False Alarms and Pseudo-Epidemics
The Limitations of Observational Epidemiology

David A. Grimes, MD, and Kenneth F. Schulz, PhD, MBA

Most reported associations in observational clinical research are false, and the minority of associations that are true are often exaggerated. This credibility problem has many causes, including the failure of authors, reviewers, and editors to recognize the inherent limitations of these studies. This issue is especially problematic for weak associations, variably defined as relative risks (RRs) or odds ratios (ORs) less than 4. Such associations, commonly reported in the medical literature, are more likely to be attributable to bias than to causal association. All observational research has bias (which can include selection, information, and confounding bias). Hence, detection of small associations falls below the discriminatory ability of observational studies. In general, unless RRs in cohort studies exceed 2 to 3 or ORs in case-control studies exceed 3 or 4, associations in observational research findings should not be considered credible. However, these guidelines are not foolproof: strong (yet spurious) associations can result when large amounts of bias are present. Only in a properly performed randomized controlled trial, free of bias, should small associations merit attention. Better training and more circumspection on the part of investigators, tougher editorial standards on the part of journals, and hefty skepticism on the part of referees and readers are necessary to avoid the dangers of false alarms, pseudo-epidemics, and their unfortunate consequences.

From the Department of Obstetrics and Gynecology, University of North Carolina School of Medicine, Chapel Hill, and FHI 360, Research Triangle Park, North Carolina.

Corresponding author: Dr. David A. Grimes, Department of Obstetrics and Gynecology, CB #7570, UNC School of Medicine, Chapel Hill, NC 27599-7570; e-mail: david_grimes@med.unc.edu.

Financial Disclosure
Dr. Grimes serves as a consultant to Bayer and Merck. The other author did not report any potential conflicts of interest.

© 2012 by The American College of Obstetricians and Gynecologists. Published by Lippincott Williams & Wilkins. ISSN: 0029-7844/12

False alarms and pseudo-epidemics are common in observational research. The net effect is a modern “Chicken Little syndrome,” an “epidemic of apprehension” about the “menace of daily life.”1 Perhaps this nervousness is justified, because none of us gets out of this alive.2 Regrettably, most published research findings, ranging from randomized controlled trials to molecular studies, are wrong.3 Indeed, “any claim coming from an observational study is most likely to be wrong.”4 Of the reported associations that are correct, most are exaggerated.5 Few appreciate how shaky is the scientific foundation on which clinicians practice and on which the public makes decisions.1 By the late 1980s, epidemiologists had noted contradictory findings in published case-control studies on 56 different topics.6 More recently, researchers identified 12 randomized controlled trials that tested 52 claims from observational studies. None of the claims could be corroborated and, ironically, for five of the 52 claims the treatment effect was statistically significant in the opposite direction.4

Some critics claim that observational epidemiology may have outlived its usefulness.7 8 This credibility problem has many causes. First, academicians have an inappropriate imperative to publish. Numbers of publications become an index of academic achievement, emphasizing quantity over quality. With approximately 25,000 biomedical journals in circulation, almost any manuscript can get published somewhere. As noted decades ago by Altman,9 what is needed is less research, better research, and research performed for the right reasons.

Second (ironically), most clinical investigators have had no formal training in what they are doing. No hospital would grant operating room privileges to a physician without satisfactory completion of a surgical residency or, for those in training, supervision by a qualified surgeon. Commercial airlines and the Federal Aviation Administration require formal training and credentialing for pilots. This is not so for medical researchers. Most physician-investigators
have a brief (and frequently inadequate) exposure to research methods too early in their medical education\textsuperscript{10,11} and no formal training thereafter in how to perform clinical research.

Third, the proliferation of large administrative databases has aggravated these problems.\textsuperscript{12} With the ubiquitous availability of computers, countless investigators around the world pore over these computerized files in search of something publishable. Entire journals are dedicated to this enterprise; several indexed in PubMed shamelessly have “data mining” in their titles. Akin to the “emperor’s new clothes,” data-dredging\textsuperscript{1} enjoys a false respectability.

