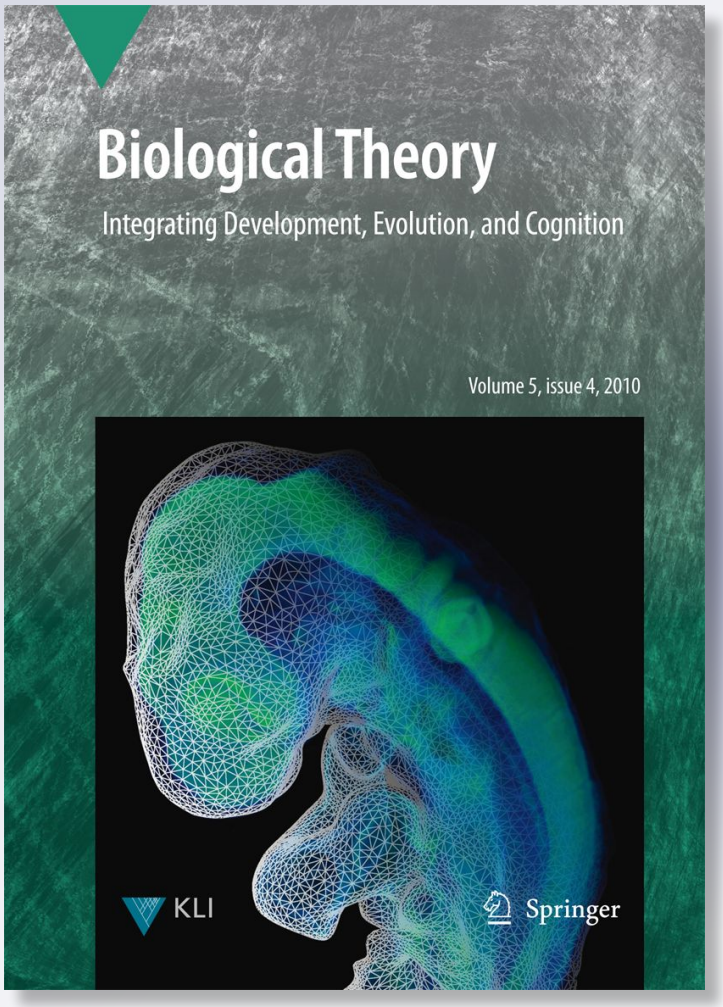
Randomization and Rules for Causal Inferences in Biology: When the Biological Emperor (Significance Testing) Has No Clothes

Kristin Shrader-Frechette

Biological Theory

ISSN 1555-5542

Biol Theory
DOI 10.1007/s13752-012-0021-y



Springer

Biol Theory
DOI 10.1007/s13752-012-0021-y

LONG ARTICLE

Randomization and Rules for Causal Inferences in Biology: When the Biological Emperor (Significance Testing) Has No Clothes

Kristin Shrader-Frechette

Received: 2 December 2011 / Accepted: 20 January 2012

Abstract Why do classic biostatistical studies, alleged to provide causal explanations of effects, often fail? This article argues that in statistics-relevant areas of biology—such as epidemiology, population biology, toxicology, and vector ecology—scientists often misunderstand epistemic constraints on use of the statistical-significance rule (SSR). As a result, biologists often make faulty causal inferences. The paper (1) provides several examples of faulty causal inferences that rely on tests of statistical significance; (2) uncovers the flawed theoretical assumptions, especially those related to randomization, that likely contribute to flawed biostatistics; (3) re-assesses the three classic (SSR-warrant, avoiding-selection-bias, and avoiding-confounders) arguments for using SSR only with randomization; and (4) offers five new reasons for biologists to use SSR only with randomized experiments.

Keywords Alcohol · Biostatistics · Causal inference · Epidemiology · Experimental study · Observational study · Randomization · Statistics · Tobacco

To assess causal hypotheses, many biologists use null-hypothesis testing (NHT) and the statistical-significance rule (SSR)—that the no-effect hypothesis ought to be rejected only if there is statistically significant evidence for some particular effect ($p \leq 0.05$). In 2011, biologists said SSR is used in over 90 % of papers in ecology/evolution and is "often required by [biology] journal editors" (Gerrodette 2011, p. 404; see Fidler 2006). Even for observational data, biologists routinely use SSR (e.g., Banebrake et al. 2010; Dietl et al. 2010; Knutsen et al. 2011) and staunchly defend it (e.g., Garamszegi et al. 2009; Legendre and Legendre 2011). Perhaps biologists want to overcome purely phenomenological methods—to promote rigorous falsificationism. As one botanist noted, although ecology used to be a field of "merry naturalists," ecologists now want to meet the "heavy demands of objectivity" (Okland 2007, p. 123).

Because biologists sometimes ignore epistemic constraints on SSR, this paper argues that their causal inferences fail. It shows that using SSR for observational research relies on false presuppositions about randomization, SSR warrants, and applications of probability laws in biology.

