Cross-contaminated cell lines weaken the US Patent Database

Iva Bojic,[1,2*] Jessica Snyder,[1] Aaron Gerow[2], Richard M. Neve[4], Carlo Ratti[1]

[1]Massachusetts Institute of Technology, SENSEable City Lab, Cambridge, Massachusetts 02139, USA
[2]Singapore-MIT Alliance for Research and Technology, SENSEable City Lab, Singapore
[3]The University of Chicago, Knowledge Lab, Chicago, Illinois 60637, USA
[4]Gilead Sciences, Foster City, California 94404, USA

*To whom correspondence should be addressed; E-mail ivabojic@mit.edu.

November 1, 2017

Including false data in a United States Patent, with or without the knowledge of the inventor, infringes the inventor's oath of truthfulness and may enable the patent to be later challenged. Inventors continue to cite contaminated or misidentified cell lines in patent application, despite contamination problems dating back to the first established cell lines in the early 1950's. Recent efforts to eradicate the use of misidentified and cross-contaminated cell lines in published research appear to be working as the percentage of research using "false" cell lines shows the tendency to drop. At the same time, possibly due to a lack of the control, citations of "false" cell lines in US approved patents has increased. The continued acceptance of patents citing false cells lines makes the system vulnerable to lengthly, costly legal battles if the citation is found and challenged by a competitor.

US patent database weakened by false cell lines

Including false data in a United States Patent, with or without the knowledge of the inventor, infringes the inventor's oath of truthfulness and may enable the patent to be later challenged.

"In the United States, in contrast to many other countries, inventors must sign a declaration affirming that everything in their application is true to the best of their knowledge. The inclusion of false data, even by mistake, could be an infringement of the oath, and thus against the law. Or it could form the basis for questioning the patent later, says Alan Grimaldi, co-chair of the intellectual-property group at Howrey law firm in Washington DC." ***From Reich, E.S. Bad data fail to halt patents. Nature. 439, 379 (26 January 2006)***

Mistaken identify compromises one of the most commonly used laboratory model system for drug manufacturers and basic medical research.

A 50+ year old issue

In 1952, after decades of attempts, George Gey successfully harvested robust, immortal cells that enabled him to establish the first human cancer cell line (*1*). Gey was generous in sharing his findings and sending extracted cells all around the world (*2*). However, exactly the same characteristics that helped Gey bring HeLa to life, were responsible for the cells rapidly overgrowing other cells, resulting in contamination in labs across the world (*3*). At first, it was difficult to distinguish between cell lines derived from different individuals of the same species, but by the early 1980s, Nelson-Rees relentlessly pursued and exposed intrahuman (HeLa and non-HeLa) and interspecies cross-contamination, publishing a comprehensive list of nearly 100 cross-contaminated cell lines (*4, 5*).

In the years that followed, many studies showed similar results, but received less attention until major world cell banks decided to act by informing clients or even withdrawing "false"

cell lines from their catalogs. The Deutsche Sammlung von Mikroorganismen und Zellkulturen pioneered the process in 1999 where they tested 252 human tumor cell lines stored in their repository and found that 18% of cell lines were cross-contaminated (*6*). This was followed by examining 550 human leukemia-lymphoma cell lines in 2003 where unequivocal evidence showed misidentification for 15% of them (*7*). Around the world, other repositories such as Cell Engineering Division of the Japanese research institution RIKEN (*8*) and National Cell Bank of Iran (*9*) followed suit finding similar results.

Measuring the problem's modern scope

We searched through manuscripts available from PubMed and SCOPUS databases as well as through the US patent database for instances of 3,508 cell lines identified in previous research (*10*) (See Supplementary Materials). During the last 15 years, the percentage of published manuscripts in which "false" cell lines were used shows the tendency to fall. For the same period, the percentage of filed patents citing misidentified and cross-contaminated cell lines has risen (Figure 1).

The most comprehensive database of misidentified and cross-contaminated cell lines lists 488 lines in which 451 cell lines were misidentified "early". For these, no known authentic stock exists. The remaining 37 cell lines were misidentified "late", where a tested sample had been overgrown, but authentic stock was found (*11*). Originally published in 2010, the database is now curated by the International Cell Line Authentication Committee. A recent AAAS/Science Magazine and Sigma-Aldrich's survey[1] found that less than half of survey respondents were familiar with the database, and only 11% had used it during the preceding year. This would not be a problem if other studies did not show that between 63% (*12*) and 69% (*13*) survey respondents obtained at least one cell line from their colleagues (i.e. other research laboratories).

