Title

Multiplicity: When many analytic plans are applied or many redundant studies are run, false-positive results are ensured

Journal

European Journal of Clinical Investigation, 52(8)

ISSN

0014-2972

Authors

Powell, Kerrington
Prasad, Vinay

Publication Date

2022-08-01

DOI

10.1111/eci.13802

Received: 29 March 2022 | Accepted: 19 April 2022

DOI: 10.1111/eci.13802

COMMENTARY

WILEY

Multiplicity: When many analytic plans are applied or many redundant studies are run, false-positive results are ensured

Kerrington Powell[1] | **Vinay Prasad**[2]

[1]College of Medicine, Texas A&M Health Science Center, Bryan, Texas, USA

[2]Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, California, USA

Correspondence
Vinay Prasad, Department of Epidemiology and Biostatistics, University of California San Francisco, 550 16th St, 2nd Fl, 94158 San Francisco, CA, USA.
Email: vinayak.prasad@ucsf.edu

Funding information
This study was funded by Arnold Ventures

Keywords: biostatistics, epidemiology, multiplicity, observational studies, oncology, randomized control trials

Multiplicity occurs when many hypotheses are tested simultaneously without consideration of one another, and often results in false-positive findings or spurious associations.[1] In retrospective observational studies on hot topics (e.g. nutritional epidemiology), thousands of independent analytic teams may approach a similar question—all with different plans. This field-wise multiple hypothesis testing has been shown experimentally to generate both positive and negative statistically significant associations, simply by analytic choices. This phenomenon has also been shown for clinical questions in cancer medicine. Even when data sets are standardized, multiple analytic approaches may yield a range of answers to a single question. Finally, randomized controlled trials, the gold standard of causal inference, have historically been immune to questions of multiple hypothesis testing, although this is increasingly being called into question with the emergence of redundant, duplicative, and large trial portfolios. In this commentary, we explore the role of multiplicity in biomedical research—a growing challenge to the interpretation of individual study results.

1 | NUTRITIONAL EPIDEMIOLOGY AND OTHER RETROSPECTIVE OBSERVATIONAL DATA SETS

Consider the case of nutritional headlines that dominate the front pages of prominent news outlets such as *The New York Times'* health section. One week, researchers may suggest that blueberries or dark chocolate have been shown to reduce your risk of cancer, but the next week, these same exposures may be found to increase your risk. What explains this phenomenon? To begin, for popular topics, it is likely that thousands of individual analyses of a data set will be performed over a relatively short period of time, each controlling for some co-variates—those that researchers believe are plausibly related to an outcome—in an effort to uncover a meaningful correlation. Each of these models will create a new relationship between the investigated variables, as Patel et al. demonstrated by simulating the research community of nutritional epidemiology.[2] The authors used the National Health and Nutrition

Examination Survey (NHANES) and probed a series of nutritional exposures, asking if they increased or decreased overall mortality. For each exposure, the researchers used baseline variables (e.g. age and sex) and the 13 most common co-variates adjusted for in the sampled literature [e.g. '(smoking, body mass index (BMI), hypertension, diabetes, cholesterol, alcohol consumption, education, income, family history of heart disease, heart disease, any cancer, physical activity) and race/ethnicity'].[2] Then, the entire research community was simulated. Over 8000 different models were generated for each exposure-mortality association by combining all conceivable combinations of the 13 modifiable demographic factors. They found that the majority of studies showed no significant association. But, what was noteworthy is that for 31% of the variables, there were *both* statistically significant positive and negative outcomes for the same hypothesis, indicating that the hazard ratio (HR) could be HR >1 or HR ≤1 with a significant *p*-value depending on the level of co-variant adjustment.[2] Researchers called this the *vibration of effects*.

