

Statistical Detection of Potentially Fabricated Numerical Data: A Case Study

By

Joel H. Pitt¹ and Helene Z. Hill²

¹ Renaissance Associates, Princeton, NJ drjhpitt@yahoo.com

² Rutgers University, NJ Medical School, Newark, NJ 07101-1709 hill@njms.rutgers.edu

Abstract

Scientific fraud is an increasingly vexing problem. Many current programs for fraud detection focus on image manipulation, while techniques for detection based on anomalous patterns that may be discoverable in the underlying numerical data get much less attention, even though these techniques are often easy to apply. We employed three such techniques in a case study in which we considered data sets from several hundred experiments. We compared patterns in the data sets from one research teaching specialist (RTS), to those of 9 other members of the same laboratory and from 3 outside laboratories. Application of two conventional statistical tests and a newly developed test for anomalous patterns in the triplicate data commonly produced in such research to various data sets reported by the RTS resulted in repeated rejection of the hypotheses (often at p-levels well below 0.001) that anomalous patterns in his data may have occurred by chance. This analysis emphasizes the importance of access to raw data that form the bases of publications, reports and grant applications in order to evaluate the correctness of the conclusions, as well as the utility of methods for detecting anomalous, especially fabricated, numerical results.

Key words: statistical forensics, data fabrication, tissue culture, triplicate colony counts, terminal digit analysis, radiation biology, cell biology

1. INTRODUCTION

During the past decade, retractions of scientific articles have increased more than 10-fold (Van Noorden 2011). At least two-thirds of these retractions are attributable to scientific misconduct: fraud (data fabrication and falsification), suspected fraud, duplicate publication, and plagiarism

(Fang, Steen et al. 2012). Techniques for early identification of fraudulent research are clearly needed. Much current attention has been focused on sophisticated methods for detecting image manipulation (Rossner and Yamada 2004) and their use is encouraged on the website of the Office of Research Integrity (ORI) of the United States Department of Health and Human Services. But statistical methods which can readily be used to identify potential data fabrication (Mosimann, Wiseman et al. 1995; Mosimann, Dahlberg et al. 2002; Al-Marzouki, Evans et al. 2005; Baggerly and Coombes 2009; Hudes, McCann et al. 2009; Carlisle 2012; Simonsohn 2012) are all but ignored by the ORI and the larger world. We believe that routine application of statistical tools to identify potential fabrication could help to avoid the pitfalls of undetected fabricated data just as tools such as, for example, CrossCheck and TurnItIn are currently used to detect plagiarism.

The first step in using statistical techniques to identify fabricated data is to look for anomalous patterns of data values in a given data set (or among statistical summaries presented for separate data sets), patterns that are inconsistent with those that might ordinarily appear in genuine empirical data. That such patterns are, indeed, anomalous may potentially be confirmed by using genuine data sets as controls, and by using simulations or probabilistic calculations based on appropriate models for the data to show that they would not ordinarily occur in genuine data.

The existence of these anomalous patterns in given suspect data sets may be indicative of serious issues of data integrity including data fabrication (Al-Marzouki, Evans et al. 2005), but they may also arise as a result of chance. Hence it is of considerable importance to have statistical methods available to test the hypothesis that a given anomalous pattern in a data set may have occurred as the result of chance.

For example, Mosimann *et al.* (Mosimann, Dahlberg et al. 2002) identified instances of fabricated data based on the observation that in experimental data sets containing count data in which the terminal (insignificant) digits are immaterial and inconsequential (hence not under the control of the investigator) it is reasonable to expect and generally the case that these inconsequential digits will appear to have been drawn at random from a uniform population. When terminal digits of the count values in a data set of this type do not appear to have been drawn from a uniform population (as may be tested using the Chi-square goodness of fit test) this may indicate that they have been fabricated.

A test like this is not entirely foolproof. Before applying it, one must ask whether there really is any evidence, beyond mere supposition, that terminal digits of data of the given kind should be random in the sense of uniform. Ideally one would like to have a probability model for the underlying randomness in the experimental data and use it to show that the distribution of terminal digits of counts values in data sets consistent with that model will be uniform (Hill and Schürger 2005). Alternately one might be able to run simulations based on an appropriate probability model and demonstrate that the terminal digits of the counts in the simulated data sets do generally appear to have been drawn uniformly. Finally, one could try to validate the assumption that terminal digits of counts in legitimate data sets are uniform, empirically, by testing the uniformity of terminal digits in indisputably legitimate experimental data sets of exactly the same type, constructed using the same protocols, as that of the suspect data.

Simonsohn (Simonsohn 2012) uncovered fabrications in several psychological research papers based entirely on the summary data available in published reports. He noted that despite the fact that the means of various variables measured in the study varied considerably, their standard

deviations were remarkably similar, and hypothesized that this would not be the case were the results derived from genuine experimental data. He confirmed his hypothesis with simulation and empirical observation of the distribution of standard deviations in comparable studies.

When we have an appropriate probability model available for the underlying experiment that purportedly produced the suspect data, we can often apply our knowledge of probability theory to determine the probability that an anomalous pattern in question may have occurred by chance in the data set under consideration. Where that probability is less than some reasonable level, we term our tests significant, and, in the absence of any alternative explanation, may find any such significant results convincing evidence that the data in question has been fabricated.

2. The Case Study: Concerns about the legitimacy of raw data generated by one Research-Teaching Specialist (RTS) in the laboratory in which one of us was a member led us investigate data sets of his which had been used in several publications, a grant application and its renewal. We also had access to data sets generated by nine other researchers in the same laboratory who followed the same or similar protocols, as well as data from three outside laboratories that employed similar techniques. By applying the same investigating techniques to their data sets, we were able to use them as controls. Copies of the laboratory notebooks containing the raw data that we analyzed were in the form of PDF files which we transferred into Excel spreadsheets (cf Supplementary Material).

