Statistical Evidence of Fraud in Department of Radiology, New Jersey Medical School

US District Court, District of New Jersey, Case # 03-4837

Plaintiffs: United States of America Ex Rel. Dr. Helene Z. Hill

Defendants: University of Medicine and Dentistry of New Jersey, Dr. Roger W. Howell and Dr. Anupam Bishayee

Expert Report prepared by Joel Pitt, PhD on behalf of Dr. Helene Z. Hill

In October, 1999, Dr. Helene Hill, a Professor of Radiology and researcher at New Jersey Medical School observed Dr. Anupam Bishayee, a post-doctoral fellow and research associate of Dr. Roger Howell making false claims about an experiment he was supposed to be performing. In March, 2001 Dr. Marek Lenarczyk, also a post-doctoral fellow in Dr. Howell's laboratory, together with Dr. Hill observed Dr. Bishayee conducting an experiment which was contaminated with micro-organisms. Although this situation would clearly invalidate the results of his experiment, Dr. Bishayee presented them as valid.

Results of Dr. Bishayee's experiments were reported in two publications and used as preliminary data for a funded grant application. The results, however, have not been replicated. Based on her understanding of the underlying science and her observations of scientific misconduct on Dr. Bishayee's part, Dr. Hill believes that it was impossible to honestly obtain the results Dr. Bishayee reported and, consequently, those results had to have been fabricated and fraudulent. In the course of her review of Dr. Bishayee's research results, Dr. Hill noticed certain unusual patterns in the data he presented. She asked us to review that data to confirm or disconfirm her belief that those irregularities are highly indicative of scientific fraud.

In reviewing Dr. Bishayee's data we used three different techniques:

1) We determined the relative frequency with which each of the digits 0-9 appear as least significant digit in Dr. Bishayee's data (a standard technique that is used by the Office of Research Integrity of the Department of Health and Human Services). Using appropriate control data to confirm assumptions about the expected relative frequencies we determined the probability that non-fabricated data could result in such frequencies is considerably less than 0.00000000001 (one in one hundred billion).

2) Data from the experiments critical in supporting the aforementioned non-replicated results is organized in groups of three measurements. We examined the frequency with which one of the three measurements is close (in a sense defined more specifically below) to the average of the three measurements and found that the frequency and pattern of closeness in Dr. Bishayee's data is completely at variance with the pattern in control data from various sources and computer simulation data. Although we cannot assign a specific probability to the results here, the distinctive pattern evident in Dr. Bishayee's data would lead any reasonable observer to conclude that Dr. Bishayee repeatedly and

deliberately invented one value in each triad to force his data to conform to the experimental results he wished to report.

3) We determined the relative frequency with which the two least significant digits in Dr. Bishayee's measurements are equal. Based on reasonable assumptions about the likelihood that the terminal digits of a non-fabricated measurement would be equal, assumptions that are borne out by our control data, we find the probability that the relative frequency of such incidents diverge from the expect frequency as much as they do in Dr. Bishayee's case is less thant 0.0000001 (one in ten million.)

In considering any claim that the probability of an outcome with certain anomalous or distinctive characteristics is miniscule it is absolutely essential to understand the assumptions on which calculation of the given probability value is based. We provide the appropriate details for the preceding assertions below.

The mere unlikelihood of an event certainly does not imply that it cannot have honestly occurred by chance. After all, lotteries regularly return winners despite the significantly low probability of winning. Nonetheless, the staggering improbability that the patterns evident in Bishayee's data would occur in the ordinary course of research leave us quite certain that much of it has, indeed, been fabricated. When our statistical results are considered in combination with direct observation of scientific misconduct by Bishayee and the irreproducibility (and apparent impossibility of reproducing) his results the conclusion that he has committed fraud seems inescapable.

Relative Frequency of Least Significant Digits

Our first review of Dr. Bishayee's data employed a technique used by the Division of Research Investigation of the Department of Health and Human Services Office of Research Integrity. The technique can be used to determine that data has been fabricated in situations where there is reason to believe that under ordinary circumstances the least significant (rightmost) digits of genuine experimental data should be uniform. If that is the case we would process the suspect data counting the number of times each of the digits 0-9 appears as the least significant digit of a data value. If these least significant digits were indeed uniform -- as they should be it if the data was truly generated experimentally -- then our counts for each of these ten digits should be roughly the same. In order to apply this approach we have to address two questions: 1) why should we believe that in the case of the experimental data we are concerned with the least significant (rightmost) digits should be uniform; and 2) how do we determine whether the actual frequencies of the data "appear roughly the same number of times?

We can address the first of these questions both *a-priori* and empirically. There are two sets of data from Dr. Bishayee's experiments. The first set consists of counts of cells obtained using a Coulter Counter -- a device for counting particles and cells. The numbers of cells that are counted in a single batch typically number in the several hundreds up to the many thousands. Since control in the process of selecting the batches of cells to be counted is far from precise enough to ordinarily extend to the last digit, it is very reasonable to assume that the last (units) digit in each count will be random. The second set consists of manual counts of colonies of cells where the numbers ordinarily vary from the mid teens to several hundreds. Although the lower numbers may suggest the possibility of somewhat greater control of the final digit, the fact that these colonies

reflect the survival of cells that were processed experimentally once again supports the belief that the units digit in each count is likely to be uniformly distributed.

Although we find the *a-priori* argument persuasive here our assumptions of uniformity is also supported empirically. As Mossiman, Wiseman et al (1995-Data Fabrication: Can People Generate Random Digits? p.34) point out:

> …a [more] direct approach is to compare the rightmost digits of unquestioned control data (from the same kind of experiment, laboratory, investigator, and time period as the questioned data) with those of the questioned data. If the rightmost digits of the questioned data depart significantly from a uniform distribution, while those of the control data do not, then there is evidence of some selective factor applying to the questioned numbers. Specifically, the selection may be due to conscious or unconscious human choice in making up numbers.

