

Publication bias *in situ*

Carl V. Phillips, MPP PhD

**University of Texas School of Public Health and Medical School,
Epiphi Consulting Group,
and
Center for Philosophy, Health, and Policy Sciences**

Poster, 36th Annual Society for Epidemiologic Research meeting

12 June 2003

Concept

The phenomenon of publication bias is well known and extensively addressed in the literature on meta-analysis and other systematic reviews: Studies that show a strong (or statistically significant result) are more likely to be published than those that do not.

But a potentially greater challenge for reviewers of the literature (systematic or otherwise) is bias that results from what researchers chose to report and highlight in their publications. Researchers have great freedom in deciding exactly what to analyze, how to analyze it, and what to report. As with traditionally defined publication bias, the analyses that are deemed unworthy of publication are largely invisible to the scientific and policy community.

While there is extensive discussion of the problems of "data dredging" and related statistical abuses, there is far more heat than light in the debate (assertions that there is or is not a problem, without formal analysis of specific problems or even clear statements of what is being debated). Proposed solutions to the challenge tend to take the form of scolding ("don't do that!"). The present analysis attempts to add some substance to the debate and some partial solutions.

Many Degrees of Freedom

Researchers' legitimate freedom about what to analyze include choices about:

- (1) Which effects and outcomes to consider in datasets with many variables.**
- (2) Whether to conduct separate analyses by subgroup.**
- (3) Which subgroup analyses to report or emphasize.**
- (4) Which functional forms to use to represent variables (e.g., how to divide continuous variables into categories).**

The categories are ordered based on how much attention they generally receive. Choice (1) (and to a lesser extent, (2)) tends to dominate the existing discussion in epidemiology. But I contend the latter entries in the list are more important, and the emphasis on the top of the list has confused the real problem. In the spirit of the order being backward, I present illustrations of (4), (3), and (2), in that order, along with a few suggestions for dealing with the resulting publication bias.

Caveats:

- This analysis is agnostic about how common these publication biases are, and about whether they should be attributed to reasonable or unreasonable choices on the part of researchers. For present purposes, I contend simply that it happens sometimes, and thus is worth thinking about.
- This analysis largely ignores all sources of error other than random sampling.

Degrees of Freedom, cont.

All research results are derived from data that can be used to test multiple hypotheses (even the most narrowly drawn clinical trial can be analyzed with the endpoint defined in different ways, stratified by age, etc.). Many epidemiologic datasets are characterized by thousands or millions of possible combinations of exposures, endpoints, and covariates (not to mention the infinite options for choices of functional forms).

There is substantial disagreement among viewpoints about how to apply statistical methods when dealing with multiple hypotheses or measurements.

At one extreme are viewpoints such as,

"We must correct test (α) values (or confidence interval widths) whenever multiple comparisons are made using the same dataset"

and even,

"Statistical analysis can only be legitimate for a short list of pre-specified hypotheses"

At the other extreme are viewpoints such as,

"Regardless of how many comparisons are examined, each can be considered and statistically tested as if it were the only one"

and,

"It does not matter if a hypothesis was proposed after looking at the data"

The participants on various sides of these debates have made valid points, but seem to have largely talked past each other rather than addressing the others' strongest arguments. The consideration of some specific cases serves to clarify the discussion.

The validity of these viewpoints depends on the circumstances and how results will be interpreted, as demonstrated by extreme examples

Example: unrelated results from the same dataset. At one extreme, consider a followup dataset originally used to report the relationship of drinking water source and *Helicobacter pylori* infection, which contains data that is later used to look at the relationship of household crowding and performance in school.

It is difficult to make sense of any argument that we must make an adjustment when doing the second analysis because we have already done the first (or, worse, an argument that we have already "used up" our .05 worth of α with the *H pylori* analysis, and so cannot test anything more with this data at all). A logical extension of that argument would be to consider the dataset that contains all quantitative human knowledge (which is a logically legitimate definition), and declare that we have to adjust for every statistical analysis ever done, effectively eliminating the legitimacy of all future statistical analysis.

Example: multiple independent measures of the "same" relationship. At the other extreme, consider a dataset with several covariates which can be used to stratify the analysis, and a researcher who decides to report the results of strata with a statistically significant relationship between exposure and disease.