Four Classic Cases

How do biologists err in making SSR-based causal inferences in observational studies? Consider four classic studies—on tobacco, alcohol, hormones, and petrochemical pollution. The Framingham (Massachusetts) biostatistical tobacco research [National Heart, Lung, and Blood Institute (NHLBI) 2009], "among the most informative in epidemiologic history" (Greenland 1990, p. 425), has a very low p value and sample sizes in the hundreds of thousands. It is the major research alleging tobacco causes heart attacks. Yet, it has been conclusively disproved (e.g., Empana et al. 2003).

Similarly, classic studies comparing biological effects of moderate alcohol intake, versus abstention, claim statistically significant, reduced risks of mortality/cardiovascular

K. Shrader-Frechette (✉)
Department of Biological Sciences, University of Notre Dame, Notre Dame, IN, USA
e-mail: kshrader@nd.edu

K. Shrader-Frechette
Department of Philosophy, University of Notre Dame, Notre Dame, IN, USA

Published online: 04 April 2012

disease, and that drinking helps prevent dementia and diabetes (e.g., Klatsky 1996; Rimm et al. 1996; McConell et al. 1997; Ruitenberg et al. 2002; Ellison 2005; Collins et al. 2009). Experts say scientific data, for alcohol's cardioprotective effects, are stronger than for any other dietary constituents (Ellison 2005). Yet, these studies also have been conclusively disproved [International Agency for Research on Cancer (IARC) 2008].

Likewise, for 50 years, classic menopausal-hormone-replacement studies touted homones' health-protective benefits—helping prevent Alzheimer's, cognitive decline, and heart disease (e.g., Grodstein et al. 2000; Sherwin 2000; Fillit 2002; Ferrara et al. 2003: Naessen et al. 2007; Tannen et al. 2007). Yet, these studies too have been conclusively disproved (e.g., Bath and Gray 2005; Strom et al. 2006). Finally, UCLA studies denied harms from ChevronTexaco petrochemical pollution in the Amazon (Kelsh et al. 2008, p. 393). They too have been conclusively disproved (e.g., O'Rourke and Connolly 2003; Hurtig and San Sebastián 2005).

What explains flaws in once-classic studies? Financial conflicts of interest (COI) may play a role, as the alcoholic-beverage industry funded/performed the pro-alcohol studies [Alcohol Beverage Medical Research Foundation (ABMRF) 2011], the pharmaceutical industry funded/performed the pro-hormone studies [Singer 2009; University of California, San Francisco (UCSF) 2011], and ChevronTexaco funded the Amazonian petrochemical studies (Kelsh et al. 2008, p. 393). Yet, the failed-replication tobacco study had no apparent COI. Indeed, it reversed years of biased tobacco-funded research (NHLBI 2009).

Epistemic Problems with the Classic Studies

What epistemic factors might explain flaws in this classic research? Consider case-specific methodological flaws. For instance, the UCLA research examined only entire cantons/counties/provinces, not oil-production areas; only dirty urban, not rural, comparison regions, and thus underestimated oil-related harms. Similarly, Framingham studies had no long-term follow-up, based on all biological endpoints (e.g., Greenland 1977; 1990).

Besides case-specific flaws, the four studies appear to share at least three shortcomings, ignoring (1) complex biological interactions, (2) selection bias, and (3) confounders. Regarding (1), because all four observational studies deal with biological effects mediated by complex interactions, not controlled by a single effect, using SSR is risky because SSR presupposes linearity and estimates parameters by quantifying an effect variation attributable to an independent covariate (Shrader-Frechette 2008a). Regarding (2), none of the studies had randomly chosen subjects. For instance, the alcohol research ignored selection biases, e.g., sick people probably do not drink; moderate drinkers likely are more socially advantaged, thus healthier, than abstainers (e.g., Fillmore 2000; Roizen and Fillmore 2000; Fillmore and Kerr 2002). Regarding (3), the studies failed to control for known confounders and confused correlations with causes. The hormone-replacement research, for instance, failed to control for wealth. Yet in the US, wealthier, thus better medically insured, thus healthier, women tended to be hormone-replacement-therapy users—not poorer, uninsured women. What methodological techniques help scientists take account of (1)–(3)? All other things being equal, large sample sizes more likely include more complex interactions; more subjects with different characteristics; thus fewer selection biases and greater ability to control for confounders. However, all four sets of studies had sample sizes in the thousands or hundreds of thousands.

What else helps scientists take account of (1)–(3)—complex interactions, selection biases, and confounders? Randomization of experimental/control data reduces selection biases and encourages random distribution of confounding/variables—rather than distribution by complex social, physical, and psychological factors (Rothman 1990, p. 417). But only experimental, not observational, studies employ randomization. Because the four classic—and many biological—studies are observational, they may fall victim to (1)–(3).

Why Biologists do Observational Research

Although few scientists initially followed Charles Sanders Peirce's example of randomizing experimental data (Peirce and Jastrow 1885), by the 1920s, historians say biologists recognized the benefits of using randomized experiments, not systematic experimental designs (Fisher 1925; Hall 2007, p. 311). If so, why do biologists often do nonrandomized, observational research?