We counted the terminal digits of the 5155 data values recorded in 171 experiments in which Dr. Bishayee used a Coulter Counter and 1121 data values from the 35 Tritium based experiments in which he counted cell colonies manually. Digit counts for the former are shown in Table 1 (on page 6) and for the latter in Table 2 (on page 7). Figure 1 below is a graph of the relative frequency of each of the ten digits in Dr. Bishayee's Coulter Counter experiments; Figure 2 is a graph of the relative frequency of these digits in the colony counts.

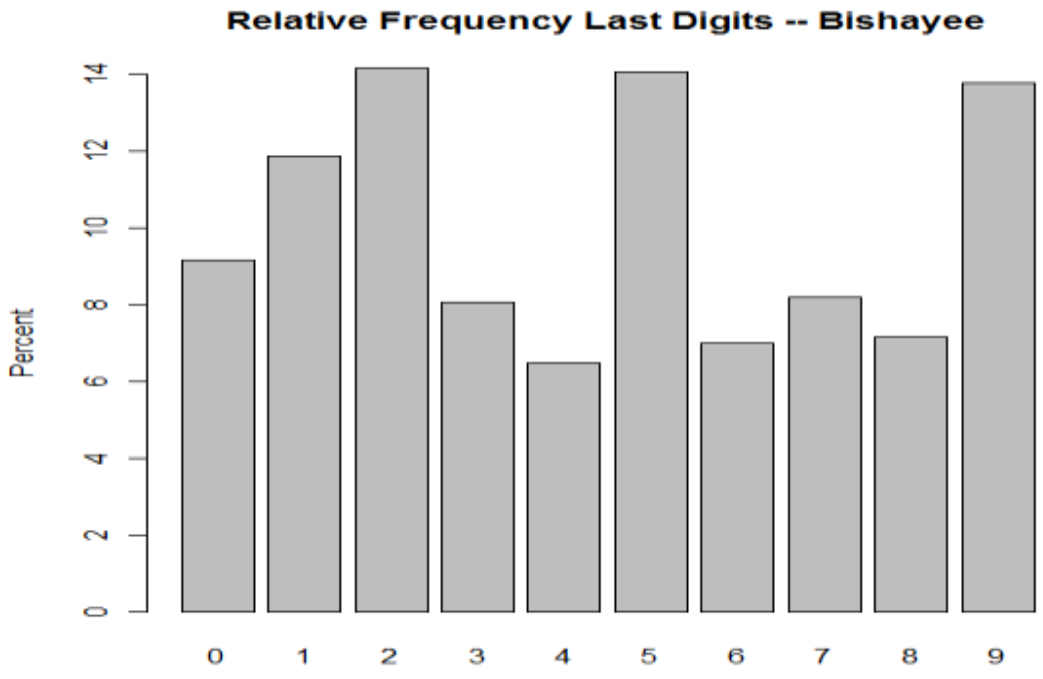

Figure 1

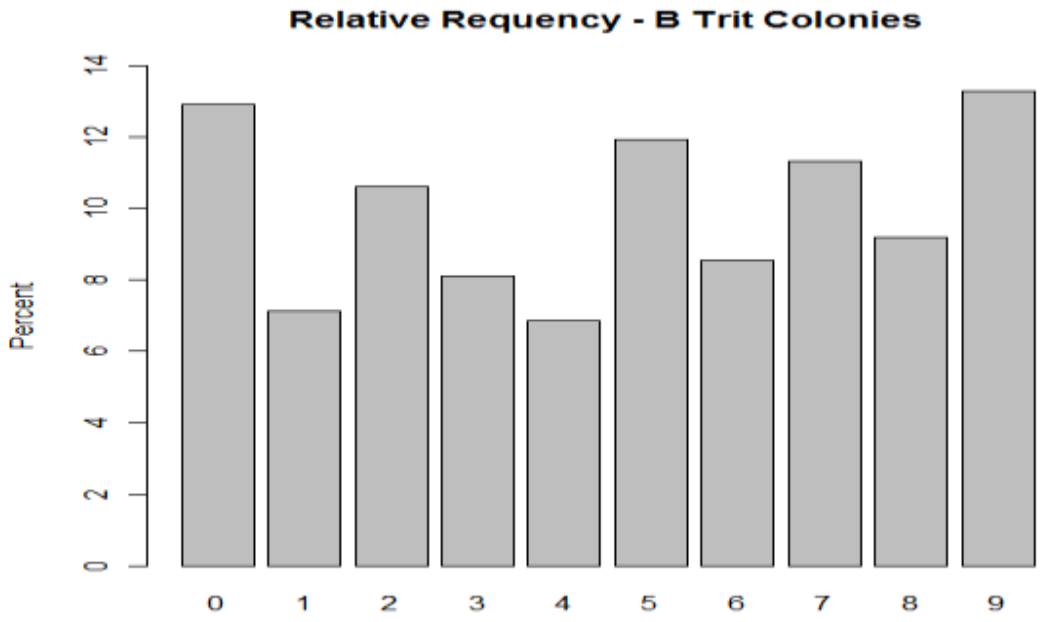

Figure 2

Both graphs clearly show strong divergence from uniformity. The conspicuously low relative frequency of the digits 3 and 4 in both graphs is particularly notable.

As was recommended by Mossiman et. al. (as quoted above), we obtained data from a variety of control sources to validate our belief that the distribution of these terminal digits should, in fact, have been uniform. In particular we counted the occurrence of the various digits as the terminal digits from 99 other experiments employing the same Coulter Counter that Dr. Bishayee used and from 59 colony counting experiments in the same laboratory. (The 99 Coulter Counter experiments included 41 by Dr. Howell, 22 by Dr. Lenarzyk, and 11 by Dr. Gerashenko, while the 59 colony counting experiments included 27 by Dr. Howell.) We solicited additional control data from other sites that employ Coulter Counters and obtained data for comparison purposes for 17 experiments conducted at the UT Southwestern Medical Center (Dallas/Fort Worth) and 11 experiments (314 data points) conducted at Case Western Reserve (360 data points). Digit counts for all of the Coulter experiments are shown in Table 1 (on page 6) and for the colony count in Table 2 (on page 7)

The graphs in Figures 3 and 4 below show the relative frequencies of the various digits in the first control group of 99 Coulter Count experiments and the second control group of 59 colony counts. (We deal with the data from UT Southwestern Medical Center and the Case Western Reserve data in the detail quantitative analysis below.)

