Manuscript Draft prepared for Nature

Presubmission Inquiry: <http://www.nature.com/nature/authors/submissions/presubs/index.html>

**Formatting:** Letters are short reports of original research focused on an outstanding finding whose importance means that it will be of interest to scientists in other fields. They do not normally exceed **4 pages** of *Nature* and, as a guideline, allow up to 30 references. They begin with a **fully referenced paragraph**, ideally of about 200 words, but certainly no more than 300 words, aimed at readers in other disciplines. This paragraph starts with a 2-3 sentence basic introduction to the field; followed by a one-sentence statement of the main conclusions starting 'Here we show' or equivalent phrase; and finally, 2-3 sentences putting the main findings into general context so it is clear how the results described in the paper have moved the field forwards. Please refer to our [annotated example](http://www.nature.com/nature/authors/gta/2c_Summary_para.pdf) to see how the summary paragraph for a Letter should be constructed. The rest of the text is typically about 1,500 words long. Any discussion at the end of the text should be as succinct as possible, not repeating previous summary/introduction material, to briefly convey the general relevance of the work. Letters typically have 3 or 4 small display items (figures or tables). Word counts refer to the text of the paper. References, title, author list and acknowledgements do **not** have to be included in total word counts.

(Summary paragraph 200-300 words: <http://s3-service-broker-live-19ea8b98-4d41-4cb4-be4c-d68f4963b7dd.s3.amazonaws.com/uploads/ckeditor/attachments/7820/2c_Summary_para.pdf>)

**Title:** Representational Bias is Human Cell Line Studies

**Date:** January 5, 2017

**Author:**  Iva Bojic1,2

Jessica Snyder1

Aaron Gerow2

Richard M. Neve4

Carlo Ratti1

**Affiliation:** 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

**Correspondence**: ivabojic@mit.edu, [snyderj@mit.edu](mailto:snyderj@mit.edu).

Summary: The inventors citing false cell lines are in breach of the truthfullness oath.

**Representational Bias is Human Cell Line Studies**

If an inventor included data found using a contaminated cell line in their U.S. patent application, then the inventor violated their truthfulness oath, thereby weakening their intellectual property claims (Reich, 2006). Likewise, we wonder if generalizing lab results to an entire patient population - without considering how the fundamental mechanism of the results may differ across the population’s spectrum of genotypes or lifestyle - also violates the truthfulness oath and consequently weakens the claims. A narrow set of legacy donors define the cell line catalog used in early cancer research. Cell lines offer the advantages of any standard, but risk biasing treatment options for the cell line donors. Results extrapolated from a cell line, donated by an individual, to an entire disease subvert the inclusive policies set up by the National Institute of Health for genders and ethnicities. To assess the extent of patent and publication’s reliance on cell lines as standards, we built an analytical tool as part of a quality control effort. First, we counted mentions of cell lines known to be contaminated using text analysis in patents and publications. The results showed action taken by editorial boards and federal funding offices over the years curbed the use of contaminated cell lines in publications: while citations in patents increased. Next, we compared representation of cell lines for each Center for Disease Control (CDC) ethnicity in the database to the prevalence of each ethnicity in the breast cancer population. Results showed a significant, but decreasing bias for cell lines derived from white patients for breast cancer research. The proposed methodology enables future work tracking the inclusivity of cell line studies. The goal being to select authenticated cell lines reflective of the patient populations. This could not only improve patient outcomes, but also expose patient groups still in need of high efficacy treatments.

# The US Population and the Human Cell Line Toolkit

The U.S. Census Bureau expanded its lexicon of race and ethnic groups to include a new distinction, Two or More Races - projected to be the fastest growing group through 2060[[1]](#footnote-1). The U.S. Census Bureau estimated ethnic minorities’ newborns have outnumbered non-Hispanic white newborns since 2013[[2]](#footnote-2). Our population reflects an increasingly global ethnic demographic, as well as newly blended profiles unmixed before. The nation’s medical community must decide if this progression necessitates a reengineering of inclusive standards for practitioners and researchers. When envisioning the coming decades of research and development, if and how laboratory research reflects the population’s -omics profile deserves data-driven evaluation. High-throughput screening of drug candidates rely on standards of human tissue, the human cell line toolkit, which as of 2018 represents cells from predominantly white donors. However, the cell line catalog began from a black woman’s biopsy and her cells remains a standard for in vitro research today.

