Article

Failure to Replicate: A Sign of Scientific Misconduct?

Helene Z. Hill 1,* and Joel H. Pitt 2

1 Rutgers New Jersey Medical School, Newark, NJ 07101-1709, USA
2 Renaissance Associates, Princeton, NJ 08540, USA; E-Mail: drjhpitt@yahoo.com

* Author to whom correspondence should be addressed; E-Mail: hill@njms.rutgers.edu; Tel.: +1-973-495-4977; Fax: +1-973-736-0738.

Received: 28 February 2014; in revised form: 12 August 2014 / Accepted: 18 August 2014 / Published: 1 September 2014

Abstract: Repeated failures to replicate reported experimental results could indicate scientific misconduct or simply result from unintended error. Experiments performed by one individual involving tritiated thymidine, published in two papers in Radiation Research, showed exponential killing of V79 Chinese hamster cells. Two other members of the same laboratory were unable to replicate the published results in 15 subsequent attempts to do so, finding, instead, at least 100-fold less killing and biphasic survival curves. These replication failures (which could have been anticipated based on earlier radiobiological literature) raise questions regarding the reliability of the two reports. Two unusual numerical patterns appear in the questioned individual’s data, but do not appear in control data sets from the two other laboratory members, even though the two key protocols followed by all three were identical or nearly so. This report emphasizes the importance of: (1) access to raw data that form the background of reports and grant applications; (2) knowledge of the literature in the field; and (3) the application of statistical methods to detect anomalous numerical behaviors in raw data. Furthermore, journals and granting agencies should require that authors report failures to reproduce their published results.

Keywords: research ethics; scientific retraction; research misconduct; fabrication; falsification; experimental replication
1. Introduction

Failure to replicate reported experimental results does not necessarily imply that scientific misconduct has occurred. It can be the result of investigator error, sampling variation, lack of statistical power or differences in experimental conditions. In any case, however, we believe that scientists have an obligation to inform readers of their reports that such failure has transpired.

As a co-author of one of two papers published in Radiation Research [1,2] that deal with the survival kinetics of Chinese hamster cells exposed to tritiated thymidine, as a member of the research group here involved and as co-investigator on the NIH grant that supported the research, one of us (Helene Z. Hill) became concerned that certain experiments carried out in the laboratory failed to confirm key experiments reported in those two papers. In reviewing the contents of notebooks in the laboratory, we discovered that there were 21 attempts (10 attempts for one type of the experiments, which were called 100% experiments, 11 for the other, which were called 50% experiments) that failed to replicate the key experiments.

The two published reports involved the killing of cultured V79 Chinese hamster cells that had incorporated the radionuclide, tritiated thymidine. In the published reports, when all of the cells incorporated the radionuclide, the killing was exponential (100% experiments). When cells that had incorporated the radionuclide were incubated for three days with naive cells that had not incorporated the radionuclide, the naive cells were also killed, even though they had never been directly exposed to the radionuclide (50% experiments), and the survival kinetics were also exponential. This latter phenomenon has come to be known as “the bystander effect”. In nuclear medicine, the specialty that involves radionuclides, understanding the bystander effect is important both therapeutically and diagnostically. Setting dose levels too high could be carcinogenic or otherwise damaging to normal tissue, and setting them too low could be therapeutically ineffective or be the cause of missed diagnoses. The two articles have been cited a total of 189 times according to the Thomson Reuters Web of Science, an indication of their interest and importance.

Earlier studies of tritiated thymidine killing of cells in culture indicate that monotonic exponential killing is only expected under certain specific circumstances, which we believe were not present in the experiments in the two publications in Radiation Research. Furthermore, because we had access to the experimental protocols and the raw data that formed the background of the graphs in the two publications, we were able to perform statistical analyses of those data, as well as of the data in the 15 experiments performed by two other individuals in the laboratory that showed very different, biphasic, survival kinetics when compared with the published graphs.

