The Current Status of Observational Studies as Scientific Evidence: A Critical Appraisal.

Steven Bratman, MD, MPH

Table of Contents

Abstract

Purely observational studies have long been considered problematic as a source of cause/effect conclusions and, for this reason, have been placed below experimental studies in the hierarchy of evidence. However, in many important areas of medicine and public health, experimental studies (specifically randomized controlled trials) are impractical to perform. In consequence, observational studies are often the primary source of public health recommendations. This practice has recently been called into question by the results of several large randomized controlled trials, which not only failed to confirm predictions based on observational studies but in some cases inverted them, transforming expectation of benefit into a discovery of harm. In this article, I interrogate the use of epidemiologic evidence as a source of "evidence-based" public health advice.

Background

In the late 1990s, while evaluating and synthesizing evidence-based information on alternative medicine, my collaborators and I came to view purely epidemiologic[a] research with great skepticism. This attitude derived in part from the history of natural supplements. A vast array of such supplements had come into popular use based on inferences drawn from large population studies but when tested in randomized controlled trials (RCTs) most proved ineffective. The subsequent findings of the Women's Health Initiative regarding hormone replacement therapy further strengthened our impression that the results of large population studies cannot be relied upon as medical evidence for the efficacy of a treatment. Therefore, in the electronic database we created, supplement recommendations grounded *only* in observational studies were classified as "Category C: Lacking Any Reliable Supporting Evidence." The only category occupying a lower position was "Category D: Meaningful Evidence Against Efficacy."

However, when I entered a Master of Public Health program in 2008, I found that observational study results were routinely used as sufficient evidence for issuing prescriptions to society at large. Initiatives such as Healthy People 2010 and the Dietary Guidelines for Americans, for example, consist in large part of recommendations based on observational studies alone. I found this startling and disturbing; it seemed that much of official public health policy was based on the type of bad science used to justify ineffective alternative medicine treatments.

Nonetheless, the fact that such studies are taken so seriously in the field of public health suggested that perhaps they had more scientific value than I had previously assumed. I, therefore, began to interrogate the issue at a deeper level. This paper is the result.

[a] When it comes to human disease and health, observational studies are generally called "epidemiological" studies. In this paper, we shall use the terms use the terms somewhat interchangeably. The related term "population study" refers to a *large* observational study.

Introduction: From John Stuart Mill to the Women's Health Initiative

The superiority of experiment over mere observation was established in John Stuart Mill's 1843 work, *System of Logic Ratiocinative and Inductive*. The modern randomized controlled trial (RCT), a form of experiment, is grounded in Mill's reasoning, as refined by the work of the statistician RA Fisher and the epidemiologist Bradford Hill. Fisher used mathematical arguments to show that randomization is necessary to allow valid calculations regarding statistical significance. Bradford Hill, on the other hand, emphasized the value of random assignment to eliminate systematic selection bias.[1]

Bradford Hill is much better known, however for his work in observational studies, where the canonical "Bradford Hill criteria" are widely utilized to draw causal conclusions in non-experimental settings. The Framingham Heart study was one of the first of the large population studies to which these principles were systematically applied. Interpretation of the Framingham findings led to an indictment of cigarettes and high blood pressure that have subsequently been accepted as correct. Many other recommendations came out of Framingham as well, such as that it is important to avoid eating eggs, to reduce salt intake and to avoid saturated fat. These conclusions, however, have proved less robust. Beginning in the 1990s, writer Gary Taubes published a series of award winning articles in the journal *Science* skewering the evidence used to support these recommendations.[2,3,4] The egg recommendation has by now fallen; that of salt and saturated fat are in the process of being quietly withdrawn.

At the same time, women's health groups such as the Women's Health Collective had begun to demand randomized controlled trials of hormone replacement therapy (HRT).[5] In those days, HRT was widely prescribed to healthy women on the theory that it reduces risk of cardiovascular disease. However, while the FDA ordinarily requires positive results in several large randomized, double-blind, placebo controlled trials prior to drug approval, HRT had been approved for cardiovascular disease prevention based on observational studies alone. It would seem, members of the collective pointed out, that a drug given to healthy people properly requires a higher rather than a lower level of evidence than a drug given to people with clear disease. These complaints finally led to the initiation of several large RCTs. In 1998, the results of the first of these became available. Although the data from the Heart and Estrogen/Progestin Replacement Study (HERS) indicated that HRT does not work, influential epidemiologists continued to defend the use of HRT on the basis of observational evidence alone.

Then came the results of the Women's Health Initiative. This mega-RCT showed that epidemiologists had not only gotten it wrong, they had gotten it exactly backwards. Rather than protecting women from cardiovascular disease, it turned out that use of HRT *increases* cardiovascular risk. The WHI at last provided a shock sufficient to provoke a vigorous debate on the value of evidence drawn from observational studies.

This field is still in flux. The current state of this dialogue is the subject of this paper.

Analysis

The purpose statement, "Evaluate the current status of observational/epidemiological studies as scientific evidence," encodes several concepts. First, it should be noted that "evidence" is more than data; it is data used to form inferences. The inferences of interest regard cause/effect relationships used to justify public health recommendations.

The term "status" implies placement in a hierarchy as well as the closely related issue of entire exclusion from that hierarchy: Aristocrats may sit at table according to their rank, while commoners cannot sit there at all. In the field under discussion, this concerns the relative position of observational studies vs. other forms of scientific evidence, together with the risk that observational studies might lose their place as scientific evidence entirely.

This unpacking in turn reveals another implicit assumption: "Science" is a uniquely privileged source of information about the physical world. It is this privilege that elevates "scientifically valid" types of evidence above others, and that gives significance to conclusions regarding the presence or absence of such status. In particular, if it were to be found that observational studies are not valid as scientific evidence, this would demote their standing and reduce their influence significantly.

In order to analyze the validity of scientific observational studies, we must first interrogate the privileged nature of science itself.

The Epistemic Prestige of Science

In March 2008, the EPA released revised "final" regulations on acceptable ozone levels, based almost entirely on the results of observational studies. Within weeks the American Lung Association, the Natural Resources Defense Council, the Environmental Defense Fund and the Sierra Club filed suit. Their claim: The Bush administration allowed business interests to override the opinion of "the EPAs own scientists."[6]

Leaving aside the merits of the claim, it clearly has some rhetorical force. A simple thought experiment shows that this force derives from a shared social judgment regarding scientists.

Consider a set of alternate phrases, as follows.

Business interests were allowed to override the opinions of:

- The EPA's own postmodernists.
- The EPA's own lawyers.
- The EPA's own economists.
- The EPA's own historians.
- The EPA's own oracle.
- The EPA's own prophets.
- The EPA's own priests.

The specific weaknesses of each of these alternate complaints illustrates and also illuminates in detail the privileged status of science. For example, "The EPA's own postmodernists" lacks any force at all; postmodernists, no matter how influential in intellectual circles, possess no authority whatever when it comes to making policy. Science, in contrast, is assumed to significantly bear on policy.

"The EPA's own lawyers" might be meaningful if the subject were law, as in "the CIA's own lawyers objected to waterboarding." Here the called-upon expertise relates to an aspect of human thought and opinion while the original phrase references something seen as a fact: Ozone at certain levels harms health. Science, in other words, has standing in extra-human reality where it is supposed to signify "what is" rather than "what people think."

Similarly evocative limitations would apply to "the EPA's own economists." These might be invoked to give valuable advice regarding an objective fact: the economic costs of regulations. However, economists are not expected to reach facts themselves, but only to opine on them. Though scientists *may* disagree, economists are *expected* to.

"The EPA's own doctors" has some apparent force, but a close look shows that their opinion is especially convincing only insofar as they speak as scientists; insofar as they speak as Marcus Welby, they lose cache.

In a previous era, "the EPA's own oracles" or their prophets might have spoken with great authority, but such founts of truth have lost ground in the last couple of millennia. While science is sometimes compared to a priesthood, it has plausibly achieved this position by *winning*; like Moses with the Pharaoh's sorcerers, science is perceived to have out-performed the competition. Of course, there remain some areas in which religious authority continues to contend against science with some success in the court of public opinion, the theory of Evolution being the prime example. However, there are no obvious voices, religious or otherwise, that offer an alternative construction of physical reality across more than a sliver of the physical world.

In other words, as these thought experiments indicate, science has no peer competitor in the area of physical reality. Granting that scientists may act on political motives, overstep their field and merely err, science is broadly accepted as possessing the best available representation of the physical world.[b] In other words, it possesses unique epistemic prestige.

But what *is* Science?

Used purely descriptively, the term "science" denotes the consensus ideas of a group of people called "scientists." These individuals can be grouped into categories, or subgenres, such as physicists, chemists, biologists, etc. However, this approach fails to capture the

[b] Even for the most extreme science skeptic, a Churchillian reformulation of the statement would at least pass muster, and serves the same point: Science possesses the least accurate representation of the physical world, except for all the others.

source of the inhering epistemic prestige. Intuitively we recognize that certain methods and means characterize practitioners of science, and it is these methods that cause us to trust what scientists say. When someone called a "scientist" fails to act within the scope of these scientific means and methods, we complain that they are not "behaving scientifically" and withhold respect. In other words, the definition given at the head of this paragraph can be inverted to "scientists are those people who use scientific means and methods." Obviously, these terms now need to be investigated.

There are several entries under "science" in Webster's dictionary.[7] The first is less than useful, but it does have the merit of getting straight to the epistemic claim. "Science: (1) Knowledge as distinguished from ignorance or misunderstanding." The problem with this definition is that it would include true knowledge regarding the Buddha nature or of the meaning of life, and such questions are currently the domain of philosophers and theologians but not scientists.

The obvious refinement here is to add a limiter, as Webster's does in a subsequent entry: "Science is concerned with the physical world and its phenomena." Yet more limitations are necessary to approach the common understanding of "science." For example, one may have definite knowledge regarding the number of apples in one's refrigerator and not, by virtue of this, qualify as a scientist. Science, Webster's helpfully adds, is "a department of systematized knowledge." This reflects the fact that science, unlike ordinary refrigerator knowledge, includes an element of systematic study. However, the same is true of stamp collecting, and while some aspects of science are fundamentally similar to stamp collecting (gathering specimens of beetles), there is often a more ambitious agenda as well (Evolution.) Science, one should properly add, generally seeks to investigate fundamental features of the world rather than minor, incidental or highly contingent ones.

This latter feature is captured in another Webster entry, in which it is stated that science seeks to discover "general truths or the operation of general laws." Such "truths" and "laws" regard regularities or patterns existing observed in objects and events. They range from simple empiric observations, such as that the ratio of the circumference of a circle to its diameter is a little greater than three, to vast theories such as General Relativity.[c]

Finally, Webster's notes that science derives these laws and truths by means of "scientific method." This is not a mere tautology because the term "scientific method" has definite content.

Scientific Method

[c] That the ratio of the circumference of a circle to its diameter equals pi may seem to be a question of mathematics rather than of science. However, the ratio was first discovered empirically, and only much later derived mathematically. As an aside, close measurement will show that the ratio in question is *not* actually pi, due to the curvature of space predicted by General Relativity. See the Appendix.

The conceptual basis of "the scientific method" was developed over centuries by such people as Francis Bacon, Galileo Galilei and John Stuart Mill. In the form commonly taught in high school, it consists of four parts: (a) Collecting data through systematic, objective observation. (b) Creating hypotheses from this data. (c) Performing experiments to test the hypotheses. (d) Refining or changing the hypothesis based on the results of experiment.

However, as Paul Feyerabend famously demonstrated in his brilliant work *Against Method*,[8] these steps fail to fully capture science as it is actually practiced. In particular, aesthetic considerations, intuitive leaps, and all other forms of human creative behavior may play a role in hypothesis formation.[d] This is especially the case for the grander type of scientific hypothesis we call a "theory."[e] Nonetheless, no matter how they are initially derived, all scientific theories must ultimately be *tested* via experiment.

What is a Scientific Experiment?

The term "scientific experiment" implies a set of distinct but related concepts.

In an experiment, the experimenter sets up an artificial situation designed to reduce extraneous variables in order to support or falsify a hypothesis. Clarity is essential; experiments whose outcomes are blurry or foggy regarding the hypothesis to be tested are not *scientific* experiments in the formal sense.[f] For example, if one wishes to test the hypothesis that a certain incantation can bring about rain, one must establish in advance criteria for determining whether the hypothesis has succeeded or failed; one may scientifically test the claim, "it will rain at least one-quarter inch within twenty-four hours," but the alternate proposition "it will rain eventually, to some extent, probably" is too vague for scientific testing. Additionally, one is not allowed to improvise post-hoc (after the fact) excuses such as

---

[d] A variation on this theme applies to mathematics itself: New areas of mathematics are not generally derived by deduction. Rather, they often involve a type of "experiment," in which one notices certain patterns while "fiddling around" and then looks for the underlying rule. Other creative thought processes are used as well. However, even though theorems are frequently conceived non-deductively, they must ultimately be proved in a rigorous, deductive manner. For a discussion of this, see Baker A. Non-Deductive Methods in Mathematics. *Stanford Encyclopedia of Philosophy*. Available at: http://plato.stanford.edu/entries/mathematics-nondeductive. Accessed 11/1/09

[e] To quote the Encyclopedia Britannica, which waxes eloquent on the subject: "[Scientific] theories are imaginative constructions of the human mind – the results of philosophical and aesthetic judgment as well as of observation – for they are only suggested by observational information rather than inductively generalized from it … Thus, whereas an empirical law expresses a unifying relationship among a small selection of observables, scientific theories have much greater scope, explaining a variety of such laws and predicting others as yet undiscovered." See: Scientific Theory. *Encyclopædia Britannica Online*. Available at: http://www.britannica.com/EBchecked/topic/528971/scientific-theory. Accessed 11/1/09

[f] The term "experiment" is also used to indicate "experimenting around," a type of free exploration to which none of these rules need apply. However, the process of experimenting around is intended to eventually create a hypothesis that can be tested in a "real" experiment.

"it would have rained if that inauspicious crow hadn't flown by." If one wishes to account for crows, one needs to do so in the fully framed hypothesis; e.g., "it will rain at least one-quarter inch within twenty-four hours so long as there are no crows." Hypotheses couched in blurry terms susceptible to post-hoc auxiliary excuses are not capable of scientific testing.

Scientific experiments also must be repeatable. If the rain incantation is tried only once or twice, it cannot be said to have passed scientific testing; it must work many times. Furthermore, experimental methods must be transparent so that other scientists can attempt to reproduce or contradict the claimed the results. (If a group of sorcerers wished to achieve scientific validation for a secret incantation, they would have to submit *themselves* for testing.)