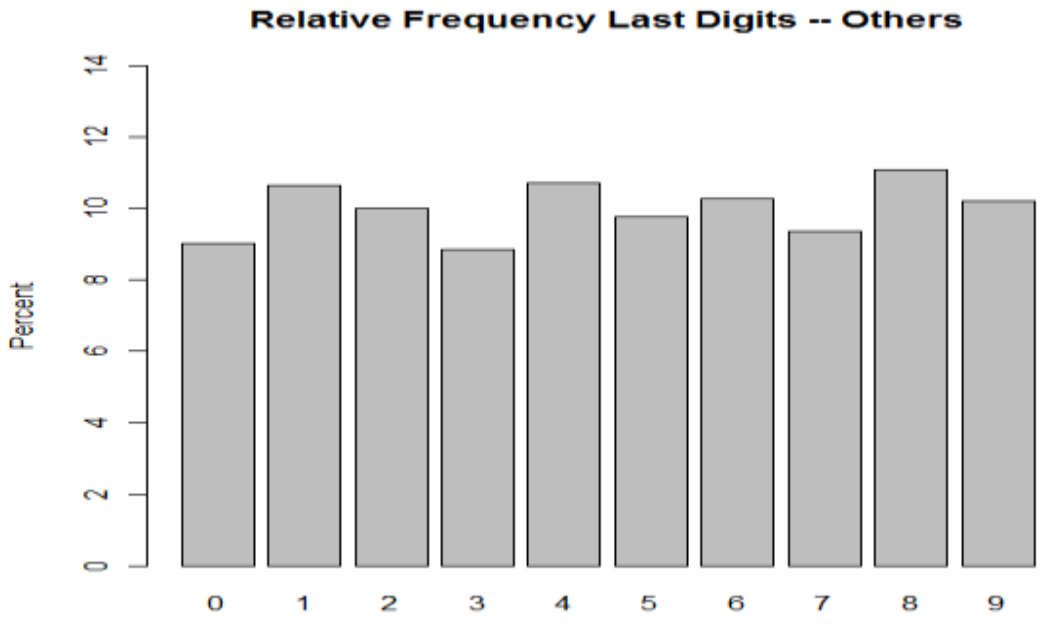

Figure 3

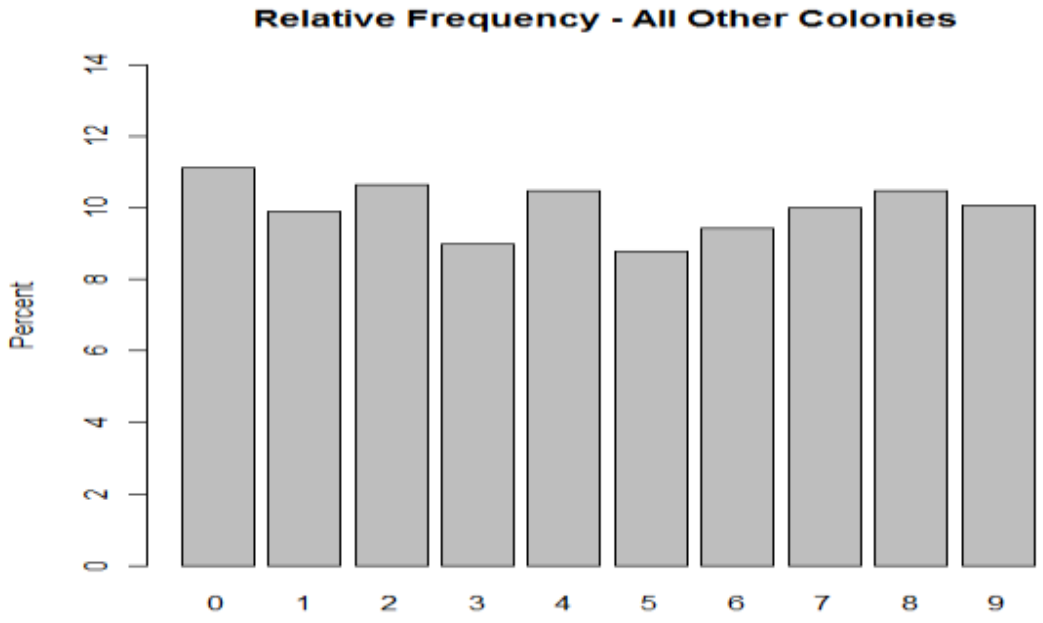

Figure 4

The relative frequencies of digit counts in the control graphs are clearly much closer to uniform than are the relative frequencies in Dr. Bishayee's data. A detailed quantitive

analysis allows us to show quite conclusively that the relative frequency differences are material and significant.

A standard mathematical calculation called the Chi-Square calculation can be used to obtain a precise numeric measure of the extent to which a given collection data fails to conform to a predicted distribution. The probability distribution of the resulting statistic is well-known and tabulated and is widely used to perform a standard statistical test known as the Chi-Square Goodness-of-Fit Test (available in virtually any standard statistics text.)

Table 1 below shows the actual frequency with which each of the digits 0-9 appeared as the terminal digit in Coulter Counter data values recorded for the 171 experiments that were conducted by Dr. Bishayee, the 99 experiments conducted by other Medical School of New Jersey researchers, the 11 experiments conducted by at Case Western Reserve, and the 17 experiments conducted at UT Southwestern. In the 12th column of the table we have recorded the calculated Chi-Square measure of the extent to which the data in question fails to conform to the predicted uniform distribution of terminal digits.

The p-Value column gives the probability that data which is really drawn uniformly could result in a set of actual digit frequencies that has Chi-Square value as large as or larger than the Chi-Square value obtained from the frequencies in that row. We performed the Chi-Square calculations using the widely used R Statistical Software and obtained the p-Values from the same source. The p-Values for the non-Bishayee data are entirely consistent with the expectation that terminal digits should be distributed uniformly, but the actual p-Value the R system returned for Dr. Bishayee's frequencies was that it was less than 2.2 time 10 to the minus 16th power. If, in fact, the terminal digits in properly conducted Coulter Count runs are genuinely random – an apriori plausible assumption that the control supports – the probability that Dr. Bishayee could have obtained the results he obtained is far less than 0.00000000001 (one in one hundred billion). (All p-Values are computed using the Chi-Square distribution with 9 degrees of freedom.)