The experiments reported in the two papers that we question were all performed, to the best of our knowledge, by a single individual in the laboratory, whom we will designate as A. The attempts to replicate the key experiments were performed by two other members of the same laboratory, whom we will designate as B and C. B performed 5 × 100% experiments and 5 × 50% experiments. C performed 2 × 100% experiments and 3 × 50% experiments. We identified 7 × 50% and 14 × 100% experiments that we believe provided the data for figures 3, 6 and 7 in the earlier paper [1] and figures 1 and 2A in the second paper [2]. All of these experiments were performed by A. Each of the 100% and the 50% protocols, respectively, for all the experiments performed by A, B and C were identical or nearly so. In fact, the protocols appeared to have been electronically copied from at least two master protocols.
While there were 21 experiments in all that failed to replicate A’s original experiments, only 15 studied the same cell line, V79 Chinese hamster cells, used by A. The replication attempts by B and C did not involve any additives (e.g., dimethylsulfoxide and lindane), thus we have compared those results by B and C, which we designate as controls only to A’s V79 results without such additives, in order to be sure that we were not inadvertently introducing any complicating or modifying factors.

In addition to the distinct differences in survival kinetics in the graphs in the two papers compared to the 15 unreported experiments, we identified two unusual numerical patterns underlying the data in figures 3, 6 and 7 of the first paper [1] and figures 1 and 2A of the second paper [2]. These unusual patterns did not appear in data sets produced in the experiments performed by B and C. We applied two different statistical challenges to test the null hypothesis that the unusual patterns in A’s data (without additives) may have been due to chance.

2. Analysis of Experiments

2.1. Biochemical and Radiobiological Considerations

Thymidine blocks the cell cycle at the entrance to and during the S (DNA synthesis) phase [3,4]. The block is reversed by deoxycytidine [5–9]. Nanomolar concentrations of radioactive tritiated thymidine (in the absence of deoxycytidine) both block the cell cycle and kill cells during the S phase [10–13]. Cells prevented from entering the S phase because of the tritiated thymidine block will not be killed. Therefore, the tritiated thymidine survival curves of asynchronous cells should be exponential in the presence of the antagonist, deoxycytidine (the block is abrogated, all cells can enter the S phase), and biphasic in its absence (the block prevents entry into the S phase, so only S-phase cells are killed). In an expanding population of V79 cells in tissue culture, about 40% are in the S phase; hence, they would be killed upon exposure to tritiated thymidine; 60% would not be killed. Hence, the survival kinetics would be biphasic, reflecting the change in slopes.

We found 12 reports in the early literature that pertain to the killing of tissue culture cells (including V79) by tritiated thymidine. In six reports, deoxycytidine was present in the medium, and the killing was exponential as a function of dose [14–19]. In six additional reports, deoxycytidine was absent, and the killing was biphasic as a function of dose [12,16,20–23]. In one additional case, the killing was exponential; however, the presence of deoxycytidine was not mentioned [24], and the specified medium does not contain it. In three of the papers, the protocols were very similar to the protocols in the two Radiation Research papers, i.e., no deoxycytidine was present, and the survival dose-response was biphasic [12,22,23]. These results were similar to those of B and C.

If the cell population has been synchronized, such that all of the cells were in the S phase at the time that the tritiated thymidine was added, this, too, would allow for exponential cell killing. Synchrony can be produced by mitotic selection, chemically blocking the cell cycle or releasing the cells from contact inhibition at confluence.

None of the protocols that we reviewed for this report called for the addition of deoxycytidine to the medium nor for synchronization of the cells before starting the experiments. The stated medium for the experiments was minimal essential medium (MEM) supplied by Life Technologies, Grand Island, NY, USA. The formulation listed on their website does not include deoxycytidine.
2.2. Replication Attempts: Is There a Bystander Effect?

2.2.1. B and C 100% Experiments

The seven experiments in Figure 1A performed by B and C show biphasic survival curves after following the same 100% protocols described in the two publications. The biphasic survival curves (symbols) indicate that about 40% of the population were rapidly killed, while 60% were killed much more slowly. A’s survival results of the experiments that followed the same protocols are exponential, without any change in slope, and killing is much more profound. Figure 1A reflects, for purposes of comparison, the 100% survival curves taken from the graphs in the two papers, as well as the survival calculated using the value for the parameter $A_1$ found in Table 1 of the 2001 [2] paper (cf. also figures 3 and 7 in the 1999 paper [1] and figure 1 in the 2001 paper [2]). A’s survivals are markedly different from those of B and C.