Finally, scientific experiments must be framed for maximum objectivity: Bias is to be sought out and eliminated, and subjective evaluation is to be rejected in favor of measurement and quantification. A small excursion is warranted for so loaded a term.

Scientific Objectivity

Webster defines "objective" as, "relating to, or being an object, phenomenon, or condition in the realm of sensible experience independent of individual thought and perceptible by all observers: having reality independent of the mind."[9] However, this is rather too high a version of the term for practical use. Not only does it involve insoluble variations on the theme of philosophical realism,[10] it flies in the face of an essential postmodern insight: that all human activity depends on contingent human facts.

Nonetheless, as a practical matter, it is possible to decrease the human contingent, and the concept of "scientific objectivity" implies full use of such methods. Perhaps the most important involves a kind of self-imposed limitation: normative scientific approaches limit their scope of allowable observations to those areas in which it seems logical that there is a heightened ability to remove the contingent. For example, we expect to achieve a higher level of objectivity when observations can be put into discrete and mutually exclusive categories than when they require description; e.g, a task involving counting the number of stones on the floor is more likely to be achieved in an objective manner than one that involves assigning adjective such as "large," "smooth" and "shapely." In general, counting, or otherwise quantifying by number, is more likely to lead to results that are widely agreed upon by a variety of observers than other methods of observation.[g,h]

---

[g] There are still issues. For example, when counting the number of species of plants per square foot, it is quite easy to unconsciously skew the results. The problem lies in a very literal border issue: how does one classify plants on the very edge of the set square? If great care isn't taken to address this systematically, the outcome of the measurement can be biased to a remarkable extent.

[h] Of course, choice of which numbers to look at, as well as how to interpret them, gives great leeway for processes that aren't objective in any sense of the word. This is obviously the case when emotional hot buttons are involved, such as (for example) sexual orientation. Regardless of emotional charge, when results must be interpreted via statistical analysis the well known problem captured in the technical phrase, "lies, damn lies and statistics" remains an issue.

Another essential limiting factor involves emotion or, rather, its absence. It can be stated without controversy that observations are more likely to have an objective character when they are conducted in areas of emotional neutrality. Thus, one may be expected to more neutrally and, hence, objectively count the number of stones in a box than the size of a crowd at a political demonstration. Additionally, in circumstances where emotional neutrality is unlikely, the well-known method of "blinding" is used to help reduce systematic bias. Thus, a crowd-counter unaware of the purpose of the gathering is more trustworthy than one who knows the political intent and may fervently agree or disagree.

Finally, the elements of clarity and transparency discussed earlier also relate to the scientific use of the term objective. Claims surrounded in a blurry cloud of possible excuses are incapable of being tested objectively: Either the rain incantation works as promised, or it does not. If the question of incantation efficacy cannot be phrased in some direct manner, it cannot be examined scientifically.

A Forensic Analogy: The Jury

Using the methods and techniques described above, scientists arrive at conclusions about the physical world. These conclusions are widely trusted, and science is ordinarily believed to discover "facts" (much of the time) rather than merely enumerate opinions. But on what is this reputation based?

Descarte, like Plato, hoped to establish science as a form of absolute truth by grounding it in deductive reasoning alone. However, the pure Cartesian/Platonic program failed, and, while Descarte did contribute significantly to mathematics, he added relatively little to science. Francis Bacon sketched out an entirely different approach to science based on inductive reasoning and careful observation, but he spent most of his life in politics. It was Isaac Newton who actually created the scientific revolution, and he did so by marrying the two methods. Though Newton relied heavily on mathematics, he was also one of history's great experimentalists. He believed that science never achieves the absolute status of mathematics but remains forever provisional, subject to revision, refinement and correction. (As an aside, even mathematical proof lacks "the status of mathematical proof." See the Appendix.)

If science remains provisional, one might fairly wonder how it can possess access to any sort of "truth" at all. Perhaps illumination can be found by comparing science not to mathematics but to the social institution of the jury.

In Anglo-American law, the primary function and unique legal capacity of a jury is to carry out a task called "finding fact." When a jury pronounces its conclusions, "juridical fact" comes into being; a person convicted of murder becomes on the instant "a murderer" in fact, and newspapers no longer need protect themselves from the charge of libel by using the term "suspect" or "accused."

Juries are called upon to determine evidentiary fact and inferential fact. If a phone record shows that the accused phoned from the victim's house ten seconds before shots were fired, the jury may make an evidentiary determination that the phone record is accurate plus an inferential finding that the accused was present when the shots were fired.

Juries are constituted and instructed to function in accordance with various traditions. Among these is the assumption that members of a jury must be sane, mature, and not overly under the influence of an intense, biasing emotion. In other words, they must be capable of rational thought. Furthermore, they are expected to rely primarily on their rational capacities during the process of deliberation, using careful reasoning and sifting of evidence rather than, for example, revelation, mood or impulse. In criminal cases, conclusions must be arrived at "beyond reasonable doubt" while in civil cases the standard is "preponderance of the evidence," but both expressions invoke rational thought. Finally, the requirement that a jury must achieve consensus symbolizes the principle that a successful legal case should be strong enough to convince almost *any* rational person.

Given this description of the functioning of a jury, one can easily identify a primary source of science's epistemic prestige: Science consists centrally of that type of rational argument most persuasive to the broadest possible jury. To adapt a phrase of philosopher of science Susan Haack, scientific fact is continuous with juridical fact, only more so.[i]

This heightened capacity to convince a rational juror is not accidental but is of the essence of science. In general, scientific arguments are assembled in such a manner that the great majority of rational people sufficiently acquainted with the evidence and the inferences would assent to the conclusions. All of the elements of scientific method described above are designed to achieve this. Consider, for example, the habit of systematic, objective, repeatable observation. Scientists, as it were, adjust the lighting, arrange their positions, and get notepads, cameras and tape recorders ready prior to the event. (Experimental scientists do all this and then get to press "repeat.") In comparison, happenstance witnesses at a murder scene are unlikely to be afforded such amenities. Common sense tells us that systematic observations designed in every way to facilitate accurate perception are inherently more reliable than those made on the fly under circumstances dictated by chance. Science further adds to its observational credibility by a deliberate self-limitation of scope; with the telling exception of psychologists, scientists excluded themselves from highly subjective issues like motive and state of mind. Thus, while both scientists and witnesses report observations, those of the former are specifically designed to inspire *universal* assent.

The preceding concerns evidentiary fact. Science also possesses a heightened persuasive ability regarding inferential fact. In general, scientific inferences involve logical arguments that, though short of mathematical proof, are much stronger than those typically available in

[i] What she actually said was, "Scientific inquiry is continuous with everyday empirical inquiry – only more so." This more general statement additionally incorporates such methods of inquiry as academic study and good journalism, all of which utilize the same general principles of sifting evidence and making careful inferences. See Haack S. *Defending Science – Within Reason. Between scientism and cynicism*. Amherst: Prometheus Books. 2003:94

a courtroom. For example, when John Snow discovered that almost everyone who drank flocculent water in London developed cholera and almost no one who drank clear water came down with the disease, he possessed a detective's dream of a case. Similarly, after extensive probing in numerous courtrooms, DNA analysis has established itself as a new type of fact rather than a mere form of expert opinion. In the hard sciences, inferential chains are commonly so persuasive that it is difficult to imagine any rational human being failing to agree with the conclusions if given the time and opportunity to follow the argument. These and other characteristics of scientific investigation result in very strong inferential cases; conversely, it would probably not be controversial to state that if a scientific argument cannot convince the great majority of rational people sufficiently acquainted with the facts, it is not a successful scientific argument.[j]

In addition to the accuracy of its observations and the clarity of its reasoning, science enjoys an additional means of impressing the jury: it can predict eclipses and produce computers, antibiotics and atomic bombs. These are performative demonstrations, like that of the expert witness on the art of shooting who asks the judge to toss up a dime and shoots it clear through from across the courtroom. At one level, scientific performances are an extension of hypothesis testing: To hazard a clear prediction of an eclipse and prove right is to perform experiment, and to repeat this successfully time and time again is to perform a repeatable experiment. However, there is also a "macro" aspect to performative success. The fact that centuries of scientific process have led to the development of bombs and computers suggests that the process has in some manner been successful; science, it would seem, must have obtained access to *something* true about the world, or it couldn't possibly have achieved such successes.

These are rational judgments. There is also a more visceral side to performative demonstration: To put it rawly, bullets triumphed over shamans. Success on the battlefield and elsewhere has plausibly contributed more than any other factor to the epistemic reputation of science in the mind of the public "jury." Conversely, because claims of magical powers have so often failed when put to the test, they have lost their former prestige. This too can be seen as a form of hypothesis testing, if a particularly vivid one: Can prayers *actually* divert bullets? Apparently not. (This type of prestige-loss through performative failure will have relevance to observational studies. More on this topic soon.)

Hierarchy of Scientific Authority

Using the jury analogy, one can create a rough hierarchy of scientific authority: Some scientific arguments are so compelling that virtually any rational juror will find them incontrovertible; others will persuade only in part while still others will (or should) encounter stony-faced skepticism.

[j] To test this principle in specific cases, Paul Feyerabend famously suggested that scientific conclusions should be subjected to evaluation by lay juries.

The characteristics that push certain scientific argument toward the "incontrovertible" end of the spectrum all lie in features of the scientific method. Table 1 lists some of the most important.

Element A. Observation: The data utilized is of a *type* that can be analyzed and collected objectively and precisely, *has* been shown to be so analyzed and collected, and by all applicable criteria appears to be reliable.
Element B. Experiment: The field of study is one that allows testing of clearly stated hypotheses in repeatable, transparently designed, and rationally convincing experiments that can either support or falsify the hypothesis; that such tests are of a nature as to strike one as objective and unbiased; that they have been repeated widely by scientists unconnected to the original researchers; and that the results of such independent and repeat testing consistently support the hypothesis.
Element C. Logical Inference: The inferential methods used are highly convincing.

Table 1. Elements of a Strong Scientific Argument

One can use this list explain the standard division of science into "hard sciences" and "soft sciences." In a classic soft science field, such as psychology, Element A is relatively weak since observations of human behavior, much less state of mind, are seldom of a type that can be observed with great precision and objectivity. This softness of observation further tends to lead to vague hypotheses and fuzzy logical inferences.

In contrast, physics is the premier hard science because it satisfies all these criteria. It makes use of observables that stand high on the scale of "objective," such as weight and length, and in the area of inferences and logical arguments, relies to a great extent on mathematics. (Use of mathematics is always a plus. Even in a courtroom, if it has been shown that $100 was received and $20 spent, the jury will tend to accept the laws of subtraction and conclude that $80 remain unaccounted for.)

While the classic soft sciences receive that description from the fuzzy nature of their observables, there is another type of soft science whose softness derives from problems with inference. Economics is a good example. While precise data can, in principle, be obtained, a natural reluctance on the part of countries to allow arbitrary experiments on their economies hamstrings the would-be "experimental economist." In turn, the absence of experiment prevents economists from testing their hypotheses, thereby requiring them to build logical cases on ideas that have not yet been proven true.

Such fields might be usefully described as "hard data, soft inference." Public Health is another example. Like economics, it relies primarily on observational (epidemiologic) rather than experimental studies. In particular, it seeks to infer cause/effect relationships from observational data. And here is where all the trouble lies.

It is notoriously difficult to make convincing cause/inferences from purely observational data. Thus, public health is caught up in soft inferences. Furthermore, these postulated cause/effect relationships are used to justify public health interventions. Interventions are a type of performance, and if such interventions fail, legitimate doubts may arise as to the epistemic status of the field that staked its credibility on that performance.

We have now arrived at the central subject of this article. To restate our research question, we shall examine the epistemic status of epidemiologic argument, and attempt to discover whether public health inferences based on observational studies deserve a high, medium or low status as scientific evidence.

Observational Studies

That the status of observational studies has come under question can be seen in the very title of an article published in the January 2001 edition of *The International Journal of Epidemiology*: "Epidemiology — Is It Time to Call It a Day?"[11]

In this article, the journal's *own* editors, George Davey Smith and Shah Ebrahim, openly consider whether their field should close up shop. They ultimately decide that epidemiology can survive, but only if it undergoes extensive self-examination and self-criticism. And they are afraid this won't happen. They worry that, instead,

> "as in many decaying research programmes, auxiliary hypothesis will be mobilized to explain each apparent 'mistake,' on a case-by-case basis rather than there being a re-evaluation of aspects of the broader paradigm within which the discipline operates."

The mistakes referred to here involve a dramatic performative failure, the repeated discovery that conclusions drawn from large population studies were wrong. To paraphrase science writer Gary Taubes, almost *every time* a large randomized controlled trial has been performed to validate conclusions drawn from observational studies, the results have overturned expectations.[12] Worse still, reliance on observational studies has resulted not only in mispredictions but also in outright public health blunders. However, rather than admitting error, epidemiologists have issued excuses in the form of auxiliary hypotheses; that is to say, they have blamed it on inauspicious crows.

Smith and Ebrahim's editorial was written prior to the major failures of epidemiological prediction to be discussed shortly. They were rather prescient; long before most of their colleagues, they were willing to call both hormone replacement therapy and antioxidants examples of failed epidemiology. They additionally raise many forgotten examples of epidemiologic errors to suggest that the field has failed far more often than it succeeded (in the area of population studies, anyway), and has gotten away with it only because of a few signal triumphs, such as the iconic linkage of cigarettes with lung cancer. Some of the colorful stories they recite include seriously faulty (and often racist, classist, sexist or

homophobic) cause/effect claims regarding the cause of pellagra, schizophrenia, Down's syndrome, peptic ulcers and HIV.

This practical, performative failure is responsible for the current crisis of epistemic status. However, these failures originate in a more fundamental scientific problem: observational studies are famously problematic as a means of arriving at cause/effect inferences.

Experiment vs. Observation: The Power of Random Assignment

In 1843, John Stuart Mill laid out fundamental empiric principles for determining cause and effect. Table 2 lists the most important of these. Taken together with the more recent concept of randomization, they demonstrate the fundamental superiority of experiment over mere observation for determining cause and effect.

To illustrate the use of these methods, consider the following scenario: Suppose one has been presented with a mysterious object from the future that possesses a button. Suppose further one discovers that whenever one presses this button, the nearest house soon catches fire. The fact that spontaneous house combustion occurs every time the button is pressed satisfies the Method of Agreement; the fact that spontaneous house combustion almost never occurs in the absence of button pushing satisfies the Method of Difference. Taken together, these findings allow one to rationally arrive at a cause/effect conclusion: Pushing the button *causes* houses to self-combust.[k]

Method of Agreement "If two or more instances of the phenomenon under investigation have only one circumstance in common, the circumstance in which alone all the instances agree, is the cause (or effect) of the given phenomenon." Method of Difference "If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance in common save one, that one occurring only in the former; the circumstance in which alone the two instances differ, is the effect, or the cause, or an indispensable part of the cause, of the phenomenon." Method of Residues

[k] We (and Mill) have assumed one more factor, that of temporal ordering. Houses must catch fire _after_ the button is pushed, rather than before.

