The Problem of Multiple Testing

Kristin L. Sainani, PhD

False-positive results that arise as the result of chance are common in the medical literature [1-3]. By chance, every study sample will have slight imbalances that don't reflect the whole population. If researchers look at enough characteristics of a given sample, they are bound to discover these quirks and conclude (mistakenly) that they have significance for the whole population. This is the problem of multiple testing—the more tests you run on a sample, the greater the likelihood of a chance finding. This article will formally describe the problem of multiple testing and give readers tools for spotting chance findings in the literature.

MULTIPLE TESTING: WHAT IS IT?

Mathematically, the problem of multiple testing can be explained as follows: every statistical test comes with an inherent false positive, or **type I error**, rate—which is equal to the threshold set for statistical significance, generally .05. However, this is just the error rate for one test; when more than one test is run, the overall type I error rate is much greater than 5%. For example, if one runs 100 independent statistical tests where it is known no effects exist, the chance of getting at least one false positive (ie, at least one *P* value less than .05) is 99.4% (see In-Depth sidebar for how this is calculated), and 5 false positives are expected (because approximately 1 in 20 tests will yield a false positive).

Type I error. In hypothesis testing, this is a false-positive error. The researcher concludes that an effect exists when it does not.

One should not extrapolate from this discussion that 1 in 20 significant findings in the literature are chance findings. A type I error can only occur if the **null hypothesis** is true—that is, if the effect being tested is not real. Therefore, if researchers only ever tested real effects, there would be no chance findings in the literature. The probability that a given significant result is a false positive also depends on the study power (the lower the study power, the more likely it is a false positive). Thus, it is impossible to accurately estimate what proportion of "significant" results ($P < .05$) in the literature are actually chance findings. On the basis of varying assumptions, researchers have estimated this proportion to be anywhere from 1.5% to 96.1%, with approximately 50% being the most likely value [2].

Null hypothesis. The hypothesis of no effect—for example, the hypothesis that 2 variables are unrelated or that 2 groups don't differ. A type I error occurs when the null hypothesis is erroneously rejected.

WHAT ARE SOURCES OF MULTIPLE TESTING?

There are many sources of multiple testing (Table 1). Besides the most obvious sources—comparing multiple groups or examining multiple outcomes—other less obvious sources include subgroup analyses, variable definitions, repeated measures, and interim analyses. Some of these sources are not always readily apparent in a published article. For example, if a data analyst tries 3 different definitions/cut-points for "moderate drinking" in regression analyses, that person has run 3 statistical tests. However, this data exploration may be hidden from the reader.

ANATOMY OF A CHANCE FINDING

Some experiments have been conducted purely to illustrate the problem of multiple testing. In these examples, the researchers knew ahead of time that the null hypothesis was true, but

K.L.S. Division of Epidemiology, Department of Health Research and Policy, Stanford University, HRP Redwood Building, Stanford, CA 94305. Address correspondence to: K.L.S.; e-mail: kcobb@stanford.edu
Disclosure: nothing to disclose

Disclosure Key can be found on the Table of Contents and at www.pmrjournal.org

PM&R
1934-1482/09/$36.00
Printed in U.S.A.

Vol. 1, 1098-1103, December 2009
DOI: 10.1016/j.pmrj.2009.10.004

IN-DEPTH: HOW IS THE PROBABILITY OF AT LEAST ONE FALSE POSITIVE CALCULATED?

If 100 statistical tests are run when: (1) there are no real effects; and (2) these tests are independent, what is the probability of at least one false positive (that is, one *P* value under .05)?

For 1 test:

If there are no real effects, the probability of a false positive arising in a given test is .05. So, the probability that a false positive does not occur is 1 - .05 = .95.

For 100 tests:

If the tests are independent, meaning they are unrelated to each other, then the probability that no false positives occur in 100 tests is: $.95^{100} = .006$. Thus, the probability that at least one false positive does occur is 1 - .006 = .994, or 99.4%.