Table 1: Frequency of Occurrence of Terminal Digits in Coulter Counter Values

Investigator	0	1	2	3	4	5	6	7	8	9	Chisq	p-Value
Bishayee	472	612	730	416	335	725	362	422	370	711	456.4	<0.00000000001
Other NJMS	249	294	276	244	296	270	284	258	306	282	13.9	0.12
Case Western	28	34	29	24	27	36	44	33	26	33	9.9	0.35
UTSouthwestern	34	38	45	35	32	42	31	35	35	33	4.9	0.84

Table 2 shows the actual frequency with which each of the digits 0-9 appeared as the terminal digit in colony counts that Dr. Bishayee recorded in the 35 Tritium related experiments for which he counted cell colonies manually, and the frequencies with which the same digits appeared as terminal digit in the colony counts reported in the 59 such experiments that other NJMS researchers conducted. In counting these digits we

specifically eliminated triads of counts in which any one of those counts was less than 10 as in such instances the actual terminal digit would be material. Here too the control data with its Chi-Square p-Value of 0.571 is supportive of our assumption that terminal digits should be uniformly distributed, while the 0.00000009 p-Value for Bishayee's data is significant reason to be concerned about its legitimacy.

Table 2: Frequency of Occurrence of Terminal Digits in Colony Count Values

Investigator	0	1	2	3	4	5	6	7	8	9	Chisq	p-Value
Bishayee	145	80	119	91	77	134	96	127	103	149	55.7	0.00000009
Other NJMS	173	154	166	140	163	137	147	156	163	157	7.6	0.57

Relative Frequency of Least Significant Digits in Individual Experiments

In the foregoing discussion we considered data collected over all experimental Coulter Counter runs. The analysis supports the belief that terminal digits should occur randomly in data from properly conducted experiments, and indicates that there is a significant reason to be concerned with the data that Dr. Bishayee reported.

We extended our analysis by examining the distribution of terminal digits in all of the 270 NJMS Coulter Count experiments for which we have data and the 28 experiments for which we received data from UT Southwestern and Case Western Reserve. We performed the same Chi-Square calculation for each of these experiments individually that we performed on the collected data discussed above, and screened the collection to determine which ones would cause a rejection of the hypothesis that terminal digits are random at the 1% level (a stringent screen condition.) There were exactly 45 such experiments, and all were experiments conducted by Dr. Bishayee.

When we looked more closely at the distribution of these suspect experiments there is a remarkable pattern over time. This can be seen quite clearly in the chart with the title "Probability of Actual Last Digit Assuming Uniform", shown as Figure 5 below (and as Figure 9 on P. 15)

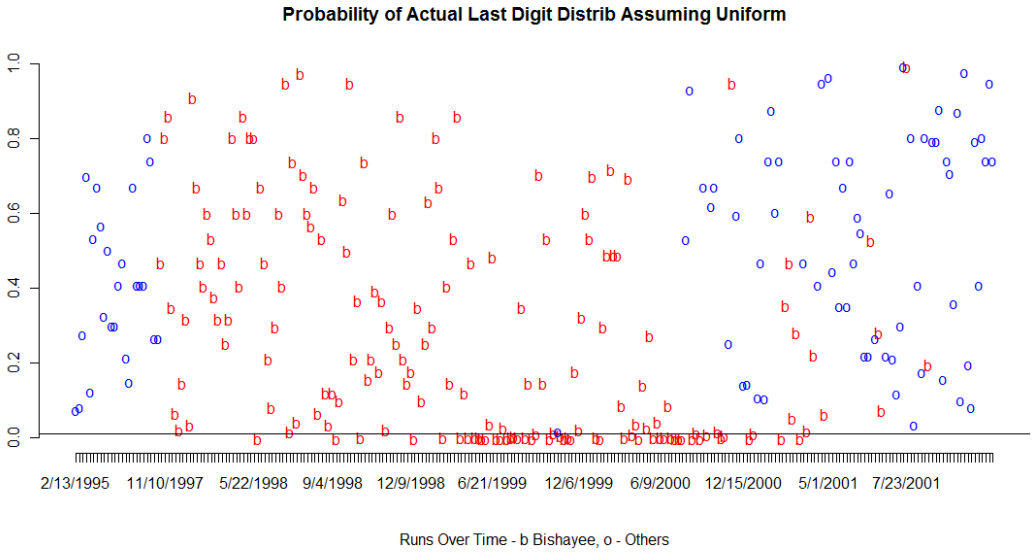

Figure 5

The lettered points on this chart correspond to the various Coulter Counter experiments conducted at NJMS. The points in red with the letter "b" correspond to experiments conducted by Dr. Bishayee and the points in blue with the letter "o" correspond to experiments conducted by other researchers. The horizontal location of each point corresponds to the date on which the experiment that corresponds to the point took place. The vertical position of each point aligns with the Chi-Square probability of the frequency with which the digit 0-9 occurred as terminal digits of the data values collected in that experiment under the assumption that the those terminal digits should be uniform (random). The horizontal line just above the horizontal axis has been drawn in line with the value 0.01 on the vertical axis. Hence, data points below that line correspond to experiments in which the frequency of terminal digits differs significantly (at the 0.01 level) from the assumption of uniformity.

As is seen clearly on the chart the frequency with which Dr. Bishayee reported experimental results that are suspect increased dramatically starting July, 1999. Even during the period in which he performed suspect experiments, there were many other experiments which were consistent with our randomness assumption. This is an important fact as it gives us strong evidence that the suspect results could not have been a result of any device malfunction.

Measurements That Are Close To The Average

Both the Coulter Counter measurements and the colony counts are reported in sets of threes. The experimenter works with three batches that are roughly the same size and in the former case uses the Coulter Counter to count the number of cells in the batch and in the latter case counts the number of colonies in each batch by hand. In looking through Dr. Bishayee's colony count data we noticed that there was a rather unusual frequency of triads in which there was a close coincidence between the actual average of the three numbers in the triad and the middle number (in size order, not position on the page) of the three.