Schoenfeld and Ioannidis extended this result in an analysis measuring 50 common ingredients randomly selected from a cookbook.[3] Then, the researchers conducted a literature search on articles that measured each ingredient's link to cancer. Most of the ingredients ($n = 40$; 80%) had articles measuring their relation to cancer risk. Despite many weak and nonsignificant relationships, most ingredients had studies with outcomes contrary to each other, showing either an increased or decreased risk of developing cancer.[3]

Zaorsky and colleagues applied the vibration of effect approach to practical questions in cancer medicine. They found that by varying other analytic choices—left truncation adjustment, propensity score matching, landmark analysis, and different combinations of co-variates—they were able to generate any desired result.[4] These are all instances of a common theme when dealing with multiplicity: studies measuring the same research question yielding opposite findings.

2 | SAME DATA, DIFFERENT INTERPRETATIONS

Work by Silberzahn et al. demonstrated a similar situation of multiplicity when they categorized the skin tone of different soccer players and included it in a data set with reports of penalties (red cards) to 29 research teams.[5] The following question was posed to the teams: Were soccer referees more likely to issue a dark-skinned player a red card, signalling a penalty, than they were a light-skinned player? Twenty of the teams reported significant evidence of bias, whereas nine teams discovered a nonsignificant relationship, with one team finding a trend in the opposite direction (i.e. bias against lighter-skinned players).[5] These different analytical strategies provide researchers with a great deal of latitude, allowing for the potential of a myriad of distinct outcomes. However, the issue intensifies when one considers that significant findings are more likely to be published,[6] resulting in a dichotomized literature devoid of a middle ground of null results.

3 | MULTIPLICITY IN RANDOMIZED TRIAL RESULTS

Randomized controlled trials (RCTs) have historically been thought to be immune to multiplicity as rarely are hundreds or thousands of studies run on a single clinical question, but this fact may be shifting. There are now four critical considerations to examine regarding the relationship between multiplicity and oncology: (1) The United States Food and Drug Administration will approve drugs based on a single positive trial, even if the primary outcome is a surrogate endpoint, and even if other trials are negative; (2) Drug approvals often generate enormous financial windfalls; (3) Pharmaceutical companies tend to conduct large, duplicative trials with little rationale; and (4) The probability value (*p*-value) is arbitrary.

First, consider neratinib, the only drug ever approved in the adjuvant setting prior to the metastatic. Approval was based on a *single* placebo-controlled Phase III trial measuring invasive disease-free survival (iDFS) as a primary composite endpoint. The magnitude of benefit was small, with 5.1% and 1.3% improvements in 5-year iDFS rates in patients with hormone receptor-positive breast cancer who began therapy with trastuzumab less than one 1-year ago or more than 1-year ago, respectively.[7] Additionally, there are occasions when a medicine, such as adjuvant sunitinib in renal cancer, is approved despite the existence of a single negative trial and a single positive trial, thus ignoring the study portfolio.[8,9]

The second and third points may be coupled; approvals of cancer drugs are anticipated to yield billion-dollar profits,[10] which encourages the conduction of duplicative studies in many tumour types, despite weak evidence. Consider the genesis of the EVOLVE-1 study, which compared everolimus with placebo in patients with advanced hepatocellular carcinoma following sorafenib failure.[11] The maximum tolerated dose and the disease control rate were tested in early Phase I and Phase I/II studies, respectively, which laid a relatively weak foundation for expediting the EVOLVE-1 trial, rather than conducting a more conservative Phase II trial.[12] Despite the negative outcome of the trial, one reason for taking such an enormous financial risk is because, despite the high upfront

costs of conducting these large trials, a far larger financial incentive remains, namely drug approval if the trial is successful.

However, the case of everolimus is just one example in the broader landscape. Consider that approximately 700 clinical studies were conducted in a single year for pembrolizumab, and that when more and more tumour types are evaluated, the risk/benefit profile of the drug deteriorates, as was shown during the development of sunitinib monotherapy.[13] Even with negative trials and worsening aggregate risk/benefit profiles, a drug approval's billion-dollar return greatly outweighs the initial expense of conducting million-dollar studies.

