

2004

# Restoring the Individual Plaintiff to Tort Law by Rejecting 'Junk Logic' about Specific Causation

Vern R. Walker

*Maurice A. Deane School of Law at Hofstra University*

Follow this and additional works at: [https://scholarlycommons.law.hofstra.edu/faculty\\_scholarship](https://scholarlycommons.law.hofstra.edu/faculty_scholarship)

---

## Recommended Citation

Vern R. Walker, *Restoring the Individual Plaintiff to Tort Law by Rejecting 'Junk Logic' about Specific Causation*, 56 Ala. L. Rev. 381 (2004)  
Available at: [https://scholarlycommons.law.hofstra.edu/faculty\\_scholarship/141](https://scholarlycommons.law.hofstra.edu/faculty_scholarship/141)

This Article is brought to you for free and open access by Scholarly Commons at Hofstra Law. It has been accepted for inclusion in Hofstra Law Faculty Scholarship by an authorized administrator of Scholarly Commons at Hofstra Law. For more information, please contact [lawcls@hofstra.edu](mailto:lawcls@hofstra.edu).

# RESTORING THE INDIVIDUAL PLAINTIFF TO TORT LAW BY REJECTING “JUNK LOGIC” ABOUT SPECIFIC CAUSATION

Vern R. Walker\*

INTRODUCTION .....	382
I. UNCERTAINTIES AND WARRANT IN FINDING GENERAL CAUSATION: PROVIDING A MAJOR PREMISE FOR A DIRECT INFERENCE TO SPECIFIC CAUSATION .....	386
A. <i>Acceptable Measurement Uncertainty: Evaluating the         Precision and Accuracy of Classifications</i> .....	389
B. <i>Acceptable Sampling Uncertainty: Evaluating the         Population-Representativeness of Samples</i> .....	396
C. <i>Acceptable Modeling Uncertainty: Evaluating the Predictive         Value of Variables</i> .....	405
D. <i>Acceptable Causal Uncertainty: Explaining the Probability         of Event Occurrence</i> .....	423
II. UNCERTAINTIES AND WARRANT IN APPLYING THE GENERALIZATION TO THE INDIVIDUAL PLAINTIFF .....	437
A. <i>Acceptable Uncertainty About Plaintiff-Representativeness:         Selecting an Adequately Representative Reference Group</i> .....	439
B. <i>Acceptable Uncertainty About Assigning a Probability to a         Specific Member of the Reference Group</i> .....	448
III. MAKING WARRANTED FINDINGS ABOUT SPECIFIC CAUSATION .....	452
A. <i>An Integrated Approach to Decision-Making About         Acceptable Residual Uncertainty</i> .....	453
B. <i>Judicial Errors in Reasoning About Specific Causation</i> .....	460
1. <i>Judges as Factfinders and the “0.5 Inference Rule”</i> .....	460
2. <i>Judges as Referees of Reasonable Inferences and Rules             on Sufficiency of Evidence</i> .....	468
3. <i>Judges as Gatekeepers of Evidence and             Rules of Admissibility</i> .....	473
CONCLUSION .....	480

---

\* Professor of Law, Hofstra University; Ph.D., University of Notre Dame, 1975; J.D., Yale University, 1980. The author is grateful for the financial support provided by a research grant from Hofstra University. He also wishes to thank Nicole Irvin and Gisella Rivadeneira for their research assistance.

## INTRODUCTION

Judges have been removing the individual plaintiff from tort cases, often as a byproduct of a campaign against "junk science."<sup>1</sup> In place of the individual plaintiff, they have been installing an abstract "statistical individual" and adopting rules that decide cases on statistical grounds. This Article argues that, ironically, the reasoning behind these decisions and rules is too often an example of the "junk logic" that judges should be avoiding. The Article analyzes the logical warrant for findings of fact about specific causation and uses that analysis to critique such rules as (1) a "0.5 inference rule" for factfinding,<sup>2</sup> (2) a "greater-than-50%" rule for evaluating the legal sufficiency of evidence,<sup>3</sup> and (3) certain rules of admissibility following the Supreme Court's decisions in *Daubert* and *Kumho Tire*.<sup>4</sup> Judges are using such rules to wrongly decide a wide variety of tort cases, from products liability cases to medical malpractice to toxic exposure cases.<sup>5</sup>

This Article demonstrates that the many kinds of uncertainty inherent in warranted findings about specific causation require the factfinder to make decisions that are necessarily pragmatic, non-scientific, and non-statistical in nature. Such uncertainties are inherent in the logic of specific causation, and are not peculiar to toxic tort cases, or to epidemiologic evidence, or even to scientific evidence. The presence of significant degrees of such uncertainty makes it impossible to prove specific causation in any factual or scientific sense. Warranted findings must rest upon the common sense, practical fairness, and rough justice of the factfinder, except in categories of cases where tort policies can justify the adoption of decision rules for the entire category. Such rules, however, should not rest on the misguided statistical reasoning of past cases, but on proper policy foundations. If this

---

1. On the campaign against "junk science," see *General Electric Co. v. Joiner*, 522 U.S. 136, 153, 154 n.6 (1997) (Stevens, J., concurring in part and dissenting in part) (distinguishing the expert "weight of the evidence" reasoning in that case from "the sort of 'junk science' with which *Daubert* was concerned"); *Amorgianos v. National Railroad Passenger Corp.*, 303 F.3d 256, 267 (2d Cir. 2002) (stating that "[t]he flexible *Daubert* inquiry gives the district court the discretion needed to ensure that the courtroom door remains closed to junk science while admitting reliable expert testimony that will assist the trier of fact"); and *Daubert v. Merrell Dow Pharm., Inc.*, 43 F.3d 1311, 1321, 1322 n.18 (9th Cir. 1995) (stating that an expert's excluded testimony in that case "illustrates how the two prongs of Rule 702 [of the Federal Rules of Evidence] work in tandem to ensure that junk science is kept out of the federal courtroom").

2. See *infra* Part III.B.1.

3. See *infra* Part III.B.2.

4. See *infra* Part III.B.3.

5. E.g., *XYZ v. Schering Health Care Ltd.*, [2002] E.W.H.C. 1420 (QB), 2002 WL 1446183 (July 29, 2002) (finding against the claimants in products liability cases against manufacturers of oral contraceptives because they failed to prove a relative risk greater than 2.0); *Fennell v. S. Md. Hosp. Ctr., Inc.*, 580 A.2d 206 (Md. 1990) (holding, in a medical malpractice case, that evidence of a loss of a 40% chance of survival was legally insufficient for satisfying the plaintiff's burden of proving that the defendant caused the plaintiff's death); *In re Hanford Nuclear Reservation Litig.*, 292 F.3d 1124 (9th Cir. 2002) (holding, in cases brought against facility operators for injuries from radioactive emissions, that the district court erred in excluding plaintiffs' expert testimony, but leaving intact on remand the district court's ruling that admissible evidence on specific causation must show that exposure to the radioactive emissions at least doubled the plaintiffs' baseline risks).

Article can clear away these logical misunderstandings, perhaps judges will provide better policy justifications and develop better rules.

Tort law uses the term "specific causation," sometimes called "individual causation," to refer to the factual issue of which particular events caused or will cause a particular injury in a specific plaintiff.<sup>6</sup> Specific causation is distinguished from "general causation," also called "generic causation," which addresses whether there is any causal relationship at all between types of events and types of injuries.<sup>7</sup> Specific causation is whether a specific event caused or will cause a specific injury, while general causation is whether such events can (ever) cause such injuries.<sup>8</sup> Usually, for a plaintiff to win damages in a tort case, the plaintiff must prove both general and specific causation.<sup>9</sup>

A finding about specific causation can be prospective and predictive, as in: "It is unlikely that the defendant's negligent conduct, which resulted in the exposure of Jessica Jones to benzene, will cause her to develop lung cancer." Or a finding might be retrospective and explanatory, as in: "It is unlikely that the defendant's negligent conduct and Jessica Jones's resulting exposure to benzene caused her lung cancer." This Article argues that both versions, despite their temporal differences, have a similar logical structure in their warrant. Therefore, the analysis provided here applies to both prospective and retrospective findings of specific causation.

The central epistemic problem posed by specific causation is justifying how a less-than-universal generalization about causation in groups can ever warrant a probabilistic finding about causation in a specific case.<sup>10</sup> When

6. *E.g.*, *DeLuca v. Merrell Dow Pharm., Inc.*, 911 F.2d 941, 957-59 (3d Cir. 1990) (reasoning from the plaintiff's burden of proving causation by "a more likely than not standard" to a requirement that epidemiologic evidence alone would be legally insufficient evidence of specific causation unless it showed a "relative risk of limb reduction defects" of at least two); *Hanford*, 292 F.3d at 1129, 1133 (using the term "individual causation" to refer to the question of "whether a particular individual suffers from a particular ailment as a result of exposure to a substance"); Michael D. Green et al., *Reference Guide on Epidemiology*, in *REFERENCE MANUAL ON SCIENTIFIC EVIDENCE* 333, 396 (2d ed. 2000) (defining "specific causation"), available at [http://www.fjc.gov/public/pdf.nsf/lookup/sciman06.pdf/\\$file/sciman06.pdf](http://www.fjc.gov/public/pdf.nsf/lookup/sciman06.pdf/$file/sciman06.pdf); Joseph Sanders & Julie Machal-Fulks, *The Admissibility of Differential Diagnosis Testimony to Prove Causation in Toxic Tort Cases: The Interplay of Adjective and Substantive Law*, 64 *LAW & CONTEMP. PROBS.* 107, 110 (2001).

7. *E.g.*, *Merrell Dow Pharm., Inc. v. Havner*, 953 S.W.2d 706, 714 (Tex. 1997) (defining "[g]eneral causation" as "whether a substance is capable of causing a particular injury or condition in the general population"); *Hanford*, 292 F.3d at 1133 (defining "general" or "generic causation" to mean "whether the substance at issue had the capacity to cause the harm alleged"); *Sterling v. Velsicol Chem. Corp.*, 855 F.2d 1188, 1200 (6th Cir. 1988) (defining "generic causation"); *In re "Agent Orange" Prod. Liab. Litig.*, 818 F.2d 145, 165 (2d Cir. 1987) (identifying the "generic causation" issue); Green et al., *supra* note 6, at 392 (defining "general causation").

8. Commentators have also accepted the usefulness of the distinction. *E.g.*, Green et al., *supra* note 6, at 381-86 (discussing the role of epidemiologic evidence in proving specific causation in addition to general causation); Bernard D. Goldstein & Mary Sue Henfin, *Reference Guide on Toxicology*, in *REFERENCE MANUAL ON SCIENTIFIC EVIDENCE* 401, 422-26 (2d ed. 2000) (discussing toxicology and causation in the individual case), available at [http://www.fjc.gov/public/pdf.nsf/lookup/sciman07.pdf/\\$file/sciman07.pdf](http://www.fjc.gov/public/pdf.nsf/lookup/sciman07.pdf/$file/sciman07.pdf).

9. *See, e.g.*, *Hanford*, 292 F.3d at 1134 (stating that plaintiffs must establish both generic and individual causation); *Havner*, 953 S.W.2d at 714-15 (discussing general and specific causation).

10. The logical structure of warrant for a direct inference, which this Article examines, is distinct

and why does statistical evidence about causation in a group warrant a finding of causation in a specific member of the group? For example, if 10% of people who experience a certain type of chemical exposure later develop cancer as a result of that exposure, what is the probability that the exposure of a specific person (for example, Jessica Jones) will cause (or did cause) *her* to develop cancer? What reason is there to place *her* in the 10% category, as opposed to the 90% category? The problem of justification exists regardless of the magnitude of the statistics. If 75% of certain types of patients at a certain stage of a disease die within five years from the normal progression of the disease, despite the best treatment, then what is the probability that a specific individual with the disease, who was misdiagnosed when her disease was at the relevant stage and who subsequently died within five years, would have died from the disease in any case, despite the misdiagnosis? Warranted findings about such questions depend upon inferences from what typically (or statistically) happens in groups of which the specific individual is a member.<sup>11</sup>

This Article shows that every warranted finding about specific causation possesses certain types of uncertainty or potential for error. Part I of the Article examines uncertainties about general causation that decrease the warrant or evidentiary support for a conclusion about specific causation. It demonstrates that there are four logically distinct types of uncertainty that are necessarily present: measurement uncertainty, sampling uncertainty, modeling uncertainty, and causal uncertainty. For each type of uncertainty, there are techniques for reducing and characterizing the extent or degree of uncertainty. In the end, however, a reasonable factfinder must decide whether the residual uncertainty of each type is acceptable or not for purposes of the tort case—that is, for warranting a conclusion of specific causation in the context of tort law.

Part II of the Article analyzes two additional uncertainties involved in drawing a conclusion about a specific individual. The first section addresses the problem of identifying the appropriate group to serve as a reference

---

from the cognitive process of producing possible generalizations on which to base a direct inference. For an example of an analysis of the latter process, see JEROME P. KASSIRER & RICHARD I. KOPELMAN, *LEARNING CLINICAL REASONING* 2-46 (1991) (analyzing the process of generating, refining, and verifying medical hypotheses for diagnosing diseases of specific patients in a clinical setting).

11. A parallel, evolving paradigm in medicine is called evidence-based medicine (EBM), in which physicians evaluate the best available scientific information and apply it to specific patients. David L. Sackett et al., *Evidence Based Medicine: What It Is and What It Isn't*, 312 *BRIT. MED. J.* 71-72 (1996) (defining EBM as “the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients”), available at <http://bmj.bmjournals.com/cgi/content/full/312/7023/71>; DAN MAYER, *ESSENTIAL EVIDENCE-BASED MEDICINE* 9-15 (2004) (discussing six steps in the process of EBM).

Another related paradigm in logic and artificial intelligence is “abduction,” or “inference to the best explanation,” which is a process of reasoning from an effect to an explanatory cause. See *ABDUCTIVE INFERENCE: COMPUTATION, PHILOSOPHY, TECHNOLOGY* 5, 17 (John R. Josephson & Susan G. Josephson eds., 1994). However, the definition of abduction sometimes excludes direct inference. See *id.* at 23-24.

group for that individual.<sup>12</sup> It examines the warrant for finding that a reference group adequately represents the specific individual—that is, that it adequately matches the plaintiff in all causally relevant variables, such as being a woman, being age forty, having no history of cancer in the immediate family, and so forth. The second section of Part II discusses the uncertainty in assigning a particular probability to the individual case, even when the reference group adequately represents the specific individual. These two major sources of uncertainty can be called, respectively, uncertainty about plaintiff-representativeness and uncertainty about assigning a probability to the individual plaintiff.

Part III of the Article summarizes all of these uncertainties into a coherent factfinding approach. It then uses this logical analysis to critique certain judicial rules that are threatening individualized factfinding in tort law. The argument is that for each type of uncertainty, as well as for the overall uncertainty, someone must decide whether the residual uncertainty is acceptable for the purposes of tort law. Such decisions cannot be purely epistemic or scientific in nature, because they involve balancing the expected risks and benefits of making findings in the face of uncertainty, as well as weighing the equitable treatment of the parties and other non-epistemic considerations.<sup>13</sup> While expertise can inform certain aspects of those decisions, there is no reason to think that experts are the optimal decision-makers. Policy-based rules are needed to determine who should decide such questions (the jury or the judge) and whether such decisions should be made on case-specific factors or on rules governing categories of cases.

Unfortunately, instead of facing such non-epistemic issues squarely and developing policies and rules to address them, many judges have relied on faulty reasoning to adopt unjustified rules that *appear* to be logically compelling. The second section of Part III examines a variety of cases in which judges have relied on such fallacious reasoning—cases involving judicial factfinding about liability for oral contraceptives, judicial rulings on sufficiency of evidence in medical malpractice cases, and judicial decisions on admissibility of expert testimony in toxic-exposure cases. Specific causation in such cases cannot be a scientific issue, however, and any decisions should be justified on substantive policy grounds, not statistical grounds. It is reasonable to hope that once such judges better understand the warrant for

---

12. For the logical literature taking formal, technical approaches to the problem of the reference class, see HANS REICHENBACH, *THE THEORY OF PROBABILITY* 372-78 (1949) (approaching the problem of the reference class by considering "the narrowest class for which reliable statistics can be compiled"); Isaac Levi, *Direct Inference and Confirmational Conditionalization*, 48 *PHIL. OF SCI.* 532 (1981) (considering three approaches to selecting reference sets for direct inference), available at <http://www.jstor.org/view/00318248/ap010194/01a00030/0>; and Henry E. Kyburg, Jr., *The Reference Class*, 50 *PHIL. OF SCI.* 374 (1983), available at <http://www.jstor.org/view/00318248/ap010201/01a00020/0>.

13. See, e.g., ARIEL PORAT & ALEX STEIN, *TORT LIABILITY UNDER UNCERTAINTY* 18-56 (2001) (evaluating various decision rules for allocating uncertainty in tort law using the non-epistemic policies of utility and fairness); RICHARD GOLDBERG, *CAUSATION AND RISK IN THE LAW OF TORTS: SCIENTIFIC EVIDENCE AND MEDICINAL PRODUCT LIABILITY* 190-212 (1999) (evaluating a probabilistic approach to causation from the standpoint of non-epistemic goals, especially economic efficiency).

finding specific causation, they will rest their rulings on a proper policy basis, and restore the individual plaintiff to tort law.

I. UNCERTAINTIES AND WARRANT IN FINDING GENERAL CAUSATION:  
PROVIDING A MAJOR PREMISE FOR A DIRECT INFERENCE  
TO SPECIFIC CAUSATION

The first step toward restoring individualized decision-making in tort law is to understand the logical role of uncertainty in inferences about specific causation. One reasonable approach to warranting a finding about specific causation is to infer it from empirical evidence about general causation—that is, from evidence about causal relationships in groups of which the specific plaintiff is a member. Logicians call such reasoning from group generalizations to specific instances a “direct inference.”<sup>14</sup> A typical direct inference has the following form:<sup>15</sup>

Most things in category *A* are also in category *B*.

This specific individual is in category *A*.

Therefore, this specific individual is probably also in category *B*.

An example is:

Most people who receive a certain dose of a particular chemical experience nausea.

This specific person will receive that dose of the chemical.

Therefore, this specific person will probably also experience nausea.

Direct inferences use information about group proportions (here, “most things in category *A*” and “most people who receive a certain dose of a particular chemical”) to help warrant findings about specific individuals.

---

14. Kyburg, *supra* note 12; Isaac Levi, *Direct Inference*, 74 J. PHIL. 5 (1977) [hereinafter *Direct Inference*], available at <http://www.jstor.org/view/0022362x/di973127/97p0003v00>; Levi, *supra* note 12; JOHN L. POLLOCK, NOMIC PROBABILITY AND THE FOUNDATIONS OF INDUCTION 108-48 (1990).

15. Instead of using category variables (variables that identify groups or sets of individuals), an equivalent formulation of direct inference can use propositional variables (variables that stand for whole propositions or statements). Using the variables *p* and *q* to stand for any two propositions or statements, the standard form of direct inference would be:

In most situations when *p* is true, *q* is also true.

In this specific situation, *p* is true.

Therefore, in this specific situation, probably *q* is also true.

The implicit quantification is not over individuals as members of sets but over situations as characterized by whether *p* and *q* are true.

Many direct inferences use somewhat vague group proportions (such as "very few," "about half of," or "almost all") to warrant findings with degrees of probability that are ordinal in nature (expressed in such terms as "unlikely," "equally likely," and "highly likely"). For example:

Very few things in category *A* are also in category *B*.

This specific thing is in category *A*.

Therefore, it is unlikely that this specific thing is also in category *B*.

Some direct inferences use cardinal quantities, and are called "statistical syllogisms,"<sup>16</sup> such as:

*X*% of things in category *A* are also in category *B*.

This specific individual is in category *A*.

Therefore, there is a probability of *X*% that this specific individual is also in category *B*.

In this quantitative formulation, a statistic from the group evidence helps warrant a mathematical probability that the conclusion is true.<sup>17</sup>

Regardless of how the direct inference is formulated, a major source of inferential uncertainty is the fact that the generalization is not universal: not all things in *A* have characteristic *B*, and only a subset of *As* are *Bs*. If every member of *A* were also a member of *B*, then an inference that any specific individual in *A* is also in *B* would be deductively valid.<sup>18</sup> Instead, even if the two premises of a direct inference are true, the conclusion can still be false. Therefore, a direct inference is defeasible, and is at best presumptively sound.<sup>19</sup>

A direct inference argument consists of two premises and a conclusion. The first or major premise is a generalization (or a statistical generalization) asserting that some proportion of things in category *A* are in fact also in

16. For examples of this terminology, see ABDUCTIVE INFERENCE: COMPUTATION, PHILOSOPHY, TECHNOLOGY, *supra* note 11, at 23-24; JOHN L. POLLOCK & JOSEPH CRUZ, CONTEMPORARY THEORIES OF KNOWLEDGE 229-30 (2d ed. 1999); and WESLEY C. SALMON, LOGIC 87-91 (2d ed. 1973).

17. The inference is called a "statistical syllogism" because a statistical premise (such as "*X* percent of *Fs* are *G*") is used instead of a universal generalization ("All *Fs* are *G*"). POLLOCK, *supra* note 14, at 75-78; SALMON, *supra* note 16, at 88-91. Toulmin calls the more general form of argument "quasi-syllogistic." STEPHEN EDELSTON TOULMIN, THE USES OF ARGUMENT 109-11, 131-34, 139-41 (1958). For an early recognition of the difficulty posed by such inferences for legal theory, see George F. James, *Relevancy, Probability, and the Law*, 29 CAL. L. REV. 689 (1941).

18. See, e.g., POLLOCK & CRUZ, *supra* note 16, at 229; Vern R. Walker, *Direct Inference in the Lost Chance Cases: Factfinding Constraints Under Minimal Fairness to Parties*, 23 HOFSTRA L. REV. 247 (1994).

19. For a discussion on defeasible or presumptive reasoning, see DOUGLAS N. WALTON, ARGUMENTATION SCHEMES FOR PRESUMPTIVE REASONING 17-45 (1996).



category *B*. Ways of expressing such a generalization include: "Most things having characteristic *A* also have property *B*"; "Most *As* are also *Bs*"; or "In most cases, if something is an *A*, then it is also a *B*." Some examples are: "The majority of adult Americans weigh over 100 pounds"; "Over half of the men in the study who took the drug developed a rash"; and "Approximately five of every 1000 persons of northern European descent are homozygous for the recessive gene for hemochromatosis."

The second or minor premise of a direct inference is a categorical assertion that a specific individual is in category *A*, is a member of group *A*, or is characterized as having property *A*. Logicians call group *A* the "reference class" or "reference group."<sup>20</sup> The minor premise identifies some specific individual (for example, "Jessica Jones," or "the current President of the United States," or "the wife of the victim" in a particular tort case) and classifies that individual as being a member of group *A*.

The conclusion of the direct inference is a probabilistic proposition stating that, with some degree of probability, the same individual identified in the minor premise is also in category *B*, or is also a member of group *B*, or also has characteristic *B*.<sup>21</sup> There are various ways of expressing degrees of probability in English, such as: "This individual is probably a *B*"; "It is probably the case that this individual is a *B*"; "More likely than not this individual is a *B*"; or "There is a 0.6 probability that this individual is a *B*."

This Article examines a particular use of direct inference or statistical syllogism—namely, to warrant conclusions about specific causation in tort cases. In this use, a causal relation to category *A* is one of the defining properties of category *B*. When used to warrant an inference to specific causation, direct inference takes a more particular form:

Most individuals in category *A* are also in category *B* as a result of being in category *A*.

This specific individual is in category *A*.

Therefore, this specific individual is probably also in category *B*, as a result of being in category *A*.

For example, category *A* might be "people exposed to chemical *C*," and category *B* might be "people who develop cancer as a result of exposure to chemical *C*." The major premise would be "most people exposed to chemi-

---

20. See, e.g., HENRY E. KYBURG, JR., *SCIENCE AND REASON* 41 (1990) [hereinafter *SCIENCE AND REASON*]; BRIAN SKYRMS, *CHOICE AND CHANCE: AN INTRODUCTION TO INDUCTIVE LOGIC* 201 (2d ed. 1975).

21. Ayer considered such a proposition to be a distinct class of judgments, which he called "judgements of credibility." A. J. AYER, *PROBABILITY AND EVIDENCE* 27-29, 54-61 (1972). As he said, "the judgement that such and such an individual smoker will probably die of lung cancer, if it is genuinely a judgement about this individual, and not just about the class of smokers to which he belongs, is a judgement of credibility." *Id.* at 28.

cal *C* are also people who develop cancer as a result of that exposure.”<sup>22</sup> Causation is a criterion for being a member of subgroup *B*. Although some concept of causal relationship or causal link helps to define category *B*, the validity of the direct inference is independent of any particular meaning given to “causation.”<sup>23</sup> A vague concept of causation may well increase the uncertainty about the membership of group *B* (as discussed below), but that problem is distinct from the more general problem of warranting direct inferences. The warranting problem analyzed in this Article is independent of the problem of how to define legal causation.

This part of the Article analyzes the sources and types of uncertainty inherent in finding the major premise of the direct inference to be true. For each kind of uncertainty, the factfinder should decide how extensive the residual uncertainty is and whether that residual uncertainty is acceptable for the purposes of tort law.<sup>24</sup> In analyzing uncertainty, the discussion introduces a number of logical and statistical concepts, such as random and biased error, statistical significance and statistical power, and relative risk and regression analysis. Such concepts, while precisely defined within science, are refinements of common-sense notions that routinely guide the reasoning of everyday life. The analysis uses these concepts to identify the kinds of uncertainty that are logically inherent in the major premise about general causation. By using scientific concepts to elucidate common-sense logic, the analysis also provides a conceptual bridge between the evidence of the expert witness and the findings of the non-expert factfinder, and it lays a foundation for investigating degrees of evidentiary support between legally available evidence and findings of fact.

#### *A. Acceptable Measurement Uncertainty: Evaluating the Precision and Accuracy of Classifications*

Knowledge about general causation rests upon observational evidence, which ranges from the casual perceptions of everyday experience to the carefully conducted measurements of scientists. The warrant for a generalization about causation is only as strong as those underlying observations. Observations or measurements, from the logical point of view, are acts of

---

22. A statistical syllogism about specific causation would have a similar structure. For example: 40% of individuals in category *A* are also in a category *B* as a result of being in category *A*. This specific individual is in category *A*. Therefore, there is a probability of 40% that this specific individual is also in category *B*, as a result of being in category *A*.

Using the chemical-exposure example in the text, such a major premise would be: “40% of people exposed to chemical *C* also develop cancer as a result of that exposure to chemical *C*.”

23. For problems with defining legal cause, see DAN B. DOBBS, *THE LAW OF TORTS* 405-41 (2001); W. PAGE KEETON ET AL., *PROSSER AND KEETON ON THE LAW OF TORTS* 263-321 (5th ed. 1984); and *RESTATEMENT (SECOND) OF TORTS* §§ 430-32.

24. For a general treatment of the epistemic role of theories of uncertainty, see Vern R. Walker, *Theories of Uncertainty: Explaining the Possible Sources of Error in Inferences*, 22 *CARDOZO L. REV.* 1523 (2001).

classifying individual objects or events into the classification categories of a predicate or variable. When people (including trained researchers) observe objects or events, gather information about them, or form opinions about them, they are classifying them into categories under some variable. When scientists record the results of such acts of classification, they call such summary reports "data." The value or score for an individual on a variable is the name of the classification category in which the individual is placed. A data set might record the heights of the students in a particular room, the LSAT scores of an entering law school class, the salaries of male and female employees in a corporation, or the symptoms of the plaintiffs in a tort suit.

Uncertainty, or the potential for error, is inherent in every observation or measurement and, therefore, in the data reporting those classifications.<sup>25</sup> If there are misclassifications in the data, this may result in erroneous conclusions about groups of individuals. Scientists generally call this the problem of measurement error or measurement uncertainty about the data.<sup>26</sup> The potential for misclassifying an individual object or event arises from many sources. For example, if classification categories are not mutually exclusive<sup>27</sup> and exhaustive,<sup>28</sup> the design of the classification system can increase the likelihood of inconsistent classifications. Inconsistencies in classification can also arise due to predicate vagueness—when there are not clear and

---

25. See, e.g., NAT'L RESEARCH COUNCIL, SCIENCE AND JUDGMENT IN RISK ASSESSMENT 161 (1994) (defining "uncertainty" as "a lack of precise knowledge as to what the truth is, whether qualitative or quantitative"). Uncertainty is distinct from variability in the data. Even if there were no uncertainty in the individual measurements, there could still be considerable variability in the data, reflecting actual differences in classification between the individuals measured. See, e.g., *id.* at 221 n.1 (discussing variability as referring to "a dispersion of possible or actual values" or to "individual-to-individual differences in quantities associated with predicted risk").

26. There are many general texts on measurement theory, including EDWARD G. CARMINES & RICHARD A. ZELLER, RELIABILITY AND VALIDITY ASSESSMENT (Michael S. Lewis-Beck ed., 1979), and EDWIN E. GHISELLI ET AL., MEASUREMENT THEORY FOR THE BEHAVIORAL SCIENCES (1981). There are also references dealing with particular techniques of measurement, such as MEASUREMENT ERRORS IN SURVEYS (Paul P. Biemer et al. eds., 1991); Theodore Peters & James O. Westgard, *Evaluation of Methods*, in TEXTBOOK OF CLINICAL CHEMISTRY 410 (Norbert W. Tietz ed., 1986); and Lloyd A. Currie, *Sources of Error and the Approach to Accuracy in Analytical Chemistry*, in 1 TREATISE ON ANALYTICAL CHEMISTRY 119-22 (I. M. Kolthoff & Philip J. Elving eds., 2d ed. 1978). Some general statistics texts have good treatments of measurement error, such as DAVID FREEDMAN ET AL., STATISTICS 90-101, 247-81, 395-411 (2d ed. 1991). For additional discussion and documentation, see J. M. Cameron, *Error Analysis*, in 2 ENCYCLOPEDIA OF STATISTICAL SCIENCES 545, 550 (Samuel Kotz & Norman L. Johnson, eds., 1982); David H. Kaye & David A. Freedman, *Reference Guide on Statistics*, in REFERENCE MANUAL ON SCIENTIFIC EVIDENCE 83, 102-04 (Fed. Judicial Ctr., 2d ed. 2000); and Vern R. Walker, *The Siren Songs of Science: Toward a Taxonomy of Scientific Uncertainty for Decisionmakers*, 23 CONN. L. REV. 567, 580-88 (1991) [hereinafter *Siren Songs*].

27. E.g., HERMAN J. LOETHER & DONALD G. MCTAVISH, DESCRIPTIVE AND INFERENTIAL STATISTICS: AN INTRODUCTION 17 (4th ed. 1993); H. T. Reynolds, *Nominal Data*, in 6 ENCYCLOPEDIA OF STATISTICAL SCIENCES 256 (Samuel Kotz & Norman L. Johnson, eds., 1985). If the variable is color, then the categories "yellow" and "non-red" are not mutually exclusive because a yellow object can be properly classified into either category.

28. E.g., LOETHER & MCTAVISH, *supra* note 27, at 17; Reynolds, *supra* note 27, at 256. If the variable is color, then the categories "red" and "blue" are not exhaustive because yellow objects would properly fit into neither category. Adding the category "other" would create an exhaustive set of categories.

operational criteria for classifying individuals into one category as opposed to another. This is a problem inherent to the predication structure of thought itself, not merely language.<sup>29</sup> When classification criteria are vague or non-existent, there are insufficient guidelines for coordinating a classification process conducted by different people, or even by the same person over time. Other sources of classification error are not due to flaws in the design of the predicates involved, but to causal aspects of the classification process. Inconsistent classifications can arise because human beings differ in their judgments, miscalculate, are inattentive, or act from ulterior motives. Instruments can malfunction, or they may operate erratically outside certain tolerances. Such causes of misclassification can sometimes be addressed by providing better training, redesigning instruments, and setting up procedures for quality control and assurance. Actual measurement processes, however, will always remain causal processes, and will produce measurements with some degree of variability.

An important objective in everyday life, in science, and in the courtroom is to identify the extent of uncertainty in those classifications that provide the evidentiary support for generalizations, in order to estimate the extent to which inferences are based upon actual differences in the world, as opposed to error in measuring the world.<sup>30</sup> Scientists have various techniques for detecting, characterizing, and reducing the measurement uncertainty due to misclassification. To the extent that there is random error ("scatter" or "noise") in the data,<sup>31</sup> experts say there is "imprecision" in the data or measurement process.<sup>32</sup> This refers to the inconsistency of results when the measurement process is used to classify the same individual a number of times. When scientists refer to a measurement process as being "reliable," they mean that it produces the same classification results on repeated measurements of the same thing.<sup>33</sup> A perfectly reliable measurement

---

29. The problem of vagueness is sometimes grouped with other concept design problems under the general umbrella of "linguistic imprecision." See, e.g., M. GRANGER MORGAN & MAX HENRION, *UNCERTAINTY: A GUIDE TO DEALING WITH UNCERTAINTY IN QUANTITATIVE RISK AND POLICY ANALYSIS* 50, 60-62 (1990). Vagueness, however, is not always simply a matter of "sloppy language." It is a problem inherent to the predication structure of thought itself, not merely one's use of language. There is no reason to be as optimistic as Morgan and Henrion, who write: "Whereas many sources of uncertainty, including lack of information and computational limitations, are often expensive or impossible to eliminate, uncertainty due to linguistic imprecision is usually relatively easy to remove with a bit of clear thinking." *Id.* at 61-62.

30. Cf. NAT'L RESEARCH COUNCIL, *supra* note 25, at 188-91 (illustrating how "uncertainty and variability can complement or confound each other").

31. In terms to be introduced in Part I.C., there is no "correlation" between errors due to unreliability and the true value. CARMINES & ZELLER, *supra* note 26, at 30; JACOB COHEN & PATRICIA COHEN, *APPLIED MULTIPLE REGRESSION/CORRELATION ANALYSIS FOR THE BEHAVIORAL SCIENCES* 68 (2d ed. 1983).

32. A retest or re-measurement of the same thing under similar conditions is sometimes called a "replication experiment." The term "precision" is sometimes used to refer to the "agreement between replicate measurements," and the degree of imprecision is the magnitude of the random scatter around the true value. See John Mandel, *Accuracy and Precision: Evaluation and Interpretation of Analytical Results*, in 1 *TREATISE ON ANALYTICAL CHEMISTRY* 256-60 (I.M. Kolthoff & Philip J. Elving eds., 2d ed. 1978); Peters & Westgard, *supra* note 26, at 412 (analytical chemistry).

33. For further discussions of reliability, see CARMINES & ZELLER, *supra* note 26, at 11-13, 29-51;

process would produce exactly the same classification every time the same individual is classified. On the other hand, an unreliable measurement process produces some degree of random error from even repeated acts of classification.

Regardless of the source, if the resulting error is symmetrical and random, then it can be expected to "cancel out" over a long run of repeated measurements.<sup>34</sup> For example, if a large number of repeated measurements are taken, a symmetrical frequency distribution around the mean value would be expected.<sup>35</sup> The degree of dispersion in the repeat-measurement data can be used to characterize the unreliability or imprecision of the measurement process.<sup>36</sup> For example, the variance or standard deviation for a set of repeated measurements indicates the amount of variability around the central value.<sup>37</sup>

Variation is not the same as uncertainty.<sup>38</sup> How much variation there is in a data set depends upon the construction of the variable and the characteristics of the individuals in the group. Uncertainty is a very different matter. It is the potential for error in a proposition or inference.<sup>39</sup> The fact that indi-

FREEDMAN ET AL., *supra* note 26, at 90-101, 395-411; GHISELLI ET AL., *supra* note 26, at 184, 191; Robert M. Groves, *Measurement Error Across the Disciplines*, in MEASUREMENT ERRORS IN SURVEYS 1-25 (Paul P. Biemer et al. eds., 1991); and Mandel, *supra* note 32, at 259.

34. The third important dimension of any data set is its *form* or *shape*. The form of the distribution is determined by how the individual values are actually distributed over the categories of the variable. If a frequency distribution is symmetrical, then the distributions on each side of the central tendency are mirror images of each other. In a perfectly symmetrical distribution, the mean and median have the same value. MICHAEL O. FINKELSTEIN & BRUCE LEVIN, STATISTICS FOR LAWYERS 3-4 (2d ed. 2001); WILLIAM L. HAYS, STATISTICS 180 (5th ed. 1994). A frequency distribution might not be symmetrical, however, but rather skewed toward one end of the scale. A skewed distribution has asymmetrical "tails," with values at one extreme disproportionate to the other. HAYS, *supra*, at 180; LOETHER & MCTAVISH, *supra* note 27, at 119-20.

35. This issue is closely related to sampling theory, discussed in Part I.B. The error terms of the repeat measurements are expected to average zero over the very long run if these errors result from a very large number of causal factors, and they are expected to be unbiased in net result. See FREEDMAN ET AL., *supra* note 26, at 90-101, 247-81, 395-411; HAYS, *supra* note 34, at 247-49. Examples are heights of human beings and other natural biological traits. See *id.* at 243-44, 247-49; FINKELSTEIN & LEVIN, *supra* note 34, at 113-15; MORGAN & HENRION, *supra* note 29, at 85-88. The normal distribution provides a convenient approximation for many sampling distributions based on large sample sizes. See HAYS, *supra* note 34, at 243-44. Many statistical texts provide basic accounts of the formal characteristics of the normal distribution. E.g., *id.* at 237-60; FINKELSTEIN & LEVIN, *supra* note 34, at 113-19; MORRIS HAMBURG, STATISTICAL ANALYSIS FOR DECISION MAKING 191-211 (3d ed. 1987); LOETHER & MCTAVISH, *supra* note 27, at 125-29; MORGAN & HENRION, *supra* note 29, at 85-88; Kaye & Freedman, *supra* note 26, at 153-59.

36. See CARMINES & ZELLER, *supra* note 26, at 43-47 (discussing Cronbach's alpha); GHISELLI ET AL., *supra* note 26, at 193-94, 204, 205-07 (discussing reliability coefficients).

37. The variance is the arithmetic average of the squared differences between the individual values and the mean. The standard deviation reverses the squaring operation by taking the square root of the variance. See, e.g., HAYS, *supra* note 34, at 182-84; LOETHER & MCTAVISH, *supra* note 27, at 137.