In order to examine this more closely we processed all of the triads that were recorded in colony count experiments performing the following calculation for each:

> we calculated the difference between the middle number and lowest number and divided that difference by the difference beween the highest number and the lowest number

If, for example, a triad contains the three numbers 28, 40, 33 the result of the calculation would be (33-28)/(40-28)=5/12=0.417. The calculation applied to the numbers 102,97,98 returns the value (98-97)/(102-97)=1/5=0.2 and applied to the number 48,56,52 gives the value (52-48)/(56-48)=4/8=0.5. As the last example illustrates, whenever the middle number is close to the average of the three numbers, the result of this calculation is a 0.5.

We would expect that when we repeatedly perform this calculation on triads chosen as they are in the colony count experiments the results will be (pretty much) uniformly random distributed between 0 and 1. In order to validate this expectation we performed a computer simulation in which we drew 500 triples of numbers randomly, and performed this calculation for each triple. A histogram of the results is shown as Figure 6. In the histogram we find the percentage of these 500 values which fall into each of the 20 intervals [0,0.05),[0.05,0.10),[0.10,0.15),…,[0.95,1.00). As is clear from the chart all of these percentages are reasonably close to 5%; the results are entirely consistent with the expectation that results should be uniformly distributed in the interval between 0 and 1.

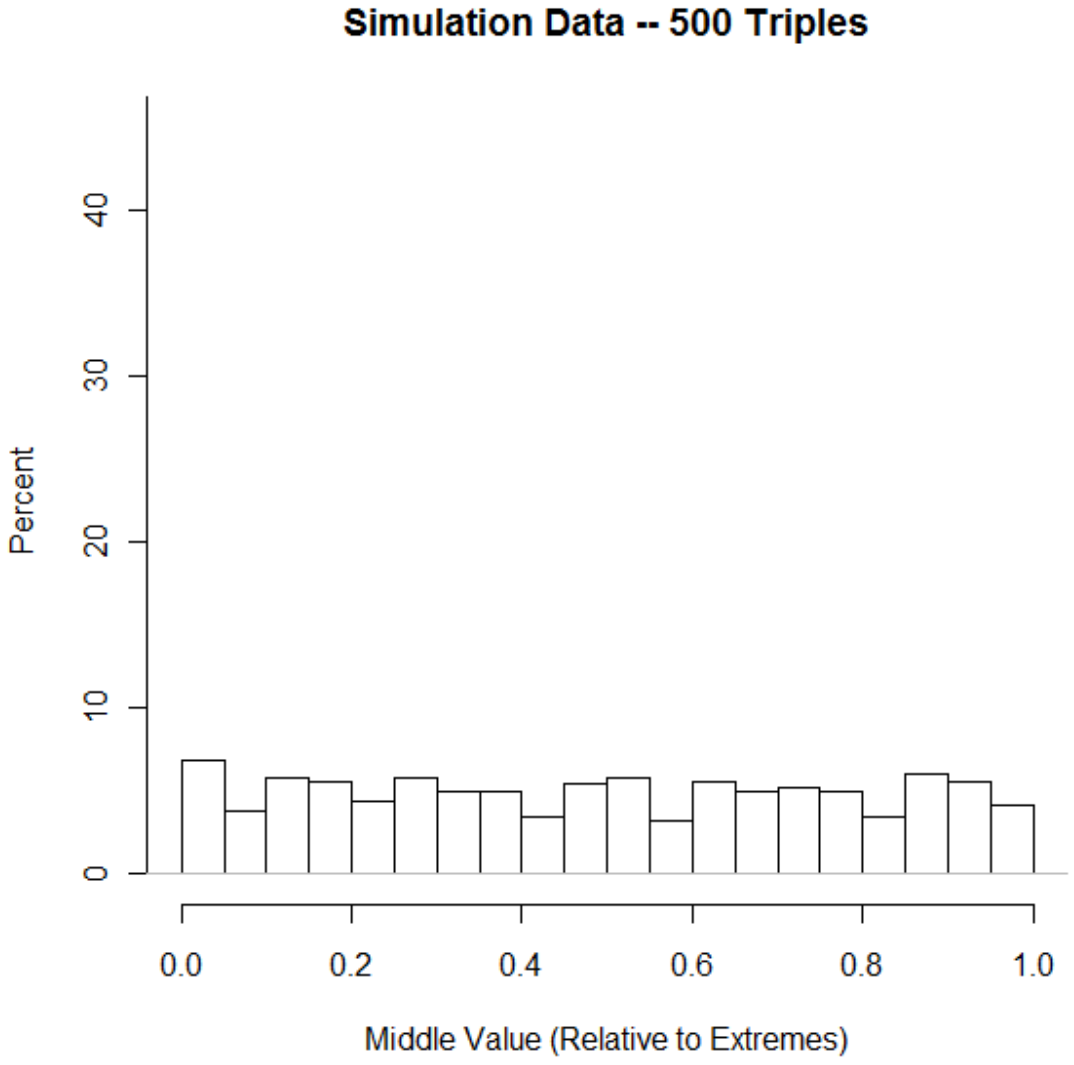

Figure 6

We performed the same calculation for the 542 complete triples that were recorded in the 59 NJMS colony count experiments that Dr. Bishayee did not perform. Once again the results were entirely consistent with our initial expectations. The histogram of the results

is shown below as Figure 7. Once again approximately 5% of the results fall into each of the 20 subintervals into which we divided the interval from 0 to 1.