[1] http//go.sigmaaldrich.com/Translational-Survey

When cell lines are obtained from colleagues, they typically lack verification or documentation about the condition or past number of the lines. There is thus an increasing chance that contamination goes unnoticed. Results from three independent studies are conclusive that nearly half of respondents never tested the identity of lines they used (*12–14*).

But is the number of available misidentified and cross-contaminated cell lines a good proxy for what scientists actually use in research? Until now, certain studies reported only on the misuse of a small subset of mostly HeLa-contaminated cell lines in scientific literature. In 2004 the authors searched through scientific literature on PubMed extending from 1969 to 2004 to find that in 220 published manuscripts one or more of 13 HeLa-contaminant cell lines were used as models for the tissue type of the original cell line (*12*). One year later, it was reported that 1,149 manuscripts published between 2000 and 2004 used one of six contaminated HeLa lines, a problem to which fewer than 10% admitted (*15*).

Let's acknowledge a misindentified or cross-contaminated cell line is scientifically credible as a general model of cell behavior or to produce virus, but not valid to represent the specific tissue of the cell line's original biopsy. Therefore, publications which mention cell lines from the misidentified and cross-contaminated database cannot be outright dismissed as bad science. In order to get the full picture, one would need to ascertain whether the use of the cell line was tissue specific or not. The results of previous studies, although limited by the number of cell lines they took into a consideration, gave insights that in a small portion of manuscripts authors acknowledged they were aware of contamination (e.g. fewer than 10% (*15*)). Due to the large number of manuscripts and patents in our database, it was impossible to go manually through each one as it was done in the previous studies. However, following the methodology described in Supplementary Materials, we estimated less than 10% of patents and 18% of manuscripts acknowledged that they knowingly used misidentified or cross-contaminated cell lines in their research. Many publications fail to use properly identified cell lines known to be free of contam-

ination, and when authors do use contaminated cell lines, most fail to cite the findings cannot be used to characterize a specific tissue.

More accepted patents cite false cell lines each year

Direct action reduced citations in publications

Since many scientific publishers have not authenticated the cell line's efficacy for over 50 years , it is not surprisingly that this problem has been described as the most compelling quality-control issue confronting the community (*16*). The "false" cell lines have already been unwittingly used in several hundreds of potentially misleading reports, including use as inappropriate tumor models and subclones masquerading as independent replicates. However, this study provides us with comforting results that over time researchers have been able to reduce the percentage of

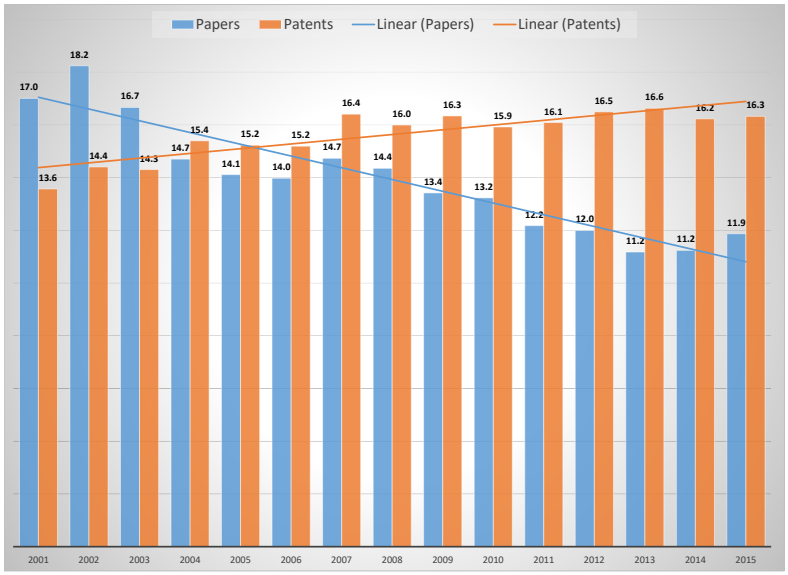

Figure 1: Percentage of manuscripts published and patents filed between 2001 and 2015 citing misidentified or cross-contaminated cell lines.

misidentified and cross-contaminated cell lines cited in manuscripts.