We believe that this was a unique situation, as we were able to review and compare essentially all the data from a single laboratory, data produced by a number of independent investigators using the same or similar research techniques, over such a long period of time. In particular it allowed us to determine whether suspect patterns that we had already noted in a limited number

of data sets from the RTS whose data had raised the initial concerns appeared in other data sets of his and whether the same patterns might be found in the data sets from the other investigators.

These other than expected patterns in the RTS's data included: (1) a non-uniform distribution of insignificant terminal digits; (2) an unusually large frequency of equal terminal digit pairs (i.e. equal right-most and second right-most digits); and (3) a surprisingly large number of triplicate colony and cell counts in which a value near the average value of the triple or even that average value appeared as one of the constituent counts of the triple.

None of these patterns were evident in any of the data sets reported by the nine other investigators in the same laboratory, or in data sets that we obtained from three other independent, outside researchers. We believe this is a matter of significant concern.

We can use the well-known chi-square goodness of fit test to determine whether non-uniformity of terminal digits can be considered significant. Additionally, a straightforward test of significance based on the binomial distribution can be used to test the significance of an unusually high frequency of equal terminal digit pairs, but there is no such standard test to determine the significance of unusually large numbers of triplicate counts containing values near their average. Random variation in these triplicate data that are common components of pharmacological, cell biological and radiobiological experimentation, can be analyzed by modeling the triples as sets of three independent, identical Poisson random variables. A major focus of this study is on developing a method to calculate bounds and estimates for the probability that a given set of n such triplicates contains k or more triples which contain their own mean. We use these bounds and estimates in tests of the hypothesis that the observed

unusually high incidence of mean containing triples in certain data sets may have occurred by chance.

Our methods should be useful to laboratory investigators in therapeutic, toxicological, cell and radiation biological studies involving evaluation of cell survival after various treatments. Much of our analyses pertain to triple replicates such as are commonly used in cell survival protocols (Bonifacino 1998; Munshi, Hobbs et al. 2005; Katz, Ito et al. 2008).

3. Experimental Protocols: The experiments we analyzed followed the same or very similar protocols and employed, with few exceptions, the same Chinese hamster cell line. The cells, harvested from mass culture, were counted, apportioned equally into culture tubes and incubated overnight with radioisotopes. They were washed free of radioactivity and transferred to new tubes for a 3-day incubation at low temperature (10.5° C) to allow for the given isotope to decay. They were then harvested, triplicate aliquots were suspended for cell counts using a Coulter ZM particle counter and aliquots were diluted and plated onto tissue culture dishes in triplicate in order that single cells could grow into colonies which were stained and counted (manually) after about a week.

4. Data sets and Probability Model: The primary data sets with which we are concerned are collection of triples of integer Coulter ZM counts and triples of colony counts. The former are copied by hand into a notebook from an LED digital readout of the Coulter ZM counter that counts single cells as they pass randomly through a narrow orifice, the latter are counted by hand. The colony triples are counts of the number of colonies formed by the surviving cells. The counts in each Coulter triple and each colony triple are modeled probabilistically as independent,

identical Poisson random variables. The Poisson parameter of these triples will, of course, vary from triple to triple.

Throughout this report, the accumulated data from the RTS's experiments are independently paralleled to the accumulated data of other investigators including nine members of the laboratory other than the RTS who utilized the same Coulter counter and/or counted colonies in the same manner, two professors from out-of-state universities who contributed triplicate data from their Coulter ZM counters, and triplicate colony counts from an additional independent laboratory.

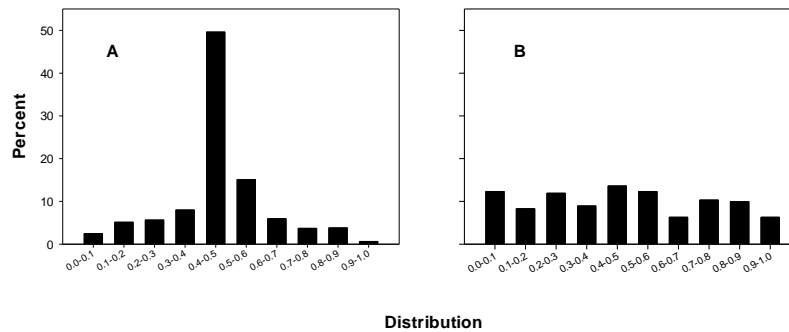
5. Analysis of Triplicate Data: Many radiobiological experiments result in data sets consisting of triplicate counts where the means of the triples are the key values that are associated with the corresponding treatments in subsequent analyses. An investigator wishing to guide the results of such experiments would have to arrange the data values in each of the triples so that their means are consistent with the desired results. The quickest and easiest way to construct such triples would be to choose the desired mean (or a close approximation) as one of the three count values and then, using two roughly equal constants, calculate the other two values as this initial value plus or minus the selected constants.

Data sets constructed in this manner might then be expected to include either (1) an unusually high concentration of triples whose **mid-ratio** (the ratio of the difference between the middle value and the smallest value to the difference between the largest value and the smallest value (the **gap**) was close to 0.5; or (2) an unusually large number of triples that actually include their own (rounded) mean as one of their values.

5.1 Initial mid-ratio review: Having observed what appeared to us to be an unusual frequency of triples in RTS's data containing a value close to their mean, we used R to calculate the mid-

ratios for all of the colony data triples that were available to us. We then constructed histograms of the resulting data sets. The results are shown in **Figures 1A** and **1B**. The histogram of mid-ratios for RTS's colony triples exhibits a distinct predominance of mid-ratios in the range 0.4 to 0.6, while the histogram of mid-ratios of the data triples recorded by the nine other members of the laboratory is fairly uniform over the ten sub-intervals. The dramatic contrast between the two histograms seems a clear indication that RTS' data may have been manipulated to guide the mean values of its triples.

