ELSEVIER

Why Are So Many Epidemiology Associations Inflated or Wrong? Does Poorly Conducted Animal Research Suggest Implausible Hypotheses?

MICHAEL B. BRACKEN, PHD

There is growing concern among epidemiologists that most discovered associations are either inflated or false. The reasons for this concern have focused on methodological issues in the conduct and publication of epidemiologic research. This commentary suggests that another reason for discrepant findings may be that animal research is producing implausible hypotheses. Many animal studies are methodologically weak, and the animal literature is not systematically reviewed and synthesized. Moreover, most bodies of animal literature may be so heterogeneous that they can be used selectively to support the plausibility of almost any epidemiology study result. Epidemiologists themselves also do not consistently conduct systematic reviews of bodies of biological evidence which might point to sources of bias in an evidence base. Animal research will likely continue to provide the biological basis for epidemiological investigation, but substantial improvement is needed in how it is conducted and synthesized to improve the predictability of animal studies for the human condition.
Ann Epidemiol 2009;19:220–224.

KEY WORDS: Animal Studies, Bias, Epidemiology Methods, Randomized Trials, Research Synthesis, Systematic Reviews.

> "Experiments should be carried out on the human body…the quality of the medicine might mean that it would affect the human body differently from the animal body"
>
> Ibn Sina 1012 CE, 402 AH

Why epidemiology has so much difficulty documenting valid and replicable associations has been widely discussed for several years (1–5). Recent observations from large randomized controlled trials (RCTs) have impressively refuted some of epidemiology's most long-standing conclusions, often made from very large and highly publicized observational studies, and the question persists as to why so much observational epidemiology is not replicated by randomized evidence (6–10).

In a recent commentary and discussion (11–14), Ioannidis suggested several reasons why most discovered true associations are inflated, including the use of thresholds of statistical significance, especially in underpowered studies; the many data manipulations used in variable construction and statistical analysis; and biases in the publication process (11). Elsewhere, Ioannidis (15) has used similar argument to suggest that most published research findings are false. These observations are important and provocative and challenge many of the assumptions underlying the validity of the epidemiologic literature. In a thoughtful response, Willett (14) suggests that "those who practice epidemiology understand that the primary research mode is still the development of testable hypotheses based on sound biological reasoning" (p. 655). This raises the question: How sound is the biological evidence from which hypotheses tested in epidemiology are derived? Are the vulnerabilities observed in epidemiologic investigations also found in the biology research base? If epidemiologists are testing implausible hypotheses derived from a poorly validated body of animal research, which is further amplified by publication bias, this may be another reason why discovered associations will be inflated or false.

DO ANIMAL STUDIES PREDICT EPIDEMIOLOGICAL ASSOCIATIONS?

The concept that animal research, particularly that relating to diet, pharmaceuticals, and environmental agents, may be a poor predictor of human experience is not new. A thousand years ago, Ibn Sina commented on the need to study humans rather than animals (16) and Alexander Pope's dictum "The proper study of mankind is man" is well known and widely cited (17). Pharmacologists in particular have long recognized the difficulties inherent in extrapolating drug data from animals to man (18, 19). Given the large number of animal studies conducted, it would be expected that some animal experiments do predict some human reactions; for example, penicillin was observed to protect both mice and humans from *Staphylococcus* infections (20), and Accutane (isotretinoin) causes birth defects in rabbits,

From the Schools of Public Health and Medicine, Yale University, New Haven, CT.
Received October 7, 2008; accepted November 29, 2008.

360 Park Avenue South, New York, NY 10010
1047-2797/09/$–see front matter
doi:10.1016/j.annepidem.2008.11.006

monkeys, and humans (but not in mice and rats) (21). However, corticosteroids are widely teratogenic in animals but not in humans (22), whereas thalidomide is not a teratogen in many animal species but it is in humans (23). Recent experience in a phase 1 study of the monoclonal antibody TGN 1412 resulted in life-threatening morbidity in all six healthy volunteers, reflecting inadequate prediction, even in non-human primates, of the human response (24).