Looking retrospectively, elevating scientific rigor took a group of brave scientists to raise their voices. Gartler was the first, when in 1967, he reported 18 human cell lines supposedly of independent origins were all, in fact, HeLa cells (*17*). However, his findings were not well received by in the scientific community, but one researcher went into open battle to expose the problem: Nelson-Rees. In the 1970s, as Nelson-Rees ran a cell bank at Berkeley under contract for the National Cancer Institute, he was ideally positioned to expand Gartler's findings in a series of publications in *Science* (*4, 5, 18*). Publishing in on this topic certainly did not help his career because in the same year, a *Nature* referred to him as a "*self-appointed vigilante*" (*19*) after which his contract was terminated by the NIH and he gave up science to open an art gallery in San Francisco.

The third crusade began in 2004 lead by Nardone, who had been educating at the NIH about cell culture techniques for more than two decades (*20*). His white paper advocated two broad changes: more regulation and increased education efforts from different professional societies. The regulation effort should come both from funding agencies, which should require cell lines authentication to receive funds, and journals, which should not publish manuscripts using cell lines not previously authenticated. Similarly, publication of new cell lines by originators, or the funding of their production should be conditional on the lines being authenticated and made freely available to other investigators (e.g. in cell banks). Over time, it was realized that resolution of the problem of misidentification and cross-contamination requires the collaboration of other stakeholders in addition to funding agencies and publishers—it should involve users (including originators) of cell lines, cell banks / distributors, reviewers for journals and funding agencies, laboratory directors, etc..

In the late 2000s, journals like *Cell Biochemistry and Biophysics*, *In Vitro Cellular & Developmental Biology* and *International Journal of Cancer* began requiring all cell lines be au-

thenticated before publication (*21*). From 1 May 2015, all authors of manuscripts involving cell lines that are submitted to Nature journals have been asked whether they authenticated their cell lines (*22*). To date, more than 70 journals have guidelines for cell line authentication prior to publication[2]. Almost simultaneously with journals changing their policies, cell banks have begun to change their attitude as well. However, reform took almost ten years, from Stacey's recommendation on a clear identification of cross-contaminated cultures in catalogue entries of culture collections (*23*), to when the database of misidentified and cross-contaminated cell lines was created in 2010, (*11*) followed by ATCC, CellBank Australia, sDSMZ, ECACC, JCRB, and RIKEN publishing the same list on their websites.

Zero tolerance for false cell lines from NIH

The slowest adopter in the process was NIH, whose role was twofold: providing proper education and adopting zero tolerance policy on funding research that uses misidentified and cross-contaminated cell lines (*24*). With its efforts to enhance reproducibility, NIH developed a training module emphasizing good experimental design, which is now incorporated into the mandatory training on responsible conduct of research for NIH intramural postdoctoral fellows (*25*). This is very much needed, as the recent study showed that when it comes to training, only 62% of survey respondents had received specific training on the problems of cell line misidentification and cross-contamination, while less than 30% were trained the importance of cell line authentication as a quality control measure for species confirmation (*14*). Finally, beginning in 2016, "*NIH expects that key biological and / or chemical resources are regularly authenticated to ensure their identity and validity for use in the proposed studies*"[3] This is a step toward a well-defined mandate for genotyping by the current authentication standard, Short Tandem

[2]http//www.celllineauthentication.com/journal-requirements.html
[3]http//grants.nih.gov/grants/guide/notice-files/NOT-OD-15-103.html

Repeats (STR) (*26*).

However, it is alarming that in the whole body of research on misidentified and cross-contaminated cell lines, only a few scholars showed concern for the usage of "false" cell lines in patents. Namely, the only occurrence found was in the manuscript from 1998 in which Markovic & Markovic said "*filing patents and product licensing may also be compromised*" (*27*). In this paper we have indeed showed that more than 15% of patents filled in 2015 cited one of misidentified and cross-contaminated cell lines, which is 5% more than for manuscripts.