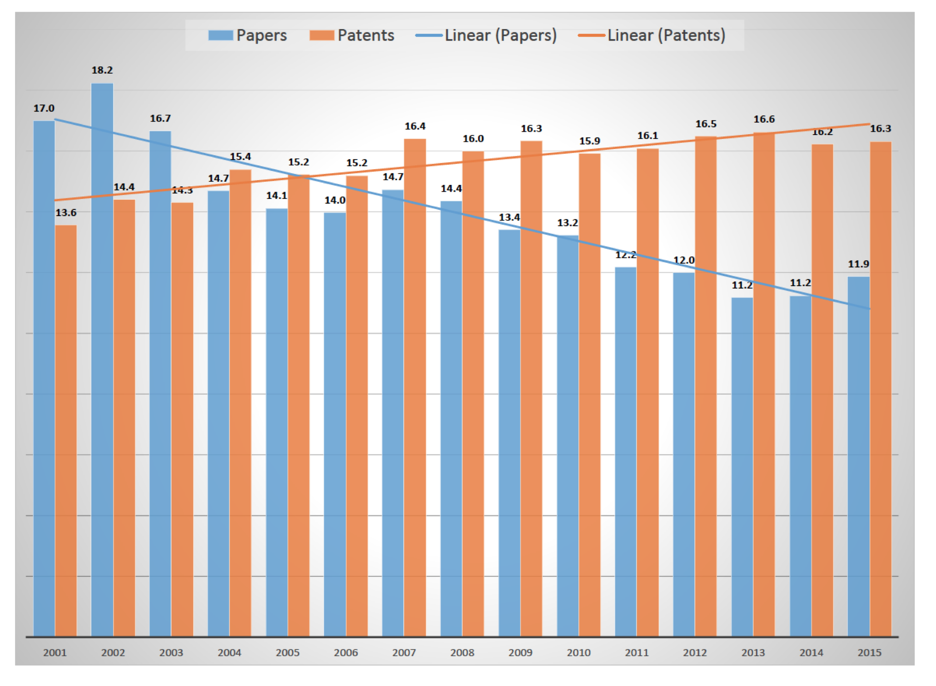
Namely, George Gey and Mary Kubicek established the first immortal human cell line in 1952 from a biopsy of a cervical adenocarcinoma taken from Henrietta Lacks (1). Gey distributed Lack’s “HeLa” to researchers globally, introducing the first standard for *in vitro* human cell research (2). However, the rapid growth that enabled Lack’s cells to grow outside her body also enabled HeLa cells to overgrow and contaminate other cultures (3). Three decades later, Nelson-Rees published the first list of nearly 100 misidentified or non-homogenous cell lines, including intra human (HeLa and non-HeLa) and interspecies contamination (4, 5). In 2010, the International Cell Line Authentication Committee (ICLAC) published the most comprehensive database of 488 misidentified cell lines, of which 451 cell lines were misidentified “early” - no known authentic stock exists. The remaining 37 cell lines were misidentified “late” - authentic stock was found (6). Despite the ICLA actively curating the list, a minority of researchers check if their cell lines are known to be false. The AAAS/Science Magazine and Sigma-Aldrich’s survey[[3]](#footnote-4) found 11% of respondents used the ICLAC’s list in the preceding year and less than half knew of the database. Other surveys found 63% (7) - 69% (8) of respondents obtained one or more cell lines from their colleagues without substantiating their authenticity using American National Standards Institute/American Tissue Culture Collection (ANSI/ATCC) or National Institutes of Health (NIH) standards.

Citing contaminated and misidentified cells comprises the scientific integrity of a researcher’s work. If that work becomes part of a U.S. patent application, the misrepresentation of cell type by using a contaminated or false cell line could violate the truthfulness oath required by the U.S. Patent Office, making the inventor’s claim vulnerable to legal challenge by outside parties. Another possible violation of the truthfulness oath might be generalizing results from a cell line models across an entire patient population without basis. Here we build a database from the CDC, PubMed, SCOPUS, and the U.S. Patent Office to define the scope of the misidentified cells and bias in cell line studies. Legal grounds to challenge the truthfulness oath is outside the scope of this work.

# Persistent Use of Contaminated Cell Lines

Misidentified and contaminated cell lines have been described as the most compelling quality-control issue confronting the research community (9). “False” cell lines have already been unwittingly used in several thousands of potentially misleading reports. In this study, we use text analysis of publications and patents to assess the scope and history of false cell line use from 2001-2015. We built a database of manuscripts from PubMed, SCOPUS, and the U.S. Patent Office, which we searched using string matching for mentions of the 3,508 cell lines named in the International Cell Line Authentication Committee (ICLAC) inventory (10) (See Supplementary Materials). The ICLAC characterized each cell line as false - if it was contaminated or misidentified - or reliable - if intact. We categorized each manuscript in the database as either false (if any false cell line was cited) or reliable (if only reliable cell lines were cited). We divided the number of false manuscripts by the total number each year to calculate the percentage of false manuscripts from 2001-2015, shown in Figure 1. The number of peer reviewed publications citing false cell lines decreased, while the percentage of U.S. patents increased.