Fourth, consider the most widely used statistical instrument, the *p*-value. If researchers run 100 trials to determine the effect of an inert drug on survival and assume a one-tail *p*-value of $p < .05$, a distribution of five trials on average will have a false-positive result. This is precisely the definition of the *p*-value—the probability of seeing this result or a more extreme result if the null hypothesis is assumed. This value is an arbitrary line in the sand, although arguably a necessary one that is admittedly susceptible for misinterpretation.

These concepts are illustrated in Figure 1. The left side (shown in blue) represents a single, large pan-tumour RCT for a novel cancer therapy that was negative. In an analysis of prespecified subgroups, there were some tumours with positive results. Is it likely that a positive subgroup finding may result in FDA approval for a particular indication? The answer to that question is no. The FDA would perform adjustments for multiplicity, and in the absence of that, the findings are, at most, hypothesis generating. Now consider that instead of conducting a single RCT, several, separate RCTs were conducted in numerous indications, approximating the mentioned subgroups (shown on the right and represented in orange). Some of these studies may be positive in the same subgroups, perhaps even by chance alone, but the overall portfolio may be the same. The difference is that now these findings will result in drug approval. The reality is that, although each of these studies in orange were performed independently, they represent a trial portfolio. Both situations are philosophically equivalent as they test a single hypothesis (i.e. does adding this drug result in clinical benefit?), and on the left, the bias is clear, but on the right, positive trials appear distinct, and the portfolio is never assessed in aggregate.

Because the trial portfolio is not considered in the present oncologic regulatory environment, multiplicity must be accounted for. One example that illustrates how statisticians and cancer doctors may view a question different is also captured in Figure 1. Some statistical experts have suggested meta-analyses be used for the figure on the Right (orange), rather than multiplicity testing.[1,14] However, these approaches fall short of answering the pertinent clinical question because the drug is considered in aggregate in multiple tumour types repeatedly, rather than identifying whether a drug works in one tumour or the other.[15] A meta-analysis or pooled estimate focuses on determining whether a drug is effective in all tumour types, rather than the cancer doctor's question of *which* tumour the treatment is effective in–a distinct difference. Because this technique does not exclude the possibility of a single positive trial leading to drug approval, multiplicity adjustment is needed to sate the doctor's and patient's question, and not a pooled estimate. This scenario also illustrates the importance of content specific experts guiding the framing of the statistical question.

Combining all the key points above, businesses are now incentivized to test drugs with marginal benefits in as many indications possible. Consider that when a pharmaceutical firm develops a drug, all translational research costs are expended, leaving just the expense of additional trials. When companies consider this sunk cost, which requires no further investment in research