Conclusions

To this end, it is important to bare in mind the impact of cell culture contamination extends far beyond the relatively narrow field of cytobiology and the researchers studying cell lines and can for example also falsely steer radiobiologists investigating certain topics in their field (*28*). The fight with misidentified and cross-contaminated cell lines so far, has certainty inspired changes in stakeholeders' attitudes and with the recent adoption of new ANSI / ATCC standards and NIH best practices, researchers will be able to allocate dedicated funds for the authenticity check. Namely, it has already been estimated that purchasing cell lines from a reputable vendor and authenticating them annually would cost only about 0.2% of the budget for an NIH funded project (*29*).

With the increasing number of scientific journals, editors requiring or recommending cell line authentication as condition for publication and availability of education materials on this matter, in the last 15 years we were able to reduce the percentage of published manuscripts using "false" cell lines to almost 10%. This still implies that more than $350 million dollars are wasted on research annually as estimated in (*14, 29*). However, what is most worrying is not that more than 15% of cell lines used in patents are "false", but also that this percentage is

growing. This trend could possible indicate that we would need to have a better control over cell lines used in research filled for patents.

1 Supplementary Materials & Methods

A set of 644,018 manuscripts published from 1970 to 2016 was selected from medline[4] and PubMed Central[5] by one of the six following Medical Subject Headings (MeSH) terms[6] *Cell Line*, *Cell Line, Transformed*, *Cell Line, Tumor*, *Breast Neoplasms*, *Prostatic Neoplasms* and *Lung Neoplasms*. The first three MeSH terms were chosen because they contained the term *cell line* and last three because breast, prostate and lung have been the top three cancer sites since the early 1990's[7]. From this set, 169,464 full text versions were available from PubMed Central or SCOPUS[8]. A second dataset of consisted of 4,568,258 patents selected from the entirety of the US Patents[9] from January 2001 to May, 2016 by searching for forms of the word *cell* in the text and supporting documentation.

A set of 3,508 cell lines of human origin (i.e. *homo sapiens*) with unique names was compiled using data from Supplementary Table 2 and Supplementary Table 6 included in (*10*). Namely, of 3,515 human cell lines, seven were listed twice: 2B8, AC-1M46, FTC-236, HKB-11, I 9.2, NCI-N417, P3HR-1. We mapped twelve potential Contaminant Statuses to either 1, denoting contaminated cell lines, or 0 denoting non-contaminated (Supplementary Table 3). *Parental*, *Parental?*, *Derivative Line* and *Derivative Line?* statuses were only considered for cell lines whose *Name* and *Canonical Name* were different, otherwise we mapped them to 0 (see Supplementary Table 2). The entire list of cell lines is shown in Supplementary Table 1.

[4] *www.ncbi.nlm.nih.gov/pubmed*; leased data, bulk download.
[5] *www.ncbi.nlm.nih.gov/pmc*; by API.
[6] *www.ncbi.nlm.nih.gov/mesh*
[7] *www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2017.html*
[8] *www.scopus.com*; proprietary API access.
[9] *www.uspto.gov*; weekly bulk downloads.

Once when we had the whole list in Supplementary Table 1, we extracted cell line Names (i.e. the first column) from the downloaded manuscripts and patents. Variation in cell line nomenclature was accommodated by allowing a space, '-', '/' or parentheses between transitions of alphabetic and numeric characters (and vice versa) and where a delimiting character occurred in the canonical name. This collapses, for example, *MCF(7)*, *MCF7*, *MCF 7* and *MCF-7*, into the canonical form *MCF-7*. This matching procedure found 136,855 manuscripts that mentioned one or more cell lines, with a total of 2,767,589 instances overall. 181,283 patents mentioned one or more cell lines and a total of 2,829,331 cell line occurrences was found. After collapsing all instances of the same cell line in a certain manuscript / patent, 552,006 and 850,116 unique cell line occurrences remained in manuscripts and patents, respectively.