Fourth, immodesty compounds the problem; most researchers are overzealous in the interpretation of their results.\textsuperscript{13} The desire to have an exciting, newsworthy discovery usually eclipses the requirement for circumspection in interpreting observational research replete with bias.\textsuperscript{8,14} The emergence of “risk factor” epidemiology, untethered by either an a priori hypothesis or a concern about biological plausibility, has damaged public health and the credibility of all medical research. “Fishing expeditions” in occupational cancer studies are three-times more likely to produce false-positive associations than are targeted studies.\textsuperscript{5} In many studies, 20\% or more of reported findings are false-positives.\textsuperscript{15} Researchers fail to account for the problem of multiple testing, bias, and multiple modeling.\textsuperscript{4}

Journals are culpable as well; uncritical acceptance of manuscripts of poor quality or with exaggerated claims is routine.\textsuperscript{10,15,16} As recently noted by Sir Iain Chalmers, medical journals today publish “a great deal of rubbish.”\textsuperscript{17} Deficiencies in both the methods of analysis and reporting of outcomes are common, even in mainstream general medical journals.\textsuperscript{16} In response, the STROBE guidelines\textsuperscript{18} are intended to improve the reporting of observational research, analogous to the CONSORT guidelines\textsuperscript{19} for randomized trials.

“Some consequences of erroneous results, especially when published in excellent journals, are truly baneful.”\textsuperscript{20} One predictable consequence is an epidemic of sensational, often misleading, media coverage of flawed studies\textsuperscript{21} (Fig. 1). The second wave of the epidemic spills over into tort law, with often tragic consequences. An example is the $2.35 billion settlement (http://www.now.org/issues/health/060104settlement.html) concerning silicone breast implants, despite the scientific evidence refuting an association with autoimmune disease.\textsuperscript{22,23}

Observational epidemiology has a checkered history. Although it has led to important discoveries, such as the protection against myocardial infarction with aspirin intake and the risk of lung cancer from smoking, it also has reported legions of dangerous false-positive associations and spurious epidemics. “There is a crisis of credibility facing epidemiology...
today, brought about by the barrage of studies with less than optimal methods and conflicting results” (Fig. 2). Some of these bogus associations persisted in the medical literature long after they were debunked. In drug–event associations, these false claims have been termed “phantom ships.” Some of these “ships” wreaked havoc before being torpedoed by better science. For example, observational epidemiology fostered a rich mythology about putative nonspecific side effects of oral contraceptives. In contrast, randomized placebo-controlled trials find these nonspecific symptoms to be approximately as common with placebos.

Well-documented errors include common exposures in daily life (Table 1). Observational epidemiology suggested that both suicide and homicide are associated with cigarette smoking. This association can be explained by the distress leading to suicide being linked with smoking. Epidemiologic studies consistently showed that beta-carotene is associated with lowering of the risk of cancer, especially lung cancer. However, that potential benefit was soundly refuted by randomized controlled trials; once again, experimentation trumped observation. Strong, consistent, observational epidemiology found menopausal estrogen therapy to be associated with a sizeable reduction in coronary artery disease. Given the large effect and the burden of disease in the population, epidemiologists gushed with enthusiasm over this news. It was wrong. The Women’s Health Initiative refuted the association; women who were healthier preferentially chose to use estrogen (healthy-user effect).

Reserpine was a useful drug for treating hypertension. In 1974, three inadequate case-control studies alleged a doubling in the risk of breast cancer among users, and reserpine was quickly driven from the market. Later case-control studies failed to corroborate the finding but, by then, the drug has been mortally wounded in the media. In 1981, a poor-quality case-control study noting an association between drinking coffee and pancreatic cancer touched off a media firestorm. Later studies denied the association. The net effect was public confusion and erosion of confidence in all medical research.

Dangerous pseudo-epidemics have been common in women’s health research as well. Case-control

Table 1. Examples of Spurious Associations in Observational Epidemiology Studies

<table>
<thead>
<tr>
<th>Exposure</th>
<th>Outcome</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cigarette smoking</td>
<td>Increased risk of suicide</td>
<td>Smoking associated with factors predisposing to mental state that increases suicide risk. Information bias and residual confounding.</td>
</tr>
<tr>
<td>Beta-carotene</td>
<td>Reduced risk of lung cancer</td>
<td>Selection bias: women who chose to use estrogen were at lower risk for coronary artery disease.</td>
</tr>
<tr>
<td>Menopausal estrogen therapy</td>
<td>Reduced risk of coronary artery disease</td>
<td></td>
</tr>
<tr>
<td>Reserpine therapy</td>
<td>Increased risk of breast cancer</td>
<td>Flawed case-control studies; findings not replicated by later, larger studies.</td>
</tr>
<tr>
<td>Drinking coffee</td>
<td>Increased risk of pancreatic cancer</td>
<td>Gravely flawed case-control study; finding refuted by later studies.</td>
</tr>
<tr>
<td>Induced abortion</td>
<td>Increased risk of breast cancer</td>
<td>Information bias; under-reporting of abortion among healthy control individuals.</td>
</tr>
<tr>
<td>Antihistamine and vitamin B6 (pyridoxine/doxylamine) exposure</td>
<td>Increased risk of birth defects</td>
<td>Junk science.</td>
</tr>
<tr>
<td>IUD use</td>
<td>Increased risk of salpingitis and infertility</td>
<td></td>
</tr>
<tr>
<td>Oral contraceptive use</td>
<td>Increased risk of pituitary adenoma</td>
<td>Confounding by indication.</td>
</tr>
</tbody>
</table>

IUD, intrauterine device; STD, sexually transmitted disease.
studies have consistently found an association between induced abortion and breast cancer. The explanation, documented first in Sweden, was systematic under-reporting of previous abortions by healthy control individuals (social desirability bias).