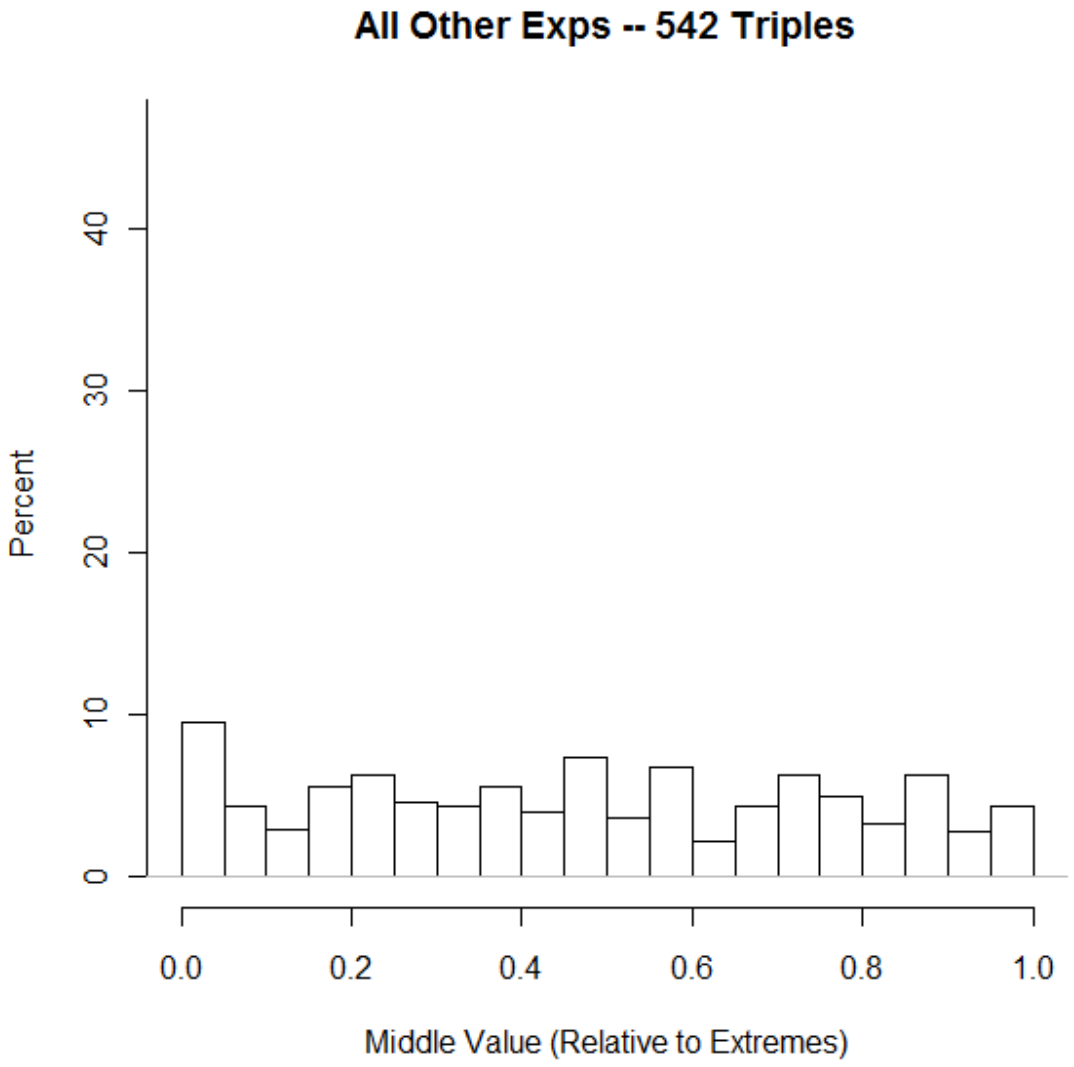

Figure 7

We then applied the same calculation to the 386 complete triples that were recorded in the 35 Tritium related experiments in which Dr. Bishayee collected colony counts. The resulting histogram appears below as Figure 8. The results of our calculation fall in the interval [0.45,0.50] for more than 45% of these triples, and they are in the interval

[0.40,0.60) for more than 65% of the triples.