**Figure 1.** Graphs of unpublished experiments performed by B and C following the protocols for exposure of 100% (A) and 50% (B) of V79 cells to tritiated thymidine.

In 1A, the experimental points marked X are estimated survival points from the 100% curves in the two published reports. The solid line was calculated using the value for $A_1$, the theoretical cross-section for survival in mBq/labeled cell, listed in Table 1 of the 2001 paper. The various symbols represent the results of seven experiments by B and C. There are more than 100-fold more survivors in the experiments by B and C than in A’s experiments at the level of 5 mBq/labeled cell. The symbols in 1B show the results of eight 50% V79 experiments of B and C. The dashed line approximates the 50% V79 survival curves depicted in the two papers. The horizontal line is at 0.70 survival, the approximate level at which the survival curves plateau. There are about 70-fold more survivors in B’s and C’s results than in those of A at the level of 20 mBq/labeled cell.
2.2.2. B and C 50% Experiments

There is little or no indication of a bystander effect in the experiments by B and C depicted in Figure 1B. About 70% of the cells are not killed. These survivors consist of the 50% naive cells that were never directly exposed to tritiated thymidine plus the survivors from the population that were so exposed, but were not killed (a crude estimate: \(0.5 \times 1.0 + 0.5 \times 0.60 = 0.80\) expected to survive; 0.70 do survive; \(\sim 0.10\) killed as bystanders?). The dashed line in Figure 2B emulates the survival of V79 in the 50% experiments in the two papers (figures 3 and 6 in the 1999 paper [1], Figure 1 in the 2001 paper [2]).

The stated purpose of the two papers was to demonstrate that a bystander effect exists: Neighboring cells that were not directly exposed to tritiated thymidine are killed when in proximity to cells that have incorporated the radionuclide. Our crude calculations do not rule out any bystander effect, but they do indicate that under the conditions of B’s and C’s experiments, the effect is much less than that displayed in the two published papers.

2.3. Statistical Analysis of the Raw Data

The Office of Research Integrity (ORI) of the NIH describes a test that can be used on numerical data based on assumptions that: (1) in experimental count data in which terminal digits are insignificant, they are likely to be relatively uniform in their distribution; and (2) humans are not generally able to generate uniformly distributed digits. The chi-squared goodness-of-fit test can be applied to the null hypothesis that the terminal digits of a data set have been drawn from a uniform distribution, with significant results indicative of potential data manipulation or systematic or unconscious rounding error. We applied the ORI technique to analyze terminal digits, and we developed a new technique, which we describe below, to distinguish data sets that may be cause for concern.

In the notebooks we examined, we were able to identify the experiments on which the graphs in figures 3, 6 and 7 in the 1999 paper and figures 1 and 2A in the 2001 paper were based. We compared these experiments performed by A to the protocols and data generated in the 15 experiments performed by B and C. As we have noted, B’s and C’s experiments deal only with tritiated thymidine killing of V79 cells following the 100% and 50% protocols. Therefore, we limited our analyses of A’s experiments to only those experiments, or parts of the experiments, that were exactly comparable to those of B and C.

We screened two types of data: (1) integer counts copied from the digital LED display of a Coulter counter that records the pulses of individual cells in liquid suspension as a fixed volume passes through an orifice, thereby causing a change in electrical resistance; and (2) counts of colonies of cells that have arisen from single cells plated and adhering to the surfaces of 60-mm tissue culture dishes overlaid with growth medium. Most experimental points were assayed in triplicate. The terminal digit analysis does not require sample replication, and the mid-ratio analysis (see below) is based on triplicate sampling.

2.3.1. Terminal Digit Distributions of the Coulter Counts

Terminal digit distributions from 15 experiments of B and C are shown in Figure 2A and those of the 17 experiments of A in Figure 2B. The actual distributions are listed in Table 1, which also shows
the chi-squared statistics and corresponding \( p \)-values. Three of the five distributions attributed to A in Table 1 have \( p \)-values of less than 0.05 (significant), and one of those is less than \( 10^{-5} \) (highly significant).

