Commentary: This study failed?

Charles Poole¹, Ulrike Peters², Dora Il’yasova³ and Lenore Arab¹

What are this study's implications? No question is asked more often of epidemiologists, or by them. In this issue of the International Journal of Epidemiology, the paper by Sesso et al.¹ on tea consumption in relation to coronary heart disease (CHD) and stroke provides an excellent illustration of several ways of framing this question.

What does it all mean?

This, of course, is the question that cartoon pilgrims climb mountains to ask their gurus. The broadest way of posing it in the present context would be with regard to what people ought to do, now that we have the results of Sesso et al.¹ about their own tea intake and that of others. These are policy decisions that require policy analyses to answer. The analyses would be complex. Health is but one of many of the considerations, CHD and stroke are just two of many health considerations, epidemiology is just one of several lines of scientific evidence pertaining to the health considerations, and the results of Sesso et al.¹ are from just one of many relevant epidemiological studies. Alternatively, “What does it all mean?” might be asked somewhat less sweepingly, with specific reference to the state of the epidemiological literature on tea in relation to CHD and stroke. A defensible answer to even this more circumscribed form of the question would require us to update the systematic review we prepared 2 years ago.² Without the benefit of a formal extension of our literature search, we are aware of at least a half dozen other studies that have appeared in the interim (references available upon request). A 1000 word commentary would be no place to try to do a good job of answering either version of the “What does it all mean?” way of asking about the implications of the Sesso et al.¹ results.

What would it all mean if this were the only study?

Suppose health were the only consideration in decisions about tea intake for individuals and populations, that CHD and stroke were the only health outcomes with hypothetical links to tea, and that the results of Sesso et al.¹ were the only available scientific information pertaining to those hypotheses. What would the study’s implications be then? Believe it or not, this ludicrously fanciful perspective is the one we are encouraged to take by the most commonplace of our received statistical methods: the null hypothesis test. The world of null hypothesis testing is an autistic world, closed off from all input but the data at hand. In that world, P < 0.05 warrants a call for action and P > 0.05 means that the status quo is to be maintained. In the latter case, it is common to say that a study has ‘failed’. Specifically, it has failed to ‘achieve significance’.

Although Sesso et al.¹ do not use the word ‘failed’, this is how they determine the implications of their results. Their take-home message is that their estimated trends and almost all of their categorical hazard ratio (HR) estimates are ‘non-significant’ because their 95% CI straddle the null value. The authors rely so heavily on null hypothesis tests that they interpret a change in an estimate from HR = 0.64 (95% CI: 0.43, 0.96) to HR = 0.67 (95% CI: 0.44, 1.02), after adjusting for a set of covariates, as though this change from ‘significant’ to ‘non-significant’ were meaningful. It is not. The two estimates are essentially identical. There is no appreciable difference in their magnitude or precision.

This reliance on null hypothesis testing almost obliges the authors to treat their own results as though they were the only information of scientific and policy relevance with regard to tea. They mention other research, of course, but their main conclusions are based exclusively on their own results:

These data provide no compelling reasons for individuals to initiate black tea consumption, which should be viewed as neither beneficial nor harmful.¹

We see no tenable rationale rationale for basing any decision about tea consumption or any judgement about whether tea is beneficial or harmful on the results of a single epidemiological study. An estimate that fails to achieve ‘significance’ could be the most valid and precise estimate in a greatly improved light. In a worthwhile systematic review, each individual study is not viewed as a failure (P > 0.05) or a success (P < 0.05). It is viewed as a contribution to a literature.

What does this study contribute to the literature?

Conducting a systematic review has the immense side benefit that one comes to view the contributions of individual studies in a greatly improved light. In a worthwhile systematic review, each individual study is not viewed as a failure (P > 0.05) or a success (P < 0.05). It is viewed as a contribution to a literature. What the study contributes is one or more estimates, in this case estimates of trend. Those estimates are not viewed by the systematic reviewer as ‘significant’ or ‘non-significant'; they are viewed as more or less precise and more or less valid.

¹ Department of Epidemiology, University of North Carolina, Chapel Hill, NC, USA.
² Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, DHHS, Bethesda, MD, USA.
³ Department of Epidemiology, Public Health Sciences, Wake Forest University School of Medicine, Winston-Salem, NC, USA.