One reason is that large, experimental studies are expensive (Shrader-Frechette 2008b). Also, controlling required conditions for experimental studies is difficult, e.g., testing hypotheses of island biogeography. Given practical difficulties, ecologists often employ "shortcut tests" that rely on assumptions such as that artificial substrates are analogous to islands, or that lists of taxa are comparable on different islands. Even one of the best "tests" of island biogeography—when E. O. Wilson and Dan Simberloff censused islands in the Florida Keys for terrestrial arthropods, then fumigated some islands to kill them (Simberloff 1976; Simberloff and Wilson 1970)—was flawed by periodic sampling, failure to test many different systems, and so on.

A third reason for observational SSR studies is that often biologists are unsure about state variables, null hypotheses, what factors must be controlled experimentally, etc. Apparent ecological patterns change, because of heritable variations and evolution (e.g., Sober 1988; Hart and Marko 2010, p. 643; Mehner et al. 2011), and biologists incompletely understand biological processes (Peters 1991; Shrader-Frechette and McCoy 1993, p. 80ff.; Mariani 2008; Roughgarden 2009; Sillero 2011). A fourth reason for observational biostatistical research is that studies often must begin after biological catastrophe, e.g., hurricane destruction. Yet only proactive studies can be randomized. Besides, for biological research on humans, observation often is necessary because classical bioethics prohibits experiments that may harm them (e.g., Beauchamp and Childless 1989; see Foot 1978).

Controversy over Using SSR in Observational Biological Research

Although many biologists do observational research that employs SSR, does SSR improve causal inferences about observational data? Dominant "black-box" biologists (BBB) answer "yes." Minority-camp "eco-biologists" (EB) answer "no."

BBB argue for methodological rigor, emphasize specific population behaviors, argue for robust statistical analysis, and pay little attention to pathogenic mechanisms or sociocultural influences on biological effects. Examples of black-box inputs include counts of diseased/non-diseased individuals; outputs include relative-risk estimates. Responding to regulators' demands for causal "proof" of biological harm, BBB argue for increased use of SSR, even in non-experimental studies (Savitz 2003; see Cranor 2006, pp. 240–241). They claim SSR is a "best case" interpretation of observational results (Rothman 1990, p. 417; see Poole and Rothman 1998).

EB, however, say observational studies fail because of (1)–(3). EB thus accuse BBB of confusing correlations with causes and using SSR, without having unbiased (i.e., random) ways to select representative samples and experimental conditions. EB agree with Fisher-randomization helps enable probability-law applications, generate rationally based causal inferences, and avoid bias (Fisher 1947; Anderson 2001; see Hacking 1990, p. 206). Who is right about randomization, BBB or EB?

Randomization as SSR Warrant?

To answer this question, consider three prominent arguments. One, made by Fisher and frequentist statisticians, is that randomization of subjects/treatments is necessary to warrant SSR logic (Fisher 1925, 1947; Pearl 2000)—i.e., necessary to avoid systemic bias, unknown correlations, and questionable assumptions that otherwise occur. Using botanical fieldwork to defend randomization, Fisher argued that if scientists' experimental schemes correlated with unobservable, systematic non-uniformities in data, SSR would fail.

> For our test of significance to be valid, the difference in fertility between plots chosen as parallels must be truly representative of…different treatments; and we cannot assume that this is the case if our plots have been chosen in any way according to a pre-arranged system…[that] may have…features in common with the systematic variation of fertility….The direct way of overcoming this difficulty is to arrange the plots wholly at random. (Fisher 1925, pp. 224–225; quoted in Hall 2007, p. 313)

Proponents say randomization helps warrant SSR because it allows causal inferences, despite heterogeneous factors. Because models (like SSR) are only as good as their underlying homogeneity assumptions (about experimental/control groups), using SSR requires justifying homogeneity assumptions. Randomization provides that justification. Otherwise, the biological emperor, SSR, has no clothes—no warrant in observational studies.

However, those opponents—who deny randomization must warrant SSR claims—say scientists can construct randomized experiments that give wrong answers (Worrall 2007, p. 468; but see, e.g., La Caze et al. 2011). Why? They say randomization works only under ideal conditions that (1) all other causes affecting some effects have identical probability distributions in experimental/control groups, and (2) assignment of individuals to experimental/control groups is statistically independent of other causally relevant features (Cartwright 1994). Because (1) and (2) are usually unmet, randomization skeptics say unrecognized heterogeneities can control outcomes (Urbach 1985), except "in the indefinite long run"; they also say there is no "rational" measure of differences between given and ideal experiments (Worrall 2007, pp. 472–473). At best, they say randomization guarantees only internal, not external, validity—it guarantees some study is valid because it works deductively, but never guarantees transferability of causal results (Cartwright 2007a, b; Worrall 2007, p. 483; Cartwright and Munro 2010).