Note that this calculation requires two key assumptions:

1. The null hypothesis is true for all 100 effects being tested.
2. The effects being tested are completely independent.

In many cases where multiple tests are run in the literature, one or both of these assumptions may not be true.

they conducted multiple statistical tests anyway to demonstrate how frequently false positives arise.

In a 1980 study in *Circulation*, researchers randomly assigned 1073 heart disease patients to two groups, group 1 and group 2, but they treated the patients exactly the same [4]. Not surprisingly, they found that survival times were similar between the 2 groups. However, when they divided the patients into 18 subgroups based on prognostic factors, they found that in a particular subgroup (those with 3-vessel disease and an abnormal left ventricular contraction), group 2 patients had a survival advantage ($P < .025$). It seems surprising that a false positive could crop up so easily, but in fact the probability of this happening was high because the authors ran 19 statistical comparisons. (It is difficult to calculate the exact probability because the subgroups overlap, but the probability would be 62% if 19 independent comparisons were run.) The lesson: if one divides the data up in enough different ways, it is easy to find a subgroup with a chance imbalance in survival.

As another example, take an informal experiment that I did in an introductory statistics class. I divided the class into 2 groups based on whether the students were born on an odd or even day and then asked them to provide data on 28 variables about themselves (such as on their likes, dislikes, and eating habits). When I compared these 28 variables between the 2 groups, I found 2 significant differences ($P = .02$, $P = .04$). Does this mean that being born on an odd or even day is really associated with these variables? Of course not; these are clearly chance findings. When no effects exist, *P* values will randomly take on any value from 0 to 1 with equal probability, meaning that when you run 28 tests a few will fall in the 0 to .05 range just by chance. Figure 1 shows the 28 *P* values from the tests that I ran—in this context it is easy to see that the 2 "significant" *P* values are not very interesting or impressive.

It's impossible to know for sure whether a particular "significant" finding in the literature is a chance finding—but the aforementioned studies illustrate a pattern that should raise a high level of suspicion. When a large number of statistical tests are run and just a few findings of modest significance arise ($.01 < P < .05$), chance should be considered as a likely explanation.

For example, consider a paper in the *Archives of Internal Medicine* in which the authors examined the relationship between caffeine consumption and breast cancer [5]. The authors found no overall association, but they reported a few significant/near-significant findings ($P = .08$, $P = .02$, $P = .02$) from subgroup analyses, which they concluded war-

Table 1. *Sources of multiple testing*

Source	Example
Multiple outcomes	A cohort study looking at the incidence of breast cancer, colon cancer, and lung cancer
Multiple predictors	An observational study with 40 dietary predictors or a trial with 4 randomization groups
Subgroup analyses	A randomized trial that tests the efficacy of an intervention in 20 subgroups based on prognostic factors
Multiple definitions for the exposures and outcomes	An observational study where the data analyst tests multiple different definitions for "moderate drinking" (eg, 5 drinks per week, 1 drink per day, 1-2 drinks per day, etc.)
Multiple time points for the outcome (repeated measures)	A study where a walking test is administered at 1 month, 3 months, 6 months, and 1 year
Multiple looks at the data during sequential interim monitoring	A 2-year randomized trial where the efficacy of the treatment is evaluated by a Data Safety and Monitoring Board at 6 months, 1 year, and 18 months

