Disclosures

September 09, 2014

In 1897, eight-year-old Virginia O'Hanlon wrote to the New York Sun to ask, "Is there a Santa Claus?"[1] Virginia's father, Dr. Phillip O'Hanlon, suggested that course of action because "if you see it in the Sun, it's so." Today many clinicians and health professionals may share the same faith in the printed word and assume that if it says it in the New England Journal of Medicine (NEJM) or JAMA or The Lancet, then it's so.

Putting the existence of Santa Claus aside, John Ioannidis[2] and others have argued that much of the medical literature is prone to bias and is, in fact, wrong.

Given a statistical association between X and Y, most people make the assumption that X caused Y. However, we can easily come up with 5 other scenarios to explain the same situation.

1. Reverse Causality

Given the association between X and Y, it is actually equally likely that Y caused X as it is that X caused Y. In most cases, it is obvious which variable is the cause and which is the effect. If a study showed a statistical association between smoking and coronary heart disease (CHD), it would be clear that smoking causes CHD and not that CHD makes people smoke. Because smoking preceded CHD, reverse causality in this case is impossible. But the situation is not always that clear-cut. Consider a study published in the NEJM that showed an association between diabetes and pancreatic cancer.[3] The casual reader might conclude that diabetes causes pancreatic cancer. However, further analysis showed that much of the diabetes was of recent onset. The pancreatic cancer preceded the diabetes, and the cancer subsequently destroyed the insulin-producing islet cells of the pancreas. Therefore, this was not a case of diabetes causing pancreatic cancer but of pancreatic cancer causing the diabetes.

Mistaking what came first in the order of causation is a form of protopathic bias.[4] There are numerous examples in the literature. For example, an assumed association between breastfeeding and stunted growth, [5] actually reflected the fact that sicker infants were preferentially breastfed for longer periods. Thus, stunted growth led to more breastfeeding, not the other way around. Similarly, an apparent association between oral estrogens and endometrial cancer was not quite what it seemed. [6] Oral estrogens may be prescribed for uterine bleeding, and the bleeding may be caused by an undiagnosed cancer. Therefore, when the cancer is ultimately diagnosed down the road, it will seem as if the estrogens came before the cancer, when in fact it was the cancer (and the bleeding) that led to the prescription of estrogens. Clearly, sometimes it is difficult to disentangle which factor is the cause and which is the effect.

2. The Play of Chance and the DICE Miracle

Whenever a study finds an association between 2 variables, X and Y, there is always the possibility that the association was simply the result of random chance.
Most people assess whether a finding is due to chance by checking if the \( P \) value is less than .05. There are many reasons why this the wrong way to approach the problem, and an excellent review by Steven Goodman[^7] about the popular misconceptions surrounding the \( P \) value is a must-read for any consumer of medical literature.

To illustrate the point, consider the ISIS-2 trial[^8], which showed reduced mortality in patients given aspirin after myocardial infarction. However, subgroup analyses identified some patients who did not benefit: those born under the astrological signs of Gemini and Libra; patients born under other zodiac signs derived a clear benefit with a \( P \) value < .00001. Unless we are prepared to re-examine the validity of astrology, we would have to admit that this was a spurious finding due solely to chance. Similarly, Counsell et al. performed an elegant experiment using 3 different colored dice to simulate the outcomes of theoretical clinical trials and subsequent meta-analysis.[^9] performed an elegant experiment using 3 different colored dice to simulate the outcomes of theoretical clinical trials and subsequent meta-analysis. Students were asked to roll pairs of dice, with a 6 counting as patient death and any other number correlating to survival. The students were told that one dice may be more "effective" or less effective (ie, generate more sixes or study deaths). Sure enough, no effect was seen for red dice, but a subgroup of white and green dice showed a 39% risk reduction (\( P = .02 \)). Some students even reported that their dice were "loaded." This finding was very surprising because Counsell had played a trick on his students and used only ordinary dice. Any difference seen for white and green dice was a completely random result.