"Deduct from any phenomenon such part as is known by previous inductions to be the effect of certain antecedents, and the residue of the phenomenon is the effect of the remaining antecedents."

Table 2. Mill's Methods for Determining Cause and Effect.[13]

The above is a description of a type of idealized experiment in which the effect of a single variable is entirely isolated. In the real world, matters are seldom so pure. When Newton set out to prove that falling bodies accelerate in exact proportion to the force applied to them, his demonstration was bedeviled by the effects of air resistance. What he wanted to show is that in an environment free of friction and of extraneous forces, objects accelerate if and only if the specified force is applied to them (just as houses self-combust when and only when the button is pressed.) In order to properly test his hypothesis, he had to struggle to eliminate these extraneous factors.

Thus, one of the main challenges in designing experiments is to determine what extraneous parameters are relevant and then to eliminate them by controlling the environment. This may be quite difficult. However, the situation is much worse when one cannot manipulate the environment at all. Suppose, for example, one is forced to study the aforementioned fire-starting object used by a ten year old who lives in a neighborhood that is being encroached upon by a wildfire, and where half the homes are made of fireproof titanium (though they appear normal.) Suppose, further, that this wayward child only pushes the button near fireproof houses or wood houses that have *already* caught fire. Mere observation of the object as it is used in such a whimsical environment would likely lead one to conclude that it is a mere toy; it does nothing at all. However, such a prediction would lead to poor performative consequences if tried.

The example just given has two essential elements: the presence of other relevant variables (fireproof status of some buildings, nearby wildfire) as well as systematic bias with respect to those variables (the child's tendency to only push the button where it has no obvious effect.) If systematic bias weren't present -- e.g., if the child flipped a coin to decide whether to push the button, rather than consulting whimsy – an observer might be able to guess the effect of the button despite the presence of the extra variables. Though some buildings would burst into flame without button pushing and others would fail to do so despite button pushing, careful analysis of proportions could demonstrate a likely cause-effect relationship.

One is enabled to draw such a conclusion because of the absence of systematic relationship between the outcome of coin flipping and the characteristics of the building in question. Mathematically speaking, the random character of coin flipping ensures that the characteristics of the building and the use or non-use of the button are statistically independent. This fact allows application of the methods of Residues: One can, in effect, "subtract" the effect of fireproof status and exposure to sparks from the observed events, and apply the methods of Agreement and Disagreement to the results. Insofar as the variables are entangled, however, there is less to subtract, and when they are perfectly entangled there is nothing to work with at all. Thus, this method is most successful when the variables are

independent; to put it another way, when their occurrence or non-occurrence together is rather random.

Effects functionally similar to such randomization occur in a so-called "natural experiment." Consider John Snow's famous description of the nature of the London water supply as he observed it during a cholera epidemic.

> "No fewer than three hundred thousand people of both sexes, of every age and occupation, and of every rank and station, from gentlefolks down to the very poor, were divided into two groups without their choice, and, in most cases, without their knowledge; one group being supplied with water containing the sewage of London …, the other group having water quite free from such impurity."[14]

The circumstance Snow faced contained two key elements that allowed him to arrive at a strong cause/effect claim: (1) Almost all cases of cholera occurred in people exposed to impure rather than pure water. (2) Due to peculiarities of the London water system of the day, cleanliness of water supply was substantially unrelated to class, wealth, overall health, social network or any other factor likely to be involved in the transmission of cholera.

It is the latter element that resembles randomization. People did not flip a coin and decide whether to drink pure or polluted water, but chance distributed the two types of water in a manner that served as the functional equivalent. Imagine that, instead, all the rich citizens of London had been supplied clean water while all of the poor citizens were forced to drink polluted water. In these (more typical) circumstances, the consistent association between cholera and consumption of polluted water would no longer strongly indicate a cause/effect relationship, as virtually any other issue incident to poverty might have been actually at fault.[1] It was the lack of tight linkage (the essentially random association) between the relevant variables that made it possible to lay blame on the water itself.

Having by luck been handed an essentially randomized situation, and being therefore able to strongly argue for a cause/effect relationship, Snow was enabled to recommend an excellent intervention: famously, "shutting off the Broad Street pump." As this classic episode in the history of epidemiology illustrates, before one can rationally choose a solution to a problem, one must first correctly identify the cause/effect relationships involved in that problem. Had circumstances been different and other aspects of poverty than water purity had been the true cause of the epidemic, shutting off the pump would have accomplished precisely nothing.

The chance combination of circumstances encountered by Snow are produced intentionally when one sets about performing an experiment. For example, if given the fire-starting object to experiment around with, one would set about comparing the effects of pushing the button and not pushing the button in a variety of circumstances. One might in particular seek to avoid linkage between button-pushing and other relevant variables by conducting these

[1] In modern terminology, polluted water might have been a "marker" for the true cause, rather than a cause itself.

experiments in circumstances free of wildfires. In other words, without any conscious invocation of the technique, even casual experimentation entails a sort of randomization.

The benefits of randomization are invoked to particular effect in a type of experiment used to evaluate medical treatments: the randomized placebo-controlled clinical trial (RCT). Medical researchers must contend with the fact that medications affect different people in different ways. To try to account for the innumerable potential variables one by one would be an impossible task. Instead, researchers utilize the procedure of "random allocation" to address all the variables, known and unknown, in a single stroke. This process involves using computer-generated random numbers or actual coin flipping to assign half the study participants to a placebo group and the other half to a treatment group.[m] Personal characteristics will still affect response to the treatment under study, but these background differences should, on average, occur to the same extent in both groups. Therefore, any difference in outcome *between* the groups can be attributed to the treatment itself: the treatment *causes* the benefit.

By analogy to these randomized controlled trials, the circumstances encountered by John Snow are described as "quasi-randomized." But Snow was lucky; it was good fortune as much as native intelligence that earned him his place in the history of epidemiology. Epidemiologists seldom encounter anything that resembles randomization when they are asked to draw cause/effect conclusions from observational studies.[n] In most cases, so to speak, the rich drink clean water and the poor drink polluted water. These and similar linkages of relevant variables greatly complicates attempts to draw cause/effect conclusions from mere observation.

While randomized experiment cuts through this Gordian knot at one stroke, epidemiologists are forced to untangle the strands in detail; they must attempt to accomplish what performers of RCTs religiously avoid, which is to say account for each variable one by one. The situation is made additionally difficult by the fact most cause/effect interactions are probabilistic rather than definite; e.g., the presence or absence of factor alters the *chance* of a subsequent effect rather than consistently producing or preventing it.[15] Such probabilistic causation causes little difficulty in experimental studies, as one can still discern differences in outcome distribution between groups. However, in observational studies, these indefinite and usually unknown statistical relationships add yet another layer of confusion to an already problematic situation.

---

[m] Here, formal rather than pragmatic randomization is essential. History has shown that no matter how hard they try to be neutral, if doctors themselves assign participants to treatment and placebo groups, unconscious factors cause them to systematically allocate such assignments in a manner that greatly alters the outcome. One estimate is that such bias can magnify the apparent benefits of a treatment by up to a factor of seven. See, for example, Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomized and non-randomized clinical trials. *BMJ*. 1998;317:1185-90.

[n] This applies, in particular, to large population studies such as those discussed in detail later in this paper. When epidemiologists seek to trace the origin of an epidemic, matters are statistically simpler.

To push our strained analogy one more lap, let us imagine that pushing the button on the future object only occasionally and with some delay causes nearby buildings to catch fire. Let us further imagine that the buildings are only erratically fireproofed, that sparks from the nearby wildfire are variably inflammatory, and that our devilish child may or may not be inclined to push the button selectively with the goal of concealing the device's real effect. In addition, we must consider the likely possibility that people who live nearer the brush-covered hillsides are more likely to fireproof their homes, and the equally likely possibility that people with relatively fireproof homes tend to become careless regarding fire danger. Finally, let it be the case, perhaps, that some people secretly possess special machines that disable the fire-starting-device, but that our device-wielding youngster has conceived a special hatred for everyone with such a protective machine and tries especially hard to burn down their houses.

Imagine how hard it would be to determine the actual effect of the button via mere observation of such a messy situation. And this, rather than Snow's London, is more typical of the situation epidemiologists find themselves in when evaluating large population studies. The only difference is that in real life there are many *more* variables involved, and their effects are much less understood.

Statistical Adjustment

Randomized controlled trials shuck off all these complications at a stroke. Observational studies, on the other hand, remain mired in them. In a labored attempt to disentangle the variables, statisticians employ numerous highly sophisticated mathematical methods, but these remain provisional and problematic, and none achieve anything close to the success of randomized controlled trials.

One standard approach involves using so-called statistical regression models. In essence, this involves using Mill's Method of Residues in a probabilistic form. However, to succeed at this, one must *know* the effects of extraneous variables, and the desired information is often only available from other observational studies that suffer from the same problems. Another method, called sensitivity analysis, attempts to estimate the maximum likely error produced by interdependence of variables. However, this itself cannot be done without considerable knowledge regarding the cause and effect relationships between the variables. And that is precisely what one usually doesn't know.

The problem ultimately rests on the presence of so-called "residual confounders," unidentified variables that lie in intermediate positions on the chain of cause and effect. To return to the cholera example, suppose again that London's water supply had systematically distributed clean water to all the rich and polluted water to all the poor (instead of crossing class lines as happened in reality.) Suppose further that, unknown to John Snow, the polluted water was actually free from cholera, but poor people were malnourished in such a way that they were extremely susceptible to developing the disease. Then, the observed cause/effect connection between polluted water and cholera would be apparent rather than real; the actual

cause/effect relationship would go from malnutrition to cholera. The association between use of polluted water and cholera would be an epiphenomenon, a factor unrelated to the disease but accidentally associated with it via intertwining of variables. The actual causal chain (commonly called a "causal diagram") would go as follows: (A) Poverty causes malnutrition. (B) Poverty causes consumption of bad water. (C) Malnutrition causes cholera. (D) Bad water has no influence on cholera.

In this example, poverty would be the devilish confounding factor; if unidentified, it would be a "residual confounder." And, contrary to widely held misconception, there is no purely mathematical way to eliminate residual confounders. None. Not statistical regression. Not stepwise or collapsibility-based methods. Not sensitivity analysis. Nothing is guaranteed to work. To quote the highly respected researcher Judea Pearl;

"[T]he issue [of unidentified confounding factors] cannot be corrected by statistical methods alone, not even by the most sophisticated techniques that purport to 'control for confounders,' such as stepwise or collapsibility-based methods … It is not determinable whether a certain variable is a confounder because this depends on all (other) confounders and biases together."[16]

Pearl goes on to prove if one wishes to entirely eliminate the effect of confounding factors, one must first fully understand the biological cause/effect relationships involved; in other words, one must be able to create a full "causal diagram" as sketched out above for cholera/malnutrition/poverty/bad water. But if one cannot eliminate confounding factors, one cannot identify cause/effect relationships. The only escape from this circular trap is to employ extra information obtained from outside the body of observational evidence: in other words, to perform experiments. Thus, it would appear that experimental studies are not only better than observational studies at discovering cause/effect relationships, experiments are a necessary prerequisite to interpreting observational studies at all.

The Value of Non-Experimental Evidence

That is too extreme a position. In my brief summary of Pearl's position, I left out an important detail: The extra information required for interpreting observational studies may come not from formal experiments, but from general knowledge of the world. When Snow assessed London's water supply as quasi-randomized, he utilized a great deal of such knowledge: his understanding of what factors are plausibly relevant to cholera susceptibility, his awareness of class structures, detailed information on personal habits, generalizations about human behavior, etc., etc. Using this information combined with an intuitive understanding of the value of quasi-randomization he was able to arrive at convincing cause/effect claim without performing any additional experiments.[o]

[o] Ultimately, however, "knowledge of the world" derives from experimentation, if only on an informal level. Pearl's expertise comes from an usual direction: the field of machine learning, where it is of interest to discover how computers may develop knowledge of the world similar to that of humans. His proof that purely statistical means cannot identify confounders derives from research in this area. It turns out that machines too must go beyond mere observation if they wish to develop

Quasi-randomized circumstances are not rare. When CDC investigators track the cause of a sudden outbreak of a rare condition, they are following in the footsteps of Snow. Consider cases of E *coli* linked to tainted spinach. Just as the London water supply cut across class lines, and exposure to polluted water was tightly associated with cholera, supply of tainted spinach was plausibly unrelated to any factors of susceptibility to E *coli* infection. Also, almost every case of E coli infection occurred in people exposed to the spinach. Thus, it was possible to establish beyond reasonable doubt that the tainted spinach caused the outbreak.

However, large population studies only seldom offer the same opportunity.[p] Sometimes, one can "slice" the available data in such a way as to induce something that resembles randomization. This approach has recently been adopted in the field of econometrics as a means of analyzing historical economic data.[17] Most of the time, though, various relatively mundane forms of argument are used to make an empirical case for the existence of a cause/effect relationship. The most famous of these approaches have been canonized as the Bradford Hill Criteria, listed in Table 3.

We've already employed almost all of these criteria already, though not by name. Returning once more to the London cholera epidemic, we can see the following: The fact that cholera occurred almost exclusively in people who drank contaminated water satisfies the criteria of "Consistency" and of "Specificity." That a high percentage of people who drank polluted water developed cholera contributes to "Strength of association." Because people did not develop cholera prior to drinking polluted water, but within a fairly predictable time afterwards, both "Temporality" and "Consistency" are satisfied. Though bacteria had not yet been discovered, doctors had already come to suspect that one would do better to avoid drinking sewage, thus satisfying "Plausibility." No other water borne diseases were well established, but the general notion of contagion via other than human-to-human contact could be imagined by "Analogy" from smallpox-laden blankets. Thus, the only two Bradford Hill criteria lacking were biological gradient and experiment.[q]

"street smarts": they must interact with the world. In other words, they must experiment. Presumably, humans too derive their general knowledge from such experiments, "conducted" throughout life. These are functionally equivalent to controlled trials, if less systematic.

