Demographic Bias in Human Cell Studies

Jessica Snyder[1], Iva Bojic[1,2], Aaron Gerow[3], Carlo Ratti [1,2]

[1]*Massachusetts Institute of Technology, SENSEable City Lab*
[2]*Singapore-MIT Alliance for Research and Technology*
[3]*University of Chicago, Computation Institute*

May 17, 2017

Breast cancer treatments are shaped by the research tools used to understand the disease. As these tools become increasingly focused, it is important to understand how inclusive they are of treatment populations. The use of human cell lines, in particular, in the development of cancer treatments poses serious questions about representation. Looking at the demographics of cell lines in published research, we find a discernible bias for the majority patient ethnicity.

From 1999 to 2013, the population of female breast cancer patients has been proportionally shrinking (figure 1A). The population-level decrease indicates that environmental stressors triggering breast cancer are likely being reduced. One explanation is that decreased prescription of a hormone-replacement therapy (HRT), which is shown to increase white and Hispanic patient's risk of breast cancer by >20% [1]. HRT was not, however, found to increase risk for black women. Because HRTs were limited in 2002, the rate of breast cancer incidence among white, Hispanic and Native American women have fallen, while the rate for black and Asian women increased over the same time. Looking simply at ethnicity, the breast cancer population is a heterogeneous disease, having sub-types with their own risk factors and pathologies.

The prognosis for women diagnosed with breast cancer has improved, in part due to advances in screening, early detection, and increasing survival rates, especially for women under 50 years old [2]. Of deaths caused by breast cancer, black women had the highest rate from 1999-2013, higher than white women, who are more likely to be diagnosed (figure 1B). This might be due to later stages of diagnosis and more aggressive tumor subtypes, both of which disproportionately affect black patients [3]. Characterization of each breast cancer sub-type provides the scientific basis for treatment options, like it already has for the most common forms as shown by the increasing survival rate. Our question is, which ethnicities, as a proxy for tumor sub-types, are represented in the medical research?

The first breast cancer cell line was donated by a 74 year old female patient in 1958, named BT-20. BT-20 remained the only breast cancer cell line for more than a decade, until other cells were cultured from white female patients in the 1970s. A list of all human cell lines – not just cancer cells – was developed in previous work to assess cell line authentication and quality control, inventory cell line features [4]. While researchers conceivable have access to the full range of cell lines, in practice, many lines are used as standards to compare new findings to previous work. The first cell line derived from a black patient's biopsy was in 1974, followed by the first Asian breast cancer cell lines 20 years later in 1994 and the first Hispanic line was developed in 1995. There are currently no Native American breast cancer cell lines currently available. To date, 134 breast cancer cell lines derived from humans are available. Of which, 68 have declared an ethnicity for the donor. Of those, the donors include a majority of 50 lines from white donors, 12 from black donors, 4 from Hispanic donors, and 2 from Asian donors.

In published research, citations to cell lines suggest that the desire for standards exerts a homogenizing force. Among 1.2 million publications from 1975 to 2016, cited a human cell line derived from breast cancer (see Appendix; Methods & Materials). 85.6% cited a cell line from a white donor. The next most frequent cell line citation was to donors of unknown ethnicity, 8.8%, then black donors at 5.0%, and at order of magnitude less publications cited cell lines from Asian donors at 0.5%, and another order of magnitude less were declared Hispanic cell lines at 0.1%. There were no citations of lines from Native

American patients as none were available (figure 1C).

Publication are the typical output of research communities, but scientific findings are also made in commercial settings. Like publications, U.S. Patents show evidence of developing treatments. From 2001 to 2015 shows, the majority of cell lines in patents were from a white donor, the second most being black. No other ethnicities were represented in patents (figure 1D).

Using statistics from the U.S. CDC, for each ethnic population, the fraction of incidence and death was calculated. Similarly, for each ethnicity, the fraction of cited cell lines in publications and patents was calculated. These proportions were used to estimate bias in research. First, there is a bias in treatment efficacy: subtracting each ethnicity's fraction in the incidence from the death population, black patient are the only group for which the death population is greater than incidence, by 0.4%.There is also bias in the publication record: blacks represent about 10% more of the death population than citations of black cell lines in publications (figure 1F). Whites are the only ethnic group represented more in publications than in the death population, by more than 15%. This finding is true also for patents, which favor the majority ethnicity (figure 1G). Comparing the representation of each ethnic group in 2013 shows a more proportional representation of the majority patient ethnicity and dependence on cell model standards, figure 1H.

