Commentary

Which of These Things Is Not Like the Others?

Jay S. Kaufman, PhD[1]; and Richard F. MacLehose, PhD[2]

In etiologic research, the goal is to estimate the causal effect of an exposure on a disease outcome, which means the result that would be obtained in a large randomized trial with perfect adherence if exposure could be assigned without regard to the baseline characteristics of the study participants.[1] But there is no reason to think that effects must be homogeneous across units in a randomized trial or in an observational study. Exposure may cause more or less disease in subgroups defined by age, sex, or any other background characteristic, whether measured or unmeasured. Indeed, the summary estimate over the population may reflect a mix of different effect magnitudes, or even a mix of subjects who are benefitted and harmed by the same treatment.[2]

Suppose that we obtain a summary estimate of causal effect from a perfectly conducted randomized controlled trial, for example a relative risk (RR) = 1.74. Across strata of baseline variables, however, the effect estimate will generally differ. Suppose that for men we observe RR = 1.87 and for an equal number of women, RR = 1.65. Now it is necessary to make a binary decision between 2 opposing views of reality. The first possibility (Fig. 1A) is that the 2 stratum-specific estimates (1.87 and 1.65) are 2 independent draws from a single underlying sampling distribution of the homogeneous effect. The difference between these 2 values is therefore due to sampling variability alone. The second possibility (Fig. 1B) is that the 2 stratum-specific estimates (1.87 and 1.65) are each a draw from their own distinct stratum-specific distribution, because men and women do not share the same common underlying effect magnitude. In this case, the unstratified value of 1.74 is guaranteed to lie somewhere between the 2 stratum-specific estimates, but the specific value it takes will depend on the proportions of men and women in the study population.

This binary decision between homogeneity and heterogeneity is therefore a fundamental step in every quantitative analysis. The appropriate analytic strategy depends on which of these 2 versions of reality is believed by the investigator. If the 2 stratum-specific values represent 2 independent draws from the same underlying sampling distribution, then it would be most efficient to combine the 2 stratum-specific estimates and report a single estimate for the effect of exposure, perhaps adjusting for confounding by the stratification variable. On the other hand, if each stratum-specific value is an estimate drawn from its own distinct sampling distribution, then it would be most advantageous to report these 2 estimates separately.[3(pp270-271)] For etiologic studies, there is no particularly useful interpretation for their combined value if they arise from distinct underlying effect magnitudes. This notion is summarized by the old saying that a man with his head in the oven and his feet in the freezer does not feel "just fine" on average.

The Role of Testing

Null hypothesis significance testing has been largely superseded in biomedical research, and modern investigators now focus more appropriately on effect and interval estimation.[4] Yet, for the binary decision between homogeneity and heterogeneity, many have argued that there still may be some valid role for the simple null hypothesis test.[5] This is because there is no causal parameter of fundamental interest to report, but instead there is simply a dichotomous modeling decision to be made. Moreover, this really is a question about sampling variability, which is exactly the question that significance tests were designed to answer. Significance tests therefore provide an easy recipe for handling the necessary question of

Corresponding author: Jay S. Kaufman, PhD, Canada Research Chair in Health Disparities, Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, 1020 Pine Avenue West, Montreal, Quebec H3A 1A2, Canada; Fax: (514) 398-4503; jay.kaufman@mcgill.ca

[1]Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, Canada; [2]Division of Epidemiology and Community Health, University of Minnesota, Minneapolis, Minnesota

DOI: 10.1002/cncr.28359, **Received:** June 14, 2013; **Revised:** August 6, 2013; **Accepted:** August 9, 2013, **Published online** September 10, 2013 in Wiley Online Library (wileyonlinelibrary.com)