Fig. 1: Distributions of the mid-ratios $(\text{middle} - \text{low})/(\text{high} - \text{low})$ for colony triples **A.** RTS, 1343 triples, 128 experiments; **B.** Other investigators, 572 triples, 59 experiments.



5.2 Appearance of the Mean in Triplicate Samples: We extended our investigation by writing an R program to identify and count triples that contained their rounded average. (**Figure 2** is a scan of a page from one of the RTS's notebooks. Triples that contain their rounded average are highlighted in blue. In this instance six of the ten triples are highlighted.) Of the 1343 complete colony triples in RTS's data, 690 (more than 50%) contained their rounded average, whereas only 109 (19%) of the 572 such triples from other investigators did.

Figure. 2: PDF Image of Colony Counts from an experiment performed by RTS. The rounded average (highlighted in blue) appears as one of the triplicate counts in 6 of the 10 samples (Ppoibin Prob = 0.00169, See Section 5.7, below.).

TABLE-4
 Expt # : 2
 Date : 12/11/18

Tube/dilution	Colony 1	Colony 2	Colony 3	Avg Colony	SE
1-2	120	111	99	116.5	
2-2	131	121	117		
3-2	104	114	96	104.6	0.8984
4-2	93	100	86	93	0.7982
5-2	78	78	63	70.3	0.6037
6-2	58	68	49	58.3	0.5007
7-2	29	22	19	23.3	0.2002
8-3	104	115	94	104	0.0895
9-4	118	126	107	116	0.0099
10-4	238	219	18	0.23	0.009

Given the marked difference between the percentage of the RTS's triples that contain their mean and the corresponding percentage of other investigators' triples that do so, and the similar disparity between the histograms of mid-ratios of the RTS's triples and those of other investigators, it is reasonable to ask whether the apparently excessive numbers of mean/near mean containing triples in the RTS's data sets might plausibly have occurred by mere chance. In order to answer that question we used a probability model for such triplicate data to calculate bounds and estimates of the probability that a given set of n such triplicates contains k or more triples. Using these estimates we are able to test the chance hypothesis.

5.3 The Model for Triplicate Data: The differences between the three actual count values in each colony count triple arise from random differences in the number of cells actually drawn and

transferred to the three dishes and the randomness in the numbers of cells that survive the treatment applied to the cells in that triple. As noted above the random variables that correspond to the number of cells that are originally in each of the three dishes can be modeled probabilistically as the values of three independent, identical Poisson random variables. The common Poisson parameter λ_0 of those three variables will be the (unknown) expected value of those cell counts.

Since the cells in the three dishes have all been exposed to the same level of radiation, the probabilities that a given cell survives to generate colonies should be the same in each of the three dishes. Accordingly, the actual number of survivor colonies in the three dishes will have a binomial distribution with the same p parameter (the common individual cell survival rate) and differing n values corresponding to the numbers of cells on each dish. It is easy to show that these resulting counts have Poisson distributions with parameter $\lambda = \lambda_0 p$.

Thus the three values in each set can be modeled as the values of three independent Poisson random variables sharing a common parameter λ . The actual value of λ varies from triple to triple as it depends both on the specific λ_0 associated with the initial cell count Poisson distribution and the specific p associated with the treatment which gave rise to the given triple. The likelihood that one of the counts in the triple is equal to or near the triple mean value depends on the value of this parameter.

Given the value of their common Poisson parameter λ a relatively straightforward calculation can be used to find the probability that a triple generated by independent, identical Poisson random variables includes its mean (see Appendix). The values of the various λ parameters of the Poisson random variables that gave rise to the triples in our data set are, of course, unknown to

us, but, in as much as actual colony count values are all less than 400 we can safely assume that the λ parameters of the underlying Poisson random variables are certainly less than 1000.

We wrote an R program to calculate the probability that a triple generated by independent, identical Poisson variables with known parameter λ includes its own (rounded) mean value and used it to calculate and create a table (referred to below as the **MidProb** table) of this probability for all integer values of λ from 1 to 2000 and, as the variation of these probabilities between successive integer values of λ greater than 2000 was negligible we extended the table by calculating the value of the probability for values of λ that were multiples of 100 between 2100 and 10000, and multiples of 1000 between 11000 and 59000 (see **Table 1** for the first 25 entries). Our calculations showed that as λ increased from 1 to 3, the probability that a randomly generated triple contains its own mean increases from about 0.27 to slightly more than 0.40 and then decreases as λ continued to increase. We were thus assured that no matter what the value of λ for the Poisson variables that generated a given triple, the probability that the triple would have included its mean as one of its three elements would not exceed 0.42.

5.5 Hypothesis testing I -- A non-parametric test: The observation that the probability that a triple generated by independent, identical Poisson variables with known parameter λ includes its own (rounded) mean value never exceeds 0.42 gives us the ability to construct a crude test of the hypothesis that an observed, suspect high number of mean containing triples in a given collection of triples may have occurred by chance. Using the number k of triples with gap two or more that contain their means and the number n of triples in the collection we need only find the binomial probability p of k or more successes in n independent Bernoulli trials where the probability of success is 0.42. If the probability p is less than the chosen α level of the test we reject the null

hypothesis at that significance level. The test is crude in the sense that the calculated p is not the p-level of the test, it is simply a (possibly gross) over estimate of the p-level.

Table 1. Partial MidProb Table. Probability that a triple generated by 3 independent Poisson random variables with parameter λ contains its mean for $\lambda = 1$ to 25. It is clear that λ continues to decrease after $\lambda = 4$.