The Frequency of False Positives

It is sometimes humbling and fairly disquieting to think that chance can play such a large role in the results of our analyses. Subgroup analyses, as shown above, are particularly prone to spurious associations. Most researchers set their significance level or rate of type 1 error at 5%. However, if you perform 2 analyses, then the chance of at least one of these tests being "wrong" is 9.75%. Perform 5 tests, and the probability becomes 22.62%; and with 10 tests, there is a 40.13% of at least 1 spurious association even if none of them are actually true. Because most papers present many different subgroups and composite endpoints, the chance of at least one spurious association is very high. Often, the one spurious association is published, and the other negative tests never see the light of day.[^10]

There is a way to guard against such spurious findings: replication. Unfortunately, the current structure of academic medicine does not favor the replication of published results,[^11] and several studies have shown that many published trials do not stand up to independent verification and are likely false positives.[^12,13] In 2005, John Ioannidis published a review of 45 highlighted studies in major medical journals. He found that 24% were never replicated, 16% were contradicted by subsequent research, and another 16% were shown to have smaller effect sizes than originally reported. Less than half (44%) were truly replicated.

The frequency of these false-positive studies in the published literature can be estimated to some degree.[^2] Consider a situation in which 10% of all hypotheses are actually true. Now consider that most studies have a type 1 error rate (the probability of claiming an association when none exists [ie, a false positive]) of 5% and a type 2 error rate (the probability of claiming there is no
association when one actually exists [i.e., a false negative] of 20%, which are the standard error rates presumed by most clinical trials. This allows us to create the following 2x2 table.

<table>
<thead>
<tr>
<th>Study</th>
<th>Truth</th>
<th>Associated</th>
<th>Not associated</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Associated</td>
<td>True positive</td>
<td>100</td>
<td>Type 1 error</td>
<td></td>
</tr>
<tr>
<td>Not associated</td>
<td>Type 2 error</td>
<td>900</td>
<td>True negative</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>900</td>
<td>1000</td>
<td></td>
</tr>
</tbody>
</table>

By plugging in the numbers above:

<table>
<thead>
<tr>
<th>Study</th>
<th>Truth</th>
<th>Associated</th>
<th>Not associated</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Associated</td>
<td>100*0.80 = 80</td>
<td>900*0.05 = 45</td>
<td>125</td>
<td></td>
</tr>
<tr>
<td>Not associated</td>
<td>100*0.20 = 20</td>
<td>900*0.95=855</td>
<td>875</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>900</td>
<td>1000</td>
<td></td>
</tr>
</tbody>
</table>

This would imply that of the 125 studies with a positive finding, only 80/125 or 64% are true. Therefore, one third of statistically significant findings are false positives purely by random chance. That assumes, of course, that there is no bias in the studies, which we will deal with presently.

3. Bias: Coffee, Cellphones, and Chocolate

Bias occurs when there is no real association between X and Y, but one is manufactured because of the way we conducted our study. Delgado-Rodriguez and Llorca[4] identified 74 types of bias in their glossary of the most common biases, which can be broadly categorized into 2 main types: selection bias and information bias.

One classic example of selection bias occurred in 1981 with a NEJM study showing an association between coffee consumption and pancreatic cancer.[15] The selection bias occurred when the controls were recruited for the study. The control group had a high incidence of peptic ulcer disease, and so as not to worsen their symptoms, they drank little coffee. Thus, the association between coffee and cancer was artificially created because the control group was fundamentally different from the general population in terms of their coffee consumption. When the study was repeated with proper controls, no effect was seen.[16]
Information bias, as opposed to selection bias, occurs when there is a systematic error in how the data are collected or measured. Misclassification bias occurs when the measurement of an exposure or outcome is imperfect; for example, smokers who identify themselves as nonsmokers to investigators or individuals who systematically underreport their weight or overreport their height. A special situation, known as recall bias, occurs when subjects with a disease are more likely to remember the exposure under investigation than controls. In the INTERPHONE study, which was designed to investigate the association between cell phones and brain tumors, a spot-check of mobile phone records for cases and controls showed that random recall errors were large for both groups with an overestimation among cases for more distant time periods. Such differential recall could induce an association between cell phones and brain tumors even if none actually exists.

An interesting type of information bias is the ecological fallacy. The ecological fallacy is the mistaken belief that population-level exposures can be used to draw conclusions about individual patient risks. A recent example of the ecological fallacy, was a tongue-in-cheek NEJM study by Messerli showing that countries with high chocolate consumption won more Nobel prizes. The problem with country-level data is that countries don't eat chocolate, and countries don't win Nobel prizes. People eat chocolate, and people win Nobel prizes. This study, while amusing to read, did not establish the fundamental point that the individuals who won the Nobel prizes were the ones actually eating the chocolate.