The antiemetic Bendectin, a combination of vitamin B6 and an antihistamine, was widely and effectively used by pregnant women in the United States. Junk science and tort law drove it from the marketplace. As wryly noted in one review, the product was not a teratogen but was a potent “tortigen.”

Women in the United States no longer have access to this drug.

Bad epidemiology nearly drove the intrauterine device from the U.S. market in the 1970s and 1980s. Many case-control studies found an association with salpingitis and later tubal infertility. Several biases accounted for this finding, including failure to control for the confounding influence of sexual behavior and sexually transmitted disease. In the first such study to gather Chlamydia serology of all case and control group participants, women who had used an intrauterine device had no increase in the risk of infertility. In contrast, women who had previous chlamydial infections had a significant increase in their risk.

Naïve epidemiologists gravely damaged women’s health for decades.

In the early 1970s, observational evidence linked use of birth control pills with pituitary adenomas. This association was attributable to confounding by indication: women with irregular menses (attributable to an unsuspected prolactin-secreting adenoma) were preferentially administered oral contraceptives to regulate their bleeding (protopathic bias). The association was real, but the outcome (pituitary tumor) led to the exposure (oral contraceptives), not the reverse.

Unlike physics, medical science lacks immutable rules. All we have to work with are theories; the more often they are tested and hold up, the more likely they are to be true. The notion of causality has its roots in philosophy; however, for pragmatic reasons, several guidelines have been developed and used over the years.

The most widely used guidelines are those of Sir A. Bradford Hill. In 1965, he proposed a list of nine considerations: strength of association; consistency of the observation; specificity of the relationship; temporal association; biologic gradient (dose-response effect); plausibility; coherence with other evidence; experiment; and analogy. One year earlier, the landmark Surgeon General’s report on smoking and health had used similar principles to evaluate the published evidence: consistency; strength of association; specific-ity; temporality; and coherence.

Susser suggested that the three critical elements are association, temporality, and direction; association is bolstered by consistency and strength of the association. Others have modified or amplified these criteria for use in infectious disease and ecology, but strength of association is dominant in all these guidelines.

Large associations are more likely to be attributable to causality than are weak associations. Stated alternatively, bias in observational studies can easily account for small associations. Observational epidemiology is somewhat adept at identifying large associations but, regrettably, not small ones. The examples provided by Hill in his landmark treatise were large associations: cigarette smoking and death from lung cancer and water source and death from cholera in London in 1854. For cigarette smoking, the relative risk (RR) was 8 for 1–14 cigarettes per day, 20 for 15–24 cigarettes per day, and 32 for 25 or more cigarettes per day (compared with none). For the cholera example, Londoners who got their water from the Southwark and Vauxhall company had an RR of death of 14 compared with those whose water was supplied by the Lambeth Company.

In contrast to strong associations like smoking and cancer, observational epidemiology is not able to evaluate weak associations, variably defined by noted epidemiologists as RR or odds ratio (OR) between 1 and 4 (Table 2). Selection bias and residual confounding may lead to these weak associations without any causal association being present. Residual confounding stems from unknown, unmeasured, or poorly measured confounding factors. This problem is heightened in studies using administrative databases, which were never designed for clinical research, but rather for billing and other uses. Only randomized controlled trials can circumvent these problems.

Because randomized controlled trials sometimes cannot be performed, many questions will remain unresolved because observational studies cannot answer them.

Several approaches have been used to reduce bias in observational epidemiology, but none eradicates it. Bigger studies are not necessarily better. Large numbers provide greater precision, but not greater validity. Very large studies commonly give precisely wrong answers, ie, high reproducibility but poor validity because of selection bias and inadequate control of confounding. Consistency is desirable only when the answer is correct. Statistically significant dif-
ferences in large studies are deceptively appealing. With a large sample size, trivial differences yield miniscule $P$ values and narrow confidence intervals (CIs).\(^4^8\) $P$-values and CIs relate to precision, not to validity, which is more important to us. Similarly, meta-analysis of observational studies provides more precision (narrower CIs) but no gain in validity.\(^2^8\) A meta-analysis can be no better than its component parts.