[p] The iconic epidemiologic success linking cigarettes and lung cancer was exceptional. Lung cancer is an otherwise rare disease, and use of cigarettes – especially in certain populations, such as the military – cut across class lines. Thus, here even population studies could be interpreted as quasi-randomized. There are other examples, such as asbestos and mesothelioma. But most of the time large population studies supply instances where variables are highly entangled rather than fortuitously held apart.

[q] Biological gradient, or dose relatedness, applies more to toxins than infections. One of drop of cholera-laden water is about as bad as a quart, because, unlike poisonous substances, cholera bacteria multiply. Experiment would remain out of reach until Robert Koch established the germ theory of disease 50 years later – inspired, in part, by Snow's research.

1. Strength of association between proposed cause and effect.
2. Consistency of findings: The association is reproducible.
3. Specificity of effect attributed to cause: Ideally, the effect *only* occurs in association with the cause.
4. Temporality: The proposed cause must occur earlier in time than the proposed effect.
5. Biological gradient: the greater the amount of the proposed cause, the more frequent (or severe) the effect.
6. Plausibility based on known science; also known as "biological mechanism."
7. Experiment, if available.
8. Analogy to other situations where cause and effect is known.
9. Coherence: The overall picture created by all of the above is consistent and convincing.

Table 3. Bradford Hill "criteria" for drawing cause-effect inferences from observational studies.

We have gone along with standard practice here and accepted the descriptive term "criteria." But Bradford Hill, himself, took pains to stress that his principles were not criteria at all, in the usual sense of a set of necessary and sufficient rules. His preferred term was "viewpoints."[r] As he wrote in 1965,

> "None of my nine viewpoints can bring indisputable evidence for or against the cause-and-effect hypothesis and none can be required as a sine qua non. What they can do, with greater or less strength, is to help us to make up our minds on the fundamental question – is there any other way of explaining the set of facts before us, is there any other answer equally, or more, likely than cause and effect?"[18]

In other words, the Bradford Hill "criteria" are not criteria at all; they are components of an empirical argument one can use to develop a convincing cause/effect claim in the absence of experimental data. However, this should come as no surprise. As Judea Pearl's work demonstrates, no criteria can *possibly* be developed to rigorously justify cause/effect claims from observational data. More broadly, this is yet another example of the epistemic limitations of science, where strong empirical arguments are the attainable goal rather than precise mathematical proof.

It so happens that one can more easily build strong cause/effect arguments from experiment than from pure observation. However, not even experiment allows us to prove the existence of cause/effect relationships.[s] The best one can ever hope to do is create a strong presumption

[r] In the succeeding text, we will still call them criteria, as that term has become part of the language.

[s] One cannot *prove* the existence of a cause/effect relationship for the simple fact that no one has ever managed to precisely define what the expression "cause/effect relationship" even means. One of the better, if still flawed, approaches involves the language of counterfactuals. As per this method, the expression "A causes B" is reinterpreted as "If A had not happened, B would not have happened."

of a cause/effect relationship. Thus, while observational studies are congenitally inferior to randomized controlled trials for this purpose, they may nonetheless at times provide sufficient evidence to establish such a claim beyond reasonable doubt, as we have already seen.

In the 2008 edition of their standard text *Modern Epidemiology,* epidemiologists Greenland and Rothman devote an entire chapter to defending observational evidence against what they see as excessive enthusiasm for experimental science.[19] Among their many clarion cries, one finds this statement:

> "The non-experimental nature of a science does not preclude impressive scientific discoveries [such as] plate tectonics, the evolution of species … and planets orbiting other stars."

This claim is uncontroversial. No one would deny that astronomy is a real science, nor that most of its claims are derived from observation alone.[t] However, a close look at these examples serves to highlight rather than overcome the problem with epidemiology.

Consider plate tectonics, a well established explanatory model that ties together innumerable observations. The case linking the Appalachian mountains with the Anti-Atlas range of northeastern Africa and the Caledonia mountains of Scotland is utterly convincing, for

---

Since, in real life, A _has_ presumably happened, the supposition is counter to the facts: it is counterfactual. This alternate way of expressing cause/effect relationships illuminates fairly deep issues regarding the value of randomized placebo-controlled trials. In brief, randomization allows one to assume that members of the treatment group are on average identical to those in placebo group in every sense other than the treatment to which they have been assigned. This allows one to test the counterfactual in real life. For example, consider the following cause/effect claim, "This treatment provides benefit. The corresponding counterfactual is "If this group of people had not taken the treatment, they would not have shown as much improvement in their condition." As it happens, this counterfactual is *instantiated* in the placebo group. It is random assignment that makes the counterfactual possible, by "cloning" the treatment group. Observational studies can seldom serve the same function due to the general non-equivalence between those people who happen to choose a treatment and those who do not. One can regard statistical means of attempting to draw cause/effect conclusions from observational studies as *simulations* instead of instantiations of the counterfactual. For more information on the counterfactual approach, see Menzies, P. Counterfactual Theories of Causation Stanford Encyclopedia of Philosophy. Available at: http://plato.stanford.edu/entries/causation-counterfactual. Accessed 11/20/09. For general issues on causation, see Dowe P. Causal Processes. Stanford Encyclopedia of Philosophy. Available at: http://plato.stanford.edu/entries/causation-process/ Accessed 11/20/09

[t] Of course, the techniques of astronomical observation often involve aspects of physics that can be thoroughly tested in experiments here on earth. But, in a sense, this simply represents means of validating observational techniques. Astronomy has an additional aspect that takes it beyond the merely observational: We can send spaceships to other planets. Such feats offer, among things, a type of performative argument for validity. Astronomy also demonstrates performative virtue by making highly precise predictions, a feature that distinguishes it from the example to follow above, that of plate tectonics.

example, and one thoroughly ready to believe these were once a single mountain range, created about 350 million years ago when Gondwana and Euramerica collided.

However, this is all description. There are no proposed interventions based on plate tectonics. The moment one attempts to contemplate any such intervention, one runs into the same problems that plague all primarily observational sciences. Suppose experts in the field announced that well-established computer models of tectonic movement predicted a 9.0 earthquake in Los Angeles on or around January 1, 2011. Such an earthquake is due eventually and is expected to cause massive destruction and loss of life, but would Homeland Security order an evacuation of Los Angeles based on this prediction? Of course not. Any such theory based on mere observation would first have to prove itself by accurately predicting *other* earthquakes.[u]

In other words, Homeland Security, popular opinion and any rational jury would demand a significant period of hypothesis testing: of experiment. If our plate tectonic experts frequently failed, and, worse, if they came up with elaborate explanations for each failure (*the earthquake in Burma took the energy from the one I predicted in Armenia*), they would rapidly fall into the category of cranks.

Therefore, the real problem with observational studies is not their validity per se, but whether conclusions (primarily, cause/effect conclusions) based on them are reliable enough to justify interventions in real life. And it is precisely in this performative arena that observational studies have run aground.

An (Ill-Timed) Counterattack

In June 2000, the New England Journal of Medicine hosted two spirited defenses of observational studies, in which authors Concato and Benson went so far as to argue that observational studies are just as good and sometimes *even better* than RCTs. At that moment in history, proponents of observational studies already begun to suffer acute questioning due to a series of failures, and it was this public doubt that provoked Concato and Benson's counterattack. However, within a year the "other shoe dropped" in the form of the Women's Health Initiative, and, in retrospect, their arguments are almost self-satirizing. Before turning to the WHI results, we shall briefly examine their brave if doomed attempt.

The focus of these articles was precisely a defense of the performative virtues of observational studies. Comparing the results of RCTs and observational studies, the authors concluded that observational studies have *in fact* proved to give results at least as reliable as those of RCTs. Authors Concato et al. reviewed all meta-analyses published in any of five major journals (*NEJM, JAMA, ANN INT Med, BMJ* and *Lancet*) that tested a clinical intervention (treatment or diagnostic technique) and included both RCTs and observational studies.[20] Utilizing rational eligibility criteria, Concato et al. identified subjects evaluated in

[u] Of note, no such skepticism would arise if astronomers warned that an asteroid was about to hit the earth. This is because astronomy possessed a large body of performative successes. We trust it.

this dual way, and reported that the RCTs and observational studies involved yielded identical results. In conclusion, these researchers concluded, "The popular belief that only randomized, controlled trials produce trustworthy results and that all observational studies are misleading does a disservice to patient care, clinical investigation, and the education of health care professionals."

In the same issue, authors Benson et al. utilized wider eligibility criteria, and found 19 issues studied both in RCTs and observational studies.[21] Again, they reported concordant results. Their conclusion was similar. "The fundamental criticism of observational studies is that unrecognized confounding factors may distort the results. According to the conventional wisdom, this distortion is sufficiently common and unpredictable that observational studies are not reliable and should not be funded. Our results suggest that observational studies usually do provide valid information."

Comment by peer reviewers in this and subsequent issues, however, found problems in these apparently robust findings.[22] The essence of the complaints focused on selection bias; essentially, that Concato and Benson had exercised a pre-publication liberty to adjust selection criteria so great that they were on this basis alone able to make the desired case. For example, not only could they choose the filter for choice of meta-analysis topic, they could also devise criteria by which studies already included in meta-analyses could be excluded. These choices, whether conscious or not, were potentially significant enough to entirely manipulate the outcome. In particular, various peer reviewers noted that by slight changes in criteria or interpretation, observational studies could easily be found that disagreed greatly with RCT results. Thus, rather than that observational studies *usually* coincide with RCT results, a more valid inference from the Concato and Benson data might be that observational studies *sometimes* coincide with RCTs.

However, as noted by several commentators, such a conclusion is simultaneously uncontroversial and unhelpful. Even if conclusions based on observational studies may at times agree with those based on RCTs, there is no means by which one can know in advance whether a certain set of conclusions based only on observational studies will coincide with as yet unperformed RCTs. And, as even Concato and Benson would concur, a sufficiently well-designed RCT can overrule findings from observational studies. (Their attempt to raise observational studies to the level of RCTs in the hierarchy of evidence primarily involved putting the best of observational studies on par with the worst of RCTs.)

But if one does not know whether the hypothesis generated by current observational studies will be overruled by future RCTs, one is in a quandary. Consider this thought experiment: Suppose, there is a 50% chance that a future RCT will discredit current observational evidence. In such a circumstance, any recommendation based on such evidence is no more reliable than flipping a coin, and the practice of issuing such a recommendation, even if it involves a form of evidence gathering, would not be what is normally thought of as "evidence based."

While predictions in physics have earned their reputation for reliability by achieving accuracies of 99.999% or more, this is obviously too high a standard for medical

recommendations. Nonetheless, to be meaningful as such, "evidence based medicine" must involve predictive values markedly beyond that of coin flipping. Exactly where to set the bar is unclear, but if even a figure as low as 80% is required, the question arises: can one, on the face of things, feel certain that purely observational medical evidence reaches that point?

The response of Concato and Benson to these critiques is, with benefit of hindsight, rather cringeworthy.

> "We sympathize with all who find intellectual security in randomization as a method of ensuring the validity of study results. Surely, however, other methods (matching, stratification, adjustment, and restriction) are available to ensure validity when randomization is absent."

If only it were so. Alas, the feeling of intellectual security provided by RCTs was just about to be justified by results, and the power of statistical methods to make observational studies reliable, thoroughly dashed.

Large Failures of Observational Studies

In this final section, we turn at last to the recent major shocks to observational evidence and the reactions these have provoked. Of these, the most dramatic and consequential involves hormone replacement therapy for postmenopausal women, and we shall begin there. ***HRT***

The 1968 publication of Robert Wilson's bestseller *Feminine Forever* initiated the era of hormone replacement therapy, in which healthy women en masse were encouraged to take estrogen.[23,24] The underlying theory was plausible: Before menopause, women have high levels of circulating estrogen, while after menopause levels of estrogen plummet. Therefore, menopause has the appearance of a deficiency disease like hypothyroidism. As such, it can be easily treated by replacing the missing hormone.

Initially, it was claimed that use of estrogen would keep women's skin looking young and "feminine." This rationale gradually declined in favor of more strictly medical claims. Of these, the most prominent was that use of HRT should dramatically reduce heart disease risk. The logic here is credible too. Prior to menopause women are less likely than men to have heart attacks than men, while after menopause the rates converge. Furthermore, observational studies showed dramatic benefit: women who did not use HRT were almost twice as likely to experience heart disease as those who did. Finally, the Bradford Hill criteria of biological plausibility could be found in the fact that estrogen has a favorable effect on lipid profile.[25] Cholesterol-lowering drugs had been approved for heart disease prevention on weaker evidence than this,[v] and in 1990 the FDA approved a label change allowing hormone replacement therapy to claim indication for heart disease prevention as well.

[v] There was not then as yet any direct evidence that people who took cholesterol lowering drugs had a lower incidence of heart disease; only that people who (for whatever reason) had lower cholesterol had less heart disease. Thus, the chain of argument for hypolipidemics was (and, except for statins, remains) weaker than that for HRT. Much more on this later.

In addition, the same observational studies that found cardiovascular benefit with HRT were seen to demonstrate that it also reduced rates of other many other conditions, including etiologically related ones such as stroke, as well as unrelated health problems such as Alzheimer's Disease and colon cancer.[26]

Still, there were some voices of dissent. Beginning in the early '80s, physicians and other activists in the Women's Health Collective (the authors of *Our Bodies Ourselves*) noted that that HRT was being recommended for *healthy* women, and that therefore the evidence for use should rather be set at a higher than a lower bar compared to the standard required when drugs are used to treat sick people. Such drugs require double-blind studies prior to approval; shouldn't the HRT recommendation require at least as much support? In fact, shouldn't it be required that HRT not only affect surrogate markers such as cholesterol, but also the intended goal of improving those surrogates, in this case reducing heart disease?

Evolving critique of the existing observational evidence buttressed these concerns. In 1994, an article was published suggesting that all observational evidence supporting HRT use might be fatally confounded by a "healthy user effect," the fact that women who happened to use HRT had many other health-positive characteristics independent of hormone use. As evidence that this confounder might produce an effect of sufficient magnitude to skew the results, the authors noted that in the same studies used to support cardiovascular benefits, women using HRT were also found to be likely than non-users to develop breast cancer. However, no one had (so far) suggested that HRT use could causally reduce breast cancer risk; rather, it was widely agreed that HRT should, if anything, slightly raise incidence of breast cancer. Therefore, the relatively lower incidence of breast cancer among women taking HRT mostly likely derived from the generally better health status of the women who chose to use HRT, and not the hormones themselves. Couldn't the same be true regarding HRT and cardiovascular disease?

This critique was immediately blasted down by one of the world's leading epidemiologists Meir Stampfer.[27] In a blistering letter to the editor, Stampfer noted that nearly all the Bradford Hill criteria were met by the observational evidence showing protection from cardiovascular disease, and that therefore it was ridiculous to doubt a true cause/effect relationship.