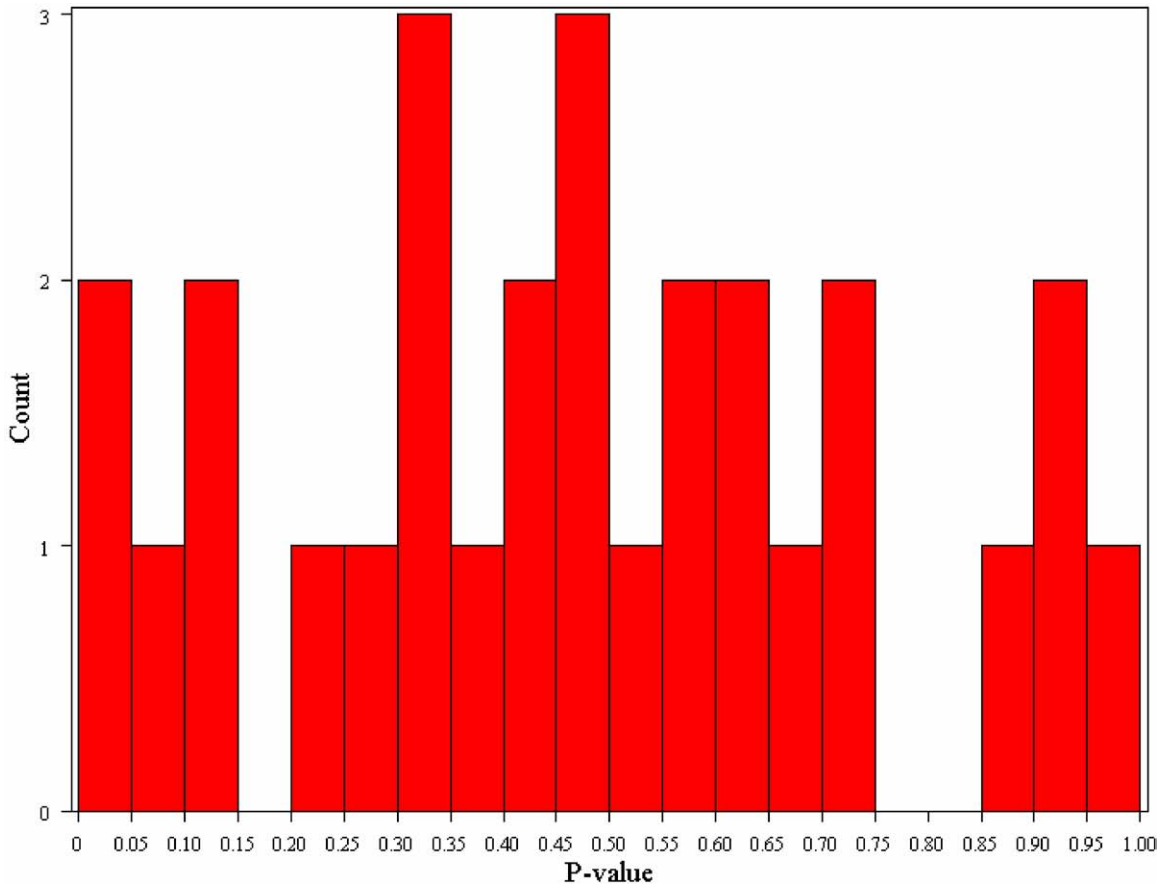

Figure 1. The distribution of the *P* values resulting from 28 statistical tests comparing 2 groups of students: those born on odd days and those born on even days. There is an equal likelihood of getting *P* values anywhere from 0 to 1, meaning that approximately 5% will come out to be less than .05. Indeed, 2 tests of 28 gave "statistically significant" results ($P < .05$).

ranted further investigation (specifically, coffee intake was linked to increased risk in those with benign breast disease, and caffeine intake was linked to increased risk of estrogen/progesterone negative tumors and tumors larger than 2 cm). The authors were appropriately cautious in their conclusions, but I would go one step further and assert that these findings are most likely the result of chance.

The authors ran 50 tests examining the relationship of breast cancer with caffeine, coffee, decaffeinated coffee, and tea intakes overall and in multiple subgroups on the basis of history of benign breast disease, body mass index (≥25, <25), menopausal status, hormone use (ever/never), hormone receptor status, tumor size (≤ or >2 cm), lymph node metastasis (yes/no), and histologic grade (well/moderate/poor). Figure 2 shows the distribution of the resulting 50 *P* values. Four *P* values were less than .10, which is consistent with chance. (In addition to the 3 *P* values mentioned previously, decaffeinated coffee was linked to protection against breast cancer in postmenopausal never hormone users, $P = .02$; interestingly, the authors did not comment on this finding.) The **effect sizes** are also consistent with chance: **risk ratios** were close to the null value of 1.0 (ranging from 0.67 to 1.79), indicated protection (<1.0) about as often harm (>1.0), and showed no consistent dose–response pattern across increasing levels of consumption. It is easy to come up with a plausible sounding biological story to explain why caffeine is important in women with benign breast disease and for certain types of tumors, but, in fact, the most likely explanation for the findings is chance alone.