λ	P	λ	P	λ	P	λ	P	λ	P
1	0.267	6	0.372	11	0.317	16	0.281	21	0.254
2	0.387	7	0.359	12	0.309	17	0.275	22	0.250
3	0.403	8	0.348	13	0.301	18	0.269	23	0.246
4	0.397	9	0.337	14	0.294	19	0.264	24	0.242
5	0.385	10	0.327	15	0.287	20	0.259	25	0.238

When we apply this test to determine how likely it is that 690 or more of the 1343 colony triples in RTS's data might have contained their rounded average by chance, we find that it is less than 2.85×10^{-12} , an extremely significant result.

Since there are only 109 mean containing triples among the 572 from other investigators, and 109 is considerably less than the expected number of successes in 572 Bernoulli trials with a success probability of 0.42 it is immediately clear that the probability of having 109 or more mean containing triples is reasonably large -- indeed it is essentially one.

5.7 Hypothesis testing II -- Using λ to obtain p-values: It is important to have a more sensitive test, as we can use it to confirm the validity of our model by applying it to what we believe to be legitimate experimental data. To do so we use a heuristic method to estimate the actual probability that a given collection of n triples includes k mean containing triples. This allows us to provide an actual p-value for the one-tailed test we apply for seemingly high numbers of mean

containing triples and thereby allows us to determine whether the numbers of mean containing triples in our controls are consistent with our model or whether they are also significantly different from what our model indicates.

We start with the observation that the results of our calculations in the **MidProb** table show that the probability that a triple of independent, identical Poisson random variables includes its own mean decreases rapidly as λ increases. For example the probability that a triple contains its mean if it is generated by Poisson random variables with $\lambda = 20$ is about 0.26, but with $\lambda = 50$ it is less than 0.18, and with $\lambda = 100$ it is less than 0.14 (it even falls to 0.032 when $\lambda = 2000$).

We applied a heuristic approach to use our table of calculated values of this probability to estimate (rather than merely bound) the probability that a given collection of n triples that are hypothesized to have been drawn as triples of independent, identical Poisson variables has as many or more than the actual number of mean containing triples than it was observed to contain. We do not know the actual λ values of the Poisson random variables that (hypothetically) generated the triples in the data sets, but the mean of any actual triple is a reasonable estimate of the λ parameter of the variables that gave rise to it. (The mean is the maximum likelihood estimator in this case.). We can then look up these (rounded) λ values in the **MidProb** table to obtain an estimate of the probability that had the triple been randomly drawn it would contain its own mean.

We are thus able to consider the events that the various individual triples of the collection contain their own means as successes in individual, independent, Bernoulli trials each with a known probability of success. The random variable (statistic) which takes as its value the number of triples in the given data set that contain their own means is the sum of the Bernoulli random variables that indicate success in the various trials. These Bernoulli trials have the known

(actually estimated) probabilities of taking the value 1 that we obtain from the **MidProb** table as described above.

Sums of such independent, not necessarily identically, distributed Bernoulli random variables are said to be Poisson binomial random variables and to have a Poisson binomial distribution. A Poisson binomial random variable that is the sum of n Bernoulli random variables can potentially take any of the values $0, 1, \dots, n$, and the probability that it takes, or is greater than, or equal to any of these potential values is completely determined by the probabilities p_1, p_2, \dots, p_n that the constituent Bernoulli random variables take the value 1.

Few (if any) standard statistical packages include functions for calculating Poisson binomial distributions. Although there is a straightforward algorithm which can, in principle, be used to calculate probabilities for the distribution function of a Poisson binomial random variable given the success probabilities of the individual Bernoulli variables p_1, p_2, \dots, p_n , issues of numerical stability in these calculations can arise for even moderately large values of n , and processing times increase exponentially as n increases. Nonetheless, we were able to take advantage of an efficient algorithm that has recently been developed and implemented as a package for R (Hong 2011) to find exact values for the tail p-values that we wish to have in testing our null hypothesis.

The function *ppoibin* in the R package *poibin* accepts as input two parameters, an integer j and a vector of probabilities p_1, p_2, \dots, p_n and returns the probability that the Poisson binomial random variable that corresponds to that vector of probabilities takes a value less than or equal to j . To use it to find the probability that there are k or more mean containing triples in a collection of triples generated by groups of three Poisson random variables with common probabilities

p_1, p_2, \dots, p_n , we execute *ppoibin* with the value $j=k-1$ as the first parameter and the given probabilities as the second and subtract the result from 1.

We applied this more refined test to the RTS collection of 1343 complete colony triples and found that, given the likely λ values that had given rise to the individual triples in the collection, the probability of the observed 690 or more mean containing triples is approximately 6.26×10^{-13} (not surprisingly an extremely significant result). Applying the same test to find the probability of finding 109 or more mean containing triples among the 572 complete colony triples that had been recorded by the other investigators in the same laboratory, we found that the probability was 0.47, and the probability of 109 or fewer such triples is 0.58; results that are entirely consistent with our hypothesis.

5.8 Hypothesis testing III -- Normal estimation of p-values: Given the success probabilities of the individual Bernoulli variables p_1, p_2, \dots, p_n the expectation of their Poisson binomial sum is $\mu = \sum_{i=1}^n p_i$ and its variance $\sigma^2 = \sum_{i=1}^n p_i (1 - p_i)$. Both are easy to calculate. When the values of the p_i 's are bounded below, the (Lindeberg-Feller) Central Limit Theorem applies and we can obtain reasonable approximations of the (upper) tail probabilities of a Poisson binomial random variable using normal probabilities.

Where an efficient implementation of an algorithm for calculating exact Poisson binomial probabilities is not available, we can use a normal approximation which with a second order correction (Volkova 1995) provides a quite precise estimate. Hong (2011) reports the results of multiple simulations that indicate that by including the second order correction the normal approximations to upper tail probabilities will usually -- but certainly not always -- return probability values marginally higher than the true tail probabilities. The normal distribution we

use to approximate a Poisson binomial is the normal with the same mean and standard deviation as the Poisson binomial.