38. See, e.g., MORGAN & HENRION, *supra* note 29, at 62-63; NAT'L RESEARCH COUNCIL, *supra* note 25, at 160-223.

39. In a sense, the dispersion or variability is an indication of the error associated with using a central tendency statistic as a predictor for the group of individual cases. As one statistics author has put it:

If central tendency measures are thought of as good bets about observations in a distribution, measures of spread represent the other side of the question: Dispersion reflects the "poor-

viduals are classified in different categories (variation) is an issue quite distinct from whether they are classified correctly (uncertainty). There can be variation without much uncertainty, and uncertainty without observed variation. Variation is a feature of objects or events (a characteristic of the world), while uncertainty is a feature of our information or inferences about those objects or events (a characteristic of our beliefs about the world). Confusion has a way of setting in, however, in part because experts sometimes use statistics of dispersion to characterize uncertainty.

If the results of a reliability experiment can be generalized to all results of that measurement process, then the standard deviation from the reliability study might warrant an estimate of the unreliability associated with any single measurement. If it is necessary or desirable to reduce the amount of error due to unreliability, however, one approach is to take a number of measurements of the subject individual and to calculate an average value to use in lieu of a single observation. Another approach is to redesign the measurement process to bring the degree of unreliability within acceptable bounds. Perhaps a redesigned measurement instrument would produce more precise data. There may also be a trade-off between the reliability of the measurement process and the unit of measurement. An instrument used to measure length to a hundredth of an inch might do so with a high degree of imprecision, but the instrument might produce far more reliable results measuring length only in feet. This merely means that retests will produce more consistent answers when measurements are taken in feet than when they are taken in hundredths of inches. The pragmatic question is whether length measured to the nearest foot provides acceptable imprecision.<sup>40</sup>

The notion of "validity" captures a very different aspect of measurement uncertainty. The validity of the results of a measurement process concerns whether the data in fact measure what they are supposed to measure.<sup>41</sup> Validity is said to address the "accuracy" of the measurement process, not its precision.<sup>42</sup> An instrument that tends to overestimate length presents a

---

ness" of central tendency as a description of a randomly selected case, the tendency of observations *not* to be like the average.

HAYS, *supra* note 34, at 182.

40. Carrying a number to decimal places well beyond what can be warranted given the level of precision of the measurement method is also potentially misleading. See, e.g., Donald A. Berry & Seymour Geisser, *Inference in Cases of Disputed Paternity*, in *STATISTICS AND THE LAW* 353, 376 (Morris H. DeGroot et al. eds., 1986) (giving the example of using a six-digit paternity index when measurement and sampling uncertainty render calculations to more than two digits "suspect").

41. For similar definitions of "validity," see CARMINES & ZELLER, *supra* note 26, at 12; GHISELLI ET AL., *supra* note 26, at 266; LOETHER & MCTAVISH, *supra* note 27, at 15, 34; Kaye & Freedman, *supra* note 26, at 103-04. In forensic science, a forensic technique's validity depends upon the percentage of cases in which the analyst can make correct determinations. E. J. Imwinkelried, *A New Era in the Evolution of Scientific Evidence—A Primer on Evaluating the Weight of Scientific Evidence*, 23 WM. & MARY L. REV. 261, 279 (1981). The term "valid" is sometimes applied by extension to the measurement instrument used to gather the data, as well as to the data.

42. See, e.g., Peters & Westgard, *supra* note 26, at 412 ("The term *inaccuracy* has been recommended to emphasize lack of agreement [between results of the method being evaluated and the criterion method] and is defined as the 'numerical difference between the mean of a set of replicate measurements and the true value.'") (quoting J. Büttner et al., *Provisional Recommendation on Quality Control in Clini-*

problem of validity. A person to whom red and orange objects appear indistinguishable may over-count the number of "red" things in her visual field. A criminal justice system that convicts defendants in part based on their race, rather than merely on the strength of the evidence, is measuring the race of the defendant, not merely guilt or innocence.

Lack of validity shows up not as random scatter in the data, but as some degree of bias or systematic error. Sometimes comparing a data set against an expected distribution produces evidence of invalidity. For example, if measurements are expected to approximate a normal distribution, then skewed results might suggest that some causal factor is at work producing the biased results.<sup>43</sup> The notion of invalidity entails a comparison between one measurement process and another measurement process that is thought to provide a standard.<sup>44</sup> Bias relative to that standard constitutes evidence that something is being measured other than what is intended or supposed.<sup>45</sup>

Errors introduced at the level of measurements or individual observations can have significant effects on predictions, theories, and the direction of later research, but they can be very difficult to correct once the data are gathered. If the source of the bias can be determined, sometimes the measurement process can be redesigned so that the distorting factors are eliminated, or at least the degree of invalidity reduced. Even if the bias cannot be eliminated or reduced, there might be a standard or criterion measure available for determining the degree of residual error. If the amount of bias is known and stable, that information might be used to adjust the measurement results. For example, if a watch faithfully "loses" a minute each hour, a per-

---

*cal Chemistry*, 22 CLINICAL CHEMISTRY 532, 538 (1976) (emphasis omitted)). This is a narrow use of the term "accuracy." In its broader use, "accuracy" refers to the degree of correspondence between any descriptive proposition and the reality it purports to describe. In this broader sense, "inaccuracy" includes total error from all sources, not just from measurement invalidity.

Measurement accuracy can be further analyzed using the model of diagnostic tests, including such concepts as forward and backward probabilities, and sensitivity and specificity. See *infra* notes 193-209 and accompanying text.

43. To understand fully the potential for error in measurement, one must model the measurement process as a causal process in which the outcome event is a good predictor of some aspect of the input event. Color vision can be the basis for measurement or classification of visible objects only if our visual experience validly measures some characteristic of visible objects. An understanding of why color vision produces valid results, however, requires an understanding of how color vision works, employing at least a rudimentary theory of cause-and-effect.

44. Regulatory agencies sometimes officially establish "reference methods" as criteria. See, e.g., Occupational Safety and Health Administration Regulations on Asbestos, 29 C.F.R. § 1910.1001 (2000); Environmental Protection Agency Regulations on National Primary and Secondary Ambient Air Quality Standards, 40 C.F.R. pt. 50 (2000); Environmental Protection Agency Regulations on Ambient Air Quality Surveillance, 40 C.F.R. pt. 58 (2000).

45. A regulatory example involves the ambient air concentrations of sulfates. Certain glass filters used in high-volume air samplers were believed to result in overestimation of true sulfate concentrations in the measured air, and the EPA noted in an interstate air pollution proceeding that the petitioning states had not corrected the sulfate data for artifact formation caused by the sampling technique. Environmental Protection Agency, Interstate Pollution Abatement, 49 Fed. Reg. 34,851, 34,863 (Sept. 4, 1984) (Proposed Determination), 49 Fed. Reg. 48,152, 48,153 (Dec. 10, 1984) (Final Determination). An example from the behavioral sciences is the California F Scale, which may be interpreted as measuring two different properties at the same time: adherence to authoritarian beliefs and the trait of tending to agree with assertions. CARMINES & ZELLER, *supra* note 26, at 15.

son can still use it to reach accurate conclusions by resetting it at a certain time each day and allowing for the "lost" time since the last resetting. Whether measurement results can be adjusted to offset a lack of validity depends upon the nature of the measurement process and our knowledge of the extent of the bias. The social sciences struggle with methods for testing validity because criterion measurement methods are rare.<sup>46</sup> For jury findings, there is seldom an independent criterion or procedure by which to determine accuracy, so it may not be possible to identify bias, or to adjust for it if it occurs.

From the standpoint of producing warranted findings, reliability and validity pose different kinds of problems. They differ in the nature of the evidence needed and in the means available to characterize the uncertainty. The evidence of unreliability is primarily "internal" to a measurement process and a data set. If reliability studies are conducted using a test-retest protocol, then the evidence of unreliability consists of inconsistent classifications exhibiting random variation. The evidence of invalidity, on the other hand, is often "external" to the data set, and involves comparing data gathered using two measurement processes for the same property or characteristic. Theories of uncertainty are needed to warrant any finding that the two measurement processes are measuring "the same thing."

The question of whether degrees of reliability and validity are acceptable requires a decision about whether these potentials for error should be tolerated. Precision and accuracy can sometimes be improved, but usually at a cost, and trade-offs are generally necessary. Scientists weigh the costs of seeking additional precision or accuracy, and they routinely make pragmatic decisions about whether to commit additional resources.<sup>47</sup> Every scientific study that is introduced into evidence resulted from pragmatic decisions about how much precision and accuracy to tolerate in the study. Decisions about acceptable reliability and validity are therefore inherently pragmatic; someone must determine whether the data are "reliable enough" and "sufficiently valid" for the purposes at hand. An institution conducting legal fact-finding might leave such decisions to the factfinder, or share them with the presiding judge as issues of law, or leave them by default with an expert witness. The pragmatic nature of the decision should be a factor in allocating such decision-making power.

---

46. On "construct validity" as an alternative to criterion validity in the social sciences, see CARMINES & ZELLER, *supra* note 26, at 22-26; GHISELLI ET AL., *supra* note 26, at 282-87. On "content validity" as another alternative, see CARMINES & ZELLER, *supra* note 26, at 20-22; and GHISELLI ET AL., *supra* note 26, at 275-77.

47. Currie, *supra* note 26, at 199-209; Peters & Westgard, *supra* note 26, at 413-15; James O. Westgard & George G. Klee, *Quality Assurance*, in TEXTBOOK OF CLINICAL CHEMISTRY 424 (Norbert W. Tietz ed., 1986).



*B. Acceptable Sampling Uncertainty: Evaluating the  
Population-Representativeness of Samples*

Sampling uncertainty is the potential for error introduced precisely because the inference proceeds from sample statistics to conclusions about population parameters.<sup>48</sup> The population is the group that is the subject of the generalized conclusion, while the sample is a subset of objects or events selected from that population.<sup>49</sup> The sample provides the actual data used to infer what measuring the entire population would show. Statisticians call a summary number characterizing the population a "parameter," while they call a summary number characterizing a sample a "statistic."<sup>50</sup> Therefore, an inference from statistics (descriptive of sample data) to parameters (characterizing a population) creates sampling uncertainty.

In general, the source of sampling uncertainty is the actual process of selecting members of the sample, which can generate a sample that is unrepresentative of the intended population. In short, the sampling process can cause bias or systematic error in the sample when that sample is evaluated from the standpoint of statistical representativeness to the target population. Some causes of unrepresentative sampling are easy to detect. For example, if for reasons of convenience researchers conduct a voting poll only in the shopping malls of major cities, then the resulting sample may not be representative of voters in rural areas or of less affluent voters. If a study sample consists of patients in hospitals, then it may not be representative of people in the general population.<sup>51</sup> Causal influences on sampling are not always easy to detect, however, and methods are needed to minimize them. This section of the Article discusses various scientific methods for reducing sampling uncertainty and various analytic techniques for characterizing the residual amount of sampling uncertainty.

Scientists approach the problem of representativeness of a sample to a population in two ways. The first is to take explicitly into account any factors thought to be important in the target population. If certain factors are known to be associated with the variables being studied, then the population might be divided into sub-populations ("strata") on the basis of those factors. Random sampling within such strata will probably produce a more representative sample than would sampling without stratification.<sup>52</sup> For example, if researchers know the demographics of the target population (such as age groupings) and expect those demographic factors to correlate with the variables being studied (such as the occurrence of a disease), then they

---

48. NAT'L RESEARCH COUNCIL, *supra* note 25, at 165 (including random sampling error and non-representativeness as species of "parameter uncertainty").

49. HAYS, *supra* note 34, at 204-06.

50. *Id.*; LOETHER & MCTAVISH, *supra* note 27, at 5.

51. See, e.g., XYZ v. Schering Health Care Ltd., [2002] E.W.H.C. 1420 (QB), 2002 WL 1446183, ¶ 292 (July 29, 2002).

52. See HAROLD A. KAHN & CHRISTOPHER T. SEMPOS, STATISTICAL METHODS IN EPIDEMIOLOGY 14-20 (1989); LOETHER & MCTAVISH, *supra* note 27, at 392-98.

might ensure that the sample mirrors the population in the proportions of those demographic factors. By creating sub-populations that are more homogeneous on a study variable than the total population is, the variability in random samples from those sub-populations can be less than it would be in a simple random sample drawn from the total population.<sup>53</sup> Even after drawing the sample, known relevant factors may be taken into account by stratifying the data during the analysis. For example, if age is a known relevant factor and age data are gathered on all the study subjects, then researchers can analyze the data within age groups.<sup>54</sup> One danger in this approach is that after the sample is drawn and the data are stratified, some strata may contain too few sample members for the desired statistical analysis.<sup>55</sup>

Stratified random sampling also helps to address the problem of confounding factors.<sup>56</sup> Confounding factors are variables that help explain observed associations (or the lack of observed associations) and whose omission from the analysis injects a potential for error.<sup>57</sup> When confounding factors are present and not controlled, they can bias a sample relative to its population, and produce statistical associations in the sample that lead to error if generalized to the population.<sup>58</sup> Controlled experiments address confounding factors by attempting to hold constant all causally relevant factors except those being studied. Observational studies can sample, measure, and statistically control any potential confounders.<sup>59</sup> Therefore, to the extent that causally relevant factors are known, stratified random sampling or stratified data analysis can increase the likelihood that the conclusion drawn will be acceptably accurate with respect to the target population.

The second scientific approach to the problem of sample representativeness is to build randomization into the sampling process.<sup>60</sup> A random

53. KAHN & SEMPOS, *supra* note 52, at 14-19; LOETHER & MCTAVISH, *supra* note 27, at 388, 393-94. For a discussion of matching controls to cases in a case-control design, see ABRAHAM M. LILIENTHAL & DAVID E. LILIENTHAL, FOUNDATIONS OF EPIDEMIOLOGY 347-52 (2d ed. 1980) (matching controls to cases in order to reduce sampling bias and increase sampling precision).

The variability discovered in the measurements can also influence the uncertainty about such central-tendency statistics as the mean. A National Research Council report gives the following example:

A group of 1000 workers observed in an epidemiologic study, for example, may have an *average* susceptibility to cancer significantly greater or less than the true mean of the entire population, if by chance (or due to a systematic bias) the occupational group has slightly more or slightly fewer outliers (particularly those of extremely high susceptibility) than the overall population. In such cases, estimates of potency or population incidence drawn from the worker study may be overly "conservative" (or insufficiently so).

NAT'L RESEARCH COUNCIL, *supra* note 25, at 238.

54. DAVID CLAYTON & MICHAEL HILLS, STATISTICAL MODELS IN EPIDEMIOLOGY 135, 141-52 (1993).

55. *Id.* at 135; LILIENTHAL & LILIENTHAL, *supra* note 53, at 348.

56. See *infra* notes 164, 178 (discussing how a confounding variable is statistically associated with one or more independent variables and with a dependent variable); Green et al., *supra* note 6, at 369; Kaye & Freedman, *supra* note 26, at 92.

57. See Green et al., *supra* note 6, at 369-70; Kaye & Freedman, *supra* note 26, at 138.

58. Green et al., *supra* note 6, at 370.

59. See, e.g., Kaye & Freedman, *supra* note 26, at 138-39.

60. For informative introductions to sampling theory, see HAYS, *supra* note 34, at 53-54; LOETHER & MCTAVISH, *supra* note 27, at 381-87; THOMAS H. WONNACOTT & RONALD J. WONNACOTT, INTRODUCTORY STATISTICS 190-94 (5th ed. 1990); Finkelstein & Levin, *supra* note 34, at 256-81; Kaye

sampling process is useful with respect to factors whose causal relevance is unknown or uncertain.<sup>61</sup> For example, a sampling process is a "simple" random procedure if on each selection of an individual to be a member of the sample, every individual in the population has an equal chance of being drawn.<sup>62</sup> After a researcher specifies a sample size (the number " $N$ " of individuals in the sample) it is possible to compute probabilities for sample types (a "sampling distribution"). If, for example, a simple random sample of 1000 individuals (" $N = 1000$ ") is to be selected from a population in which 40% would vote for Carter, there would be a probability of less than 0.01 that the sample would have fewer than 350 individuals (less than 35%) voting for Carter.<sup>63</sup>

Randomly drawing a sample can then support the following hypothetical reasoning about an unknown parameter in the population:

*If the parameter equals  $X$ , then randomly selecting a sample of size  $N$  with statistic  $S$  has a very low probability of occurrence.*<sup>64</sup>

A sample of size  $N$  with statistic  $S$  was selected in a random manner from the population.

Therefore, *probably* the parameter does *not* equal  $X$ .

The warrant for this inference rests on the hypothesis about the population parameter, the random nature of the sampling method, the sample size, and probability theory. Randomly drawing a sample that has little likelihood of being drawn is itself good evidence that the population parameter does not have the hypothesized value. If the population percentage of voters for Carter *were* 40%, then it would be very unlikely that one would draw by pure chance a sample of 1000 members with less than 35% voting for

---

& Freedman, *supra* note 26, at 115-33, 153-59.

61. LOETHER & MCTAVISH, *supra* note 27, at 394.

[O]ne of the strengths of the simple random sample is that when one is ignorant of the relevant variables other than the independent and dependent variables, one can execute a simple random sample with some confidence that the unknown but relevant variables will be sampled in approximately the proportions in which they occur in the total population.

*Id.*

62. On this meaning of simple random sampling, see HAYS, *supra* note 34, at 53-54; LOETHER & MCTAVISH, *supra* note 27, at 381-82.

63. The supporting calculation is as follows. With simple random sampling, the expected value or mean for the sampling distribution of the proportion is 0.4, and the standard error is approximately 0.015. With a simple random sample of this size and such a hypothetical parameter, the sampling distribution for the percentage closely approximates a normal distribution. See HAMBURG, *supra* note 35, at 185-94; HAYS, *supra* note 34, at 128-48, 177-79, 189-90; LOETHER & MCTAVISH, *supra* note 27, at 417-27. In the example in the text, the form of a two-tailed sampling distribution is approximately normal, and less than 1% of samples are expected to fall outside 2.58 standard errors on either side of the mean. Therefore, the probability of randomly selecting a sample with less than 35% voters for Carter is less than 0.01.

64. This Article adopts a "classical" approach to sampling theory, which means that probabilities have the relative-frequency interpretation that is familiar from gambling. See, e.g., WONNACOTT & WONNACOTT, *supra* note 60, at 70-101; Kaye & Freedman, *supra* note 26, at 117 n.112.

Carter. Drawing a sample with a very low probability is therefore good grounds for rejecting the hypothesis.

Such calculations of sample probability can then lead to decision rules or inference rules about hypothesis rejection. Scientists typically consider acceptable evidence for rejecting an hypothesis to be drawing one of the 5% least likely samples—one out of the set of samples that has an aggregate probability less than 0.05.<sup>65</sup> Randomly drawing one of these 5% least likely samples is so improbable an event that doing so is called “statistically significant,”<sup>66</sup> and warrants rejecting the hypothesis as probably false.<sup>67</sup> Of course, the hypothesis might still *be* correct, and it is said to be a “Type I error” to reject a true hypothesis.<sup>68</sup>

Scientists also report the statistical significance of sampling results in terms of “p-values,” which is short for “probability values.”<sup>69</sup> The p-value for a statistic is the probability of drawing that statistic value or a more extreme value, given the hypothesis. Thus, instead of saying “the sample percentage of 35% was statistically significant at the 0.05 level,” a scientist might report the same information in terms of a p-value: “the sample percentage was 35% ( $p < 0.05$ ).” The parenthetical “( $p < 0.05$ )” asserts the proposition that the p-value for this sample result is less than 0.05. The conventional decision rule is therefore to reject the hypothesis if sample results have a p-value less than 0.05. Using this 0.05 probability convention as the basis for rejecting hypotheses, however, means that Type I errors may occur in about 5% of sampling cases. P-values state probabilities for drawing samples of certain types, not probabilities about the truth or falsehood of the hypothesis.<sup>70</sup> But drawing a sample result that has a low p-value furnishes

65. MARTIN BLAND, AN INTRODUCTION TO MEDICAL STATISTICS 152 (1987) (medical sciences); COHEN & COHEN, *supra* note 31, at 20-21 (behavioral sciences); LOETHER & MCTAVISH, *supra* note 27, at 484 (sociology); Michael Cowles & Caroline Davis, *On the Origins of the .05 Level of Statistical Significance*, 37 AM. PSYCHOLOGIST 553, 553 (1982); James H. Ware et al., *P Values, in MEDICAL USES OF STATISTICS* 181, 185-88 (John C. Bailar III & Frederick Mosteller eds., 2d ed. 1992) (medical literature).

66. BLAND, *supra* note 65, at 148-62; HAYS, *supra* note 34, at 270-84; LOETHER & MCTAVISH, *supra* note 27, at 480-93.

67. Courts have adopted similar statistical reasoning to determine whether a plaintiff has presented a prima facie case of unlawful discrimination. See, e.g., *Wards Cove Packing Co. v. Atonio*, 490 U.S. 642, 650-55 (1989); *Hazelwood Sch. Dist. v. United States*, 433 U.S. 299, 307-13 (1977); *Castaneda v. Partida*, 430 U.S. 482, 492-99 (1977); *Palmer v. Shultz*, 815 F.2d 84, 90-97 (D.C. Cir. 1987).

68. HAYS, *supra* note 34, at 282; LOETHER & MCTAVISH, *supra* note 27, at 489-90; WONNACOTT & WONNACOTT, *supra* note 60, at 302-03.

69. On p-values, see WONNACOTT & WONNACOTT, *supra* note 60, at 293-301; Kaye & Freedman, *supra* note 26, at 121-23; and Ware et al., *supra* note 65, at 181. Wonnacott and Wonnacott describe the p-value for a sample result relative to the null hypothesis as a measure of the “credibility” of the null hypothesis, and if “this credibility falls below” the selected level of significance, then the null hypothesis can be rejected. *Id.* at 301. With this in mind, the “p-value is the lowest that we could push the level  $\alpha$  [the level of significance] and still be able (barely) to reject  $H_0$  [the null hypothesis].” WONNACOTT & WONNACOTT, *supra* note 60, at 301 n.7.

70. In classical sampling theory, it is a mistake to think that the p-value is the probability that the hypothesis is true. As Kaye and Freedman note: “The significance level tells us what is likely to happen when the null hypothesis is correct; it cannot tell us the probability that the hypothesis is true.” Kaye & Freedman, *supra* note 26, at 125. Or as they say in another place: some statements in cases “confuse the probability of the kind of outcome observed, which is computed under some model of chance, with the

evidence that mere chance due to the sampling probably does not explain the difference between the expected result based on the hypothesis and the actual result obtained.<sup>71</sup>

Using the set of the 5% least likely samples to reject hypotheses is a convention among scientists, involving a potential for random sampling error that is deemed acceptable for purposes of scientific research.<sup>72</sup> Like many conventions, the threshold of acceptable uncertainty could be drawn somewhat higher or lower, and within a rough range one could plausibly draw any line for rejecting and not rejecting the hypothesis.<sup>73</sup> When probabilities are interpreted as relative frequencies, then the convention on the level of significance for rejecting hypotheses also implicitly accepts a long-run error rate for such rejections.<sup>74</sup> If a large number of samples were drawn from a population in which the hypothesis is true, then for a 0.05 level of significance one expects to draw statistically significant results about 5% of the time.<sup>75</sup> Hypothetical reasoning from the population to the probability of

probability that chance is the explanation for the outcome." *Id.* at 122 n.132. Because this Article adopts classical thinking about probability, p-values are probabilities assigned to samples, not to the hypothesis itself. Moreover, it is important to distinguish between assertions assigning a mathematical probability to a sample and an inference rule that warrants rejecting hypotheses as "improbable" on the basis of p-values for actual data. These different uses of the term "probable" invite the very confusion against which Kaye and Freedman warn. *Id.*

71. See Environmental Protection Agency, Guidelines for Carcinogen Risk Assessment, 51 Fed. Reg. 33,992, 33,997 (Sept. 11, 1986). As Kaye and Freedman have stated, "significant differences are evidence that something besides random error is at work, but they are not evidence that this 'something' is legally or practically important." Kaye & Freedman, *supra* note 26, at 124.

72. See CARL F. CRANOR, REGULATING TOXIC SUBSTANCES: A PHILOSOPHY OF SCIENCE AND THE LAW 29-48 (1993); HAYS, *supra* note 34, at 267-82, 302-03 ("[C]onventions about significant results should not be turned into canons of good scientific practice"); Neil B. Cohen, *Confidence in Probability: Burdens of Persuasion in a World of Imperfect Knowledge*, 60 N.Y.U. L. REV. 385, 409-17 (1985); Cowles & Davis, *supra* note 65, at 553 (detailing the history behind the convention, and suggesting that the choice was related to the earlier concept of "probable error"); Michael O. Finkelstein, *The Application of Statistical Decision Theory to the Jury Discrimination Cases*, 80 HARV. L. REV. 338, 364 (1966); Michael D. Green, *Expert Witnesses and Sufficiency of Evidence in Toxic Substances Litigation: The Legacy of Agent Orange and Bendectin Litigation*, 86 NW. U. L. REV. 643, 683-84 (1992) ("[T]he choice of .05 is an arbitrary one," and "[u]ltimately, the relative costs and benefits of the types of errors must be compared to decide the appropriate level of statistical significance"); Donald N. McCloskey, *The Loss Function Has Been Mislaid: The Rhetoric of Significance Tests*, 75 AM. ECON. REV. 201, 201 (1985), available at <http://links.jstor.org/sici?sici=0002-8282%28198505%2975%3A2%3C201%3ATLFBM%3E2.0.CO%3B2-9> ("Roughly three-quarters of the contributors to the *American Economic Review* misuse the test of significance."); Jeffrey H. Silber & Herbert Kaiser, *Loss Weighting and the Human Cost of Experimentation*, 38 J. CHRONIC DISEASES 507 (1985) (arguing that the choice of the level of statistical significance in medical clinical studies is a function of implicit decisions concerning the relative importance of future versus present patients); Ware et al., *supra* note 65, at 186 (the popular scientific convention has a disadvantage of suggesting "a rather mindless cut-off point, which has nothing to do with the importance of the decision to be made or with the costs and losses associated with the outcomes"). On the importance of considering practical consequences in selecting a level of significance, see CRANOR, *supra* note 72, at 31-48; HAYS, *supra* note 34, at 283-84; and LOETHER & MCTAVISH, *supra* note 27, at 499-501.

73. In some areas or applications, scientists use other probabilities to define a "very low probability," such as 0.01 (the 1% least likely samples). See, e.g., Kaye & Freedman, *supra* note 26, at 124; HAYS, *supra* note 34, at 283-84; LOETHER & MCTAVISH, *supra* note 27, at 484.

74. On error rates and the two types of errors, see HAYS, *supra* note 34, at 279-84; FINKELSTEIN & LEVIN, *supra* note 34, at 120-22; and LOETHER & MCTAVISH, *supra* note 27, at 479-93.

75. See LOETHER & MCTAVISH, *supra* note 27, at 489-90.

samples, therefore, raises the question of what error rate is acceptable for rejecting hypotheses. A decision is needed concerning what degree of uncertainty is acceptable in risking a Type I error.

Hypothetical reasoning is also used in another facet of sampling uncertainty. A scientist is concerned not only with the probability of wrongly rejecting a true hypothesis (making a Type I error), but also with the probability of correctly rejecting a false hypothesis. The "statistical power" of a sampling procedure is the probability of correctly concluding that a false hypothesis is in fact probably false.<sup>76</sup> It is the probability of drawing a statistically significant result (leading to rejection of the hypothesis), or a sample with a p-value more extreme than the chosen level of significance. Statistical power is the answer to the question: "How probable is it that a statistically significant result will be drawn if the hypothesis being tested is *not* correct?" This probability obviously depends upon what the true value is. For example, if the true value is exactly the critical value needed to reject the hypothesis, and any sample drawn is as likely as not to be statistically significant (that is, fall above or below that critical value), then the statistical power of the study would be 0.5 or 50%. In such a case, the power would increase as the difference between the hypothesis and the true value increases, and a hypothesis that is "way off" the true value is more likely to be rejected than one that is "close."<sup>77</sup>

Statistical power is therefore a function of the hypothesis chosen, the critical value of the sample statistic for rejecting that hypothesis, the true value in the population, and the probability distribution for the statistic based on the true value.<sup>78</sup> In a specific study, the researcher determines the hypothesis value and calculates the critical value for rejecting it, given a level of significance (for example, 0.05). Both the critical value and the probability distribution based on the true value depend upon the sample size  $N$ , with power increasing as  $N$  increases. The third factor, however, the true population value, is unknown. The power for a study is therefore a set of probability distributions based on a range of possible true values.

---

76. For treatments of statistical power, see BLAND, *supra* note 65, at 159-60; CLAYTON & HILLS, *supra* note 54, at 206-09; COHEN & COHEN, *supra* note 31, at 59-61, 162, 166 & app. G.2; HAYS, *supra* note 34, at 284-95, 328-34; FINKELSTEIN & LEVIN, *supra* note 34, at 82-87, 509-510; STEVE SELVIN, STATISTICAL ANALYSIS OF EPIDEMIOLOGIC DATA 71-89 (1991); Rebecca DerSimonian et al., *Reporting on Methods in Clinical Trials*, in MEDICAL USES OF STATISTICS 333, 343 (John C. Bailar III & Frederick Mosteller eds., 2d ed. 1992); Jennie A. Freiman et al., *The Importance of Beta, the Type II Error, and Sample Size in the Design and Interpretation of the Randomized Controlled Trial: Survey of Two Sets of "Negative" Trials*, in MEDICAL USES OF STATISTICS 357, 358-64 (John C. Bailar III & Frederick Mosteller eds., 2d ed. 1992); Kaye & Freedman, *supra* note 26, at 125-26; and Ware et al., *supra* note 65, at 195-96.

77. The power of a sampling protocol can be compared to the notion of detection. Power is the probability of detecting a difference between the hypothetical value and the true value, which is of course unknown. See SELVIN, *supra* note 76, at 71; Freiman et al., *supra* note 76, at 359-60. As stated by Hays, power is "not unlike the power of a microscope. Power reflects the ability of a decision rule to detect from evidence that the true situation differs from a hypothetical one." HAYS, *supra* note 34, at 287-89.

78. COHEN & COHEN, *supra* note 31, at 59; HAYS, *supra* note 34, at 284-93; SELVIN, *supra* note 76, at 72-73; Freiman et al., *supra* note 76, at 359-62.

As long as the true value remains unknown, and so the real statistical power of the study remains unknown, of what use is a power determination to legal factfinding? One answer is that the power of the study can influence how a factfinder should interpret the study results.<sup>79</sup> For example, in the early 1980s there was a dispute over whether the Environmental Protection Agency should find that there was "a reasonable basis to conclude" that formaldehyde presents a significant risk of serious or widespread harm to humans from cancer.<sup>80</sup> Part of the dispute involved an epidemiologic mortality study that did *not* show a statistically significant increase of risk of dying from cancer among exposed workers compared to unexposed workers.<sup>81</sup> From the standpoint of statistical power, however, the study had only a 4% chance of rejecting the "null" hypothesis that there was no increase in risk for cancer of the pharynx or of the larynx (assuming a 0.05 level of significance), even if there had been in reality a two-fold increase in risk for those exposed. Mayo reports that to have "a fairly high probability (.8)" of obtaining a statistically significant difference in those types of cancer in this study, the actual increase in risk would have to have been on the order of forty or fifty times.<sup>82</sup> The absence of statistically significant results in such a study cannot warrant a finding that there was *no* increased risk to exposed workers. Low statistical power to detect even a substantial increased risk is evidence that the sample involved is too small to warrant a finding that such a risk increase is unlikely. Lack of power cautions a reasonable factfinder against drawing inferences of no general causation.

As discussed above, a Type-I error is the error of getting a "false-positive" sampling result—that is, of obtaining statistically significant results by pure chance and wrongly rejecting a correct hypothesis. A "false-negative" error (or "Type-II error") consists of failing to reject a false hypothesis.<sup>83</sup> Statistical power is the probability that a sampling procedure will generate a sample with statistically significant results, leading to a correct rejection of the false hypothesis, and therefore causing researchers to *avoid* a Type-II error. Therefore, when a study reports results that are not statistically significant, the higher its power for "detecting" a population parameter

---

79. See Freiman et al., *supra* note 76, at 361-62; Ware et al., *supra* note 65, at 195. Scientists use power determinations prior to undertaking a study in order to decide whether to undertake it at all or in order to determine the size of the sample to draw. CLAYTON & HILLS, *supra* note 54, at 209; COHEN & COHEN, *supra* note 31, at 60-61 & app. G.2; Ware et al., *supra* note 65, at 195-96.

80. This case study is presented in Deborah G. Mayo, *Sociological Versus Metascientific Views of Risk Assessment*, in ACCEPTABLE EVIDENCE: SCIENCE AND VALUES IN RISK MANAGEMENT 247, 261-75 (Deborah G. Mayo & Rachel D. Hollander eds., 1991).

81. See *id.* at 272.

82. *Id.* at 273.

83. COHEN & COHEN, *supra* note 31, at 166; CRANOR, *supra* note 72, at 32; FINKELSTEIN & LEVIN, *supra* note 34, at 81-88, 120-22; HAYS, *supra* note 34, at 282; Cohen, *supra* note 72, at 410-17; William E. Feinberg, *Teaching the Type I and Type II Errors: The Judicial Process*, 25 AM. STATISTICIAN 30 (1971), available at <http://links.jstor.org/sici?sici=0003-1305%28197106%2925%3A3%3C30%3ATTIAT%3E2.0.CO%3B2-Z>; Freiman et al., *supra* note 76, at 359. The probability of making a Type II error is often designated as  $\beta$ , and the probability of avoiding a Type II error (statistical power) as  $1-\beta$ . On inconsistency of notation, see Kaye & Freedman, *supra* note 26, at 125 n.144.

of  $X$ , the more likely it is that the study would have produced statistically significant results if  $X$  were in fact true. If the study had a low power to "detect"  $X$  and reject the hypothesis, then it is not surprising that the study failed to produce statistically significant results.

Unlike the scientific conventions for acceptable rates of Type I error, the decision rules for when the power of a study is acceptable are not as widely agreed among scientists. "Fairly high" power for clinical studies in the medical area is on the order of 0.95, and 0.9 is the usually accepted standard for clinical trials.<sup>84</sup> That is, for any true parameter value that is likely to have clinical significance, the probability of obtaining statistically significant results relative to the null hypothesis should be in the range of 0.9 to 0.95. Medical researchers want a high probability of positive results (statistically significant results) if the true value in the population is medically important. On the other hand, behavioral scientists are often forced to accept less power for their studies, which might be considered acceptable in the 0.7 to 0.9 range, with 0.8 often being accepted as adequate power.<sup>85</sup>

Scientists often provide "confidence intervals" as a convenient means of characterizing a number of the aspects of random sampling uncertainty. A confidence interval is constructed around a central value, which is usually the "maximum-likelihood estimate" (MLE) for the parameter. The MLE is the population value that maximizes the likelihood or probability of obtaining the actual sample, or the parameter value that makes the statistical results have the highest probability of occurrence.<sup>86</sup> Intuitively, the MLE is the "best bet" for the true parameter value because it is the value that, if it is true, would make the obtained sample results have the highest probability of being drawn.<sup>87</sup> A sampling distribution for the statistic is then constructed around the MLE value, instead of around an arbitrarily chosen hypothesis or around the null hypothesis. Moreover, just as significance testing requires selecting a level of statistical significance for rejecting an hypothesis, constructing confidence intervals requires selecting a degree of confidence in order to calculate the boundary limits for the interval. A 95% confidence interval is constructed using methods similar to those used in significance testing at the 0.05 level.<sup>88</sup>

Confidence intervals efficiently provide information about the potential for Type I errors. A confidence interval divides all possible parameter val-

---

84. See BLAND, *supra* note 65, at 161; Freiman et al., *supra* note 76, at 369.

85. COHEN & COHEN, *supra* note 31, at 162.

86. The "maximum-likelihood estimate" is the parameter value that "makes the occurrence of the actual result have greatest *a priori* likelihood." HAYS, *supra* note 34, at 211. It is "the hypothetical population value that maximizes the likelihood of the observed sample." WONNACOTT & WONNACOTT, *supra* note 60, at 568.

87. See generally HAYS, *supra* note 34, at 208-11; WONNACOTT & WONNACOTT, *supra* note 60, at 564-81.

88. The upper and lower boundaries or "limits" of the 95% confidence interval leave outside the interval the 5% least likely samples. On calculating confidence intervals for proportions, see HAYS, *supra* note 34, at 258-60; LOETHER & MCTAVISH, *supra* note 27, at 452-58; and WONNACOTT & WONNACOTT, *supra* note 60, at 273-74.



ues into two categories. As a close approximation, the sample results are statistically significant at the 0.05 level for all parameter values lying *outside* the 95% confidence interval.<sup>89</sup> The values lying *inside* the confidence interval, however, cannot be rejected at the 0.05 level on the basis of the sample.<sup>90</sup> For most practical purposes, therefore, a confidence interval can be used to characterize the statistical significance of the sample results for all possible parameter values.

With respect to avoiding Type II errors, a confidence interval also provides some information about the power of the sample. A very wide confidence interval suggests a small sample size and correspondingly low power.<sup>91</sup> With everything else equal, increasing sample size will increase statistical power and decrease the width of the confidence interval.<sup>92</sup> The width of the confidence interval provides some indication of the magnitude of difference the study is capable of detecting and, therefore, of the power to detect that the true value is not equal to the null hypothesis.<sup>93</sup>

Sampling uncertainty (the potential for error due to sampling) therefore has several aspects. A major potential for error is due to causal factors that influence the sampling process and cause it to produce samples that are biased in relation to the target population. Although strictly implemented sampling protocols can reduce this uncertainty, the factfinder should decide whether the risk of bias due to the sample-selection process is acceptable. Even if known biasing factors are controlled and randomization is built into the sampling protocol, there is still sampling uncertainty due to chance alone. The amount of random uncertainty in a study or sample can be divided into the complementary risks of using statistically significant results to reject a true hypothesis (a Type-I error) or relying on a lack of statistically significant results in not rejecting a false hypothesis (a Type-II error). Before relying on even a scientifically drawn random sample, the factfinder should decide whether the risk of either type of random sampling error is acceptable.

Thus, even when scientists have a sound basis for calculating statistical significance and statistical power, this does not eliminate the need to decide whether the residual sampling uncertainty is acceptable. Whether in the form of systematic bias or in the form of random difference, the risk re-

---

89. COHEN & COHEN, *supra* note 31, at 63; HAYS, *supra* note 34, at 221-24, 254-58; WONNACOTT & WONNACOTT, *supra* note 60, at 288-92. For a properly constructed 95% confidence interval, the probability is at least 0.95 that the interval actually covers the true population value. Over all possible samples of the same size, about 95% of these confidence intervals will include the true population value. If one of those samples is selected at random, the probability is 0.95 that its 95% confidence interval covers the true value. See LOETHER & MCTAVISH, *supra* note 27, at 455-56; WONNACOTT & WONNACOTT, *supra* note 60, at 254-59.