Effect size: A measure of the magnitude of an observed effect—for example, how big the difference between groups is.

Risk ratio: A measure of relative risk formed by dividing the risk in one group by the risk in a reference group. Values of 1.0 indicate no difference in risk; values >1.0 indicate increased risk; and values < 1.0 indicate decreased risk.

MULTIPLE TESTING: WHAT TO LOOK FOR

In judging whether a given finding in the literature is likely to be to the result of chance, readers should consider whether the analyses were hypothesis-driven or exploratory, how many tests were run, the size of the *P* values, the pattern of effect sizes, and whether *P* values were adjusted for multiple comparisons (Table 2).

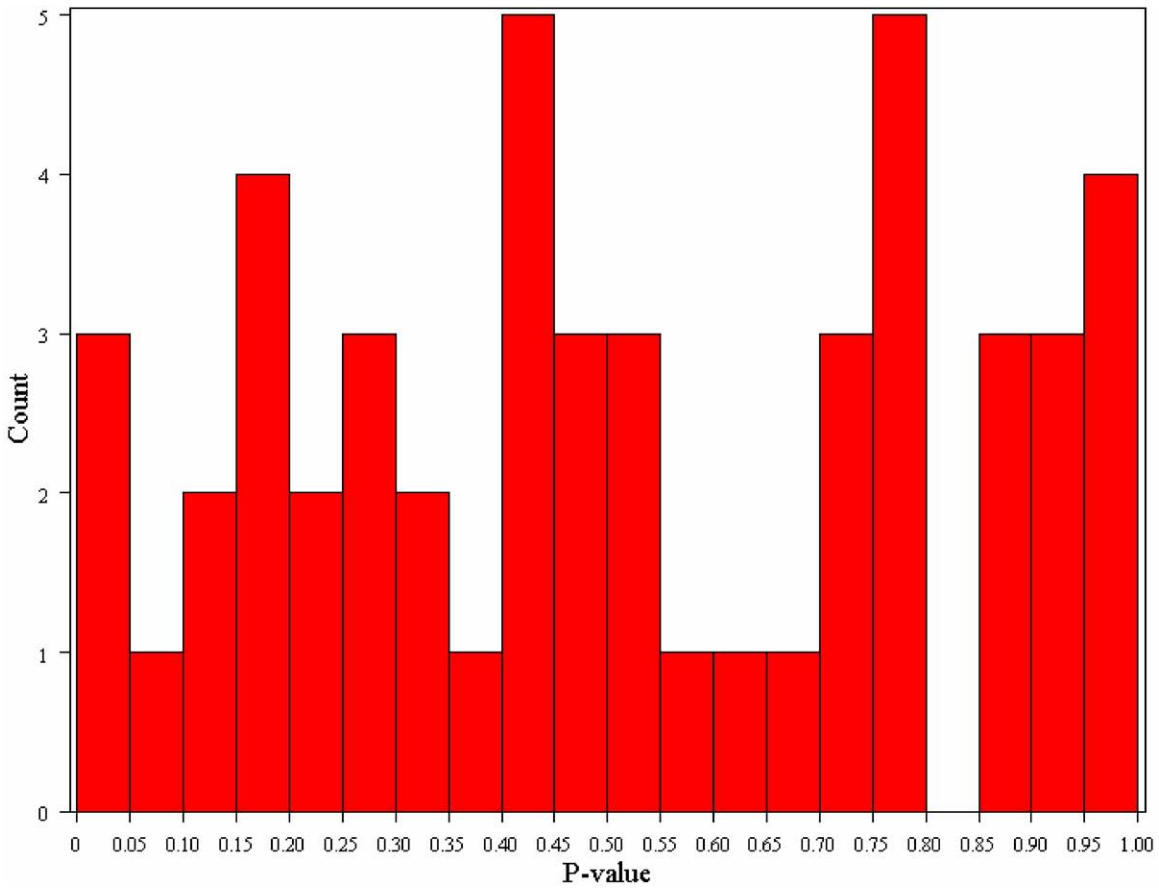

Figure 2. The distribution of *P* values from 50 statistical tests examining the relationship between caffeine, coffee, and tea intakes and breast cancer (5). *P* values were taken from tests for trend only and from the most adjusted model when more than one model was presented (Tables 2-5 from Ishitani, et al (5)). Four *P* values fall below .10, which would be expected due to chance.