Previous studies, although limited by the number of cell lines they took into a consideration, show a <10% of false manuscripts authors acknowledged they were aware of contamination (11). 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.



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

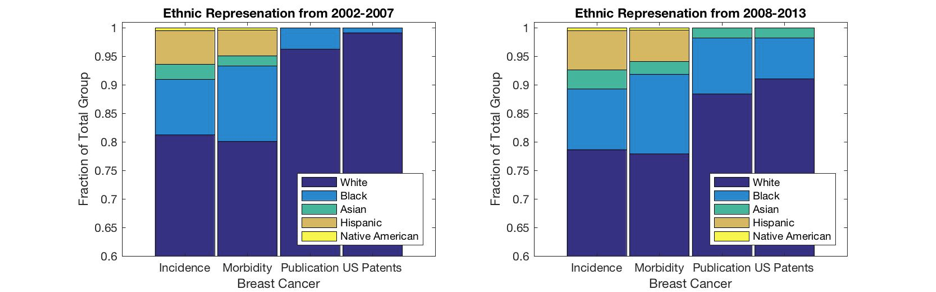
# Ethnic Representation in Human Cell Line Studies

We compared the demographics of the breast cancer patient population to the demographics of the breast cancer cell lines cited in publications and patents. Using the database built to quantify contaminated cell lines and the ICLAC - which characterized the cell line donor’s gender and ethnicity, if known - all the manuscripts citing a breast cancer cell line were categorized by the ethnicity of the cell line’s donor.

Breast cancer research relied on a standardized set of less than 70 cell lines since the late 1950’s. The individual genomic sensitivities of those few donors appreciably shaped the global, industrial pharmaceutical enterprise, responsible for treating the nearly 300k patients diagnosed with breast cancer in the U.S. alone. The efficacy of breast cancer treatment can be found in the U.S. CDC reports, which reports a sustained increasing trend of survival, the number of women whose cause of death is breast cancer has decreased each year since 1999.

However, when survival rates are grouped by race, Caucasian woman are most likely to survive, while black women are least likely to survive a breast cancer diagnosis (12) (13). Of the numerous factor contributing to this inequality, we investigated the preclinical models used to build treatments, finding more than 70% of the breast cancer cell lines were from white donors. We ask if there is a historical bias to use breast cancer cell lines derived from Caucasian donors. Using the database, we found <15% of breast cancer cell lines cited in peer-reviewed publications are non-white, even less in U.S. patents. We compared the demographics of cell lines cited in manuscripts 20% non-white in the CDC patient population reported to be 1.2 million people. Does this bias meaningfully bias pharmacologic formulation for genomes similar to a small group of Caucasian donors? These cells for analysis of the genetic profiles - accessible through tools such as the Cancer Cell Line Encyclopedia (14), enable researchers to assess the genomic sensitivities implicit in a cell line choice, exposing unrepresented profiles in preclinical study models.

From 2002-2007, the U.S. CDC reports 1.2 million people diagnosed with breast cancer diagnosis. More women were diagnosed from 2008-2013, while the number of woman who during the same interval of time decreased. From 2008-2013, breast cancer patients numbered 1.3 million people being treated Using the ethnicities as defined by the CDC, patients included 78-79% White, 11-14% Black, 2-3% Hispanic, .05-.07% Asian/Pacific Islander, and .004-.005% American Indian/Native American. More recent CDC data was not available to extend the study further. However, the tools to build the dataset will be available to continue this accounting.



**Figure 2. Representation of CDC defined ethnic groups in the breast cancer patient population, publication record and U.S. patent database for two 6-year periods.**

# 4. Looking forward

We recognize the selection criteria for in vitro cell models affects the final efficacy and cost of breast cancer treatments (28). Since the founding of cell lines as an *in vitro* model more than 60 years ago, the research community has fought to eradicate misidentified and cross-contaminated cell lines. ANSI/ATCC and NIH standards enforced by journal editors and funding agencies oblige researchers to authenticate cell lines as reliable - estimates for purchasing cell lines from a reputable vendor and annual authentication would cost ~0.2% of the budget for an NIH funded project (15).

# 

These efforts have driven down the percentage of false cell line citation in patents. However, according to the U.S. patent database, inventors continue to cite erroneous cell lines in patent applications. The continued acceptance of patents citing false cells lines makes the system vulnerable to lengthy, costly claim disputes based on breaches of the truthfulness oath signed by the inventors - whereby if breached even unknowingly - could invalidate the claim.