Correspondence: Charles Poole, Department of Epidemiology (CB 7435), University of North Carolina School of Public Health, Chapel Hill, NC 27599-7435, USA. E-mail: cpoole@unc.edu

Dr Poole is supported in part by a grant from the National Institute of Environmental Health Sciences (P30 S10126).
Unfortunately, Sesso et al.\textsuperscript{1} report for each outcome only the ‘P for trend’, and not what the systematic reviewer needs: the estimated ‘coefficient for trend’ and its estimated standard error (or a CI from which the standard error can be deduced). Consequently, the systematic reviewer must use meta-analytical methods\textsuperscript{4} to find out what this study is contributing to the literature. For CHD and a 3-cup/day increment of tea intake, we estimate HR = 0.94 (95% CI: 0.86, 1.04). This estimate is identical to the most precise one in our previous review,\textsuperscript{2} from a report by Klatsky et al.,\textsuperscript{5} but the Sesso et al. estimate is considerably more precise; its inverse-variance weight is more than three times that of the Klatsky et al. estimate. For stroke, we estimate from the Sesso et al. results HR = 0.95 (95% CI: 0.78, 1.13) for an increment of 3 cups/day. This estimate is vastly more precise than any estimate for stroke in our previous review.\textsuperscript{2} It is also much closer to the null than any of those estimates. Therefore, somewhat ironically, the publication of this new study has almost certainly increased the evidence of publication bias in the literature as a whole, as very precise estimates very close to the null value almost always do. The influence of the Sesso et al. results on the evidence of overall heterogeneity and on associations between characteristics of the studies and their trend estimates is more difficult to discern, especially considering the fact that they are not the only results to appear in the interim.

From the validity standpoint, the Sesso et al.\textsuperscript{1} results share the strengths and limitations of most of the other cohort studies. The obvious strength, of course, is that control-selection bias is not an issue, as it is in the case-control studies. The most obvious drawback is that the outcomes are highly aggregated, presumably because they were self-reported by the participants. We would prefer to have results for specific outcomes, such as myocardial infarction, from all studies; but only the case-control studies provide that information. We have agreed\textsuperscript{6} that it would be highly desirable to distinguish haemorrhagic stroke from ischaemic stroke.\textsuperscript{7} This splitting is presumably not possible in most cohort studies.

We do not share the strength of conviction of Sesso et al. that clinical confirmation of 47 of 49 self-reported cases of CHD and of 12 of 15 self-reported cases of stroke means that the self-reporting ‘has been shown to be valid’.\textsuperscript{1} Forty-nine and 15 are small sample sizes for estimating proportions. Moreover, separating the false positives from a group of reported positives tells us nothing whatsoever about false negatives. Nor are we prepared to assume that the sensitivity of self-reported outcome ascertainment is perfect. Even non-differentially imperfect sensitivity biases relative measures of effect if the specificity is imperfect as well.\textsuperscript{8}

Each study provides something novel. In the case of the Sesso et al.\textsuperscript{1} results, it is the incorporation of a second questionnaire assessment of tea intake for a subset of the cohort. HR estimates farther from the null are obtained, which the authors interpret solely from the standpoint of reduced measurement error, even while contending that a validation study showed a single questionnaire measurement to be ‘reasonably reliable and valid’ because it explained 59% of the variance among tea intakes measured by dietary records. Another way of looking at the two-measurement results is that they might reflect some aspect of exposure-response or induction time that is not captured very well by a single, baseline measurement of tea intake.

### Conclusion

In summary, the results of Sesso et al.\textsuperscript{1} constitute one of many contributions to the literature on tea in relation to CHD and stroke. The great strength of these most recent contributions is their precision. The validity considerations are mixed, as the study shares the strengths and the limitations of most of the other cohort studies in comparison with the case-control studies. Of course these results have no potential by themselves to constitute an exclusive basis for aetiological judgements or policy recommendations. Given that the follow-up period closed in 1995, 6 years before our review\textsuperscript{2} was published, we conjecture (perhaps self-flatteringly) that we might have played some small role in bringing these results to light. If so, we are grateful, for we suspect that the published results on tea in relation to CHD and stroke remain but a small and possibly unrepresentative fraction of all the results on this topic that lie resting in the figurative file drawers of epidemiological researchers around the world.

### References

3. Poole C. Low P-values or narrow confidence intervals: Which are more durable? Epidemiology 2001;12:291–94.