Four dichotomous covariates (less information than most datasets contain), would divide the observations into $2^4 = 16$ non-overlapping strata. If the covariates are independent of the exposure-disease association, they create 16 independent comparisons (if they are not independent, the results are similar, but the math is more complicated). For a one-tailed significance level of .025, if there were no relationship between exposure and disease (the usual null hypothesis), there would be a $1-(1-.025)^{16} = .33$ chance of getting at least one significant positive association.

Further (non-independent) strata can be defined by ignoring the value for one or more covariates (so the value for each covariate in a stratum is either 1, 0, or "either"), for a total of 3^4 strata. Looking at these additional 65 strata further increases the chance of finding a significant association. Still more different strata definitions result from different categorizations of the covariates.

Clearly finding a single significant stratum-specific result and reporting it as if it passed the test, "is it incredibly unlikely this result was due to chance alone?", is misleading when there is a 1-in-3 or greater chance of getting such a result due to chance alone.

The key difference is in the plain language interpretation. If researchers are engaged in the standard practice of reporting a finding as positive (real, etc.) if and only if it passes a test of statistical significant, then:

In the second example, it is fairly likely there will be at least one statistically significant association even if there is no true association. Any such result will probably be reported by saying "the exposure is associated with the disease" or "the exposure is associated with the disease for some subgroup(s)." Thus for the 16 independent tests, there is a 33% chance that *a particular conclusion* will be made about an association, even if no such association exists.

In the first example, the conclusion that would result from finding one of those unrelated associations or any of 14 others we might think of would be quite different depending on which association it was. The chance of making *some conclusion* based on a significant association would be the same as in the other example (33% for 16 comparisons), but the chance of *any particular conclusion* being made due to random error alone would be .025.

Example of (4) – Choosing your cutpoint (a stylized example)

This example also appears as Example 3(b) in my adjacent poster, "On the nature of random error." See also Example 3(a) in that poster for a problem that can lead to either traditional publication bias or publication bias in situ.

When multiple versions of the comparison of interest are considered, there is a surprisingly large upward bias due to random error. Consider a case-control study with 100 cases and 100 noncases, and an exposure variable measured as 10 ordered categories (i.e., values 1,2,...,10, with larger numbers representing greater exposure). Assume that each of the 10 values is equally likely for cases and noncases in the population (which means, in particular, there is no systematic relationship between the exposure and the disease; the standard null hypothesis is correct).

What is the chance of observing a statistically significant positive relationship between exposure and disease, at a one-tailed significance level of .025? For any particular dichotomous definition of exposed and unexposed (e.g., exposure variable of 5 or greater is exposed), the chance is (by definition) 2.5%.

But if there is no clear cutpoint for defining exposure, there are many apparently legitimate comparisons. There are 9 possible cutpoints that divide the observations into two categories, defining those above the cutpoint to be exposed and those below unexposed. (Other options include comparing a group of highest categories to a group of lowest categories, leaving out the middle, such as 8-10 versus 1-3. These yield an additional 36 possibilities.)

Considering only the 9 dichotomizations, we can look at the results of repeated samples from the true underlying values. For each sample, one of the dichotomizations will produce the largest OR. If we collect those largest ORs from each sample, the median value of that set will be 1.5. So, if the "right" cutpoint is selected based on looking for which cutpoint shows the largest effect of the exposure, the median "effect" will be substantially elevated above 1.0 (the true effect). If we pre-specified a particular dichotomization, the median would be 1.0.

For the 9 dichotomizations, the chance of seeing at least one statistically significant result is 13%. This is still unlikely, but considerably more likely than the $\leq 2.5\%$ reported for the p-value when the result is reported. (The 13% is less than the 20% chance of seeing a significant result for 9 *independent* tests because the results are correlated, so when one is significant, others likely are. Thus, the same total number of significant results leads to fewer simulated samples with at least one such result. The correlations are complicated, so the results were calculated based on simulating repeated study samples.)

This is not to suggest that researchers typically analyze their data using every possible cutpoint and consider it legitimate to report whichever one has the statistically strongest result. But it does illustrate that the false positive rate is far higher than implied by the test statistics whenever there are multiple functional forms, any of which might be reported. Unbiased random errors, combined with picking and choosing functional forms, lead to biases in reported results.

Summary: For any dataset and comparison, different functional forms will show different levels of association. Choosing to report results from the functional forms that result in stronger associations will bias results upward.