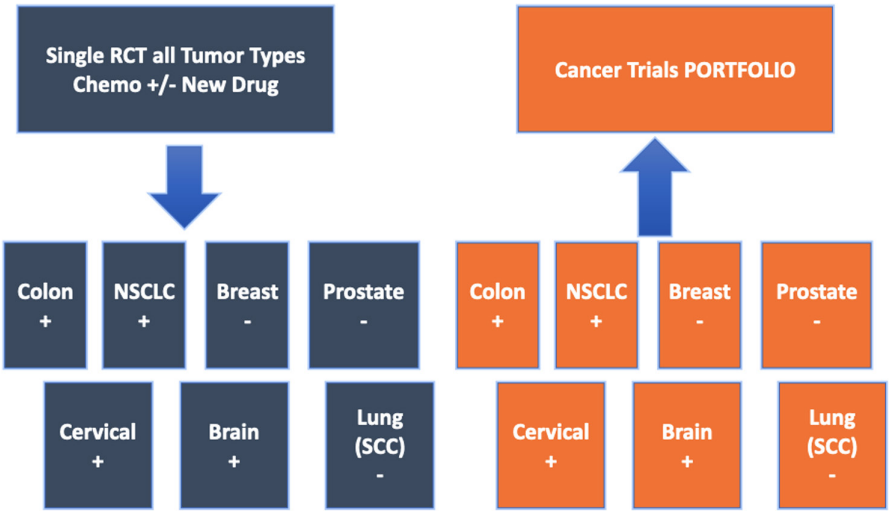

FIGURE 1 Schematic illustrating the requirement for multiplicity adjustment. Abbreviations: RCT, randomized controlled trial; Chemo, chemotherapy; +, positive trial results; −, negative trial results; NSCLC, non-small cell lung cancer; SCC, small cell lung cancer

and development but just the expense of the additional trial at the end, it incentivizes the company to test it in every single tumour type as many times as possible. When combined with the low bar for drug approval, the considerable post-approval revenue, and a generation's threshold of significance, pharmaceutical companies stand to profit enormously. Because these therapies are likely more effective than an inert substance, both true and false positives are obtained, which when averaged, results in a highly profitable approach. We see this with immune checkpoint inhibitor trials. There are now thousands of studies of largely similar molecules with massive duplication in the same or similar cancer settings,[16] often yielding conflicting results.[17]

4 | SOLUTIONS

As with many challenges, recognition and awareness is a prerequisite for thoughtful solutions. While preregistration of observational research may be beneficial, incentives are needed to ensure uniformity and consistency. Some researchers have outlined viable reform strategies to address this issue in the existing state of biomedical research.[18] Moreover, registration of observational studies is different than prospective research as it can be performed after the analysis is run.

Studying the impact of the policy is also warranted. When it comes to redundant and duplicative clinical trials, we must exercise caution to avoid waste. Patients are the most valuable and scarce resource, and they deserve the opportunity to contribute to solutions of the most pressing clinical questions; repeatedly conducting duplicative trials falls short of this aim.

Lastly, regulatory oversight might be needed for popular drug classes to prevent competing research agendas. Too many trials in the same cancer setting run the danger of creating an environment in which no trial fully accrues. Elsewhere, we have proposed statistical corrections for the results of individual trials run in settings with many duplicative results.[1] The purpose of biomedical research is to provide new information and results that lead to improved patient outcomes. Alignment with this goal will become increasingly difficult unless we confront the intrusion of multiplicity and establish higher standards.

AUTHOR CONTRIBUTION

VP conceptualized study design; KP reviewed the literature; VP reviewed and confirmed abstracted data; KP wrote the first draft of the manuscript; and all authors reviewed and revised subsequent and finalized draft of the manuscript.

CONFLICT OF INTEREST

Vinay Prasad's Disclosures. (Research funding) Arnold Ventures (Royalties) Johns Hopkins Press, Medscape (Honoraria) Grand Rounds/lectures from universities, medical centers, non-profits, professional societies, Youtube, and Substack. (Consulting) UnitedHealthcare. (Speaking fees) Evicore. (Other) Plenary Session podcast has Patreon backers. All other authors have no financial nor nonfinancial conflicts of interests to report.

ORCID

Kerrington Powell https://orcid.org/0000-0001-7067-3559
Vinay Prasad https://orcid.org/0000-0002-6110-8221