The use of cell models may hinder generalization of results. More than 50,000 women will be diagnosed with breast cancer between 2010-2020 in the U.S. alone, at a medical cost estimated at $48 billion [5, 6]. The National Institutes of Health (NIH) and the National Cancer Institute (NCI) invested a combined $4.1 billion in breast cancer research from 2012 through 2014. Assessments of disease burden and treatment effectiveness are used to guide funding for medical research and policy toward the most productive outcomes [7]. To this end, as black patients comprise more of the death population than incidence, CDC statistics suggests that breast cancer treatment is less effective for black women than women of all other races. Despite the fact that previous studies showed black breast cancer patients presented elevated risk for aggressive, intrinsic factors of breast cancer [8, 9], instead of assigning more resources into research using cell lines from black donors, the opposite is the case: whites are over-represented in research by more than 15%.

Today, the only available regulative apparatus is found in the US Food and Drug Administration (FDA) which in 1998, when pharmaceutical efficacy was found to be sensitize to genetic factors, responded by implementing the Demographic Rule. The Rule established guidelines for racial and ethnic representativeness during clinical trails. However, clinical trials are one of the final stages of work that often begins with the use of cell models, animal models, and finally human models. Should such guidelines be extended to the preclinical stage? Recent efforts from the NIH to balance sex in cell and animal studies suggests so [10]. The 1993 NIH Revitalization Act aimed to increase representation of women and minorities in clinical trials, but did not address earlier stages of research. Despite multiple calls for action, the publication record continues to neglect ethnic considerations, the effects of which surely permeate later-stage research. While legislative solutions may be too restrictive for all research, prioritizing inclusiveness and diversity in the use of cell lines could be feasibly integrated into review and editorial processes.

Initiatives such as Cancer Moon Shot's database, which catalogs tumor sub-types from the patient populations, could help reduce dependence on standard models, the use of which passively perpetuates the bias in early-stage research, by highlighting patients without effective treatment options [11]. We hope similar such collaborative approaches can populate the human cell line bank and medical research community to discover the multitude of pathologies involved in breast cancer, alleviating the community's dependence on standards as we realize the potential of personalized medicine.

References

[1] Million Women Study Collaborators et al. Breast cancer and hormone-replacement therapy in the million women study. *The Lancet*, 362(9382):419–427, 2003.

[2] Ruth Etzioni, Nicole Urban, Scott Ramsey, Martin McIntosh, Stephen Schwartz, Brian Reid, Jerald Radich, Garnet Anderson, and Leland Hartwell. The case for early detection. *Nature Reviews Cancer*, 3(4):243–252, 2003.

[3] Nataliya G Batina, Amy Trentham-Dietz, Ronald E Gangnon, Brian L Sprague, Marjorie A Rosenberg, Natasha K Stout, Dennis G Fryback, and Oguzhan Alagoz. Variation in tumor natural history contributes to