Proposed partial solution: Authors should clearly acknowledge when there are multiple functional forms they thought plausible and calculated results for; ideally they should make some effort at specifying these *a priori*. Results should be reported for several of these (or all of them, if practical), with a discussion of which one(s) the authors think are most useful. Researchers can report standard test statistics or confidence intervals, but they should also report them based on the probability that *any of the functional forms considered* had a certain test value or deviates from the null by a certain amount. (Such values can be estimated using bootstrapping or other simulation methods.)

Example of (3) – Choosing Subgroups to Report (from literature on smokeless tobacco and oral cancer)

Background: Smokeless tobacco use (in particular, snuff dipping) is widely believed to be a proven major cause of oral cancer. However, the evidence overall does not support this claim. As reported in more detail in my poster yesterday, some studies have supported the claim of a relationship and the Surgeon General's Report [1986] declared the relationship to be proven. But there is currently much stronger evidence supporting the claim that there is no detectable relationship between modern smokeless tobacco products and oral cancer.

In the last 22 years, since Winn [1981] (see below), there have been over 20 published articles and abstracts of epidemiologic studies of this relationship. The two highest quality studies [Lewin et al., 1998; Schildt et al., 1998] provide strong evidence that the association is null or very close to it, but about 1/3 of the publications report a significant positive association. Several of these publications clearly report stratum-specific results that overstate the overall study findings.

Example, Blot et al. [1988]: From a case-control study that primarily reported on smoking and alcohol use, and was underpowered to study smokeless tobacco, the authors reported only one OR for smokeless tobacco use, 6.2 (95% CI 1.9-19) for nonsmoking females based on 10 exposed subjects. The authors say nothing to suggest that this subgroup's result is not representative of the association in the whole study population. Using other information from the article, it is possible to approximate values for the omitted subgroups. The OR for the whole population is almost exactly 1, which implies an unreported protective effect for everyone but female nonsmokers. The ORs for everyone but female nonsmokers and for all males are about 0.8.

Example, Kabat et al. [1994]: From another case-control study of various risk factors with few snuff users, the only OR reported for snuff is 34.5. This appears to be for nonsmoking females, and is actually infinity (no exposed controls, 4 exposed cases), with the finite result generated by replacing the cell count of 0 with 0.5 (The relevant paragraph of the article lacks clarity, but this seems to be the source of the OR.) Not reported is the OR for nonsmoking men, which is 0.0, based on the same number of exposed subjects (no exposed cases, 4 exposed controls).

Example, Winn et al. [1982]: This was published only as an SER abstract (something I can hardly criticize, eh? :-). The language is similar to the previous

example, concluding that smokeless tobacco use is associated with oral and possibly digestive organ cancers. But the only OR reported is for "very light smokers," which leaves the reader wondering why that stratum was chosen and what the results for other strata looked like.

Example, Winn et al. [1981]: This is the major study that is primarily responsible for the belief that smokeless tobacco causes oral cancer [see, Phillips et al., yesterday's poster]. It was a matched case-control study of 757 women, mostly elderly and Appalachian, with 231 exposed (mostly to products very different from modern snuff), so there was decent statistical power for subgroup analysis. To their credit, the authors did not report numbers for one subgroup while completely omitting those for the complementary subgroups. However, in the abstract (and, as a result, in a large portion of the writing about the topic to this day), only two numbers are mentioned: the 4.2 relative risk for nonsmoking white women, and the "approach[ing] 50-fold" relative risk for one subgroup. The widely quoted 4.2 is the largest among the smoking/race strata, and no rationale is presented for leaving out the 1/6 of the population that was black (who had a barely elevated cancer rate among nonsmokers).

The widely quoted 50-fold statistic (actually 47.5) is for nonsmokers with ≥ 50 years of exposure, restricted to cases with cancer at two anatomic subsites (and their matched controls), further restricted by eliminating subjects not meeting certain data requirements. The restrictions created the huge OR by

omitting all but 2 of the 41 total nonsmoking unexposed cases, and thus creating a very small denominator (34 out of 169 unexposed controls remained). Furthermore, we have re-analyzed the Winn data, and found that other stratifications of exposure period produce much lower risk ratios (indeed, the 47.5 may be the highest that is possible to get), which relates to the previous section of this poster.