90. *E.g.*, WONNACOTT & WONNACOTT, *supra* note 60, at 288-318.

91. See LOETHER & MCTAVISH, *supra* note 27, at 503; SELVIN, *supra* note 76, at 177.

92. See, *e.g.*, HAYS, *supra* note 34, at 256-60.

93. See LOETHER & MCTAVISH, *supra* note 27, at 503. Constructing a confidence interval does not create better evidence of the true parameter. The empirical evidence still consists simply of the actual sample results, and any of these methodologies for characterizing sampling uncertainty is a way of presenting that evidence.

mains that the sample drawn is not representative of the target population. One danger with using scientific terminology and calculating mathematical probabilities is being misled into thinking that doing so eliminates decisions about accepting uncertainties.<sup>94</sup> Mathematical probabilities do not change the sampling evidence or change the fact that it is (merely) sampling evidence. Mathematical probabilities characterize the decision options in quantitative terms, but they do not eliminate the necessity for making those decisions or turn those pragmatic decisions into scientific issues of fact.

*C. Acceptable Modeling Uncertainty: Evaluating  
the Predictive Value of Variables*

The analysis in the two previous sections applies even to generalizations from data taken on single variables and, therefore, to findings involving only single variables. But, the major premise of a direct inference to specific causation states a causal relationship between two variables, *A* and *B*. The warrant for that premise is usually a statistical association between *A* and *B*, which warrants using information on the variable defining the reference group (*A*) to predict outcomes on another variable that identifies a subgroup (*B*). Scientists often develop models that express the values or statistics of variable *B* as a mathematical function of the values or statistics of variable *A*. Such models can help warrant predictions about one variable on the basis of information about the others.<sup>95</sup> Modeling uncertainty is the potential for error created by selecting a particular model.

Everyday life, science, and tort law are all concerned with determining risk. The public may be concerned with the risk of harm in commercial air travel, the risk of disease from exposure to chemicals in food or the environment, or the increased health risks for people who have identifiable genetic factors. The factfinder in a tort case often determines the existence and magnitude of such risks. Risk is in part a function of the expected rate of occurrence (incidence) of a harm in a group of individuals, such as the expected percentage of new injuries within some period of time or out of some

---

94. The technical terminology can also obscure the fact that only random sampling uncertainty is being characterized, not other types of uncertainty. See, e.g., Green, *supra* note 72, at 667-68, 681 (criticizing a court decision as "flat wrong in its notion that statistical significance or confidence intervals reflect anything about possible sources of error in an epidemiologic study other than sampling error").

95. The traditional use of regression analysis in the behavioral sciences was for making predictions, with only incidental attention to explanation or causal analysis. See COHEN & COHEN, *supra* note 31, at 41. Selvin describes two principal uses of mathematical models, contrasting two "types of mathematical structures" and referring to both as "models":

Both employ mathematical expressions to describe relationships within a set of data but with different goals. One attempts to reflect biological or physical reality, whereas the other is essentially a mathematical convenience to make predictions or to represent a set of relationships in a parsimonious way. . . . Models used in most statistical analyses are carefully constructed to reflect the observed data, providing a mathematically convenient way to deal with complex issues without detailed knowledge of underlying mechanisms.

SELVIN, *supra* note 76, at 37.

number of exposed people.<sup>96</sup> A standard study design for scientific investigations into risk is the controlled experimental study, which can furnish evidence that one variable is a risk factor for another variable.<sup>97</sup> These variables might be either categorical or quantitative. Examples of categorical variables are ingesting or not ingesting a specified drug, and developing or not developing a particular disease. Examples of quantitative variables are level of exposure and kidney dysfunction as measured by increased excretion of a protein in the urine.<sup>98</sup> In a controlled study, a sufficiently large number of relatively homogeneous subjects are randomly assigned to the test and control groups, and they are carefully monitored so that all likely causal factors are controlled physically.<sup>99</sup> If over the course of the study there is a statistically significant difference in disease incidence between the test and control groups, then the study provides some evidence that the exposure being tested is a risk factor for the disease.

Such a controlled experimental design may furnish a paradigm for how to conduct a study about risk, but scientists cannot always perform such studies. A controlled study may not be methodologically feasible, as when the subjects' behavior must be studied in natural circumstances, not in the laboratory. It may not be economically feasible if a large, long-term laboratory study is needed to study a rare disease, and it may not be ethically ac-

---

96. In epidemiologic terminology, "incidence" is the number of new cases of a disease occurring in some group during a specified period of time, while "prevalence" is the number of cases present in the group at a specified time. See, e.g., LILIENFELD & LILIENFELD, *supra* note 53, at 139. Those authors state that the "incidence rate is a direct estimate of the probability, or risk, of developing a disease during a specified period of time." *Id.* A broader conception of "risk" would include not only expected incidence, but all of the epistemic uncertainty associated with expected incidence. For a discussion of the evaluative judgments involved in risk assessment that are not acknowledged in the standard definition of "risk," see K.S. SHRADER-FRECHETTE, *RISK AND RATIONALITY: PHILOSOPHICAL FOUNDATIONS FOR POPULIST REFORMS* 58-63 (1991). On the need to re-conceive risk characterization as more than the summary of a technical process, see NAT'L RESEARCH COUNCIL, *UNDERSTANDING RISK: INFORMING DECISIONS IN A DEMOCRATIC SOCIETY* (Paul C. Stern & Harvey V. Fineberg eds., 1996).

97. A risk factor is a variable whose values or categories are statistically associated with an increased incidence of disease or injury in a population. The etiologic or causal agent for the disease or injury may be unknown. See, e.g., LILIENFELD & LILIENFELD, *supra* note 53, at 259. What is known is that having a certain score on the variable warrants a prediction that the occurrence of the disease or injury is more likely. An assertion that a variable is a risk factor usually also suggests that there exists some causal connection between the risk factor and the dependent variable, even if the precise causal relationship is unknown. See *id.*

98. For example, worker exposure to cadmium might be measured by the amount of cadmium in the workers' blood or urine, or by linking job histories with air measurements of cadmium in various work departments over time. Kidney dysfunction in workers might be measured by the concentrations of  $\beta$ -2-microglobulin in their urine. See Michael Thun, *Kidney Dysfunction in Cadmium Workers*, in *CASE STUDIES IN OCCUPATIONAL EPIDEMIOLOGY* 105-26 (Kyle Steenland ed., 1993).

99. For a discussion of a controlled study design for toxicological investigations, especially using laboratory test animals, see Norton Nelson, *Toxicology and Epidemiology: Strengths and Limitations*, in *EPIDEMIOLOGY AND HEALTH RISK ASSESSMENT* 37 (Leon Gordis ed., 1988). For discussions of controlled study designs in clinical trials involving human subjects, see BLAND, *supra* note 65, at 6-25; CHARLES H. HENNEKENS & JULIE E. BURING, *EPIDEMIOLOGY IN MEDICINE* 178-212 (1987); LILIENFELD & LILIENFELD, *supra* note 53, at 256-72; Philip W. Lavori et al., *Designs for Experiments—Parallel Comparisons of Treatment*, in *MEDICAL USES OF STATISTICS* 61-82 (John C. Bailar III & Frederick Mosteller eds., 2d ed. 1992); and Lincoln E. Moses, *Statistical Concepts Fundamental to Investigations*, in *MEDICAL USES OF STATISTICS* 5-25 (John C. Bailar III & Frederick Mosteller eds., 2d ed. 1992).

ceptable if it requires experimentation on healthy human beings. Although health care researchers can sometimes approximate such a controlled protocol in a clinical study on medical patients, they often must resort to epidemiologic studies. An epidemiologic study draws its data from observations of human beings in natural and uncontrolled settings.<sup>100</sup> In lieu of a test group and a control group, an epidemiologic or observational study will identify a study group and a comparison group, based on differences in exposure.<sup>101</sup> However, it is not possible to assign subjects randomly to exposed or unexposed groups, or to ensure that such independent variables as genetics, diet, and environment are physically controlled and uniform for all subjects. At best, researchers must identify potentially relevant variables, take measurements on those variables for each subject, and search for statistical relationships among those variables. If such statistical relationships can be generalized from the study sample to the population, then they might warrant predictions for some variables based on other variables, provided the factfinder can take the inherent uncertainties into account.

The kinds of causal chains of importance in tort law often involve many variables (such as genetic, developmental, and environmental factors) that interact in complicated ways. A study design might exclude from the sample types of individuals who have complicating factors (such as certain diseases or genetic histories). A stratified random sampling protocol can ensure that the stratification variables are proportionally reflected in the sample.<sup>102</sup> Researchers might physically control certain variables (such as diet or environmental conditions), ensuring that these variables have identical values for both test and control groups. To the extent, however, that such direct techniques are unavailable, researchers can "statistically control" relevant variables by gathering data upon them and using models to take them into account when calculating risk.<sup>103</sup> The potential for predictive error that is due to leaving important variables out of account in modeling is a first major source of modeling uncertainty.

For models to be useful, statistics must characterize the strength of association and predictive power between studied variables in a way that supports causal interpretations. Relative risk (*RR*) is one statistic used to characterize the statistical association between two variables. Relative risk is useful when the predictor variable (such as exposure) and the injury variable are categorical variables. A general format for characterizing relative risk is a "2 x 2 table," as in Table 1. The values in the cells of this table (namely, *w*, *x*, *y*, and *z*) are the frequencies or numbers of individuals in the study who satisfy that combination of classification categories. For example, *w* is the number of individuals who are in the study group (for example, people exposed to an environmental agent) and who suffer the relevant injury dur-

---

100. LILIENTHAL & LILIENTHAL, *supra* note 53, at 4.

101. *See id.* at 191-255.

102. *See supra* text accompanying notes 52-59.

103. *E.g.*, KAHN & SEMPOS, *supra* note 52, at 85-113.

ing the study period. The value  $y$  is the number of individuals in the comparison group (unexposed) who suffer the injury. The values in the other two cells of the table reflect similar calculations. The table therefore displays a frequency distribution over the combinations of categories for the two categorical variables.<sup>104</sup> The frequency statistics in this table can be used to calculate the incidence of the injury in the study group,  $w / (w + x)$ , and in the comparison group,  $y / (y + z)$ . These rates are obviously affected by any measurement errors in classifying individuals in the study<sup>105</sup> and by the sample of individuals in the study.

TABLE 1

	<u>Injury or Dependent Variable:</u>		<u>Totals:</u>
	<u>Yes</u>	<u>No</u>	
<u>Predictor or Independent Variable, used to identify:</u>			
<u>Study Group (Exposed):</u>	w	x	w+x
<u>Comparison Group (Unexposed):</u>	y	z	y+z
<u>Totals:</u>	w+y	x+z	w+x+y+z

The relative risk is the ratio of the injury incidence in the study group (exposed group) to the injury incidence in the comparison group (unexposed group).<sup>106</sup>

104. Epidemiologists usually list the exposure or independent variable down the left (vertical) side of the table and the disease or injury variable across the top (horizontally). See, e.g., KAHN & SEMPOS, *supra* note 52, at 45-50; SELVIN, *supra* note 76, at 345. On the other hand, the table can be set up with the dependent variable down the left side and the independent variable across the top. See, e.g., LOETHER & MCTAVISH, *supra* note 27, at 165-76 (the tradition in sociology is to use a dependent variable as a row variable, down the stub or side of the table).

105. Measurement validity for dependent or independent variables might be adversely affected by the knowledge of those taking the measurements about whether particular study subjects are in the experimental group or the control group. When such measurement bias is a possibility, standard protocols include single-blind studies (the primary observer cannot tell whether the subjects are in the experimental group or the control group) and double-blind studies (both subjects and observing researchers do not know the subject's group membership). See, e.g., BLAND, *supra* note 65, at 20-22; LILIENTHAL & LILIENTHAL, *supra* note 53, at 265; SELVIN, *supra* note 76, at 49; DerSimonian et al., *supra* note 76, at 342-44.

106. For definitions of relative risk, see Green et al., *supra* note 6, at 348-49; KAHN & SEMPOS, *supra* note 52, at 45-47; and LILIENTHAL & LILIENTHAL, *supra* note 53, at 209-16.

$$w / (w + x)$$

$$RR = \frac{w / (w + x)}{y / (y + z)} .$$

For example, in a study about the risks due to exposure to a particular chemical, relative risk would compare the rate of injury among those exposed with the rate among those not exposed.<sup>107</sup> If exposure to the chemical is not associated in the study with an increased incidence of injury (that is, if the incidences in the two groups are identical), then the relative risk is equal to one. There is no statistical association between the predictor and injury variables in the study sample. If a higher incidence of injury is observed in the study group than in the comparison group, the relative risk is greater than one. For example, a doubling of incidence in the exposure group would result in a relative risk of two. A lower incidence of injury would be reflected in a relative risk less than one.

Relative risk in a sample can be used to make predictions about association in the relevant population, provided sampling uncertainty is taken into account. The null hypothesis is that there would be no statistical association in the population if every individual could be measured on the two variables. That is, the null hypothesis is that *RR* in the population would have a value of one. Even if a population does have *RR* = 1, samples selected on the basis of chance alone could still exhibit relative risks not equal to one. In order to allow for sampling uncertainty, researchers conduct significance testing for *RR* using the same principles discussed in the previous section. It might be very unlikely that the *RR* actually observed in the sample would be drawn randomly from a population with a *RR* = 1. If the sampling is conducted in a suitably random manner, statisticians can identify a sampling

---

107. An alternative statistic used to characterize the degree of association is the odds ratio, defined using the notation of Table 1 as  $(w/x)/(y/z)$ , or the mathematically equivalent cross-product form  $wz/xy$ . See, e.g., FINKELSTEIN & LEVIN, *supra* note 34, at 37-38; KAHN & SEMPOS, *supra* note 52, at 51-54; Leon Gordis, *Estimating Risk and Inferring Causality in Epidemiology*, in EPIDEMIOLOGY AND HEALTH RISK ASSESSMENT 51, 51-52 (Leon Gordis ed., 1988). As a general principle, if the incidence rate of injury is relatively small (*w* is small relative to *x*, and *y* relative to *z*), then the odds ratio will closely approximate the relative risk. See FINKELSTEIN & LEVIN, *supra* note 34, at 37-38; KAHN & SEMPOS, *supra* note 52, at 55; LILIENTHAL & LILIENTHAL, *supra* note 53, at 209-10. In case-control studies, where the study group is identified as having an injury or disease and a comparison group is a suitably matched group, the odds ratio provides a stable and unbiased estimate of the relative risk. HENNEKENS & BURING, *supra* note 99, at 79-81.

Another advantage of the odds ratio is that it avoids the dependence of relative risk on the choice of reference class. A low injury rate (such as 5% of exposed cases, 1% of unexposed cases) may translate into a high relative risk (five times the unexposed risk). If the same percentages, however, are used to calculate non-injury cases (95% for exposed, 99% for unexposed), the "relative safety" seems reasonably high  $(.95/.99 = 95.95\%)$ . The odds ratio components, however, remain the same whether "risk" or "safety" is being described, only the numerator and denominator are reversed: the odds ratio of injury is equal to  $99/19$  (that is,  $(5/95)/(1/99)$ ) and the odds ratio of non-injury is equal to  $19/99$  (or  $(95/5)/(99/1)$ ). See, e.g., FINKELSTEIN & LEVIN, *supra* note 34, at 37-38; KAYE & FREEDMAN, *supra* note 26, at 109-10. Thus, an odds ratio is less subject to rhetorical manipulation.

distribution for  $RR$ , and can use that distribution and the selected level of significance to find the critical value for  $RR$  in the study. If the sample  $RR$  is more extreme than that critical value, then the study results are statistically significant for purposes of rejecting the null hypothesis.<sup>108</sup> If, however, the sample results are not statistically significant, the convention on statistical significance does not warrant rejecting the null hypothesis. Confidence intervals can also be constructed for  $RR$ . Suppose that the sample has a  $RR = 2.5$  and that the 95% confidence interval based on the sample is 1.7-3.7.<sup>109</sup> This means that, using a convention of statistical significance at the 0.05 level, the sample results are consistent with values for the population  $RR$  ranging from 1.7 to 3.7. Since the null hypothesis of  $RR = 1$  lies outside the specified confidence interval, a factfinder would be warranted in rejecting the null hypothesis as probably not true. That is, he would be warranted in concluding that the sample  $RR$  of 2.5 probably did not result merely by chance from a population with  $RR = 1$ .

Relative risk is used as a predictive model for the relative frequency of events within groups of individuals—for example, to predict the number of injuries in a group of people exposed to a chemical. Suppose that, with measurement and sampling uncertainties at acceptable levels, the real relative risk for exposed groups in the population is 2.5 and that about ten injuries would occur per 100,000 people without exposure. The relative risk model predicts that the rate for an exposed group would be approximately twenty-five per 100,000, or an increase of fifteen cases per 100,000 people exposed. The proportional increase in the exposed group is generally called the “attributable risk”—an estimate of the proportion or number of injuries in the exposed group that might be attributed to the exposure and not to baseline causes.<sup>110</sup> Attributable risk therefore estimates that proportion of

---

108. For discrete data, such as that presented in the 2 x 2 table in Table 1, the chi-square test is commonly used to determine whether the difference in incidence rates between the study and comparison groups is statistically significant. See, e.g., HENNEKENS & BURING, *supra* note 99, at 249-52.

109. This example, with calculations, can be found in Vern R. Walker, *The Concept of Baseline Risk in Tort Litigation*, 80 KY. L.J. 631, 661 n.92 (1991-1992). For discussions on how to construct confidence intervals for relative risk or the odds ratio, see HENNEKENS & BURING, *supra* note 99, at 252-58; KAHN & SEMPOS, *supra* note 52, at 45-69; LILIENFELD & LILIENFELD, *supra* note 53, at 343-46; and SELVIN, *supra* note 76, at 344-47 (odds ratio).

110. Attributable risk is defined as the difference between the incidence rates in exposed and unexposed groups (incidence rate if exposed minus incidence rate if not exposed), as a proportion of the incidence rate in the exposed group (the “attributable fraction” or “rate fraction”). Green et al., *supra* note 6, at 351-52; Sander Greenland, *Relation of Probability of Causation to Relative Risk and Doubling Dose: A Methodologic Error That Has Become a Social Problem*, 89 AM. J. PUB. HEALTH 1166, 1167 (1999). Attributable risk is also defined as “the maximum proportion of a disease that can be attributed to a characteristic or etiological factor,” LILIENFELD & LILIENFELD, *supra* note 53, at 217, and sometimes as “the difference between the incidence rates in the exposed and nonexposed groups.” HENNEKENS & BURING, *supra* note 99, at 87-95.

It is a fallacy, however, simply to equate attributable risk and probability of causation, determining the latter requires “a specific biologic model for the disease process.” See Greenland, *supra*, at 1167. Unfortunately, some courts too readily interpret attributable risk as causal. E.g., Merrell Dow Pharm., Inc. v. Havner, 953 S.W.2d 706, 721 (Tex. 1997) (citing the “attributable proportion of risk” as “[p]erhaps the most useful measure,” saying that “it reflects the percentage of the disease or injury that could be prevented by eliminating exposure to the substance”).

cases in the exposed group that is in excess of baseline, assuming that the incidence in the comparison group is a good estimate of the baseline incidence.<sup>111</sup>

Leaving important variables out of account may produce only a crude estimate of true relative risk, and taking those variables into account might increase or decrease the “refined” or “adjusted” relative risk.<sup>112</sup> Regression models refine or adjust statistical measures of association by explicitly taking multiple variables into account, both categorical and quantitative variables.<sup>113</sup> The predicted variable is generally called the “dependent variable.” An “independent variable” is any variable used to make the prediction, and is therefore part of the evidence that warrants the prediction. Particular fields of science may have different terminology for this same evidentiary relationship. For example, the medical literature often refers to the dependent variable as the “outcome variable” or “response variable” and to an independent variable as a “predictor variable.”<sup>114</sup> The terminology used to describe regression models, however, should not have a causal connotation. Causal explanations impose causal interpretations on predictive models of the sort discussed in the next section. This section discusses the merely associational models that support predictions and the uncertainty inherent in using such models.

A simple model involving only two variables will introduce the basic concepts underlying regression models. Assume that a study analyzes data for a pair of quantitative variables, such as a quantitative measure of exposure to a chemical and the concentration of protein in a person’s urine. In a given context, one of these variables is the independent or predictor variable, and the other is the dependent or predicted variable. For each individ-

---

111. KAHN & SEMPOS, *supra* note 52, at 72-81.

112. *E.g.*, *id.* at 85-113 (discussing techniques for adjusting odds ratios without using multivariate models). For discussions on the effect on relative risk, see *infra* notes 160-64, 176-80, 227 and accompanying text.

113. For discussions of linear regression models in various areas of science, see WILLIAM D. BERRY & STANLEY FELDMAN, *MULTIPLE REGRESSION IN PRACTICE* (1985); BLAND, *supra* note 65, at 188-211 (medicine); COHEN & COHEN, *supra* note 31, at 68-71 (behavioral sciences); HAYS, *supra* note 34, at 597-809 (experimental psychology); LOETHER & MCTAVISH, *supra* note 27, at 230-48, 314-57 (sociology); LARRY D. SCHROEDER ET AL., *UNDERSTANDING REGRESSION ANALYSIS: AN INTRODUCTORY GUIDE* (1986); and Katherine Godfrey, *Simple Linear Regression in Medical Research*, in *MEDICAL USES OF STATISTICS 201-35* (John C. Bailar III & Frederick Mosteller eds., 2d ed.1992) (focusing on examples from articles published in the *New England Journal of Medicine*).

For discussions of regression analysis in a legal context, see FINKELSTEIN & LEVIN, *supra* note 34, at 350-479; Michael O. Finkelstein, *The Judicial Reception of Multiple Regression Studies in Race and Sex Discrimination Cases*, 80 COLUM. L. REV. 737 (1980); Michael O. Finkelstein, *Regression Models in Administrative Proceedings*, 86 HARV. L. REV. 1442 (1973) [hereinafter *Regression Models*]; Franklin M. Fisher, *Multiple Regression in Legal Proceedings*, 80 COLUM. L. REV. 702 (1980); and Daniel L. Rubinfeld, *Reference Guide on Multiple Regression*, in *REFERENCE MANUAL ON SCIENTIFIC EVIDENCE* 179 (Fed. Judicial Center ed., 2d ed. 2000), available at [http://www.fjc.gov/public/pdf.nsf/lookup/sciman00.pdf/\\$file/sciman00.pdf](http://www.fjc.gov/public/pdf.nsf/lookup/sciman00.pdf/$file/sciman00.pdf). For cases involving regression analysis, see *McCleskey v. Kemp*, 481 U.S. 279 (1987) (death penalty); *Bazemore v. Friday*, 478 U.S. 385 (1986) (employment discrimination); and *Campos v. City of Baytown*, 840 F.2d 1240 (5th Cir. 1988) (voting dilution), *cert. denied*, 488 U.S. 1002 (1989).

114. See, *e.g.*, BLAND, *supra* note 65, at 190; Godfrey, *supra* note 113, at 201-02.



ual involved in the study, there is a measurement value for each of the two variables. A graph can chart the possible values of the independent variable along the horizontal axis and the possible values of the dependent variable along the vertical axis, as in Figure 1. The hypothetical data for Figure 1 are similar to those in a study of kidney dysfunction in workers exposed to cadmium.<sup>115</sup> An individual's exposure is measured in milligrams of cadmium per cubic meter of air, multiplied by the number of work days exposed to such an air concentration.<sup>116</sup> The level of kidney dysfunction is measured by the small protein  $\beta$ -2-microglobulin in each worker's urine, as a ratio to serum creatinine.<sup>117</sup> A point is entered on the graph for each individual, using the value on each variable to determine the appropriate location within the area of the graph. This produces a "scatterplot" depicting all individuals and their measurements—a two-dimensional map locating every individual on the scales of the two variables. The dashed horizontal line, with a value of about 700 on the vertical axis, is the arithmetic mean for the values of the dependent variable ( $\beta$ -2-microglobulin in urine).

The values of the independent variable (exposure) could be used to predict the values of the dependent variable. A mathematical model is linear if the formula used to make those predictions identifies a straight line.<sup>118</sup> Figure 2 shows such a straight line drawn through the scatterplot. Moving along the line from the lower left corner of the graph to the upper right corner, the values of the independent and dependent variables both increase. Because the line is straight, these values for the line increase at a constant rate relative to each other.<sup>119</sup> That is, for each unit of change in the independent variable, there is a constant amount of change in the dependent variable, all along the line.<sup>120</sup>

---

115. This simplified example with hypothetical data is derived from Thun, *supra* note 99, at 105-26.

116. *Id.* at 105-26.

117. *Id.*

118. See LOETHER & MCTAVISH, *supra* note 27, at 232-34.

119. See, e.g., COHEN & COHEN, *supra* note 31, at 27; *Regression Models*, *supra* note 113, at 1448; Godfrey, *supra* note 113, at 202.

120. See COHEN & COHEN, *supra* note 31, at 27.

FIGURE 1

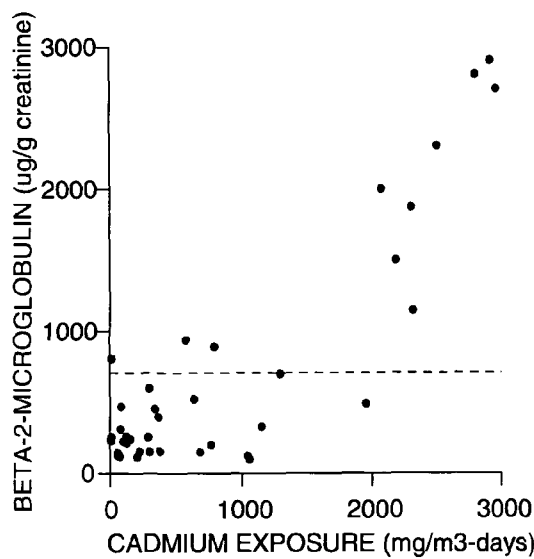
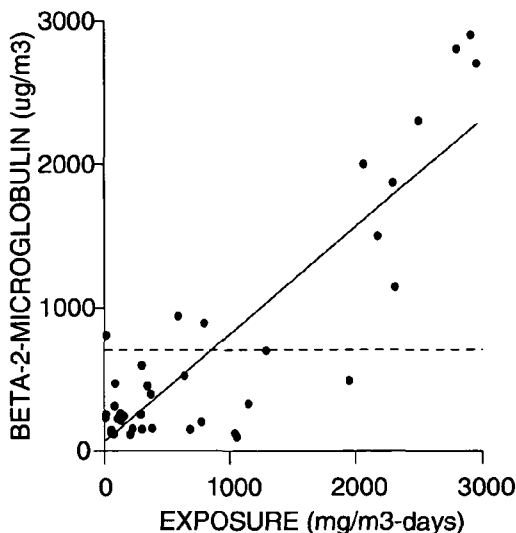


FIGURE 2



For every straight line that can be drawn through the scatterplot of data, there is an algebraic formula that calculates the point on the line associated with any value of either variable.<sup>121</sup> A straight line identifies the values of the dependent variable as some constant multiple of the independent variable, adjusted by another constant, as shown by the following general algebraic formula:<sup>122</sup>

$$Y = A + BX.$$

The constant  $A$  is the point on the vertical axis where the line intersects that axis; that is, it is the  $Y$ -value on the line when  $X = 0$ . The constant  $A$  is called the “ $Y$ -intercept” because it is the point where the line crosses the  $Y$  axis (vertical axis).<sup>123</sup> The constant  $B$  is the rate of change in the  $Y$ -value on the line for every unit of change in the  $X$ -value.<sup>124</sup> In Figure 2,  $B$  is the constant quantity by which the line rises vertically as the value of  $X$  increases

121. See LOETHER & MCTAVISH, *supra* note 27, at 232-34.

122. See *id.*

123. See *id.* at 233.

124. See *id.*

horizontally. The constant  $B$  is therefore called the “slope” of the line.<sup>125</sup> Given these two constants, a point of  $Y$ -axis intercept and a constant rate of change, the resulting model identifies a unique straight line through the scatterplot. Each choice of different values for either  $A$  or  $B$  defines a different straight line through the data.

Such a linear model can be used to make predictions about risk—that is, about rates of occurrence within groups of individuals. To predict a value for the dependent variable based on a value for the independent variable, one would find the relevant point on the independent axis ( $X$ ), find or calculate the point on the line directly above that point, and use the  $Y$  coordinate for that point on the line as the predicted value for the dependent variable. In the kidney dysfunction example in Figure 2, an exposure of 2000 mg/m<sup>3</sup>-days would lead to a prediction (based on the displayed linear model) of about 1700 µg of β-2-microglobulin per gram of creatinine in a worker’s urine. Predictive error would occur whenever an individual’s true value on the dependent variable is different than the value given by the prediction line.<sup>126</sup> When a regression model has exposure as an independent variable and injury as a dependent variable, then predicted value can be interpreted as an estimate of risk based on exposure. Obviously, the prediction line will generate the same predicted value for any one value of  $X$ . The model will have predictive error to the extent that there is variability among the real  $Y$  values for different individuals who have the same  $X$  value.

The amount of residual predictive error can be minimized by carefully choosing the prediction line.<sup>127</sup> Some straight lines through the data points are better predictors than others. Scientists minimize predictive error by using the line that is identified by the “least squares” technique.<sup>128</sup> The method of least squares identifies the particular line with the least predictive error around it, using as the measure of predictive error the average (arithmetic mean) of the squared differences between the individual values and the mean value.<sup>129</sup> This line is called the “linear regression line,”<sup>130</sup> and its algebraic form is simply a straight line:

$$Y_i = A + B(X_i),$$

125. *Id.*

126. If there is measurement error, then the observed  $Y$ -value for an individual might not lie on the prediction line, whereas the real  $Y$ -value does. There would be no modeling error, but this fact would be masked by the measurement error. Conversely, due to measurement error, the observed  $Y$ -value might lie on the prediction line, but the real value does not. In such a case, the model would not be as accurate in predicting real  $Y$ -values as it appears.

127. See BLAND, *supra* note 65, at 191; LOETHER & MCTAVISH, *supra* note 27, at 238, 246-47.

128. See BLAND, *supra* note 65, at 191; LOETHER & MCTAVISH, *supra* note 27, at 238, 246-47.

129. See BLAND, *supra* note 65, at 191-94; COHEN & COHEN, *supra* note 31, at 42-43, 50, 77; FINKELSTEIN & LEVIN, *supra* note 34, at 358-61; LOETHER & MCTAVISH, *supra* note 27, at 246-48; SCHROEDER ET AL., *supra* note 113, at 17-23.

130. LOETHER & MCTAVISH, *supra* note 27, at 233.

where  $X_i$  is the value of the independent variable  $X$  for any particular individual  $i$ , and  $Y_i$  is the predicted value of the dependent variable  $Y$  for that same individual.<sup>131</sup> The constant  $A$  is now called the "regression constant," which is the predicted value of  $Y$  when  $X = 0$ .<sup>132</sup> The constant  $B$  is called the "regression coefficient," the constant amount by which the predicted value of  $Y$  increases (or decreases) for every unit of increase in  $X$ .<sup>133</sup> Using least squares as the measure of predictive error, the regression line is, by definition, the best linear predictor for the dependent variable within a given set of data. Of course, even the regression line can still have considerable dispersion around it, and it might be a very poor predictor for  $Y$ . However, it is still the *best* linear predictor, in the sense that any other straight line through these data, on these two variables, would have even more dispersion around it and would lead to even more predictive error.

It is therefore important to ask how good the regression line is as a predictive model for a set of data, or how well the model "fits" the data. Statisticians construct direct measures of dispersion around the regression line.<sup>134</sup> However, scientists usually find more useful one or more indirect measures of how well the model fits the data, such as comparing the predictive performance of the regression line to that of the arithmetic mean of the dependent variable. For example, to predict the heights of individuals in a group of people with the least amount of signed error on average, one would make predictions using the arithmetic mean of the heights of the people in the group.<sup>135</sup> Over many such predictions for randomly drawn individuals, the expected error would be lower using the arithmetic mean than using any other single value. But suppose it is known how each individual in the group scored on a second variable, such as weight or age. Should the information about that second variable be used to predict height, or should one continue simply to use the group mean for height and ignore the second variable? Would using the regression line as the predictor decrease the predictive error as compared to using the mean? The answer is that predictions should be based on the regression line if the error around it (measured as

---

131. There exists some minor ambiguity in this notation. The individual denoted by  $i$  may be unique, but that individual's  $X$ -score ( $X_i$ ) is not, nor is the predicted  $Y$ -score ( $Y_i$ ). Therefore,  $X_i$  denotes a classification category on the  $X$  variable, and  $Y_i$  denotes the classification category on the  $Y$  variable that corresponds to the point  $\{X_i, Y_i\}$  on the regression line.

The regression equation is sometimes written to yield the observed values of individuals, which may not lie directly on the regression line, but rather above or below the line by some "error value"  $e_i$ :

$$Y_i = A + B(X_i) + e_i.$$

WONNACOTT & WONNACOTT, *supra* note 60, at 373; 400-01.

132. COHEN & COHEN, *supra* note 31, at 42.

133. *Id.* at 11-12, 41-44; LOETHER & MCTAVISH, *supra* note 27, at 234-39; SCHROEDER ET AL., *supra* note 113, at 11-17; Godfrey, *supra* note 113, at 202-15.

134. See, e.g., COHEN & COHEN, *supra* note 31, at 46-49, 126-30, 354-55. Two direct measures of residual modeling uncertainty are the variance of residual error (or "variance of residuals") and the standard deviation of residual error (or "standard deviation of residuals"). See *id.* The variance of residual error is the average of the squared differences between observed scores and predicted scores, and the standard deviation of residual error is the square root of that average. *Id.*

135. For any given group, the amount of signed error around the arithmetic mean is zero. HAYS, *supra* note 34, at 172-73.

average squared difference from the mean) is less than the variance of the dependent variable alone (that is, the average squared difference around that variable's mean).

The coefficient of determination, or  $r^2$ , reports the proportion of average predictive error that can be eliminated by using the regression line, as compared to using simply the mean of the dependent variable.<sup>136</sup> It is the proportion of the variance of the dependent variable that can be eliminated by using the regression line as the predictor.<sup>137</sup> The coefficient  $r^2$  ranges from zero to one.<sup>138</sup> When  $r^2 = 0$ , the regression line provides no more predictive success than the mean alone.<sup>139</sup> If  $r^2 = 1$ , then there is no predictive error using the regression line: that is, all the data points fall precisely on the regression line. As  $r^2$  increases from zero to one, it is a useful index of scatter reduction and predictive success, because it is the proportion of variance eliminated by using the regression line. In that sense, it is also a measure of the strength of linear statistical association between the two variables.

The square root of  $r^2$  (simply " $r$ ") is Pearson's correlation coefficient.<sup>140</sup> Like  $r^2$ ,  $r$  ranges in absolute value from zero to one:  $r = 0$  if there is no correlation between the variables, and  $r = 1$  if there is perfect correlation.<sup>141</sup> So like  $r^2$ ,  $r$  measures the degree or strength of correlation, and  $r$  obtains its intuitive meaning through  $r^2$ .<sup>142</sup> For example, an  $r = 0.5$  corresponds to an  $r^2 = 0.25$ , which in turn means that 25% of the variance of the dependent variable is eliminated using the regression line. However, one advantage of  $r$  over  $r^2$  is that  $r$  can be positive or negative, ranging from negative one (perfect negative or inverse correlation) to positive one (perfect positive correlation).<sup>143</sup>

It is one thing to know what these statistical measures mean, and another to appreciate when the results have practical significance. Is reducing the squared error by 25% a good performance for a mathematical model? The reasonable expectations of scientists vary depending on the context. For example, in the behavioral sciences, where many factors can influence human behavior and researchers seldom expect any single independent variable to be a very good predictor, an  $r = 0.1$  is sometimes considered a small

136. *Id.* at 613-14; LOETHER & MCTAVISH, *supra* note 27, at 239-41; SCHROEDER ET AL., *supra* note 113, at 26; *Regression Models*, *supra* note 113, at 1448-53.

137. HAYS, *supra* note 34, at 614.

138. WONNACOTT & WONNACOTT, *supra* note 60, at 487.

139. *See id.*

140. HAYS, *supra* note 34, at 608-13; LOETHER & MCTAVISH, *supra* note 27, at 239-40. The correlation coefficient is different from, but related to, the regression coefficient. HAYS, *supra* note 34, at 608-13. The correlation coefficient is a symmetrical measure of association between the two variables, while the regression coefficient is asymmetrical because it predicts the dependent variable using the independent variable. COHEN & COHEN, *supra* note 31, at 34-44. The value of the correlation coefficient would be identical to that of a regression coefficient calculated for standardized scores of the two variables. *Id.* at 30-36; *see* HAYS, *supra* note 34, at 608-11; LOETHER & MCTAVISH, *supra* note 27, at 241-46; SCHROEDER ET AL., *supra* note 113, at 28-29; Godfrey, *supra* note 113, at 215-17.

141. HAYS, *supra* note 34, at 608-13; LOETHER & MCTAVISH, *supra* note 27, at 239-40.

142. HAYS, *supra* note 34, at 608-14; LOETHER & MCTAVISH, *supra* note 27, at 239-40.

143. *See* LOETHER & MCTAVISH, *supra* note 27, at 240.

degree of correlation,  $r = 0.3$  a medium degree, and  $r = 0.5$  a large degree.<sup>144</sup> So, for behavioral studies, a population  $r^2 = 0.25$  indicates quite a good model fit. Predictive expectations in medicine can be higher, and scientists in the physical sciences dealing with relatively simple phenomena might expect and achieve even better fits from their models. Ultimately, the question of whether the "goodness of fit" of a mathematical model is acceptable depends upon what is at stake in the particular pragmatic context. Even in legal factfinding, the fit of a model might be "good enough" for some purposes (such as regulatory measures to protect public health) but not for others (such as imposing a fine on a defendant). The acceptability of model fit in factfinding about specific causation in tort cases is only one particular application within law.