Using the normal approximation has a second advantage, in as much as the z-values we calculate in order to look up normal probabilities are informative without recourse to an actual table of normal probabilities. Virtually all students of statistics learn that in normal populations upper tails corresponding to z-scores of 2 or more or 3 or more are quite unlikely -- with the first having a probability of less than 0.025 and the second having a probability of less than .0015.

To use this approach to approximate the probability of the 690 or more mean containing triples among the RTS' 1343 complete triples, we first obtain (to two decimal places) $\mu=220.31$ and $\sigma=13.42$. Using a standard correction for continuity, the z-value we use to find the probability of 690 or more mean containing triples is $\frac{689.5-220.31}{13.42} = 34.97$ so large that the upper tail probability is effectively indistinguishable from 0, hence significant at virtually any level.

It is important to keep in mind that the normal distribution probabilities are approximations, not exact values, of the Poisson binomial probabilities. Unfortunately the normal approximations of upper tail Poisson binomial probabilities are generally less than the true values. In this instance, however, the aforementioned Volkova correction provides the same estimate.

5.9 Application to Coulter Counts: While the means of colony triples are the key values of interest to investigators, means of Coulter triples are not as significant. Thus there is less reason to believe that an investigator wishing to guide results might be inclined to construct Coulter triples that include their own means as one of their values. Nonetheless we extended our investigation and counted the number of mean contain triples in both the RTS' Coulter triples and

those from other investigators. The results are interesting and illustrate the power and importance of the more sensitive tests we discussed in 5.7 and 5.8 above.

Coulter data from the RTS included 1717 complete triples, 173 of which included their rounded mean, while we had 929 complete Coulter triples from other investigators in the same lab, 36 of which included their rounded means. Application of the crude test described in 5.6 gives no reason for concern as in both cases the numbers of mean containing triples are consistent with our belief that the probability that any given triple includes its mean will be less than 0.42.

When, however, we apply the more refined analysis introduced in sections 5.7 and 5.8, we find reason once again to question RTS' data. Coulter count values are in a much higher range than colony count values, thus the Poisson random variables that give rise to them have λ values in a higher range and probabilities that Coulter triples include their means tend to be lower. Using our table of probabilities, triples of independent Poisson random variables with given λ parameters that contain their own mean, we found that were we to randomly generate 1717 Poisson triples with respective λ parameters set equal to the means of the RTS' actual triples the expected number of mean containing triples would be 97.74 and the standard deviation 9.58. Given this (and using the normal approximation to the Poisson binomial) the 7.80 z-score that corresponds to the actual number of 173 mean containing triples in the RTS' data makes it immediately clear that it is exceedingly unlikely we might have encountered such a large number of mean containing triples by chance. The actual Poisson binomial tail probability is 6.26×10^{-13} .

When we apply the same analysis to the Coulter triples we obtained from other investigators in the same lab the results are well within the expected range. According to our calculations the expected number of mean containing triples would be 39.85 and the standard deviation is 6.11.

Hence the z-value corresponding to the actual number of 36 mean containing triples is -0.71 and the actual p-value is 0.76, entirely consistent with our model.

We applied the same analysis to the triplicate Coulter count data sets we had from two investigators in other labs and triplicate colony counts from an investigator in another lab and the results for all of these sets are summarized in **Table 2** below.

Table 2: Summary results for analysis of mean containing triples for colony and Coulter count triples from RTS, 9 other investigators from the same lab, and investigators in outside labs

TYPE	INVESTIGATOR	NO. EXPS	NO. COMPL/TOTAL	NO. W MEAN	NO. EXPECTED	STD	Z-VALUE	P _≥ K
COLONIES	RTS	128	1343/1361	690	220.3	13.42	34.97	0
COLONIES	Others	59	572/591	109	107.8	9.23	0.08	0.466
COLONIES	Outside lab 1	1	49/50	3	7.9	2.58	-2.11	0.991
COULTER	RTS	174	1716/1717	173	97.7	9.58	7.80	6.26x10 ⁻¹³
COULTER	Others	103	929/929	36	39.9	6.11	-0.71	0.758
COULTER	Outside lab 2	11	97/97	0	4.4	2.03	-2.42	1.00
COULTER	Outside lab 3	17	120/120	1	3.75	1.90	-1.71	0.990

5.10 Probability model for Mid-Ratios: We took a similar approach to evaluating the significance of the occurrence of high percentages of triples having mid-ratios close to 0.5 to that which we used when dealing with triples that contain their mean. In like manner, we wrote an R function to calculate the probability that the mid-ratio of a triple with a given parameter λ falls within the interval [0.40,0.60]. Using this function we calculated these probabilities for each of

the integer values of λ from 1 to 2000 and stored them in a table. The results of these calculations showed that as λ increases from 1 to 10 the probability that a triple has a mid-ratio in the interval $[0.40, 0.60]$ increases from about 0.184 to slightly more than 0.251 and decreases thereafter. Thus our calculated results tell us that for every value of λ , the probability that the mid-ratio is in the interval $[0.40, 0.60]$ is less than 0.26. Hence, given a collection of n triples the probability that k or more of those triples have mid-ratios in the interval $[0.40, 0.60]$ cannot be greater than the probability of k or more successes in n independent Bernoulli trials in which the probability of success is 0.26. As was the case when we considered triples which contain their mean, these Binomial probabilities can be used to provide a crude but potentially useful test of significance.

We used the same heuristic approach that we had used to develop a more refined significance test for the occurrence of triples that contain their own means to develop a more refined significance test for the incidence of mid-ratios in the $[0.40, 0.60]$ interval. This test could be of use in detecting instances in which an investigator wishing to guide the mean values of triplicates employs a reasonably subtle technique.

6. Terminal Digit Analysis: J. E. Mosimann and colleagues (Mosimann, Wiseman et al. 1995; Mosimann, Dahlberg et al. 2002) recommend a technique for identifying aberrant data sets based on the observation that under many ordinary circumstances the least significant (rightmost) digits of genuine experimental count data can be expected to be uniformly distributed and the further observation that when people invent numbers they are generally not uniform.