The preceding criticisms of randomization as SSR warrant have at least four problems. First, they misunderstand the mathematical power randomization confers on inferences. Skeptics forget that Fisher required randomization for SSR to ensure that,

> any observation was interchangeable with any other in the analytic expressions … [and] when the null hypothesis was true … the variance calculated for the

> group means was on the average identical with that within the [experimental/control] groups… Thus randomized experiments could be analyzed as if the observations were roughly normally distributed and independent. (Box 1978, pp. 148–149; quoted in Hall 2007, p. 311)

Indeed, experiments show repeated randomization generates z or F distributions (Mayo 1987, p. 593). Skeptics also forget the randomization rationale of Papineau (1994): When the probability that effect E is greater, given F, than not given F—$p(E/F) > p(E/\sim F)$—then either (1) F causes E, or (2) F is correlated with one or more other factors that cause E. To eliminate (2), the data showing $p(E/F) > p(E/\sim F)$ must be randomized, to try to ensure that (a) subjects have equal probability of being in either an experimental or control group and (b) confounders have equal probability of being in either group. Without ensuring (a) and (b) through randomization there is less reason to believe SSR yields causal connections. Although hidden differences in experimental/control groups are always possible, and re-randomizing might avoid them, randomizing is better than not-randomizing because—by severing many existing causal/confounder connections—randomization helps reveal connections being tested.

A second problem with skeptics' "warrant" objections is their forgetting that alternatives to randomization require reliance on subjective assumptions about soil heterogeneity, fertility, etc. Given different assumptions, standard error could be calculated in different ways. Because not all calculations are correct, randomized-experimental design (viz., blocking, i.e., grouping material into homogeneous subgroups of known confounders; then covariance, i.e., analytical removal of influences that cannot be randomized out) helps avoid many flawed assumptions (Fisher 1925, pp. 224–225; Hall 2007, p. 314; see Mayo 1987). Skeptics, however, fail to show how alternatives to randomization avoid unknown subjective assumptions.

A third problem with skeptics' "warrant" objections is their attacking a straw man. Most proponents agree—randomization is insufficient for causal validity—and instead argue that because randomizing is better than not randomizing, SSR ought to be used only under these better conditions. Obviously almost nothing is sufficient to justify causal inferences, or Hume would have no problems with induction. Because randomization is not sufficient, however, does not mean randomization is not necessary for avoiding many causal-inference problems. Moreover, contrary to Worrall (2007), just because one cannot quantitatively estimate randomization's benefits does not mean they are negligible. People should not ignore whatever cannot be quantified.

A fourth problem with skeptics' "warrant" objections is that they set the bar too high. They admit randomization guarantees internal validity (Worrall 2007; Cartwright and Munro 2010), but reject it for not guaranteeing external validity. Yet almost nothing guarantees external validity—transferability of results. Indeed, randomization critics themselves provide no externally guaranteed method.

Randomization Helps Avoid Selection Bias

Proponents say random allocation to experimental/control groups helps avoid selection bias, promotes impartiality, and thus improves causal inferences (Fisher 1947, p. 19; Byar et al. 1976; Papineau 1994). Because randomization helps ensure that subjects have the same probability of being in experimental/control groups, causal results more likely arise from experimenter-imposed differences, not hidden confounders (Fisher 1947, p. 19; Papineau 1994). This equal probability helps provide representative, homogeneous samples—even for poorly understood populations—something non-randomized methods cannot do. Because homogeneity relationships make statistical inferences "work," either randomization or knowing all homogeneity relationships is required. Why? Trying to define non-experimental populations more precisely, with descriptors like "diabetic," Framingham studies illustrate nontransferable conclusions (Empana et al. 2003). But trying to ensure transferability—by avoiding narrow-parent-population definitions—fails to ensure representativeness. Because non-experimental studies can have transferability or representativeness, not both, they ought to avoid SSR.

In response to the preceding arguments, philosophers of science tend to agree. Yet, they say selection bias (in observational SSR studies) is likely small, and can "be eliminated by means" other than randomization (Kadane and Seidenfeld 1990; Worrall 2007, p. 458). However, both responses err. Private-interest science contradicts claims about small selection biases, e.g., studies showing pharmaceutical industry bias in clinical-drug trials (Krimsky 2003). Similarly, claims about alternatives to eliminate selection bias appear naïve. Urbach (1985, p. 265), for instance, proposes Knut Vik squares, saying it is "a mystery" why Fisher "should have objected" to this alternative to randomization, where plants in a field are "separated from one of like variety by a knight's chess move." Yet statisticians like Tedin showed Urbach wrong. Knut Vik methods cause experimental-error overestimation; other systems cause underestimation, and randomization causes neither (Mayo 1987, p. 594). Because no randomization skeptics appear to have provided alternatives that are precise, impersonal, and more successful than randomization, it seems needed to help warrant SSR.