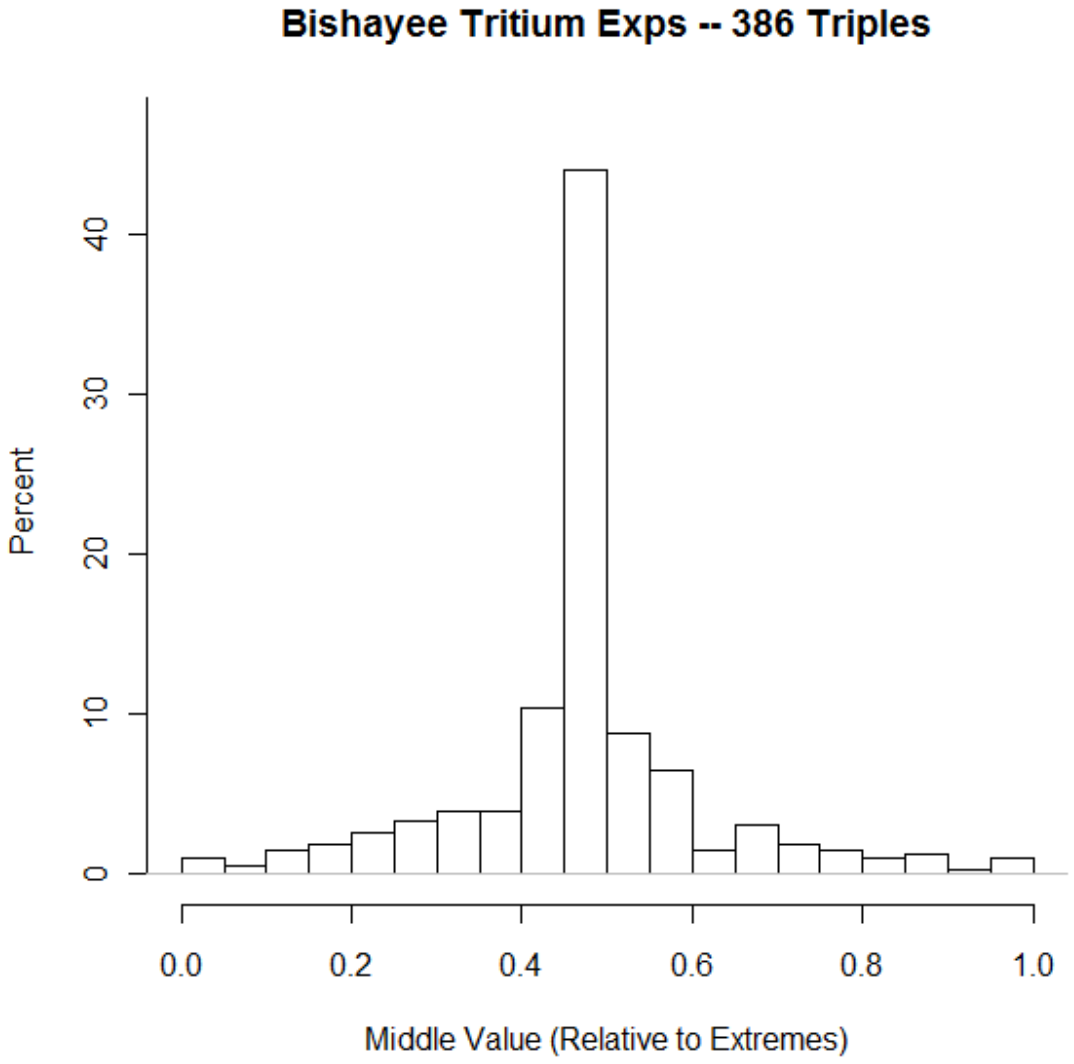

Figure 8

Although the results documented in earlier sections of this report provide strong statistical support for the claim that Dr. Bishayee fabricated significant amounts of the experimental data he reported, we find the results of this analysis particularly dramatic and compelling. The distinctive pattern evident in Dr. Bishayee's data would lead any reasonable observer to conclude that Dr. Bishayee repeatedly and deliberately invented at least one value in each triad to force his data to conform to the experimental results he wished to report.

Relative Frequency of Data Values in Which Last Two Digits Are Equal

As the final step of our review of Dr. Bishayee's data we revisited his Coulter Counter data and determined the relative number of times that the last two digits of numbers that

he recorded were equal. Of the 5155 numbers that were recorded in the 171 experiments using Coulter Counter experiments that were performed by Dr. Bishayee there were 636 (about 12.3%) in which the last two digits were equal. Of the 2759 values recorded in the other 99 NJMS Coulter Counter based experiments there were 280 (about 10.1%) data values in which the last two digits are equal.

It is reasonable to assume that the last two digits of three plus digit experimental data (in which the terminal digits are relatively immaterial) will be equal about 10% of the time. Once again we view the data for the 99 trials conducted by investigators other than Dr. Bishayee as providing a reasonable control with which to test this hypothesis. The fact that about 10.1% of the 2759 terminal digit pairs that were generated in these 99 trials turned out to be equal certainly appears to be consistent with it, but there is a very specific statistical test that we can apply here.

Under the assumption that last two digits of data values are equal with probability 0.10 in a sequence of n independent experimental trials the actual number of data values in which the last two digits are equal will have a binomial distribution with parameters p=0.10 and n=the number of trials. Consequently, to test the null-hypothesis that probability that terminal digits of any given data value is equal to 0.10 against the alternative that the probability is actually greater than 0.10 we need only find the probability that a binomial random variable with parameters p=0.10 and n=2759 takes a value that is greater than or equal to 280. (This is what is technically referred to as a one-tailed test of null-hypothesis. In this case the one tailed test would be appropriate as the alternative that is suggested by both sets of actual data is simply that the probability of occurrence of equal pairs may actually be greater than 0.10.) As reported by the R statistical program – responding to the command pbinom(280,2759,0.10,lower.tail=FALSE) -- this probability is about 0.38; our control does not lead us to reject the null hypothesis.

We apply the same one-tailed test to Dr. Bishayee's data. In this instance we need to obtain the probability that a binomial random variable with parameters p=0.10 and n=5155 takes a value that is greater than or equal to 636. Once again we employ R to obtain the appropriate probability – using the command pbinom(636,5155,0.10,lower.tail=FALSE) – to obtain the value 2.579297e-08 (i.e. .0000000279297) a probability that is far less than one in ten million. The z-score that corresponds to the 636 of Dr. Bishayee's data values in which the last two digits are equal is 5.59. The corresponding p-value is .000000023. The strong unlikelihood that the number of data items in which the last two digits equal could differ as dramatically as the number in Dr. Bishayee's data did casts further doubt on the legitimacy of his data.

Conclusion

Although one hopes that the incidence of the use of fabricated data in research is low (and some studies do seem to support the believe that it is – Taylor, et.a. 2001), it is clear that however low it is, it does exist. Researchers may choose to fabricate data as an expedient to justify scientific results that they believe (or would like to claim) to be true, or they may simply do so to relieve themselves of the burden of honest toil. In either case the consequences are seriously damaging to science.

As Mossiman, et. all (2002) point out, a "useful way to assess questioned data is to examine inconsequential components of data sets that are not directly related to the scientific conclusions of the purported experiment…[if] the allegation is true and the data are falsified, the falsifier typically devote[s] attention to numbers that establish the desired scientific outcome. Properties of the numbers that are not directly related to the desired outcome are less likely to receive consideration by the falsifier" In our study of Dr. Bishayee's experimental data we have found ample indications of such a failure to pay attention to the "inconsequential components" of his data sets. Having done so certain patterns of regularity that are seen to be present in the comparable control data sets that we studied are absent in the data that Dr. Bishayee presented. These regularities included:

1) *the non-significant low order digits of data are ordinarily uniformly distributed*: control from multiple sources and settings displayed this regularity, it was easily seen from graphs of Dr. Bishayee's that it did not display this regularity. Applying the standard statistical Chi-Square goodness-of-fit test to Dr. Bishayee's data showed that the probability of obtaining the actual distribution of low order digits that occurred in his data assuming they were truly uniform was less than one in one hundred billion.

2) *when data is collected in triads with elements having essentially the same distribution, the median value of the triad is equally likely to be as close to the smallest or the largest as it is to the mean of the two:* data from our controls and a computer simulation clearly showed this regularity while in Bishayee's data the median was extremely close to the mean in more than 60% of his trials – a pattern that seems clearly to indicate that Dr. Bishayee was deliberately inventing results to justify an assumed result

3) *when the two lower order of digits are non-significant they will be equal in about 10% of data values*: our control data exhibited this regularity while in a test of the null-hypothesis that Dr. Bishayee's data does we reject the null hypothesis with a p-value that is less than one in ten-million.