As discussed earlier, Bradford Hill himself only regarded his categories of evaluation as "viewpoints," not as criteria at all. Nonetheless, they had by this time attained the status of doctrine. Stampfer specifically cited the following:

- Strength of Association: Studies showed a robust risk reduction of 50%.
- Consistency of Findings: The evidence from observational studies consistently showed benefit.
- Biological Gradient: The more women were "compliant" with HRT, the less likely they were to develop heart disease.

- Biological Mechanism: Estrogen not only lowers cholesterol, it is also an antioxidant. (The collapse of the antioxidant hypothesis had not yet occurred. See below.)

In fact, of all the relevant criteria, *only* experiment was lacking, and, in Stampfer's opinion, that hardly seemed necessary in the face of such overwhelming evidence. Against the charge that reduced cancer rates suggested confounding, Stampfer offered biological arguments to the effect that perhaps HRT prevents breast cancer too.

More difficult to counter, however, was another finding that came to light just after Stampfer's rebuttal: In at least one observational study, HRT had proved to be protective against accidental deaths and homicides to the same extent that it reduced cardiovascular death.[28] This seemed rather difficult to attribute to the drug itself. However, reduced homicide rates were perfectly consistent with a healthy user effect based on socioeconomic status. Despite these remarkable findings, the public health recommendation held, and by the mid 1990s it was common among doctors to describe HRT as the single best thing a woman can do to protect her health.[29]

Nonetheless, RCTs of hormone replacement therapy had at last been designed, funded and instituted. The results of the first such study appeared in 1998. This randomized controlled trial enrolled 2,763 postmenopausal women with heart disease, gave half HRT and half placebo, and followed them for four years.[30] The results: use of hormone replacement therapy as opposed to placebo provided *no* cardiovascular benefit.

This would seem to have been an experimental invalidation of the HRT hypothesis. However, just as rainmakers may invoke inauspicious crows, epidemiologists found excuses; in the terminology of Smith and Ebrahim, they deployed auxiliary hypotheses to defend a failing research program. For example, perhaps it was too late for HRT to help women who already had cardiovascular disease; perhaps they required stronger medications, such as hypolipidemic drugs.[w]

These arguments succeeded in maintaining the HRT hypothesis for some years, up through the Concato/Benson defense and past the Smith and Ebrahim expression of concern for their own field. Then, in 2001 came the bombshell.

The Woman's Health Initiative enrolled 27,500 healthy women. Its participants were given either HRT (in one of two forms) or placebo. The trial was scheduled to run for nine years, but it was cut short at just over five years when the results showed that HRT *increased* the risk of coronary artery disease, stroke, invasive breast cancer and pulmonary embolism.[31]

---

[w] The same argument would be used when studies of antioxidants in smokers failed to find benefit. It was even suggested that such studies had been designed to fail, by picking participants too sick to benefit. Actually, however, it is generally easier to show benefit in populations at higher risk of a negative outcome, and that, not desire to fail, was the reason for the study. (It is also generally regarded as more ethical to give treatments to people who are sick than to those who are well, something that should have been considered by those prescribing HRT.)

Subsequent analysis additionally showed that HRT substantially increased rather than reduced risk of dementia.[32]

The results of the WHI could be argued against, but only a minority of epidemiologists essayed the attempt. The consensus response was to take the WHI results as an overwhelming experimental disproof of the HRT hypothesis. Within a year, prescription of HRT dwindled away almost to nothing. The fact that such prescription had gone on so long —that a hypothesis based on observational evidence had been promulgated and defended by leading epidemiologists for decades, only to at last prove incorrect – was devastating. Epidemiology had staked its credibility and lost. This at last caused a true crisis of faith, provoking a re-evaluation of the very process of drawing conclusions from observational studies.

In the language used in this paper, performative failures gave rise to a re-examination of the arguments used to support the predictions. The first proposed explanation of the error invoked the healthy user effect: Women who used HRT were perhaps healthier for other reasons than those who did not. In particular, HRT users, on average, belonged to a higher socioeconomic class than non-users, and socioeconomic status is well known to correlate closely with overall health. Therefore, researchers looked back at the original observational data and used statistical methods to adjust for socioeconomic status and/or education level. (Rather amazingly, this had not been done before.)

This attempt at post-hoc (after the fact) data adjustment was successful, but only in part. Removing the effect caused by socioeconomic status yielded a match with the WHI results regarding heart disease but not regarding stroke.[33] Apparently, unequal "allocation" to socioeconomic status was not the sole confounding factor. Researchers than tried other factors, giving them various weightings, and using every possible statistical technique available. Despite the fact that they knew where they were going, it took several years before statisticians could successfully reinterpret past observational data to fit current knowledge.[34]

Whether this final statistical correction was a true discovery of past error or simply the result of number massage remains unclear. But one thing is absolutely clear: proper statistical adjustments of the observational data could not be made until *after* the desired results of those adjustments were made known via the WHI results.[35]

Diane Petitti, the researcher who picked up on the apparent prevention of homicide by HRT, notes a series of lessons to be learned from the HRT debacle. [36] Of these, perhaps the most interesting "lesson" is, "Do not be seduced by mechanism." This statement stands in direct contradiction with the "biological plausibility" element of Bradford Hill's criteria. She additionally advises researchers to keep on guard against three other major Bradford Hill criteria, "strength of evidence," "consistency," and "coherence." For, prior to the WHI results, strong, consistent, coherent and biologically plausible evidence indicated that HRT reduces cardiovascular risk; while, as we now know, HRT raises cardiovascular risk, and by about as much as it was thought to reduce it. Thus, epidemiologists had used piles of data and satisfaction of numerous Bradford Hill criteria to arrive at a conclusion that was exactly upside down.

The repercussions from the prolonged prior defense of this error continue. Attempts have been made to determine in what circumstances observational studies might be reliable, but a note of sober self-examination pervades the writings of many serious epidemiologists.[37] This is a healthy sign. One hopes the error will not be forgotten too quickly.

Cholesterol

Presumably, the erroneous conclusions regarding HRT derived from various confounding factors that skewed the results seen in observational studies. The net result was an erroneous assignment of cause and effect: Prior to the WHI, it was thought HRT protects against heart disease. After the WHI results, it now appears that people who were protected against heart disease by other factors (such as high socioeconomic status) happened to be more likely to use HRT. Apparently, their overall health advantage was sufficient to offset the harm caused by HRT leaving a net benefit compared to others who did not harm themselves by taking HRT but were generally less healthy for other reasons.

Thus, in real life, women who took HRT did not do so badly, but they would have done even better if they had avoided hormones. The others who did not take HRT were fortunate in their decision, for use of HRT would have harmed them further. However, physicians, relying on their mistaken sense of cause and effect, had been working hard to get *more* women to take HRT. It now appears that to the extent they succeeded, they caused harm. This illustrates the close connection between proper assignment of cause/effect relationships and proposed interventions, and the inherent problems with using observational studies as a basis for such interventions.

It also leads one to question other supposed cause/effect relationships based entirely on observational evidence, especially those used to justify interventions. One of these involves serum cholesterol. While there is no doubt that higher cholesterol levels (or, more precisely, certain ratios) are strongly associated with higher rates of heart disease, this by itself does not prove that high cholesterol *causes* heart disease. And if cholesterol does not cause heart disease, using drugs to lower cholesterol makes no sense.

The association between high serum cholesterol and high rates of heart disease was initially established in the Framingham Heart Study. From this uncontroversial but non-causal finding, the leap to causality was made with little discussion. At the time, the argument seemed open and shut: (1) Atherosclerotic plaques contain cholesterol. (2) Atherosclerosis causes heart disease. (3) High serum cholesterol is strongly associated with high rates of heart disease. (4) The higher the cholesterol the more heart disease. (5) Therefore, high serum cholesterol causes heart disease.[x]

What no one suspected at the time is that the cholesterol in atherosclerotic plaques does not arise due to a piling up of cholesterol in the blood. Rather, as radio-tracing and other careful

[x] Here we are invoking several of the Bradford Hill criteria: plausibility (biological mechanism), strength of evidence, consistency of findings, biological gradient and coherence.

biological methods have shown, plaque cholesterol is created from scratch by the cells of the arterial wall. Serum cholesterol is an innocent bystander (or, at most, an irritant that causes the cells to behave in this way.) Thus, the argument just given is fundamentally flawed.

So were the interventions based on this argument. The first step doctors took in response to Framingham was to announce that people should not eat eggs. Eggs are high in cholesterol, and so the logic seemed correct; that millions of people were employed in the egg industry seemed irrelevant in the face of the great benefits to be reaped.

As it happens, we now know that serum cholesterol is almost entirely produced in the liver, and that dietary cholesterol intake is virtually irrelevant. Therefore, this recommendation made no sense at all. Furthermore, reanalysis of Framingham showed an entire absence of association between egg intake and heart disease.[38] Thus, the "stop eating eggs" recommendation was at best nonsensical.

Along with eggs, doctors and public health officials also advised the public to cut down on dairy, "red meat" and fat in general. The argument regarding fat in the diet seemed plausible on similar grounds: atherosclerotic arteries are clogged by fatty plaques; grease clogs pipes; therefore, dietary fat clogs arteries. Observational studies seemed to support this argument, especially regarding saturated fat. As in many countries, the higher the saturated intake, the higher the heart disease.[y] This, in turn, led to the creation of a huge industry devoted to low-fat foods (in which sugar made up the missing calories) and the widespread substitution of transfat-rich margarine for butter. Also, two or three generations were raised to believe that it is almost as unhealthy to eat red meat as to smoke cigarettes.[z]

But none of this makes sense. For one, the fat in atherosclerotic arteries is not derived from fat in the blood; like cholesterol, it is created in place. For another, increased intake of fat, or of saturated fat, does not raise cardiovascular risk as measured by current tests of cholesterol levels. True, the more fat one eats, the higher the total serum cholesterol, and it was this relationship that set the "low-fat" recommendation in motion. However, we now know that it is the ratio of certain subtypes of cholesterol that are associated with increased cardiovascular risk, not the level of total cholesterol. And here's the kicker: when one cuts down on total fat intake, HDL ("good") cholesterol falls along with LDL ("bad") cholesterol, and serum triglycerides (somewhat surprisingly) rise. The net result is no change at all in cardiovascular

[y] This wasn't true everywhere. In particular, it wasn't true in France, where higher saturated fat intake correlated with lower rates of heart disease. This has come down in history as "The French Paradox," and the auxiliary hypotheses used to explain it are widely thought of as fact: EG, the French drink red wine, and that protects them. But as the saturated fat hypothesis began to break down, it became clear, in retrospect, that there was no French Paradox at all. France didn't fit the hypothesis because the hypothesis was wrong. It should have been called "The French Counterexample," or simply "evidence against the theory." Bratman S. Controlling Dietary Total and Saturated Fat To Reduce Cardiovascular Disease: A Classic Public Health Error. Final paper. Epidemiology, MPH program, Fall Semester 2008. Medical College Wisconsin.

[z] Many people believe it today, as public health authorities are retreating from that position very quietly, in order to avoid admitting a mistake.

risk as currently calculated.[39, 40,41] Furthermore, reanalysis of observational data now shows that amount or type of fat consumption is unrelated to cardiovascular risk, whether in France or elsewhere, the only exception being trans-fats (whose widespread use is largely a result of public health officials pushing margarine.)[42] Overall cardiovascular risk as currently determined only decreases if one loses weight – and the benefits are seen to the same extent regardless of whether one uses a high-fat Atkins diet or a low-fat Weight Watchers one.

One wonders how much objective harm was done by these suggestions; deaths due to increased consumption of trans fats, obesity contributed to by the high fructose syrup that replaces the missing calories in low fat foods, loss of jobs in the egg and beef industry, not to mention widespread shaming for innocuous habits and tastes. Nonetheless, one result of the cholesterol-causality assumption has been highly positive: the invention of the right drugs for the wrong reasons.

It is easy to design RCTs to test whether medications reduce cholesterol, and if one believes that high cholesterol causes atherosclerosis, any drug that reduces cholesterol should be health promoting. Or so it seemed. Thus, the premature logical leap from the Framingham findings facilitated drug development for high cholesterol, just as a similarly premature leap from the Framingham findings regarding hypertension led to the development of antihypertensives. However, the history of these two interventions has diverged.

Numerous classes of antihypertensives were developed, based on entirely different mechanisms of action. Decades after they were put in use based on purely observational evidence, antihypertensives were finally tested in regards to the intended actual endpoint (reduced cardiovascular disease), rather than just the surrogate measure (reduced blood pressure.) The results have been uniformly positive: every single class of antihypertensive dramatically reduces cardiovascular disease, and all are about equally effective.[43] This can be regarded as proof of causality by therapeutic trial. Since the various anti-hypertensives are so different from one another, the fact that they are almost identically effective strongly supports the hypothesis that hypertension causes heart disease.[aa] But the opposite is the case with cholesterol.

There are, similarly, several classes of drugs used to lower cholesterol. But, here, studies of the actual endpoint have shown that of all these drugs only one class is beneficial, the statins.[44,45] Statin drugs markedly decrease cardiovascular disease, as hoped. But the other drugs not only fail to improve cardiovascular mortality, some seem to increase *non*-cardiovascular mortality.

[aa] This is a form of quasi-randomized natural experiment. Since the mechanism of action of the various classes of antihypertensives vary widely, it is reasonable to assume that their effects on physiologic processes beyond that of hypertension also vary widely. Taking all the studies of antihypertensives together, then, all variables other than blood pressure are being affected in a quasi-random way. Thus, the single variable in question has been effectively isolated, and the consistent effectiveness of antihypertensives strongly suggests that blood pressure is in fact causally related to cardiovascular outcomes.

Looking back, the entire history of cholesterol drug development appears to have been based on an error. Few seem to have paused to wonder whether high cholesterol is merely a marker. One does not treat a marker: people with prostate cancer have higher levels of PSA, but draining PSA from the blood would not plausibly reduce rates of prostate cancer. And yet, it now appears plausible that measurement of serum cholesterol merely analyzes an epiphenomenon, and that the marked reduction of death produced by statin drugs is a *coincidence*.[bb]

Most of these findings were already known by the time Concato and Benson defended observational studies, but they had not created much of a stir, perhaps because the truly effective statin class of drugs had largely replaced the older medications anyway. In 2008, the results of a large study demonstrated that a newly invented class of hypolipidemics failed to improve actual endpoints.[46] Ezetimibe markedly improves lipid levels, but it does not appear to affect rates of cardiovascular disease. Given that of all drugs that lower cholesterol, only the statins reduce heart disease, it is logical to conclude that the entire fifty plus year effort focus on reducing cholesterol may have been based on a faulty cause/effect claim derived from observational data.[47,48,49] These are early days – 1994 in HRT time – and further findings will be necessary before the cholesterol hypothesis actually falls.[cc]

Also of interest is the question of how statin drugs produce their undoubted benefit. One hypothesis commonly mentioned invokes their antioxidant effects. But the antioxidant hypothesis is (or should be) at death's door too.