Randomization Helps Avoid Confounders

A third proffered benefit of randomization is helping shield studied effects from confounders, because experimental/control groups are matched for known confounders, then randomized (Fisher 1947, p. 19; Giere 1979, p. 296; Schwartz et al. 1980, p. 7; Papineau 1994). Randomization helps "break the mechanism" between earlier causes (e.g., healthy immune systems) and effects (e.g., recovered health), so new cause-effect relationships (e.g., cures from pharmaceuticals) can be assessed (Pearl 2000, p. 348; Woodward 2006, pp. 56–57). Without randomization, it is less possible to know whether test statistics reflect bias/confounders, or the investigated effect (Wing 2003, p. 1815).

Skeptics, however, say randomization is not needed to avoid confounders (e.g., Urbach 1985; Kadane and Seidenfeld 1990; Howson and Urbach 1993; Worrall 2007; Cartwright and Munro 2010; see Eberhardt and Scheines 2007), because randomization is merely a "fallible mechanical procedure" (Cartwright 2007a, p. 19). However, this response attacks a straw man. Randomization is neither perfect nor always needed, as some early theorists claimed (Gore 1981; Tukey 1977, p. 684). Rather, without randomization, SSR is less able to avoid confounders.

Instead of randomization, skeptics propose alternatives such as (1) avoiding poor observational studies (Worrall 2007, p. 58; Benson and Hartz 2000, p. 1878; Concato et al. 2000, p. 1887; Cartwright and Munro 2010); (2) matching members of experimental/control groups (Worrall 2002, 2007); (3) using "historical controls" and expert-opinion matching (Howson and Urbach 1993, pp. 378–379; see pp. 279–280), and (4) developing antecedent causal knowledge—so information about capacities is the "conduit" warranting causal inferences (Cartwright and Munro 2010).

Unfortunately, none of these alternatives (1–4) appears as good as randomization. Matching known traits (1–3) does not help avoid unknown confounders/selection bias, but randomization does, ensuring that (a) subjects have equal probability of being in experimental/control groups; (b) possible confounders have this equal probability; and (c) after repeated randomizations, (a) and (b) are ensured. Regarding (4), by definition, biostatistics is used when causal/physical information is inadequate. Thus (4) begs the question. Moreover, all alternatives to randomization face subjective/systematic assumptions (see above); changes in natural history, historical populations, experimental populations; inaccurate, incomplete historical data; and selection/researcher bias—as revealed by invalid historical/observational "trials," e.g., treating cancer with interferon, ulcers with stomach freezing, or myocardial infarction with hydrocortisone (e.g., Grage and Zelen 1982, p. 37).

New Arguments for Randomization with SSR in Biology

Besides classic rationales, at least five new arguments suggest SSR requires randomization. The first begins with randomization-skeptics' admissions that "randomization can do no epistemological harm," assuming known causal factors are balanced between experimental/control groups (Worrall 2007, pp. 484–485). If so, the prevent-the-greater-harm argument supports randomization, because—unlike the tobacco, alcohol, hormone, and Amazon research—randomization avoids SSR observational studies' misleading guarantees of causal reliability.

A second new argument—the "appearance-of-bias argument"—is that randomization tells others that systemic research bias is less likely to have occurred. Even Bayesians say randomization helps convince others that research is not "rigged" (e.g., Kadane and Seidenfeld 1990).

Third, randomizing with SSR helps thwart special-interest "bending science" (McGarity and Wagner 2008). Top journal editors warn against sponsor censoring of science (Davidoff et al. 2001, p. 826). Some require that "for industry-sponsored studies" data analysis must be done by an "independent statistician at an academic institution … not employed by the sponsor" (JAMA, Editorial Policies for Authors—JAMA, p. 4). Given such requirements, skeptics—who propose randomization alternatives—appear naive about science funding/control. Special interests fund 75 % of science (Koizumi 2005). Frequently they misuse biology (Shrader-Frechette 2004). They can misuse randomized research (Lexchin et al. 2003; Carpenter 2010), but doing so obviously is more difficult than misusing non-randomized research.

A fourth reason to use SSR with randomizing is that it promotes transparency. It helps standardize biological research, ensure using similar methods for similar studies, and facilitate research cross-comparisons. Otherwise, randomization alternatives—systematic experimental designs—would vary among researchers and reduce transparency (Hall 2007, p. 214).

A fifth new argument is that randomization helps democratize biological experimentation, so non-geniuses can do quality research. Randomization provides reliable ways for average biologists to take account of unavoidable population, sample, plot, etc. differences. As Fisher notes, Darwin's nonrandomized botanical experiments "worked" only because his brilliance, careful matching, and controlling for confounders avoided "the sole source" of experimental error—unknown heterogeneities, like "differences in soil fertility" (Fisher 1947, pp. 47–48; quoted in Hall 2007, pp. 315–316). If Darwin had randomized, however, Fisher says he would have avoided "the anxiety

of considering … the innumerable causes" that could disturb his data (1947, p. 9). Besides, because biological data vary more than those in physics/chemistry, randomizing with SSR seems especially important in biology—a point missed by randomization skeptics. They sometimes suggest biological heterogeneities are no greater than in physics/chemistry (e.g., Urbach 1985; Mayo 1987).