Dating back almost three decades, numerous authors have suggested thresholds for identifying associations worthy of further consideration for possible causality (Table 2). These range from 2 to 4. The more liberal guideline (RR or OR of two) suggests that results between 1 and 2 can be discounted. Such results generally represent noise, not signal. The more cautious guidelines would reject results between 1 and 3 or 4 as noninformative. The suggestion of the pioneering epidemiologist Sir Richard Doll was that the lower 95% confidence limit should not overlap 3, which means that the point estimate would be much higher.\(^7\) In summary, experienced epidemiologists and editors (Table 2) are cautious in the interpretation of weak associations.

Figures 3 and 4 reflect the guidelines suggested by leaders in epidemiology and evidence-based medicine (Table 2).\(^4^9\) Because case-control studies are especially vulnerable to bias,\(^4^9\) these studies need more stringent safeguards against fallacious findings: an OR of 3 or more for meriting interest. In Figure 3, ORs more than 1 and less than 1 are reciprocally related: an OR of 3 is equal in strength but opposite in direction to an OR of 0.33 (the reciprocal of 3, or $1 \div 3$). Thus, research findings with ORs between 0.33 on the protective side and 3 on the adverse side should be discounted (zone of potential bias). Results 3 or more or 0.33 or less (zones of potential interest) would merit further consideration. Some (Table 2) suggest more conservative OR thresholds of 4 and 0.25 (dotted lines in Fig. 3). Because cohort studies are less prone to bias, the recommended RR threshold is 2 or more and, correspondingly, 0.5 or less on the

---

**Table 2. Recommended Minimum Effect Sizes* for Observational Studies by Type of Study**

<table>
<thead>
<tr>
<th>Study Type</th>
<th>Minimum Effect Size</th>
<th>Author</th>
<th>Year of Recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative risk</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cohort</td>
<td>3</td>
<td>Khoury(^6^0)</td>
<td>1992</td>
</tr>
<tr>
<td>Cohort</td>
<td>3</td>
<td>Angell(^7)</td>
<td>1995</td>
</tr>
<tr>
<td>Cohort</td>
<td>3 or 4</td>
<td>Temple(^7)</td>
<td>1995</td>
</tr>
<tr>
<td>Cohort</td>
<td>2</td>
<td>Shapiro(^4^5)</td>
<td>2000</td>
</tr>
<tr>
<td>Cohort</td>
<td>2</td>
<td>Shapiro(^4^5)</td>
<td>2004</td>
</tr>
<tr>
<td>Cohort</td>
<td>3</td>
<td>Straus(^4^9)</td>
<td>2005</td>
</tr>
<tr>
<td>Cohort</td>
<td>2 to 3</td>
<td>Shapiro(^4^7)</td>
<td>2008</td>
</tr>
<tr>
<td>Odds ratio</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Case–control</td>
<td>2</td>
<td>Wynder(^6^4)</td>
<td>1990</td>
</tr>
<tr>
<td>Case–control</td>
<td>3</td>
<td>Khoury(^6^0)</td>
<td>1992</td>
</tr>
<tr>
<td>Case–control</td>
<td>4</td>
<td>Straus(^4^9)</td>
<td>2005</td>
</tr>
<tr>
<td>Case–control</td>
<td>3</td>
<td>Parascandola(^4^1)</td>
<td>2006</td>
</tr>
<tr>
<td>Not specified</td>
<td>3 to 4</td>
<td>Lilienfeld(^4^5)</td>
<td>1983</td>
</tr>
<tr>
<td>Not specified</td>
<td>2 to 3</td>
<td>Sackett(^4^6)</td>
<td>1991</td>
</tr>
<tr>
<td>Not specified</td>
<td>Lower limit of 95% confidence interval more than 3</td>
<td>Doll(^7)</td>
<td>1995</td>
</tr>
<tr>
<td>Not specified</td>
<td>4</td>
<td>Trichopoulos(^7)</td>
<td>1995</td>
</tr>
</tbody>
</table>

* To avoid weak associations.
protective side (Fig. 4). Results 2 or more would merit further consideration, as would those 0.5 or less. Again, some advise more conservative RR thresholds of 3 and 0.33 (dotted lines in Fig. 4). As seen in Figures 3 and 4, the zones of potential bias are large. Danger lurks within this broad range of possible results.