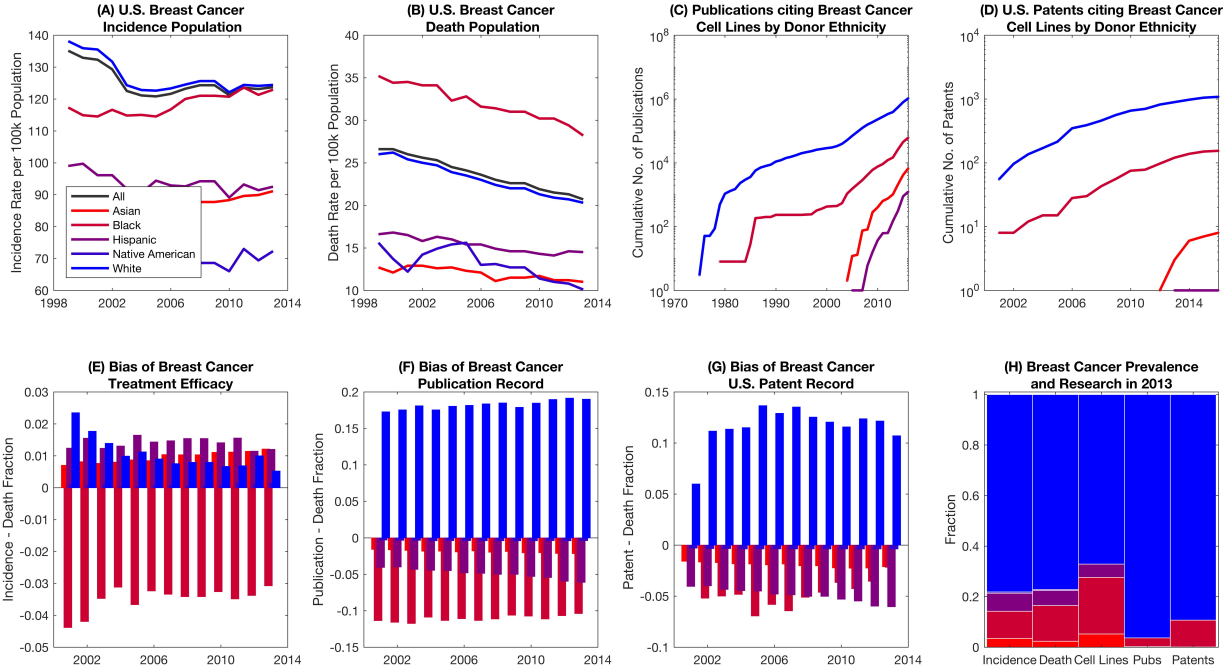

Figure 1: Breast cancer prevalence and research.

racial disparities in breast cancer stage at diagnosis. *Breast Cancer Research and Treatment*, 138(2):519–528, 2013.

[4] Mamie Yu, Suresh K Selvaraj, May MY Liang-Chu, Sahar Aghajani, Matthew Busse, Jean Yuan, Genee Lee, Franklin Peale, Christiaan Klijn, Richard Bourgon, et al. A resource for cell line authentication, annotation and quality control. *Nature*, 520(7547):307–311, 2015.

[5] Angela B Mariotto, K Robin Yabroff, Yongwu Shao, Eric J Feuer, and Martin L Brown. Projections of the cost of cancer care in the United States: 2010–2020. *Journal of the National Cancer Institute*, 2011.

[6] Hannah K Weir, Trevor D Thompson, Ashwini Soman, Bjørn Møller, and Steven Leadbetter. The past, present, and future of cancer incidence in the United States: 1975 through 2020. *Cancer*, 121(11):1827–1837, 2015.

[7] Jane J Kim, Anna NA Tosteson, Ann G Zauber, Brian L Sprague, Natasha K Stout, Oguzhan Alagoz, Amy Trentham-Dietz, Katrina Armstrong, Sandi L Pruitt, Carolyn M Rutter, et al. Cancer models and real-world data: Better together. *Journal of the National Cancer Institute*, 108(2):djv316, 2016.

[8] Dezheng Huo, Francis Ikpatt, Andrey Khramtsov, Jean-Marie Dangou, Rita Nanda, James Dignam, Bifeng Zhang, Tatyana Grushko, Chunling Zhang, Olayiwola Oluwasola, et al. Population differences in breast cancer: Survey in indigenous African women reveals over-representation of triple-negative breast cancer. *Journal of Clinical Oncology*, 27(27):4515–4521, 2009.

[9] Kerryn W Reding, Christopher S Carlson, Orsalem Kahsai, Christina C Chen, Andrew McDavid, David R Doody, Chu Chen, Kimberly Lowe, Leslie Bernstein, Linda Weiss, et al. Examination of ancestral informative markers and self-reported race with tumor characteristics of breast cancer among black and white women. *Breast Cancer Research and Treatment*, 134(2):801–809, 2012.

[10] Janine A Clayton, Francis S Collins, et al. NIH to balance sex in cell and animal studies. *Nature*, 509(7500):282–283, 2014.

[11] Douglas Lowy, Dinah Singer, Ron DePinho, Gregory C Simon, and Patrick Soon-Shiong. Cancer moonshot countdown. *Nature Biotechnology*, 34(6):596–599, 2016.

Appendix

The same analysis was conducted from prostate and lung cancer.