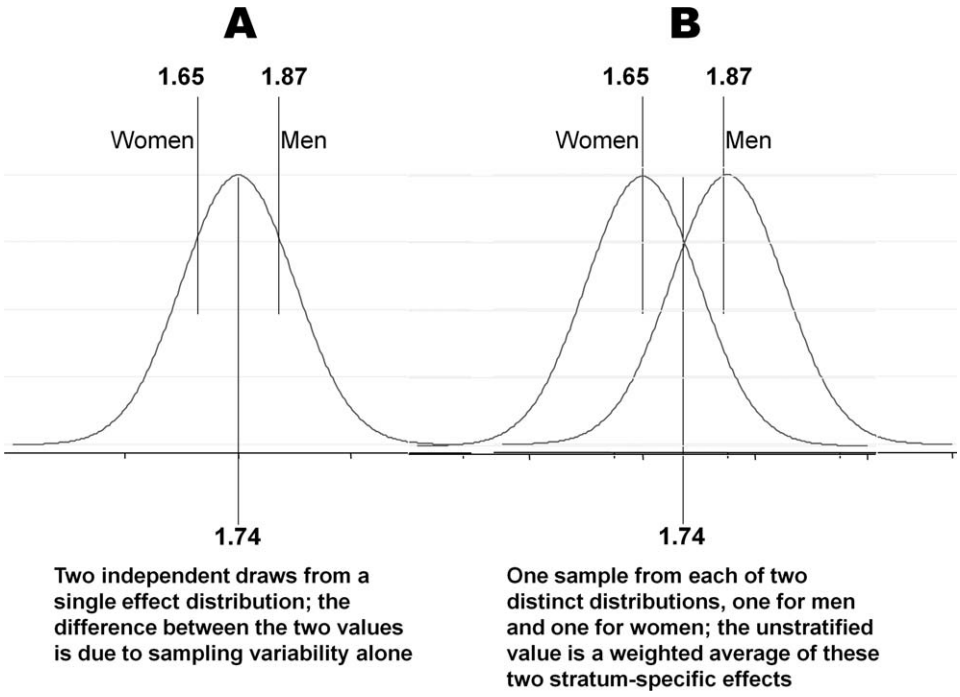

Figure 1. Graphs depict binary decision between homogeneity and heterogeneity in 2 opposing views of reality.

reporting a single homogeneous effect versus heterogeneous stratum-specific effects, but this approach also leads to all of the usual concerns about testing, especially issues of power and multiplicity.

When making a binary decision between the null hypothesis H_0 (homogeneity) and the alternate hypothesis H_A (heterogeneity), there are 2 ways to be right and 2 ways to be wrong. If H_0 is true and the test fails to reject H_0, or if H_0 is false and is correctly rejected, then no error has occurred. However, the probability of failing to reject when H_0 is false is called β (Type II error), and $(1-\beta)$ is the power of the test (ie, the probability of correctly rejecting H_0 when it is false). Tests with low power will tend to err on the side of decisions of homogeneity because of the inability to distinguish truly heterogeneous values. To guard against this implicit bias toward homogeneity, some authors recommend more liberal *P* value cutpoints for defining significance in such tests, such as 0.10 or 0.15.[6] The alternative kind of error is rejecting H_0 when it is true, a Type I error, and the probability of making this kind of error is referred to as the α level of the test. For homogeneity tests, the Type I error rate determines the proportion of the tests will indicate heterogeneity when homogeneity is in fact true.

The multiplicity problem is that because α is the proportion of tests that will yield a Type I error, an investigator who conducts $1/\alpha$ tests of a truly null hypotheses can expect a rejection, and therefore the apparent discovery of some heterogeneity even when none is present. This has led to a pervasive phenomenon in which investigators would fail to find an overall effect, but because they were desperate for a "significant" result, would sift through a multitude of baseline strata in search of some subgroup with an exposure *P* value $< .05$. Despite many sincere efforts to guard against this practice, often referred to as "fishing" or "dredging," this problem persists in the scientific literature. For example, a large clinical trial of a new vaccine against HIV infection in 1998 and 1999 found no overall effect, but "exploratory" analyses suggested that there might be a protective effect of the vaccine in "nonwhites."[7] VaxGen, the company that held the patent for the drug, issued a press release in 2003 that tried to spin the null overall finding into a promising result based on a smaller *P* value in the nonwhite subgroup.[8] A biological effect of this vaccine was never established and the protective subgroup effect is now thought to have been merely a Type I error.