If results are in the zone of potential interest and merit further consideration, then readers should still carefully consider the potential of large biases leading to strong spurious associations. In women’s health, two notorious examples are the putative relationship between the intrauterine device and salpingitis, and that between fertility drugs and ovarian cancer. In the former case, an RR of 10.550 in a famous British cohort study was later found to be attributable to bias.51 The most egregious example, however, may be the reported RR of 2752 in a meta-analysis of fertility drugs and ovarian cancer, which was later refuted.53 These examples yielded results far outside the zone of potential bias. Only after thorough consideration of large biases as an explanation should readers consider the possibility for causation by usual guidelines.41

Much of the published literature has results that lie within the zone of potential bias (Fig. 1); according to contemporary epidemiologic standards (Table 2), these results are not worthy of serious consideration. In one survey of published reports, three-quarters of the associations had RRs of less than 2.5. Approximately half were between 1.4 and 2.5, and large associations were uncommon.54 Another author examined the effect sizes reported in abstracts at one meeting of the International Society for Pharmacoepidemiology. He found that 55% of abstracts featured effect sizes less than 2 (weak associations).55

**DISCUSSION**

Because this problem has several causes, several remedies are needed. Clinical researchers should be trained to perform research. Learning medicine by apprenticeship was abandoned more than a century ago and replaced by formal instruction and credentialing. Why lesser standards prevail in medical research remains unclear. Badly performed research (eg, “stupid epidemiology”)55 harms patients and wastes resources. Thus, it is also unethical.56

The “low-hanging fruit” in observational epidemiology has been picked by now. Most of the big associations have already been found, eg, smoking and cancer or asbestos and mesothelioma.55,57 Researchers now are left to sift through countless weak associations that, because of bias, are difficult to interpret.58 The tendency of “risk factor epidemiology” [also termed “risk-factorology”] to search mindlessly for any activity of daily life that might be dangerous to health has been roundly criticized. Newspapers and broadcast media, however, thrive on their weekly dose of this silliness,59 as do plaintiff’s lawyers.35 According to Dimitrios Trichopoulos, former head of epidemiology at the Harvard School of Public Health, “We are fast becoming a nuisance to society…. People don’t take us seriously anymore, and when they do take us seriously, we may unintentionally do more harm than good.”7

Journals need to be less tolerant of poor research and its overzealous interpretation. If weak associations are to be published, then authors should be required to provide explicit warnings that the results are more likely attributable to bias than to causality. Further, readers should be warned that no clinical decisions should be based on the results. Cigarettes boxes have warnings about the dangers of the product. Journals might consider putting similar box warnings on observational epidemiology reports alerting readers to the potential dangers within. Weak associations are routinely reported without appropriate caveats. As noted by Shapiro, many investigators claim, “Our findings suggest that x may increase (or decrease) the risk of y,” when what should be said is, “We have identified a small association, and have not the foggiest notion of what it means.”55

Readers need to know that weak associations are simply beyond the discrimination ability of observational studies. An analogy would be trying to read the
name of the airline on a jet flying overhead at 10,000 m. Even with binoculars, the task is beyond the discrimination ability of the eyes. Or, to use another analogy of Shapiro, observational epidemiology is like a light microscope; a randomized controlled trial is akin to an electron microscope. The former has less discrimination ability. One cannot discern DNA with a light microscope, even at its highest magnification. Similarly, no observational study can identify an RR of 1.7, even with the best of methods (seldom used). Indeed, “… RR point estimates from nonexperimental studies are, by their nature, crude and imprecise.”

To claim that a weak association of 1.7 is valid and reliable is inconsistent with both the epidemiology literature and common sense. We are aware of no published epidemiologic guideline suggesting that RR or OR less than 2 in observational research should be deemed credible (Table 2). Khoury et al suggest that recurrent weak associations should prompt exploration of unmeasured confounding, exposure misclassification, etiologic heterogeneity in outcomes, biologic interactions, and differential survival as possible explanations. Skepticism is appropriate; for example, in two landmark Surgeon General’s reports on smoking and cancer, no causal claim was made for ORs less than 2.

Strong associations in the zones of potential interest merit attention, especially if many studies consistently display strong associations. Those associations are more likely to be real and perhaps causal. However, investigators, readers, referees, and journal editors should still closely scrutinize the methods behind those results. Strong associations consistent across many studies are “persuasive only if the studies use different architectures, methodologies, and subject groups and still come up with the same results.” If the studies use similar designs, and if an inherent bias exists, then David Sackett says, “it wouldn’t make any difference how many times it’s replicated. Bias times 12 is still bias.”

In conclusion, when readers encounter observational epidemiology reports with effect sizes less than 3 or 4, the results are frequently wrong. These associations do not warrant news coverage or changes in clinical practice. Observational epidemiology is a blunt tool, often wielded clumsily by the naïve. Caution lector.

REFERENCES