**Figure 2.** Terminal digit distributions for coulter counts. (A) B’s and C’s Coulter digit distributions, chi-square for fit to uniform, \( p = 0.462 \); (B) A’s Coulter digit distributions, chi-square for fit to uniform, \( p = 4.49 \times 10^{-4} \).

**Table 1.** Terminal digit analysis of coulter counts of experiments (Exps) represented in the published figures. The numbers of the various digit distributions were compared to the corresponding uniform distributions to calculate the chi-squared statistic and the corresponding probabilities for nine degrees of freedom.

<table>
<thead>
<tr>
<th>Figure</th>
<th>Type</th>
<th># of Exps</th>
<th>Terminal Digit</th>
<th>Total</th>
<th>Chi-square</th>
<th>( p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Controls</td>
<td>---</td>
<td>100%</td>
<td>7</td>
<td>17 18 16 18 31 16 22 22 23 20 203</td>
<td>9.2</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>---</td>
<td>50%</td>
<td>8</td>
<td>17 22 22 23 20 23 25 33 24 21 230</td>
<td>6.8</td>
<td>0.66</td>
</tr>
<tr>
<td>1999 3 and 6</td>
<td>50%</td>
<td>2</td>
<td>5 4 12 5 2 6 3 7 13 3 60</td>
<td>21.0</td>
<td>( 1.3 \times 10^{-2} )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>100%</td>
<td>2</td>
<td>6 9 3 4 11 4 7 4 3 6 57</td>
<td>11.3</td>
<td>0.26</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>100%</td>
<td>3</td>
<td>10 3 4 1 3 7 4 4 4 2 42</td>
<td>14.2</td>
<td>0.12</td>
</tr>
<tr>
<td>2001</td>
<td>1</td>
<td>100%</td>
<td>7</td>
<td>16 22 27 17 8 24 14 12 11 23 174</td>
<td>20.7</td>
<td>( 1.4 \times 10^{-2} )</td>
</tr>
<tr>
<td></td>
<td>2A</td>
<td>50%</td>
<td>3</td>
<td>4 11 12 4 0 12 1 8 3 17 72</td>
<td>89.7</td>
<td>( 8.7 \times 10^{-6} )</td>
</tr>
</tbody>
</table>

### 2.3.2. Frequency of the Mean in Triplicate Counts

The second technique we applied was based on the observation that an unusual percentage of A’s colony count triplicates included a value that was close to the triplicate’s mean. Colony means are the data points used to construct the graphs in the papers. They provide the basis for conclusions regarding survival and bystander effects.

We first noticed the unusual frequency of near mean-containing triplicates in A’s data sets when we calculated and drew a histogram of a statistic associated with data triples, the mid-ratio, the number obtained by sorting the values of the triplicates in increasing order as \( a \), \( b \) and \( c \) and dividing \( (b-a) \) by \((c-a)\). In triplicates with a mid-ratio close to 0.5, the middle value \( b \) will be quite close to the mean.
Figure 3 shows histograms of the colony mid-ratios of B and C (3A) and of A (3B). The differences in the relative numbers of mid-ratios near 0.5 is dramatic and highlights an exceptional discrepancy between A’s colony data mid-ratios and those of B and C.

Figure 3. Mid-ratio distributions for colony counts. (A) B’s and C’s colonies, 29 of 88 in the 0.4–0.6 interval, 35.4 ± 5.19 expected ($p \sim 0.91$); (B) A’s colonies, 83 of 111 in the 0.4–0.6 interval, 34.2 ± 5.1 expected, $p \sim 3.73 \times 10^{-18}$.

Eighty-five of the 138 triplicates (61.6%) in the 17 of A’s experiments that we analyzed included their (rounded) means as one of the three members of the triplicate, whereas only 21 of the 136 triplicates (15.4%) in the 15 of B’s and C’s experiments that we analyzed did so. This disparity led us to ask whether the large number of mean-containing triplicates in A’s data might have occurred as a result of ordinary chance variation.