The temptation to sift through many baseline strata in search of a "significant" exposure effect has led to profound suspicion over subgroup-specific effects in the clinical trials literature, particularly if the main effect is null.[9] Recommendations for trialists are to avoid subgroup-

specific *P* values, but instead focus on interaction tests between the treatment and the baseline covariate. Even then, many authors caution that additional criteria such as biological plausibility should be considered, as well as the number of subgroup analyses and whether they were prespecified. Rothwell,[10(p181)] for example, notes that "Selective reporting of post hoc subgroup observations, which are generated by the data rather than tested by them, is analogous to placing a bet on a horse after watching the race."

The Role of the Effect Measure Scale

Another crucial point is that apparent heterogeneity of exposure effects can be expected when different groups of patients have very different absolute risks with or without treatment. This is because heterogeneity of the effect measure across strata depends on the choice of the effect measure, such as the risk difference (RD), risk ratio (RR), or odds ratio (OR). When there is a non-null overall effect, then homogeneity of the RD generally forces RR and OR heterogeneity as a mathematical necessity. Likewise, homogeneity of the RR or the OR generally implies RD heterogeneity. Therefore, whenever there is some non-null effect, there must usually be heterogeneity on some scale. Whether this will be detected statistically is a question of the modeling strategy and the statistical power of the test. White and Elbourne provide an example from the UK Hip Trial in which effect of ultrasound on treatment modality is estimated, with possible stratification by an important baseline covariate.[11] One outcome ("operative treatment") is uncommon and the exposure effect is small, and there is no evidence of heterogeneity of effects no matter which effect measure is used. An alternate outcome ("any treatment") is common and the exposure effect is large, leading to highly significant heterogeneity of effect estimates for the RD and OR, but not for the RR. This is because the pattern of joint effects between the exposure and the covariate was approximately multiplicative in the risks, but not in the odds.

The Current Situation in Observational Studies of Disease Etiology

Although the clinical trials community has dealt head-on with the subgroup analysis problem, the world of observational epidemiologists has not yet arrived at any similarly coherent consensus. Some have advocated that all observational studies must also be preregistered, just like trials, so that planned comparisons can be declared a priori and be accounted for formally.[12] Given the extensive use of secondary data and the reliance on serendipity and iterative processing of data, however, this proposal seems hopelessly unrealistic. At the very least, however, we can expect that in observational epidemiology, decisions about heterogeneity would be based on substantive knowledge and valid heterogeneity tests, that authors would honestly report the number of such tests performed, and that data sharing and replication would help to quickly identify chance results that are of no etiologic importance. Examination of current practice in our journals suggests that we still have some way to go in improving reporting standards. For example, Knol and colleagues examined reporting in a large number of case-control and cohort studies in selected biomedical journals and found that less than half presented stratum-specific estimates and appropriate tests for interaction when asserting evidence of heterogeneity of effects.[13] Nieuwenhuis et al examined a series of articles in 5 leading journals (including *Science* and *Nature*) and reported that approximately half of the articles employed incorrect procedures to test for subgroup heterogeneity.[14]

One of the most egregious improprieties is to assert heterogeneity of the effect on the argument that the exposure has a "significant" effect in one stratum of the baseline covariate, but not in another. This is because the power of the test will generally differ. For example, suppose that we obtain a summary RR = 1.74 ($P = .01$), and when stratified by sex, we observe RR = 1.87 ($P = .03$) for men and RR = 1.65 ($P = .06$) for women. One should not conclude that the exposure has an effect in men but not in women, because each of these null-hypothesis *P* values conflates effect magnitude and statistical power.[15] Suppose that the study population was composed mostly of men. In that case, it would require a much larger magnitude of effect in women to generate a *P* value $< .05$. To compare each estimate to the null is a very different question than to compare the 2 estimates to one another, or to the uniform value that would be reported if they were considered to be homogeneous.[16]