Are the Analyses Hypothesis Driven or Hypothesis Generating?

When evaluating the literature, readers should distinguish between **hypothesis-driven analyses** and hypothesis-generating or **exploratory analyses**. When researchers specify a priori (before the study is conducted) a small number of hypotheses that they are planning to test, including clear definitions of the predictors and outcomes, this is hypothesis-driven research. This approach limits the number of statistical tests run, thus controlling the overall type I error. In contrast, when researchers test a large number of hypotheses after the data have been collected—essentially searching through the data to find associations—this is hypothesis-generating, or exploratory, research, and the type I error rate is likely to be high. This is not to say mining the data in this manner is wrong or bad. Often researchers collect large amounts of data in the course of a focused study, and it would be a waste not to examine these data. But "statistically significant" results that come out of such an exploratory analysis should be regarded with a greater level of scrutiny.

Hypothesis-driven analyses. When researchers run a limited number of prespecified statistical tests.

Exploratory analyses. When researchers run a large number of unplanned statistical tests looking for patterns in their data.

For example, if a randomized trial tests the effect of a drug versus a placebo in stroke patients, the *P* values associated with the primary hypothesis (the difference in recovery between drug-treated and placebo-treated patients) can be taken at face value. However, if in that same study, the researchers explored the associations between a large number of nutrition variables (that happened to be collected on a food frequency questionnaire as part of the study) and stroke recovery, then these results should be clearly identified as exploratory and interpreted cautiously.

How Many Tests Were Run?

Readers can count the number of tests reported in a paper and multiply it by .05 to get a rough idea of the number of *P* values less than .05 that would be expected to arise by chance alone (if no effects being tested were real). Of course, the data presented in a paper usually represent a subset of all the statistical tests run (particularly for exploratory analyses), so

Global test. A statistical test that compares multiple groups simultaneously, generating only one *P* value. For example, ANOVA tests the global null hypothesis that several groups' means are equal. If the null hypothesis is rejected, this indicates that at least one mean differs.

keep in mind that the number of tests run may be much larger than the number of tests reported.

There are a number of statistical approaches available that can reduce the number of tests being run. **Global tests** such as analysis of variance (ANOVA) and repeated-measures ANOVA compare multiple groups or multiple time points simultaneously, thus generating only one *P* value. For example, rather than running 3 *t* tests (and thus inflating the type I error) to compare the means in 3 groups, one can instead conduct a single ANOVA analysis that compares all three means at once. Statisticians also create **composite outcomes** to reduce the number of outcomes being tested.

Composite outcome. Where multiple endpoints are combined into a single outcome measure.

How Significant Is the *P* Value?

The strength of the evidence against the null hypothesis increases with smaller *P* values. Therefore, a *P* value of $<.001$ is less likely to be a chance finding than a *P* value of $<.05$. In one paper, researchers estimated—based on certain assumptions about the literature—that when a significance threshold of $P < .05$ is used, about 1 in 2 significant findings in the literature will be the result of chance, but when a threshold of $P < .001$ is used, only 1 in 56 significant findings will be the result of chance [2].

What Is the Pattern of Effect Sizes?

P values do not tell the whole story, and readers should always also consider effect sizes. If a particular association (eg, caffeine and breast cancer) has been evaluated across multiple different tests, the pattern of the resulting effect sizes can be informative. Real effects may be missed if a study has low **statistical power**. Therefore, if there is a consistent pattern of effect sizes, for example, if all the risk ratios are in the direction of harm (>1.0), but only a few results achieve moderate statistical significance, one may suspect that low statistical power is a factor rather than chance. On the other hand, if the effect sizes show no consistent pattern—as in the caffeine/breast cancer study mentioned previously—then chance may be a more likely explanation.