We developed a model, to be reported elsewhere, that allowed us to estimate the expectation and standard deviation of the number of mean-containing triplicates that could occur by chance in a given collection of triplicates. For collections of (moderately) large numbers of triplicates, the distribution of the number of mean-containing triplicates is approximately normal; thus, our estimates for expectation and standard deviation allow us to obtain reasonable estimates of the probability that a given collection might have included as many or more mean-containing triplicates than it actually did: we simply look up the tail probability corresponding to the z-score of that actual count corrected for continuity (i.e., ((actual-0.5) –expected)/standard deviation). The results of applying our model to the 136 control (data of B and C) triplicates and triplicates identified as being represented in the five figures in the two papers are shown in Table 2. The numbers of mean-containing triplicates in B’s and C’s experiments are well in line with expectation, while those for A’s experiments are all significantly larger than could reasonably be expected. The extremely significant normal tail probabilities highlight the unlikelihood that the unusually large numbers of mean-containing triplicates could have simply occurred by chance.
Table 2. Analysis of the frequency of the occurrence of the rounded mean in the triplicate samples of colony counts. Column 6 is the number N of triplicate samples that contain the rounded mean of the three counts. Column 7 is the number K that is expected based on our model. The z-scores incorporate the correction for continuity. The numerator in Column 5 is the number of qualifying triplicates (triplicates for which the gap (c-a) is greater than two). The denominator in Column 5 is the total number of triplicates.

<table>
<thead>
<tr>
<th>Figure Type</th>
<th># of Exps</th>
<th># of Triplicates</th>
<th>N = No. with Mean</th>
<th>K = No. Expected</th>
<th>SD</th>
<th>Z-Score</th>
<th>p (N ≥ K)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Controls</td>
<td>100%</td>
<td>7</td>
<td>65/67</td>
<td>14</td>
<td>12.1</td>
<td>3.2</td>
<td>0.44</td>
</tr>
<tr>
<td>---</td>
<td>50%</td>
<td>8</td>
<td>71/71</td>
<td>7</td>
<td>11.7</td>
<td>3.1</td>
<td>−1.7</td>
</tr>
<tr>
<td>1999 3&amp;6</td>
<td>50%</td>
<td>2</td>
<td>19/19</td>
<td>14</td>
<td>3.0</td>
<td>1.6</td>
<td>6.7</td>
</tr>
<tr>
<td>3</td>
<td>100%</td>
<td>2</td>
<td>20/20</td>
<td>12</td>
<td>3.1</td>
<td>1.6</td>
<td>5.2</td>
</tr>
<tr>
<td>7</td>
<td>100%</td>
<td>3</td>
<td>13/15</td>
<td>9</td>
<td>3.2</td>
<td>1.5</td>
<td>3.4</td>
</tr>
<tr>
<td>2001 1</td>
<td>100%</td>
<td>7</td>
<td>63/66</td>
<td>37</td>
<td>11.2</td>
<td>3.0</td>
<td>8.4</td>
</tr>
<tr>
<td>2A 50%</td>
<td>3</td>
<td>23/24</td>
<td>13</td>
<td>3.9</td>
<td>1.8</td>
<td>4.8</td>
<td>1.9 × 10^{-5}</td>
</tr>
</tbody>
</table>

3. Discussion

The bystander effect has had a large impact on the field of radiation biology. The Web of Science tallies 7159 related records in their core collection. The experiments that we analyzed that are the focus of the two reports [1,2] published in *Radiation Research*, purport to extend the bystander effect to nuclear medicine, potentially a significant contribution to the radiobiological literature.

The kinetics of all of the experiments in the reports are based on the kinetics of the 100% survival curves. If the 100% survival curves are exponential, then survival curves of mixtures of 50% and of 90% bystanders considered in the reports can also be exponential. If the 100% survival curves are biphasic, then it follows that survival curves of mixtures of 50% and of 90% bystanders should also be biphasic. If the kinetics of the 100% experiments are in error, then experiments involving bystander effects (50% and 90% bystanders), must also be in error. Any modifications of purported bystander effects by dimethylsulfoxide and/or lindane must also be in error (*cf.* figures 6 and 7 in the 1999 paper and figures 1, 2A and 2B in the 2001 paper).

3.1. Can the Discrepancies Be Related to the Times of the Experiments?

B’s and C’s 100% experiments have very different outcomes from those of A. Which are correct? Can they both be correct? Is it possible that conditions and circumstances could have changed so much that such different kinetics could have occurred?