Nonetheless, this wholly inappropriate practice is all too commonly encountered in the cancer literature. For example, Moore et al considered the potentially protective effect of physical activity on incident prostate cancer in black and white men over a 7-year follow-up.[17] They observed a statistically significant protective effect in black men who reported higher levels of physical activity when they were in their 20s (RR = 0.65; 95% confidence interval [CI] = 0.43-0.99) and when they were in their 30s (RR = 0.59; 95% CI = 0.36-0.96). No statistically significant effects were observed for white men. Although the authors cautioned that this finding should be considered exploratory on the basis that only one of

the heterogeneity tests had "borderline statistical significance" ($P = .10$ for men in their 20s), they went on to explain in the discussion section that "there are several potential biological reasons why physical activity may reduce prostate cancer risk among black men but not among white men." They then listed several hypotheses involving genetic variants, different underlying biological susceptibilities, inflammation and immune function differences, and higher levels of testosterone among young black men. But the general conclusion asserted that "physical activity may reduce prostate cancer risk among black men but not among white men" requires evidence of heterogeneity between blacks and whites, not within each group against the null. This is because the primary analytic decision is either to stratify on race or not stratify on race. It is the choice between reporting one common effect estimate for all men combined, versus reporting a separate estimate for whites and for blacks. Therefore, the question of interest is whether the stratum-specific estimates differ from the common effect estimate, not whether they each differ from the null.

How to Test for Heterogeneity

Many valid tests of heterogeneity exist in the statistical literature, including the Breslow-Day test, Wald test, and regression-based test of interaction. The Breslow-Day test is appropriate for the OR measure only[18] and has been improved by Tarone.[19] The corrected version is now implemented in most commercial software such as SAS and Stata, but the procedure is not as convenient for multivariable analysis, and is therefore decidedly less common in current practice than the other options. The Wald test works with any effect measure and can also be estimated using covariate-adjusted effect measures.[3(pp279-280)] This procedure is also readily available in commercial software packages, such as PROC FREQ in SAS and epitab in Stata. The formula is simply:

$$\chi^2_{Wald} = \sum_i \frac{(U_i\text{-}U)^2}{V_i}$$

Where U_i is the i-th stratum-specific estimate of the effect measure, V_i is the associated variance for estimate U_i, and U is the common value of the measure under the null hypothesis of homogeneity. This common value should be adjusted for any confounding by the stratification variable. The resulting statistic has a chi-squared distribution with degrees of freedom equal to the number of strata minus 1. For ratio measures such as the RR and OR, one should use the natural logarithm of the measure in the formula above.

Regression-based tests are based on the notion that a regression model with no product interaction terms implies homogeneity on the specific scale of the model. For example, logistic regression entails an assumption of homogeneous ORs over all strata when there is no product interaction term included in the model. Therefore, if a logistic regression is specified for 0/1-coded binary variables X, Y, and Z as $\text{logit}(Y) = \beta_0 + \beta_1 X + \beta_2 Z$, then the model form dictates linearity of the log-odds, which imposes homogeneous OR values for all contrasts. This means that across levels of Z, the exposure X effect will have to be the single homogeneous value $OR = \exp(\beta_1)$. To relax this assumption of equal Z-stratum-specific OR values, one would have to introduce a product interaction term so that the model would be: $\text{logit}(Y) = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ$. The extra term allows the 2 Z-stratum-specific OR values to differ. Specifically, the OR values will be $\exp(\beta_1)$ in the $Z = 0$ stratum and $\exp(\beta_1 + \beta_3)$ in the $Z = 1$ stratum. Because these will be equal only in the special case that $\beta_3 = 0$, this equality can be taken as a null hypothesis for testing homogeneity. Similar reasoning applies for other effect measures. For the RR, the relevant model would be a modified Poisson regression,[20] or a generalized linear model (GLM) with a log link and binomial distribution,[21] whereas for the RD it would be a linear probability model or a GLM with an identity link and binomial distribution.[22]

One word of caution about these homogeneity tests is that they generally have very low statistical power, often approximately 25% of the power of main effects tests.[23] This means that many true instances of heterogeneity will not be detected. As noted above, one commonly proposed solution is to conduct the test using a higher Type I error criterion, such as $\alpha = 0.10$ or 0.15.[6] Although this makes it easier to detect real heterogeneity, it also naturally raises the proportion of instances that heterogeneity will be declared erroneously due to a Type I error. Thus, it is trading in one kind of error for another kind of error. The question of which α level to use, therefore, is the question of which kind of error one would most prefer to minimize.

