Part 2—Unusual clustering of coefficients of variation in published articles from a medical biochemistry department in India

Joyce C. McCann,[1] Mark L. Hudes, and Bruce N. Ames[1]
Children's Hospital Oakland Research Institute, Oakland, California, USA

Key Words: lipoic acid · oxidative stress · antioxidants · aging · data variability · data manipulation

Drs. Varalakshmi, Panneerselvam, and Sakthisekaran (VPS) have not responded to the substance of our article (1). In fact, their intended rebuttal merely restates what we did and, evidently unintentionally, highlights a major strength of our analysis. Their major critique is that "the statistical methodology . . . seems problematic since heterogenous data have been arbitrarily pooled and a single set of 50% boundaries for the pooled coefficients of variation (CVs) have been constructed" assuming that "these data derive from a single normally distributed population." VPS point out that assuming the data derive from a single normally distributed population is "unacceptable" because of the "heterogeneous biochemical experiments in terms of animal models (species, age, sex, body weight, number of animals in each group), pathologies, treatment groups, types of biochemical analyses, assay protocols, instrumentation, *etc.*"

We couldn't agree more that the assumption that all data in a study measuring heterogeneous variables derive from a single normally distributed population is extremely improbable. The improbability of this assumption is in fact one of the key elements of our analysis. Thus, results in our article illustrate that, despite measuring multiple heterogeneous variables, the overall variability of results reported by VPS in the majority of their articles published after January 2000 is *even less* than would have been expected had they measured only a single normally distributed variable.

The extraordinarily tiny variability of CVs in articles of VPS compared to samples of articles from other laboratories measuring similar variables and to their own articles published prior to 2000 is illustrated in **Table 1**.

As shown in Table 1, CVs were aberrantly clustered in only one of the 26 articles we analyzed from other laboratories. This article, which turned up in a key word search we used to select control articles, was from a laboratory located on the same campus as VPS. Also, none of the articles we analyzed that were published by VPS prior to 2000 contained CVs that were aberrantly clustered. However, 61% of articles analyzed that were published by VPS after January 2000 were aberrantly clustered. Moreover, the clustering was exceptionally tight in many of these articles, as indicated by the extremely small *P* values. Thus, the discrepancy in these highly clustered articles on the one hand, compared to articles from other laboratories and even to VPS's own earlier work on the other, is dramatic.

In the remainder of this reply, we address two issues. First, the methodology may be unfamiliar to some readers, so we try to explain it in nontechnical language. Second, we present the results of a simulation, in addition to those presented in our article, that evaluates two statistical aspects of our analysis that could theoretically result in increased likelihood of CVs falling within the 50% limits.

METHODOLOGICAL EXPLANATION IN NONTECHNICAL LANGUAGE

If any measurement is taken multiple times, say, for example, if the activity of a particular enzyme is measured in the blood of 6 rats, the measurement will not be the same for every rat; multiple factors will contribute to some variability. Results of this kind of experiment are often presented as a mean (*i.e.*, an average) and a measure of variability, often the SD. The coefficient of variation (CV), when expressed as a proportion, is defined as the ratio of the SD to the mean; this normalizes the SD to the mean, thus making the CV a powerful tool to compare standardized variability among different experimental systems. In the rat experiment, the CV would be the ratio of the SD to the mean of the 6 enzyme activity measurements.

If multiple sets, each $n = 6$, of these means and SDs are sampled from a single population of rats, the resulting collection of CVs will have its own distribution. From such a distribution, using assumptions we discuss below, we computed a range of values (limits) around the mean CV within which sample CVs would

[1] Correspondence: Children's Hospital Oakland Research Institute, 5700 Martin Luther King Jr. Way, Oakland, CA 94609, USA. E-mail: J.C.M., jmccann@chori.org; B.N.A., bames@chori.org

doi: 10.1096/fj.08-123117

0892-6638/09/0023-0706 © FASEB

TABLE 1. *Aberrant clustering ($P<0.01$) of CVs in articles from VPS compared to other laboratories*

"*P* value"	Other laboratories	Laboratories of VPS: Published before January 2000	Laboratories of VPS: Published after January 2000
$<10^{-8}$			16
$<10^{-4}$	1[a]		17
0.0001–0.0009			3
0.001–0.009			4
0.01–0.049		1	5
0.05–0.09			1
0.1–0.49	4	2	7
>0.5	21	15	13

"*P* values" were calculated as described in Hudes *et al.* (1). [a] From a laboratory on the same campus as VPS.

be expected to fall a certain percentage of the time. We could have chosen any percentage, but 50% was a convenient number. Hence we call this range of values the "50% limits."