In the years that followed, more studies exposed a worrying level of cell line contaminations and misidentification, but generally fell on deaf ears until major cell line repositories began to inform clients of “false” cell lines, even withdrawing them from distribution. The Deutsche Sammlung von Mikroorganismen und Zellkulturen (DSMZ) repository 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 (16). This was followed by examining 550 human leukemia-lymphoma cell lines in 2003 where unequivocal evidence showed misidentification for 15% of them (17). Around the world, other repositories such as Cell Engineering Division of the Japanese research institution, RIKEN (18), and National Cell Bank of Iran (19) followed suit finding similar results. In 2017, an internal audit determined nearly half of the 278 cell lines used by more than 20 Chinese institutions were cross-contaminated or misidentified (20). A majority of those misidentified cell lines allegedly established by a Chinese lab were genuinely HeLa cells or HeLa hybrids.

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 (15, 21). 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 filed for patents.

# 5. References

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. A. Capes-Davis, et al., International journal of cancer 127, 1 (2010).
7. G. C. Buehring, E. A. Eby, M. J. Eby, In Vitro Cellular & Developmental Biology-Animal  
   40, 211 (2004).
8. M. Shannon, et al., International Journal of Cancer 138, 664 (2016).
9. M. Yu, et al., Nature 520, 307 (2015).
10. J. R. Masters, In Vitro Cellular & Developmental Biology-Animal 41 (2005).
11. DeSantis CE, et al., CA: A cancer journal for clinicians 66, 1 (2016).
12. Iqbal J, et al., Jama, 313, 2 (2015).
13. Barretina J, et al., Nature 483, 7391 (2012).
14. L. P. Freedman, I. M. Cockburn, T. S. Simcoe, PLoS Biol 13 (2015).
15. R. A. MacLeod, et al., International Journal of Cancer 83, 555 (1999).
16. H. Drexler, W. Dirks, Y. Matsuo, R. MacLeod, Leukemia 17, 416 (2003).
17. K. Yoshino, et al., Human Cell 19, 43 (2006).
18. S. Azari, N. Ahmadi, M. J. Tehrani, F. Shokri, Biologicals 35, 195 (2007).
19. Huang, Yaqing, et al., PloS One 12, 1 (2017).
20. L. P. Freedman, et al., BioTechniques 59, 189 (2014).

# 6. Supplemental Materials

A set of 644,018 manuscripts published from 1970 to 2016 was selected from Medline[[4]](#footnote-9) and PubMed Central[[5]](#footnote-10) by one of the six following Medical Subject Headings (MeSH)[[6]](#footnote-11) terms *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]](#footnote-12). From this set, 169,464 full text versions were available from PubMed Central or SCOPUS[[8]](#footnote-13). A second dataset consisted of 4,568,258 patents selected from the entirety of the U.S. patents[[9]](#footnote-14) 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 and P3HR-1. The table describes characteristics of each cell line, those being cell line name, canonical name, diagnosis, gender, ethnicity, and contaminant status. Indexing all options for each characteristics allowed for like cell lines to be grouped during analysis; for example take the characteristic gender, options include female, male and unknown, which were indexed as 1, 2, and 3 respectively. Diagnosis and ethnicity were similarly indexed. 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.

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 these 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 (6) HeLa is the most common. Here, it is important 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 citing 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.

The CDC published the United States Cancer Statistics (USCS), which registered the incidence and death counts for each cancer type by race (age-adjusted). The three cancer sites analyzed are consistently the most prevalent, those being female breast, prostate, and lung and bronchus. The Census Bureau began providing prevalence metrics for races other than White and Black after 1990. The most recent year available is 2014. The CDC National Program of Cancer Registries (NPCR) and National Cancer Institute (NCI) Surveillance, Epidemiology, and End Results (SEER) Public Use Data recorded the number of people alive who have survived a diagnosis or are living with the diagnosis; complete prevalence.

1. <https://census.gov/content/dam/Census/library/publications/2018/demo/P25_1144.pdf> [↑](#footnote-ref-1)
2. <http://www.pewresearch.org/fact-tank/2016/06/23/its-official-minority-babies-are-the-majority-among-the-nations-infants-but-only-just/> [↑](#footnote-ref-2)
3. http//go.sigmaaldrich.com/Translational-Survey [↑](#footnote-ref-4)
4. [www.ncbi.nlm.nih.gov/pubmed](http://www.ncbi.nlm.nih.gov/pubmed) [↑](#footnote-ref-9)
5. [www.ncbi.nlm.nih.gov/pmc](http://www.ncbi.nlm.nih.gov/pmc) [↑](#footnote-ref-10)
6. [www.ncbi.nlm.nih.gov/mesh](http://www.ncbi.nlm.nih.gov/mesh) [↑](#footnote-ref-11)
7. [www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2017.html](http://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2017.html) [↑](#footnote-ref-12)
8. [www.scopus.com](http://www.scopus.com) [↑](#footnote-ref-13)
9. [www.uspto.gov](http://www.uspto.gov) [↑](#footnote-ref-14)