The last experiment performed by A for the two papers was completed in February of 2000, while the first experiment performed by B was in mid-October of 2000, and the first performed by C was in mid-April of 2001. It is, therefore, possible that some major change in the experimental conditions occurred between February and October of 2000. There are no indications in any of the protocols as to what this change, if such one did occur, could have been. Cells do change in tissue culture, but the V79 cultures were frequently renewed from frozen stocks to avoid genetic drifting. We also found two 100% experiments in the laboratory notebooks with two other cell lines, WB mouse cells and AG1522
human cells that both showed biphasic survivals, reinforcing our notion that the cell cycle blocking effect of tritiated thymidine is a universal phenomenon.

3.2. Can the Discrepancies Be Related to Changes in the Two Protocols?

The 100% protocols and the 50% protocols respectively followed by A, B and C are essentially the same, except for differences in the tritiated thymidine concentrations, even though they span a period of more than 2.5 years. Protocol examples are available in the Supplementary Materials.

3.3. 100% Survival Results of B and C Are Compatible with the Experiments of Others

As noted in the Introduction, in six out of seven reports that we were able to find in the literature, tritiated thymidine survival curves were biphasic when no deoxycytidine was present, similar to the seven 100% experiments performed by B and C. Two of those reports were actually published after the two papers analyzed here. In six additional earlier reports, survival was exponential in the presence of deoxycytidine, while A’s survival curves were exponential in its absence. Only one earlier report [24] is compatible with A’s results: exponential survival in the apparent absence of deoxycytidine.

3.4. Statistics Can Be a Powerful Tool to Test Numerical Results

Use of statistics has been recommended as a useful tool for examining and verifying numerical results in clinical and preclinical trials [25], although it has not been put to use in either very often. The power of statistics was highlighted recently in a series of papers in the journal *Anaesthesia* that question the integrity of 168 randomized controlled trials reported by a single individual [26]. In another study, Hudes *et al.* questioned the unusual clustering of coefficients of variation in preclinical studies by three members of the same medical biochemistry department [27]. Al-Marzouki *et al.* raised questions about data obtained in a clinical dietary trial [28], and Baggerly and Coombes were able to uncover multiple errors in clinical studies purporting to predict the responses of certain cancers to chemotherapy [29]. Simonsohn used statistics alone to identify fraud by analyzing means and standard deviations derived from raw data [30]. Each statistical method employed by these various investigators was tailored to the specific situation, unlike image analysis used to detect manipulations for which one size fits all [31].

4. Conclusions

(1) A review of the literature involving the killing of tissue culture cells that have incorporated tritiated thymidine into their DNA predicts exponential survival as reported in two papers published in the journal *Radiation Research* [1,2] under conditions that were likely not present in the survival experiments performed by A.

(2) Likewise, the early and recent literature predicts biphasic survival curves that were not seen in the two papers under conditions that were likely to have been present in the experiments performed by A.

(3) The expected biphasic survival was seen in the 15 experiments performed by B and C following the same or similar protocols listed by A. These results of B and C appear to be reliable and indicate
that under the experimental conditions that prevailed in all of the studies, there is only a small bystander effect.

(4) Statistical analyses of the numerical data that form the background of survival experiments performed by A and reported in the two papers are not consistent with the expectation of the random distributions of numbers.

(5) Statistical analyses of the numerical data that form the background of survival experiments performed by B and C are consistent with the expectation of random distributions of numbers.

(6) This report emphasizes the importance of familiarity with the published literature in the field in question and of access to the raw data on which scientific reports rely.

Acknowledgments

The authors are grateful to Michael E. Robbins (deceased), a highly respected Radiation Biologist, whose analysis of the radiation biology in the two papers in *Radiation Research* reinforced the concerns regarding the survival kinetics raised regarding this report.

Author Contributions

Helene Z. Hill is Professor of Radiology, a co-investigator on the NIH grant (R01 CA083838) that supported this research and a co-author of one of the papers [2]. As a Radiation Biologist, she is responsible for the literature search, summary and analysis of the radiation biology data. Joel H. Pitt is a statistician and is responsible for the numerical analysis of the data.

Conflicts of Interest

The authors declare no conflict of interest.

References


©2014 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/3.0/).