows the most questionable expert opinions to be admitted with little scrutiny whereas more reliable scientific evidence gets the full Daubert treatment.").

[193] *See* Kaye & Freedman, *supra* note 40, at 211.

[194] *See generally* NAS REPORT, *supra* note 134.

Most people do not understand statistics—even those people, like scientists, who often use them in their work.[195] Even though judges have become far more sophisticated in their approach to scientific testimony in the years since *Daubert* required their gatekeeping attention, statistical inference drawing has not yet penetrated most courts. And there are powerful incentives for judges not to look carefully at forensic techniques. But that is not an excuse to throw up our hands in despair. Rather, to encourage judges to take *Daubert* and the federal rules seriously, the following guidelines ought to help.

A. General Causation Issues

First, the purpose of epidemiologists' use of statistical concepts like relative risk, confidence intervals, and statistical significance are intended to describe studies, not to weed out the invalid from the valid. If the methodology is otherwise sound, small studies that fail to meet a P-level of 5, say, or have a relative risk of 1.3 for example, or a confidence level that includes 1 at 95% confidence, but relative risk greater than 1 at 90% confidence ought to be admissible. And understanding that statistics in context means that data from many sources need to be considered in the causation assessment means courts should not dismiss non-epidemiological evidence out of hand.

Some judges are quite capable of such analysis. The neurontin litigation is instructive in the care with which both sets of judges (federal and state) approached their gatekeeping duties.[196] At issue was the general causation testimony of whether neurontin can cause suicidal behavior.[197] The plaintiffs' general causation testimony was based on the theory that neurontin altered brain chemistry by increasing the amount of gamma-amino-butyric acid ("GABA") in the brain, leading to a decrease of other neurotransmitters like serotonin and norepinephrine, which prompts behavioral disturbances,

[195] *See id.* at 5 ("The length of the congressional charge and the complexity of the material under review made the committee's assignment challenging. In undertaking it, the committee first had to gain an understanding of the various disciplines within the forensic science community . . . .").

[196] *See In re* Neurontin Mktg., 612 F. Supp. 2d 116, 131 (D. Mass. 2009). Not only were federal and state court claims involved, but the federal court admissibility claims were resolved under a reliability (Daubert) analysis, while the state claims' admissibility determinations were resolved under a Frye standard. *Id.* at 122 n.3. Recognizing the importance of coordination of related claims, the admissibility determinations were conducted jointly, and, in a separate opinion, the state court adopted the federal court's analysis and findings on reliability. *Id.*

[197] *Id.* at 123.

depression, and suicidal behavior.[198] Plaintiffs based their theory on a Food and Drug Administration ("FDA") meta-analysis, animal studies, in vitro studies of human and animal tissue, and case reports,[199] all of which were hotly contested by defendants as unreliable bases for the plaintiffs' experts' conclusions.[200] Turning frequently to the Federal Judicial Center's Reference Manual on Scientific Evidence, the court painstakingly analyzed each of the contested studies, and how they related to each other and the plaintiffs' causation theory, before concluding that the plaintiffs' expert causation testimony was reliable.[201]

In each of the usually troubled areas of statistical significance, relative risk, animal studies, and in vitro studies, the court displayed active and thoughtful analytic prowess.[202] The most important study for the plaintiffs' causation theory was the FDA meta-analysis,[203] and the defendants raised a number of difficult issues.[204] The statistical significance problem arose because while neurontin was positively associated with "suicidality events" in the study, the associations were not statistically significant.[205] The court did not find this dispositive, both because suicide is a rare event (making it difficult to obtain the large number of subjects that would be required to produce statistical significance), and because the studies relating to the class of drugs to which neurontin belongs did demonstrate statistical significance.[206] As for relative risk, a relative risk of 1.57 and an attributable risk of 0.28 were enough to demonstrate a positive association, again because of the rarity of the event and the exclusion of high-risk subjects from the study.[207]

In addition, the defendants contested the chemical and pharmacological similarity of neurontin/gabapentin to the drugs in the FDA study.[208] However, because the plaintiffs' experts were able to explain what gabapentin does in the brain (increase the quantity of GABA), the analogy was strong enough to render extrapolation from other GABA-increasing drugs reliable, even if the

[198] *Id.* at 124.

[199] *Id.* at 132–56.

[200] *Id.* at 129–30.