It is has been known for some time that many animal experiments are poorly designed, conducted, and analyzed and that this may be one reason why they often do not translate into replication in human therapeutic trials (25–27) or into cancer chemoprevention. Some human carcinogens were predicted in animal studies (aflatoxins, benzene, diethylstilbestrol, vinyl chloride), but other agents were positive in animal studies but not in human studies (acrylamide, alar, cyclamate, red dye #2, saccharin) (28–30). It has only recently been observed that most of the animal literature is also inadequately reviewed and summarized and this too may contribute to failure to replicate animal research in humans. In one survey, only 1 in 10,000 MEDLINE records of animal studies were tagged as being meta-analyses versus 1 in 1,000 for human research (31). However, this research often provides the rationale for hypotheses studied by epidemiologists. In recent reports, the poor quality of research synthesis was documented by a comprehensive search of MEDLINE, which found only 25 systematic reviews of animal research despite there being several million individual studies in citation databases (32). Other recent studies similarly found only 30 (33) and 57 (34) systematic reviews of any type of animal research. One recent study of the health effects associated with low-dose Bisphenol A in human urine (35) conflicts with the systematic review of the rodent studies that found little evidence for any health associations (36).

Systematic review of animal studies is well advanced in the field of stroke research (37), an area where almost no new human therapies have been developed despite decades of experimental and human study. In one systematic review of FK506 used for experimental stroke, in which 29 separate studies were found in the literature, only one study blinded investigators to the intervention and two blinded them for the outcome assessment; none met all 10 quality criteria established by the reviewers (one study met no criteria and the highest score was 7). Meta-analysis of the animal FK506 studies demonstrated a strong trend for the methodologically weakest studies to show the strongest protective effects and the methodologically strongest studies to show no (or weak) protective effects (38).

The limited number of systematic reviews of the animal literature that have been done point to the poor quality of animal research and the difficulty of extrapolating from it to humans (39), a concern increasingly being made in other fields of drug discovery (40, 41). Some key problems are summarized from Pound et al. (32):

- Disparate animal species and strains, with a variety of metabolic pathways and drug metabolites, leading to variation in efficacy and toxicity
- Different models for inducing illness or injury with varying similarity to the human condition
- Variations in drug dosing schedules and regimen of uncertain relevance to the human condition
- Variability in animals for study, methods of randomization, choice of comparison therapy (none, placebo, vehicle)
- Small experimental groups with inadequate power, simple statistical analysis that does not account for confounding, and failure to follow intention to treat principles
- Nuances in laboratory technique that may influence results (e.g., methods for blinding investigators) may be neither recognized nor reported
- Selection of outcome measures, may be disease surrogates or precursors, of uncertain relevance to the human clinical condition
- Length of follow up varies and may not correspond to disease latency in humans

The quality of in-vitro research and review, much of which is closely tied to animal experimentation, has been even less formally studied. In one rare study of how in-vitro research is reviewed, a total of only 45 systematic reviews of any type of bench study was found (33).

The poor quality of much animal and in-vitro research poses substantial difficulty for epidemiologists who use "biologic plausibility" as one of their guidelines for inferring causality (42). A discussion of biological mechanisms, usually relying on animal research, is quite common in reports of epidemiological association. However, it seems that animal research on almost any topic of epidemiologic interest is so heterogeneous and inadequately synthesized that it is possible to selectively assemble a body of evidence from the animal and in-vitro studies that support almost any epidemiologic result.

PUBLICATION BIAS

In contrast to large epidemiological projects, the smaller scale of animal experiments, often from individual laboratories, would suggest greater opportunity for publication bias. Publication bias has been well documented in the randomized trial literature and has been attributed to a range of biases: authors being more likely to write up positive results and to send their manuscripts reporting positive results to higher profile journals, to journal editors being more likely to accept positive results and to publish them early (43,

AEP Vol. 19, No. 3
March 2009: 220–224

44). There has been little formal study of publication bias in observational epidemiology, or in animal and in-vitro research (45, 46) and the few studies that have been done have not found evidence of publication bias (47, 48). Nonetheless, its documentation in the more transparent circumstances of RCTs suggests that publication bias must be a common phenomenon in observational epidemiology, animal, and in-vitro studies. Failure to systematically review these bodies of evidence together with publication bias in the literature base provide the opportunity for substantial bias and misleading results in the animal literature used to create hypotheses for testing in epidemiological studies.