Two key assumptions were made in calculating the 50% limits that highlight why we think results reported by VPS are so aberrant. Both of these assumptions are conservative in that they yield 50% limits that will generally result in an *underestimation* of the actual amount of variability—*i.e.,* in real experiments, we expected that <50% of CVs would be likely to fall within the limits.

The first assumption was that the underlying data in an article were normally distributed. However, in most biological experiments, data are not normally distributed. Most biological variables have distributions with heavier tails than a normal distribution (*i.e.,* proportionately more measurements are further from the mean than if the data were normally distributed). In our article (1), we conducted simulations to confirm that ~50% of CVs calculated from normally distributed data fell within the 50% limits, and that CVs from distributions with heavier tails were generally more likely to fall outside the 50% limits.

The second assumption, which was the focus of the critique of VPS, was that all variables in any article came from the same single distribution. But, as VPS point out, any collection of different variables (*e.g.*, different enzyme activities) are extremely unlikely to come from the same single distribution. Even if different biological variables have normal distributions, it is highly unlikely they will have the same mean and SD (or the same CV). The violation of this assumption, as was uniformly the case in all articles we analyzed, also leads to fewer CVs falling within the 50% limits.

Thus, in our analysis, we purposely, as VPS point out, "pooled CVs of heterogeneous variables having different distributions". In fact, we included all CVs that could be calculated from data presented in data tables in each article. We therefore anticipated that the assumption of a single normal distribution of data would be violated, and we expected that the percentage of CVs falling within the limits would be unlikely to significantly exceed 50%, and would in almost all cases be significantly less than 50%.

As discussed above, this expectation was borne out for articles from other laboratories and also for all articles of VPS analyzed that were published prior to 2000. In contrast, the majority of articles analyzed from the laboratories of VPS that were published after 2000 showed aberrant clustering of CVs. *The key point is that, since the 50% limits were constructed so conservatively, more CVs in these aberrantly clustered articles fell within the limits even than would have been expected had the underlying distribution of the multiple variables in each article actually been a single normal distribution.*

ADDITIONAL SIMULATION

This simulation was performed at the suggestion of a *FASEB Journal* reviewer to evaluate two statistical aspects of our analysis that, as indicated in our article (1), could theoretically increase the likelihood of CVs falling within the 50% limits: *1*) measuring multiple end points (*e.g.*, different enzyme activities) on the same individuals, which characterized all of the articles we analyzed; and *2*) using the sample median CV from each article as an estimate of the population CV. For reasons discussed in our article, we expected these 2 effects to be small, but an additional simulation was performed to verify this expectation.

We examined the NHANES data (2003–2004 release) from a single lab (lab 13), which had measurements on 27 different variables. We included only subjects that had complete information on all 27 variables. This resulted in over 3000 subjects. We generated 1,000 data sets, each consisting of 4 groups of 6 subjects each. For each of these groups, we computed the CV for all 27 variables. Thus, for each data set there were 108 CVs computed. Then, 50% intervals were constructed, using the single normal distribution assumption and using the median of the 108 CVs as an estimate of the population CV for each data set. Out of a total of 108 CVs for each data set, the number of CVs that fell within the 50% limits ranged from 9 to 34 (8.3 to 31.5%). Thus, all 1000 data sets had fewer than 50% of CVs inside the limits. All computed "*P* values" were >0.9. These results suggest that the effect of measuring different variables, which would tend to decrease the number of CVs inside the 50% interval, more than compensates for effects of either nonindependence or use of the sample median CV, which could tend to increase the number of CVs in the interval.

CONCLUSION

Our study demonstrates that the variability of CVs in the majority of articles published by VPS after 2000 is very small in comparison with theory, with articles from

other laboratories, and with their own articles published prior to 2000. The obvious question is why, but unfortunately, their response does not offer an explanation.

Over the course of this project we have been frustrated by the lack of any independent entity that could have provided assistance and guidance. The National Institutes of Health Office of Scientific Integrity was unable to help because they did not fund the work in question. As far as we have been able to discern, the primary oversight responsibility on published articles lies with individual journal editors, and we have shared our analysis with the editors of journals in which articles with aberrant clustering appeared. We also shared information with the Committee on Publication Ethics (COPE) (an organization for journal editors), and with Elsevier, which publishes the majority of the >50 journals in which articles from VPS have appeared. Based on our experience, we think an international organization is needed to serve as a clearing-house specializing in assessment and follow-up of cases such as this. Perhaps an organization such as COPE could take on this function. FJ

REFERENCE

1. Hudes, M. L., McCann, J. C., and Ames, B. N. (2009) Unusual clustering of coefficients of variation in published articles from a medical biochemistry department in India. *FASEB J.* **23,** 689–703

Received for publication September 25, 2008.
Accepted for publication September 25, 2008.