[201] *Id.* at 130–59.

[202] *See id.* at 130–58.

[203] *See id.* at 133–37.

[204] *See id.* at 137–40.

[205] *Id.* at 137–38.

[206] *Id.* at 141.

[207] *Id.* at 138–39. It probably didn't hurt that a "blue ribbon committee" at the FDA had reviewed the FDA statistics and conclusions and found them sound. *Id.* at 140.

[208] *Id.* at 129 (noting the defendant's argument that the drugs were "distinct" from each other).

exact mechanism was unknown.[209] Finally, rather than reject out of hand the plaintiffs' animal, in vitro (tissue and cell), and case studies, the court carefully examined how each related to the plaintiffs' overall causation theory, and how, in conjunction, they were sufficient to demonstrate the reliability of the expert causation testimony.[210]

Cases like this demonstrate the feasibility of teaching non-scientist judges how to be intelligent consumers of scientific information. Thinking about causation in biological systems may not be taught in law school, but it should be. Thinking "like a lawyer" in our modern world includes more than teaching Newtonian-type causation; it requires addressing probabilistic causation issues. It requires an understanding of how complex systems work.

B. Specific Causation

Specific causation testimony should be limited to what the medical witness actually knows and practices. Testimony about what injury the plaintiff suffered, and what tests were performed to diagnose that injury is undoubtedly helpful to the jury. But as for causation, once the medical testimony has established the injury diagnosis, testimony should be limited to group data. Permitting medical witnesses to opine on individual causation goes far beyond their expertise. Experts are neither trained nor practiced in determining causation (outside the courtroom); their inference drawing offers no special insight to the jury. Because experts are not any more adept at such inference drawing, it should be left to the jury to draw inferences from the general data.

An example of how medical diagnostic testimony could be validly presented to jury can be derived from the way the COVR instrument presents the probable risk of future violence.[211] Rather than claiming that the tested individual will or will not commit another act of violence (or even the percentage of risk the individual poses), the instrument presents its findings as a probability statement that the tested individual is within a group with a certain percentage risk of recidivism.[212] By analogy, medical witnesses could similarly testify that the plaintiff, diagnosed with a particular disease (based

[209] *Id.* at 141–44.

[210] *Id.* at 130–58.

[211] Barbara E. McDermott et al., *Predictive Ability of the Classification of Violence Risk (COVR) in a Forensic Psychiatric Hospital*, 62 PSYCHIATRIC SERVICES 430, 430 (2011) (discussing how COVR "predict[s] community aggression among civilly committed psychiatric patients").

[212] Robert J. Snowden et al., *Assessing Risk of Future Violence Among Forensic Psychiatric Inpatients with the Classification of Violence Risk (COVR)*, 60 PSYCHIATRIC SERVICES 1522, 1522 (2009).

on particular factors, tests, family history, etc.), falls within a group (because of exposure to a particular chemical) that has a (say) 10% increased risk of injury.

This kind of limitation is not without precedent. Judge Jed Rakoff limited expert testimony in the ephedra litigation[213] to conclusions that could validly be drawn from the relied-upon studies. No definitive epidemiological study existed for ephedra, which, as a nutritional supplement, did not need FDA approval for marketing.[214] In the ephedra litigation, Judge Rakoff permitted the plaintiffs' witnesses to testify based on animal studies, analogous human studies, and biologically plausible theories of the mechanisms involved, but only to opine that "there is a reliable basis to believe that ephedra may be a contributing cause of cardiac injury and strokes in people with high blood pressure, certain serious heart conditions, or a genetic sensitivity to ephedra" rather than to the probability that ephedra causes heart attacks or strokes.[215] Most of the ephedra cases settled shortly after this decision.[216] By limiting the testimony to those inferences that could legitimately be drawn from the available evidence, Judge Rakoff balanced the realities of imperfect scientific knowledge with the requirements of sufficiently proving causation to achieve sound gatekeeping.

C. Criminal Identification Techniques

The flaw in criminal identification and arson testimony is the absence of general data from which anyone—expert or layperson—can draw a rational inference. The problem with criminal identification techniques is that (with the exception of DNA identification testimony) there is no general data from which to draw inferences.[217] DNA testimony is presented as general probability statements regarding the likelihood of randomly finding a similar configuration (a match) in the general population.[218] Most other criminal identification techniques are presented categorically (as a match), usually to the exclusion of all other individuals.[219] Rather than question the empirical

[213] *See* Rakoff, *supra* note 27, at 1392.