REFERENCES

1. Prasad V, Booth CM. Multiplicity in oncology randomised controlled trials: a threat to medical evidence? *Lancet Oncol*. 2019;20(12):1638-1640. doi:10.1016/s1470-2045(19)30744-2
2. Patel CJ, Burford B, Ioannidis JP. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *J Clin Epidemiol*. 2015;68(9):1046-1058. doi:10.1016/j.jclinepi.2015.05.029
3. Schoenfeld JD, Ioannidis JP. Is everything we eat associated with cancer? A systematic cookbook review. *Am J Clin Nutr*. 2013;97(1):127-134. doi:10.3945/ajcn.112.047142
4. Zaorsky NG, Wang X, Lehrer EJ, et al. Retrospective comparative effectiveness research: will changing the analytical methods change the results? *Int J Cancer Jan 31*. 2022;150:1933-1940. doi:10.1002/ijc.33946
5. Silberzahn R, Uhlmann EL, Martin DP, et al. Many analysts, one data set: making transparent how variations in analytic choices affect results. *Adv Methods Pract Psychol Sci*. 2018;1(3):337-356.
6. Thornton A, Lee P. Publication bias in meta-analysis: its causes and consequences. *J Clin Epidemiol*. 2000;53(2):207-216. doi:10.1016/s0895-4356(99)00161-4
7. Chan A, Moy B, Mansi J, et al. Final efficacy results of neratinib in HER2-positive hormone receptor-positive early-stage breast cancer from the phase III ExteNET trial. *Clin Breast Cancer*. 2021;21(1):80-91.e7. doi:10.1016/j.clbc.2020.09.014
8. Haas NB, Manola J, Uzzo RG, et al. Adjuvant sunitinib or sorafenib for high-risk, non-metastatic renal-cell carcinoma (ECOG-ACRIN E2805): a double-blind, placebo-controlled, randomised, phase 3 trial. *Lancet*. 2016;387(10032):2008-2016. doi:10.1016/s0140-6736(16)00559-6
9. Ravaud A, Motzer RJ, Pandha HS, et al. Adjuvant sunitinib in high-risk renal-cell carcinoma after nephrectomy. *N Engl J Med*. 2016;375(23):2246-2254. doi:10.1056/NEJMoa1611406
10. Tay-Teo K, Ilbawi A, Hill SR. Comparison of sales income and research and development costs for FDA-approved cancer drugs sold by originator drug companies. *JAMA Netw Open*. 2019;2(1):e186875. doi:10.1001/jamanetworkopen.2018.6875
11. Zhu AX, Kudo M, Assenat E, et al. Effect of everolimus on survival in advanced hepatocellular carcinoma after failure

of sorafenib: the EVOLVE-1 randomized clinical trial. *Jama*. 2014;312(1):57-67. doi:10.1001/jama.2014.7189
12. Prasad V. Translation failure and medical reversal: two sides to the same coin. *Eur J Cancer*. 2016;52:197-200. doi:10.1016/j.ejca.2015.08.024
13. Carlisle B, Demko N, Freeman G, et al. Benefit, risk, and outcomes in drug development: a systematic review of sunitinib. *J Natl Cancer Inst*. 2016;108(1):djv292. doi:10.1093/jnci/djv292
14. Gates S. Statistical significance and clinical evidence. *Lancet Oncol*. 2020;21(3):e118. doi:10.1016/s1470-2045(19)30854-x
15. Prasad V, Booth CM. Statistical significance and clinical evidence - Authors' reply. *Lancet Oncol*. 2020;21(3):e119. doi:10.1016/s1470-2045(20)30092-9
16. Upadhaya S, Neftelinov ST, Hodge J, Campbell J. Challenges and opportunities in the PD1/PDL1 inhibitor clinical trial landscape. *Nat Rev Drug Discov*. 2022. doi:10.1038/d41573-022-00030-4. Online ahead of print.
17. Gill J, Cetnar JP, Prasad V. A timeline of immune checkpoint inhibitor approvals in small cell lung cancer. *Trends Cancer*. 2020;6(9):736-738. doi:10.1016/j.trecan.2020.05.014
18. Ioannidis JP, Khoury MJ. Assessing value in biomedical research: the PQRST of appraisal and reward. *Jama Aug 6*. 2014;312(5):483-484. doi:10.1001/jama.2014.6932

How to cite this article: Powell K, Prasad V. Multiplicity: When many analytic plans are applied or many redundant studies are run, false-positive results are ensured. *Eur J Clin Invest*. 2022;52:e13802. doi:10.1111/eci.13802