So far this discussion of regression analysis has addressed only the modeling uncertainty that is due to the choice of a particular *straight* line as the prediction line. A particular line is chosen when values are assigned to the two constants  $A$  and  $B$  in the mathematical formula  $Y = A + BX$ . Such modeling error is minimized by finding and using the values for  $A$  and  $B$  that produce the regression line. The correlation coefficient for a linear regression model measures the improvement in predictive success using the best *linear* model. A low value for  $r$ , for example, means that the best linear model does not substantially improve predictive success over the simple mean. But such a low value for  $r$  does not evaluate another kind of modeling uncertainty—the uncertainty created by the selection of the mathematical form to be used. It may be that for a particular set of data a *nonlinear* prediction line would yield less predictive error than any straight line.<sup>145</sup> Even if  $r$  is low for the linear regression model, there may still be a statistical association between the two variables, only the association is not a linear one.<sup>146</sup> For example, for a particular data set, a quadratic equation of the form  $Y = A + BX + CX^2$  might be a better fitting model than any linear equation.<sup>147</sup> Figure 3 illustrates a nonlinear model. Regression techniques can be used to find best-fitting nonlinear models, and the kinds of issues discussed above for linear models have their counterparts with nonlinear models.<sup>148</sup>

---

144. COHEN & COHEN, *supra* note 31, at 59-61.

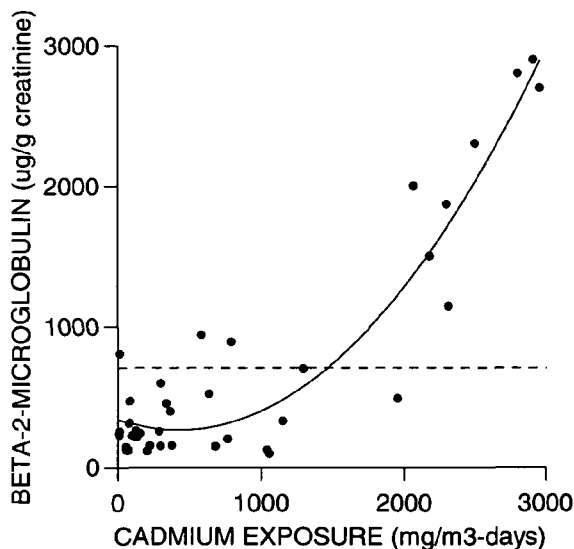
145. LOETHER & MCTAVISH, *supra* note 27, at 240-41, 247.

146. *Id.*

147. BERRY & FELDMAN, *supra* note 113, at 60-64.

148. *Id.* at 9-18, 51-64; COHEN & COHEN, *supra* note 31, at 223-74; WONNACOTT & WONNACOTT, *supra* note 60, at 449-51.

FIGURE 3



The predictive uncertainty due to the decision to use a particular form of mathematical model is therefore a second major type of modeling uncertainty. Every mathematical model has formal assumptions or conditions to be met, analyzes the observational data in certain ways, and produces statistics that have a potential for predictive error when applied.<sup>149</sup> Before relying upon any modeling statistics, a reasonable factfinder would evaluate the evidence that warrants that the formal conditions are sufficiently satisfied, that the resulting statistics are appropriately applied, and that the residual predictive uncertainty is within acceptable bounds for the legal purposes at hand.

The kind of predictive error measured by  $r^2$  and  $r$  is only the uncertainty due to mathematical modeling, for these statistics characterize the decreased variability due to using the regression line. They do not address the measurement uncertainty inherent in the data, because the mathematical model uses the data (scatterplot points) as given.<sup>150</sup> In addition, sampling uncer-

149. For a discussion of formal assumptions or conditions for applying regression analysis, see KAHN & SEMPOS, *supra* note 52, at 140-43. See also *infra* note 172.

150. However, linear regression analysis is commonly used to assess the validity of measurement techniques. There may be a criterion or reference method for taking a measurement on a variable, such as a laboratory method for measuring blood glucose. If a new, more convenient, and less costly measurement technique is developed, medical scientists will want to know how well the measurements using



tainty only occurs if a statistical association observed in a sample is used to make inferences about the statistical association in the population. When experts wish to generalize from study samples to populations, they can conduct significance testing for hypotheses about the population coefficient of determination and the population correlation coefficient,<sup>151</sup> and they can use sample data to construct a confidence interval for parameter values consistent with the sample results.<sup>152</sup> They can also determine the power of a study to detect a true population  $r^2$  or  $r$  of a given magnitude.<sup>153</sup> But even a regression model constructed on accurate and complete population data (not simply sample data) would still introduce predictive uncertainty, as long as there is residual variability among individuals in that population.

An important advantage of a regression model is that it takes variables into account transparently by including them as independent variables.<sup>154</sup> Moreover, *multivariate* regression models can incorporate several predictor variables and can model the increase in predictive success when multiple kinds of information are taken into account.<sup>155</sup> A multiple regression model adds new variables to the right side of the prediction equation, as in the following equation for two independent variables  $X$  and  $Z$ :<sup>156</sup>

$$Y = A + BX + CZ.$$

The prediction for  $Y$  is calculated using not only the value on variable  $X$ , but also the value on variable  $Z$ . The  $Y$ -intercept  $A$  is the value on the  $Y$  variable when both  $X$  and  $Z$  equal zero. In a multivariate model there is also a coeffi-

the new technique correlate with measurements using the criterion method. In such an application, the correlation coefficient is often referred to as a validity coefficient and it provides a measure of the validity of the new method relative to the criterion method. CARMINES & ZELLER, *supra* note 26, at 17-18; GHISELLI ET AL., *supra* note 26, at 269. The question of when the degree of criterion validity is acceptable depends upon the pragmatic context, which includes clinical concern for the patient's welfare if there are extreme under-predictions or over-predictions. Certain directions and degrees of bias might be more medically troubling than others.

151. On significance testing and confidence intervals for  $r^2$  or  $r$ , see COHEN & COHEN, *supra* note 31, at 51-59, 62-65; HAYS, *supra* note 34, at 620-31, 644-52; LOETHER & MCTAVISH, *supra* note 27, at 600-04; *Regression Models*, *supra* note 113, at 1449-53; and Fisher, *supra* note 113, at 716-20.

152. See *supra* note 151.

153. COHEN & COHEN, *supra* note 31, at 59-61.

154. See WONNACOTT & WONNACOTT, *supra* note 60, at 397.

155. On multiple regression analysis, see BERRY & FELDMAN, *supra* note 113, at 1-71; COHEN & COHEN, *supra* note 31, at 1-120, 300-50; HAYS, *supra* note 34, at 673-809; WONNACOTT & WONNACOTT, *supra* note 60, at 397-514; and Rubinfeld, *supra* note 113, at 181-85.

156. The general form for a multiple linear regression model is:

$$Y_i = A + B_1(X_{1i}) + B_2(X_{2i}) + \dots + B_k(X_{ki}),$$

where  $Y_i$  is the predicted value of  $Y$  for individual  $i$ ,  $X_1$  through  $X_k$  are  $k$  independent variables and  $X_{ki}$  is the value of variable  $X_k$  for individual  $i$ . COHEN & COHEN, *supra* note 31, at 81-85.  $A$  is the  $Y$ -intercept or regression constant, and  $B_1$  through  $B_k$  are the "partial regression coefficients" for the independent variables. *Id.* The meaning of the regression constant here is similar to the meaning in the bivariate model:  $A$  is the predicted value of  $Y$  when the value of every independent variable in the model is zero. A partial regression coefficient  $B_j$  is the (constant) increase in the predicted value of  $Y$  for a unit increase in the independent variable  $X_j$ , when the values of all the other independent variables are held constant. See COHEN & COHEN, *supra* note 31, at 81-100; HAYS, *supra* note 34, at 673-80, 687-92; WONNACOTT & WONNACOTT, *supra* note 60, at 396-414.

cient for each independent variable.<sup>157</sup> In the formula above, the constant  $B$  is the coefficient for variable  $X$  and  $C$  is the coefficient for  $Z$ . These constants, called partial regression coefficients, state the direct contribution of each independent variable to the predicted value of  $Y$ , *after* the contributions of all the other independent variables have been taken into account.<sup>158</sup> A partial regression coefficient states the incremental contribution to the prediction attributable to that specific independent variable, compared to what the prediction would have been with only the other independent variables included in the regression model.<sup>159</sup>

A multiple regression model finds the best-fitting predictor for the *combination* of independent variables. The best predictor uses the independent variables *as a group* to minimize the mean squared differences for predictions on the dependent variable. Because a multiple regression model is that combination of coefficients that produces the least amount of predictive error, some partial regression coefficients might change if independent variables are added to the model or some are taken away. If an independent variable that is statistically irrelevant is added to the model (that is, a variable whose partial regression coefficient in the population is zero), this will tend to increase the standard error (and hence widen the confidence intervals) for the partial coefficient of another independent variable in the model that is correlated with it.<sup>160</sup> Therefore, adding variables that turn out to be irrelevant tends to increase the measures of sampling uncertainty.<sup>161</sup> Doing so may also increase the risk of observing statistical significance that occurs merely by chance (Type I error), because the number of variables with partial correlation coefficients is increased.<sup>162</sup>

On the other hand, if a statistically relevant independent variable is omitted from the model (a variable whose partial regression coefficient in the population is not zero), then this will bias the coefficient of an independent variable in the model that is correlated with the omitted variable, and may do so seriously.<sup>163</sup> This means that the long-run expected value of sample coefficients for the included independent variable will be higher or

---

157. See LOETHER & MCTAVISH, *supra* note 27, at 332-33.

158. The partial regression coefficient for an independent variable  $X_i$  describes the change in the dependent variable "that accompanies a unit change in the regressor  $X_i$ , if all the other regressors remain constant." WONNACOTT & WONNACOTT, *supra* note 60, at 413. Each partial regression coefficient "represents the relative amount of contribution of that variable [to the overall prediction], after contributions of the other variables included in the regression equation are taken into account." LOETHER & MCTAVISH, *supra* note 27, at 332. A partial *correlation* coefficient can also be computed for each independent variable in the model, a conceptual counterpart to the bivariate correlation coefficient. See COHEN & COHEN, *supra* note 31, at 91-92; LOETHER & MCTAVISH, *supra* note 27, at 317-18, 333-34.

159. See WONNACOTT & WONNACOTT, *supra* note 60, at 400-14.

160. See BERRY & FELDMAN, *supra* note 113, at 12-14, 18-20.

161. The standard error for the coefficient of the irrelevant variable is probably not zero, and it will be increased by correlation with another independent variable in the model. *Id.* The increase in this standard error means that there is a greater probability of drawing a sample with a non-zero coefficient for the irrelevant variable. *Id.*

162. See *supra* text accompanying notes 74-75.

163. See BERRY & FELDMAN, *supra* note 113, at 20-21; WONNACOTT & WONNACOTT, *supra* note 60, at 397-400, 406-09, 417-20; Rubinfeld, *supra* note 113, at 188-89.

lower than its population coefficient, perhaps by a substantial amount. This potential for bias makes it important to include such "confounding variables" in the model and to control statistically for any relevant independent variables.<sup>164</sup>

Thus, a multiple regression model is the best-fitting model using a particular *set* of variables. The multiple correlation coefficient  $R$  is a measure of how well the predictive model performs as a whole.<sup>165</sup>  $R$  is, for the multivariate model, the counterpart to the correlation coefficient  $r$  of a bivariate model.<sup>166</sup> It is an index of the strength of linear association between the independent variables as a set and the dependent variable.<sup>167</sup> The statistic  $R$  varies from zero to one: when  $R = 0$ , the model predictions for the dependent variable are no better than the arithmetic mean of the dependent variable, and when  $R = 1$  there is perfect correlation, with no predictive error at all.<sup>168</sup> The multiple correlation coefficient  $R$  can also be squared, yielding  $R^2$ , called the coefficient of multiple determination.<sup>169</sup>  $R^2$  is the proportion of the variance in the dependent variable that is eliminated by making predictions using the multiple regression model instead of the dependent variable's arithmetic mean.<sup>170</sup> As with bivariate regression, the statistics  $R$  and  $R^2$  for linear regression models provide measures of predictive success only for the best-fitting linear model. It might be that, for a particular set of variables and a particular data set, a nonlinear model would reduce modeling uncertainty better than the best linear model would.

The relevance of modeling uncertainty to direct inference can now be summarized. The major, statistical premise of a direct inference to specific causation asserts a causal connection between two variables,  $A$  and  $B$ . Normally, a statistical association between  $A$  and  $B$  warrants using  $A$  to predict  $B$  and furnishes empirical evidence of the causal relationship. The statistical model supporting the major premise is, however, a source of uncertainty that can undermine any predictions, any causal interpretation, and ultimately any direct inference. There are two major sources of modeling uncertainty. The first is the specification of variables within the model—especially

---

164. A definition of "confounding factor" is provided in Green et al., *supra* note 6, at 369-73, 389. A confounding factor is "both a risk factor for the disease and a factor associated with the exposure of interest. Confounding refers to a situation in which the effects of two processes are not separated." *Id.* at 389. See *infra* note 173. See also HENNEKENS & BURING, *supra* note 99, at 35-37; Kaye & Freedman, *supra* note 26, at 138-39. On the possible effect of adding new variables to the model, see *supra* notes 160-64; *infra* notes 176-80, 227 and accompanying text.

165. See LOETHER & MCTAVISH, *supra* note 27, at 334-35.

166. *Id.*

167. *Id.* at 332-35.

168. COHEN & COHEN, *supra* note 31, at 86-88; LOETHER & MCTAVISH, *supra* note 27, at 334-35. Once the multiple regression model is used to calculate the predicted values of  $Y$ , the multiple correlation  $R$  is simply the correlation  $r$  between these predicted values and the observed  $Y$  values. See WONNACOTT & WONNACOTT, *supra* note 60, at 496-97. Unlike the bivariate Pearson correlation  $r$ , the value of  $R$  cannot be negative. HAYS, *supra* note 34, at 698.

169. COHEN & COHEN, *supra* note 31, at 86-88; LOETHER & MCTAVISH, *supra* note 27, at 334-35.

170. COHEN & COHEN, *supra* note 31, at 86-88, 100; HAYS, *supra* note 34, at 696-700; LOETHER & MCTAVISH, *supra* note 27, at 334-35; WONNACOTT & WONNACOTT, *supra* note 60, at 497.

whether the model takes into account enough of the relevant variables, so that the model's risk statistics have acceptable accuracy relative to the target population. The second major source of uncertainty is the form of model used, including the conditions to be met and the kinds of statistics generated. Both sources of modeling uncertainty are in addition to measurement uncertainty and sampling uncertainty,<sup>171</sup> and both sources contribute to the residual potential for predictive error, which can undermine any premise that most things in category *A* are also in category *B* as a result of being in category *A*.

Relative risk calculated from a regression analysis can quantify the strength of association within a general causal relationship, but use of such a statistic is always subject to modeling uncertainty. A reasonable factfinder must decide whether the model takes into account an adequate number of the relevant variables, whether the formal conditions of the model have been adequately satisfied,<sup>172</sup> whether the particular model has an acceptable fit to the data, and whether the residual degree of predictive error is acceptable. In the context of direct inference in torts, there are two distinct objectives to be considered. The first is model acceptability for purposes of finding general causation. The next section of the Article discusses such a causal interpretation of the model. The second is model acceptability for purposes of drawing a direct inference of specific causation. Part II of the Article discusses the role of modeling in warranting that inference.

#### *D. Acceptable Causal Uncertainty: Explaining the Probability of Event Occurrence*

This section completes the analysis of the uncertainties that are inherent in a major, statistical premise about general causation within groups of individuals. General causal propositions relate two or more variables by stating what types of events are causally linked to other types of events. Examples are whether the ingestion of Bendectin during pregnancy can cause the fetus to develop abnormally, whether ingesting water with a certain lead concentration can have harmful developmental effects on children, or whether the use of chlorofluorocarbon propellants in metered-dose inhalers for asthma

---

171. See *supra* text accompanying notes 150-53. Modeling error is in addition to measurement error and sampling error. As stated in Rubinfeld:

If the expert calculated the parameters of a multiple regression model using as data the entire population, the estimates [coefficients] might still measure the model's population parameters with error. Errors can arise for a number of reasons, including (1) the failure of the model to include the appropriate explanatory variables; (2) the failure of the model to reflect any nonlinearities that might be present; and (3) the inclusion of inappropriate variables in the model.

Rubinfeld, *supra* note 113, at 198.

172. The appropriate use of multiple regression models involves meeting or respecting more conditions than those discussed here. Such conditions involve additivity and linearity, collinearity, scedasticity, and autocorrelation. See generally BERRY & FELDMAN, *supra* note 113, at 37-88; LOETHER & MCTAVISH, *supra* note 27, at 329-30.

patients can damage the ozone layer. Such conclusions about general causation assert more than mere predictions based on a correlation. Causal explanations describe the causal tendencies or influences that underlie the observed associations and account for them. Causal accounts provide the warrant for predictions about unobserved events, and also for explanations why such events are likely or unlikely to occur in circumstances beyond those studied in the sample data. Causal explanations supply the evidentiary warrant for direct inferences to specific causation in the particular case.

The ideal warrant for a general causal explanation is a controlled experiment designed to keep causal uncertainty to acceptable levels. Causal uncertainty is the potential for error created by imposing a causal interpretation on a predictive model. The ideal experimental protocol has several features:<sup>173</sup> (1) it uses test subjects (such as laboratory animals) that are relatively homogeneous with respect to all characteristics known to be risk factors for the dependent variable (such as age or genetic makeup); (2) the subjects are randomly assigned to a test group and to a control group; (3) the test group is exposed to the test agent (such as a drug) but the control group is not; (4) the physical environment, diet, and other external factors for the two groups are maintained in identical fashion, with the exception of the exposure of the test group to the agent being tested; (5) the two groups are carefully monitored for all independent variables that might prove to be causally relevant; and (6) the two groups are carefully monitored for the dependent variable (such as a disease) to determine whether there is a difference in incidence between the two groups.<sup>174</sup> All measurement techniques and data should be acceptably valid and reliable. The sample should be acceptably large and random, and random assignment to the test and control groups warrants that any group differences are not due to the group-assignment process. If the study sample is large enough, there might even be warrant for thinking that the two groups are probably comparable to each other on all relevant independent variables other than the test exposure. Any residual variability can be measured and modeled with regression techniques, looking for any potentially explanatory variables other than exposure. Such a design therefore addresses both sampling and modeling uncertainties. And any statistically significant outcomes between the test and control groups constitute evidence that the exposure variable is probably associated with the outcome variable in the general population, not just in the sample. At some point, especially after other controlled studies replicate those results, there would be warrant for finding that the exposure variable is also causally related to the outcome variable and that it helps explain the outcome. Additional controlled experiments might even be able to model

---

173. See BLAND, *supra* note 65, at 6-25; HENNEKENS & BURING, *supra* note 99, at 178-212; LILIENTHAL & LILIENTHAL, *supra* note 53, at 256-73; Lavori et al., *supra* note 99, at 61-82; Moses, *supra* note 99, at 5-25; Nelson, *supra* note 99, at 37-48.

174. When there is a possibility of measurement bias due to knowledge of the group membership, standard protocols include blind studies. See *supra* note 105 and accompanying text.

the chain of causal events at the cell or biochemical levels, and intervention at those levels (such as vaccination) might be able to alter outcomes.

Such an ideal controlled study has become the paradigm of warrant for findings of general causation. The strategy behind a controlled experiment is to create a situation in which, *if* there is a causal connection between two variables, *then* it is likely to be detected as a statistically significant association, manifested in a statistically significant relative risk or correlation coefficient. Moreover, *if* a statistically significant association appears in the study data, *then* it is unlikely to be due to chance or to some type of study error, and the difference in outcome between the groups is probably due to the one known difference between them—the difference in exposure or predictor variable. The controlled experimental design warrants the conclusion that statistically significant results probably will be observed *if, but only if*, there is a true causal connection between the types of events being studied.

Much of legal factfinding still involves the kinds of general causation familiar from ordinary experience: falling heavy objects can crush other objects in their paths, metal automobiles with momentum can severely injure human bodies in a collision, bullets fired from guns can kill people. No controlled experiments are needed because individual and collective experiences amply warrant the general causal conclusions. However, the machinery of legal factfinding is aimed increasingly at causal claims whose warranting evidence is not available to the casual observer—claims about the metabolism of pharmaceuticals in the human body, the biological effects of electromagnetic fields around electric transmission lines, or the influence of the job applicant's sex on an organization's hiring decisions. In such cases, untrained observations leave too much causal uncertainty.

Causal uncertainty is the additional uncertainty created precisely because the conclusion is about general causation and is based on a causal interpretation of a predictive model. The primary source of causal uncertainty is that the observed or predicted association is neither sufficient nor necessary for general causation. A statistically significant association does not always warrant making a causal connection, and the absence of a statistically significant association does not always warrant asserting a lack of causal connection. First, although the presence of a statistically significant association in a study sample may be good evidence that a real statistical association probably exists in the population, this study result might be "causally spurious"—that is, it can lead to false causal conclusions.<sup>175</sup> A real statistical association between *A* and *B* in a population does not entail that *A* causes *B*. *B* may in fact cause *A*, and the direction of the purported causal influence should be reversed. Or, the observed statistical association might result because some third factor *C* has a causal influence on both *A* and *B*. The influence of *C* might not be detected by a study design in which

---

175. See COHEN & COHEN, *supra* note 31, at 359; JAMES A. DAVIS, *THE LOGIC OF CAUSAL ORDER* 16-27 (1985); DAVID A. KENNY, *CORRELATION AND CAUSALITY* 4 (1979); LOETHER & McTAVISH, *supra* note 27, at 292-99; WONNACOTT & WONNACOTT, *supra* note 60, at 487-89.

C is not controlled either physically or statistically. Therefore, even a real statistical association between A and B does not always warrant inferring any causal action between A and B. As discussed in the previous section, a relative risk of B given A, if calculated using an inadequately specified model, might disappear altogether once additional variables are controlled.

Second, the absence of a statistically significant association between A and B can lead to a false conclusion that there is no causal relationship between these types of events. This may be due merely to having statistical power that is too low to detect the association and is, therefore, a problem of sampling uncertainty. Even where sampling uncertainty is acceptable, however, and there is no statistical association to be detected in the normal population, there might still be a causal relationship that would come to light if the normal course of events were manipulated in different ways. For example, in the complex causal systems studied in genetics, ecology, and the medical and behavioral sciences, many causal influences are antagonistic, with one or more events counteracting, masking, or "suppressing" the would-be direct effects.<sup>176</sup> True causal relationships may be masked by the causal influence of other events and not revealed unless researchers manipulate and monitor the masking events. The study protocol might limit the ability to detect the causal influences of some variables, and the statistical model might leave some variables out of the analysis.

A major source of causal uncertainty is, therefore, incompleteness in the set of variables studied. In the language of multiple regression analysis, this kind of causal error is induced by "premature closure" or "under-specification" of the regression model.<sup>177</sup> If a causally relevant factor is not included in the model as an independent variable, observed correlations may be causally spurious and a lack of observed correlation may be causally misleading. Moreover, the addition of a new independent variable to a regression model might either increase or decrease a correlation coefficient already in the model.<sup>178</sup> For example, an observed correlation between A (an

---

176. On suppressor variables generally, see COHEN & COHEN, *supra* note 31, at 94-96; DAVIS, *supra* note 175, at 32-33; and LOETHER & MCTAVISH, *supra* note 27, at 299-301. "Suppression is a plausible model for many homeostatic mechanisms, both biological and social, in which force and counterforce tend to occur together and have counteractive effects." COHEN & COHEN, *supra* note 31, at 96.

177. See BERRY & FELDMAN, *supra* note 113, at 18-26; Fisher, *supra* note 113, at 708-09; Graham & Garber, *Evaluating the Effects of Automobile Safety Regulation*, 3 J. POL'Y ANALYSIS MGMT. 206, 211-12 (1984).

178. As stated by Kenny:

Too often researchers examine the simple, or raw, correlation coefficient as an indication of causal effects. The naive logic is that if X causes Y, then X and Y should be correlated, and if X does not cause Y, they should be uncorrelated. Neither statement is true. After controlling for other exogenous [causal or independent] variables, a strong relationship can vanish and a zero relationship can become strong.

KENNY, *supra* note 175, at 62. As Davis puts it: "Absent variables might do anything. . . . [W]e cannot always make the conservative assumption that additional variables would result in lower values [coefficients] for our [causal] arrows. The missing variables might be suppressors." DAVIS, *supra* note 175, at 65-66. The results of omitting relevant variables are potentially serious for the model. See, e.g., BERRY & FELDMAN, *supra* note 113, at 20-25.

Relative risk estimates can increase or decrease once additional variables are taken into account.

independent variable) and *B* (a dependent variable) might be partially or entirely eliminated when a new independent variable *C* is added and correlations appear between *C* and *A* and between *C* and *B*. The partial correlation coefficient for *A* might approach zero once *C* is introduced into the model, providing evidence that the original correlation between *A* and *B* was causally spurious. This explains why scientists are reluctant to infer causation merely on the basis of epidemiologic evidence, where many causally relevant variables may be unstudied.<sup>179</sup> A fortiori, mere reports of observations from individual cases usually provide even weaker evidence of causation.<sup>180</sup> Moreover, if the occasion of an observation is emotionally charged or if the observer has a strong interest in a particular interpretation, then the observation may lead to a superstitious belief in a causal connection, whereas a carefully controlled investigation would prove that the causal interpretation is erroneous.

Various factors affect the weight of the evidence for placing a causal interpretation on a statistical association. The weight of evidence or evidentiary support for a true causal relationship can have degrees, and finding general causation may be more or less warranted.<sup>181</sup> The first major factor affecting the weight of evidence for general causation is, therefore, the extent to which the set of study variables is sufficiently complete. The degree of warrant increases if there is good evidence that enough of the important and causally relevant variables are included in the study, so that the observed association is unlikely to be spurious. In the language of multiple

---

See, e.g., Greenland, *supra* note 110, at 1168 (stating that “[i]ndividual and population risks vary with factors other than the exposure in question”; that “[a]s a result of the inevitable complex interactions among risk factors,” the variation can be large; and that “it is possible for the variation to be in either direction,” either increasing or decreasing); Irva Hertz-Picciotto, *Shifting the Burden of Proof Regarding Biases and Low-Magnitude Associations*, 151 AM. J. EPIDEMIOLOGY 946, 947 (2000) (suggesting that researchers should ask “what evidence exists that upward biases are present and that they outweigh biases in the other (downward) direction”); Samuel Shapiro, *Bias in the Evaluation of Low-Magnitude Associations: An Empirical Perspective*, 151 AM. J. EPIDEMIOLOGY 939 (2000) (giving examples from data on oral contraceptives and breast cancer). Epidemiologists recount episodes where documented and confirmed relative risks of two to three may vanish altogether when researchers conduct larger studies or control for confounding variables. *XYZ v. Schering Health Care*, [2002] E.W.H.C. 1420 (QB), 2002 WL 1446183, ¶¶ 288-89 (July 29, 2002).

For further discussion of the statistical effects of including an irrelevant variable in a regression model or of excluding a relevant variable, see *supra* notes 112, 160-64 and accompanying text.

179. Green et al., *supra* note 6, at 335-38, 374-79.

180. Mary Sue Henifin et al., *Reference Guide on Medical Testimony*, in REFERENCE MANUAL ON SCIENTIFIC EVIDENCE 439, 474-75 (2d ed. 2000), available at [http://www.fjc.gov/public/pdf.nsf/lookup/sciman00.pdf/\\$file/sciman00.pdf](http://www.fjc.gov/public/pdf.nsf/lookup/sciman00.pdf/$file/sciman00.pdf).

181. Various areas of science have developed their own guidelines for causal inference—guidelines adapted to the characteristics of the area. For example, epidemiology uses the Henle-Koch-Evans Postulates to guide an inference from statistical associations to biological causation. See LILJENFELD & LILJENFELD, *supra* note 53, at 292-95, 316-18 (also discussing Henle-Koch Postulates in the context of determining the causal role of a microorganism in an infectious disease); Bert Black & David E. Lilienfeld, *Epidemiologic Proof in Toxic Tort Litigation*, 52 FORDHAM L. REV. 732, 762-64 (1984). The guidelines used by the United States Environmental Protection Agency in inferring cancer causation were modeled on criteria developed by Bradford Hill in examining cigarette smoking and lung cancer. See Proposed Guidelines for Carcinogen Risk Assessment, 61 Fed. Reg. 17,960, 17,974-75 (proposed Apr. 23, 1996).



regression analysis, warrant requires good evidence that the regression model is adequately specified. Another way to pose the question is whether the available studies have taken into account enough of the factors that might be confounding. A variable might be confounding either by creating an observed association that is causally spurious or by masking an association that would otherwise appear.<sup>182</sup>

In principle, if a study has taken into account all of the variables that might be causing the observed associations, then those associations that are observed are more likely to reflect real causal connections, and an absence of statistically significant associations is more likely to reflect an absence of causation. In reality, however, studies almost never include *all* of the variables that might causally influence an outcome or dependent variable. Not only observational studies on human subjects, but even controlled studies on laboratory animals, must often leave out of account variables that can affect or refine the study outcomes. Given the realities of causal complexity and the limitations of human knowledge, we often accept a great deal of uncertainty about general causation. The meaning of "sufficiently complete" here is, therefore, that enough important causal variables have been taken into account given the pragmatic outcomes at issue in the legal context. Even two or three independent variables might explain enough of the variation in the dependent variable to satisfy the policies at work in a given factfinding context.<sup>183</sup> Or perhaps any additional variables would be unlikely to change the observed associations so drastically as to undermine the ultimate legal conclusion. Therefore, deciding that a particular degree of incompleteness is acceptable is necessarily a practical issue to which non-epistemic considerations are relevant.

The importance of a control group is to provide a similar comparison group in which the unknown (but causally relevant) factors can bring about similar effects in the absence of the test variable (exposure). Ideally, the control group reproduces those unknown causal influences in the same proportions and to the same extent that they occur in the test group. In a controlled study, therefore, the study design may warrant the finding that a particular causal model is probably adequately specified. If the test group and the control group are identical on all variables that are probably causally relevant (other than the difference in exposure), if the two groups are large enough to capture the combinations of unknown causal factors in similar proportions, and if there occurs a statistically significant relative risk in the test group, then this study design warrants the conclusion that the difference between rates of occurrence is probably due to the exposure of the test group. Therefore, a study has a stronger design, from the standpoint of supporting causal inferences, if all potentially causal factors (other than the

---

182. Green et al., *supra* note 6, at 369-73; see *supra* notes 155-64 and accompanying text.

183. See, e.g., *Bazemore v. Friday*, 478 U.S. 385, 397-404 (1986) (noting that while the omission of variables from a regression analysis may render the analysis less probative, an analysis which accounts for "major factors" is normally admissible in Title VII pattern and practice cases).

variable being investigated) are held constant for the test and control groups—including genetic factors, developmental factors for the individual subjects, and environmental factors. The evidence of causal connection is strengthened to the extent that there is good evidence that enough of the potentially relevant causal factors (even if unknown) have been physically or statistically controlled.

For the same reason, randomization within the study design strengthens the warrant for causal inference. If the subjects in the study are assigned to the test or control groups by a truly random method, then this combats any potential for bias or confounding that might result from the process of group assignment.<sup>184</sup> The rationale parallels that for random sampling. Through random sampling, researchers try to eliminate any statistical association between sample statistics and any factor that might have influenced sample selection. Through randomization, they try to eliminate any statistical association between the outcome variable and any factor that might have influenced group assignment. Randomization becomes especially important if the assignment process could be influenced by any feature of the individual subjects that is also causally relevant to the variables being studied. The same reasoning supports the desirability of blind studies, in which researchers conducting measurements or performing other actions in connection with the study do not know which subjects are in which study groups. If those conducting the study were to know the group assignment, this knowledge might influence how they take measurements or perform other tasks, and such knowledge might causally influence the reported results. In other words, randomization and blind protocols are designed to keep researchers from introducing any confounding causal influences. If randomization and blind protocols are not used, then the added causal uncertainty about confounding design will weaken the warrant for any causal conclusions based on the study results.

But the fact that randomization can eliminate *one* source of bias does not mean that it eliminates *all* sources of bias. Even with randomization, there is no guarantee that the groups are comparable to each other or to the population on *all* causally relevant factors: this depends on the complexity

---

184. LILIENFELD & LILIENFELD, *supra* note 53, at 257; Lavori et al., *supra* note 99, at 61-69. Randomization is sometimes extolled as the near-panacea for eliminating inference problems due to confounding variables or non-comparability between the experimental group and the control group. For example, Bland states that if we randomize, "[t]he only differences between the groups will be those due to chance. . . . Any difference between the groups which is larger than [the likely effects of chance] is likely to be due to the treatment, since there will be no other differences between the groups." BLAND, *supra* note 65, at 8-9. As another example, Lilienfeld & Lilienfeld state:

The epidemiologist can achieve comparability [between the experimental and control groups] on factors that are known to have an influence on the outcome, such as age, sex, race, or severity of disease, by matching for these factors. But one cannot match individuals for factors whose influence is not known or cannot be measured. This problem can be resolved by the random allocation of individuals to the experimental and control groups, which assures the comparability of these groups with respect to *all* factors—known and unknown, measurable and not measurable—except for the one being studied.

LILIENFELD & LILIENFELD, *supra* note 53, at 257.

of the underlying causal process, the prevalence of relevant causal factors within the population, and the adequacy of the sampling.<sup>185</sup> Before drawing an inference of specific causation, the factfinder will face the decision whether the study design (including physical controls, random sampling, randomization, and model specification) warrants the conclusion that any residual uncertainty about causal completeness is acceptable. The residual risk of unknown confounding factors must be acceptable for the purposes of tort law.

After completeness of the causal model, a second major factor affecting the weight of evidence on causation is the strength of the statistical association itself, or the degree to which the outcome variable varies with the input variable.<sup>186</sup> The strength of association can be measured by the magnitude of the relative risk or of the correlation coefficient. Using current scientific conventions, random sampling uncertainty alone is within acceptable limits if the strength of association clears the threshold of statistical significance. Beyond that threshold, however, associations that are statistically significant can still vary in strength, and the stronger the association, the more likely it is that there is some underlying causal relationship. A very strong association makes it less likely that unknown but causally relevant variables would explain away the association if only they were taken into account.<sup>187</sup> In the terminology of regression analysis, as the population correlation coefficient approaches one, the predictive power of the independent variable increases and the residual scatter of predictive error approaches zero. If the population correlation coefficient were ever equal to one, then the independent variable would be a perfectly accurate predictor for the dependent variable. As the correlation coefficient approaches one, therefore, at least in large samples, it is less likely that the addition of a new independent variable to the model would make that correlation disappear altogether.

The mechanistic ideal is to predict outcomes accurately and completely and to control the occurrence of events, at least under experimental conditions. With relatively closed causal systems and completely specified causal models, such as a kinetic model for balls on a billiard table, it may be possible to predict and explain nearly all variability in events. It is often possible in law, when dealing with macro events, to approximate a mechanistic explanation, for which a set of input conditions completely determines the

---

185. See COLIN HOWSON & PETER URBACH, *SCIENTIFIC REASONING* 143, 152 (1989) ("[R]andomization cannot possibly guarantee that the [experimental] groups will be free from bias by unknown nuisance factors").

186. See Proposed Guidelines for Carcinogen Risk Assessment, 61 Fed. Reg. at 17,974; HENNEKENS & BURING, *supra* note 99, at 39-40; LILIENFELD & LILIENFELD, *supra* note 53, at 300-02; Green et al., *supra* note 6, at 376-77; see also JOHN STUART MILL, *A SYSTEM OF LOGIC* 260-63 (Longmans, Green & Co., 8th ed. 1900) (1843) (discussing the method of concomitant variation).

187. E.g., Shapiro, *supra* note 178, at 939 (concluding that "if an association is of relatively low magnitude (defined here as a relative risk estimate of less than 2.0), it may not be possible to judge whether or not it can be entirely accounted for by bias," as opposed to causation due to exposure); Hertz-Picciotto, *supra* note 178, at 947 (arguing that causal inferences should draw upon more than the magnitude of association, and urging that hypotheses of bias be evaluated just like other causal hypotheses).

outcome events. For example, hard metal objects moving with high energy can seriously injure unprotected human bodies through impact. Increasingly, however, factfinders in tort cases must decide general causation within open-ended causal systems, within incomplete causal models, and with a great deal of individual variability that the available model cannot explain. It is not reasonable to expect every causal model to display anything approaching a perfect correlation or to be causally complete. Fortunately, warranted factfinding is possible without having a complete, mechanistic explanation. Other things being equal, the higher the strength of association in a given study and the smaller the amount of residual unexplained variability, the greater the warrant for a causal conclusion. Once the weight of all of the evidence on causation reaches an acceptable level (given the legal context), then a conclusion of causation may be warranted even if the strength of association is relatively low.

The third major factor that can strengthen a causal inference is consistency of results among multiple studies.<sup>188</sup> If an observed association appears consistently in different studies by different investigators with different samples drawn from diverse populations, then it becomes far more likely that there is an underlying causal process producing that association. It is less likely that the observed association is coincidental, or that confounding factors explain the results. At the same time, lack of consistency among well-conducted studies may suggest that the association occurred by chance (however unlikely that was) or that other variables causally explain the observed association. If statistical results are inconsistent, then there should be a causal explanation for the inconsistency. Thus, replicating a suggestive lead study or conducting follow-up studies is often a high priority in science, because those studies can contribute substantially to the warrant for causal conclusions.

Finally, the degree of warrant for a causal connection depends upon the physical and biological plausibility of the causal model, including temporal directionality from the cause to the effect. A proposed causal model must be scientifically plausible from the standpoint of other well-founded causal theories and principles, as well as our general experience.<sup>189</sup> In epidemiology and toxicology, for example, the weight of evidence for a causal relationship between exposure and response variables increases if there is a plausible mechanism for the causal action and the causal theory is consistent

---

188. Proposed Guidelines for Carcinogen Risk Assessment, 61 Fed. Reg. at 17,974; HENNEKENS & BURING, *supra* note 99, at 41-42; LILIENFELD & LILIENFELD, *supra* note 53, at 298-300; Green et al., *supra* note 6, at 377-78.

189. The role of theory remains paramount even when empirical studies are available, for statistical analyses of data are not sufficient to determine whether a model is adequately specified. For example, a low  $R^2$  does not necessarily mean that a causally relevant variable has been omitted from the model. The  $R^2$  may also be low due to measurement error or due to the form of the model selected. See BERRY & FELDMAN, *supra* note 113, at 25. Second, although statistical tests are helpful in identifying included but seemingly irrelevant variables, if those variables have theoretical relevance they should be removed only with great caution. See *id.* at 25-26; WONNACOTT & WONNACOTT, *supra* note 60, at 408-09.

with accepted theories of pharmacokinetics and metabolism.<sup>190</sup> The evidence for causation is also stronger if there is a dose-response relationship between the variables that is positive and mechanistically plausible under the circumstances.<sup>191</sup> Beyond being consistent with the data used to support it, a causal model must also be plausible given other scientific theories about the underlying phenomenon.