Another common ecological fallacy is the association between height and mortality. There are a number of reviews suggesting that shorter stature is associated with a longer life span. However, most of these studies looked at country-level data. Danes are taller than Italians and also have more coronary heart disease. However, if you look at twins or individuals within the same country, you see the opposite association -- namely, it is the shorter individuals who have more heart disease. Again, the fault lies in looking at countries rather than individuals.

4. Confounding

Confounding, unlike bias, occurs when there really is an association between X and Y, but the magnitude of that association is influenced by a third variable. Whereas bias is a human creation, the product of inappropriate patient selection or errors in data collection, confounding exists in nature.

For example, diabetes confounds the relationship between renal failure and heart disease because it can lead to both conditions. Although patients with renal failure are at higher risk for heart disease, failing to account for the inherent risk of diabetes makes that association seem stronger than it actually is.

Confounding is a problem in every observational study, and statistical adjustment cannot always eliminate it. Even some of the best observational trials fall victim to confounding. Hormone replacement therapy was long thought to be protective for cardiac disease until the Women’s Health Initiative randomized trial refuted that notion. Despite the best attempts at statistical adjustment, there can always be residual confounding. However, simply putting more variables into a multivariate model is not necessarily a better option. Overadjusting can be just as problematic, and adjusting for unnecessary variables can lead to biased results.
Real-World Randomization

Confounding can be dealt with through randomization. When study subjects are randomly allocated to one group or another purely by chance, any confounders (even unknown confounders) should be equally present in both the study and control group. However, that assumes that randomization was handled correctly. A 1996 study sought to compare laparoscopic vs open appendectomy for appendicitis. The study worked well during the day, but at night the presence of the attending surgeon was required for the laparoscopic cases but not the open cases. Consequently, the on-call residents, who didn't like calling in their attendings, adopted a practice of holding the translucent study envelopes up to the light to see if the person was randomly assigned to open or laparoscopic surgery. When they found an envelope that allocated a patient to the open procedure (which would not require calling in the attending and would therefore save time), they opened that envelope and left the remaining laparoscopic envelopes for the following morning. Because cases operated on at night were presumably sicker than those that could wait until morning, the actions of the on-call team biased the results. Sicker cases preferentially got open surgery, making the outcomes of the open procedure look worse than they actually were. So, though randomized trials are often thought of as the solution to confounding, if randomization is not handled properly, confounding can still occur. In this case, an opaque envelope would have solved the problem.

5. Exaggerated Risk

Finally, let us make the unlikely assumption that we have a trial where nothing went wrong, and we are free of all of the problems discussed above. The greatest danger lies in our misinterpretation of the findings. A report in the New England Journal of Medicine reported that African Americans were 40% less likely to be sent for an angiogram than their white counterparts. The report generated considerable media attention at the time, but a later article by Schwartz et al. pointed out that the results were overstated. Had the authors used a risk ratio instead of an odds ratio, the result would have been 7% instead of 40%, and it's unlikely that the paper would have been given such prominence. Choosing the correct statistical test can be difficult. Nearly 20 years ago, Sackett and colleagues proclaimed "Down with odds ratios!" and yet they remain frequently used in the literature.

Another major problem is the use of relative risks vs absolute risks. Although the latter are clearly preferable, one review of almost 350 studies found that 88% never reported the absolute risk. Furthermore, overreliance on relative risks can be very misleading. Baylin and colleagues reported that the relative risk for myocardial infarction in the hour after drinking a cup of coffee was 1.5 (ie, a 50% increase). This rather concerning finding was taken up by Poole in a bitingly satirical letter to the editor, where he calculated that the relative risk of 1.5 translated to an absolute risk of 1 heart attack for every 2 million cups of coffee. Clearly, well-done studies have to be put in clinical context, and it is paramount to remember that statistical significance does not imply clinical significance.

Why Bother?

With all of the different ways that clinical trials can go wrong, one might wonder why we bother at all. Unlike little Virginia, who was prepared to believe whatever she saw in the newspaper, we
have become, if not cynics, then at least skeptics when it comes to our published research. But skepticism is a good thing and makes us challenge what we think we know in favor of what we can prove. Without this skepticism, we would still be prescribing hormone replacement therapy to prevent heart disease in women, giving class I anti-arrhythmics to cardiac patients after myocardial infarction, and prescribing COX-2 inhibitors with reckless abandon.

As Dr. Fiona Godlee summed up in her BMJ editorial on evidence-based medicine, “[it’s a] flawed system but still the best we’ve got.”