The methods described above are adequate when one has the individual level data, but it is often necessary to judge the heterogeneity of effect estimates from aggregated estimates, such as those that would be available from published reports. Typical examples would include taking stratum-specific estimates and their standard errors (or confidence interval limits) from the table of a single

published study, or else comparing estimates from 2 or more distinct publications. There are 2 simple ways to approach this problem: either as a simple difference of independent parameters[24] or using the heterogeneity test developed for meta-analysis.[25] To test that the difference of parameters is equal to zero, it is necessary to have the variance of the difference, which is the sum of the 2 variances under the assumption of independent samples. If there are 2 mutually exclusive strata, such as men and women, then this assumption will be met. The other crucial assumption is that the parameters are estimated on a sufficiently large number of observations so that the sampling distribution will be approximately normal (ie, at least 20 or 30 observations per stratum). This normality assumption will also be abetted by conducting the operations on the log-scale for ratio measures. For example, when the OR is the chosen effect measure, use the ln(OR), which is the logistic regression coefficient.

Then for effect estimates β_1 and β_2 with standard errors SE(β_1) and SE(β_2), the difference is $\delta = (\beta_1 - \beta_2)$, and $SE(\delta) = \sqrt{(SE(\beta_1))^2 + (SE(\beta_2))^2}$, which is to say that the variance of the difference is the sum of the individual variances. Armed with the estimate σ and its standard error, one can now conduct a simple Wald-style test in which the ratio $\sigma/SE(\sigma)$ is a Z-score. The square of this test statistic is equivalent to the Wald test shown above, with $\sigma = U_i$ and $U = 0$. Although this test statistic could be squared and *P* values based on the χ^2 distribution with 1 degree of freedom, it is more common to compute the Z-statistic. Returning to the previous example, say that stratification by sex yielded RR = 1.87 (95% CI = 1.39, 2.52) for men and RR = 1.65 (95% CI = 1.29, 2.12) for women. This is ln(1.87) = 0.6259 for men and ln(1.65) = 0.5008 for women, and the respective standard errors SE(β_1) and SE(β_2) are computed by inverting the confidence intervals as [ln(UL) − ln(RR)]/1.96 to give 0.1531 and 0.1276, where UL refers to the upper limit of the 95% CI. Thus, $\delta = (\beta_1 - \beta_2) = 0.1252$ and $SE(\delta) = \sqrt{(0.1531)^2 + (0.1276)^2} = 0.1992$. The test statistic is therefore 0.1252/0.1992 = 0.6282. The right-tail probability demarcated by a Z-score of 0.6282 in the normal distribution is 0.26. This is a 1-sided *P* value, and can be doubled to provide a 2-sided *P* value of 0.53. Most people would consider this to constitute very little evidence against the null of homogeneity, and therefore in the absence of any substantial prior substantive information that the effects should be heterogeneous, one should prefer to report the summary RR = 1.74 rather than the 2 sex-specific effects.

The second approach in this situation would be to consider RR = 1.87 (95% CI = 1.39, 2.52) for men and RR = 1.65 (95% CI = 1.29, 2.12) for women as though they were the estimates from 2 separately published studies, and we were interested in whether a meta-analysis of these values would provide for a single pooled summary measure, which requires homogeneity. The variances of the stratum-specific ln(RR) estimates are the squares of the standard errors shown above, or $VAR(\beta_1) = 0.1531^2 = 0.0234$ and $VAR(\beta_2) = 0.1276^2 = 0.0163$. The pooled summary estimate is then just the weighted average of the 2 log-scale estimates, weighted by the inverse of the variance. This allows the more precise stratum to count more than the less precise stratum.

$$pooled\ \beta = \frac{\left(\frac{\beta_1}{VAR(\beta_1)} + \frac{\beta_2}{VAR(\beta_2)}\right)}{\left(\frac{1}{VAR(\beta_1)} + \frac{1}{VAR(\beta_1)}\right)} = \frac{\left(\frac{0.6259}{0.0234} + \frac{0.5008}{0.0163}\right)}{\left(\frac{1}{0.0234} + \frac{1}{0.0163}\right)} = \frac{57.4983}{104.15} = 0.5521$$