Publications in genetic epidemiology, where up to a million single nucleotide polymorphism associations are examined (49), have provided an opportunity to observe how publication bias operates in this area of observational epidemiology. Ioannides et al (50) have documented the early publication of extreme genetic associations (those suggesting both higher risk and protective genes), whereas later studies, often of higher quality and on larger samples, report smaller effects or do not show any association with the same genotype. In another comparative study, 20 candidate genes previously significantly associated with atorvastatin could not be replicated in a genome-wide association study. Only one was found to be statistically significantly associated and eight showed opposite directions of effect (51). Not only do candidate genes represent a very small fraction of the genome, they are often based on animal models which, while they may represent genes conserved in humans, have different RNA, proteins, gene interactions, and other epigenetic characteristics (52). Moreover, the animal phenotype is not always analogous to the human phenotype (53), all of which may make animal genetic studies uncertain predictors of human genetic associations. Systematic review of the murine models for amyotrophic lateral sclerosis and other neurodegenerative diseases have recently identified major design flaws (54), including genetic heterogeneity even in inbred littermates so that the designed phenotype may be lost (55).

OUTCOME REPORTING BIAS

Bias in reporting the primary outcome is a recently documented phenomenon in randomized trials. Chan et al. (56) showed major discrepancies between declared primary outcomes in randomized trial protocols from what was published as the primary outcome in the same study. Overall, 62% of trials were discrepant between the protocol and the published primary outcome, with trials changing the proposed primary to secondary, completely ignoring (and not mentioning) the proposed primary outcome in the publication, introducing a primary outcome that was a protocol secondary outcome, or reporting a primary outcome not mentioned in the protocol. Outcome reporting has not been systematically studied in observational epidemiology or in animal and in-vitro experimentation, but, given the absence of specificity often found in observational epidemiology or animal protocols and the lack of registration of protocols (compared to what is now expected of randomized trial protocols (57)), it seems highly likely that outcome bias is a problem in these areas of research. The influence of choice of referent group (or "comparator") has also been studied most formally in the RCT literature (58, 59), raising concern that similar sources of bias occur in observational epidemiology and in animal research.

CONCLUSIONS

Animal research will likely continue to be an important component of the biological underpinnings of hypothesis development in epidemiology; therefore epidemiologists have a vested interest in ensuring that the research they rely on is as valid as possible and that it has been systematically reviewed. Given the likelihood that some epidemiology studies may be testing implausible hypotheses, what measures can be taken to improve this aspect of our science?

- More rigorous animal experiments and their systematic review should lead to more valid hypotheses for epidemiological investigation. While one would hope that bench scientists would learn to do systematic reviews themselves, they are likely to need the help of epidemiologists trained in systematic reviewing. Epidemiologists who depend on animal research may themselves need to conduct systematic reviews of the animal research they rely on both for hypothesis development and when using animal research to understand the biological plausibility of their research findings. All too often, animal research may be selectively reported to support epidemiological observations rather than by reference to a systematic review of the totality of animal evidence.
- Ensure that systematic reviewing methodology is part of the education of epidemiologists and that it is routinely practiced. This will lead to more valid and unbiased summaries of the state of biological and epidemiological knowledge.

I am grateful to Iain Chalmers for his comments on an early draft of this paper. Arienne Hoey provided technical assistance with the manuscript.

REFERENCES

1. Taubes G. Epidemiology faces its limits. Science. 1995;269:164–169.
2. Bracken MB. Alarums false, alarums real: challenges and threats to the future of epidemiology. Ann Epidemiol. 1998;8:79–82.