Antioxidants

Beginning in the early '90s, accumulating evidence from observational studies including Nurses Health strongly suggested that dietary antioxidants could reduce both cardiovascular and cancer risk. In particular, intake of vitamin E and beta carotene separately were each shown to have dose related beneficial associations.

As with HRT and cholesterol, biological plausibility presented no obstacle: investigation of free radicals and the need to quench them has been elaborated in detail. It seemed possible that these safe, "natural" treatments could provide marked public health benefits. And soon (but not before recommendations in favor of antioxidants had already become ubiquitous) the

[bb] But there are numerous examples in medicine of successful treatments whose mechanism of action turns out to be entirely different from the theory about their mechanism that brought them into use. SSRI antidepressants are one example. They were designed to suppress serotonin uptake on the theory that depression is a serotonin deficiency disease; however, their effect on serotonin levels is transient, and fades by about the time that their antidepressant effect begins. Presumably their actual mechanism of action is something different.

[cc] The fact that numerous pharmaceutical companies have non-statin hypolipidemics under development is a particular obstacle.

largest double-blind studies in history were set in motion to verify the expected causal connection.

By now, more than 60 large RCTs have reported their results. All together, more than 200,000 people have participated in these trials, making this the first time that RCTs have involved as many people as the largest population studies. The findings from these studies have been consistent: antioxidant supplements do not work.[50,51] In fact, supplementation with vitamin E or beta carotene may *increase* risk of cardiovascular disease and various forms of cancer.

The mere title of one article by a major researcher in the field expresses the mood of those who had hoped for something better: "*Nutrient supplements and cardiovascular disease: a heartbreaking story*."[52] Once more, inferences from observational studies appear to have foundered on residual confounders.

Again, Ebrahim and Smith's prediction regarding behaviors typical of a "decaying research program" were proved correct when antioxidant supporters instantly deployed a plethora of post-hoc alternative hypotheses. Perhaps the wrong form of vitamin E was used. Perhaps mixed carotenes are beneficial rather than isolated beta-carotene. Perhaps antioxidants must be mixed in a great big mixed basket. Perhaps anything whatsoever, but the fact is, currently there is no good evidence for benefit in heart disease or cancer with any antioxidant, and there is some evidence against these alternative hypotheses.[reference] Using Occam's Razor, the simplest explanation of the events is that the antioxidant hypothesis is wrong, and the body's intrinsic antioxidant system may be sufficiently effective that adding supraphysiologic doses of vitamins may fails to confer any added benefit.

Problems with the antioxidant hypothesis have been developing over an even longer course than those with HRT; the first major negative RCT appeared in 1994.[53] However, the great hope placed in this approach, and its character as "natural medicine," have served to maintain faith in antioxidants, despite the truly stupendous body of evidence against them. Everyone "knows" that antioxidants are good for you, even though we do not know this at all.

Discussion

This article set out to evaluate the current status of observational/epidemiological studies as scientific evidence in the field of public health. As we have now seen, the validity of scientific evidence is established in much the same manner as a legal case is made to a jury. The most trusted scientific arguments, however, possess a type of heightened credibility compared to the average juridical argument, deriving, as we have discussed, from features of the scientific method. It is fair to ask whether inferences based on observational/epidemiologic studies possess a similar character.

Clearly, there are well-known logical problems that stand in the way of drawing cause/effect conclusions from mere observation. These problems were noted as early as 1843 by John Stuart Mill and have since been emphasized by RA Fisher (the "father" of randomization) and even by Bradford Hill himself, the author of the canonical Bradford Hill "criteria" for deriving cause/effect claims from non-experimental studies. Most recently, the work of Judea Pearl and others has shown that not even the most advanced statistical techniques can systematically eliminate confounding factors. As there are plausibly thousands or millions of interdependent factors that influence human health, the issue of confounders arises with considerable force in observational studies of human health.

Occasionally, circumstances of quasi-random allocation occur, such as when tainted spinach produces an epidemic. In such cases, it may be possible to derive a highly convincing argument for causality from mere observation. However, in most other circumstances, such as those involving large population studies, nothing that even faintly resembles random allocation has taken place, and one can more plausibly argue for than against the existence of powerful residual confounders. Thus, it is fair to label any cause/effect claim based on associations seen in such studies as weak juridical evidence rather than persuasive juridical evidence, much less the heightened type of persuasive argument typically labeled "scientific."

Compounding the weakness of the inferential case is a record of performative failure. The recent results described here as well as the decades of errors surveyed in Smith and Ebrahim's article indicate that causal reasoning from observational studies is (except in the special case of quasi-randomized circumstances) is highly *unreliable* as a means of prediction. Recent history can be summarized as follows: Epidemiologists made predictions based on large population studies. (HRT prevents heart disease.) They were so incautious as to go ahead and implement interventions based on these predictions. (We recommend use of HRT.) They did so without waiting for or even particularly encouraging confirmatory experiment. (The observational evidence is so strong, we already know.) When questioned for acting precipitately, they ridiculed their questioners. (All this talk of a "healthy user" effect is absurd.) When experimental results running counter to prediction began to trickle in, epidemiologists raised auxiliary hypotheses rather than admitting there might be a problem. (HRT did not help women in the HERS study because they already had heart disease.)

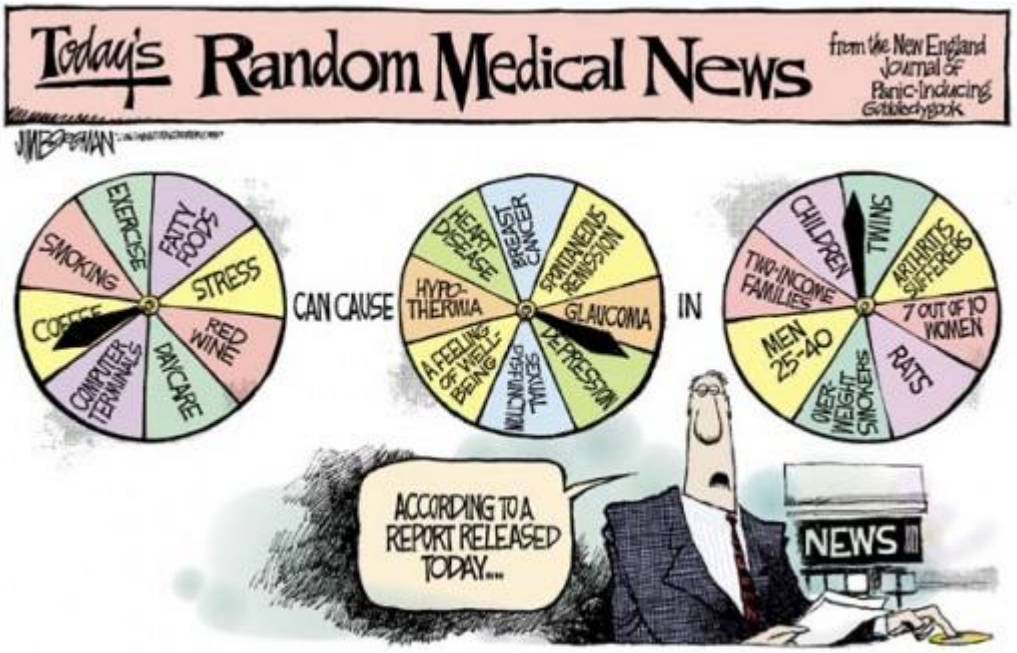
Finally, experimental results arrived that not only failed to verify the prediction, but actually proved that the epidemiologists had gotten it exactly backwards. (Rather than reducing cardiovascular risk by 50%, use of HRT raises it by 50%.) In retrospect, epidemiologic advice proffered as in the public good actually caused public harm. (Use of HRT actively killed rather than saved women.)

Worse still, this was not the first time epidemiologic errors had caused widespread harm. (It was epidemiologists who promoted margarine.) Nor was it only the fourth or fifth time that interventions based on epidemiologic claims had been proved without foundation. (A short list would include the erroneous labeling of eggs, red meat and dairy products as unhealthy; and the unfounded enthusiasm for low-fat and low-salt foods.)

When an expert expends considerable energy defending a claim, and the claim proves to be wrong, that expert loses credibility. If the same expert repeats the process with another claim, and is again proven wrong, one may arguably estimate the retained credibility as near zero. This is especially the case when the claim leads to an intervention, and the intervention kills people. Applying these principles to epidemiologic cause/effect argument from large population studies, a rational jury would not be willing to extend any trust at all, much less the heightened respect accorded to a form of science.

Conclusion

As proven/demonstrated above, epidemiologic claims do not qualify as scientific evidence except in certain limited circumstances, primarily those of the quasi-randomized natural experiment. However, it would appear that epidemiologists do not strictly limit themselves in this way, as the following widely reproduced cartoon suggests.



Cartoon by Jim Borgman, first published by the Cincinnati Inquirer and King Features Syndicate 1997 Apr 27; Forum section: 1 and reprinted in the New York Times, 27 April 1997, E4.

It is often suggested that the media is at fault for twisting non-causal statements into causal ones. Certainly causal statements make for jazzier stories. Consider this headline from a recent news article on data from the Nurse's Health Study: "Fresh Vegetables, Fruits Reduce Diabetes Risk."[54] In fact, the evidence does not support any such causal statement. A healthy user effect is much more plausible, as the same people who ate more fruits and vegetables also exercised more and less commonly smoked cigarettes compared to those who ate less in the way of fruits and vegetables. Thus, a more accurate headline would read, "Higher Intake of Fruits and Vegetables Common in Healthier People but May Contribute Little or Nothing to their Overall Health."

It is safe to say that any article making so equivocal a claim would produce yawns rather than attentive readers, yet, in this case anyway, blame cannot be placed on the media. Consider this direct quote from the lead researcher: "Based on the results of our study, people who have risk factors for diabetes may find it helpful to fill up on leafy greens like lettuces, kale and spinach and whole fruits, like apples, bananas, oranges and watermelon …" The expression "may find it helpful" leaves some wiggle room, but this is still a direct recommendation.

Researchers may also feel tempted for professional or merely emotional reasons to leap over the limited conclusions warranted by evidence and make far more exciting causal claims. Considering this natural tendency, perhaps a new term should be added to the standard narrative of evidence-based medicine, that of "evidence related." This proposed fresh bit of jargon would be utilized in circumstances where the evidence is good but the inferences are weak: that second type of soft science referred to earlier as "hard data, soft inferences."

Still, one is led to wonder why, given the record of failure, epidemiologic pronouncements are taken seriously at all. Various plausible explanations are not difficult to find. For example, there is the fact that statistical inference is notoriously counterintuitive; the public "jury" can scarcely be blamed for failing to notice the problem of residual confounding. In addition, one must keep in mind the indubitable history of success on the part of epidemiologists, from the iconic discovery of the harm of cigarettes to the semi-annual discovery of tainted food sources. That these accurate pronouncements are limited to quasi-randomized circumstances is a fine point again not easily recognized by the public jury.

Finally, perhaps people merely *want* to be given advice on how to stay healthy, and are therefore willing to overlook the failures of those who give them advice. This it is probably impossible to arrive at scientifically verifiable advice on lifestyle issues is an unpleasant truth,[55] and we may tend to prefer scientific fibs that satisfy the emotional need for guidance.

But is there any harm in such fibs?

No one seems to have essayed a careful analysis of deaths caused by prescription of HRT to

otherwise healthy women, but the findings of the WHI suggest that the advice to use HRT may have killed thousands of women. The cholesterol hypothesis was more of a mixed bag, causing a vast increase of transfats in our diet in the form of margarine and other hydrogenated oils, but also leading to the invention of the truly (if mysteriously) effective statin drugs. The antioxidant hypothesis, on the other hand, seems to have entirely failed to harm or hurt on a physical level, and it may have been economically beneficial by creating jobs in the supplement and publishing industries.

However, beyond the specifics of harm done to people, there is another risk in the overuse of observational studies: the loss of credibility that results from "crying wolf." As the Borgman cartoon indicates, the public has certainly developed a certain skepticism regarding the advice promulgated by public health experts. One wonders to what extent the current resistance to swine flu and other vaccinations has been facilitated by a learned lack of trust in public health authorities. In another arena, some portion of the doubt regarding anthropogenic global warming may result from a semiconscious awareness that global warming predictions (necessarily) lack any experimental support. If epidemiologists continue to "deploy auxiliary hypotheses" instead of learning to couch their recommendations more chastely, one might expect that further errors will occur, followed by further loss of credibility, followed by loss of trust when it is needed most.

And yet, there is a need to make public policy decisions even when hard evidence is lacking. For example, (to return to an issue mentioned near the beginning of this paper) *some* standard setting for air quality is necessary. We do know that very high levels of ozone are toxic, and while residual confounders bedevil all attempts to accurately quantify the risks of low-level ozone, we need to draw the line somewhere.

When there are no valid scientific arguments to be made, vaguely evidence-related arguments are certainly as valid as anything else on offer. This itself is unproblematic. However, there often seems to be an irresistible temptation to exploit the reputation of good science by waving about merely evidence-related claims as if they too possessed heightened forms of persuasion. This is ultimately no more than propaganda, and, like all propaganda, it debases the truth. To preserve science's hard earned epistemic prestige, perhaps the label "scientific" should be reserved for cases when traditional scientific norms are achieved.

Much of epidemiology fails to reach this standard. Therefore, it should not appropriate the terminology.

Appendix

Contingency and Consensus in Mathematics and Logic

If science possesses epistemic prestige, mathematics commands something like epistemic adulation, as "mathematical proof" is widely regarded as virtually synonymous with "revealed truth." The field of logic is epistemically prior to mathematics, and its laws are often regarded as true by *necessity*. However, a close look shows that even mathematics and logic depend on contingent facts and human consensus. In other words, they too depend on appeal to a human jury.

Logic

Consider the following a variant of a simple logical principle known as modus ponens:[dd]

Proposition 1: The set A is a subset of the set B.
Proposition 2. Object *a* belongs to the set A.
Inference: Object *a* is an element of the set B.

E.g., if all Corvettes are cars, we can infer that the particular red Corvette sitting there by the curb is a car. But *why* are we allowed to infer this? The obvious answer is, "Because it's obvious!"