To exclude false positives, we examined the frequency distribution of cell lines occurring in a manuscripts and patents. Approximately 50% of our records are of a cell line found only once in a particular manuscript / patent (Supplementary Table 4). There is a good chance these mentions are false positives, which motivated removing mentions of cell lines with fewer than two mentions. This means we only consider instances that repeated two or more times in the same manuscript or patent. After applying this criteria, 52% and 45% of unique cell lines occurrences in manuscripts and patents remained, respectively. Because the chance a record is a false positive is correlated with the length of cell line name (44 cell line names have only two characters), we used a threshold of three. The final dataset included 233,656 unique manuscripts and 315,502 patents.

From the set of 259 parental cell lines, we chose 26 of cell lines whose parental line was HeLa (Supplementary Table 5) to assess the percentage of research acknowledging the contamination problem. We chose HeLa as a parental cell line because according the database of cross-contaminated or misidentified cell lines (*11*) HeLa is the most common. Here, it is im-

portant to note that if research requirement was for any human cell line, then it is not important whether it was HeLa or another cell line. However, in those cases where it was assumed that a specific tissue origin of the cell line was used and cell line was in fact cross-contaminated or misidentified, the work is dubious. We found 3,423 unique manuscripts and 4,579 patents cite one of 26 cell lines from Supplementary Table 5. However, only 610 manuscripts and 478 patents also mentioned HeLa. Although, the authors could acknowledge they were aware of contamination without explicitly mention HeLa, this is likely a small portion of research.

References and Notes

1. G. Gey, W. D. Coffman, M. T. Kubicek, *Cancer research* **12**, 264 (1952).

2. B. J. Culliton, *Science* **184**, 1058 (1974).

3. R. Chatterjee, *Science* **315**, 928 (2007).

4. W. A. Nelson-Rees, R. R. Flandermeyer, *Science* **191**, 96 (1976).

5. W. Nelson-Rees, D. Daniels, R. Flandermeyer, *Science* **212**, 446 (1981).

6. R. A. MacLeod, *et al.*, *International Journal of Cancer* **83**, 555 (1999).

7. H. Drexler, W. Dirks, Y. Matsuo, R. MacLeod, *Leukemia* **17**, 416 (2003).

8. K. Yoshino, *et al.*, *Human Cell* **19**, 43 (2006).

9. S. Azari, N. Ahmadi, M. J. Tehrani, F. Shokri, *Biologicals* **35**, 195 (2007).

10. M. Yu, *et al.*, *Nature* **520**, 307 (2015).

11. A. Capes-Davis, *et al.*, *International journal of cancer* **127**, 1 (2010).

12. G. C. Buehring, E. A. Eby, M. J. Eby, *In Vitro Cellular & Developmental Biology-Animal* **40**, 211 (2004).

13. M. Shannon, *et al.*, *International Journal of Cancer* **138**, 664 (2016).

14. L. P. Freedman, *et al.*, *BioTechniques* **59**, 189 (2014).

15. J. R. Masters, *In Vitro Cellular & Developmental Biology-Animal* (2005), vol. 41, pp. 6A–6A.

16. R. M. Nardone, *Biotechniques* **45**, 221 (2008).

17. S. M. Gartler, *National Cancer Institute Monograph* **26**, 167 (1967).

18. W. A. Nelson-Rees, R. R. Flandermeyer, P. K. Hawthorne, *Science* **184**, 1093 (1974).

19. J. Maddox, *Nature* **289**, 212 (1981).

20. R. M. Nardone, *Cell biology and toxicology* **23**, 367 (2007).

21. A. T. C. C. S. D. O. W. ASN-0002, *et al.*, *Nature Reviews Cancer* **10** (2010).

22. Editorial, *Nature* **520**, 264 (2015).

23. G. Stacey, *Nature* **403**, 356 (2000).

24. J. R. Lorsch, F. S. Collins, J. Lippincott-Schwartz, *Science* **346**, 1452 (2014).

25. F. S. Collins, L. A. Tabak, *Nature* **505**, 612 (2014).

26. J. Masters, *et al.*, *ATCC® Standards Development Organization* (2012).

27. O. Markovic, N. Markovic, *In Vitro Cellular & Developmental Biology-Animal* **34**, 1 (1998).

28. B. P. Lucey, W. A. Nelson-Rees, G. M. Hutchins, *Archives of pathology & laboratory medicine* **133**, 1463 (2009).

29. L. P. Freedman, I. M. Cockburn, T. S. Simcoe, *PLoS Biol* **13**, e1002165 (2015).