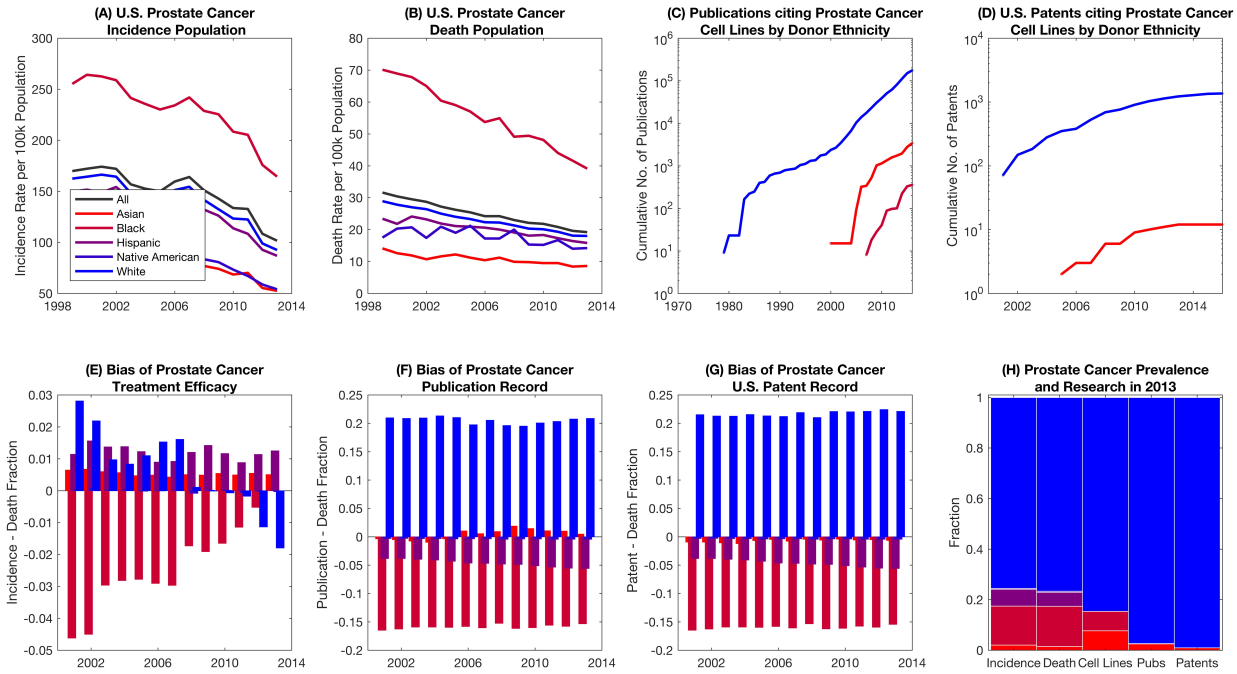

Figure 2: Prostate cancer

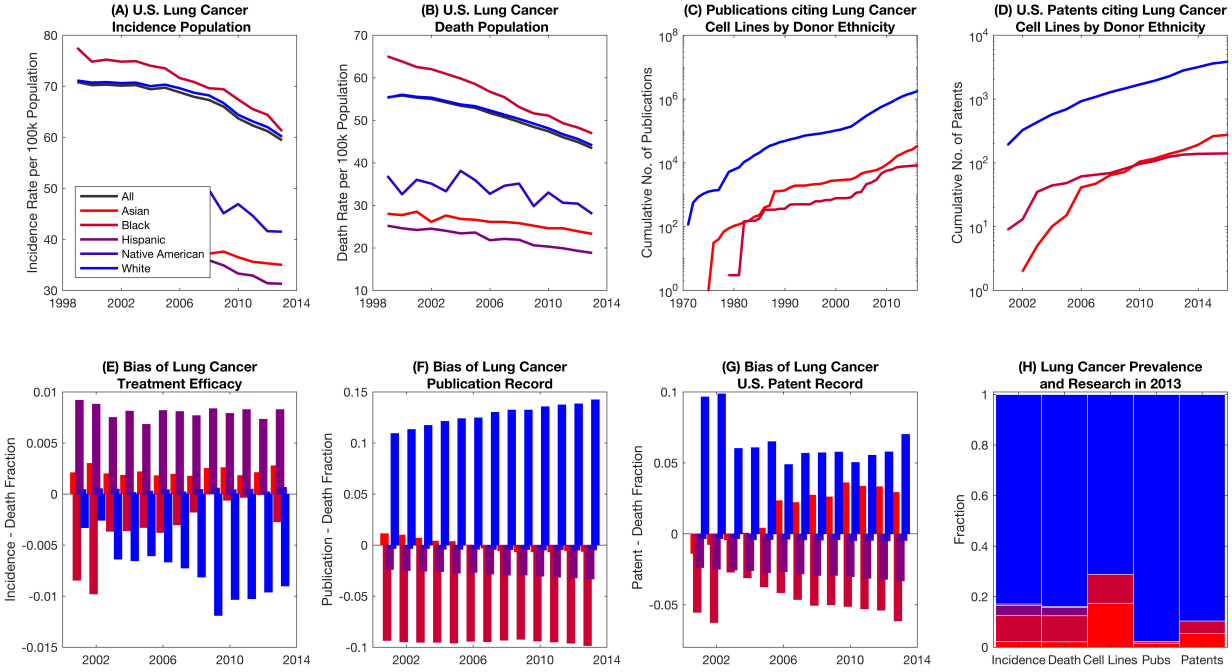

Figure 3: Lung cancer.

Methods & Materials

A set of 644,018 manuscripts published from 1970 to 2016 was selected from medline[1] and PubMed Central[2] by one of the six following Medical Subject Headings (MeSH) terms[3] *"Cell Line"*, *"Cell Line, Transformed"*, *"Cell Line, Tumor"*, *"Breast Neoplasms"*, *"Prostatic Neoplasms"* and *"Lung Neoplasms"*. The first three MeSH terms were chosen because they were the only ones containing *cell line* term and last three because female breast, prostate and lung have been the top three cancer sites since the early 1990's[4]. From this set of 644,018 manuscripts, 169,464 full text versions were available for download from PubMed Central or SCOPUS[5]. Moreover, a set of 4,568,258 patents was selected from the entirety of the US Patents[6] from January 1990 to May, 2016 by performing a string matching for word *"cell"* in the patent text (including their supporting documentation).

A set of 3,508 cell lines of human orign (i.e. *homo sapiens*) with unique names was compiled using data from Supplementary Table 2 and Supplementary Table 6 included in the supporting data for [4]. Namely, out of 3515 human cell lines from Supplementary Table 2, seven of them were listed twice 2B8, AC-1M46, FTC-236, HKB-11, I 9.2, NCI-N417, P3HR-1. Moreover, we did mapping between their twelve Contaminant Statuses and our two Contaminant Indexes ("1" denoting contaminated cell lines) as it is shown in our Supplementary Table 3. When it comes to their *"Parental"* and *"Parental?"* statuses, we consider that only those cell lines whose *"Name"* and *"Canonical name"* are different, are contaminated. Otherwise we would map them into "0" (please refer to our Supplementary Table 2). We need explanation about other mapping, i.e. Dis, Gen, Eth...

The whole list of cell lines we used is shown in our Supplementary Table 1.

Once when we had the whole list in Supplementary Table 1, we extracted cell lines Name (i.e. the first column) from the sets of downloaded manuscripts and patents. Variation in cell line nomenclature was accommodated by

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

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 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 lines in a certain manuscript/patent, we were left in total with 552,006 and 850,116 unique cell line occurrences in manuscripts and patents, respectively.

In order to exclude false positive records, we made a histogram showing how often the same cell line is occurring in a manuscript/patent. Results showed in Supplementary Table 4 confirmed that approximately 50% of our records is of cell lines that were found in a certain manuscript/patent only once. Since for those records, there is a higher chance to be false positives, we decided to use a threshold of two. This means that we are considering only those instances of cell lines that repeated at least for two times in the same manuscript/patent. By imposing this rule we were left with 52% and 45% of unique cell lines occurrences in manuscripts and patents, respectively. Moreover, since the chance that a record is a false positive is correlated with the length of cell line name, for 44 cell lines whose name length is two, we used a threshold of three. Finally, we omitted all cell line occurrences of "false" cell lines. In the end we were left with 182,829 and 225,828 unique records in manuscripts and patents, respectively.