# A Wake-up Call to Statistical Consultants

Jonathan J. Shuster, PhD. Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida and Chris Delcher, PhD, Department of Pharmacy Practice and Science, College of Pharmacy, University of Kentucky. Correspondence: shusterj@ufl.edu

Based on a random survey of American Statistical Association (The ASA) members by Wang et al. (2018), one can infer that surprisingly many collaborator (client or colleague) requests for analysis should have aroused suspicions of possible misconduct. The goals of this follow-up analysis, using the actual survey data supplied by the corresponding author Dr. Ralph Katz, are to answer the following questions (1) How many ASA members have received at least one of three specific inappropriate requests (cited below) in the past five years? and (2) How many episodes of these requests collectively occurred in the past five years? Neither the Wang et al. (2018) article nor the accompanying editorial by Localio et al. (2018) addressed these critical questions. Briefly, the results of our analysis conservatively suggest that over 1,800 ASA members, covering over 3,000 episodes in the past five years (or 600 episodes per year), have received what some may call "nefarious-looking" requests because they seem to be intended to deceive. To illustrate the potential severity of these numbers, consider that in 2017 the Office of Research Integrity (ORI) received 215 new cases (phone, E-mail, or institutional), which may have qualified as nefarious-looking. Even if half of the consulting cases merit reporting to ORI, funder of the Wang at al. (2018) study, this would more than double their caseload.

Given the magnitude and implications of our estimates, we recommend new procedures for consultants, their institutions, and The ASA to follow to help maintain high integrity of statistical science.

## The Wang et al. (2018) Survey

The Office of The ASA Director provided the Wang team with a random sample of 4,000 ASA members. The team screened 126 of these members as ineligible because they were not primarily involved in biostatistical consulting or data analysis, leaving a frame of 3,874 members. By random sampling in sixteen 50-person batches of e-mails, the team requested a final sample of 800 members to complete the survey. Four hundred attempted to complete the survey but ten of these were excluded, leaving a final sample of 390.

The survey asked the respondent two questions about each of eighteen scenarios of analytic requests they received for "inappropriate" action: (a) frequency in the last five years: 0, 1, 2-4, 5-9, or 10+ episodes and (b) the consultant's perceived seriousness of the apparent "bioethical violation" on a scale of 0-5 with 5 being the most serious.

In our judgment, only three of the 18 questions would require immediate action to resolve possible misconduct ("nefarious-looking"). The other questions seemed mostly to deal with competency issues on the part of the client. For a full list of their 18 questions, see Table 1 of Wang et al. (2018). Survey

respondents agreed with our assessment, as these were the only three questions where 80% or more put the question as being very inappropriate (4 or 5). All other questions scored below 70%.

**The nefarious-looking questions:** How many times in the last five years were you asked to (1) "falsify the statistical significance (such as P-value) to support a desired result"? (2) "change data to achieve the desired outcome (such as prevalence rate of cancer or other disease)"? and (3)" remove or alter some data records (Observations) to better support the research hypothesis"?

While the data are insufficient to estimate the actual number of members who have experienced these episodes or total number of episodes, we can impute a meaningful and conservative (stochastically lower) outcome as follows:

For the 390 who completed the survey, we impute an outcome for each question as 0, 1, 2 (if 2-4 episodes), 5 (if 5-9 episodes) and 10 (if 10+episodes). Since for each respondent, the same episode may be reported under multiple questions, we imputed the overall episode count conservatively as the maximum imputed response for the three questions. For the 410 members sampled, who either refused to participate (400) or had a non-evaluable response (10), we conservatively imputed the response as zero. Table 1 gives the distribution of the 800 episode-count outcomes. Table 2 provides summary estimates for the analysis in the next section. The true outcomes for actual counts under a 100% survey response rate is stochastically larger than our derived outcomes. Non-responders will include an unknown but large number of members who actually do no consulting whatsoever, certainly far greater than 126/4,000. Only about 10% of the members belong to the Consulting Section of The ASA. Hence, a missing-at-random approach could well overestimate both endpoints.