Where We Go from Here

To encourage scientists to use SSR under conditions (like randomization) that help warrant causal inferences, perhaps journals could encourage researchers who use SSR for non-randomized data (e.g., Banebrake et al. 2010; González et al. 2011; Knutsen et al. 2011) to support their causal inferences by additional means. Additional causal-warrant alternatives include replication (e.g., Nickerson 2000), weight-of-evidence arguments (MacNeil 2008), and inferences to the best explanation (Shrader-Frechette 2011, chap. 4).

Our lessons about causal inferences? Techniques—like using SSR with observational data—-cannot substitute for good scientific analysis. "Powerful tools [like SSR] must be used carefully" (Davidoff et al. 2001).

Acknowledgments The author thanks the US National Science Foundation (NSF) for the History and Philosophy of Science research grant, "Three Methodological Rules in Risk Assessment," 2007-2009, through which work on this project was done. All opinions and errors are those of the author, not the US NSF.

References

Alcohol Beverage Medical Research Foundation (ABMRF) (2011) http://www.abmrf.org/. Accessed 24 Jan 2011

Anderson G (2001) Genomic instability in cancer. Curr Sci 81(5): 501–550

Banebrake TC, Christensen J, Boggs CL, Ehrlich PR (2010) Population decline assessment, historical baselines, and conservation. Conserv Lett 3(6):371–378

Bath PM, Gray LJ (2005) Association between hormone replacement therapy and subsequent stroke: a meta-analysis. Br Med J 330(7487):342

Beauchamp T, Childless J (1989) Principles of biomedical ethics. Oxford University Press, New York

Benson K, Hartz AJ (2000) A comparison of observational studies and randomized controlled trials. N Engl J Med 342:1878

Box JF (1978) R. A. Fisher: the life of a scientist. Wiley, New York

Byar DP, Simon RM, Friedewald WT, Schlesselman JJ, DeMets DL, Ellenberg JH, Gail MH, Ware JH (1976) Randomized clinical trials—perspectives on some recent ideas. N Engl J Med 295:74–80

Carpenter D (2010) Reputation and power. Princeton University Press, Princeton

Cartwright N (1994) Nature's capacities and their measurement. Clarendon Press, Oxford

Cartwright N (2007a) Are RCTs the gold standard? Biosocieties 2(1):11–20

Cartwright N (2007b) Hunting causes and using them. Cambridge University Press, Cambridge

Cartwright N, Munro E (2010) The limitations of randomized controlled trials in predicting effectiveness. J Eval Clin Pract 16(2):260–266

Collins MA, Neafsey EJ, Mukamal KJ, Gray MO, Parks DA, Das DK, Korthuis RJ (2009) Alcohol in moderation, cardioprotection, and neuroprotection: epidemiological considerations and mechanistic studies. Alcohol Clin Exp Res 33:1–14

Concato J, Shah N, Horwitz RI (2000) Randomized controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med 342:1887

Cranor C (2006) Toxic torts. Cambridge University Press, New York, pp 240–241

Davidoff F, DeAngelis CD, Drazen JM, Hoey J, Højgaard L, Horton R, Kotzin S, Nicholls MG, Nylenna M, Overbeke AJPM, Sox HC, Van Der Weyden MB, Wilkes MS (2001) Sponsorship, authorship, and accountability. N Engl J Med 345:825–826

Dietl G, Durham S, Kelley P (2010) Shell repair as a reliable indicator of bivalve predation by shell-wedging gastropods in the fossil record. Palaeogeogr Palaeoclimatol Palaeoecol 296(1/2):174–184

Eberhardt F, Scheines R (2007) Interventions and causal inference. Philos Sci 74(5):981–995

Ellison RC (2005) Importance of pattern of alcohol consumption. Circulation 112:3818–3819

Empana JP, Ducimetière P, Arveiler D, Ferrières J, Evans A, Ruidavets JB, Haas B, Yarnell J, Bingham A, Amouyel P, Dallongeville J, PRIME Study Group (2003) Are the Framingham and PROCAM coronary heart disease risk functions applicable to different European populations? The PRIME Study. Eur Heart J 24:1903–1911

Ferrara A, Quesenberry CP, Karter AJ, Njoroge CW, Jacobson AS, Selby JV, Northern California Kaiser Permanente Diabetes Registry (2003) Current use of unopposed estrogen and estrogen plus progestin and the risk of acute myocardial infarction among women with diabetes: the Northern California Kaiser Permanente Diabetes Registry, 1995–1998. Circulation 107:43–48

Fidler F, Burgman M, Cumming G, Buttrose R, Thomason N (2006) Impact of criticism of null-hypothesis significance testing on statistical reporting practices in conservation biology. Conserv Biol 20(5):1539–1544