3. Davey Smith G, Ebrahim S. Epidemiology—is it time to call it a day? Int J Epidemiol. 2001;30:1–11.
4. von Elm E, Egger M. The scandal of poor epidemiological research. BMJ. 2004;329:868–869.
5. Shapiro S. Looking to the 21st century: have we learned from our mistakes, or are we doomed to compound them? Pharmacoepidemiol Drug Saf. 2004;13:257–265.
6. Bonovas S, Filioussi K, Sitaras NM. Statin use and the risk of prostate cancer: a metaanalysis of 6 randomized clinical trials and 13 observational studies. Int J Cancer. 2008;123:899–904.
7. Furlan AD, Tomlinson G, Jadad AA, Bombardier C. Methodological quality and homogeneity influenced agreement between randomized trials and nonrandomized studies of the same intervention for back pain. J Clin Epidemiol. 2008;61:209–231.
8. Martinez ME, Marshall JR, Giovannucci E. Diet and cancer prevention: the roles of observation and experimentation. Nat Rev Cancer. 2008 Aug 7 [Epub ahead of print].
9. Rosano GM, Vitale C, Lello S. Postmenopausal hormone therapy: lessons from observational and randomized studies. Endocrine. 2004;24:251–254.
10. Wolfe F, Michaud K, Dewitt EM. Why results of clinical trials and observational studies of antitumour necrosis factor (anti-TNF) therapy differ: methodological and interpretive issues. Ann Rheum Dis. 2004;63(Suppl 2):ii13–ii17.
11. Ioannidis JP. Why most discovered true associations are inflated. Epidemiology. 2008;19:640–648.
12. Kraft P. Curses—winner's and otherwise—in genetic epidemiology. Epidemiology. 2008;19:649–651; discussion 657–668.
13. Senn S. Transposed conditionals, shrinkage, and direct and indirect unbiasedness. Epidemiology. 2008;19:652–654; discussion 657–658.
14. Willett WC. The search for truth must go beyond statistics. Epidemiology. 2008;19:655–656; discussion 657–658.
15. Ioannidis JP. Why most published research findings are false. PLoS Med. 2005;2:e124.
16. Ibn Sina from the James Lind Library. Available at: http://www.jameslindlibrary.org/. Accessed August 2008.
17. Gold H. The proper study of mankind is the man. Am J Med. 1952;12:619–620.
18. Lasagna L. The diseases drugs cause. Perspect Biol Med. 1964;7:457–470.
19. Brodie BB. Symposium on clinical drug evaluation and human pharmacology. VI. Difficulties in extrapolating data on metabolism of drugs from animal to man. Clin Pharmacol Ther. 1962;3:374–380.
20. Florey HW, Abraham EP. The work on penicillin at Oxford. J Hist Med Allied Sci.. 1951;6:302–317.
21. Nau H. Teratogenicity of isotretinoin revisited: species variation and the role of all-*trans*-retinoic acid. J Am Acad Dermatol. 2001;45:S183–187.
22. Needs CJ, Brooks PM. Antirheumatic medication in pregnancy. Br J Rheumatol. 1985;24:282–290.
23. Lepper ER, Smith NF, Cox MC, Scripture CD, Figg WD. Thalidomide metabolism and hydrolysis: mechanisms and implications. Curr Drug Metab. 2006;7:677–685.
24. Stebbings R, Findlay L, Edwards C, Eastwood D, Bird C, North D, et al. "Cytokine Storm" in the Phase I trial of monoclonal antibody TGN1412: better understanding the causes to improve preclinical testing of immunotherapeutics. J Immunol. 2007;179:3325–3331.
25. Hackam DG, Redelmeier DA. Translation of research evidence from animals to humans. JAMA. 2006;296:1731–1732.
26. Perel P, Roberts I, Sena E, Wheble P, Briscoe C, Sandercock P, et al. Comparison of treatment effects between animal experiments and clinical trials: systematic review. BMJ. 2007;334:197.
27. Roberts I, Kwan I, Evans P, Haig S. Does animal experimentation inform human healthcare? Observations from a systematic review of international animal experiments on fluid resuscitation. BMJ. 2002;324:474–476.
28. Corpet DE, Pierre F. How good are rodent models of carcinogenesis in predicting efficacy in humans? A systematic review and meta-analysis of colon chemoprevention in rats, mice and men. Eur J Cancer. 2005;41:1911–1922.
29. Corpet DE, Pierre F. Point: From animal models to prevention of colon cancer. Systematic review of chemoprevention in min mice and choice of the model system. Cancer Epidemiol Biomarkers Prev. 2003;12:391–400.
30. America's War on Carcinogens: reassessing the use of animal tests to predict human cancer riskAmerican Council for Science and Health; 2006.
31. Sandercock P, Roberts I. Systematic reviews of animal experiments. Lancet. 2002;360:586.
32. Pound P, Ebrahim S, Sandercock P, Bracken MB, Roberts I. Where is the evidence that animal research benefits humans? BMJ. 2004;328:514–517.
33. Mignini LE, Khan KS. Methodological quality of systematic reviews of animal studies: a survey of reviews of basic research. BMC Med Res Methodol. 2006;6:10.
34. Peters JL, Sutton AJ, Jones DR, Rushton L, Abrams KR. A systematic review of systematic reviews and meta-analyses of animal experiments with guidelines for reporting. J Environ Sci Health B. 2006;41:1245–1258.
35. Lang IA, Galloway TS, Scarlett A, Henley WE, Depledge M, Wallace RB, et al. Association of urinary bisphenol A concentration with medical disorders and laboratory abnormalities in adults. JAMA. 2008;300:1303–1310.
36. Goodman JE, McConnell EE, Sipes IG, Witorsch RJ, Slayton TM, Yu CJ, et al. An updated weight of the evidence evaluation of reproductive and developmental effects of low doses of bisphenol A. Crit Rev Toxicol. 2006;36:387–457.
37. Sena E, van der Worp HB, Howells D, Macleod M. How can we improve the pre-clinical development of drugs for stroke? Trends Neurosci. 2007;30:433–439.
38. Macleod MR, O'Collins T, Horky LL, Howells DW, Donnan GA. Systematic review and metaanalysis of the efficacy of FK506 in experimental stroke. J Cereb Blood Flow Metab. 2005;25:713–721.
39. Bebarta V, Luyten D, Heard K. Emergency medicine animal research: does use of randomization and blinding affect the results? Acad Emerg Med. 2003;10:684–687.
40. Kenter MJ, Cohen AF. Establishing risk of human experimentation with drugs: lessons from TGN1412. Lancet. 2006;368:1387–1391.
41. Sundstrom L. Thinking inside the box. To cope with an increasing disease burden, drug discovery needs biologically relevant and predictive testing systems. EMBO Rep. 2007;8 Spec No:S40–43.
42. Hill AB. The environment and disease: association or causation? Proc R Soc Med. 1965;58:295–300.
43. Stern JM, Simes RJ. Publication bias: evidence of delayed publication in a cohort study of clinical research projects. BMJ. 1997;315:640–645.
44. Ioannidis JP, Contopoulos-Ioannidis DG. Reporting of safety data from randomised trials. Lancet. 1998;352:1752–1753.
45. Neitzke U, Harder T, Schellong K, Melchior K, Ziska T, Rodekamp E, et al. Intrauterine growth restriction in a rodent model and developmental programming of the metabolic syndrome: a critical appraisal of the experimental evidence. Placenta. 2008;29:246–254.
46. Macleod MR, O'Collins T, Howells DW, Donnan GA. Pooling of animal experimental data reveals influence of study design and publication bias. Stroke. 2004;35:1203–1208.
47. Juutilainen J, Kumlin T, Naarala J. Do extremely low frequency magnetic fields enhance the effects of environmental carcinogens? A meta-analysis of experimental studies. Int J Radiat Biol. 2006;82:1–12.
48. Dirx MJ, Zeegers MP, Dagnelie PC, van den Bogaard T, van den Brandt PA. Energy restriction and the risk of spontaneous mammary tumors in mice: a meta-analysis. Int J Cancer. 2003;106:766–770.
49. Bracken MB, DeWan A, Hoh J. Genome wide association studies. In: Rebbeck TR, Ambrosome CB, Shields PG, eds. Fundamentals of molecular epidemiology. New York: Taylor & Francis; 2008 pp. 225–238.

50. Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG. Replication validity of genetic association studies. Nat Genet. 2001;29:306–309.
51. Thompson JF, Man M, Johnson KJ, Wood LS, Lira ME, Lloyd DB, et al. An association study of 43 SNPs in 16 candidate genes with atorvastatin response. Pharmacogenomics J. 2005;5:352–358.
52. Williams SM, Haines JL, Moore JH. The use of animal models in the study of complex disease: all else is never equal or why do so many human studies fail to replicate animal findings? Bioessays. 2004;26:170–179.
53. Wojczynski MK, Tiwari HK. Definition of phenotype. Adv Genet.. 2008;60:75–105.
54. Scott S, Kranz JE, Cole J, Lincecum JM, Thompson K, Kelly N, et al. Design, power, and interpretation of studies in the standard murine model of ALS. Amyotroph Lateral Scler. 2008;9:4–15.
55. Schnabel J. Neuroscience: standard model. Nature. 2008;454:682–685.
56. Chan AW, Hrobjartsson A, Haahr MT, Gotzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. JAMA. 2004;291:2457–2465.
57. Krleza-Jeric K, Chan AW, Dickersin K, Sim I, Grimshaw J, Guud C. Principles for international registration of protocol information and results from human trials of health related interventions: Ottawa statement (part 1). BMJ. 2005;330:956–958.
58. Montori VM, Jaeschke R, Schunemann HJ, Bhandara M, Brozek JL, Devereaux PJ, et al. Users' guide to detecting misleading claims in clinical research reports. BMJ. 2004;329:1093–1096.
59. Mann H, Djulbegovic B. Why comparisons must address genuine uncertainties. James Lind Library. Available at: www.jameslindlibrary.org/essays/bias/comparator_bias.html. Accessed Jan 7, 2009.