The pooled summary estimate of the RR is therefore exp(0.5521) = 1.74. Using β_P to refer to the pooled effect, Cochran's Q test statistic is then computed as:

$$Cochran's\ Q = \left[\frac{(\beta_1 - \beta_P)^2}{VAR(\beta_1)} + \frac{(\beta_2 - \beta_P)^2}{VAR(\beta_2)}\right] = \left[\frac{(0.6259 - 0.5521)^2}{0.0234} + \frac{(0.5008 - 0.5521)^2}{0.0163}\right] = 0.3946$$

Note that this expression for Cochran's Q is identical to the expression given above for the Wald χ^2 test, but with the pooled estimator replacing *U*. This is also a χ^2 test statistic with degrees of freedom equal to the number of strata minus 1. With 2 strata in our example, therefore, we need only look up the tail probability for the χ^2 distribution with 1 degree of freedom that falls to the right of the value 0.3946. This is 0.53, and because the χ^2 distribution has only a single tail, this is not doubled. Note that the *P* values are similar and inferences are identical whether the Z-score test or the Cochran's Q test is used. The slight discrepancy between the 2 reflects the different ways of expressing the null hypothesis. In the Z-score test, the null hypothesis is that the difference of the stratum-specific estimates is zero. For Cochran's Q, the null hypothesis is that the both estimates are equal to the pooled estimate.

The use of these simple formulae obviates the need to compare overlap of confidence intervals or engage in any other similar guesswork. When considering whether to assert that one subgroup has a larger effect than another, one can always verify that this apparent difference is larger than what one might expect from sampling variability alone. Of course there are other alternate hypotheses beyond sampling variability. One stratum may be more confounded than another, or more severely mismeasured. As always, the null hypothesis significance test considers only sampling variability as the possible explanation for the apparent heterogeneity. All other potential improprieties would have to be pursued in additional analyses.

An Example in the Cancer Literature

To demonstrate a real-life example, consider once again racial heterogeneity in the protective effect of physical activity on prostate cancer. As noted above, Moore et al reported a protective effect in black men, but not white men, and advanced several biological hypotheses to explain why it makes sense that this protection against prostate cancer should be seen only in black men.[17] In the same journal 4 years later, Singh et al reported a protective effect of physical activity on prostate cancer, but only in white men.[26] These authors also assembled a list of biological explanations in the discussion, but this time for why it would make sense for this protection to occur only in white men. The observed disparity in the Singh et al article was that the adjusted effect of continuous exercise on cancer for white men was OR = 0.90 (95% CI = 0.82-1.00; $P = .041$); whereas for black men, OR = 1.02 (95% CI = 0.91-1.13; $P = .76$). Repeating the calculations illustrated above, one can quickly determine that the pooled uniform OR would be 0.96. Therefore, the authors must decide whether it makes sense to provide the readers with this single uniform effect, or whether there is compelling evidence of distinct effects in the 2 groups. The value of σ (ie, the difference in ln(OR) values) is equal to -0.125, 95% CI = -0.272, 0.022. The Wald test therefore yields a test statistic of -1.670, which demarcates a tail probability of 0.047. Doubling this to form a 2-sided P value yields 0.095. This matches the Cochran Q result, with $\chi^2 = 2.79$, and thus $P = .095$.

As noted above, because of the lower power of the test, some authors recommend trading in Type II error for Type I error by using a higher significance criterion, such as 0.10 or 0.15. Use of a more liberal cutpoint in this instance could render this result to be labeled as "statistically significant." However, the judgment of heterogeneity versus homogeneity still has to be made in substantive terms, not only statistical terms. A 2011 meta-analysis by Liu and colleagues considered 88,294 prostate cancer cases in 43 published studies from countries around the world, and reported protective summary effects for total activity (RR = 0.90; 95% CI = 0.84-0.95), occupational activity (RR = 0.81; 95% CI = 0.73-0.91) and recreational activity (RR = 0.95; 95% CI = 0.89-1.00).[27] When stratified by race, there were similar magnitude effects for the effect of total physical activity for whites (RR = 0.86; 95% CI = 0.77-0.97) and blacks (RR = 0.74; 95% CI = 0.57-0.95). Checking the heterogeneity P value between racial groups from the meta-analysis yields $P = .29$. To assert heterogeneity from a much smaller study, like the one reported by Singh et al, should therefore require very compelling evidence.