In his hilarious dialogue of Achilles and the Tortoise, Lewis Carrol introduces a recalcitrant tortoise who refuses to accept the inference as obvious.[56] The Tortoise demands that Achilles provide an intellectually compelling justification for what he sees as a vast leap of reasoning. But, as Achilles discovers, there is no way to justify a modus ponens inference except via a type of argument that *uses* the rule of modus ponens. The implication is therefore that if one were to run across a class of intelligent beings for whom this most basic of logical principles was not acceptable, one could do nothing to persuade them. To put it another way, modus ponens is acceptable because a broad consensus of people *say that it is*.

But why is it that people agree on this principle?

One hypothesis is that we learn modus ponens and other such basic logical rules through experience. Growing up as children we are told that all Corvettes are cars; furthermore, each time we happened across a specific Corvette we were told this Corvette is, in fact, a car; therefore, we come to learn by experience that the general proposition "all Corvettes are

[dd] The actual principle of "modus ponens" is more general than the example given, but this subtype presents the same issues, and is easier to make vivid.

cars" is related to the repeated instances of "this Corvette is a car;" finally, we apply induction to these and similar experiences and come to accept modus ponens as correct.

But this analysis is hardly satisfying. It would seem that the rule of modus ponens is true *by definition*, and that no induction from experience should be necessary. EG. the very meaning of "The set A is a subset of the set B" is that each of the elements of A is an element of B. Therefore, it is not an inference at all to note that if *a* is an element of A, then *a* is an element of B; it is simply an instance of the meaning of the term "subset."

There are two problems here. The first involves the circularity explained in the Achilles/Tortoise argument: the transition from the first half of the last sentence to the concluding half implicitly invokes the rule of modus ponens. Another and more fertile problem lies in the concept of "set."

For "The set of all Corvettes is a subset of the set of all cars" to be a meaningful expression, its component reference to the "the set of all cars" must itself be a meaningful expression. But what exactly are the elements of the set of all cars? Obviously, a car without a headlight is still a car. What about a car without an engine? One without its front half? A car that has been crushed into scrap metal? A drawing of a car?

In general, any real world example of a set is full of boundary problems like these.[ee] The image of a child learning to make set/subset inferences shows that this is simply elided, not solved. Infants very much enjoy engaging in such behaviors as pointing to a broken car and inquiring, "Is that a car?" Their purpose in making such inquiries may involve both genuine curiosity about the conceptual issue and a raw pleasure in driving the nearest adult mad. The fact that the question *is* maddening, however, derives from the fact that it is hard.

In practice, the question "what is a car" is resolved via an issue of contingent fact. The answer to the question "Is that a car?" revolves most urgently around the issue of physical danger. Therefore, the implied question is, "am I in the presence of an object that might run me over?" Even an infant knows that a drawing of a car is safe to walk past. With increasing sophistication, a child may infer that cars set up on blocks are almost equally benign. But the reason a child learns to conceptualize the term "car" is that the concept has survival value. For the same reason, children conceptualize "snake" and "stranger" and "food." It's only when asked to think carefully about such terms (perhaps under pressure of an inquiring infant mind) that people notice the fuzziness of these categories.

For most practical purposes, such fuzziness doesn't matter. However, this is a contingent fact, not a logical truth. Imagine a community of intelligent beings who evolved on world composed entirely of miscible liquids. Unlike us, such beings might have no incentive to invent the notion of a set, as it would lack utility in their fluid world. Were we to suggest to them the concept of a set as part of an attempted dialogue on universal truths, they might fail

[ee] To put this in terms of another logical principle, it is not true that an object is either a car or it is not a car. The law of the excluded middle fails in the real world. Here on earth, every dichotomy is a false dichotomy.

to understand what we were getting at. To them, the very proposal of a set might trigger on a serious intellectual level the same questions that an infant asks to madden. From this reaction, we might be forced to conclude that all propositions and inferences based on the concept of set are not matters of logical truth but mental approximations contingent to safe life on earth.

To put it another way, the fact that we intuitively find the concept of "set" meaningful depends on a subconscious waving away of a whole range of difficulties and problems. Therefore, our supposedly "rigorous" and "absolute" logical conclusions only appear so because we exercise a kind of selective neglect, the option to avoid looking closely when it serves us not to.

The ordinary approach to this problem is to treat "real" sets as idealized mental objects. Thus, though we may not know exactly what the set of all cars may be, we have an abstract notion of a perfect set regardless of its specific content. Thus, we can all agree that the set {1,2} is a subset of the set {1,2,3} and contains as one of its elements the object "1."

But *why* is this clear? Is it because humans developed a rough notion of a set to facilitate survival? Or do sets "actually exist" on some "abstract plane?" The latter "platonic" description claims that abstract objects such as sets and their relations possess a type of prior existence independent of instantiations in real life.[57]

Whether or not this may be "true," the platonic description succeeds because, and only because, people widely find it agreeable. Thus, it is contingent upon consensus.

Mathematics

The stubborn boundary problem inherent in defining the set of all cars may be unimportant when it comes to crossing the street safely, but it has consequences for the presumed absolute truth of mathematics. Ever since David Hilbert set out his program of axiomatization in 1899, mathematics has been grounded in set theory.[ff] But if sets are a fuzzy abstraction derived from contingency and consensus, it would seem that any field founded on them must also invoke contingency and consensus. This squares poorly with the near universal belief that mathematics touches on absolute truth.

Consider the most basic form of mathematics, arithmetic. When we say "1 + 1 = 2," we implicitly engage in an activity very like pretending we know how to define "the set of all cars." Thus, while it is frequently the case that one car plus one car equals two cars, what do we say about circumstances where two cars strike one another and one of breaks in half? To say, "well, 1 car plus 1 car equals two cars providing nothing exceptional occurs," is to

[ff] Hilbert's program ran into many problems, including Russel's paradox and Goedel's indecidability theorem, but these have been addressed in part, and it is nonetheless the case that mathematics is currently taught in the language of set theory.

invoke the principle of selective neglect: we shall ignore all cases where the answer doesn't come out the way we want.

The problem become even more vexed if the subject of discussion expands from cars to includes boundary-less objects like gobs of water. In such cases, we tend to abandon the notion of number and shift conceptual ground to the idea of volume. Thus, even though one gob of water when combined with a second gob forms a single new gob, thus failing to satisfy arithmetic when considered in terms of discrete objects, one cup of water plus another cup does equal two cups of water, recovering arithmetic on new grounds.

However, this solution is still contingent. If we combine one cup of water with a cup of alcohol the result will not measure out at two cups of liquid as alcohol and water partially dissolve in one another. Perhaps, then, weight is the proper subject of measurement. But what about in cases when combining two liquids causes a chemical reaction in which much of the mass escapes as gas? In such cases, the apparent failure of arithmetic led early chemists to discover and weigh the lost gas. Again, arithmetic was saved. Weight (or, more precisely, mass) adds perfectly in all closed systems – except when it doesn't. In circumstances where nuclear fission or fusion takes place, mass no longer adds properly. Here, a new principle is invoked to rescue arithmetic, that of conserved addition of mass/energy. Only, this breaks down at high speeds, because in special relativity mass/energy is no longer conserved. Instead, conservation applies to a rather different sort of object, the mass/energy/momentum 4-vector. But even this breaks down over long distances when gravity is involved, as there is no equivalent conservation law at all in general relativity.

Thus, the very concept of number as an experimental principle is really rather sloppy. We seem to invoke excuses frequently and/or change ground precisely in order that we may keep on saying "1 + 1 = 2." An alien race that didn't happen to share the concept of number could be excused for thinking that our behavior suggests an attempt to fudge.

But to most humans, the expression "1 + 1 = 2" seems to represent something *real*, regardless of problems of instantiation. Possibly, numbers have a platonic reality, just like sets. Or, perhaps we "believe" in number because we find it useful to do so, just as we find the concept of "car" useful, despite the boundary problems that arise if we think about it too closely. After all, addition works in a repeatable way in widely varying circumstances in real life, from counting votes (except when chads are involved) to baking cakes. Such marked utility serves to persuade us of the "truth" of arithmetic; this is a performative demonstration akin to the way in which the existence of cell phones persuades us as to the truth of electronics. We are in fact so persuaded that we actively engage in selective neglect of exceptions in order to preserve the concept.

The performative power of mathematics is truly remarkable. Whole books have been written on the subject; the brief discussion that follows is intended primarily to expand the theme of contingency and consensus.

Under many circumstances, weight is perfectly conserved. If we weigh objects separately and add their weights we arrive at the same number as if we were to weigh them together. Except

that it doesn't always work, due to the problem of rounding. In the real world, measuring scales always have limited precision. A scale may be able to distinguish weights to a tenth of a pound, a hundredth of a pound, or a millionth of a pound, but no scale can give us the weight of an object to an unlimited number of decimal points. To put this another way, there is no such thing as a 1 lb weight; there may be a 1.0001 pound weight, or a 1.0000001 lb weight, but nowhere can we find an object that weighs exactly one pound. The moment one thinks one might have stumbled on one, the invention of a more precise measuring instrument will prove this erroneous.

Consider, however, the following question: "Is it *possible* that a found object might by chance, weigh exactly one pound no matter how precise the instrument?" The obvious answer is either "no" or "vanishingly unlikely" depending on the interpretation, but let us look at a prior issue: Is the question at the head of this paragraph even intelligible?

It certainly *seems* to be a reasonable question. On the face of it, "Are there objects that weigh exactly one pound, or aren't there any such objects?" falls in the logical category of A or Not-A. By the standard rules of logic, one or the other but not both should be true. However, a close analysis shows that within the seemingly simple phrasing of the question the highly abstract concept of infinity is encoded, and therefore before we can attempt to answer "yes" or "no" we must first discover what it is that we are talking about.

When we say that an object weighs exactly one pound, we mean that its actual weight in pounds is 1.0000000 …, where the ellipses indicate that the string of 0s is meant to go on *forever*. Clearly, we are now referring to something outside the real world. None of us have ever encountered an infinite sequence of anything. To explore the concept of infinity we must take a long excursion away from the real world and into some kind of abstract plane of reality; what is perhaps most amazing about mathematics is that after taking this excursion we return equipped with real world powers we did not previously possess. That this is possible reinforces our intuitive sense that the abstract world of mathematics is in some sense *true*. But just how it is true -- in what manner such abstract concepts truly "exist" -- remains beyond human capacity to pin down.

For a demonstration of the practical powers to be gained by dwelling for a time in the questionably real world of the abstract, consider the number we call pi. The ancient Egyptians experimentally discovered that if they multiplied the diameter of a circle by 3.14 the resulting number closely matched the measurement that would be obtained by measuring the circumference of the circle directly. However, with large enough circles, or accurate enough measures, we soon find that 3.14 isn't exactly right. The practical solution would be to carefully draw and measure lots of circles and thereby refine the ratio.[gg] Another approach would be to use Calculus, a field of mathematics that employs highly abstract concepts. By means of Calculus, we can calculate pi to any desired number of decimal points, and, what's more, if we try it experimentally we find that Calculus has given us the right answer.

[gg] This assumes that the ratio is a single number; that no matter what the circle, the ratio will remain the same. Again, we can discover this experimentally, by drawing lots of circles. We can also "prove" it mathematically, whatever "proof" means.

Calculus, however, depends on the concept of infinity, and infinity does not exist. Why, then, does it give us the right answer?

Perhaps the concept of "infinity" is not really needed; perhaps we can get by with "very, very big." Regarding calculating pi to a fixed number of decimal points, this is perhaps true.[hh] If so, mathematicians have not actually departed from reality, but only examined it closely.

However, close inspection shows that the minimum level of abstraction necessary to analyze what we mean by "pi" involves abstractions much thornier than "infinity." Calculus, it turns out, makes extensive use of the misnamed "real" numbers, which include both typical numbers (defined in a moment) and "irrational" numbers whose decimal expansion goes on forever without any definite repeats.[ii] In what sense do such numbers "exist?"

For the purposes of discussion, we shall switch to a simpler irrational number, the one commonly referred to as "the square root of 2." The expression in quotes refers to a purported number which, when multiplied by itself, becomes two. I write purported, because the implicit existence-claim is highly problematic. All measurements we make in the world are discovered in terms of ratios of whole numbers: 3/8ths of an inch, 99/100ths, etc. Such typically encountered numbers are called "rational" numbers. But, as far back as Pythagoras, it has been known that it is impossible for any rational number to have the property that when multiplied by itself the result is two.[jj] Therefore, in what sense does sqrt(2) exist?

Perhaps it exists in the sense that, as with pi, it is possible to compute sqrt(2) to any desired number of decimal points. But this turns out to be true for only a small subset of the irrational numbers, the so-called "computable numbers." It can easily be shown that most of the irrational numbers cannot be computed at all. Worse still, they can't even be *described*.[kk]

[hh] Except that such calculation involves concepts such as "lines," "circles" and "length," all of which are abstractions of real world objects that involve infinite perfection. So, maybe not.

[ii] Fractions may have infinite decimal expansions, but after some number of terms they consist of a fixed set of numbers that repeats endlessly. This is less than trivial to prove.

[jj] Sketch of proof, with many details left out: Suppose a/b were such a number. Then, a^2 / b^2 must equal two. A^2 must have an even number of twos in its prime factorization, twice the number of twos in the factorization of a. The same is the case of b^2. Dividing them therefore results in a fraction with an even number of twos in the numerator and in the denominator. Proceed to cancel out the twos in pairs. One must ultimately reach a point where canceling stops. The result will be a factorization of the numerator in which the number of twos remaining is an even number (since canceling involves subtraction of number of factors, and an even number minus an even number is an even number) But the number 2, which is the result of the fraction, has only a single 2 in its factorization. This leads to a contradiction.

[kk] Both computing and describing must be done using a language; the total of all possible expressions in any language with a finite number of symbols is infinite but of a type of infinity called countable; the total "number" of irrational numbers is of a type of infinity called "uncountable." A famous proof by George Cantor (1845-1918) demonstrates that uncountable infinities are much, much larger than countable ones in all important senses of the expression "much, much larger." Therefore, virtually no

If virtually all irrational numbers cannot even be described, it would seem problematic to assert that they exist. However, when mathematicians use Calculus to figure out ways of computing such numbers as pi and sqrt(2), they unavoidably make use of the full realm of irrational numbers. The reason that the set of rational and/or computable numbers must be augmented by these problematic objects is that to perform the calculations in Calculus one must operate on an infinitely smooth line.[ll] The set of rationals is not smooth enough, in this sense, for Calculus to work. However, after working for a while in this imponderable world of non-describable objects, one can obtain formulas that apply to the inherently non-smooth measurements we make in the actual world. The fact that it works at all is rather remarkable.