Statistical power. The probability of finding a real effect if it's there (that is, of rejecting the null hypothesis when one should). Statistical power may be low if sample size is small, if variability (of the things being measured) is high, or if the true effect size is small (and thus difficult to detect).

Has the *P* Value Been Adjusted for Multiple Comparisons?

Statisticians have devised ways to "adjust" *P* values or confidence intervals to account for the number of tests run. The basic idea is to preserve the overall type I error rate at .05 by lowering the threshold for statistical significance (to lower than $<.05$) or widening the confidence interval. For example, the simplest approach is the Bonferroni correction: When k tests are run, only *P* values under .05/k are deemed significant (eg, if 5 tests are run, only *P* values under .01 are reported as significant). Although easy to understand and conduct, the Bonferroni correction is overly conservative—it represents a "worst-case" scenario where all the tests being conducted are completely independent (which is usually not the case). Thus, many less conservative (but more mathematically intensive) methods have been developed. Applying these requires statistical software and/or consultation with a statistician.

Formal corrections for multiple comparisons are most often used in the context of hypothesis-driven research. For example, if the authors plan to look at multiple outcomes, look at a limited number of planned subgroups, or engage in interim analyses, they may build in a correction for these multiple comparisons. For exploratory analyses, formal adjustment of *P* values (and confidence intervals) is usually impractical. In these contexts, it is difficult to precisely quantify the total number of tests run and their interrelatedness; and, because of the large number of tests run, the adjusted threshold for statistical significance may be so small that it

Table 2. *Summary of factors that may be indicative of chance findings*

1. Analyses are exploratory.	The authors have mined the data for associations rather than testing a limited number of a priori hypotheses.
2. Many tests have been performed, but only a few *P* values are "significant."	If there are no associations present, .05*k significant *P* values ($P < .05$) are expected to arise just by chance, where k is the number of tests run.
3. The "significant" *P* values are modest in size.	The closer a *P* value is to .05, the more likely it is a chance finding. According to one estimate (2), about 1 in 2 *P* values $<.05$ is a false positive, 1 in 6 *P* values $<.01$ is a false positive, and 1 in 56 *P* values $<.0001$ is a false positive.
4. The pattern of effect sizes is inconsistent.	If the same association has been evaluated in multiple ways, an inconsistent pattern of effect sizes (eg, risk ratios both above and below 1) is indicative of chance.
5. The *P* values are not adjusted for multiple comparisons	Adjustment for multiple comparisons can help control the study-wide false-positive rate.

may be unreachable (and the false-negative rate will be extremely high). For exploratory analyses, it is more important to judge *P* values cautiously than to try to formally determine their true significance level.

CONCLUSION

Multiple testing is a major source of false positives in the medical literature. Exploratory analyses are particularly prone to this type of error and should be interpreted cautiously. When a few moderate size "significant" *P* values arise in the course of a large number of exploratory analyses, these likely reflect chance rather than real associations. Precise adjustment of *P* values and confidence intervals is often impractical in the context of exploratory research, but can be useful for hypothesis-driven research.

REFERENCES

1. Ioannidis JP. Why most published research findings are false. PLoS Med 2005;2:124.
2. Sterne JA, Smith GD. Sifting through the evidence—what's wrong with significance tests? BMJ 2001;322:226-231.
3. Boffetta P, McLaughlin JK, La Vecchia C, Tarone RE, Lipworth L, Blot WJ. False-positive results in cancer epidemiology: A plea for epistemological modesty. J Natl Cancer Inst 2008;100:988-995.
4. Lee KL, McNeer JF, Starmer CF, Harris PJ, Rosati RA. Clinical judgment and statistics: lessons from a simulated randomized trial in coronary artery disease. Circulation 1980;61:508-515.
5. Ishitani K, Lin J, PhD, Manson JE, Buring JE, Zhang SM. Caffeine consumption and the risk of breast cancer in a large prospective cohort of women. Arch Intern Med 2008;168:2022-2031.