An essential aspect of causal plausibility is the requirement that a cause must precede its effect in time, with enough time intervening to allow a causal mechanism to produce the result.<sup>192</sup> Disregarding the temporal directionality of causation can also lead to serious inferential errors about probability of occurrence. Calculating probabilities of occurrence accurately becomes especially critical when the risk estimate for general causation plays a warranting role in direct inference to specific causation. The conditional probability of *B* given *A* ("*Prob(B|A)*") quantifies the likelihood that *B* will occur once *A* occurs. "Forward" conditional probabilities condition events that occur later in time on events that occur earlier in time.<sup>193</sup> In Figure 4, in which the arrows indicate both the direction of time and the direction of causal influence, the probability of *B* given *A* and the probability of *E* given *B* ("*Prob(E|B)*") are examples of forward conditional probabilities.<sup>194</sup> For example, diagnostic tests are designed so that a pre-existing condition causes some positive test result to occur. Two forward conditional probabilities that help characterize important uncertainties for a diagnostic test are its sensitivity and specificity.<sup>195</sup> The sensitivity of a diagnostic test (such as a medical test) is the proportion of all affected individuals (those who have some condition or characteristic) who correctly test positive (the "true positive rate"). An affected person who receives a negative result on the test has a "false negative" result. The specificity of a test is the propor-

---

190. See Proposed Guidelines for Carcinogen Risk Assessment, 61 Fed. Reg. at 17,975; HENNEKENS & BURING, *supra* note 99, at 40-41; LILIENTHAL & LILIENTHAL, *supra* note 53, at 297-98, 315-16; Green et al., *supra* note 6, at 378.

191. See Proposed Guidelines for Carcinogen Risk Assessment, 61 Fed. Reg. at 17,974; HENNEKENS & BURING, *supra* note 99, at 43; LILIENTHAL & LILIENTHAL, *supra* note 53, at 309-15; Green et al., *supra* note 6, at 377.

192. See DAVID HUME, A TREATISE ON HUMAN NATURE 467 (T.H. Green & T.H. Grose eds., Longmans, Green & Co. 1874) (1738) (discussing Rule 2 for determining cause-and-effect: "The cause must be prior to the effect."). See Regulations Restricting the Sale and Distribution of Cigarettes and Smokeless Tobacco to Protect Children and Adolescents, 61 Fed. Reg. 44,396, 44,476 (Aug. 28, 1996); Proposed Guidelines for Carcinogen Risk Assessment, 61 Fed. Reg. at 17,974; DAVIS, *supra* note 175, at 11-16; HENNEKENS & BURING, *supra* note 99, at 42-43; KENNY, *supra* note 175, at 2-3; Green et al., *supra* note 6, at 376.

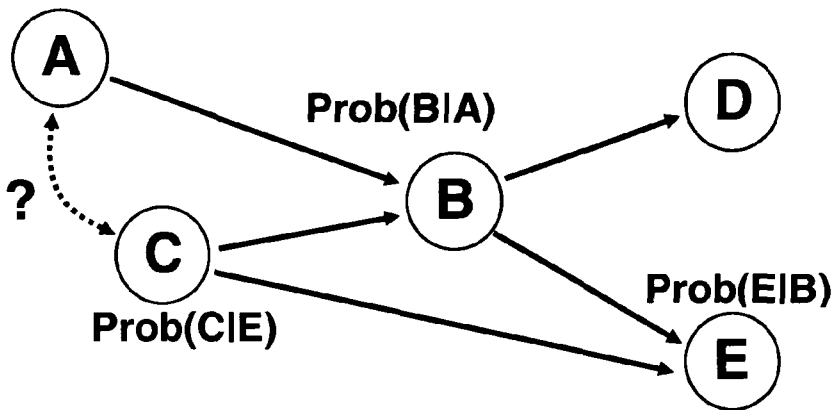
193. For a similar use of the terms "forward" and "backward," see NORMAN ABRAMSON, INFORMATION THEORY AND CODING 99 (1963).

194. For discussions of path analysis, which places the results of a multiple regression analysis in a form conducive to a causal interpretation, see COHEN & COHEN, *supra* note 31, at 79-123, 353-78; DAVIS, *supra* note 175, at 7-69; KENNY, *supra* note 175, at 22-44; LOETHER & MCTAVISH, *supra* note 27, at 338-43; and WONNACOTT & WONNACOTT, *supra* note 60, at 396-433. On the theory of causal modeling generally, see JUDEA PEARL, CAUSALITY: MODELS, REASONING, AND INFERENCE (2000); and GLENN SHAFER, THE ART OF CAUSAL CONJECTURE (1996).

195. See FINKELSTEIN & LEVIN, *supra* note 34, at 82; KASSIRER & KOPELMAN, *supra* note 10, at 18-22; Henifin et al., *supra* note 180, at 465-67.

tion of all unaffected individuals who correctly test negative on the test (the "true negative rate"). They do not have the condition and are correctly diagnosed or classified as not having it. A person who does not have the condition but receives a positive result on the test constitutes a "false positive." Sensitivity and specificity together are the "operating characteristics" of the test, and do not depend on the prevalence rate of the condition in the population.<sup>196</sup> Sensitivity and specificity reflect the forward conditional probabilities of the test results given the input condition.<sup>197</sup>

FIGURE 4



"Backward" conditional probabilities quantify the probability that a type of event occurred earlier in time given that a type of event occurs later

196. FINKELSTEIN & LEVIN, *supra* note 34, at 82.

197. It is sometimes said that a factor affecting weight is the "degree of specificity" between a causal variable and an effect variable. If an association is "highly specific" in this sense, then the input variable is associated only (or mostly) with a particular output variable, and the input-output relationship has a high "true negative rate." See LILIENFELD & LILIENFELD, *supra* note 53, at 302; Green et al., *supra* note 6, at 379. An association is more likely to have a causal basis if the exposure is associated only with a single type of disease or outcome. Green et al., *supra* note 6, at 379. The reason for thinking that highly specific associations are indicative of causation is that "[t]he vast majority of agents do not cause a wide variety of effects." *Id.*

This generalization is related to the notion that a diagnostic test for a disease is highly specific if it has a high rate of detection for "true negatives," meaning that a high proportion of all individuals who do not have the disease fail to cause a positive result on the test and therefore correctly test negative for the disease. The specificity of the test "is defined as the percent of those who do *not* have the disease and are so indicated by the test." LILIENFELD & LILIENFELD, *supra* note 53, at 150-51. By comparison, sensitivity "is defined as the percent of those who have the disease, and are so indicated by the test." *Id.*; HENNEKENS & BURING, *supra* note 99, at 331-35. In complex causal systems, specificity should not be demanded, or even expected. When specificity is in fact high, however, it can strengthen the case for causation.

in time. A backward conditional probability uses the occurrence of the effect as evidence of the prior occurrence of the cause. In Figure 4,  $Prob(C|E)$  is an example of a backward conditional probability. The formula for computing values for backward conditional probabilities is Bayes' Theorem,<sup>198</sup> which can be written as:

$$Prob(C|E) = \frac{Prob(C) \times Prob(E|C)}{Prob(E)},$$

in which  $C$  is the (earlier) cause and  $E$  is the (later) effect. The occurrence of  $E$  is some evidence that  $C$  had occurred earlier in time. The conditional probability  $Prob(C|E)$  is called the "posterior probability" of  $C$  conditioned on  $E$ , or the probability of  $C$  after taking the occurrence of  $E$  into account. It expresses the expected relative frequency for a prior occurrence of  $C$ , given that  $E$  occurs.<sup>199</sup> Although there are many controversies over the proper interpretation and use of Bayes' Theorem,<sup>200</sup> the theorem itself is undoubtedly deducible within the probability calculus.<sup>201</sup>

---

198. For discussions of Bayes' Theorem, see C.G.G. AITKEN, STATISTICS AND THE EVALUATION OF EVIDENCE FOR FORENSIC SCIENTISTS 31-56 (1995); L. JONATHAN COHEN, AN INTRODUCTION TO THE PHILOSOPHY OF INDUCTION AND PROBABILITY (1989), §§ 3, 9, 19; MICHAEL O. FINKELSTEIN, QUANTITATIVE METHODS IN LAW: STUDIES IN THE APPLICATION OF MATHEMATICAL PROBABILITY AND STATISTICS TO LEGAL PROBLEMS 87-89 (1978); GOLDBERG, *supra* note 13, at 38-48; ALVIN I. GOLDMAN, KNOWLEDGE IN A SOCIAL WORLD 109-30 (1999); HOWSON & URBACH, *supra* note 185; GUDMUND R. IVERSEN, BAYESIAN STATISTICAL INFERENCE (1984); JOSEPH B. KADANE & DAVID A. SCHUM, A PROBABILISTIC ANALYSIS OF THE SACCO AND VANZETTI EVIDENCE (1996); HENRY E. KYBURG, JR., PROBABILITY AND INDUCTIVE LOGIC 19-20, 68-74 (1970) [hereinafter PROBABILITY AND INDUCTIVE LOGIC]; POLLOCK & CRUZ, *supra* note 16, at 92-119; DAVID A. SCHUM, THE EVIDENTIAL FOUNDATIONS OF PROBABILISTIC REASONING 41-54, 213-222 (1994); Lea Brilmayer & Lewis Kornhauser, Review: *Quantitative Methods and Legal Decisions*, 46 U. CHI. L. REV. 116 (1978); Richard D. Friedman, *Assessing Evidence*, 94 MICH. L. REV. 1810 (1996); David Kaye, *Probability Theory Meets Res Ipsa Loquitur*, 77 MICH. L. REV. 1456 (1979) [hereinafter *Probability Theory*]; D. H. Kaye, *What Is Bayesianism?: A Guide for the Perplexed*, 28 JURIMETRICS J. 161 (1988); Richard O. Lempert, *Modeling Relevance*, 75 MICH. L. REV. 1021 (1977) [hereinafter *Modeling Relevance*]; Richard O. Lempert, *The New Evidence Scholarship: Analyzing the Process of Proof*, 66 B.U. L. REV. 439 (1986) [hereinafter *Evidence Scholarship*]; *Statistical Approaches, Probability Interpretations, and the Quantification of Standards of Proof*, THE EVOLVING ROLE OF STATISTICAL ASSESSMENTS AS EVIDENCE IN THE COURTS, 191-205 (Stephen E. Fienberg ed., 1989); Laurence H. Tribe, *Trial by Mathematics: Precision and Ritual in the Legal Process*, 84 HARV. L. REV. 1329 (1971) [hereinafter *Trial by Mathematics*]; and Vern R. Walker, *Language, Meaning, and Warrant: An Essay on the Use of Bayesian Probability Systems in Legal Factfinding*, 39 JURIMETRICS J. 391, 397-404 (1999).

199. FINKELSTEIN & LEVIN, *supra* note 34, at 75-81.

200. The debate over the application of Bayes' Theorem has been vigorous among legal scholars. Representative of this debate are: Ronald J. Allen, *The Nature of Juridical Proof*, 13 CARDOZO L. REV. 373 (1991); Ronald J. Allen, *A Reconceptualization of Civil Trials*, 66 B.U. L. REV. 401, 401-15 (1986); Lea Brilmayer, *Second-Order Evidence and Bayesian Logic*, 66 B.U. L. REV. 673 (1986); Craig R. Callen, *Kicking Rocks with Dr. Johnson: A Comment on Professor Allen's Theory*, 13 CARDOZO L. REV. 423 (1991); David L. Faigman & A. J. Baglioni, Jr., *Bayes' Theorem in the Trial Process: Instructing Jurors on the Value of Statistical Evidence*, 12 LAW & HUM. BEHAV. 1 (1988); Stephen E. Fienberg, *Gatecrashers, Blue Buses, and the Bayesian Representation of Legal Evidence*, 66 B.U. L. REV. 693 (1986); Edward Gerjuoy, *The Relevance of Probability Theory to Problems of Relevance*, 18

Having a warranted value for  $Prob(C|E)$  depends upon having warranted values for the three probabilities on the right-hand side of the equation. First, the conditional probability  $Prob(E|C)$  is the forward probability that  $E$  will occur given that  $C$  occurs (conditional on  $C$ 's occurring). This is often called the "likelihood" of  $E$  given  $C$ .<sup>202</sup> The existence of a non-zero value for the likelihood  $Prob(E|C)$  is warranted by the evidence that there is a general causal relationship running from  $C$  to  $E$ . Any particular value assigned to  $Prob(E|C)$  can be warranted by a partial correlation coefficient in an adequately specified regression model, subject to the types of uncertainty discussed in previous sections of this Article. Second, if adequate studies are available, the prevalence of  $C$  and  $E$ , or the relative frequencies with which they occur in the population, may be used to estimate the unconditioned probabilities  $Prob(C)$  and  $Prob(E)$ . The probability  $Prob(C)$  is called the "prior probability" of  $C$ , determined without taking  $E$  into account. The unconditioned probability  $Prob(E)$  is the probability that the effect  $E$  will occur at all, through all possible causal sequences.<sup>203</sup>

The example of a diagnostic test illustrates Bayes' Theorem and the importance of prevalence in calculating backward conditional probabilities. As stated just above, the sensitivity and specificity of a diagnostic test are forward conditional probabilities. Using test results as the basis for diagnosing the test subject's condition that caused those results, however, requires

JURIMETRICS J. 1, 9-28 (1977); John Kaplan, *Decision Theory and the Factfinding Process*, 20 STAN. L. REV. 1065, 1083-91 (1968); D. H. Kaye, *The Probability of an Ultimate Issue: The Strange Cases of Paternity Testing*, 75 IOWA L. REV. 75 (1989); *Evidence Scholarship*, *supra* note 198; and Glanville Williams, *The Mathematics of Proof—II*, 1979 CRIM. L. REV. 340, 340-50 (1979).

For arguments on the usefulness of Bayes' Theorem in legal decisionmaking, see GOLDBERG, *supra* note 13, at 38-48; Michael O. Finkelstein & William B. Fairley, *A Bayesian Approach to Identification Evidence*, 83 HARV. L. REV. 489 (1970) [hereinafter *Bayesian Approach*]; Michael O. Finkelstein & William B. Fairley, *The Continuing Debate over Mathematics in the Law of Evidence: A Comment on "Trial by Mathematics"*, 84 HARV. L. REV. 1801 (1971); *Probability Theory*, *supra* note 198; David Kaye, *The Paradox of the Gatecrasher and Other Stories*, 1979 ARIZ. ST. L.J. 101 (1979); Jonathan J. Koehler & Daniel Shaviro, *Veridical Verdicts: Increasing Verdict Accuracy Through the Use of Overtly Probabilistic Evidence and Methods*, 75 CORNELL L. REV. 247 (1990); Daniel J. Kornstein, *A Bayesian Model of Harmless Error*, 5 J. LEGAL STUD. 121 (1976); and *Modeling Relevance*, *supra* note 198.

For arguments against at least certain uses of Bayesian analysis in a legal context, see L. JONATHAN COHEN, *THE PROBABLE AND THE PROVABLE* (1977) [hereinafter *THE PROBABLE AND THE PROVABLE*]; Brillmayer & Kornhauser, *supra* note 198; Craig R. Callen, *Notes on a Grand Illusion: Some Limits on the Use of Bayesian Theory in Evidence Law*, 57 IND. L.J. 1 (1982); L. Jonathan Cohen, *The Logic of Proof*, 1980 CRIM. L. REV. 91; Laurence H. Tribe, *A Further Critique of Mathematical Proof*, 84 HARV. L. REV. 1810 (1971); and *Trial by Mathematics*, *supra* note 198.

201. E.g., FINKELSTEIN, *supra* note 198, at 87-89; PROBABILITY AND INDUCTIVE LOGIC, *supra* note 198, at 19-20; *Bayesian Approach*, *supra* note 200, at 498-99; *Probability Theory*, *supra* note 198, at 1468-71; *Trial by Mathematics*, *supra* note 198, at 1351-53; Walker, *supra* note 18, at 265 n.40.

202. KADANE & SCHUM, *supra* note 198, at 122-26; SCHUM, *supra* note 198, at 49-50, 146, 218-22.

203. See, e.g., FINKELSTEIN & LEVIN, *supra* note 34, at 75-81. Relative to only a single causal event  $C$ ,

$$Prob(E) = Prob(E|C) Prob(C) + Prob(E|not-C) Prob(not-C),$$

where  $Prob(E|not-C)$  is the probability that  $E$  will occur given that  $C$  does not occur. See *id.*  $Prob(E)$  is a function of all of the forward conditional probabilities leading to  $E$ , each weighted by the probability that the corresponding causal sequence will occur. It is therefore equal to the sum of  $Prob(E|C)$  and  $Prob(E|not-C)$ , where  $Prob(E|not-C)$  is the total probability that  $E$  will occur as the result of any causal sequence in which  $C$  does not occur.



backward conditional probabilities, or the “predictive values” of the test.<sup>204</sup> The positive predictive value is the proportion of individuals who test positive that truly have the condition or characteristic.<sup>205</sup> The negative predictive value is the proportion who test negative that truly do not have the condition.<sup>206</sup> While sensitivity and specificity do not depend on the prevalence or rate of occurrence of the condition in the population, the predictive values of the test do. For example, for a test with sensitivity and specificity each equal to 99%, the probability that a person who tests positive is actually affected is one-half when the prevalence of the condition in the population is one per 100—that is, among all those with positive results on the test, 50% are expected to have the condition.<sup>207</sup> This is because those who do not have the condition are ninety-nine times more numerous than those who do have the condition, and those without the condition still test positive about one time in 100.<sup>208</sup> For a prevalence of one per 1000, however, the positive predictive value of the test falls to about 9%.<sup>209</sup>

In sum, the above factors—namely, model completeness, strength of association, consistency among studies, scientific plausibility and temporal directionality—all affect the weight of evidence for a general causal relationship between two variables. They also reflect the human desire for a single causal account of reality. The causal ideal is a single causal model for the entire universe, which would explain everything that occurs, and we

204. *Id.* at 82-83.

205. *Id.* at 82 (defining the “[p]ositive predictive value” as “the proportion of all test-positive people who are truly affected”).

206. *Id.* at 83 (defining the “[n]egative predictive value” as “the proportion of all test-negative people who are truly unaffected”).

207. *See id.* at 82-83 (calculating the odds of being affected given a positive test outcome). If  $C+$  = the condition of being affected,  $not-C+$  = the condition of being unaffected, and  $E+$  = the event of having a positive test result, then the odds on being truly affected given a positive test result are:

$$\frac{Prob(C+|E+)}{Prob(not-C+|E+)} = \frac{Prob(E+|C+)}{Prob(E+|not-C+)} \times \frac{Prob(C+)}{Prob(not-C+)}$$

The odds on being truly affected given a positive test result also equal:

$$(Sensitivity / 1 - Specificity) \text{ Prevalence Odds,}$$

where the *Prevalence Odds* =  $Prob(C+)/Prob(not-C+)$ . *See id.* at 83.

208. Bayes' Theorem shows how prevalence comes into play in both  $Prob(C)$  and  $Prob(E)$ , the unconditioned rates of occurrence in the population. If the probability of testing positive is  $Prob(E+)$  and the probability of being affected by the condition is  $Prob(C+)$ , then the conditional probability of having the condition if the patient tests positive on the diagnostic test is  $Prob(C+|E+)$ . The sensitivity of the test is the forward conditional probability  $Prob(E+|C+) = 0.99$ . The prevalence of the condition is  $Prob(C+) = 0.01$ . The numerator of Bayes' Theorem is therefore  $Prob(C+) Prob(E+|C+) = (0.01)(0.99) = 0.0099$ . The denominator is  $Prob(E+) = Prob(E+|C+) Prob(C+) + Prob(E+|not-C+) Prob(not-C+)$ . *Supra* note 203. The first term in the denominator sum is the same as the numerator. The latter term in the denominator equals the product of the conditional probability of a positive test result for an unaffected patient without the condition (that is, one minus the specificity of the test, or  $1 - 0.99 = 0.01$ ) and the prevalence of unaffected people in the population (that is,  $1 - 0.01 = 0.99$ ). Therefore, the denominator  $Prob(E+) = (0.99)(0.01) + (0.01)(0.99)$ , which is twice the numerator  $(0.01)(0.99)$ . The probability of being affected given a positive test result is therefore  $1/2$ .

209. A comparable calculation to that in note 208 shows how the positive predictive value falls to about 9% for a prevalence of one per 1000. With the forward conditional probabilities for sensitivity and specificity remaining at 0.99, and the prevalence of the condition ( $Prob(C+)$ ) falling to 0.001, the calculation for Bayes' Theorem becomes  $Prob(C+|E+) = (0.001)(0.99) / (0.99)(0.001) + (0.01)(0.999) = 0.00099 / (0.00099 + 0.00999) = 0.00099 / 0.01098$  = about 9%.

therefore expect all valid causal models to cohere with each other. In the unified causal account of reality, we expect completeness, consistency, and coherence: all phenomena must be explained, all studies must have consistent results, and all theories must fit nicely together. Causal generalizations must support law-like (“nomic”) probabilities, which provide explanations as well as predictions.<sup>210</sup> We expect causal models to explain not only why individuals are *similar* (why the data have the central tendency they do), but also why individuals are *different* from each other (why the data show the variability that they do). Complete causal explanations would explain the variability found in identifiable subgroups of cases, and would explain the differences among individual cases. It is the concept of causation, applicable to every individual case, that warrants inferences from statistics about groups to conclusions about individual members of those groups. A model that is truly causal—and not merely statistical—explains events in every individual case, and every individual deviation from a generic prediction deserves a causal explanation.

## II. UNCERTAINTIES AND WARRANT IN APPLYING THE GENERALIZATION TO THE INDIVIDUAL PLAINTIFF

This part of the Article analyzes two additional uncertainties introduced by using a major statistical premise to draw a probabilistic conclusion about a specific member of the reference group. That is, an assertion about a general causal relationship connecting two variables may have acceptable uncertainty considered as a stand-alone assertion, but the assertion may be less acceptable when it is used to draw a direct inference in a particular tort case. The first section discusses the problem of identifying the appropriate group to serve as the reference group *A*. It examines the warrant for finding that any particular reference group (such as women over age forty, who have no history of cancer in the immediate family) adequately represents the specific plaintiff in all causally relevant variables. It also discusses what “adequately represents” means in this context. The second section addresses the additional uncertainty introduced by finding a probability that a specific member of *A* is also a member of a subgroup *B*. Therefore, the two major sources of additional uncertainty in drawing the direct inference are (1) identifying a sufficiently representative reference group for the specific plaintiff, and (2) using statistical evidence about that reference group to assign a probability for classifying that specific plaintiff.<sup>211</sup>

---

210. See POLLOCK, *supra* note 14, at 32-38, 42-43, 81-86, 132-40; Walker, *supra* note 18, at 286-89. In order to justify direct inference, the indefinite probability that forms the evidentiary basis for the inference must be a “nomic probability,” or a “law” that supports counterfactual assertions about what would be the case in alternative possible worlds. See POLLOCK, *supra* note 14, at 32-38, 42-43, 81-86, 132-40.

211. For the logical literature discussing formal, technical approaches to the problem of identifying the reference class, see REICHENBACH, *supra* note 12, at 372-78 (approaching the problem of the reference class by considering “the narrowest class for which reliable statistics can be compiled”); Levi,

The following hypothetical illustrates a typical direct inference to specific causation. Assume that scientific research warrants the finding that exposure to a particular chemical has a causal link to a certain kind of cancer in humans. Assume further that there are studies with acceptable measurement, sampling, modeling, and causal uncertainty that show that the baseline risk of the cancer in the general population is about four percent and that a particular degree of exposure to the chemical can cause the cancer in an additional six percent of exposures (the attributable risk).<sup>212</sup> That is, ten percent of all people exposed are expected to develop the cancer, although four percent of them would have developed the cancer even in the absence of exposure. There is additional evidence supporting the causal relevance of other risk factors, such as age, sex, a history of certain other diseases, and a certain genetic makeup. In addition, there is good evidence that the known risk factors do not explain all of the variability observed in actual cases of exposure. Because the causal mechanism linking the exposure to the production of cancerous cells is unknown, there is no explanation why ninety percent of those exposed do not develop the cancer at all, or why six percent develop exposure-caused cancer. The question in this part of the Article is when such warranted statistical findings about groups also warrant a probabilistic finding of specific causation about a particular member of the reference group—for example, a plaintiff named Jessica Jones.

In tort law, the cases that present problems in finding specific causation fall into two categories. In the first category, prospective specific causation is the issue. Examples are cases claiming risk or fear of cancer.<sup>213</sup> The plaintiff Jessica has been exposed to the chemical but has not yet developed the cancer. The reference group *A* consists of people with exposure similar to Jessica's, and the legally important subgroup *B* consists of those who will develop cancer as a result of the exposure. In the hypothetical above, those members of *A* who are also members of *B* comprise six percent of *A*. The second category of tort cases presents the issue of retrospective specific causation. Not only has Jessica Jones been exposed, but she has also developed the relevant kind of cancer. The reference group *A* becomes those who have been exposed *and* who later develop cancer, and the legally important subgroup *B* is the group of *exposure-caused* cancer cases.<sup>214</sup> In the hypo-

---

*supra* note 12 (considering three approaches to selecting reference sets for direct inference); *Direct Inference*, *supra* note 14, at 14, 20 (arguing that direct inferences should always rest on using the "narrowest reference class"); and Kyburg, *supra* note 12.

212. On attributable risk, see *supra* note 110 and accompanying text.

213. *E.g.*, *Potter v. Firestone Tire and Rubber Co.*, 863 P.2d 795, 816 (Cal. 1993) (holding that, as a general rule, in the absence of a present physical injury, a plaintiff can recover damages for fear of future cancer only if the plaintiff proves, among other things, that "the plaintiff's fear stems from a knowledge, corroborated by reliable medical or scientific opinion, that it is more likely than not that the plaintiff will develop the cancer in the future due to the toxic exposure"); *Mauro v. Raymark Indus., Inc.*, 561 A.2d 257, 264 (N.J. 1989) (allowing damages for an enhanced risk of developing cancer in the future only if the plaintiff establishes "the future occurrence of cancer as a reasonable medical probability").

214. For technical discussions of models with "latent" or unobservable variables, see JOHN C. LOEHLIN, *LATENT VARIABLE MODELS: AN INTRODUCTION TO FACTOR, PATH, AND STRUCTURAL ANALYSIS* (1998); and PEARL, *supra* note 194, at 41-64.

thetical above, the members of *A* who are also members of *B* comprise sixty percent of this refined reference group *A*. On average, out of every ten people who are exposed *and* later develop the cancer, four are baseline cancer cases and six have exposure-caused cancer. The question is under what conditions a reasonable factfinder can use such causal generalizations to warrant a direct inference about Jessica Jones.

Beyond the uncertainties inherent in finding the exposure to be causally relevant to the cancer at all, and inherent in quantifying any increased risk for those exposed, there are additional uncertainties created by using such generalizations to make a direct inference. First, there are uncertainties associated with identifying an appropriate reference group *A* to represent the specific plaintiff ("plaintiff-representativeness"). Second, even assuming an acceptable reference group, there is a potential for error in making an inference from acceptable statistics for that reference group to a probability for the particular plaintiff's situation. The next section addresses the first kind of uncertainty, while the subsequent section addresses the second kind of uncertainty.

*A. Acceptable Uncertainty About Plaintiff-Representativeness:  
Selecting an Adequately Representative Reference Group*

In a direct inference, the major premise makes an assertion about a general causal relationship between the reference group identified as *A* and a subgroup identified as *B*.<sup>215</sup> It also asserts the relative frequency or proportion of *B*s within the reference group *A*. Using a relative-frequency interpretation of conditional probability,  $Prob(B|A)$  refers to the expected relative frequency with which individuals within the reference group *A* are also members of *B*.<sup>216</sup> Different reference groups provide different denominators for the ratio  $B/A$ , and therefore may warrant different probabilities in the conclusion.<sup>217</sup> For example, the probability of exposure-caused cancer (*B*) brought about by a particular level of exposure to a certain chemical (*E*) may be higher or lower for men taking the drug (*M*) than for women taking the drug, and higher or lower for people who have a certain gene (*G*), or who have a certain disease in their medical history (*D*). Different combinations of these characteristics produce different reference groups—for example, exposed men with the gene (*E & M & G*), exposed men with the back-

---

215. The *Reference Guide on Epidemiology* for federal judges makes a similar point: However, before an association or relative risk is used to make a statement about the probability of individual causation, the inferential judgment . . . that the association is truly causal rather than spurious is required: "[A]n agent cannot be considered to cause the illness of a specific person unless it is recognized as a cause of that disease in general."

Green et al., *supra* note 6, at 383-84 (quoting Philip Cole, *Causality in Epidemiology, Health Policy, and Law*, 27 ENVTL. L. REP. 10,279, 10,281 (1997)).

216. See SKYRMS, *supra* note 20, at 201.

217. To simplify the symbolism, "*B*" denotes the subgroup of *A* that is of interest in the direct inference. For example, *B* is not the group characterized simply by developing cancer, but rather the subgroup of exposed people with exposure-caused cancer.

ground disease ( $E \& M \& D$ ), or exposed women with the gene and the disease ( $E \& \text{not-}M \& G \& D$ ). In addition, the selection of the reference group to use may produce different probabilities of occurrence (for example,  $\text{Prob}(B|C \& M \& G) = 0.07$ ,  $\text{Prob}(B|C \& M \& D) = 0.08$ ,  $\text{Prob}(B|E \& \text{not-}M \& G \& D) = 0.05$ ).<sup>218</sup>

For example, in the case of human cancer caused by exposure to carcinogens, there can be a high degree of variability in response due to different environmental factors and to variability in human susceptibility.<sup>219</sup> Although differences in susceptibility are known to correlate with such factors as age, sex, race, and ethnicity, the causal reasons are not well understood, and scientists must estimate "the form and breadth of the distribution of interindividual variability" by combining data on combinations of particular factors (a "bottom-up" method) and by using heterogeneity-dynamics models to explain demographic data (a "top-down" approach).<sup>220</sup> Some causal factors that affect susceptibility are prevalent in the population, while others are rare.<sup>221</sup> Some factors have only marginal effects on relative risk, while others substantially increase relative risk.<sup>222</sup> Some factors play minor roles in combination with a large number of other genetic, environmental, and lifestyle influences, with the net result being an essentially continuous, random distribution, while other factors or combinations tend to bias susceptibility upwards.<sup>223</sup> Some factors in combination might increase risk additively, while others might increase it multiplicatively.<sup>224</sup>

The general problem is providing a criterion for the set of factors that should define the reference group for the direct inference.<sup>225</sup> In a gambling context, there is prior agreement on the reference group of events—that is, on what will count as a valid lottery draw or a valid coin toss. In a transpar-

218. Relative risk would vary as well. As one researcher has stated:

In theory, variation in relative risks with background risk could be examined with epidemiologic data if the data were so extensive and accurate that one could validly estimate variation in background risk across the myriad subgroups of risk factors (age, sex, occupation, genetic susceptibility, etc.). Unfortunately, epidemiologic data are rarely so extensive and accurate, and, as a consequence, they rarely indicate the potential range of variation in relative risks.

Greenland, *supra* note 110, at 1168.

219. See NAT'L RESEARCH COUNCIL, *supra* note 25, at 196-203, 206 (stating that "[v]ariability affects each step in the carcinogenesis process (e.g., carcinogen uptake and metabolism, DNA damage, DNA repair and misrepair, cell proliferation, tumor progression, and metastasis)").

220. See *id.* at 200-03, 206-10; Greenland, *supra* note 110, at 1168 (stating that "epidemiologic data cannot establish the absence of individuals who are exceptionally vulnerable to exposure effects and who constitute a subgroup with an exceptionally high relative risk"); Muin J. Khoury et al., *On the Measurement of Susceptibility in Epidemiologic Studies*, 129 AM. J. EPIDEMIOLOGY 183 (1989) (analyzing the complexities of estimating the proportion of persons in a population who are "susceptible" to a risk factor).

221. NAT'L RESEARCH COUNCIL, *supra* note 25, at app. H-2.

222. *Id.*

223. See *id.* at 201-02, 206-09 (concluding that "[a] 10-fold [upward] adjustment [of risk] might yield a reasonable best estimate of the high end of the susceptibility distribution for some pollutants when only a single predisposing factor divides the population into normal and hypersusceptible people").

224. See *id.* at 226-29; Henifin et al., *supra* note 180, at 476 & n.136.

225. See ABDUCTIVE INFERENCE COMPUTATION, PHILOSOPHY, TECHNOLOGY, *supra* note 11, at 27; PROBABILITY AND INDUCTIVE LOGIC, *supra* note 198, at 78-82; REICHENBACH, *supra* note 12, at 372-78.

ent gambling setup, the probability (expected relative frequency) of each type of outcome is available equally to all players. Moreover, the gaming mechanism is calibrated to produce sequences of outcomes that are unbiased with respect to those relative frequencies and which have an acceptable range of random variation around those relative frequencies. Also, there is negligible uncertainty about how to classify the individual outcome events into the categories that determine the payoffs. Too much subjectivity in declaring the winner could bias the results, assuming self-interested behavior by the referees or judges. In a gambling context, therefore, participants often try to create and operate a gaming mechanism that ensures that direct inferences to specific outcome events are warranted. Put another way, in a fair and transparent gambling context, the gaming process is designed precisely to warrant direct inferences to specific outcomes.

In tort litigation, however, such an approach to warrant is not possible. The causal systems at issue cannot be manipulated to warrant direct inferences, nor can prior agreements precisely identify the reference group and outcome classes. In particular, it is not the reference group that is a "given," but rather the individual plaintiff. In tort cases, the reference group must be selected to match the plaintiff, not the plaintiff selected to match the reference group. The task is identifying a reference group that will help explain the causal connection (if any) between the specific plaintiff and the plaintiff's identified type of injury (such as a particular type of cancer).<sup>226</sup> The appropriate reference group to use for a direct inference about a specific individual is a function of which additional characteristics of that individual are causally relevant to being in subgroup *B*. As new variables are added to refine the reference group, the expected relative frequency of exposure-caused cancer in that group might fluctuate, either increasing or decreasing.<sup>227</sup> If the expected relative frequency of exposure-caused cancers in a reference group is biased due to confounding causal variables, then any probability in the conclusion of the direct inference will also be biased. There may be confounding unless the reference group adequately represents Jessica on all of the variables that in Jessica's case are causally relevant to Jessica's type of harm. To the extent that a possible reference group does not adequately represent Jessica's measurement values on enough of the

---

226. For the same degree of exposure, the relative risk may well depend on the nature of the injury. See, e.g., David A. Freedman & Philip B. Stark, *The Swine Flu Vaccine and Guillain-Barre Syndrome: A Case Study in Relative Risk and Specific Causation*, 64 LAW & CONTEMP. PROBS. 49, 54-55 (2001) (for the association between swine flu vaccination and Guillain-Barre syndrome, a data analysis showed a strong association for cases with extensive paralysis, but little evidence of association for cases with limited paralysis).

227. See *supra* notes 112, 160-64, 176-80 and accompanying text; Freedman & Stark, *supra* note 226, at 54-57 (discussing how, when the data relating swine flu vaccination and Guillain-Barre syndrome are "stratified" both on time of vaccination and time since vaccination . . . the relative risk for late-onset cases is well above 2.0"); Goldstein & Henifin, *supra* note 8, at 422-31 (discussing factors to consider concerning a "[s]pecific [c]ausal [a]ssociation [b]etween an [i]ndividual's [e]xposure and the [o]nset of [d]isease"); Henifin et al., *supra* note 180, at 461-78 (discussing factors to consider in evaluating causation in a specific case); *Daubert v. Merrell Dow Pharm., Inc.*, 43 F.3d 1311, 1321 n.16 (9th Cir. 1995) (giving the example of refining a reference group and adjusting the pertinent relative risk).

causally relevant risk factors, both known and unknown, then those statistics may provide an inaccurate and biased probability for Jessica's case.<sup>228</sup> The possible reference groups for which statistics happen to be available may not include the reference group that is appropriate for a direct inference to Jessica's case. The adjusted statistics for a more refined reference group, one that is more representative of Jessica, might be either higher or lower than the crude relative risk for a group identified without the added variables.<sup>229</sup> A reasonable factfinder must decide whether a proposed reference group is acceptably representative of the particular plaintiff and whether uncertainty on this issue is within acceptable bounds.

This notion of "refining" the reference group plays a critical role in the analysis of direct inference. A crude reference group is refined by adding causally relevant variables and values of variables that are significant factors in the particular plaintiff's case. Refining the reference group to represent the plaintiff could proceed by creating a reference-group profile for the plaintiff. The first step in creating such a profile is identifying the variables that are known or suspected to be causally relevant to the type of injury relevant in the plaintiff's case. For example, if the plaintiff Jessica has a particular type of cancer, then her reference-group profile is a list of all known or suspected risk factors for that type of cancer. Each assertion of general causal relevance for any particular risk factor is subject to all of the inherent uncertainties discussed in Part I. The next step is classifying or measuring the plaintiff under the appropriate category for each of the causally relevant variables. The plaintiff's combination of values or classifications for the set of variables is the reference-group profile for the plaintiff.<sup>230</sup> That profile then defines or identifies a reference group that is repre-

---

228. See Freedman & Stark, *supra* note 226, at 50, 60-62 (discussing the evidence in a particular case, and arguing that "the proof of specific causation, starting from a relative risk of four, seems unconvincing" once individuating factors are taken into account).

229. A committee of the National Research Council, in a statutorily mandated review of the EPA's carcinogenic risk assessment methods for hazardous air pollutants, urged that a research priority should be "[t]o explore and elucidate the relationships between variability in each measurable [cancer-susceptibility] factor (e.g., DNA adduct formation) and variability in susceptibility to carcinogenesis," so that the research results could be used "to adjust and refine estimates of risks to individuals (identified, identifiable, or unidentifiable) and estimates of expected incidence in the general population." NAT'L RESEARCH COUNCIL, *supra* note 25, at 207.

230. It is important to distinguish the causally relevant variable (such as sex, age, or genotype) from the plaintiff's value or score in the classification categories of that variable. See, e.g., *Siren Songs*, *supra* note 26, at 574-80. If the variable is sex, then the plaintiff's classification under that variable might be "female." For the variable age, the plaintiff's value might be "fifty-four" or "fifty to fifty-five," depending upon the measurement categories used in the relevant studies. If the variable is genotype, then the plaintiff's classification or score might be "negative" on the test for a particular allele.

If a plaintiff's value on a causally relevant variable is negative, indicating an absence of a causally relevant event or characteristic, then the refined reference group would include only individuals who, like the plaintiff, do not have that characteristic. Many commentators and courts conceptualize this not as selecting a representative reference group, but as "ruling out" or eliminating alternative causes. See, e.g., Henifin et al., *supra* note 180, at 468-78 (discussing steps in "[d]etermining external causation" of a specific individual's medical condition); Sanders & Machal-Fulks, *supra* note 6, at 122-25 (discussing judicial rulings concerning "ruling out" possible causes); Alani Golanski, *General Causation at a Crossroads in Toxic Tort Cases*, 108 PENN. ST. L. REV. 479, 500-04, 522-23 (2003) (arguing that a differential diagnosis "is even more likely than an epidemiological study to be accepted as the sole

sentative of the individual plaintiff relative to the particular type of injury. A reference-group profile for the plaintiff on exposure-caused cancer of a certain type, therefore, consists of people with values similar to the plaintiff's values on all of the variables that are causally relevant to developing that type of cancer after an exposure like the plaintiff's.