Improving Practice in the Future

Many have argued that epidemiologic research is largely a matter of pattern recognition, and one of the most fundamental of such exercises is the distinction between similar and dissimilar. When viewing numbers, it is trivial to assess whether $v_1 = v_2$ or $v_1 \neq v_2$, and so if we were able to view the true parameters of interest, there would never be any confusion. The problem is only that we view each value indistinctly, filtered through multiple sources of distortion.[28] Among these many sources of distortion, including confounding, information bias, and selection bias, the easiest one to assess is sampling variability. Under the premise that our data arise from a random sample of a target population, we draw a parameter from its sampling distribution (ie, the distribution of its value over hypothetically repeated iterations of the study). Classical statistical theory allows us to define the mean and the variance of this sampling distribution, and therefore to assess quantitatively whether 2 values that are drawn might differ reasonably just because of the random process of making these draws. If the observed difference is greater than one would expect from a series of random picks from a common distribution, then we need another explanation for this apparent heterogeneity. If other perturbations of the values can be dismissed (eg, confounding, selection, and information biases) then we might be persuaded that the true population effects are distinct. Because subgroup effects are commonly reported, careful thinking about this logic will eliminate many ultimately false assertions that currently appear in the literature. Even under ideal circumstances, a proportion of tests (α or β, depending on the true state of nature) would yield erroneous results. Given the presence of other biases in observational data, this proportion is certainly larger. But the proportion of

erroneous claims that are made by improper assessment of sampling variability is entirely avoidable. The calculations described in this commentary can be performed in a matter of minutes, or if incorporated into a spreadsheet or computer program, a matter of seconds. There is therefore no longer any excuse for haphazard or naive approaches to this problem. Obviously incorrect techniques, such as testing each subgroup-specific effect against the null, and then asserting heterogeneity if $P > .05$ in one group and $P < .05$ in another subgroup, should be relegated to an ignoble past to which we will never deign to return.

CONFLICT OF INTEREST DISCLOSURE

Dr. Kaufman was supported by the Canada Research Chairs program. Dr. MacLehose was supported by NIH grant 1U01-HD061940.