Table 1: Conservative Episode Count of the 800 Sampled Individuals

| Episodes | 0 | 1 | 2 | 5 | 10 |
|---|---|---|---|---|---|
| Count (N=800) | 699 | 73 | 15 | 7 | 6 |

Table 2: Conservative Point Estimates of Mean or Proportion

| Meets Eligibility for Survey | Fraction with Episodes | Mean Number of Episodes |
|---|---|---|
| 3874/4000=0.97 | 101/800=0.13 | 198/800=0.25 |

## Point Estimates and 95% Lower confidence Intervals for Members and Episodes.

According to figures provided by The ASA, there are 17,400 current members (M).

To estimate a product of parameters, let $\hat{\theta}_1$ and $\hat{\theta}_2$ be asymptotically independent non-negative sample means (or proportions) from random samples of size $n_1$ and $n_2$ respectively from a population with M members.

Denote $E(\hat{\theta}_j) = \theta_j$ (j=1,2).

Then it follows from the central limit theorem for finite populations and the delta method that $M\hat{\theta}_1\hat{\theta}_2$ is asymptotically normally distributed with mean $M\theta_1\theta_2$ and asymptotic variance

$\hat{V}$ = (M $\hat{\theta}_2 SE_1)^2$ + (M $\hat{\theta}_1 SE_2)^2$

where $SEj^2 = Sj^2 (M- n_j)/\{(M-1) n_j\}$ with $S_j^2$ =the sample variance from the observations from sample j.

$M\hat{\theta}_1\hat{\theta}_2$ and $M\hat{\theta}_1\hat{\theta}_2$ -1.645sqrt ($\hat{V}$ ) respectively are the point estimate and 95% one-sided lower confidence limit for $M\theta_1\theta_2$ the population total.

Table 3 provides the results of our conservative estimation process.

Table 3: Conservative Estimates of members experiencing and total count of inappropriate statistical consultation episodes

| | Point Estimate | Lower 95% Confidence Limit |
|---|---|---|
| Total Members with Episodes | 2127 | 1808 |
| Total Episodes | 4171 | 3179 |

## Implications for the Consulting Community

We view these estimates (over 1,800 members experiencing over 3,000 episodes) as unacceptably high and a wake-up call for action by all of us engaged in statistical team science. We must be more proactive to greatly reduce or eliminate this behavior, not only for our clients, but for the integrity of our profession. It is not appropriate for us to resolve suspicious requests in a vacuum without a thorough and discrete assessment by university or organizational ethics officers.

We are obligated professionally to reach out when faced with a request for aiding and abetting potentially unethical conduct. However, we must not make any direct accusations of intellectual misconduct on the part of our colleagues. Examples of effective approaches for handling this type of apparent misconduct, which we presume might be a form of academic cheating for the purposes of "winning" for professional advancement, can be gleaned in other environments. For example, the American Contract Bridge League (ACBL) which has faced cheating in tournament play has a two-pronged approach that The ASA would be well-advised to consider. First, if during a tournament a potential irregularity has occurred, the tournament director is summoned. The complainant (analog of statistician) is simply requesting that the director review the facts and make a ruling. A party who is dissatisfied with the ruling can appeal it to a higher authority (for Bridge aficionados this is called an "Appeals Committee"). Second, any player (statistician) can have an episode recorded by the National Recorder. The recorder must record all requests, whether deemed frivolous or not, and have them adjudicated by experts. A single recording in of itself may have no significance but repeated recordings may result in disciplinary action. For example, a statistician who was asked to analyze a completed study that lacked any statistical input until that point, would judge whether to accept the challenge. Simultaneously, s/he would counsel the colleague against such a request in the future, suggesting ways to integrate a statistician from the study's inception. It would be ideal to document these requests so that each statistician will have proper documentation in case a collaborator tries to engage in "statistician shopping." The initial request was not unethical, but preventable subsequent requests under the same circumstances would be.

The Wang et al. (2018) article delivers a dire message about the "better scientists" who seek out expert statistical collaboration. It could not address the scarier issue involved with those scientists who do not

seek our advice. Nonetheless, the Wang team carried out an important survey about both statistical consultants and the clients that we serve. The papers are complementary, as Wang et al. consider this from the client perspective and this paper considers this from the consultant perspective.

## Google Survey of The ASA Consulting Section

In order to gain some insight into The ASA Consulting Section Members' attitudes in reporting these potential requests to higher authorities, we conducted a non-scientific Google Survey of the 1558 members of The ASA Consulting Section in four waves of requests to the Electronic Consultants' Forum. We received only 52 responses (3.3% response rate). The results of six of its seven questions are tabulated in Table 4. The seventh question is not reported per a request of Dr. Katz, Corresponding author of Wang et al. (2018).