The added uncertainty about plaintiff-representativeness, therefore, is whether the reference group used in the major premise adequately matches the plaintiff's reference-group profile. This problem is reminiscent of the completeness problem of general causation, discussed in Parts I.C and I.D, but it is not the same problem. The problem of completeness in general causation is whether taking some confounding variable into account would affect the statistical significance of a relative risk or partial regression coefficient, or affect the strength of association to such an extent that it would undermine a finding of general causation. The problem of plaintiff-representativeness is whether the reference group adequately represents the plaintiff on enough of the causally relevant variables and has suitable statistics for warranting a direct inference (that is, with acceptable measurement, sampling, modeling, and causal uncertainty). The reference group *A* must *both* adequately represent the plaintiff *and* have acceptable measurement, sampling, modeling, and causal uncertainty in a general causal sense. Of course, it is possible that the studies that happen to be available do provide warranted causal statistics for such a plaintiff-representative group, but in practice this may rarely be true. The available studies seldom have a large enough sample to support a warranted probability about people like a particular individual. For example, even a large epidemiologic cancer study might have too few people like Jessica in the study to provide acceptable statistical power—that is, too few women of Jessica's age, genetic makeup, medical history, and other causally relevant characteristics. Moreover, in many cases, enough is known about which factors are causally relevant for the factfinder to conclude that the available statistics are *not* plaintiff-representative. The argument in this Article is that in such cases, any direct

---

evidence of causation in a case," and that even when a "valid" differential diagnosis fails to "eliminate all other causal factors that may have contributed to a plaintiff's disease," it should overcome a court's "reservations about the admissibility or probativeness of epidemiological studies finding relative risk ratios greater than 1.0 but not greater than 2.0"); *Stevens v. Sec'y of HHS*, No. 99-594V, 2001 WL 387418, at \*26 (Fed. Cl. Mar. 30, 2001) (proposing and using a requirement that a claim under the National Vaccine Injury Compensation Program will not be successful unless "petitioners . . . affirmatively demonstrate by a preponderance of the evidence that there is no reasonable evidence that an alternative etiology is the more probable cause of the alleged injury").

Courts frequently call this process of ruling out alternative possible causes "differential diagnosis" or "differential etiology." *Henifin et al.*, *supra* note 180, at 470 n.112; *Sanders & Machal-Fulks*, *supra* note 6, at 120-29 (discussing judicial rulings on the admissibility of differential diagnosis testimony); Gary Sloboda, *Differential Diagnosis or Distortion?*, 35 U.S.F. L. REV. 301, 308-23 (2001) (surveying federal cases ruling on the admissibility of expert testimony on causation that is based on differential diagnosis). As explained in this Article, however, this process of considering causes other than one for which the defendant is responsible is merely one aspect of selecting an acceptably representative reference group.



inference at all must rest on non-scientific decisions about the acceptability of the uncertainty about plaintiff-representativeness.

One extreme position would be that no direct inference is ever warranted unless the statistical generalization takes into account *all* of the factors that are causally relevant in the particular plaintiff's case.<sup>231</sup> That is, a direct inference must be based on consideration of the total or complete evidence that is relevant.<sup>232</sup> Of course, if there were a complete list of causally relevant variables, and if the plaintiff's values or classifications on every variable were known, and if that complete reference-group profile identified a reference group for which a methodologically acceptable study had taken all of the causally relevant factors and values into account, then the statistical evidence would warrant a direct inference.<sup>233</sup> Some causal processes are in fact so simple, well-known, and stable that generalizations assigning probabilities of occurrence are warranted. For example, a highly and acutely poisonous compound may be so disruptive of human metabolism that it is (almost) universally fatal if ingested at a certain dose. Other causal mechanisms are so well understood that we do not even require empirical support from scientifically designed studies. An example is the inability of humans to survive for very long without oxygen. But, many causal processes currently at issue in tort cases exhibit substantial, unexplained variability in outcome, especially in humans. A lifetime of heavy cigarette smoking, for example, may or may not cause lung cancer in individual cases, despite the existence of sound evidence of general causation. In such cases, if courts were to require plaintiffs to produce evidence that meets the "complete evidence" requirement, then no direct inferences about specific causation would be warranted. The substantive objectives of tort law could never be achieved under such a decision rule. In practice, courts consider many direct inferences sufficiently warranted for legal purposes, even though the evidence is not known to be complete or is known to be incomplete.

An alternative to the "complete evidence" decision rule is to develop a criterion for comparing the completeness of two alternative reference-group profiles and then use the more complete profile as the evidentiary basis of the inference. In this way, there could at least be marginal progress toward adequacy and a basis for determining which direct inferences are more warranted than others. For example, Reichenbach proposed making the infer-

---

231. Pollock posits the "total evidence" requirement that a direct inference must take into account every relevant proposition about the specific individual that the factfinder is warranted in believing. See POLLOCK & CRUZ, *supra* note 16, at 98-104; POLLOCK, *supra* note 14, at 132-34, 136-37.

232. SALMON, *supra* note 16, at 90-91.

233. Another line of reasoning infers the probable cause by eliminating or "ruling out" all other causes, thereby conducting a "differential diagnosis." See *supra* note 230. The degree of warrant in such an inference depends upon the completeness of the list of possible causes. See Susan R. Poulter, *Science and Toxic Torts: Is There a Rational Solution to the Problem of Causation?*, 7 HIGH TECH. L.J. 189, 232-35 (1992) (arguing that unless "most causes of the disease in question are known . . . the elimination of other risk factors would not significantly increase the likelihood that the toxic exposure was the cause of the plaintiff's disease"); Sanders & Machal-Fulks, *supra* note 6, at 133-34.

ence on “the narrowest class for which reliable statistics can be compiled.”<sup>234</sup> A “narrower” class would be constructed simply by adding more variables to the profile that decrease the number of individuals who satisfy the profile. For example, the class of “women over forty years of age who have a particular gene” is a narrower class than the class of “women over forty years of age.”

There are several epistemic problems, however, with such a rule. First, Reichenbach’s rule provides, at best, a purely comparative criterion. A more complete reference-group profile yields a direct inference that has a higher degree of warrant than a direct inference has from a less complete reference group. But even if this is true, such a comparison leaves unsolved the problem of identifying a minimal threshold that a reference class must cross before it can warrant a direct inference with acceptable uncertainty. This “threshold problem” is deciding when the reference-group profile takes into account sufficient numbers or types of relevant factors, such that the reference class *adequately* represents the specific individual and warrants a direct inference. In other words, the threshold problem is deciding when the reference-group profile contains enough causally relevant variables to warrant any direct inference at all. Whether a causal model is adequately representative of a particular plaintiff depends upon the degree of reference-group uncertainty that is acceptable in the particular legal context.

A second epistemic difficulty with Reichenbach’s rule is that adding variables that turn out to be causally irrelevant is not always harmless. That is, merely adding variables to the profile does not always produce more accurate statistics and more appropriate direct inferences. As discussed in Part I.C,<sup>235</sup> adding variables that are in fact irrelevant to the dependent variable tends to increase the random sampling uncertainty for the variables that are relevant and may increase the likelihood of Type I errors.

A third difficulty, as Reichenbach recognized, is that “reliable statistics” are often unavailable when reference groups become narrower.<sup>236</sup> There are often additional variables that are known to be causally relevant, but their addition to the profile would define a reference group for which there are no warranted statistics about risk. So Reichenbach’s rule is little more than a pragmatic maxim to take into account as many factors as possible, without outrunning the available statistics that have acceptable uncertainty. According to this rule, whenever adequate data come to an end, the factfinder should use the last, generally acceptable estimate of risk, even though there is good evidence that the estimate is inaccurate in the plaintiff’s case. But if

---

234. REICHENBACH, *supra* note 12, at 374.

235. See *supra* notes 112, 160-62, 178, 227 and accompanying text.

236. REICHENBACH, *supra* note 12, at 375. Even if there is no necessary connection between the size of the reference group and the acceptability of the statistics for that group, in practice there may be a correlation. See ABDUCTIVE INFERENCE COMPUTATION, PHILOSOPHY, TECHNOLOGY, *supra* note 11, at 27 (stating that “[t]here is almost always a certain arbitrariness about which reference class is chosen as a base for the probabilities; the larger the reference class, the more reliable the statistics, but the less relevant they are; whereas the more specific the reference class, the more relevant, but the less reliable”).

Reichenbach's maxim is used as a rule of decision in tort cases, then available statistics should always win out over individualized factfinding. The "statistical individual" would always trump individualized decision-making. As a factfinding rule, however, it is unclear what the rule's policy justification would be. As argued in Part I, leaving causally relevant variables out of account creates uncertainty about the stability of any observed relative risk or partial regression coefficient. Adding a causally relevant variable to the model could either increase or decrease the prior estimate of relative risk. When these lessons are applied to a reference group whose function is to adequately represent a specific plaintiff, then leaving causally relevant variables out of account undermines the warrant for making any direct inference. In the context of warranting direct inference, statistics based on an inadequately representative reference group do not warrant a direct inference at all. If the reference group "for which reliable statistics can be compiled,"<sup>237</sup> to use Reichenbach's phrase, is known not to be adequately representative of Jessica Jones, then those statistics alone cannot warrant any direct inference about her.

If the only acceptable statistics about general causation describe a reference group that is not (or is not known to be) adequately representative of a particular plaintiff relative to the relevant type of injury, then any justification for finding specific causation must rest on substantive policy grounds, not on logical or epistemic grounds. The issue of specific causation cannot be purely factual or scientific. This is true regardless of the magnitude of the relative risk that happens to occur in the available statistics. To the extent that a reference group is unrepresentative of the plaintiff, no relative risk estimate based on that group is warranted when applied to the plaintiff.<sup>238</sup> A relative risk of ten is no more warranting than a relative risk of 1.5, *unless* there is good evidence that it is also a good estimate of relative risk in a reference group that is adequately representative. The degree of support for direct inference is directly related to plaintiff-representativeness, not to the magnitude of the relative risk estimate for the reference group.<sup>239</sup> Unless a finding of plaintiff-representativeness is warranted, then there is no epistemic justification for relying on any particular relative risk at all, regardless of its magnitude. In cases involving substantial uncertainty about plaintiff-representativeness, any finding of probability about specific causation must rest on non-epistemic as well as epistemic grounds.

The ideal evidence to support a direct inference about Jessica Jones would be well-designed studies with subjects and controls matched to her and her circumstances on all variables that are causally relevant to the out-

---

237. REICHENBACH, *supra* note 12, at 374.

238. Cf. Freedman & Stark, *supra* note 226, at 61-62 (concluding, in a case where "the plaintiff is in crucial detail remarkably unlike the other GBS victims" that were studied, "[e]pidemiologic data cannot determine the probability of [specific] causation in any meaningful way because of individual differences").

239. The magnitude of the relative risk estimate is, however, one factor among many relevant in evaluating general causal uncertainty. See *supra* Part I.D.

come. The possibility of conducting such an ideal study, however, is limited by time, resources, and politics, as well as by the ethical constraints on human experimentation. Such an ideal study also faces methodological problems. While the causal relevance of some variables may be known and the relevance of others suspected, there may also be good evidence that there exist many unknown and unstudied causes. Moreover, the factfinder usually has, or can obtain, a great deal of information about the specific individual—what the courts call “particularistic evidence.”<sup>240</sup> Such evidence may describe observable physical features, medical factors (such as metabolic data), behavioral characteristics (such as dietary patterns), personal history (such as environmental exposures and history of disease), family history (such as genetics), and so forth. The question is which of this particularistic evidence is causally relevant to the outcome in the specific case. The answer to this question is generally unavailable. Considering the extent of human variability, it is unlikely that the factfinder will have adequate causal evidence on a group of people sufficiently similar to the plaintiff on even known or suspected causal variables, let alone on unknown causal variables.

A reasonable factfinder must decide whether the residual uncertainty in selecting a representative reference group is acceptable in the particular tort case, after evaluating the statistical evidence that is available. But, in law—as in science and everyday life—factfinders need not always suspend fact-finding until they have a *complete* reference-group profile and *complete* knowledge of general causal factors, provided they can decide that a causal model is *sufficiently* complete and representative for the practical purposes at hand. With some kinds of phenomena, the unexplained individual variability may be minimal. With other injuries, experts may determine that a few causal factors are so dominant in the individual case that refining the reference group further is unlikely to change the legal finding. In other situations, the causal understanding may be so incomplete and the individual variability so pronounced that any direct inference to the specific case would incur a very high degree of uncertainty. Where along this continuum of uncertainty the factfinder should draw the line of acceptability must be a matter of common sense and public policy.<sup>241</sup> When the epistemic ideal of a completely representative reference group and complete empirical evidence is unattainable, substantive legal policies might also justify rules allocating the burden of proving or disproving reference-group acceptability. Once the plaintiff does the best she reasonably can to reduce reference-group uncer-

---

240. See, e.g., *In re Joint E. & S. Dist. Asbestos Litig.*, 52 F.3d 1124, 1130 (2d Cir. 1995).

241. In clinical medicine, a parallel problem is deciding whether to treat a patient's condition using a diagnosis based on incomplete information or whether to wait pending more tests. See KASSIRER & KOPELMAN, *supra* note 10, at 24-27 (stating that “[t]he trade-offs between the risks and benefits of tests and treatments are embodied in the threshold concept,” and analyzing medical decision rules for testing and treating particular patients in a clinical setting). In law, a factfinder's decision might be influenced by the likelihood that additional evidence would alter the provisional inferences that rest on the currently available evidence, and by whether the lack of additional evidence is fairly chargeable to a particular party in the litigation.

tainty, there may be good policy reasons to shift away from chasing an unattainable epistemic ideal and toward achieving non-epistemic objectives.

*B. Acceptable Uncertainty About Assigning a Probability to a Specific Member of the Reference Group*

Even if the reference group in the statistical major premise is refined to the point where it is acceptably representative of the specific plaintiff, there will be additional uncertainty in using the statistics about that group as the basis for assigning a probability to the specific case—at least, as long as the premise does not assert a universal causal relationship from being *A* to being *B* (that is, “all *As* are also *Bs* as a result of being *As*”).<sup>242</sup> Even if 90% of the members of reference group *A* have characteristic *B* as a result, this might not warrant assigning a 0.9 probability to the proposition that a specific member of *A* is also a *B*. This section examines individual-probability uncertainty, or the residual potential for error in making such an assignment of probability to a specific plaintiff’s case. One question is whether there is any epistemic basis at all for warranting a probability assignment to the individual case.<sup>243</sup>

The first, intuitive response might be that this inference problem is no different than assigning a probability to the next outcome in a game of chance. Imagine a lottery machine in which a forced air stream mixes lightweight plastic balls in a chamber until a vacuum mechanism selects one of them. Suppose there are one hundred colored balls in the lottery machine, and that six of the balls are red (6%), four are white (4%), and the remainder yellow (90%), to use the same percentages as in the illustration involving the plaintiff Jessica Jones. The probability of drawing a red ball on the next draw is 0.06 if the chance setup gives every individual ball an equal chance of being drawn. When this condition is met, the probability of drawing a red ball on the next draw is equal to the proportion of red balls in the lottery machine. This reasoning employs a “probability-of-selection” warrant for assigning a probability of occurrence to the next outcome event. It treats the

242. See Greenland, *supra* note 110, at 1168-69 (arguing that “[a]ll epidemiologic measures (such as rate ratios [relative risks] and rate fractions [attributable risks]) reflect only the net impact of exposure on a population, rather than the total number of persons affected by exposure,” and that the attributable risk is not generally equal to the “probability of causation”); James Robins & Sander Greenland, *The Probability of Causation Under a Stochastic Model for Individual Risk*, 45 *BIOMETRICS* 1125, 1128, 1133, 1134-35 (1989) (arguing that the “probability of causation” in a population is not identifiable from epidemiologic data in the presence of heterogeneity in background disease risks).

243. Some authors who take a causal approach to warranting direct inference describe it as an inference from “indefinite physical probabilities” to “definite probabilities.” See POLLOCK & CRUZ, *supra* note 16, at 98-100; POLLOCK, *supra* note 14, at 20-22 (distinguishing between “indefinite probabilities” (such as “the probability of a smoker getting lung cancer, or the probability of its raining when we are confronted with a low pressure system of a certain sort”) and “definite probabilities” (such as “the probability that Jones will get cancer, or the probability that it will rain tomorrow”). Others follow Karl Popper in interpreting probabilities as characterizing the “propensities” of objects or types of objects to behave in certain ways under certain conditions. See, e.g., *THE PROBABLE AND THE PROVABLE*, *supra* note 200, at 21-24, 295-309; *PROBABILITY AND INDUCTIVE LOGIC*, *supra* note 198, at 46-51; *SCIENCE AND REASON*, *supra* note 20, at 39; POLLOCK, *supra* note 14, at 23-32; SKYRMS, *supra* note 20, at 199-205.

next draw as a sample of one and uses reasoning similar to the rationale behind sampling theory (discussed in Part I.B). This rationale rests upon having a warranted probability of selecting an individual or sample from the reference group.

There are two elements to this type of warrant. First, there must be warrant for the proportion or percentage of red balls in the machine. This information about the group of balls may incur the kinds of statistical uncertainty discussed in Part I, such as measurement and sampling uncertainties. Second, there must be warrant for the premise that every individual ball has an equal chance (or some specified probability) of being selected on the next draw. Some selection processes might be designed to ensure drawing a white ball or might happen to be biased in favor of drawing red balls, and so forth. If nothing is known about the drawing mechanism, then there is no warrant for assigning a probability of selection. In order for a probability-of-selection approach to provide warrant, there must be an adequately specified causal model for the ball-selection process that warrants assigning a probability to the event of being drawn. Under a probability-of-selection pattern of warrant, the central issues therefore become: (1) the accuracy of, and warrant for, the descriptive statistics about the reference group from which the specific individual is drawn; and (2) the accuracy of, and warrant for, the probabilities assigned to the causal process of selecting individuals from this group.

The mathematical probabilities involved in probability-of-selection reasoning can be interpreted as expected relative frequencies. On a relative-frequency interpretation, the finding that "there is a 0.06 probability that the next ball drawn will be red" is really a statement about the expected relative frequency of drawing red balls on repeated tries of the same process. If the individuals in the group have an equal chance of being selected, then the proportion of individuals of each type in the group is the probability to assign to the next outcome of the selection process. Relative-frequency probabilities treat the selection process as a repeatable procedure, and any prediction of a specific selection event is interpreted as a prediction about the relative frequencies of the types of selection outcome. This interpretation works well when applied to gambling strategies because betting on the outcome events of a gambling mechanism can be a repeatable process. Lottery machines are designed and calibrated so that the probabilities in the premise will be acceptably accurate predictions for long series of outcomes. By design, therefore, there are no predictive factors for the next outcome that can improve on the probabilities in the premise. The objective behind a *fair* gambling game is to have the direct inference about the next event be available to all players and be the best available prediction for the next outcome. The direct inference in such a case is warranted by the way the gaming machine is built and operated, as well as by the relevant statistics on past series of outcomes.

Factfinders in law can sometimes employ a probability-of-selection rationale to warrant the last step in a direct inference. The composition of the

reference group may be known, and the process of selecting members from that group may have an adequately specified causal model that warrants assigning forward conditional probabilities to selection events. However, this rationale may also have limited usefulness for specific causation in typical tort cases. The evidence in a tort case rarely warrants the finding that the causal process of selecting plaintiffs *from* the acceptably representative reference group is either random or adequately modeled. When an individual plaintiff files a tort claim, causal factors in the plaintiff's decision process might bias the selection process and introduce uncertainty about confounding into the selection process.<sup>244</sup> For example, alarming publicity about an exposure event (such as a chemical release) might be positively correlated with anxiety in those people potentially exposed, causing an increase in the number of self-reports of subjective symptoms and a willingness to file tort suits claiming damages for emotional distress. Plaintiffs reporting such symptoms may not be "random draws" from a reference group identified merely by exposure without attendant publicity.<sup>245</sup> Although such publicity might introduce bias into plaintiff selection for emotional distress, it might not affect the risk of actually developing cancer. When courts adopt rules that limit certain types of damages depending on the level of exposure,<sup>246</sup> they may be reducing the risk of bias in plaintiff selection. Class action lawsuits can reduce plaintiff-selection uncertainty by using the reference group *A* to define the plaintiff class, which avoids selection bias by including all members of *A* as potential plaintiffs.<sup>247</sup>

An advantage of the probability-of-selection warrant is that it can ignore the causal model for creating members of the subgroup *B*. In the probability-of-selection rationale, the warrant depends only on the composition of the reference group and on the process of selection *from* the reference group. In the lottery machine example, the process of physically producing red balls is of no importance in assigning probabilities to selecting red balls. An important disadvantage, however, of the probability-of-selection rationale is that it may distort the justification behind the legal judgment. The typical tort case starts with the plaintiff as a given and tries to select an adequately representative reference group and to explain the causal connection between members of that group and the type of alleged injury. By contrast,

---

244. A confounding variable in this context would be correlated with both the event of becoming a plaintiff and the injury claimed in the lawsuit. See *supra* notes 56, 164, 178 and accompanying text.

245. Another way to view the problem is that when the alleged injury is emotional distress, then an adequate causal model for that injury (as opposed to cancer) will include publicity as a relevant causal factor.

246. E.g., *Mauro*, 561 A.2d at 260-67 (allowing damages for emotional distress and the cost of future medical surveillance, provided the jury finds that the plaintiff sustained an "asbestos-related injury," but distinguishing and not allowing damages for an enhanced risk of developing cancer in the future unless the plaintiff establishes "the future occurrence of cancer as a reasonable medical probability").

247. Plaintiffs who "opt out" of the class, however, will still encounter plaintiff-selection uncertainty. Moreover, a class action might merely defer plaintiff-selection uncertainty to the findings used to award damages to particular plaintiffs. In a settlement, the negotiated agreement might resolve the issue of plaintiff-selection uncertainty in a manner negotiated among plaintiffs themselves.

the probability-of-selection rationale takes the reference group as a given and studies the selection of the plaintiff from that group.

One suggested alternative to the probability-of-selection rationale is the "principle of indifference."<sup>248</sup> The principle of indifference is a decision rule that assumes that any unmodeled factors that are causally relevant to the specific individual's being a member of *B* are equally likely to be true or false for that specific individual. According to the principle of indifference, the factfinder should assume that such factors will offset each other in any particular case—that is, that unmodeled factors tending to increase the probability in the conclusion will offset unmodeled factors tending to decrease it. Proponents of this rationale, therefore, conclude that all unmodeled factors can be justifiably ignored for purposes of the direct inference.

The analysis in Parts I.C and I.D shows that this is a suspect principle. First, in general, the addition of a causally relevant factor to the model could adjust the relative risk either up or down.<sup>249</sup> Moreover, in many tort cases, there is positive evidence that using the principle of indifference is unwarranted. A significant amount of residual unexplained variability may demonstrate that the causal model is incompletely specified and that the reference-group profile for the plaintiff is incomplete. There may also be evidence to identify causally relevant factors in the particular plaintiff's case, but inadequate data to determine the magnitude of their causal influence, or incomplete statistical evidence of the plaintiff's combination of values on those factors. For example, there may be sufficient evidence to indicate that age and a history of diabetes are causally relevant factors for the alleged injury and that they probably change the relative risk, but insufficient evidence about how much those factors would adjust the relative risk for this plaintiff.<sup>250</sup> Moreover, there may be good evidence that unknown causal factors exist and that they do make a difference in individual cases. Adopting the principle of indifference is therefore often a decision simply to ignore the direct-inference problem.<sup>251</sup> It is also a decision to ignore the very factors that distinguish one individual from other individuals and may, in fact, be a policy of indifference toward individual litigants.

---

248. See *THE PROBABLE AND THE PROVABLE*, *supra* note 200, at 43-47; JOHN MAYNARD KEYNES, *A TREATISE ON PROBABILITY* 44 (1948); PORAT & STEIN, *supra* note 13, at 46-47, 178.

249. *Supra* notes 112, 160-64, 176-80, 227 and accompanying text.

250. For an example of a probabilistic attempt to adjust relative risk on the basis of individuating causal factors, see Freedman & Stark, *supra* note 226, at 60-61. That attempt also illustrates the additional uncertainty that is created.

251. One effect of the policy-based "eggshell skull" rule of tort law (that the defendant "takes the plaintiff as she finds him," even if the plaintiff is unusually susceptible to injury) is that the factfinder need not quantify, for purposes of finding liability, the causal contribution of particular risk factors that are internal to the plaintiff. See, e.g., DOBBS, *supra* note 23, § 177, at 433-34; KEETON ET AL., *supra* note 23, at 291-92. In calculating damages, however, a jury might deal with the residual epistemic uncertainty very differently after finding the defendant liable than it would have in the liability phase. Indeed, some judicial rules impose on the defendant the burden of proving that a portion of damages is not chargeable to the defendant because it is due to a preexisting condition of the plaintiff. See, e.g., RESTATEMENT (THIRD) OF TORTS: APPOINTMENT OF LIABILITY § 26 cmt.h (2000).



The appropriate solution, however, is to face squarely the logic of direct inference. There is often residual uncertainty in assigning any particular probability to a specific member of the reference group, even after deciding that the uncertainties about general causation and plaintiff-representativeness are within acceptable bounds. In a particular case, evidence might in fact support using the probability-of-selection rationale or the principle of indifference, and warrant assigning a probability with acceptable uncertainty. Whether the conditions of either of those rationales are sufficiently satisfied for tort purposes will depend on the goals and objectives of tort law. In the absence of such acceptable evidence, however, any probability assignment must be justified entirely by non-epistemic considerations.

That ultimate decision to use statistics from the reference group to assign a particular probability to the individual case, however, is *epistemically unwarranted* unless decisions are made about acceptable uncertainty. Unless the uncertainties are acceptable for the general causal link between the exposure variable and the alleged type of injury, unless the reference-group profile is acceptably complete and the reference group adequately represents the plaintiff, and unless the measurement, sampling, modeling, and causal uncertainties are within acceptable bounds for that representative reference group, there is no warrant for assigning any probability to a particular plaintiff based on that reference group. If the reference group is acceptably homogeneous with the plaintiff on enough causally relevant factors, and the statistical uncertainties are within acceptable bounds, then non-epistemic policies might support rules or decisions to fill the ultimate inferential gap. Courts are less likely to turn to such policy-based decision rules, however, if they fail to acknowledge that logical gap.

### III. MAKING WARRANTED FINDINGS ABOUT SPECIFIC CAUSATION

Parts I and II identified the kinds of uncertainty that are logically inherent in any direct inference to a conclusion about specific causation. This part of the Article develops those elements into a single, coherent approach for a reasonable factfinder to follow.<sup>252</sup> For each kind of uncertainty, a factfinder should decide how extensive the residual uncertainty is and whether that uncertainty is acceptable for the purposes of the tort case. Decisions about acceptability are necessarily pragmatic. Because they are not purely factual or scientific, they should be justified by the non-epistemic goals of tort law.<sup>253</sup> The first section of this part summarizes the decision structure for a warranted finding of specific causation. The second section illustrates

---

252. For a parallel approach in evidence-based medicine, see *supra* note 11.

253. Even with regard to the uncertainties inherent in findings of general causation (the uncertainties discussed in Part I), there may be good reasons why courts should not merely adopt scientific conventions. See Carl F. Cranor et al., *Judicial Boundary Drawing and the Need for Context-Sensitive Science in Toxic Torts after Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 16 VA. ENVTL. L.J. 1, 21-25 (1996) (discussing epistemological differences between tort law and science).

the practical implications of this analysis by discussing three types of cases in which judges have been making logical errors about specific causation. As a result of this mistaken reasoning, judges have been deciding these cases in ways that effectively remove the individuality of the plaintiff from the factfinding. A proper understanding of direct-inference warrant, however, shows that the solution is to restore the individuality of the plaintiff to the case through an affirmation of the jury's role as pragmatic decision-maker and an insistence on better policy rationales for judicial decision rules.

*A. An Integrated Approach to Decision-Making About  
Acceptable Residual Uncertainty*

Every direct inference rests upon a major premise about general causation.<sup>254</sup> That premise rests in turn upon empirical observations or data. The data result from classifying individual objects or events into categories of variables. Measurement uncertainty is the potential for error due to misclassification—that is, the possible error from placing an observed individual into the wrong category of a variable. As discussed in Part I.A, distinguishing biased measurement error from unbiased measurement error allows a reasonable factfinder to deal with the potentials for such errors in different ways. For example, if measurements are known to be unreliable to some degree (subject to random error), then taking a sufficiently large sample of repeat measurements and using an average of those measurements can increase the precision of the measurement. If measurements are known to be invalid or biased, then there may be some basis for adjusting measurements before relying upon them. For either kind of measurement uncertainty, however, a reasonable factfinder must decide whether the residual degrees of uncertainty are acceptable for the purposes of tort law. If the extent of reliability and validity is unknown, this is also an uncertainty that should be taken into account in deciding what inferences to make.

A major premise about general causation rests upon data drawn from a mere sample of all of the observable cases. As discussed in Part I.B, sampling uncertainty is the potential for error introduced by using data and statistics from a sample to warrant generalizations about all objects or events,

---

254. This Article does not directly engage the argument by some commentators that, for reasons of policy, plaintiffs should not have to prove general causation (at least in certain types of cases). For such arguments, see Margaret A. Berger, *Eliminating General Causation: Notes Towards a New Theory of Justice and Toxic Torts*, 97 COLUM. L. REV. 2117, 2117 (1997) (arguing for “abolishing proof of general causation” in toxic tort cases, because the causation model “is inconsistent with notions of moral responsibility underlying tort law”); and Sloboda, *supra* note 230, at 302, 315-16, 323 (arguing that eliminating the general causation requirement and “imposing strict specific causation standards” based on differential diagnosis could produce “a reliable, useful, and fair standard by which to appraise the admission of medical and scientific causation evidence in federal courts”). This Article examines the logic of a direct-inference warrant for specific causation and does not develop arguments based on tort policies. Of course, it is fair to expect those advocating the elimination of general causation on policy grounds to explain the logic of warranting specific causation without direct inference and general causation.

whether past, present, or future. The question is whether an actual sample is adequately representative of the population that becomes the subject of the generalization. The distinction between biased and random error applies to the sampling process as well as to the measurement process. Researchers may be able to ensure that samples are adequately representative of variables known to be important in the target population. Moreover, a researcher can use a randomizing procedure to eliminate any causal factors that could bias the sampling process itself, although such a procedure cannot guarantee the representativeness of every sample drawn. However, even if a sampling process were acceptably unbiased, chance alone could still produce a sample that is to some degree unrepresentative of the target population. Statistical significance, statistical power, and confidence intervals can characterize the residual, random sampling uncertainty. A large and randomly drawn sample, with statistically significant results, can help warrant a finding of general causation. A small sample, however, with low statistical power and no statistically significant results may possess too much sampling uncertainty for warranting any finding about general causation. A reasonable factfinder must decide whether the sampling process is acceptably unbiased, whether the risk of chance variations between the population and the sample is acceptable, and whether the total sampling uncertainty is acceptable.

Generalizations about causal influences rest not only on measurements of single variables for samples of individuals, but also on statistical associations among variables and categories of individuals. The major premise of a direct inference of specific causation asserts that some proportion of things in category *A* are also in category *B* as a result of being in category *A*. It rests upon evidence that being in category *A* is statistically associated with being in category *B*, and that knowing that an individual is a member of *A* increases the accuracy of a prediction that the individual is also a member of *B*. For purposes of direct inference to specific causation, the relative risk statistic is a useful measure of the strength of association between *A* and *B*. When generalized to a population, relative risk estimates the difference in risk of injury associated with an independent variable (such as exposure). Regression models can take many independent variables into account, and a partial regression coefficient of one variable (such as exposure) estimates the contribution of that variable to the overall model prediction, after the contributions of all the other independent variables in the model have been factored in. With each model and set of statistics, however, there are inherent uncertainties due to the modeling. Before relying upon statistical models in drawing conclusions, the factfinder should first decide whether the uncertainty created by using them is acceptable. One major source of uncertainty is whether the model takes into account enough of the variables that are causally relevant for the adverse event. When new variables are taken into account, statistics such as relative risk or regression coefficients may change. Another major source of uncertainty is the form of the model used. If the formal conditions of the model are not sufficiently satisfied, or if the

model fits the data poorly and leaves a substantial amount of individual variability unpredicted, then the model statistics may be misleading. Modeling error can lead to false predictions and a false conclusion about how being in category *A* causally relates to being in category *B*. A reasonable factfinder would evaluate the extent of modeling uncertainty and decide whether the residual uncertainty is acceptable for the purposes of tort law.

Finally, in order to warrant a conclusion about a general causation connection between *A* and *B*, it is not sufficient to conclude that groups of individuals characterized by *A* and *B* are statistically associated. The reasonable factfinder must add to the statistical model a causal interpretation. Causal relationships explain why some types of events occur after other types of events—not simply in the sample of events already observed, but also in the future and even in possible worlds that may never actually occur. Generalizations about a causal relationship between *A* and *B* are supported by an observed association between *A* and *B*, provided causal uncertainty is within acceptable bounds. The weight of evidence for a causal connection is influenced by several considerations. The degree of warrant increases as the statistical model takes more causally relevant factors into account and thereby adjusts for potentially confounding variables. The weight of evidence for causation also increases as the strength of the statistical association increases, as multiple studies produce consistent results, and as a mechanism of causation becomes more plausible. At a minimum, the temporal directionality from cause to effect must leave room for a more refined causal model to elicit a causal mechanism. In sum, a reasonable factfinder must decide whether the empirical studies and causal model take enough causally relevant factors into account, so that the relative risk or partial regression coefficient is sufficiently accurate and stable for legal purposes. Whenever residual causal uncertainty exists, the factfinder must decide how complete a model needs to be, or how much support the evidence needs to provide, before a finding of general causation is warranted in a legal context. In law, as in everyday life, how much evidentiary support is needed depends upon what is at stake and how difficult it is to obtain more evidence. Scientists can provide guidance on how much causal uncertainty there is and how additional research might reduce it, but substantial causal uncertainty is so pervasive that deciding when it is acceptable is rarely a purely scientific decision.

The four kinds of uncertainty just discussed (measurement, sampling, modeling, and causal) are inherent in any major premise that asserts a general causal connection between two variables. They are the kinds of uncertainty inherent in any finding that a particular variable is a risk factor for an injury or disease. They, therefore, necessitate decisions about acceptability for any finding that any particular variable is causally relevant to variable *B*. But a direct inference does not end with a finding of general causation. The next task for the factfinder is to identify all of the variables known or sus-

pected to be causally relevant to the plaintiff's type of injury and to gauge the extent of unexplained variability due to unknown causes.<sup>255</sup> The plaintiff's combination of values for all of the causally relevant variables constitutes the reference-group profile, which defines a reference group that is somewhat representative of the plaintiff.<sup>256</sup> The addition of causally relevant variables to the profile may change the relevant statistics of variables already in the profile (such as relative risk due to chemical exposure)—both substantially and in either direction. In many tort cases, the only available reference group with any statistics at all is known to be incomplete, and the statistics are known to rest on significant uncertainties. A reasonable factfinder must then decide whether the evidentiary warrant for directly inferring specific causation is "good enough" for the purposes of tort law. Deciding to base a direct inference on a particular reference group that is incomplete, but which has acceptable uncertainty in its statistics, is a decision that cannot be purely factual or scientific.

Once the factfinder identifies the most representative group for the particular plaintiff and decides that the group is representative enough for fact-finding purposes, the factfinder must construct and evaluate a major premise for the direct inference using the selected reference group as category A. That is, the factfinder should evaluate the available statistics for the selected reference group considering all of the types of general uncertainty discussed in Part I. Even if there is adequate scientific evidence to find a general causal relationship between some type of exposure and the plaintiff's type of injury, that does not resolve the issue of acceptable uncertainty for a reference group that adequately represents the particular plaintiff. The scientific study that is adequate to establish general causation for a particular risk factor might not provide a reference group that is adequately matched to the plaintiff on enough causally relevant variables. A warranted direct inference to specific causation must rest on statistics with acceptable uncertainty for a plaintiff-representative group.

If a factfinder decides that the reference group adequately represents the individual plaintiff and that the statistics causally linking that reference group to subgroup *B* have acceptable levels of uncertainty, then there is still uncertainty created by using those statistics as the basis for assigning a probability to the plaintiff's injury, as long as the causal model falls short of establishing a mechanism that completely explains every individual case. In some cases, information about the plaintiff-selection process might warrant

---

255. Many courts have insisted, of course, that the studies offered into evidence must be relevant to the plaintiff's situation. *See, e.g., Merrell Dow Pharm., Inc. v. Havner*, 953 S.W.2d 706, 720 (Tex. 1997) (ruling that "to survive legal sufficiency review," a "claimant must show that he or she is similar to those in the studies"). By focusing on the admissibility or sufficiency of the scientific evidence, and not on the logical structure of the direct inference itself, it is possible to miss the central point that the issue of specific causation—given significant uncertainties—can never be entirely scientific, or even entirely factual.

256. On the usefulness of causal modeling in clinical medicine, see KASSIRER & KOPELMAN, *supra* note 10, at 28-31.

treating the plaintiff as a purely random selection from the reference group. Then the probability to be assigned to the plaintiff's case can be warranted by the composition of the reference group and the random selection process from that group. This probability-of-selection warrant cannot, however, cure any lack of representativeness in the reference group itself. The warrant for applying any group statistics to the plaintiff's case must include the proposition that the reference group is acceptably representative of that individual case. Moreover, if the reference group is acceptably representative of, and adequately matched to, the specific plaintiff, then this reduces the need to rely on randomness at this stage of the direct inference. Randomization becomes less important as the reference group becomes more representative of the plaintiff. Although non-epistemic decisions or rules must often fill the inferential gaps created by incomplete causal models, the more representative the reference group, the more reasonable such reliance might appear.

The next section uses this analysis to critique the reasoning of courts in a variety of cases and procedural contexts. Too many courts have reasoned that, as long as there is no known mechanism that completely explains the plaintiff's injury, the plaintiff must prove that subgroup *B* of the reference group is larger than the baseline-risk subgroup. If the subgroup *B* were smaller or of equal size, so this reasoning goes, then a simple random draw from the reference group does not have a probability greater than 0.5 of selecting a defendant-created injury. Those courts therefore adopt a quantitative rule for plaintiff success: the plaintiff must prove that subgroup *B* is greater than 50% of the reference group *A* (a "greater-than-50%" rule), or (alternatively) that the relative risk of a defendant-created injury over all other relevant causes must be greater than 2.0 (a " $RR > 2.0$ " rule).<sup>257</sup> This quantitative test becomes a necessary condition for the plaintiff's success.<sup>258</sup> As the next section illustrates, courts have used this reasoning to make findings of fact in cases tried without a jury, to adopt rules about the legal sufficiency of evidence, and to develop exclusionary rules of evidence for proffered expert testimony.