Summary: With enough different ways to look at subgroups, it is frequently possible to select findings that support a given hypothesis. It is methodologically invalid to observe that smokeless tobacco almost certainly does not protect against oral cancer (except by reducing smoking), and therefore protective associations can be ignored, while positive associations are reported. Such practice will obviously bias reported results, even if some result from every study is published (i.e., even if there is no publication bias in the standard sense). This pattern of publication bias *in situ* in the smokeless tobacco literature is not unique; for many possible causal relationships, there has been substantial effort to find a relationship in the data, and often the contrary evidence is ignored. (See also the study of phenylpropanolamine and hemorrhagic stroke I have discussed at length previously, where major policy decisions were made based on the positive association for women while the protective effect for men was ignored.)

Proposed Partial Solution: Unless the underlying study question is focused on a particular subgroup, if a stratum-specific result is reported, results for all complementary strata should be reported, as well as the result for the whole population. This applies to the text and abstract (as well as press releases and such). It is probably not possible for researchers to keep partisans from willfully misinterpreting study results, such as by highlighting certain subpopulation results that may reflect random variation, but the research reports and journals should at least not actively encourage such misinterpretations.

Example of (2) – Choosing to do Subgroup Analysis (VaxGen HIV vaccine trial)

Background: Earlier this year, VaxGen Corporation released results of an HIV vaccine trial. The result was a trivial reduction in incidence among the treatment group compared to the placebo group, but the three non-white racial groups (black, Asian, and "other") each showed a substantial reduction in incidence (see table). VaxGen reported the overall failure of the trial to the popular and business press [e.g., Pollack and Altman, 2003], but tried to salvage some hope for the drug by pointing out the results for non-whites, suggesting that maybe it held promise of a vaccine for some populations. VaxGen's search for a silver lining resulted in rash of criticism from the research community (focused on the reporting of a result that was not a pre-specified hypothesis and the failure to correct for multiple hypothesis testing) and a shareholder lawsuit, alleging that the statistically illicit reporting of the subgroup result artificially inflated share prices.

	<u>total</u>	<u>vaccine</u>		<u>placebo</u>		<u>RR</u>
	subjects	subjects	incidence	subjects	incidence	
white	4511	3003	179	1508	81	1.11
nonwhite	498	327	12	171	17	0.37
black	314	203	4	111	9	0.24
asian	73	53	2	20	2	0.38
other	111	71	6	40	6	0.56
total	5009	3330	191	1679	98	0.98

So, what should VaxGen have done with the results? One simple choice is to report as they did, and another is to throw out the tantalizing findings about non-whites because there was no *a priori* hypothesis that the vaccine would work for particular racial groups. To follow a general rule of ignoring interesting but surprising findings would be a huge waste of information. At the same time, if we search for biological plausibility to explain whatever associations appear in the data, we can usually come up with something. (In this case, it has been speculated that some genotypes get a benefit from the vaccine, and the frequency of those genotypes is strongly correlated with race.)

The plausibility of the biological claim can be debated and errors not due to random sampling should be considered (including allegations of some irregularities in the data). Setting these aside, what can we say about random chance alone?

To look at the result for non-whites, with its p-value of .004, ignoring the fact that the subgroup analysis was clearly data-driven would overstate the finding, as suggested in the preceding examples.

A naive standard correction for multiple hypothesis testing would make the opposite error. Setting aside the possibility that other covariates would have been used to stratify the data had they produced subgroups with positive findings, the combination of the 4 racial groups implies a test of $2^4 - 1 = 15$ different hypotheses. To apportion the α among those 15 or make a similar adjustment makes it vanishingly unlikely for any to be statistically significant, even for a strong association.

An alternative is to ask the question, "if the vaccine has no effect, what are the chances of seeing, *in any racial group or combination thereof*, a result as extreme as the observed 63% reduction?" Asked that broadly, the answer is about 20%. (This was calculated using Monte Carlo simulation, by generating a large number of random draws based on the assumption of no effect. Because the hypotheses are non-independent in a complicated way, it would be quite difficult to calculate analytically.)

However, most of the 20% comes from the unstable results for the two smallest groups, Asians and "other". Restricting the analysis to combinations of racial groups that contain black (with or without Asians or other), the probability is only a 2.1%. (Note that the probability of seeing a 63% reduction by chance alone for any group that includes the whites is vanishingly small and can be ignored.)