Table 4 (Respondent Count Distribution N=50 evaluable of 52 responses)

| Question | Yes | No | Unsure |
|---|---|---|---|
| 1.(Wang #1) "Falsify the statistical significance (such as P-value) to support a desired result." Would you report the client requesting this to university or company officials? | 30 | 8 | 12 |
| 2. (Wang #2) "Change data to achieve the desired outcome (such as prevalence rate of cancer or other disease)". Would you report the client requesting this to university or company officials? | 33 | 7 | 10 |
| 3. (Wang #3) "Remove or alter some data records (Observations) to better support the research hypothesis". Would you report the client for requesting this to university or company officials? | 24 | 8 | 18 |
| 4. Would you consider a Consultant who failed to report any of these violations to be guilty of scientific misconduct? | 23 | 11 | 16 |
| 5. Do you think a Consultant can infer a motive on the part of the client in Wang Survey Question #3, when you find out data were changed? | 16 | 16 | 18 [1] |
| 6. Do you think these three questions were properly worded, so that the Respondent understood his/her response? | 26 | 19 | 5 |

[1] One respondent did not fill in a response and was put into unsure; Two more did not answer any question but had comments. These were excluded from the table.

Note: 20(5) of the 50 Respondents said yes(no) uniformly to the first three questions (Wang #1, Wang #2, and Wang #3) respectively.

Although the low response rate may indicate considerable apathy to the issues we raise in the Wang et al. (2018) survey as they apply to the Statistical Consulting community, those responding for the most part do not support giving the client a free pass on the three nefarious looking requests. Furthermore, we are alarmed that about one-fifth of respondents would **not** consider failure to report the incident as misconduct on the part of the statistical community (Question 4).

## Limitations

Our analysis has two limitations that are beyond anything mentioned in either of the parent articles. First, because the survey retrospectively requested respondents to estimate their five-year experience, respondents could well have had recall bias, especially with respect to number of episodes and whether

they occurred within the five-year window. However, it seems likely that the estimate of whether or not the member had at least one of these potentially nefarious requests should be viewed as a still more conservative estimate of their career-long experiences. The second limitation is of greater concern. The questions seem to require an inference on the part of the consultant about the purpose of the nefarious looking request. For example, if you were asked to remove or alter some records (an affirmative answer to the first part of Question 3), how does the consultant infer that the purpose was "to better support the research hypothesis"? The ability of the respondent to understand intent seems uncertain at best. Question 6 of our Google survey supports the foundation of this concern with only about one-half of respondents finding the Wang questions clearly worded.

## Acknowledgements:

We sincerely appreciate the help of The ASA Webmaster, Ryan Bell, for getting us the data on the total numbers of The ASA and member of The ASA Statistical Consulting Section. Thanks go to Dr. Christine Ring, Health Scientist Administrator, Office of Research Integrity. She supplied the 2017 ORI caseload statistic. We appreciate the help of Dr. Ralph Katz, Professor of Epidemiology & Health Promotion, New York University College of Dentistry, for supplying de-identified respondent information on Wang #1-Wang #3 questions as well as constructive suggestions on our manuscript. Thanks also go to the 52 members of the American Statistical Association Consulting Section who completed our Google Survey. Finally, thanks go to Dr. Janet Turk Wittes, Statistics Collaborative Inc. for her review of an early draft of the paper and for her suggestions.

## Funding:

This project was entirely self-funded, and where opinions are stated, they are strictly those of the authors.

## References

Wang M.Q., Yan A.F., Katz R.V. 2018. Researcher Requests for Inappropriate Analysis and Reporting: A U.S. Survey of Consulting Biostatisticians. *Ann Intern Med.* 169(8):554-558. doi: 10.7326/M18-1230. Epub 2018 Oct 9. PubMed PMID: 30304365.

Localio A.R., Stack C..B., Meibohm A.R., Ross E.A., Guallar E., Wong J.B., Cornell J.E., Griswold M.E., Goodman S.N. 2018. Inappropriate Statistical Analysis and Reporting in Medical Research: Perverse Incentives and Institutional Solutions. *Ann Intern Med*. 169(8):577-578. doi: 10.7326/M18-2516. Epub 2018 Oct 9. PubMed PMID: 30304363.