For a number of reasons, however, it is misguided to impose a quantitative threshold on the basis of such reasoning, and courts should re-examine the basis for such quantitative rules.<sup>259</sup> First, as long as a significant amount

---

257. For surveys of court decisions using such rules, see Russell S. Carruth & Bernard D. Goldstein, *Relative Risk Greater Than Two in Proof of Causation in Toxic Tort Litigation*, 41 JURIMETRICS J. 195 (2001); Golanski, *supra* note 230, at 488-504; Green, *supra* note 72 (surveying the legacy of the Agent Orange and Bendectin litigation); and Lucinda M. Finley, *Guarding the Gate to the Courthouse: How Trial Judges Are Using Their Evidentiary Screening Role to Remake Tort Causation Rules*, 49 DEPAUL L. REV. 335, 347-64 (1999).

258. I have argued elsewhere that the preponderance standard of proof itself does not require such a quantitative test, and that the preponderance standard is perfectly intelligible without converting it into a greater-than-50% rule. Vern R. Walker, *Preponderance, Probability and Warranted Factfinding*, 62 BROOK. L. REV. 1075, 1094-1100 (1996) [hereinafter *Preponderance, Probability*].

259. Courts and commentators have long debated the use of such quantitative rules applied to particular types of evidence, such as epidemiology. See, e.g., *Merrell Dow Pharm. Inc.*, 953 S.W.2d at 715-21; Finley, *supra* note 257. The argument in this Article, however, is much broader. First, the uncertainties inherent in direct inference are logical in nature and are not limited to a particular type of evidence,

of individual variability remains unexplained by available causal models, there may be no epistemic warrant for relying on *any* statistics that happen to be available. The residual unexplained variability is good evidence that the causal model is incomplete and that there are causally relevant factors yet to be identified by science. If more of those factors were known, and the reference group for the plaintiff were appropriately refined to take them into account, then the relative risk within the reference group for exposed versus unexposed people could either increase or decrease.<sup>260</sup> An underlying causal mechanism may function as a confounder to explain many of the risk factors previously identified.<sup>261</sup> Unless the factfinder first determines that a proposed reference group is acceptably representative of the specific plaintiff, any relative risk for that group has no determinable probative value in the specific case. In tort cases where causal mechanisms are not completely understood, non-epistemic decisions are needed about whether the degree of plaintiff-representativeness is acceptable and whether the uncertainty in assigning a probability to the specific case is acceptable.

Second, any proposed reference group must have not only acceptable plaintiff-representativeness, but also acceptable levels of the general uncertainties discussed in Part I. That is, even an adequately representative reference group must have relative risk statistics that reflect the combination of risk-factor values present in the plaintiff's case. For example, if the plaintiff's genetic background and medical history present known risk factors, then the relevant relative risk is not the risk in the general population, but the risk in a susceptible sub-population of people sufficiently like the plaintiff. A relative risk for a single factor in the general population may be relevant to establishing general causation for the defendant-created exposure, but its probative value for drawing a direct inference in the particular case may be unknown. A factfinder would be guessing in drawing a direct inference about a specific plaintiff using such a general-population relative risk, unless he or she first decided that the general population was a sufficiently representative reference group for the purposes of tort law.

Third, decisions about the acceptability of the various kinds of inherent uncertainty are not scientific issues. Many of the types of uncertainty are unquantifiable, and scientists have not even established conventions for their own purposes.<sup>262</sup> Furthermore, the uncertainties are not confined to

---

such as epidemiologic evidence. Second, this logical analysis provides a systematic approach to all of the types of uncertainty involved. Therefore, this Article reaches a much broader and stronger conclusion about when legal decision rules concerning specific causation should be policy-based.

260. Any statistical value that happens to occur in the available studies is open to significant revision, either up or down, as knowledge of causal factors increases. Knowing neither the direction nor the magnitude of those adjustments, it may be judicially arbitrary to require the available evidence at any single point in time to satisfy any statistical threshold.

261. For discussions of confounding factors, see *supra* notes 56, 164, 178 and accompanying text.

262. A notable exception is the scientific convention to adopt a 0.05 level of statistical significance as the decisional probability for Type I errors that are due to random sampling uncertainty. See *supra* notes 72-75 and accompanying text. See also *Merrell Dow Pharm., Inc.*, 953 S.W.2d at 724 (deciding that it is "unwise to depart from the methodology that is at present generally accepted among epidemiologists,"

certain areas of science, such as epidemiology.<sup>263</sup> As this Article demonstrates, the types of uncertainty are logically inherent to direct inference based on any kind of knowledge about general causation. Factfinders may have to make decisions about acceptable uncertainty on a case-by-case basis, applying common sense and a rough sense of justice, unless courts can establish principled rules for particular categories of cases. The “lost-chance” cases, discussed in Part III.B.2, provide examples of policy-based rules that respect the uncertainties inherent in finding specific causation. Once the extent of inherent uncertainty is clear (as well as the non-scientific nature of decisions about that uncertainty), then the need for non-epistemic decisions and policy-based rules also becomes clear.<sup>264</sup>

Finally, the non-epistemic character of any bright-line, quantitative threshold is highlighted by its arbitrary nature. On the one hand, a greater-than-50% rule seems to set the quantitative threshold too high. In a particular case, there may be a minimally acceptable but incomplete causal model with a calculated  $RR < 2.0$ , but there may also be other known risk factors present in the plaintiff’s case that probably increase the risk by some unquantified amount. In such a case, a reasonable factfinder could find that the relative risk in the reference group probably sets a floor for the plaintiff’s individual risk, but that the plaintiff’s individual risk is probably higher. One might argue that a bright-line test of 50% or  $RR > 2.0$  for the *quantified portion* of the evidence is too high, when considered in combination with such unquantified evidence of additional risk factors. On the other hand,

---

and that the court would not “acknowledge a statistically significant association beyond the 95% level to 90% or lower values”). The pragmatic nature of even this convention, however, has been widely acknowledged. *See supra* note 72.

263. As discussed in Part III.B.3, some courts have mistakenly thought that specific causation poses special problems for epidemiologic evidence. *Cf. Green et al., supra* note 6, at 381 (stating that “[e]pidemiology is concerned with the incidence of disease in populations and does not address the question of the cause of an individual’s disease,” a question that is “beyond the domain of the science of epidemiology”); Freedman & Stark, *supra* note 226, at 61-62 (concluding that when epidemiologic data provides the evidence, “[t]he scientific connection between specific causation and a relative risk of 2.0 is doubtful”).

264. An example of such a policy-based decision rule is presented by *Stevens v. Sec’y of HHS*, No. 99-594V, 2001 WL 387418 (Fed. Cl. Mar. 30, 2001), involving a claim under the National Vaccine Injury Compensation Program. Chief Special Master Golkiewicz proposed and applied a rule requiring only “[r]easonable efforts” by the petitioner “to rule out known alternate causes,” and rejected the argument that “a petitioner must eliminate potential *unknown, unidentified, speculative, unapparent, or spontaneous causes*,” because the latter rule “would necessarily prevent any petitioner from prevailing” in a vaccine case. *Id.* at \*26.

Commentators have also proposed policy-based rules concerning specific causation. *See, e.g., Berger, supra* note 254, at 2117-20 (proposing to abolish the causation requirement in toxic tort cases in pursuit of various policy objectives); Gerald W. Boston, *A Mass-Exposure Model of Toxic Causation: The Content of Scientific Proof and the Regulatory Experience*, 18 COLUM. J. ENVTL. L. 181, 187-91, 363-82 (1993) (arguing for different rules on the sufficiency of evidence of causation in mass-exposure tort cases and in isolated-exposure cases, on the grounds of achieving consistency of outcomes across cases, achieving optimal levels of deterrence, avoiding unlimited liability, harmonizing public health regulation and tort law, and promoting corrective justice); Finley, *supra* note 257, at 366 (arguing against admissibility rules requiring epidemiologic evidence of relative risk greater than 2.0, and arguing that by making “individualistic causation judgments” in products liability cases courts are “making policy judgments about which party should bear the responsibility for causal uncertainty, and which party is in the best position to learn more about and absorb or spread the costs of the risks”).



given the substantial uncertainty inherent in applying any relative risk from incomplete causal evidence, one might argue that a greater-than-50% rule sets the quantitative threshold too low. The argument would be that specific causation requires a finding that the *real RR* is probably greater than 2.0, after taking into account measurement uncertainty, sampling uncertainty, modeling uncertainty, causal uncertainty, and uncertainty about plaintiff-representativeness. If all uncertainty is chargeable to the plaintiff, then an appropriate "cushion" above 50% should be built into any quantitative rule. Therefore, taking the two branches of argument together, any rule establishing a bright-line, statistical threshold for all cases is arbitrarily too high or too low, if its rationale is the logically flawed reasoning about direct inference to specific causation.

The appropriate response to these various problems is to abandon altogether both the flawed reasoning and its quantitative progeny. In the face of significant uncertainty, decisions to find specific causation cannot be epistemically warranted and must be justified on non-epistemic grounds. The next section extends this general criticism further, by addressing a variety of quantitative rules that courts have applied in particular cases.

### *B. Judicial Errors in Reasoning About Specific Causation*

The previous section of the Article used the analysis in Parts I and II to outline a factfinding approach to warranting conclusions about specific causation. This section examines the practical implications of this analysis by critiquing judicial errors in various kinds of tort cases. Those cases also illustrate how the presence of uncertainty undermines the notion that specific causation is a factual or scientific issue. Recognizing the extent of that uncertainty should lead courts to consider more appropriate policy justifications and to adopt better decision rules.

#### *1. Judges as Factfinders and the "0.5 Inference Rule"*

Under the traditional standard of proof in tort litigation, a finding of specific causation must be warranted by "a preponderance of the evidence."<sup>265</sup> Courts have interpreted this phrase as meaning "more likely than not," "probably true," or "more probably true than false."<sup>266</sup> The "weight" or "convincing force" or "probative value" of the evidence supporting the finding must be "greater than" the weight of evidence against the finding.<sup>267</sup> Such statements about the standard of proof are uncontroversial. Many courts and theorists, however, re-formulate the preponderance standard as a

---

265. See, e.g., 2 MCCORMICK ON EVIDENCE § 339 (John W. Strong ed., 5th ed. 1999).

266. See, e.g., *id.*; FINKELSTEIN, *supra* note 198, at 65-66; David Kaye, *Naked Statistical Evidence*, 89 YALE L.J. 601, 603 (1980) (book review); J.P. McBaine, *Burden of Proof: Degrees of Belief*, 32 CAL. L. REV. 242, 247, 260-61 (1944); *Preponderance, Probability*, *supra* note 258, at 1076-78.

267. See, e.g., FLEMING JAMES, JR. ET AL., CIVIL PROCEDURE § 7.14 (4th ed. 1992); McBaine, *supra* note 266, at 247.

quantitative rule: the finding must have a probability of being true that is greater than 1/2, 0.5, or 50%.<sup>268</sup> This re-formulation can be called the "0.5 inference rule" for factfinding.<sup>269</sup> As applied to findings about specific causation, this quantitative rule requires a finding of specific causation if, but only if, the probability of specific causation is greater than 0.5.<sup>270</sup> I have argued elsewhere that such a quantitative interpretation of the standard of proof is unnecessary, misleading, and unwise, when we evaluate the rule using the traditional policy rationales for the preponderance standard.<sup>271</sup> The present Article, however, attacks the error of applying the 0.5 inference rule to findings of specific causation. Some courts, discussed below, have compounded the adoption of the 0.5 inference rule with a misunderstanding about the logic of specific causation, with unfortunate results for tort plaintiffs.

A judgment in the High Court of Justice, Queens Bench Division, illustrates judicial misunderstandings about the structure of warranted factfinding about specific causation.<sup>272</sup> The judge acted as factfinder in that case,

268. See, e.g., *United States v. Shonubi*, 895 F. Supp. 460, 471 (E.D.N.Y. 1995), *vacated by* 103 F.3d 1085 (2d Cir. 1997); *United States v. Fatico*, 458 F. Supp. 388, 403 (E.D.N.Y. 1978), *aff'd*, 603 F.2d 1053 (2d Cir. 1979); *United States v. Schipani*, 289 F. Supp. 43, 55-56 (E.D.N.Y. 1968), *aff'd*, 414 F.2d 1262 (2d Cir. 1969); *Cooper v. Hartman*, 533 A.2d 1294, 1299-1300 (Md. 1987) (holding that the plaintiff must prove the patient "had a better than 50% chance of full recovery absent the malpractice" and stating that "probability" means "greater than 50% chance" and "possibility" means "less than 50% chance"); *Cooper v. Sisters of Charity, Inc.*, 272 N.E.2d 97, 103-04 (Ohio 1971) (stating that "probable" in connection with standard of proof "is more than 50% of actual"). For theoretical discussions by commentators, see PORAT & STEIN, *supra* note 13, at 18; Steve Gold, *Causation in Toxic Torts: Burdens of Proof, Standards of Persuasion, and Statistical Evidence*, 96 YALE L.J. 376, 378, 384-86 (1986); David Kaye, *The Limits of the Preponderance of the Evidence Standard: Justifiably Naked Statistical Evidence and Multiple Causation*, 1982 AM. B. FOUND. RES. J. 487, 493; Talbot Page, *On the Meaning of the Preponderance Test in Judicial Regulation of Chemical Hazard*, 46 LAW & CONTEMP. PROBS. 267, 269-71 (1983); David Rosenberg, *The Causal Connection in Mass Exposure Cases: A "Public Law" Vision of the Tort System*, 97 HARV. L. REV. 851, 857 (1984); and Ralph K. Winter, Jr., *The Jury and the Risk of Nonpersuasion*, 1971 LAW & SOC'Y REV. 335, 336-39.

269. See, e.g., PORAT & STEIN, *supra* note 13, at 18 (discussing the preponderance of the evidence standard in tort cases, and combining a probabilistic factfinding rule with a damage award rule of full recovery); Gold, *supra* note 268, at 386, 395.

270. See, e.g., *Hartman*, 533 A.2d at 1299-1300; *Sisters of Charity, Inc.*, 272 N.E.2d at 103-04. See also Ronald J. Allen, *A Reconceptualization of Civil Trials*, 66 B.U. L. REV. 401, 405 (1986); Cohen, *supra* note 72, at 394; Kaye, *supra* note 268, at 493; Koehler & Shaviro, *supra* note 200, at 249-52; *Modeling Relevance*, *supra* note 198, at 1033-34; *Evidence Scholarship*, *supra* note 198, at 451-52, 454. For a contrary view, see Charles Nesson, *The Evidence or the Event? On Judicial Proof and the Acceptability of Verdicts*, 98 HARV. L. REV. 1357, 1359 (1985).

271. *Preponderance, Probability*, *supra* note 258, at 1097-1120. The preponderance standard itself, without the quantitative interpretation, serves the traditional objectives of treating parties in a fair and unbiased fashion and of creating an incentive on all parties to produce adequate evidence if they can do so. *Id.* at 1121. I have argued that what it means to say that a finding is "probably true" and "warranted by a preponderance of the legally available evidence" is that it satisfies the formal requirements of consistency and coherence, that it has sufficient support in the legally available evidence (in the sense of being reasonably inferable from that evidence), and that the factfinder anticipates a fairly wide scope of agreement on the findings, at least by reasonable people who weigh the same evidence. *Id.* at 1120-21. The analysis of direct inference to specific causation in the present Article constitutes a step toward developing useful warrant rules in a particular area of tort law.

272. *XYZ v. Schering Health Care Ltd.*, [2002] E.W.H.C. 1420, 2002 WL 1446183 (Q.B. July 29, 2002).

and the judgment contains an account of the factfinder's reasoning.<sup>273</sup> The case involved seven lead claims (out of a total of ninety-nine claims) brought against three pharmaceutical manufacturers and alleging products liability for oral contraceptives.<sup>274</sup> The claimants alleged injuries related to venous-thromboembolism (VTE), which includes such disorders as deep vein thrombosis, pulmonary embolism, and cerebral venous thrombosis.<sup>275</sup> The allegedly defective products were "third generation combined oral contraceptives," which contained a synthetic oestrogen in a dose equal to or less than thirty  $\mu\text{g}$ , combined with a "third generation" synthetic progestogen.<sup>276</sup> Claimants sought to prove that the third generation products put them at greater risk of VTE than a second generation product would have.<sup>277</sup> A "second generation combined oral contraceptive" was defined as a product with a synthetic oestrogen dose of no more than fifty  $\mu\text{g}$ , combined with one of the progestogens used in the "second generation" of the product.<sup>278</sup> The claimants therefore needed to prove, in part, specific causation: that *the* third generation products that they ingested caused *their* particular VTE injuries.<sup>279</sup>

Despite the individual nature of the claimants' cases, the judge and (apparently) all of the parties considered the determining issue to be whether the available epidemiologic studies established that third generation products at least doubled the true risk of VTE, compared to the risk posed by second generation products.<sup>280</sup> Before other issues were tried, the judge conducted a forty-two-day trial devoted to this issue alone,<sup>281</sup> and the trial

---

273. See *id.* ¶ 19.

274. *Id.* ¶ 1. The claimants alleged that the defendants' products were defective under the provisions of the Consumer Protection Act 1987 and Product Liability Directive 85/374/EEC. *Id.* ¶ 2. The judgment occurred in the context of highly publicized concern in the United Kingdom over the safety of oral contraceptives. *Id.* ¶ 3. See also *infra* note 278.

275. *Schering Health Care*, 2002 WL 1446183, ¶¶ 1, 4.

276. See *id.* ¶¶ 4-8.

277. See *id.* ¶¶ 20, 339, 343. Although the "first generation" of combined oral contraceptives contained high doses of synthetic oestrogen (150  $\mu\text{g}$  or more), concern over a possible increased risk of VTE led to development of a "second generation" product with lower doses of oestrogen (on the order of 50  $\mu\text{g}$  or less), combined with a variety of progestogens. *Id.* ¶ 10.

278. *Id.* ¶ 10. Products with lower doses of oestrogen had been introduced to lower any risk of certain cardio-vascular diseases, including not only VTE, but also "acute myocardial infarction [and] cerebral infarction." See *id.* ¶ 52. There was surprise, therefore, and an ensuing "pill scare" among the general public, when the UK Committee on the Safety of Medicine "circulated a warning" in October 1995 that recent studies indicated that third generation products were associated with "around a two-fold increase in the risk" of VTE, compared with second generation products. See *id.* ¶¶ 3, 11.

279. As the judge stated:

I have to decide on the balance of probabilities on the evidence presented before me whether each of these Claimants has established her case that these products were defective, that the defective nature of them caused or contributed to her injury, and therefore that she is entitled to damages as a result.

*Schering Health Care*, 2002 WL 1446183, ¶ 19.

280. *Id.* ¶¶ 20-21 (stating that "all agree that if the Claimants fail to prove this the action should go no further as it could not succeed"). While arguing that the factfinding logic in that case was flawed, this Article takes no view on whether the Court had a responsibility to use correct reasoning even if the claimants' attorneys did not.

281. *Id.* ¶ 22.

ended with a judgment for defendants that disposed of the cases.<sup>282</sup> In arriving at this judgment, the judge found that there probably was no increased risk at all,<sup>283</sup> because he was persuaded by a controversial application of a regression model to the data in a particular study.<sup>284</sup> In the alternative, however, without relying on that regression model, he found that there was probably a (general) causal connection, but that the increased risk was most likely around 1.7.<sup>285</sup> The court held that a finding of increased risk of about

282. *Id.* ¶¶ 339-45.

283. The analysis in this Article also lays a foundation for further research into the causal role of gender bias—bias in deciding, for example, whether or when a causal model about risk takes enough causally relevant variables into account. If, as this Article establishes, decisions about the acceptability of uncertainty are not scientific in nature, then such decisions might exhibit biases along demographic lines, such as gender. Cf. Vern R. Walker, *Consistent Levels of Protection in International Trade Disputes: Using Risk Perception Research to Justify Different Levels of Acceptable Risk*, 31 ENVTL. L. REP. 11,317, 11,319-24 (2001) (summarizing scientific literature on demographic biases in risk perception, including gender bias in estimating degrees of risk).

284. *Schering Health Care*, 2002 WL 1446183, ¶¶ 121-63, 339. The controversy involved whether a “Cox regression analysis with time-dependent covariates” should be used in a case-control study, and whether the researchers applied it correctly. *Id.* ¶¶ 124-26. The hypothesis behind using the model was that the higher risk for the third generation products might be explained by a combination of (a) a “healthy user” effect or “depletion of susceptible[s]” among second generation users, whereby long-term users who tolerate the products well tend to continue to use them, while those who do not tolerate them well tend to switch to third generation products, and (b) a loading of less healthy or more problematic users into the third generation group, by physicians who prescribe the “safer” drug to the patients that appear to be at higher risk. Michael A. Lewis et al., *The Increased Risk of Venous Thromboembolism and the Use of Third Generation Progestagens: Role of Bias in Observational Research*, 54 CONTRACEPTION 5, 9 (1996) [hereinafter “TNS 2”], available at <http://www.sciencedirect.com>; Michael A. Lewis et al., *The Differential Risk of Oral Contraceptives: The Impact of Full Exposure History*, 14 HUM. REPROD. 1493, 1493 (1999) [hereinafter “TNS 3”], available at <http://humrep.oupjournals.org>; *Schering Health Care*, 2002 WL 1446183, ¶¶ 106-07, 122. The Cox model with time-dependent covariates uses the product-exposure history of users to adjust a “hazard ratio” that measures statistical association. TNS 3, at 1494; *Schering Health Care*, 2002 WL 1446183, ¶¶ 124-25.

For each case and control, the Transnational Study (TNS) recorded data (by month) on which product the participant had used, although the validity of such self-reported data was controversial. *Id.* ¶¶ 127-31. The Cox model contained a variable for prior product exposure (a “time-dependent covariate”) that could take on various values over time as the participant started, stopped, or switched products. *Id.* ¶¶ 144-45. The Cox model in effect compared different exposure groups over time and generated their respective hazard rates. *Id.* ¶ 145. The court found

that Cox was an appropriate model to apply to this dataset, was applied to it correctly, that it did . . . make effective adjustment for the effect of lifetime duration of [combined oral contraception] use, and therefore as a matter of probability there is no true relative risk of VTE attaching to [third generation products] as against [second generation products].

*Id.* ¶ 162. This finding of no increased risk would mean that the claimants’ cases failed. *Id.* ¶ 163. However, in the alternative and in case of appellate reversal on this use of Cox, the judge concluded that “I think it sensible to go on to reach conclusions on the other points outstanding, despite my finding on Cox.” *Id.*

285. *Id.* ¶¶ 341-44. The judge wrote:

As the above findings [about increased risk statistics] both dispose of the first issue in a way which means that the claim must fail, it is not strictly necessary for me to make a finding as to whether the RR of 1.7 itself translates into a relationship of true cause and effect or is a merely statistical appearance. If I had to do so I would incline to a finding that there is an underlying causal connection at about that level of increased risk.

*Id.* ¶ 344.

The judge carefully noted the difference between relative risk and an odds ratio, and decided that for his purposes he would use them “interchangeably albeit inaccurately.” *Id.* ¶ 26-27. As discussed in Part I.C, this can be a reasonable approach when assessing the strength of association for purposes of deciding general causation.

1.7 disposed of the claimants' cases because, with a relative risk falling short of 2.0, no claimant could prove that the third generation product (even if defective) *probably* caused *her* VTE.<sup>286</sup> A product posing such an increased risk might be defective, but each individual case would fail on the proof of specific causation.

This Article does not second-guess the judge's findings of fact, nor does it re-evaluate the evidence linking oral contraceptives to VTE. However, using the analysis in this Article would have produced a very different fact-finding pattern in that case. First, with regard to a major premise in a direct inference of causation, the evidence supported a finding of general causation between use of a third generation product and VTE. In so finding, the judge evaluated the evidence for uncertainties due to measurement, sampling, and associational modeling.<sup>287</sup> He further considered the weight of evidence for finding causation and not merely association.<sup>288</sup> The portions of the judgment addressing those uncertainties illustrate the various aspects of the analysis in Part I. When controversies arose over the acceptability of various kinds of uncertainty, he made decisions about acceptability and gave his reasons.<sup>289</sup> He at least implicitly considered the levels of all of these uncertainties to be acceptable for purposes of the case when he decided that, on balance, the evidence favored a finding of general causation.

The judge also concluded that the epidemiologic evidence supported finding "a *RR* of about 1.7."<sup>290</sup> Thus, if there is a group of women similar to

---

286. *Id.* ¶¶ 21, 345.

287. Concerning measurement uncertainty, see *id.* ¶¶ 128-31 (discussing validity of interview data on personal histories of oral contraception use); *id.* ¶¶ 194-97, 219-24 (discussing whether, for purposes of matching controls to cases, the age of the study participant should be categorized by "year of birth" or by "5 year bands"); and *id.* ¶¶ 279-86 (discussing the diagnostic uncertainty for VTE, and the possible referral-bias and diagnostic-bias in the case group because women perceived to be at risk might have a greater chance of both referral and diagnosis).

Concerning sampling uncertainty, see *id.* ¶¶ 64-67 (discussing which data from which study centers to include in the WHO study); and *id.* ¶¶ 258-61, 279-86 (discussing referral-bias and diagnostic-bias as possible causes of lack of sample representativeness). The court's judgment in effect combined measurement and sampling uncertainty into a single concern over "bias" as "anything which will distort the study by making the sampling process on which it is based unrepresentative or skewed in favour of or against a particular side of the equation." *Id.* ¶ 258.

Concerning modeling uncertainty, see *id.* ¶¶ 194-97, 219-24 (discussing how best to control for the age of the study participant so as to minimize possible confounding effects). See also *supra* note 284 (discussing the modeling uncertainty in applying a Cox regression model).

288. Concerning causal uncertainty, see *id.* ¶¶ 24, 302-08, 344. Cf. *id.* ¶¶ 262-78 (discussing prescriber-bias as a possible causal confounder—that is, the possibility that doctors would "preferentially prescribe" third generation products to patients who are already "at an elevated risk of VTE").

289. The judgment discusses the acceptability of various kinds of uncertainty inherent in the major premise of general causation. See, e.g., *id.* ¶¶ 64-67 (discounting the "much higher" risk estimates from non-Oxford centers in the WHO study because of "the sparsity of the non-Oxford data" and their resulting "very wide [confidence intervals]"); *id.* ¶¶ 128-31 (finding "a body of good quality evidence as to total contraceptive history" for participants in the TNS study); *id.* ¶¶ 209-24 (finding that "[t]he right value to give [to the odds ratio in] the GPRD studies collectively is one falling in the area between 1.5 and 1.8," despite the "impossibility" of saying "as a matter of probability that one of these studies should be accepted in total as being right and the other rejected in total as being wrong"); and *id.* ¶ 297 (declining to reduce previous estimates of relative risk, despite the possibility of bias). See also *supra* note 284 (discussing the modeling uncertainty in applying a Cox regression model).

290. *Schering Health Care*, 2002 WL 1446183, ¶¶ 93, 341, 343; see also *supra* note 286 and accompanying text.

those in the study and who are taking the third generation combined oral contraceptives, and if in that group about 170 women are expected to develop VTE, then about 100 of those 170 women would have developed VTE even using the second generation product, while the other seventy cases are expected to develop VTE as a result of taking the third generation product.<sup>291</sup> In the standard formulation of direct inference used in this Article, the women who develop VTE as a result of taking a third generation combined oral contraceptive comprise subgroup *B*.

Given a determination of general causation for the product, however, the court should have turned its attention to particular plaintiffs instead of ending all of their cases based on the estimate of general increased risk of 1.7. The next task was to derive a reference-class profile for the particular type of VTE injury of each individual plaintiff, using all factors known or suspected to be causally relevant in each case. This the court did not do. A short opening paragraph of the judgment summarizes minimal information about the exposure and symptoms of the claimants,<sup>292</sup> but the remainder of the long and detailed judgment makes no use of this information, nor does it at any time discuss the individual evidence of specific causation for these claimants.<sup>293</sup> Yet the seven claimants, who suffered different types of injuries, may have differed from each other on variables causally relevant to their injuries.<sup>294</sup> The judgment does not discuss the medical histories of these individual claimants, their general health conditions, or information on other variables thought to be causally relevant—such as obesity, varicose veins, or their histories of thrombosis, rheumatic heart disease, or hypertension in pregnancy.<sup>295</sup> Such plaintiff-specific information clearly played no role in the judge's reasoning.

---

panying text.

291. One estimate of the baseline incidence of VTE in healthy non-pregnant women taking second generation combined oral contraceptives is about fifteen cases per 100,000 woman-years. *Schering Health Care*, 2002 WL 1446183, ¶ 17 (summarizing new warnings required to be on the Summary of Product Characteristics for third generation combined oral contraceptives); cf. World Health Organization Collaborative Study of Cardiovascular Disease and Steroid Hormone Contraception, *Effect of Different Progestagens in Low Oestrogen Oral Contraceptives on Venous Thromboembolic Disease*, 346 LANCET 1582, 1587 (1995) [hereinafter "WHO Study II"] (providing another estimate of the incidence of idiopathic VTE), available at <http://www.sciencedirect.com>. Therefore, the group of healthy women that would yield about seventy cases of third generation-caused-VTE would have about 667,000 woman-years of exposure. Such a group could expect about 100 baseline cases of VTE and about seventy excess cases that are third generation-caused.

292. *Schering Health Care*, 2002 WL 1446183, ¶ 4 (summarily describing claimants).

293. The cases of the claimants failed before the judge heard any evidence relating to individual claimants. *Id.* ¶ 22; Mark Mildred, *Case Comment*, 4 J. PERS. INJ. L. 4, 428-30 (2002).

294. Three suffered deep vein thrombosis, reporting such symptoms as "leg pain and loss of mobility"; two suffered pulmonary embolisms, with symptoms such as chest pain; one suffered cerebral venous thrombosis, reporting "very severe headaches," "nausea and giddiness"; and one suffered "a stroke as a result of a paradoxical embolism." *Schering Health Care*, 2002 WL 1446183, ¶ 4. With respect to this last claimant, the judge wrote: "I do not as yet have full details of her current symptoms." *Id.*

295. On causally relevant factors, see *id.* ¶ 11 (quoting the Committee on Safety of Medicines as listing risk factors for VTE); *id.* ¶ 17 (discussing new warnings for third generation combined oral contraceptives); World Health Organization Collaborative Study of Cardiovascular Disease and Steroid Hormone Contraception, *Venous Thromboembolic Disease and Combined Oral Contraceptives: Results of International Multicentre Case-Control Study*, 346 LANCET 1575, 1579-80 (1995) [hereinafter "WHO

Once general causation is found for the product, however, each claimant deserves to have her own case decided on its own merits. A reasonable factfinder should determine, for each plaintiff taken individually, a profile of causally relevant characteristics that identifies a reference group for that plaintiff and her relevant injury—that is, a group of people who are similar to the particular plaintiff in her type of injury and in all known or suspected causally relevant factors for that type of injury. Then the factfinder should use each plaintiff's reference-group profile to evaluate the available evidence in the record for any warranted generalizations for women satisfying this profile. If the totality of the epidemiologic evidence contains data gathered on enough of those variables, then there may be acceptable relative risk estimates for such a reference group.<sup>296</sup> For example, there may be relative risk estimates for women with a history of hypertension in pregnancy, who are obese, and who suffer deep vein thrombosis while using a particular third generation product.<sup>297</sup> Moreover, the toxicological evidence or clinical studies might contain evidence of bias for additional variables, which medical experts could use to adjust these relative risk estimates either up or down for application to each specific claimant.<sup>298</sup> If sufficient data are available, a reasonable factfinder might estimate, with acceptable uncertainty, the relative risk for women similar to the specific plaintiff—in this litigation, the relative risk of the injury for women-like-the-plaintiff who use the third

---

Study I"], available at <http://www.sciencedirect.com>; and Walter O. Spitzer et al., *Third Generation Oral Contraceptives and Risk of Venous Thromboembolic Disorders: An International Case-Control Study*, 312 BRIT. MED. J. 83, 84, tbl.1 (1996) [hereinafter "TNS I"], available at <http://www.bmj.bmjournals.com>.

"[M]ajor risk factors for VTE [include] trauma, surgery, immobilisation, and pregnancy . . . ." WHO Study II, *supra* note 291, at 1587. The WHO case-control study was designed to investigate only "idiopathic" VTE, and women with these major risk factors were excluded from the study. See WHO Study I, *supra* note 295, at 1576-77 (listing exclusion criteria for cases and controls); WHO Study II, *supra* note 291, at 1587 (describing the study design); *Schering Health Care*, 2002 WL 1446183, ¶ 292 (discussing the potential bias introduced by studying cases of idiopathic VTE in hospital patients, if the "target population" is "healthy women in the community"). While these risk factors are relevant for each specific claimant, it is possible that the lead claimants were chosen in part because these major causes were absent in their cases. A thorough approach to developing a reference-group profile would list all known or suspected risk factors for the claimant's injury and would classify the claimant on every one of them.

296. Unfortunately, when plaintiffs attempt to refine the reference group and re-analyze the available data, they often encounter judicial skepticism, especially if the raw data and re-analysis are unpublished. *E.g.*, *Merrell Dow Pharm., Inc. v. Havner*, 953 S.W.2d 706, 725-27 (Tex. 1997). Such refinement, however, is precisely what plaintiffs and factfinders should do, and it should come as no surprise that refinement almost certainly leads to increased statistical uncertainties about the plaintiff's reference group. The problem of deciding what to do about those uncertainties has no scientific or epistemic solution.

297. There is a good likelihood, of course, that the available studies would have insufficient statistical power to provide a statistically significant relative risk for such a refined reference group. See *supra* Part I.C (discussing statistical power and sampling uncertainty). A major thesis of this Article is that there is every reason to think that in toxic tort cases, a reference group thought to be adequately representative will turn out to be inadequately studied. Given this central fact about specific causation, there should be policy grounds for deciding when claimants recover damages and when they do not.

298. The possible relevance of different types of evidence (epidemiologic, toxicological, and even case reports) to a plaintiff-specific reference group should make judges cautious about adopting rules that exclude studies from evidence one-by-one, in isolation from other evidence in the case. See *infra* Part III.B.3.

generation oral contraceptive that she did, compared to the risk for women-like-the-plaintiff who use second generation oral contraceptives. The court, however, did not evaluate the evidence for such claimant-specific reference groups, but held against all the claimants on the single finding that the general increased risk was less than 2.0. This finding was not about any individual claimant but, in reality, about a "statistical woman."

The magnitude of the increased risk for the "average woman" in an available study, however, would seem an arbitrary rule of decision in such a case were it not for the logical fallacy behind it. In a case such as *Schering Health Care*, where it is demonstrable that no available causal model even approaches the completeness needed to explain or predict individual cases,<sup>299</sup> assigning *any* probability of *specific* causation on the basis of such evidence cannot be epistemically warranted. Justice MacKay's primary logical error was thinking that *if* the general studies had in fact shown a relative risk greater than 2.0, *then* this would be sufficient evidence to warrant finding specific causation.<sup>300</sup> His error was thinking that finding specific causation on the *Schering Health Care* evidence could *ever* be a factual issue.<sup>301</sup> But, as a logical matter, even if the evidence had shown a general increased risk of 3.0, a finding of specific causation in a particular claimant's case would have been just as speculative as it would have been with a relative risk of 1.7. The critical question is not the magnitude of the general relative risk in some study group compared to the baseline rate in that group, but the acceptability of the uncertainties inherent in estimates of relative risk for claimant-specific reference groups that are defined by the causally relevant factors in each specific case.

In tort cases like *Schering Health Care*, specific causation cannot be a factual or scientific issue. Given the incompleteness of the best available causal model for the relevant injuries, whether any particular plaintiff ever prevails must be a matter of common sense, fairness, and policy. The important policy decisions are who should bear the cost of uncertainty about causation, and whether factfinders should make such decisions on a case-by-

299. XYZ v. Schering Health Care Ltd., [2002] E.W.H.C. 1420 (QB), 2002 WL 1446183, ¶ 31 (July 29, 2002) (stating that "[t]he condition under consideration in this case is on the face of it classically suited for an epidemiological investigation," because "haematology is very far from reaching a full understanding of what causes blood clots to form in the venous system").

300. This was clearly a *logical* error on the judge's part, because he gave the following as his reason why the claimants *must* fail unless they prove a general relative risk for VTE that is greater than 2.0:

If factor X increases the risk of condition Y by more than 2 when compared with factor Z it can then be said, of a *group* of say 100 with both exposure to factor X and the condition, that as a *matter of probability* more than 50 would not have suffered Y without being exposed to X. If medical science cannot identify the members of the group who would and who would not have suffered Y, it can nevertheless be said of *each member* that *she was more likely than not* to have avoided Y had *she* not been exposed to X.

*Id.* ¶ 21 (emphasis added). The emphasized words are those with logical significance for direct inference. The analysis in Parts I and II explains why this reasoning is fallacious.

301. This logical error also cost the litigants and the court a substantial amount of time and money spent litigating a precise point-estimate of relative risk. See, e.g., *id.* ¶ 32 (listing the subtle differences in relative risk values espoused by the claimants' and defendants' expert witnesses). If this inquiry is doomed to an inconclusive end, then such expenditures are very inefficient.



case basis or whether courts should adopt uniform decision rules for entire categories of cases. The fallacy of thinking that a general relative risk greater than 2.0 would obviate the problem of plaintiff-representativeness has led courts to the two further mistakes of thinking that relative risk greater than 2.0 is *sufficient* evidence for a finding of specific causation and that such a relative risk is *necessary* for such a finding. Logically speaking, none of these three propositions is true. In the *Schering Health Care* case, these fallacies led the judge to decide the fates of ninety-nine claims (and in effect uncounted others) on a single finding about a statistical woman, whereas a clearer understanding of the logic of specific causation might have led the court to adjudicate individual cases on a broader policy basis. These same fallacies have also led courts to adopt unjustified decision rules for ruling on motions about the sufficiency of the evidence and about the admissibility of particular items of evidence. The next two sections illustrate these two types of error.

## 2. *Judges as Referees of Reasonable Inferences and Rules on Sufficiency of Evidence*

Once judges believe that reasonable factfinders operating under the preponderance standard of proof will make findings about specific causation in accordance with the 0.5 inference rule, then it is a short step for those judges to adopt corresponding rules for evaluating the legal sufficiency of the supporting evidence.<sup>302</sup> Judicial misunderstandings about the logical warrant for specific causation compound the confusion. Judges who think that specific causation is always a factual issue, on which scientific experts have the dominant voice, are prone to drawing arbitrary lines in the shifting statistical sands and to dismissing cases for the wrong reasons. The so-called "lost-chance" cases in the area of medical malpractice provide one illustration of these judicial errors.