So, what is the right answer? There is never a single Right Answer from a study; the quality of the answer always depends on the question being asked. Have we found a successful vaccine? Clearly not, as the relative risk for the whole population shows. Should the result for non-whites be considered unlikely to be due to random chance (i.e., statistically significant)? It depends on whether you consider it the answer to the question "does the vaccine show a result for blacks/non-whites?", in which case the answer is 'yes' (though the effective study size is quite small), or "does the vaccine show a result for any racial group?" in which case the answer is 'it is fairly likely we would see such a result due to chance alone.' (*Aside: Surprise! Frequentist hypothesis testing is not an objective exercise – it depends on subjective decisions about what to test!*)

Proposed Partial Solution: Since one of the above statistics, or something else, might be the right measure, depending on the question, and readers may be interested in various questions, the results should be reported with many relevant statistics. This will allow readers (researchers working on related projects, policy makers, investors) to focus on what they consider to be the answer to their own questions.

Conclusions

A few common themes emerge from this collection of observations:

Publication bias *in situ* can produce very misleading results in the scientific literature, leading to widespread misperceptions and misguided policies.

The debate about corrections for multiple comparisons should be reformulated with an eye to publication bias. Multiple measures from the same data present no inherent problem. But picking and choosing among many comparisons and then reporting one result as if it were the only one calculated is likely to substantially misrepresents the findings.

In particular, if there are multiple results in a study (e.g., separately analyzed strata) that could lead to the same plain-language conclusion (e.g., this study showed an association between smokeless tobacco and oral cancer), then it is highly misleading to statistically analyze or report a subgroup result as if it were the only comparison being made.

Neither the problem nor the solution is about statistics *qua* statistics; no solution will be found by appealing to statistical theory. The critical issue is the plain-language interpretation of results.

- If particular results for two or more comparisons would lead to the same plain-language conclusion (e.g., "smokeless tobacco is associated with oral cancer"), then an acknowledgment of and appropriate statistical adjustment for the multiple comparisons should be made when reporting that conclusion.
- If some of those results support a contradictory conclusion, they should also be reported. Whether those contradictory results are statistically significant or not is of no consequence; if they would have been reported had they shown a positive association, they should be reported.

- However, if the multiple comparisons would not be translated into the same plain-language conclusion, then none of the problems discussed here will occur. Such scenarios (like the very first example about *H. pylori*) are sometimes used in strawman arguments against these concerns, but they are not relevant. (Dredging for associations in huge datasets remains a problem for traditionally defined publication bias. If there are many other datasets that do not show the given association, those results might never be published. But this problem is not the same as the multiple comparisons problem that leads to publication bias *in situ*.)

Publication bias *in situ* can and does bias scientific conclusions.

Furthermore, in cases like some of the examples given here, where it appears to be willfully misleading, it tends to diminish the credibility of epidemiology in general.

REFERENCES

Blot WJ, McLaughlin JK, Winn DM, Austin DF, Greenberg RS, Preston-Martin S, Bernstein L, Schoenberg JB, Stemhagen A, Fraumeni JF. Smoking and drinking in relation to oral and pharyngeal cancer. *Cancer Research*. 1988;48:3282-3287.

Kabat GC, Cheng CJ, Wynder EL. The role of tobacco, alcohol use, and body mass index in oral and pharyngeal cancer. *International Journal of Epidemiology*. 1994;23:1137-1143.

Lewin F, Norell SE, Johansson H, Gustavson P, Wennerberg J, Biorklund A, Rutqvist LE. Smoking tobacco, oral snuff, and alcohol in the etiology of squamous cell carcinoma of the head and neck; a population-based case-referent study in Sweden. *Cancer*. 1998;82:1367-1375.

Pollack A, Altman LK. Large trial finds AIDS vaccine fails to stop infection. *The New York Times*. February 24, 2003, p. A1.

Schildt EB, Eriksson M, Hardell L, Magnuson A. Oral snuff, smoking habits and alcohol consumption in relation to oral cancer in a Swedish case-control study. *International Journal of Cancer*. 1998;77:341-346.

Surgeon General's Advisory Committee on the Health Consequences of Using Smokeless Tobacco. The health consequences of using smokeless tobacco. U.S. National Institutes of Health Publication No. 86-2874. Washington: United States Public Health Service, 1986.

Winn DM, Blot WJ, Shy CM, Pickle LW, Toedo A, Fraumeni, JF. Snuff dipping and oral cancer among women in the southern United States. *New England Journal of Medicine*. 1981;304:745-749.

Winn DM, Walrath J, Blot W, Rogot E. Chewing tobacco and snuff in relation to cause of death in a large prospective cohort. Abstract. *American Journal of Epidemiology*. 1982;116:567.