As per our introductory remarks it is important to confirm the applicability of this expected uniformity in any context in which we hope to use it. The fact that, in as much as the cells counted in a single batch by the Coulter counter typically number in the several hundreds up to

the many thousands, control in selecting the batches of cells to be counted is far from precise enough to extend to the last digit, lends some *a priori* support to the expectation that terminal digits will be uniform. But we also ran simulations generating data sets of triples of independent identical Poisson random variables with comparable means, and the distributions of terminal digits in these sets were consistent with the hypothesis of uniformity.

Based on these considerations we believe it is reasonable to suppose that the Mossiman technique applies to the various Coulter count data sets under consideration. The fact that we are able to apply our tests of uniformity to what we believe to be uncontested experimental data in the course of our test provides further of empirical confirmation of the applicability of the Mossiman test.

6.1 Application of terminal digit analysis to the data sets: We counted the number of times each of the digits 0,1,...,9 occurred as the rightmost digit of counts copied from the Coulter ZM counter screen and from colony counts. (Note that these analyses do not require that the data be arranged in triplicate sets.) If these least significant digits were indeed uniform -- as they should be if the data was truly generated experimentally -- then our counts for each of these ten digits should be roughly the same.

We obtain a more precise measure of the degree to which these distributions diverge from the expected uniform by applying the Chi-square test for goodness of fit. We show the actual distribution of terminal digits for the various full data sets we considered in **Table 3**, along with the computed Chi-square statistics and the associated p-values. The p-values for RTS's terminal digit sets result in our rejecting the null hypothesis of uniformity at any reasonable level (and even unreasonable levels) of significance; results for all other investigators' data sets are consistent with our null hypothesis.

Table 3. Terminal digit analysis of Coulter and colony counts. “Others” refers to other investigators in the laboratory. Outside labs contributed two sets of Coulter data and one set of colony data. Probabilities of 0 were too small to estimate.

Type	Investigator	Digit										Total	Chi-sq	P-value
		0	1	2	3	4	5	6	7	8	9			
Coulter	RTS 174 exps	472	612	730	416	335	725	362	422	370	711	5155	456.4	0
Coulter	Others 103 exps	261	311	295	259	318	290	298	283	331	296	2942	16.0	0.067
Coulter	Outside lab 11 exps	28	34	29	24	27	36	44	33	26	33	314	9.9	0.36
Coulter	Outside lab 17 exps	34	38	45	35	32	42	31	35	35	33	360	4.9	0.84
Colonies	RTS 128 exps	564	324	463	313	290	478	336	408	383	526	3501	200.7	0
Colonies	Others 59 exps	187	180	193	178	183	173	176	183	183	178	1814	1.65	0.996
Colonies	Outside lab 1 exp	21	9	15	16	19	19	9	19	11	12	150	12.1	0.21

7. Equal Digit Analysis: Just as it is reasonable to expect that insignificant terminal digits in experimental data would be approximately uniform, it also seems reasonable to expect that the last two digits of three plus digit experimental data (in which the terminal digits are relatively immaterial) will be equal approximately 10% of the time. We used R to count the number of terminal digit pairs in the RTS' and other investigators' Coulter count data and found that there were 291 (9.9%) equal pairs of rightmost digit pairs among the 2942 Coulter count values produced by investigators in the laboratory other than the RTS, while there were 636 (12.3%) such pairs in the RTS's 5155 recorded Coulter counts. Assuming that these right-most pairs were generated uniformly, the probability of 636 or more equal pairs in 5155 Coulter values is less

than 3.3×10^{-8} , which significantly contraindicates their expected randomness. In contrast, the probability of 291 or more equal pairs among 2942 Coulter values for the other researchers is 0.587 which is consistent with our randomness hypothesis.

8. Summary

1. In the RTS's experiments, the averages of triplicate colony counts appear as one of those counts at improbably high levels based on our model. The rates at which triplicate colony counts reported by other investigators include their averages is consistent with our model.

2. In the RTS's experiments, the mid-ratio values of triplicate colony counts fall in the interval $[0.4, 0.6]$ at improbably high levels based on our model. The rates at which mid-ratios of triplicate colony counts reported by other investigators fall in that interval is consistent with our model.

3. Distributions of terminal digits of values in the RTS's Coulter counts and colonies differ significantly from expected uniformity. This does not hold for the colony and Coulter terminal digits of other workers.

5. Significantly more than the expected one tenth of the data values the RTS recorded in his Coulter counts have equal terminal digits. This does not hold for the occurrences and distributions of terminal doubles in the Coulter counts of other workers.

9. Discussion

9.1 Limitations In most case studies, the number of controls is either equal to or greater than the number of test values. Since this is a *post hoc* study, we had no control over the numbers of data we analyzed. To address our concern about smaller control sample sizes in one such instance, we randomly selected 314 terminal digits from the RTS's Coulter results and ran chi-square

analyses 100,000 times to test for uniformity. All of the runs would have rejected the null hypothesis for uniformity at the 0.00001 level; one run rejected the hypothesis at the 0.000000001 level. The value of 314 was selected because it is the total number of digits supplied by one of the two outside contributors and was the smallest of the Coulter sample sets with which we worked (cf **Table 3**).

During the time that the RTS was working in the laboratory, few experiments were being performed simultaneously by others, which resulted in some temporal disparity. However, the protocols that we analyzed were followed almost identically by all of the members of the laboratory. There is no *a priori* evidence that the cells, instrumentation, equipment and consumable supplies used by the other researchers were any different from those utilized by the RTS. There is also no evidence that different operators could influence the terminal digits seen on the display of the Coulter counter. All of the investigators used similar techniques to stain and count the colonies.