The thinking behind the 0.5 inference rule has played a decisive role in confusing the courts over the lost-chance cases.<sup>303</sup> In a typical lost-chance case, at the time the defendant's negligent conduct occurred, the plaintiff already had a pre-existing illness with a higher-than-50% baseline risk of death or further bodily injury.<sup>304</sup> The defendant's negligence then caused the

---

302. See, e.g., Green, *supra* note 72, at 691 (concluding that "any relative risk less than two would be inadequate to support a plaintiff's verdict," at least "in the absence of other evidence enabling a more refined assessment with regard to the plaintiff").

303. For discussion of these cases, see Walker, *supra* note 18, at 248-56, 297-307.

304. E.g., *Falcon v. Mem'l Hosp.*, 462 N.W.2d 44 (Mich. 1990) (62.5% chance of dying); *Kallenberg v. Beth Israel Hosp.*, 357 N.Y.S.2d 508 (App. Div. 1974), *aff'd*, 337 N.E.2d 128 (1975) (60-80% chance of dying); *Herskovits v. Group Health Coop.*, 664 P.2d 474 (Wash. 1983) (61% chance of dying).

When lost-chance situations involve baseline risks less than 50% (that is, success rates over 50%), courts have routinely sent the cases to the jury. E.g., *Rewis v. United States*, 503 F.2d 1202 (5th Cir. 1974) (survival more likely than not if proper diagnosis and treatment); *Glicklich v. Spievack*, 452 N.E.2d 287, 291 (Mass. App. 1983) (94% chance of surviving 10 years with proper diagnosis, reduced to "a 50% or less chance of ten year survival"); *Hamil v. Bashline*, 392 A.2d 1280 (Pa. 1978) (75% chance

plaintiff's risk to increase. However, when the plaintiff subsequently died or suffered further injury, the high baseline risk prevents experts from testifying that the defendant's negligence, and not just the pre-existing condition, was a cause of the death or further injury.<sup>305</sup> Most judges considering these cases implicitly adopt the 0.5 inference rule and conclude that a warranted inference depends entirely on whether the proportion of negligence-caused injuries, out of the total number of injuries, is greater than 0.5.<sup>306</sup> They think that if the excess risk due to the defendant's negligence exceeds—ever so slightly—the baseline risk, then the plaintiff is entitled to recover.<sup>307</sup> Moreover, they do not understand how a plaintiff could prove specific causation as more likely than not if that proportion is less than or equal to 0.5 (50%).

Without an appreciation of the non-scientific nature of any decisions for dealing with the uncertainties that are necessarily inherent in the lost-chance cases, and misled by the mistaken logic behind the 0.5 inference rule, courts that acknowledge the fairness issue in those cases have been reluctant to decide them on an appropriate policy foundation. Few courts have even considered re-thinking the 0.5 inference rule itself.<sup>308</sup> Some courts, in order to avoid the seemingly inexorable and harsh result of that rule, have replaced the traditional “but-for” concept of causation with a vague “substantial factor” concept.<sup>309</sup> Many courts have adopted a “loss of chance” as a new kind of compensable injury, despite their misgivings about the implications of doing so for “mere-risk” cases where a physical injury has not occurred.<sup>310</sup> Tragically, still other courts have refused to assist lost-chance

of survival if properly treated).

305. See, e.g., *DeBurkate v. Louvar*, 393 N.W.2d 131, 136-37 (Iowa 1986) (citing nine prior cases); *Falcon*, 462 N.W.2d 44, 56; *Kallenberg*, 357 N.Y.S.2d 508, 511; *Herskovits*, 664 P.2d 474, 477.

306. For a noteworthy awareness of the fallacy in this, see *Falcon*, 462 N.W.2d at 47 (acknowledging that “[t]o say that a patient would have had a ninety-nine percent opportunity of survival if given proper treatment, does not mean that the physician's negligence was the cause in fact if the patient would have been among the unfortunate one percent who would have died”).

307. *Dumas v. Cooney*, 1 Cal. Rptr. 2d 584, 589 (Dist. Ct. App. 1991) (stating that where testimony establishes a better-than-even chance of survival absent negligence, “a finding for the plaintiff is consistent with existing principles of proximate cause”); *Cooper v. Hartman*, 533 A.2d 1294, 1299 (Md. 1987) (holding that, under “traditional rule” governing standard of proof, the plaintiff has a burden of proving that the patient “had a better than 50% chance of full recovery absent the malpractice”).

308. For an example of a court probing the correct logic, see *Rewis*, 503 F.2d at 1205-11 (holding in a case involving misdiagnosis of aspirin poisoning in a child that it was essential that the factfinder examine assumptions concerning aspirin ingestion rate, absorption rate into the bloodstream, and elimination rate from blood in the particular patient; if the patient did not fit the assumed characteristics on these factors, then “there is at least an equal chance that she would have fallen below the line indicated [the line above which fatalities are likely to occur], in which event she would have shown up on the chart as one patient who survived”).

309. See e.g., *Evers v. Dollinger*, 471 A.2d 405, 413-15 (N.J. 1984) (adopting “substantial factor” causation); *Herskovits*, 664 P.2d at 477-78 (same). Cf. *Werner v. Blankfort*, 42 Cal. Rptr. 2d 229, 232-39 (Dist. Ct. App. 1995) (discussing the range of positions on causation taken by the courts); *DeBurkate*, 393 N.W.2d at 137 (stating that by viewing the bodily injury as the compensable harm and by allowing the plaintiff to recover full damages for that injury, courts “effectively allow[] a jury to speculate on causation because expert testimony that a physician's negligence probably caused the total damages is not required,” and thereby adopt “an extreme position [that] clearly distorts the traditional principles of causation”); *Walker*, *supra* note 18, at 249-51.

310. E.g., *Wollen v. DePaul Health Ctr.*, 828 S.W.2d 681, 685 (Mo. 1992); *Herskovits*, 664 P.2d 474, 477.

plaintiffs at all and have kept them from reaching the jury by reasoning that the 0.5 inference rule requires a holding that the plaintiff's evidence is legally insufficient.<sup>311</sup>

The analysis in this Article, however, leads to the conclusion that a finding of specific causation in a lost-chance case is not a factual or scientific matter, and that the issue should never depend solely or even principally upon the magnitude of the percentage in the major premise. Whether the available evidence is legally sufficient should depend largely on what kinds and levels of uncertainty are present in the evidence and in any direct inference from that evidence, and on whether those uncertainties are acceptable given the tort context. What has happened in these cases is that many courts have recognized the fairness of letting the jury decide each case on its own evidence.<sup>312</sup> Unfortunately, some of these courts have modified the concepts of causation or compensable injury in order to do so. But it is the mistaken 0.5 inference rule applied to specific causation that brought about these conceptual changes and made judges reluctant to adopt new rules about specific causation that are justified squarely on policy grounds.

Moreover, in the lost-chance cases, clearly there are policy rationales for helping the plaintiff get to the jury. For example, in most cases, the defendant had a physician-patient relationship with the plaintiff. Some courts have held that the defendant therefore undertook a duty of care to protect the plaintiff even from increased risk.<sup>313</sup> This special relationship supports a fairness argument for placing the cost of particular types of uncertainty on the defendant. Another fairness rationale for assisting the plaintiff is the notion that the defendant's negligence caused the lack of probative evidence.<sup>314</sup> The reasoning is that if there had been no negligence, then the factfinder would know that the plaintiff must be a baseline case. Although this Article cannot evaluate such policies and rules, they do serve to illus-

---

311. See, e.g., *Cooper*, 533 A.2d at 1299 (holding that the plaintiff failed to prove a "probability" of causation, where the chance of recovery was only "possible" and the court defined "probability" as a "greater than 50% chance": "[p]robability exists when there is more evidence in favor of a proposition than against it (a greater than 50% chance that a future consequence will occur)") (quoting *Pierce v. Johns-Manville Sales Corp.*, 464 A.2d 1020, 1026 (Md. 1983)); *Fennell v. S. Md. Hosp. Ctr., Inc.*, 580 A.2d 206 (Md. 1990) (holding that a 40% lost chance of survival was insufficient evidence that the defendant caused the plaintiff's death); *Cooper v. Sisters of Charity, Inc.*, 272 N.E.2d 97, 104 (Ohio 1971) (holding that expert opinion that expectation of survival was "[m]aybe . . . around 50%" was legally insufficient to create a jury issue).

312. *Chudson v. Ratra*, 548 A.2d 172, 179-80 (Md. Ct. Spec. App. 1988) (stating that "courts . . . have allowed juries to determine probabilities based directly or indirectly on universal statistics or even on the expert's general experience with other patients," either out of necessity or "based upon a tacit recognition that even estimates of probabil[ities] tailored specifically to a given patient are ultimately derived from generally accepted statistical norms").

313. E.g., *Thompson v. Sun City Cmty. Hosp., Inc.*, 688 P.2d 605, 615-16 (Ariz. 1984); *Falcon*, 462 N.W.2d at 51-52. See RESTATEMENT (SECOND) OF TORTS § 323 (1965).

314. E.g., *Thompson*, 688 P.2d at 616; *Falcon*, 462 N.W.2d at 49-51. Two commentators have proposed imposing liability if the defendant wrongfully caused "evidential damage." PORAT & STEIN, *supra* note 13, at 73-76, 160-84, 195-201. For a critique of this proposal, see Vern R. Walker, *Uncertainties in Tort Liability for Uncertainty*, 1 LAW, PROBABILITY & RISK 175, 179-84 (2002).

trate the kinds of non-epistemic rationales used to justify rules dealing with specific causation in lost-chance cases.

The analysis in this Article not only legitimates the adoption of policy-based rules, but also provides more nuanced predicates for those rules based on the layers of uncertainty involved in specific causation. In the case of any specific plaintiff, there may be some known risk factors for which there is adequate statistical information, some known risk factors with inadequate statistical information, some suspected risk factors with little or no quantitative information, and other pieces of information of unknown causal relevance. There may also be significant measurement, sampling, modeling, and causal uncertainties underlying the relative risk in the major premise. This array of uncertainties may be due in part to societal decisions about research funding, in part to a party's failure to generate additional information, in part to a party's failure to produce the evidence available to it, and in part to inherent uncertainties that are unlikely to be eliminated even if more scientific evidence were available. Judicial rules could use such distinctions to allocate the burdens of production and persuasion differently, depending on the type of uncertainty involved. For example, a court could impose on the plaintiff the burden of proving, with acceptable uncertainty, general causation between the defendant's negligence and the plaintiff's type of injury in a reference group roughly representative of the plaintiff. That is, the plaintiff must prove general causation, but need not prove specific causation. The burden might then shift to the defendant to prove a direct inference of specific causation.<sup>315</sup> That is, if the defendant wants the factfinder to rely on particular statistics to draw a direct inference to a probability for the plaintiff, then the defendant must prove that the reference group and its statistics adequately represent the individual plaintiff. When the defendant urges the factfinder to infer that the plaintiff probably would have died even absent the defendant's negligence, the defendant is urging the factfinder to draw a direct inference to the specific case. This allocation of burdens of proof allows meritorious cases to go to the jury without the need to alter traditional concepts of causation or compensable injury.

Broader institutional policies (not specific to lost-chance cases) can also come into play. To the extent that inherent uncertainties remain after all reasonable efforts have been taken to reduce or eliminate them, courts can look to non-epistemic grounds to justify decision rules for dealing with such cases. For example, in *Liriano v. Hobart Corp.*, Judge Calabresi wrote, for a court applying the tort law of New York, that "[w]hen a defendant's negligent act is deemed wrongful precisely because it has a strong propensity to cause the type of injury that ensued," the plaintiff has satisfied his burden of producing legally sufficient evidence of specific causation.<sup>316</sup> Judge Calabresi also stated that "[i]n such situations, rather than requiring the

---

315. See Walker, *supra* note 18, at 303.

316. 170 F.3d 264, 271 (2d Cir. 1999).

plaintiff to bring in more evidence to demonstrate that his case is of the ordinary kind, the law presumes normality and requires the defendant to adduce evidence that the case is an exception."<sup>317</sup> In *Zuchowicz v. United States*, Judge Calabresi wrote (for a court applying the tort law of Connecticut) that

when a negative side effect is demonstrated to be the result of a drug, and the drug was wrongly prescribed in an unapproved and excessive dosage (*i.e.* a strong causal link has been shown), the plaintiff who is injured has generally shown enough to permit the finder of fact to conclude that the *excessive dosage* was a substantial factor in producing *the harm*.<sup>318</sup>

Such a rule protects patients against the often unknown risks of ingesting prescription drugs at higher-than-approved dosages and is in line with a broader rule that presumes a causal connection between violation of a statute and the very type of accident the statute was intended to prevent.<sup>319</sup> In certain kinds of cases, therefore, the evidence of general causation might satisfy the plaintiff's ordinary burden on specific causation.

Another line of cases (besides the lost-chance cases) that illustrates judicial errors of logic about specific causation deals with sufficiency-of-evidence rules about epidemiologic evidence. Some courts have adopted the rule that if the available evidence is entirely epidemiologic in nature, then a finding of specific causation is not warranted unless the relative risk of the type of outcome over the baseline risk is greater than 2.0.<sup>320</sup> They have applied this rule even outside the context of a patient's lawsuit against her doctor.<sup>321</sup> When relative risk is less than 2.0, the number of exposure-caused

317. *Id.*

318. 140 F.3d 381, 391 (2d Cir. 1998) (emphasis added).

319. *See id.* at 390-91.

320. *See, e.g., DeLuca v. Merrell Dow Pharm., Inc.*, 911 F.2d 941, 957-59 (3d Cir. 1990); Carruth & Goldstein, *supra* note 257; Finley, *supra* note 257, at 347-64. In *DeLuca*, the Court of Appeals reasoned from the plaintiff's burden of proving causation by "a more likely than not standard" to a requirement that epidemiologic evidence alone would be legally insufficient evidence of specific causation unless it showed a "relative risk of limb reduction defects" of at least two, quoting with approval the following passage from *Manko v. United States*, 636 F. Supp. 1419, 1434 (W.D. Mo. 1986), *aff'd in relevant part*, 830 F.2d 831 (8th Cir. 1987):

A relative risk of "2" means that the disease occurs among the population subject to the event under investigation twice as frequently as the disease occurs among the population not subject to the event under investigation. Phrased another way, a relative risk of "2" means that, on the average, there is a fifty per cent likelihood that a particular case of the disease was caused by the event under investigation and a fifty per cent likelihood that the disease was caused by chance alone. A relative risk greater than "2" means that the disease more likely than not was caused by the event.

*DeLuca*, 911 F.2d at 958-59. *See also* Black & Lilienfeld, *supra* note 181, at 767 ("If, in an exposed population, more than half the cases of a disease can be attributed to the exposure . . . then absent other information about a diseased individual, it is more likely than not that his or her illness was caused by the exposure."). Moreover, epidemiologic evidence cannot be sufficient to establish causation unless the relative risk is "greater than 2." *Id.* at 769.

321. *E.g., DeLuca*, 911 F.2d at 941 (the claim alleged that Bendectin was a defective pharmaceutical

cases in the reference group is estimated to be less than the number of baseline cases. These courts consider it therefore unreasonable for a factfinder to infer that the specific plaintiff is more likely to be an exposure-caused case than a baseline case.<sup>322</sup> The analysis in this Article demonstrates, however, that without a warranted finding that the reference group in the major premise is acceptably complete and representative of the particular plaintiff, then the magnitude of the relative risk is of unknown relevance to the plaintiff. Even a relative risk much higher than 2.0 would be of unknown relevance.<sup>323</sup> Therefore, when there are significant uncertainties inherent in any direct inference, then the cases should be decided on grounds of fairness or other substantive policies. Courts cannot avoid adopting appropriate policies in these cases by relying instead on mistaken reasoning about the probative value of the relative risk.

### 3. Judges as Gatekeepers of Evidence and Rules of Admissibility

A series of cases in the federal courts of the Ninth Circuit illustrates how judges have used flawed reasoning about specific causation to adopt rules about the admissibility of expert testimony. The Supreme Court's decision in *Daubert v. Merrell Dow Pharmaceuticals, Inc.* ("*Daubert*") was on a writ of certiorari to the United States Court of Appeals for that circuit.<sup>324</sup> In that case, the plaintiffs brought suit against the marketer of the prescription drug Bendectin, claiming that their serious birth defects were caused by their mothers' ingestion of the drug.<sup>325</sup> The Supreme Court held that Federal Rule of Evidence 702, which governs the admissibility of expert testimony in federal courts, sets out a two-prong test.<sup>326</sup> First, for any assertion of the expert to be admissible as "scientific knowledge" under the rule, it must meet a standard of "evidentiary reliability" by being "derived by the scien-

---

product).

322. Some judicial opinions struggle to sort out the logical issues involved, but fail. One court, while recognizing "that there is not a precise fit between science and legal burdens of proof," remained persuaded "that there is a rational basis for relating the requirement that there be more than a 'doubling of the risk' to our no evidence standard of review and to the more likely than not burden of proof." *Merrell Dow Pharm., Inc. v. Havner*, 953 S.W.2d 706, 717 (Tex. 1997). Although the court explicitly did not hold "that a relative risk of more than 2.0 is a litmus test," it nevertheless used that test repeatedly in rejecting the proffered epidemiologic evidence as insufficient. *Id.* at 718, 724-28.

323. Some authors have suggested that a relative risk greater than 2.0 is a "threshold" that would permit an inference to specific causation, provided that individuating and biasing factors are taken into account. *E.g.*, Green et al., *supra* note 6, at 384-86 (stating that such "threshold" reasoning is conditioned on a number of assumptions, including that "the relative risk found in the [epidemiologic] study is a reasonably accurate measure of the extent of disease caused by the agent" and that "the plaintiff in a given case is comparable to the subjects who made up the exposed cohort in the epidemiologic study and that there are no interactions with other causal agents"). This Article provides the logical foundation for such suggestions, and guidance for developing rules in pursuit of them.

324. 509 U.S. 579, 585 (1993).

325. *Id.* at 582.

326. *See id.* at 589-92.

tific method.”<sup>327</sup> Second, the testimony must “assist the trier of fact” by being relevant to the issues in the particular case.<sup>328</sup>

Upon remand, the court applied this standard to the proffered expert testimony in the case.<sup>329</sup> “*Daubert II*” states that only one of the plaintiffs’ experts, Dr. Palmer, offered an opinion of specific causation based on plaintiff-specific information.<sup>330</sup> The court held, however, that “Dr. Palmer offer[ed] no tested or testable theory to explain how,” from his review of the plaintiffs’ medical records, “he was able to eliminate all other potential causes of birth defects.”<sup>331</sup> The court held as a matter of law that Dr. Palmer’s testimony failed *Daubert*’s first prong requiring “sound science.”<sup>332</sup>

The court divided the remainder of the plaintiffs’ expert testimony into three categories based on the nature of the supporting evidence—namely, statistics drawn from epidemiologic studies, conclusions based on causal studies in laboratory animals, and conclusions based on similarity of chemical structure between Bendectin and other drugs suspected of causing birth defects.<sup>333</sup> With regard to the last two categories, the court probably doubted that the testimony could pass the first-prong test of scientific soundness, especially given the lack of scientific consensus about causation in humans.<sup>334</sup> Those doubts, however, would normally require a remand to the district court for further proceedings under the new standard.<sup>335</sup> The testimony based on animal studies and chemical structure, however, had a fatal flaw as a matter of law under the second prong.<sup>336</sup> The test under the second prong is whether there is a relevance “fit” between the proffered testimony and specific causation.<sup>337</sup> The animal-study and chemical-structure experts, however, only testified “to a possibility rather than a probability” and did not “quantify this possibility.”<sup>338</sup> Therefore, their conclusions failed to fit

---

327. *Id.* at 589-90.

328. *Id.* at 591-92.

329. *Daubert v. Merrell Dow Pharm., Inc.*, 43 F.3d 1311, 1314-16, 1322 (9th Cir. 1995) (affirming the district court’s grant of summary judgment without remand after applying “the new standard announced by the Supreme Court”). This case will be referred to as “*Daubert II*”.

330. *Daubert II*, 43 F.3d at 1319 (describing Dr. Palmer as “the only expert willing to testify ‘that Bendectin did cause the limb defects in each of the children’”) (emphasis added); *id.* at 1320 n.12 (stating that “[u]nlike the other experts, who speak in terms of probabilities, Dr. Palmer goes so far as to conclude that plaintiffs’ injuries were in fact caused by Bendectin rather than another cause”).

331. *Id.* at 1319 (agreeing “with the Sixth Circuit’s observation that ‘Dr. Palmer does not testify on the basis of the collective view of his scientific discipline, nor does he take issue with his peers and explain the grounds for his differences’”).

332. *Id.* at 1316, 1319.

333. *Id.* at 1314.

334. *See id.*

335. *Id.* at 1320 (concluding that “[w]here [the first-prong test] the only question before us, we would be inclined to remand to give plaintiffs an opportunity to submit additional proof that the scientific testimony they proffer was ‘derived by the scientific method’”).

336. *See id.* at 1322.

337. *Id.* at 1320 (holding as the “traditional burden” of the plaintiffs that “they must prove that *their* injuries were the result of the accused cause *and not some independent factor*”) (emphasis added).

338. *Id.* at 1322 (citing *Turpin v. Merrell Dow Pharm., Inc.*, 959 F.2d 1349, 1360 (6th Cir. 1992)).

the plaintiffs' need to prove specific causation by a preponderance of the evidence.<sup>339</sup>

The only remaining testimony was the epidemiology-based testimony. Here again, the court saw no need for a remand, either under the first prong of *Daubert*<sup>340</sup> or under the "fit" requirement of the second prong.<sup>341</sup> The court did not remand for a determination of "fit" because it applied a rule something like the following:<sup>342</sup>

*If the plaintiff has the burden of proving specific causation by a preponderance of the evidence,*

*and if the only or principal evidentiary support for causation consists of statistics derived from epidemiologic evidence,*

*then expert testimony on causation can satisfy the "fit" requirement of Daubert only if there is sufficient evidence for a reasonable factfinder to find that the relative risk created by the hazard for the plaintiff's type of injury is greater than 2.0 when compared to the baseline risk for that same type of injury.*

Although *Daubert II* leaves some uncertainty about the scope of the precise rule, this formulation reflects the understanding of later judges or courts that attempted to apply or distinguish the rule.<sup>343</sup> The only rationale that the court gives for adopting this rule is the preponderance standard of proof.<sup>344</sup> The apparent reasoning is that a factfinder, in the face of such evidence, should employ a 0.5 inference rule, and the evidence would, therefore, have to show a doubling of the risk in order to satisfy the "fit" test of admissibility. The court then reviewed the proffered expert opinions that were based on epidemiology, found that none of them even *claimed* that ingesting Bendectin during pregnancy more than doubled the risk,<sup>345</sup> and held as a matter

---

339. *Id.*

340. *Id.* at 1314-16, 1319-20 (stating that the court's "inclination" to remand also covered the statistical testimony based on epidemiologic studies).

341. *Id.* at 1320-22.

342. *Daubert II* contains several statements of the rule, including the following two versions: "In terms of *statistical proof*, . . . plaintiffs must establish not just that their mothers' ingestion of Bendectin increased somewhat the likelihood of birth defects, but that it *more than doubled* it—*only then* can it be said that Bendectin is *more likely than not* the source of their injury." *Id.* at 1320 (emphasis added). "For an *epidemiological study* to show causation under a *preponderance standard*, the *relative risk* of limb reduction defects arising from the epidemiological data . . . will, at a *minimum*, have to *exceed '2.'*" *Id.* at 1321 (internal quotations omitted and emphasis added).

343. *E.g., In re Hanford Nuclear Reservation Litig.*, 292 F.3d 1124, 1136 (9th Cir. 2002) (interpreting *Daubert II* as adopting a "doubling of the risk" test for a case in which "there was no definitive evidence that Bendectin is a substance capable of causing birth defects," and therefore "statistical epidemiological evidence" was "necessary," and "plaintiffs relied primarily on epidemiological evidence").

344. *See Daubert II*, 43 F.3d at 1320 (giving as the basis for its rule that "California tort law requires plaintiffs to show not merely that Bendectin increased the likelihood of injury, but that it more likely than not caused *their* injuries").

345. *Id.* at 1320-21.



of law that the proffered testimony was inadmissible because it failed the second prong of *Daubert*.<sup>346</sup>

From a purely epistemic standpoint, the reasoning behind this admissibility rule requiring  $RR > 2.0$  has several logical flaws.<sup>347</sup> First, the probative value of the relative risk is unknown without a warranted finding that it describes a reference group that is adequately representative for the individual subject of the direct inference. One might argue that the mere relevance of a relative-risk value is established if the data are about human beings, a group of which the plaintiffs are members. But the *magnitude* of the relative-risk statistic has *unknown probative* value for a direct inference unless the evidence warrants a finding that the reference group is adequately representative of the subject of that inference, and there is a decision that any incompleteness in defining the reference group is acceptable. Without acceptable uncertainty about plaintiff-representativeness, the stability of the relative-risk statistic is unknown, and that statistic might increase or decrease as additional factors are taken into account.<sup>348</sup>

Second, the evidence in the *Daubert* case convinced the court that scientists do not know the causal mechanisms that bring about the kind of limb reduction suffered by the plaintiffs.<sup>349</sup> In fact, the court thought that there was ample evidence in the record for finding that most causes of limb reduction are unknown.<sup>350</sup> Therefore, there was good evidence that *no* reference-group profile could be complete and that *any* direct inference to a specific case would be epistemically unwarranted. If this is true, then it is mistaken to conclude (or suggest) that if the relative-risk values from the same epidemiologic studies had happened to be higher than 2.0, then the probability in the individual case would have been over 0.5. Such reasoning is misleading not only to plaintiffs, but also to judges, who might feel relieved of the burden of giving policy justifications for their admissibility rules.

Third, once the logic of direct inference is understood, it should be clear that there is nothing peculiarly different about epidemiologic evidence. It does not matter which scientific methodology establishes the general causal relevance of the various risk factors. General causal relationships might be established by toxicology using animal models, epidemiology using human data, mechanistic experiments at the subcellular level, or other methods. The evidence does not even need to be scientific in any strict sense, for many areas of specialized activity rely upon accepted generalizations about

---

346. *Id.* at 1320-22.

347. For a critique of simplified admissibility rules, including relative risk rules, see Cranor et al., *supra* note 253, at 25-62 (arguing that it is "important to avoid the temptation to adopt overly stringent admissibility rules for scientific evidence").

348. The court also erred in stating that "[a] relative risk of less than two . . . actually tends to disprove legal causation, as it shows that Bendectin does not double the likelihood of birth defects." *Daubert II*, 43 F.3d at 1321.

349. *Id.* at 1313-14.

350. *Cf. id.* at 1320 (stating that "we know that some [birth] defects—including limb reduction defects—occur even when expectant mothers do not take Bendectin, and that most birth defects occur for no known reason").

causation.<sup>351</sup> The plaintiff-specific reference-group profile should include all known or suspected causally relevant factors, regardless of the type of evidence that establishes the causal relevance. The decisions that factfinders must make about uncertainty are not created because the available evidence happens to be epidemiologic. Unfortunately, once the doubling-of-risk admissibility rule is adopted as a rule about *epidemiologic* evidence, courts and parties may start looking for *better kinds* of evidence than epidemiology. Such a quest, however, misunderstands the nature of the logical problem.

Finally, because the court thought that its admissibility rule followed logically from the preponderance-of-the-evidence standard of proof, it did not bother to explore policy justifications for its rule. Such mistaken reasoning by the court is not harmless error, because the court ended up adopting a doubling-of-risk admissibility rule without giving any appropriate policy grounds for doing so. The court failed to appreciate that the epistemic reasoning it gave was flawed and also failed to supply any other justification. Perhaps if the court had analyzed the direct inference correctly, it would have sought out more appropriate policy rationales and better justifications for its rules of admissibility.

These logical errors have continued to infect cases in the Ninth Circuit, as illustrated by the legal fate of claims concerning exposure to radioactive emissions from the Hanford Nuclear Weapons Reservation in southeastern Washington.<sup>352</sup> Thousands of plaintiffs alleged that they were harmed by exposure to such emissions and sued the parties that operated the Hanford facility when the emissions occurred.<sup>353</sup> In 1991, the district court consolidated all of the actions and later divided discovery into three phases.<sup>354</sup> Phase I consisted of document production and interrogatories about the operating and emissions history and about the “plaintiffs’ exposures, medical histories, and relevant illnesses and injuries.”<sup>355</sup> Phase II was to “focus on causation,” including expert witness reports.<sup>356</sup> The court later bifurcated this second phase into two parts, dealing with generic causation and specific causation.<sup>357</sup> The final Phase III would include general liability and other remaining issues.<sup>358</sup>

---

351. Federal Rule of Evidence 702, the rule at issue in *Daubert* and *Daubert II*, applies to “scientific, technical, or other specialized knowledge.” FED. R. EVID. 702. The logical analysis of this Article applies to all direct inferences to specific causation, and therefore to all expert testimony about specific causation, whether scientific or not. *E.g.*, *Kumho Tire Co., Ltd. v. Carmichael*, 526 U.S. 137, 143-47, 153-58 (1999) (involving expert testimony on the cause of a specific tire blow-out).

352. *In re Berg Litig.*, 293 F.3d 1127, 1129 (9th Cir. 2002); *In re Hanford Nuclear Reservation Litig.*, 292 F.3d 1124, 1126-27 (9th Cir. 2002).

353. *Hanford*, 292 F.3d at 1126-27.

354. *Id.* at 1128-29; *In re Hanford Nuclear Reservation Litig.*, No. CY-91-3015-AAM, 1998 WL 775340, \*2 (E.D. Wash. Aug. 21, 1998).

355. *Hanford*, 292 F.3d at 1129; *In re Hanford Nuclear Reservation Litig.*, 1998 WL 775340, at \*2.

356. *Hanford*, 292 F.3d at 1129.

357. *Id.*

358. *Id.*

In March and June of 1997, while discovery was still in the generic-causation stage of Phase II, and while discovery on individual medical causation was not yet permitted under the district court's discovery management order, the defendants filed motions for summary judgment.<sup>359</sup> In August of 1998, the district court entered a 762-page order that largely granted the defendants' motion.<sup>360</sup> The few claims that survived summary judgment were those that passed muster under the district court's interpretation and application of the doubling-of-risk rule from *Daubert II*.<sup>361</sup> Plaintiffs who were dismissed from the litigation by the district court's summary judgment appealed, contending both procedural error in granting summary judgment at the generic causation phase of discovery and substantive error in applying the doubling-of-risk rule.<sup>362</sup> In June 2002, the Court of Appeals reversed the lower court's grant of summary judgment and remanded for further proceedings.<sup>363</sup> The litigation therefore raised two issues: how to distinguish between general causation and specific causation in practice, and how to interpret and apply the admissibility rule adopted in *Daubert II*.

The distinction between general causation and specific causation can begin to blur in toxic tort cases because the dose actually received by a particular plaintiff appears relevant to deciding whether the toxic agent *can* cause adverse effects at that dose.<sup>364</sup> The reasoning and the doubling-of-risk rule of *Daubert II* compounds the confusion by testing the admissibility of expert testimony in part by the "fit" of the testimony to proving specific causation. The district court's partial summary judgment in *Hanford* excluded expert testimony on generic causation that did not pass this test of "fit."<sup>365</sup> The Court of Appeals reversed the district court's application of the *Daubert II* admissibility rule because in *Hanford*, unlike *Daubert II*, there was ample evidence in the record that radiation can cause a broad range of injuries even at low doses.<sup>366</sup> In reaching this decision, however, the appellate court may have sown even more confusion with the following comment on the *Daubert II* rule: "It is critical to stress that the plaintiffs in *Daubert II* had no scientific evidence that Bendectin was capable of causing birth defects (generic causation), and therefore were required to produce epidemiol-

---

359. *Id.* at 1129-30, 1134-35; *In re Hanford Nuclear Reservation Litig.*, 1998 WL 775340, at \*9.

360. *Hanford*, 292 F.3d at 1131.

361. *Id.* at 1131-32; *In re Hanford Nuclear Reservation Litig.*, 1998 WL 775340, at \*328-32.

362. *Hanford*, 292 F.3d at 1133-37.

363. *Berg*, 293 F.3d at 1133; *Hanford*, 292 F.3d at 1138-39.

364. See, e.g., *Sterling v. Velsicol Chem. Corp.*, 855 F. 2d 1188, 1200 (6th Cir. 1988) (stating that "generic and individual causation may appear to be inextricably intertwined" when dealing with "a kind of generic causation—whether the combination of the chemical contaminants and the plaintiffs' exposure to them had the capacity to cause the harm alleged"); *Hanford*, 292 F.3d at 1130-35 (describing the confusion and arguments over how to classify contested issues of fact and law when discovery was bifurcated into generic causation and individual causation).

365. *Hanford*, 292 F.3d at 1132 ("Expert testimony indicating only that the radiation emitted from Hanford was capable of causing a disease was excluded as irrelevant unless it also passed muster under the 'doubling of the risk' standard, i.e., unless the expert opined that the radiation emissions amounted to a 'doubling dose.'"); *In re Hanford Nuclear Reservation Litig.*, 1998 WL 775340, at \*13, \*330.

366. *Hanford*, 292 F.3d at 1137.

ogical studies to prove that Bendectin more likely than not caused their own particularized injuries (individual causation)."<sup>367</sup>

This reasoning raises a number of questions. First, the plaintiffs in *Daubert II* were in fact trying to introduce not only epidemiologic evidence, but also evidence from laboratory animal studies and chemical structure analyses.<sup>368</sup> The court in *Daubert II* excluded that proffered scientific evidence as inadmissible, not because it was irrelevant to general causation, but precisely because it lacked the required "fit" to specific causation and therefore would not be helpful to the trier of fact.<sup>369</sup> Therefore, it would seem that the district court in *Hanford* was interpreting *Daubert II* correctly. On remand, therefore, the district court may conclude that its only error was applying the *Daubert II* rule too early in the proceeding, but that using the rule against the plaintiffs would be appropriate during the specific-causation stage of discovery.

Second, the passage reinforces the suggestion that epidemiologic studies are not "scientific" and have some lower epistemic status. This reinforces the notion that the *Daubert II* admissibility rule should apply only to epidemiologic evidence and has no relevance to (real) "scientific evidence." For reasons argued above, this belief that the problem of warranting a direct inference to specific causation is due to the nature of epidemiologic evidence is a mistake of logic. And it is not a harmless mistake, because the belief may place needless pressure on future litigation to police a bright line between "epidemiological studies" and "scientific evidence."

A logical analysis of direct inference can clarify these confusions and put tort law about specific causation on a proper policy foundation. General causation is indeed a logically prior and distinct issue from specific causation. A reasonable factfinder would first determine the list of causally relevant variables for the plaintiff's kind of injury. In toxic tort cases where extent of exposure is critical, general causation includes characterizing the available dose-response knowledge, especially within the range of exposures allegedly relevant to the plaintiff's case. If exposure to the hazard is found to be a causally relevant factor for the plaintiff's type of injury, then the investigation of specific causation can begin by establishing the plaintiff's reference-group profile on all of the variables known or suspected to be causally relevant. The specific-causation inquiry can then proceed to the questions of adequate completeness of the causal model and the uncertainty associated with the plaintiff's relative risk. Therefore, the line between general and specific causation can be reasonably bright if drawn along logical

---

367. *Id.* at 1136-37.

368. *Daubert II*, 43 F.3d at 1314-16, 1319-22.

369. If the decision in *Daubert II* had not rested on "fit" to specific causation, the court would have ordered a remand to the district court. *See supra* text accompanying notes 333-39. As the Supreme Court reasoned in *Daubert*, the study of the phases of the moon may well provide valid scientific knowledge, but "(absent creditable grounds supporting such a link), evidence that the moon was full on a certain night will not assist the trier of fact in determining whether an individual was unusually likely to have behaved irrationally on that night." *Daubert*, 509 U.S. at 591.

lines. Courts can adopt decision rules within those boundaries on appropriate policy grounds, and judges can manage complex tort cases using those rules and boundaries.

Once the conceptual boundaries are clear, this sharpens the issue of adopting a doubling-of-risk rule of admissibility. The analysis in this Article leads to the conclusion that the rule, as applied in *Daubert II* and by the district court in *Hanford*, rests upon an error about the logic of warranted findings about specific causation. On remand, the district court in *Hanford* should resolve those logical issues correctly. A relative risk value is of unknown probative value to any particular plaintiff unless the reference group is acceptably representative of that plaintiff. In the case of some plaintiffs, with certain exposures and injuries, the causal model may be well enough understood that an expert opinion about specific causation is admissible under *Daubert*. In many other cases, however, there may be good evidence that the reference-group profile is substantially incomplete or that the available statistics for the reference group are unacceptably uncertain. In such cases, when plaintiffs have taken all reasonable steps to produce sufficient evidence, the courts should decide on policy grounds when to dismiss the cases and when to give them to the jury. It seems unlikely, however, that a bright-line rule based on the mere magnitude of relative risk will be justifiable on policy grounds. What the courts are not free to do is to take refuge in the illusion that this is a question of logic, not policy. Logic dictates that it must be a matter of policy.

### CONCLUSION

The traditional tort requirement that the plaintiff must prove specific causation by a preponderance of the evidence has led to mistaken judicial reasoning in recent cases. Courts have used faulty logic about the warrant for specific causation to adopt rules for factfinding, rules for deciding that the totality of admitted evidence is legally insufficient, and rules for excluding expert testimony from the case. In mistaking the logical nature of the problem, they have sometimes limited these rules to toxic tort cases, or to epidemiologic evidence, or to scientific evidence. In so reasoning and ruling, they have removed individual plaintiffs as the subjects of factfinding and have substituted instead an abstract, "statistical individual." This Article, by analyzing the logic of warranted direct inference about specific causation, shows the way out of these errors. Direct inferences are warranted by findings that six types of inherent uncertainty are within acceptable bounds: measurement uncertainty, sampling uncertainty, modeling uncertainty, causal uncertainty, uncertainty about plaintiff-representativeness, and uncertainty about assigning a probability to a specific plaintiff. In cases involving significant uncertainty—whether in products liability, medical malpractice, or toxic torts—decisions about acceptable uncertainty are neither factual nor scientific in nature. Those decisions should depend, not upon mistaken statistical reasoning, but upon common sense, fairness, and justice, as well as

on the substantive and process policies of tort law. It is time to correct the central fallacy, reject the “junk logic” that has wrongly driven the cases, and restore the individual plaintiff to the factfinding process.