Fillit HM (2002) The role of hormone replacement therapy in the prevention of alzheimer disease. Arch Intern Med 162: 1934–1942

Fillmore KM (2000) Is alcohol really good for the heart? Addiction 95(2):173–174

Fillmore KM and Kerr W (2002) A Bostrom abstinence from alcohol and mortality risk in prospective studies. Nordic Stud Alcohol Drugs 19(4):295–296

Fisher RA (1925) Statistical methods for research workers. Oliver and Boyd, Edinburgh

Fisher RA (1947) The design of experiments. Oliver and Boyd, Edinburgh

Foot P (1978) Virtues and vices and other essays in moral philosophy. University of California Press, Berkeley; Blackwell, Oxford

Garamszegi LZ, Calhil S, Dochtermann N, Hegyi G, Hurd PL, Jorgensen C, Kutsukake N, Lajeunesse MJ, Pollard KA, Schielzeth H, Symonds MRE, Nakagawa S (2009) Changing philosophies and tools for statistical inferences in behavioral ecology. Behav Ecol 20(6):1363–1375

Gerrodette T (2011) Inference without significance. Marine Ecol 32:404–418

Giere RN (1979) Understanding scientific reasoning. Holt Rinehart and Winston, New York, p 296
González AL, Fariña JM, Kay AD, Pinto R, Marquet PA (2011) Exploring patterns and mechanisms of interspecific and intraspecific variation in body elemental composition of desert consumers. Oikos 120:1247–1255
Gore SM (1981) Assessing clinical trials: why randomize? Br Med J 282:1958
Grage TB, Zelen M (1982) The controlled randomized trial in the evaluation of cancer treatment. UICC Tech Rep Ser 70:23–47
Greenland S (1990) Randomization, statistics, and causal inference. Epidemiology 1(6):421–429
Greenland S (1977) Response and follow-up bias in cohort studies. Am J Epidemiol 106:184
Grodstein F, Manson JE, Colditz GA, Willett WC, Speizer FE, Stampfer MJ (2000) A prospective, observational study of postmenopausal hormone therapy and primary prevention of cardiovascular disease. Ann Intern Med 133:933–941
Hacking I (1990) The taming of chance. Cambridge University Press, New York
Hall NS (2007) R. A. Fisher and his advocacy of randomization. J Hist Biol 40:295–325
Hart M, Marko P (2010) It's about time: divergence, demography, and the evolution of developmental modes in marine invertebrates. Integr Comp Biol 50(4):643
Howson C, Urbach PM (1993) Scientific reasoning: the Bayesian approach. Open Court, Peru
Hurtig AK, San Sebastián M (2005) Epidemiology vs epidemiology: the case of oil exploitation in the Amazon basin of Ecuador. Int J Epidemiol 34(5):1170–1172
International Agency for Research on Cancer (IARC) (2008) Monograph: overall evaluations of carcinogenicity to humans, Supplement 7, 2008. http://monographs.iarc.fr/ENG/Monographs/suppl7/Suppl7-5.pdf. Accessed 27 Aug 2008
Kadane JB, Seidenfeld T (1990) Randomization in a Bayesian perspective. J Stat Plan Inference 25:329–345
Kelsh M, Morimoto L, Lau E (2008) Cancer mortality and oil production in the Amazon region of Ecuador. Int Arch Occup Environ Med 82(3):381–395
Klatsky AL (1996) Alcohol and hypertension. Clin Chim Acta 246:91–105
Knutsen H, Olsen EM, Jorde PE, Espeland SH, André C, Stenseth NC (2011) Are low but statistically significant levels of genetic differentiation in marine fishes "biologically meaningful"? A case study of coastal Atlantic cod. Mol Ecol 20(4):768–783
Koizumi K (2005) R&D trends and special analyses, AAAS Report XXIX, XXVII. AAAS, Washington
Krimsky S (2003) Science in the private interest. Rowman and Littlefield, Lanham
La Caze A, Djulbegovic B, Senn S (2011) What does randomization achieve? Evidence Based Med. doi:10.1136/ebm.2011.100061
Legendre P, Legendre L (2011) Numerical ecology. Elsevier, Amsterdam
Lexchin J, Bero L, Djulbegovic B, Clark O (2003) Pharmaceutical industry sponsorship and research outcome. Br Med J 326(7400): 1167–1170
MacNeil MA (2008) Making empirical progress in observational ecology. Environ Conserv 35(3):193–196
Mariani S (2008) Through the explanatory process in natural history and ecology. Hist Philos Life Sci 30(2):159–178
Mayo O (1987) Comments on "Randomization and the design of experiments" by P. Urbach. Philosophy of Science 54(4):592–596
McConell MV, Vavouranakis I, Wu LL, Vaughan DE, Ridker PM (1997) Effects of a single, daily alcoholic beverage on lipid and hemostatic markers of cardiovascular risk. Am J Cardiol 80:1226–1228
McGarity T, Wagner W (2008) Bending science. Harvard University Press, Cambridge
Mehner T, Freyhof J, Reichard M (2011) Summary and perspective on evolutionary ecology of fishes. Evol Ecol 25(3):547–556
Naessen T, Lindmark B, Lagerström C, Larsen HC, Persson I (2007) Early postmenopausal hormone therapy improves postural balance. Menopause 14:14–19
National Heart, Lung, and Blood Institute (NHLBI) (2009) Framingham Heart Study. NIH, Bethesda
Nickerson RS (2000) Null hypothesis significance tests: a review of an old and continuing controversy. Psychol Methods 5(2): 241–301
Okland RH (2007) Wise use of statistical tools in ecological field studies. Folia Geobot 42:123–140
O'Rourke D, Connolly S (2003) Just oil? The distribution of environmental and social impacts of oil production and consumption. Annu Rev Environ Resour 28:587–617
Papineau D (1994) The virtues of randomization. Br J Philos Sci 45:437–450
Pearl J (2000) Causality. Cambridge University Press, New York
Peirce CS, Jastrow J (1885) On small differences in sensation. Mem Natl Acad Sci 3:73–83 (Reprinted in Burks AW (ed) (1958) Collected papers of Charles Sanders Peirce, vol 7. Harvard University Press, Cambridge, pp 13–34
Peters RH (1991) A critique of ecology. Cambridge University Press, Cambridge
Poole C, Rothman KJ (1998) Our conscientious objection to the epidemiology wars. J Epidemiol Community Health 52:612–618
Rimm E, Klatsky AL, Grobbee D, Stampfer MJ (1996) Review of moderate alcohol consumption and reduced risk of coronary heart disease: is the effect due to beer, wine, or spirits? Br Med J 312:731–736
Roizen R and Fillmore KM (2000) The coming crisis in alcohol social science, nordic studies on alcohol and drugs (English Supplement) 17:91–104
Rothman KJ (1990) Statistics in non-randomized studies. Epidemiology 1:417–418
Roughgarden J (2009) Is there a general theory of community ecology. Biol Philos 24(4):521–529
Ruitenberg A, van Swieten J, Witteman J, Mehta K, van Duijn C, Hofman A, Breteler MMB (2002) Alcohol consumption and risk of dementia: the Rotterdam study. Lancet 359(9303):281
Savitz DA (2003) Interpreting epidemiologic evidence. Oxford University Press, New York
Schwartz D, Flamant R, Lellouch J (1980) Clinical trials. Academic Press, London
Sherwin B (2000) Mild cognitive impairment: potential pharmacological treatment options. J Am Geriatr Soc 48(4):431–441
Shrader-Frechette K (2004) Measurement problems and Florida panther models. Southeast Nat 3(1):37–50
Shrader-Frechette K (2008a) Statistical significance in biology: neither necessary nor sufficient for hypothesis-acceptance. Biol Theory 3(1):12–16
Shrader-Frechette K (2008b) Evidentiary standards and animal data. Environ Justice 1(3):139–144
Shrader-Frechette K (2011) Fighting climate change with renewable energy, not nuclear power. Oxford University Press, New York
Shrader-Frechette K, McCoy ED (1993) Method in ecology. Cambridge University Press, Cambridge
Sillero N (2011) What does ecological modelling model? A proposed classification of ecological niche models based on their underlying method. Ecol Model 222(8):1343–1346
Simberloff D (1976) Species turnover and equilibrium island biogeography. Science 194:572–578
Simberloff D, Wilson EO (1970) Experimental zoogeography of islands. Ecology 51:934–937