When we consider the staggering improbability that the patterns evident in Bishayee's data would occur in the ordinary course of research in combination with the known direct observations of scientific misconduct by Bishayee and the irreproducibility (and apparent impossibility of reproducing) his results the conclusion that he fabricated his results and has committed fraudulent research is inescapable.

References

Al-Marzuki, S., Evans, S., et al. (2005) Are these data real? Statistical methods for the detection of data fabrication in clinical trials, BMJ Volume 331, 30

Chapanis, A. (1953), "Random-number Guessing Behavior," American Psychologist 8, 332.

Hill, T.H. (1999) "The difficulty of faking data" Theodore Hill, Chance 26, 8-13

Mossiman, J., Dahlberg, J., et al. (2002) "Terminal Digits and the Examination of Questioned Data" Investigating Research Integrity Proceedings of the First ORI Research Conference on Research Integrity pp. 269-290

Mossiman, J., Wiseman, C., et al. (1995) "Data Fabrication: Can People Generate Random Digits?" Accountability in Research, Vol. 4, pp. 31-55

Preece, D.A. (1981) “Distribution of Final Digits in Data”, *The Statistician*, Vol. 30, No. 1. (Mar., 1981), pp. 31-60.
R Version 2.7.2 (2008) The R Foundation for Statistical Computering, www.r-project.org
Taylor, Rosemary N. et. al. (2001) Statistical techniques to detect fraud and other data irregularities in clinical questionnaire data, University of Sheffield, Department of Probability and Statistics
Yule, G. U. (1927). “On reading a scale.” *Journal of the Royal Statistical Society,* 90, 570-87.

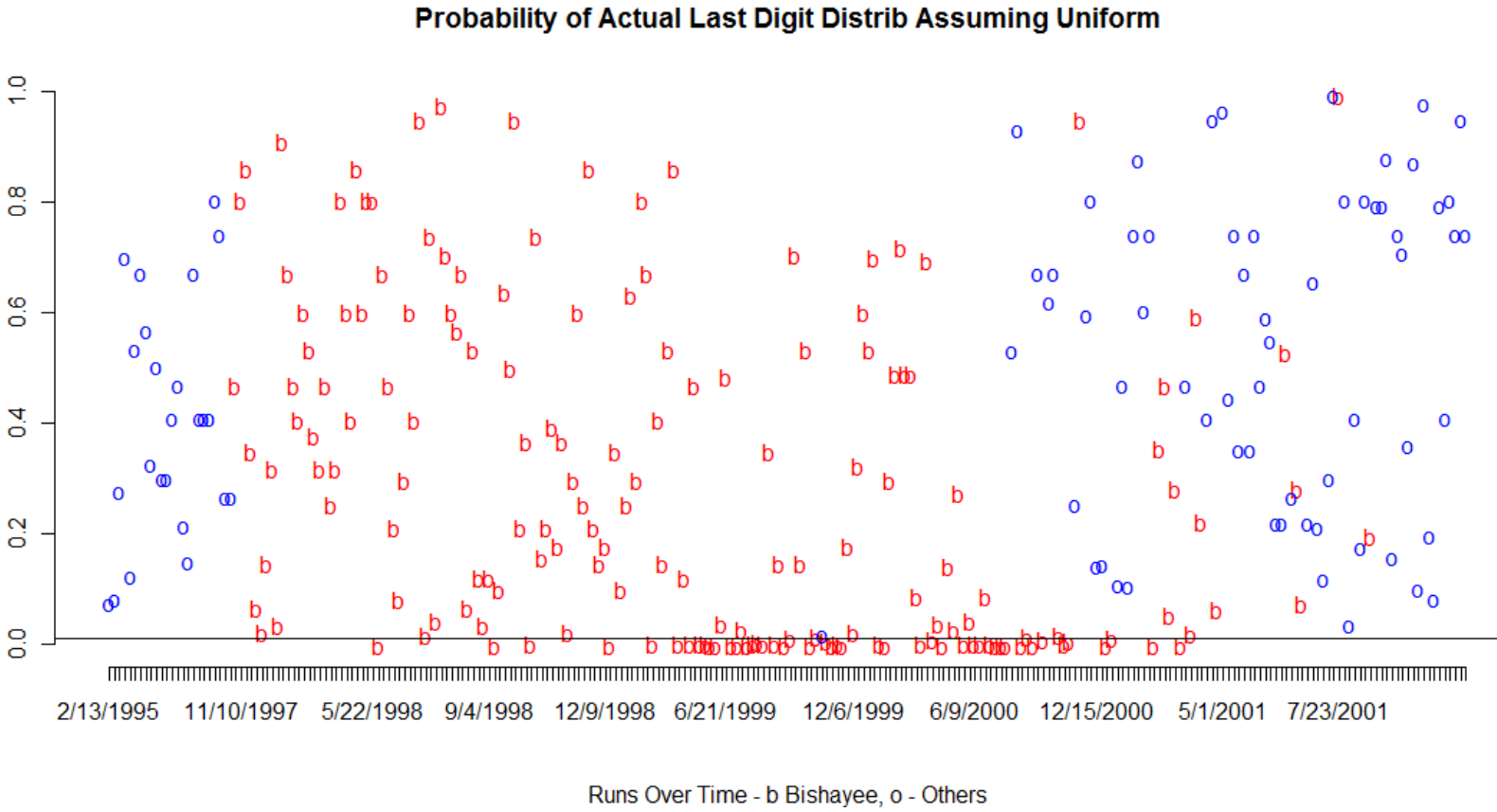
Probability of Actual Last Digit Distrib Assuming Uniform
0.0
0.2
0.4
0.6
0.8
1.0
2/13/1995
11/10/1997
5/22/1998
9/4/1998
12/9/1998
6/21/1999
12/6/1999
6/9/2000
12/15/2000
5/1/2001
7/23/2001
Runs Over Time - b Bishayee, o - Others

Figure 9