9.2 Power of statistics: In a recent editorial in *Science*, Davidian and Louis emphasize the increasing importance of statistics in science and in world affairs as a “route to a data-informed future” (Davidian and Louis 2012). Statistical analysis of numerical data can be used to identify aberrant results (Tomkins, Penrose et al. 2010; Postma 2011; Tomkins, Penrose et al. 2011), even in esoteric studies (Brown, Cronk et al. 2005) (Trivers, Palestis et al. 2009). Recently, a rigorous statistical analysis of data that purported to predict the responses to chemotherapeutic agents of human lung, breast and ovarian cancers demonstrated the erroneous nature of the results (Baggerly and Coombes 2009; Baggerly and Coombes 2011) and led to several retractions (Baggerly and Coombes 2010; Goldberg 2010; 2011; 2011) and a resignation. In this

case, patients were potentially directly affected by the use of the wrong drug and/or the withholding of the right drug.

Statistics were used to uncover fraudulent behavior on the part of Japanese anesthesiologist Y. Fujii who is believed to have fabricated data in as many as 168 publications (Carlisle 2012). In like manner, Al-Marzouki, et al. (Al-Marzouki, Evans et al. 2005) used statistics to implicate R.B. Singh for fabricating data in a clinical trial involving dietary habits. Their control, like our controls, was a similar trial performed using comparable methods by an outside group. Of interest is the fact that Singh was unable to produce his original data for re-examination because it had been, he alleged, consumed by termites. Hudes, et al. and McCann, et al. (Hudes, McCann et al. 2009) used statistics to detect unusual clustering of coefficients of variation in a number of articles produced by members from the same biochemistry department in India. The controls for these studies were obtained by searching for similar studies in PubMed. Once data manipulation is suspected, it is up to the statistician to find the proper test(s) to reveal discrepancies – to “let the punishment fit the crime”, so to speak.

9.3 Are the RTS 's data real: The consistent and highly significant improbability that any of the multiple anomalies observed in the RTS's data sets are likely to have occurred by chance, and the fact that none of these anomalies appear in either the many data sets we examined from the nine other investigators in the same laboratory, working under the same conditions with the same equipment or in the comparable data sets we obtained from investigators outside the laboratory, leaves us with no alternative than to believe that the RTS's data is simply not genuine experimental data.

10. Remedies

10.1 Automated analysis can deter tampering with results: Automatic colony counters are commercially available, and their use in colony survival and other such studies should be encouraged. The counts from particle counters such as the Coulter ZM should be recorded on a printer.

10.2 Journals should require the availability and archiving of raw data. Many now do. This will permit verification, help to avoid unnecessary duplication of experimental results and facilitate interactions and interchanges among researchers.

10.3 An Excel spreadsheet, available on request to perform the calculations that we have proposed in this article, understanding that most researchers performing these types of survival and related experiments are not versed in the use of the statistical program R. The spreadsheet is available from Dr Pitt on request.

Appendix

Calculating the probability that a Poisson triple contains its rounded mean:

As a preliminary to determining the probability that a triple contains its rounded mean, we first calculated the probability that a triple randomly generated by three independent Poisson random variables with a given λ has a gap of two or more and contains its own mean. This event is the union of the infinite collection of mutually exclusive events:

A_j = the event that the gap is j and the triple contains its own (rounded) mean, for $j = 2, 3, 4, 5, \dots$

Hence its probability is, the sum of the separate probabilities of the A_j 's, $\sum_{j=2}^{\infty} P(A_j)$

For each j the event A_j is itself the union of the infinite collection of mutually exclusive events:

$A_{j,k}$ = the event that the largest value in the triple is k (hence the smallest is $k-j$) and the triple includes as one of its elements its own (rounded) mean

where, for any given j , the admissible values of k are $j, j+1, j+2, j+3, \dots$. Hence $P(A_j) =$

$$\sum_{k=j}^{\infty} P(A_{j,k})$$

To calculate $P(A_{j,k})$ we observe that in order for the event $A_{j,k}$ to occur, the smallest of the three elements of the triple must be $k-j$, and, of course, the largest must be k , but depending on the parity of j there may be one or two different possible values completing the triple. When j is even the third must be $k-j/2$ as it is easy to see that this is the only integer value that can complete a triple $\{k-j, n, k\}$ that has mean n . However, when j is odd, there are two distinct integer values that can complete the triple $\{k-j, n, k\}$ so that its mean is n , these are: $k-[j/2]$ (where $[x]$ is the greatest integer function, i.e. $[x] =$ greatest integer less than or equal to x) and $k-[j/2]-1$.

Since the elements of our triples are assumed to be independently generated Poisson random variables with common parameter λ we can obtain formulas in terms of Poisson probabilities for $P(A_{j,k})$. We first consider the case j even. Writing $p(n, \lambda)$ for the Poisson probability ($e^{-\lambda} \frac{\lambda^n}{n!}$) of obtaining the value n from a Poisson random variable with parameter λ , the probability that a triple consists of the values $\{k-j, k-[j/2], k\}$ in any one of the six different orders in which these numbers can be permuted is $p(k-j, \lambda)p(k-[j/2], \lambda)p(k, \lambda)$ and hence the probability of obtaining the triple for j even is

$$P(A_{j,k}) = 6p(k-j, \lambda)p(k-[j/2], \lambda)p(k, \lambda)$$

Applying a similar analysis with the two distinct triple types that could result in the event A_{jk} when j is odd we get for odd j

$$P(A_{j,k}) = 6p(k-j, \lambda)(p(k - [j/2], \lambda) + p(k - [j/2] - 1, \lambda))p(k, \lambda)$$