Singer N (2009) Medical papers by ghostwriters pushed therapy. New York Times. 5 Aug 2009; Section A, p 1

Sober E (1988) The principle of the common cause. In: Fetzer J (ed) Probability and causality: essays in honor of W. C. Salmon. Reidel, Boston, pp 211–228

Strom BL, Schinnar R, Weber AL et al (2006) Case–control study of postmenopausal hormone replacement therapy and endometrial cancer. Am J Epidemiol 164:775–786. doi:10.1093/aje/kwj316

Tannen RL, Weiner MG, Xie D, Barnhart K (2007) Estrogen affects postmenopausal women differently than estrogen plus progestin replacement therapy. Hum Reprod 22:1769–1777

Tukey JW (1977) Some thoughts on clinical trials. Science 198:684

University of California, San Francisco (UCSF) (2011) Drug industry document archive (Dida). http://dida.library.ucsf.edu. Accessed 24 Jan 2011

Urbach PM (1985) Randomization and the design of experiments. Philos Sci 52:256–273

Wing S (2003) Objectivity and ethics and environmental health science. Environ Health Perspect 111(14):1809–1818

Woodward J (2006) Invariance, explanation, and understanding. Metascience 15:56–57

Worrall J (2002) What evidence in evidence-base medicine? Philos Sci 69(S3):S316–S330

Worrall J (2007) Why there's no cause to randomize. Br J Philos Sci 58:453–454