REFERENCES

1. Hernán MA, Robins JM. Estimating causal effects from epidemiological data. *J Epidemiol Community Health.* 2006;60:578-586.
2. Greenland S, Robins JM. Identifiability, exchangeability, and epidemiological confounding. *Int J Epidemiol.* 1986;15:413-419.
3. Greenland S, Rothman KJ. Introduction to stratified analysis. In: Rothman KJ, Greenland S, Lash TL, eds. Modern Epidemiology, 3rd ed. Philadelphia, PA: Lippincott, Williams & Wilkins; 2008.
4. Vandenbroucke JP, von Elm E, Altman DG, et al;STROBE Initiative. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. *Epidemiology.* 2007;18:805-835.
5. Weinberg CR. It's time to rehabilitate the P-value. *Epidemiology.* 2001;12:288-290.
6. Marshall SW. Power for tests of interaction: effect of raising the Type I error rate. *Epidemiol Perspect Innov.* 2007;4:4.
7. Flynn NM, Forthal DN, Harro CD, Judson FN, Mayer KH, Para MF;rgp120 HIV Vaccine Study Group. Placebo-controlled phase 3 trial of a recombinant glycoprotein 120 vaccine to prevent HIV-1 infection. *J Infect Dis.* 2005;191:654-665.
8. Watanabe ME. Skeptical scientists skewer VaxGen statistics. *Nat Med.* 2003;9:376.
9. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet.* 2000; 355:1064-1069.
10. Rothwell PM. Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *Lancet.* 2005;365:176-186.
11. White IR, Elbourne D. Assessing subgroup effects with binary data: can the use of different effect measures lead to different conclusions? *BMC Med Res Methodol.* 2005;5:15.
12. Bracken MB. Preregistration of epidemiology protocols: a commentary in support. *Epidemiology.* 2011;22:135-137.
13. Knol MJ, Egger M, Scott P, Geerlings MI, Vandenbroucke JP. When one depends on the other: reporting of interaction in case-control and cohort studies. *Epidemiology.* 2009;20:161-166.
14. Nieuwenhuis S, Forstmann BU, Wagenmakers EJ. Erroneous analyses of interactions in neuroscience: a problem of significance. *Nat Neurosci.* 2011;14:1105-1107.
15. Lang JM, Rothman KJ, Cann CI. That confounded P-value. *Epidemiology.* 1998;9:7-8.
16. Gelman A, Stern H. The difference between "significant" and "not significant" is not itself statistically significant. *Am Stat.* 2006;60: 328-331.
17. Moore SC, Peters TM, Ahn J, et al. Age-specific physical activity and prostate cancer risk among white men and black men. *Cancer.* 2009;115:5060-5070.
18. Breslow NE. Statistics in epidemiology: The case-control study. *J Am Stat Assoc.* 1996;91:14-28.
19. Tarone RE. On heterogeneity tests based on efficient scores. *Biometrika.* 1985;72:91-95.
20. Zou G. A modified poisson regression approach to prospective studies with binary data. *Am J Epidemiol.* 2004;159:702-706.
21. McNutt LA, Wu C, Xue X, Hafner JP. Estimating the relative risk in cohort studies and clinical trials of common outcomes. *Am J Epidemiol.* 2003;157:940-943.
22. Wacholder S. Binomial regression in GLIM: estimating risk ratios and risk differences. *Am J Epidemiol.* 1986;123:174-184.
23. Greenland S. Tests for interaction in epidemiologic studies: a review and a study of power. *Stat Med.* 1983;2:243-251.
24. Altman DG, Bland JM. Interaction revisited: the difference between two estimates. *BMJ.* 2003;326:219.
25. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ.* 2003;327:557-560.
26. Singh AA, Jones LW, Antonelli JA, et al. Association between exercise and primary incidence of prostate cancer: does race matter? *Cancer.* 2013;119:1338-1343.
27. Liu Y, Hu F, Li D, et al. Does physical activity reduce the risk of prostate cancer? A systematic review and meta-analysis. *Eur Urol.* 2011;60:1029-1044.
28. Maclure M, Schneeweiss S. Causation of bias: the episcope. *Epidemiology.* 2001;12:114-122.

Erratum

Erratum: Kaufman JS and MacLehose RF. Which of these things is not like the others? *Cancer* doi: 10.1002/cncr.28359.

The formula for the pooled effect estimate on page 5 contains an incorrect subscript. The formula should read $pooled = \frac{\frac{\beta_1}{Var(\beta_1)} + \frac{\beta_2}{Var(\beta_2)}}{\frac{1}{Var(\beta_1)} + \frac{1}{Var(\beta_2)}}$. This error does not affect any of the calculations that follow.

We also wish to clarify that the Cochran's Q statistic and Wald's Z-statistic reported on page 5 are equivalent tests, and will therefore always provide the same p-values when assessing homogeneity of stratum specific effects.

Neither of these errors alters the substantive recommendations of our commentary.

The authors regret these errors.

DOI: 10.1002/cncr.28496, wileyonlinelibrary.com **Published online:** December 2, 2013.

Erratum: Shanafelt T, Lanasa MC, Call TG, Beaven AW, Leis JF, LaPlant B, Bowen D, Conte M, Jelinek DF, Hanson CA, Kay NE and Zent CS. Ofatumumab-based chemoimmunotherapy is effective and well tolerated in patients with previously untreated chronic lymphocytic leukemia (CLL). *Cancer*. 2013;119:3788-96.

The authors noticed an error in the Materials and Methods section, second paragraph, which describes the ofatumumab dose.

The ofatumumab dose was not based on body surface area and the first sentence should read "**. . .**ofatumumab (cycle 1: 300 mg on day 1, 1000 mg on day 2; cycles 2–6: 1000 mg on day 1) given intravenously every 21 days."

The authors regret this error.

DOI: 10.1002/cncr.28502, wileyonlinelibrary.com **Published online:** December 2, 2013.