[214] *Id.* at 1390.

[215] *Id.* at 1391.

[216] *Id.*

[217] *See, e.g.*, NAS REPORT, *supra* note 134, at 149 (observing that forensic science experts present their conclusions categorically and without reference to population data).

[218] *See* People v. Nelson, 185 P.3d 49, 87 (Cal. 2008) (discussing statistical methods in DNA testimony).

[219] *See, e.g.*, NAS REPORT, *supra* note 134 at 127–82.

basis for this testimony, courts simply accept expert claims regarding their ability to make particular judgments.[220]

These criminal identification techniques differ from medical causation testimony in that there are no general data from which to draw inferences. Without a general population database, criminal identification techniques simply cannot meet relevance and reliability standards.[221] All suffer from the fallacy of attempting to draw individual conclusions without any data on the prevalence of the relevant characteristics (striations in latent fingerprints, bruises in bitemarks, etc.) in the population. Limiting testimony here to inferences that can validly be drawn from general data will not work in the absence of general population data. From an epistemic point of view, such testimony is worthless and undermines the concept of justice.

If courts were willing to take *Daubert* and Fed. R. Evid. 702 seriously, these techniques would not be admissible.[222] The fact that non-DNA evidence continues to be a mainstay of criminal prosecutions is a travesty.[223] As noted above, Congress attempted to ameliorate this problem by charging the National Academy of Science to examine the scientific validity of forensic science disciplines.[224] When the NAS did so, it found them wanting.[225]

The NAS Report carefully examined each of the criminal identification techniques commonly used in our courts and found them lacking scientific basis.[226] It proposed a number of solutions, but the key was increased research.[227] If such research were to be performed, the NAS Report was optimistic that the forensic science disciplines "might have the capacity (or the potential) to provide probative information to advance a criminal investigation."[228] The courts, however, have overwhelmingly ignored the NAS Report and continue to admit forensic testimony, *Daubert* or no

[220] *See* Faigman et al., *supra* note 2, at 438–39 (noting the statistical inference problem in criminal identification testimony).

[221] *See* NAS REPORT, *supra* note 134, at 114–15 (finding that the FBI used a population database for DNA analysis and that this led to reliable evidence).

[222] *Id.* at 53.

[223] *See id.* at 41 (noting that only 10% of crime labs' caseloads consist of DNA evidence).

[224] *See supra* note 140.

[225] *See* NAS REPORT, *supra* note 134, at 127–82 (castigating the lack of scientific foundation for pattern identification analysis in all disciplines in the forensic sciences).

[226] *Id.* at 149–83.

[227] *See id.* at 190 (calling for reform of the forensic sciences and outlining an agenda to expand independent research into the "accuracy, reliability and validity in the forensic science disciplines").

[228] *Id.* at 127.

Daubert.[229] I am not aware of a single case that has relied on the NAS Report in excluding forensic science testimony.[230]

In all these instances, experts could legitimately testify about the general data, to the extent it exists. In most of these instances, however, there is no such data. But even where data exist, for example, in fingerprint databases, forensic scientists must be candid about the limitations of the data. In order for such testimony not to be misleading, factfinders must be told how little is known about the prevalence of various patterns. And it must be left to the jury to draw inferences from the general data, since the expert has no special knowledge in this regard.[231] Which is, of course, precisely the opposite of what actually happens in our criminal courts, where the expert testifies that there is a match in the patterns, excluding the possibility that anyone other than the defendant could be the source.

Excluding pattern identification testimony that has failed to obtain the necessary data is the only proper solution for the courts. A system that requires a higher standard for expert testimony in civil than criminal cases does not appear to comport with our idea that criminal cases should be held to a higher standard of proof. Permitting into evidence testimony without empirical basis makes our justice system an empty promise.

[redacted]CONCLUSION

Translation of key statistical concepts is important in achieving fair and accurate admissibility decisions. Understanding the ideas represented by concepts such as statistical significance, relative risk and confidence intervals is essential if expert testimony is to be evaluated rationally. Even without a deep knowledge of statistics, judges should be able to realize that these concepts are not items to be checked off some list of factors for scientific validity, but a way of describing the results of scientific studies.