We combine the preceding observations to obtain a formula for the probability $P(A)$ that a triplet of numbers chosen independently from the same Poisson distribution contains its (rounded) mean. We get

$$P(A) = 6\left(\sum_{\text{odd } j=3}^{\infty} \sum_{k=j}^{\infty} p(k-j, \lambda)(p(k - [j/2], \lambda) + p(k - [j/2] - 1, \lambda))p(k, \lambda) + \sum_{\text{even } j=2}^{\infty} \sum_{k=j}^{\infty} p(k-j, \lambda)p(k - [j/2], \lambda)p(k, \lambda)\right)$$

And writing $\text{odd}(x)$ for the function that is 1 when x is odd and 0 when x is even we can rewrite this as the single double sum:

$$P(A) = 6\left(\sum_{j=2}^{\infty} \sum_{k=j}^{\infty} p(k-j, \lambda)(p(k - [j/2], \lambda) + \text{odd}(j)p(k - [j/2] - 1, \lambda))p(k, \lambda)\right)$$

Since we wish to obtain decimal values for these probabilities for various values of λ we note

that if, for a given λ we choose N such that $\sum_{j=N+1}^{\infty} p(j, \lambda) < 10^{-9}$ or, equivalently,

$\sum_{j=0}^N p(j, \lambda) \geq 1 - 10^{-9}$, then we can obtain a value of $P(A)$ accurate to 5 decimal places using

the formula:

$$P(A) = 6 \left(\sum_{j=2}^N \sum_{k=j}^N p(k-j, \lambda) (p(k - \lfloor j/2 \rfloor, \lambda) + \text{odd}(j)p(k - \lfloor j/2 \rfloor - 1, \lambda)) p(k, \lambda) \right)$$

Using this formula, we wrote an R program to calculate the probability that a triple of independent Poisson random variables with a common parameter λ includes its mean as one of its three elements. We ran this program to create a table of the values of this probability for each of the integer values of λ from 1 to 2000. As a double check on the applicability of our calculation, we performed bootstrap calculations of selected probabilities using R to perform sets of 200,000 trials. The results were consistent with our calculations.

References

- Al-Marzouki, S., S. Evans, et al. (2005). "Are these data real? Statistical methods for the detection of data fabrication in clinical trials." BMJ **331**(7511): 267-270.
- Baggerly, K. A. and K. R. Coombes (2009). "Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology, ." The Annals of Applied Statistics **3**(4): 1309-1334.
- Baggerly, K. A. and K. R. Coombes (2010). "Retraction based on data given to Duke last November, but apparently disregarded." The Cancer Letter **36**(39): 1,4-6.
- Baggerly, K. A. and K. R. Coombes (2011). "What Information Should Be Required to Support Clinical "Omics" Publications?" Clin Chem **57**: 688-690.
- Bonifacino, J. S. (1998). Current protocols in cell biology. New York, John Wiley: v. (loose-leaf).
- Brown, W. M., L. Cronk, et al. (2005). "Dance reveals symmetry especially in young men." Nature **438**(7071): 1148-1150.

Carlisle, J. B. (2012). "The analysis of 168 randomised controlled trials to test data integrity."

Anaesthesia **67**(5): 521-537.

Davidian, M. and T. A. Louis (2012). "Why statistics?" Science **336**(6077): 12.

Fang, F. C., R. G. Steen, et al. (2012). "Misconduct accounts for the majority of retracted scientific publications." Proc Natl Acad Sci U S A **109**(42): 17028-17033.

Goldberg, P. (2010). "Nevins retracts key paper by Duke group, raising question of harm to patients." The Cancer Letter **36**(39): 1-4.

Hill, T. P. and K. Schürger (2005). "Regularity of digits and significant digits of random variables. ." Stoch. Proc. Appl. **115**: 1723-1743.

Hong, Y. (2011). "Technical Report No. 11-2." Technical Reports On Computing the Distribution Function for the Sum of Independent and Non-identical Random Indicators.

Hudes, M. L., J. C. McCann, et al. (2009). "Unusual clustering of coefficients of variation in published articles from a medical biochemistry department in India." Faseb J **23**(3): 689-703.

Katz, D., E. Ito, et al. (2008). "Increased efficiency for performing colony formation assays in 96-well plates: novel applications to combination therapies and high-throughput screening." Biotechniques **44**(2): ix-xiv.

Mosimann, J. E., J. E. Dahlberg, et al. (2002). "Terminal digits and the examination of questioned data." Accountability in Research **9**: 75-92.

Mosimann, J. E., D. V. Wiseman, et al. (1995). "Data fabrication: Can people generate random digits?" Accountability in Research **4**: 31-55.

Munshi, A., M. Hobbs, et al. (2005). Clonogenic Cell Survival Assay. Methods in Molecular Medicine. R. D. Blumenthal. Totowa, NJ, Humana Press. **1**: 21-28.

Postma, E. (2011). "Comment on "Additive genetic breeding values correlate with the load of partially deleterious mutations"." Science **333**(6047): 1221.

Rossner, M. and K. M. Yamada (2004). "What's in a picture? The temptation of image manipulation." J Cell Biol **166**(1): 11-15.

Simonsohn, U. (2012). "Just post it: The lesson from two cases of fabricated data detected by statistics alone." Available at SSRN: <http://ssrn.com/abstract=2114571> or <http://dx.doi.org/10.2139/ssrn.2114571>

Tomkins, J. L., M. A. Penrose, et al. (2010). "Additive genetic breeding values correlate with the load of partially deleterious mutations." Science **328**(5980): 892-894.

Tomkins, J. L., M. A. Penrose, et al. (2011). "Retraction." Science **333**(6047): 1220.

Trivers, R., B. G. Palestis, et al. (2009). The Anatomy of a Fraud: Symmetry and Dance.
Antioch, CA 94509, TPZ Publishers.

Van Noorden, R. (2011). "Science publishing: The trouble with retractions." Nature **478**(7367):
26-28.

Volkova, A. Y. (1995). "A refinement of the Central Limit Theorem for Sums of Independent Random Indicators." Theory Probab. Appl. **40**(4): 791-794.