What if results did not agree with the math? Would we reject the math? Or would we reject our measurements?

Obviously, the latter. If, for example, we measured the circumference of a circle and divided it by the diameter, and the result didn't match the computed value of pi to the expected number of decimal places, we would confidently assume something went wrong with the measurement process. This shows that we regard math as "true" in a fundamental way. Is this simply because it is so often right? If so, mathematics would have the status of a science, in which inductions are made from observation. But, surely mathematics is something other than an experimental science. The abstruse deliberations of mathematics remain in a world of abstraction for a long time, as it were, and do not require any "checking in" with the actual world. The rules of behavior in the land of mathematics, so to speak, are self-contained, and could be performed by a brain in a vat (or, more practically speaking, a computer.) It is we in the real world who make use of mathematics; mathematics doesn't need the real world. This remarkable and apparently inexplicable convergence of pure thought and practical observation is certainly persuasive as to the validity of that form of thought.

But what if we ran into a circumstance where mathematics failed to correlate with the real world? This is not a purely hypothetical concept. It turns out that, if measured carefully enough, the ratio of the circumference of a circle to its diameter is *not* pi for any circle drawn in the vicinity of the earth. The origin of the discrepancy lies in the fact that the mathematical derivation of the formula for pi invokes an imaginary concept, a "perfectly flat surface." The ratio will not turn out to be pi on any non-flat surface, such as the surface of the earth. Furthermore, even if we make our best efforts to create a flat surface separate from the

---

irrational numbers are computable/describable. See, for example, Cantor's Diagonal Argument. Available at http://en.wikipedia.org/wiki/Cantor%27s_diagonal_argument. Accessed 11/1/09

[ll] This is an oversimplification of a complex subject. Calculus as it is normally performed requires a set that contains all its limit points. One can create sequences of rational or computable numbers that converge in the limit on a number that is neither computable or rational However, by definition, the set of real numbers contains all its own limit points. If it were found that some sequence of reals converged on a number not included in the reals, that new number would automatically become part of the reals. To figure out what this definition even means remains something of a challenge, considering that it regards objects that can't be described in any possible language.

surface of the earth, the ratio of the circumference of a circle drawn on that surface to its diameter will still deviate from pi because space (or, rather, "spacetime") is curved. But we do not take this as disproving math. Rather, we see it is seen as confirmation of a somewhat more mathematically complex model of reality, that of General Relativity.[mm]

However, we happen to know that general relativity isn't correct. In many real situations, quantum mechanics gives different, and more accurate predictions, and quantum mechanics is entirely incompatible with general relativity. (Conversely, quantum mechanics falsely predicts that the diameter of a circle divided by its diameter will never vary from pi.) However, since the days of Newton in the late 17th century, physicists have repeatedly experienced the satisfaction of reducing ever larger portions of physical reality to elegant mathematical formulae. This historically contingent fact has provided physicists with a semi-religious faith: that a single elegant mathematical theory will some day be found that represents the whole universe. Consistent with this recognizable human emotion, physicists continue to strive to combine quantum mechanics and general relativity into a single larger theory. String theory is the current candidate. But they may or may not succeed. They might never find the formulae; alternatively, perhaps no such set of formulae exist. Nothing obvious about the universe, nor about mathematics, guarantees that the former can be represented by the latter. And, significantly, nothing about this process strikes fear into the heart of mathematicians, or even ruffles their feathers to the slightest degree. Application to the real world is of no particular significance to mathematics. If it ultimately proves a contingent fact that physicists fail to develop a unified mathematical "theory of everything," mathematicians will scarcely notice.

Thus, we are left with a peculiar set of irreconcilables. On the one hand, mathematics has proved to be a remarkably effective tool for dealing with the world. On the other hand, mathematics apparently invokes a plane of reality entirely separate from the world. Regardless of where or even whether such a plane of reality might exist, human beings have shown a remarkable and entirely cross-cultural ability to agree with one another on how to operate within it.[nn] Thus, even – or, perhaps, especially – in mathematics, human consensus is fundamental, while at the same time its application to the world remains contingent and inexplicable.

[mm] Remarkably, the extent of spacetime curvature near the earth's surface is large enough that current techniques can measure the deviation of real measurements from pi. Perhaps even more remarkably, the true ratio of the circumference of a circle to its diameter in the vicinity of earth can be predicted by very carefully measuring the acceleration of a falling body and applying Einstein's equations.

[nn] This is the case even though it has proved quite difficult to systematically enumerate the "rules of behavior" in mathematics. As it happens, all attempts to enumerate the fundamental principles of mathematics have run into serious problems, and even the most successful ultimate rely on undefinable objects and relations (such as sets and membership in sets) that for no obvious reason make perfectly good sense to human beings.

Endnotes

[1] Armitage P. Fisher, Bradford Hill, and randomization. *International Journal of Epidemiology.* 2003;32:925-928.
[2] Taubes G. Nutrition. The soft science of dietary fat. *Science*. 2001;291:2536-45.
[3] Taubes G. The (political) science of salt. *Science*. 1998;281:898-901, 903-7.
[4] Taubes G. Epidemiology faces its limits. *Science*. 1995;269:164-9.
[5] [No authors listed.] Unique to Women. Menopause and Hormone Therapy. Available at: http://www.ourbodiesourselves.org/book/companion.asp?id=28&compID=101. Accessed 11/1/09
[6] Bratman S. Reflections on the uses of Scientific Authority
(as illustrated by the EPA Standards-Setting Process for Ozone). Final paper. Environmental Medicine, MPH program, Spring Semester 2009. Medical College Wisconsin.
[7] [No Authors Listed.] Entry: Science. Available at http://www.merriam-webster.com/dictionary/science. Accessed 11/1/09.
[8] Feyerabend F. *Against Method*. New York: Verso; 1984.
[9] [No Authors Listed.] Entry: Objective. Available at http://www.merriam-webster.com/dictionary/objective. Accessed 11/1/09.
[10] See, for example:
Miller A. Realism. *Stanford Encyclopedia of Philosophy*. Available at: http://plato.stanford.edu/entries/realism Accessed 11/1/09
Boyd R. Scientific Realism. *Stanford Encyclopedia of Philosophy*. Available at: http://plato.stanford.edu/entries/scientific-realism Accessed 11/1/09
Ladyman J. Structural Realism http://plato.stanford.edu/entries/structural-realism/
[11] Davey Smith G, Ebrahim S. Epidemiology--is it time to call it a day? *Int J Epidemiol*. 2001;30:1-11.
[12] Taubes, G. Do we really know what makes us healthy? *NY Times Magazine*. September 30, 2007. Available at http://www.nytimes.com/2007/09/16/magazine/16epidemiology-t.html. Accessed 11/1/09.
[13] Mill JS. System of Logic, Ratiocinative and Inductive. Public Domain. 1843. Available online at http://www.gutenberg.org/etext/27942. Accessed 11/1/09
[14] Snow J. On the Mode of Communication of Cholera. (Public Domain) 1855. Available at: http://www.ph.ucla.edu/EPI/snow/snowbook.html. Accessed 11/1/09.
[15] Hitchcock C. Probabilistic Causation. *Stanford Encyclopedia of Philosophy*. Available at: http://plato.stanford.edu/entries/causation-probabilistic. Accessed 11/1/09
[16] Pearl J. Statistics and Causal Inference: A Review. *Test.* 2003;12:281–345
[17] Angrist J. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press. 2008
[18] Austin Bradford Hill, "The Environment and Disease: Association or Causation?,"
Proceedings of the Royal Society of Medicine, 58 (1965), 295-300. http://www.edwardtufte.com/tufte/hill
[19] Rothman K, Greenlan S, Lash T. *Modern Epidemiology, Third Edition*. Philadelphia: Lippincott, Williams & Wilkins. 2008:24
[20] Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med*. 2000;342:1887-92.
[21] Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. *N Engl J Med.* 2000;342:1878-86.
[22] The following were published together"
Friedman HS. Observational studies and randomized trials. *N Engl J Med*. 2000;343:1195-6; author reply 1196-7.
Kunz R, Khan KS, Neumayer HH. Observational studies and randomized trials. *N Engl J Med.* 2000;343:1194-5; author reply 1196-7.
Pocock SJ, Elbourne DR. Randomized trials or observational tribulations? *N Engl J Med.* 2000;342:1907-9. Author reply 1196-7
In addition, there were multiple letters to the editor on this topic in the same issue.

[23] Wilson R. *Feminine Forever*. NY: M. Evans and Company, Inc. 1968.
[24] Stevens R. Rosenberg C, Lawton Burns. *History and health policy in the United States: putting the past back in*. New Brunswick: Rutgers University Press. 2006:162.
[25] The Writing Group for the PEPI Trial. Effects of estrogen or estrogen/progestin regimens on heart disease risk factors in postmenopausal women. The Postmenopausal Estrogen/Progestin Interventions (PEPI) Trial. *JAMA*. 1995;273:199-208.
[26] Barrett-Connor E, Stuenkel CA. Hormone replacement therapy (HRT)--risks and benefits. *Int J Epidemiol*. 2001;30:423-6.
[27] Stampfer M, Grodstein F. Cardioprotective effect of hormone replacement therapy. Is not due to selection bias. *BMJ*. 1994;309:808-9.]
[28] Petitti DB, Perlman JA, Sidney S. Postmenopausal estrogen use and heart disease. *New Engl J Med* 1986;315:131–32.
[29] Stevens R. Rosenberg C, Lawton Burns. History and health policy in the United States: putting the past back in. New Brunswick: Rutgers University Press. 2006:161-171
[30] Hulley S, Grady D, Bush T, et al. Randomized Trial of Estrogen Plus Progestin for Secondary Prevention of Coronary Heart Disease in Postmenopausal Women. *JAMA*. 1998;280:605-613.
[31] Rossouw JE, Anderson GL, Prentice RL et al. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results From the Women's Health Initiative randomized controlled trial. *JAMA*. 2002;288:321-33.
[32] Craig MC, Maki PM, Murphy DG. The Women's Health Initiative Memory Study: findings and implications for treatment. *Lancet Neurol*. 2005;4:190-4.
[33] Prentice RL. Data analysis methods and the reliability of analytic epidemiologic research. *Epidemiology*. 2008;19:785-8; discussion 789-93.
[34] Vandenbroucke JP. The HRT controversy: observational studies and RCTs fall in line. *Lancet*. 2009;373:1233-5.
[35] Prentice R. Data Analysis Methods and the Reliability of Analytic Epidemiologic Research. *Epidemiology*. 2008;19:785-788
[36] Petitti D. Commentary: Hormone replacement therapy and coronary heart disease: four lessons. *International Journal of Epidemiology*. 2004;33:461–463
[37] Vandenbroucke JP. When are observational studies as credible as randomised trials? *Lancet*. 2004;363:1728-31
[38] Dawber T, Nickerson R, Brand F, et al. Eggs, serum cholesterol, and coronary heart disease. *American Journal of Clinical Nutrition*. 1982;36:617-625.
[39] Mensink RP, Zock PL, Kester AD et al. Effects of dietary fatty acids and carbohydrates on the ratio of serum total to HDL cholesterol and on serum lipids and apolipoproteins: a meta-analysis of 60 controlled trials. *Am J Clin Nutr*. 2003;77:1146-55.
[40] Sanders TA. High- versus low-fat diets in human diseases. *Curr Opin Clin Nutr Metab Care*. 2003;6:151-5.
[41] Hu FB, Willett WC. Optimal diets for prevention of coronary heart disease. *JAMA*. 2002;288:2569-78.
[42] Mente A, de Koning L, Shannon HS et al. A systematic review of the evidence supporting a causal link between dietary factors and coronary heart disease. *Arch Intern Med*. 2009;169:659-69.
[43] Turnbull F. Effects of different blood-pressure-lowering regimens on major cardiovascular events: results of prospectively-designed overviews of randomised trials. *Lancet*. 2003;362:1527-35
[44] Studer M, Briel M, Leimenstoll B, et al. Effect of different antilipidemic agents and diets on mortality: a systematic review. *Evid Based Cardiovasc Med*. 2005;9:237-40.
[45] Muldoon MF, Manuck SB, Mendelsohn AB, et al. Cholesterol reduction and non-illness mortality: meta-analysis of randomised clinical trials. *BMJ*. 2001 Jan 6;322(7277):11-15.
[46] Taylor AJ. Given the ENHANCE trial results, ezetimibe is still unproven. *Cleve Clin J Med*. 2008;75:497-8, 502, 505-6.
[47] Pauriah M, Struthers AD, Lang CC. Biomarkers and surrogate endpoints in cardiovascular therapeutics research: under scrutiny following results of the ENHANCE Study. *Cardiovasc Ther*. 2008;26:85-8.
[48] Suckling K. The ENHANCE Study: an unusual publication of trial data raises questions beyond ezetimibe. *Expert Opin Pharmacother*. 2008;9:1067-70
[49] Mandell BF. A drug, a concept, and a clinical trial on trial. *Cleve Clin J Med*. 2008;75:470

[50] Bjelakovic G, Nikolova D, Gluud LL, Simonetti RG, Gluud C. Antioxidant supplements for prevention of mortality in healthy participants and patients with various diseases. *Cochrane Database of Systematic Reviews.* 2008;(2):CD007176.
[51] Bardia A, Tleyjeh IM, Cerhan JR et al. Efficacy of antioxidant supplementation in reducing primary cancer incidence and mortality: systematic review and meta-analysis. *Mayo Clin Proc*. 2008;83:23-34.
[52] Lichtenstein AH. Nutrient supplements and cardiovascular disease: a heartbreaking story. *J Lipid Res*. 2009;50 Suppl:S429-33.
[53] Albanes D, Heinonen OP, Huttunen JK, et al. Effects of alpha-tocopherol and beta-carotene supplements on cancer incidence in the Alpha-Tocopherol Beta-Carotene Cancer Prevention Study. *Am J Clin Nutr*. 1995;62(6 suppl):1427S-1430S.
[54] Vann M. Fresh Vegetables, Fruits Reduce Diabetes Risk. http://tulane.edu/news/newwave/080108_diabetes_risk.cfm
[55] Vandenbroucke JP. When are observational studies as credible as randomised trials? *Lancet*. 2004;363:1728-31.
[56] Math Academy Online™ / Platonic Realms™. Carroll's Paradox. Available at: http://www.mathacademy.com/pr/prime/articles/carroll/index.asp. Accessed 11/1/09
[57] See, for example, Miller A. Realism. Stanford Encyclopedia of Philosophy. Available at http://plato.stanford.edu/entries/realism/. Accessed 11/1/09