[229] See, for example, the following cases rejecting the argument that the NAS Report warrants exclusion of forensic testimony: United States v. Casey, 928 F. Supp. 2d 397, 400 (D.P.R. 2013) (rejecting the argument that the NAS Report warrants exclusion of firearm testimony); Commonwealth v. Joyner, 467 Mass. 176, 181, 181 n.6 (2014) (fingerprint testimony); State v. McGuire, 16 A.3d 411, 436 (N.J. Super. Ct. 2011) (toolmark analysis); Commonwealth v. Edmiston, 65 A.3d 339, 567, 571 (Pa. 2013) (microscopic hair analysis); Coronado v. State, 384 S.W.3d 919, 927 (Tex. App. 2012) (bitemark testimony); People v. Luna, 989 N.E.2d 655, 675 (Ill. App. Ct. 2013) (fingerprints).

[230] A search of the Westlaw Database "all courts" on October 7, 2014, using the search terms "NAS w/5 Report & forensic" yielded 38 cases, not one of which cited the NAS Report as a basis for excluding forensic testimony.

[231] *See* Commonwealth v. Gambora, 933 N.E.2d 50, 61 n.22 (Mass. 2010) (attempting to sidestep the issues raised in the NAS Report and explaining that expert fingerprint testimony is admissible as "opinion, not a fact").

Getting the courts to abandon their insistence that experts testify beyond their capabilities in medical causation and criminal identification cases may be more difficult. Individuation testimony gives the courts the alluring illusion of certainty. But courts are not unfamiliar with limiting expert testimony, knowing that experts, probably like all of us, tend to expound on matters beyond the boundaries of their expertise. A good example of judicial awareness of this propensity can be found in the In re: *Welding Fume Products Liability Litigation Daubert* order.[232] There, noting that all the experts were "to varying degrees, guilty of the same fault: they reach outside their area of expertise to opine about the ultimate issue of, for example, whether exposure to manganese in welding fumes can cause Parkinson's Disease," the court limited the scope of nearly all the parties' experts.[233]

On the other hand, "simply because a witness is not an expert about everything does not mean he is unqualified to offer expert opinion about anything."[234] Medical experts certainly know some things that would be helpful to a jury. They can ably testify about the process of diagnosis, what tests were run, and what signs and symptoms of disease were considered in arriving at a diagnosis. That is, a medical expert can testify about how the diagnosis of cancer, say, was made, and what alternative diseases were considered and ruled out—the process of differential diagnosis. Certainly that information would be helpful to the jury.

What is not helpful to the jury (because it is beyond the scope of medical experts' capabilities) is determining causation in the individual case. Doctors are no better at drawing individual inferences about causation than is the jury (or the judge). Their testimony is not helpful; it merely pretends to a certainty that does not exist. Judges should therefore limit the scope of medical testimony to what can legitimately be determined. The experts should testify to what they legitimately know, and the jury should draw the inference regarding causation.

The same is true of criminal identification evidence. The National Academy of Sciences laid out the problems and the solution. For forensic science to be scientific—and therefore admissible—research must be done to establish general population data. Then, as with DNA testimony, the jury must be told what the population statistics are and left to draw its own inference. There is no way around it: if we actually care about justice and we wish to use science to achieve it, it has to be real science, not its bogus imitation. At the core is also the issue that judges, litigators, and juries think

[232] No. 1:03–CV–17000, 2005 WL 1868046, at *6 (N.D. Ohio August 8, 2005).

[233] *Id.* (permitting only the neurologists to testify to whether manganese could cause Parkinson's).

[234] *Id.*

that science is about absolute answers, cut-offs, and bottom lines. The *Daubert* decision, citing Popper's philosophy, includes the idea that science provides only an approximation of the truth, and by repeated testing, comes closer to the truth, but never reaches "the absolute" truth.[235] Applied to statistical reasoning, that means an understanding that statistics is a form of logic, not a mechanistic dichotomy.

Thinking logically about science and causation issues is well within the capabilities of judges and the lawyers who inform them. One need not have taken courses in statistics, epidemiology, or medicine to understand sound reasoning. Law schools can help, by emphasizing probabilistic thinking as part and parcel of "thinking like a lawyer," a goal all legal educators claim. Even without law school preparation, some judges have mastered these concepts. Emphasis on reasoning, and better translation of key statistical concepts, is key to this enterprise.

[235] *See* Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 579, 593 (1993).