Analysis of *Nature* news reveals gender and regional disparities in scientific coverage

This manuscript (permalink) was automatically generated from greenelab/nature_news_manuscript@4e652d8 on June 12, 2021.

Authors

- **Natalie R. Davidson**
  0000-0002-1745-8072 · nrosed · n_rose_d
  University of Colorado School of Medicine, Aurora, Colorado, United States of America; Center for Health AI · Funded by Grant XXXXXXXX

- **Casey S. Greene**
  0000-0001-8713-9213 · cgreene · GreeneScientist
  University of Colorado School of Medicine, Aurora, Colorado, United States of America; Center for Health AI · Funded by The Gordon and Betty Moore Foundation (GBMF 4552)

Abstract

Scientific news coverage shapes the public's view of the current state of scientific findings and legitimizes experts. Through researching a story, journalists identify and interview a limited number of sources. These sources may come from a journalist's research or through recommendations by other scientists. In either case, unconscious biases may influence who is identified as an expert to interview, possibly skewing the selection of interviewees. We analyzed more than 16,000 news articles published by *Nature* to quantify possible disparities. Our analysis considered three possible sources of bias: gender, name origin, and country affiliation. To explore these sources of bias, we extracted cited authors' names and affiliations, as well as extracted names of quoted speakers. We then used the names to predict gender and name origin of the authors and speakers. In our analysis, we found a bias towards male quotation, but quotation is trending toward equal representation at a faster rate than academic publishing. Interestingly, we found that the gender disparity in quotes was column-dependent, with the "career-features" column reaching gender parity. Our name origin analysis found a significant over-representation of names with Celtic/English origin and under-representation of names with an East Asian origin. This finding was observed both in extracted quotes and citations, but dampened in citations. Finally, we performed an analysis to identify how countries vary in the way that they're described in the news. We found a set of countries that are typically mentioned in the text of the article, but whose academic output is not highly cited, and their counterpart, a set of countries that are highly cited, but not commonly mentioned. We found that the articles in which the less cited countries occur tend to have more agricultural terms, whereas articles including highly cited countries have broader scientific terms. This discrepancy indicates a possible lack of regional diversity in the reporting of scientific output.

Introduction

News coverage of science shapes who both peers and the public consider a scientist and field expert. This indication of legitimacy can either help recognize persons who are typically overlooked due to systemic biases or intensify biases. Journalistic biases have been observed by journalists themselves [1,2,3,4], as well as by independent researchers [5,6,7,8,9,10]. Researchers found a gap between male and female speakers or quotes, with independent studies finding that between 17-40% of total subjects were female across multiple news outlets between 1985 and 2015 [5,6,10]. One study found 27-35% of total subjects in science-related news were female between 1995 and 2010 [10]. However, news coverage is not the only source of bias. Both gender and racial disparities already exist in science as observed in differences in citation [11,12], funding [13,14,15,16], and publication rates [17,18,19].

Therefore, it is crucial to ensure that science coverage does not solely focus on a few well-known scientists but expands our shared view of an expert scientist. One may believe that science coverage would simply reflect the most current and groundbreaking findings. Still, there are many ways gender, racial, or regional biases can unknowingly seep into coverage. In researching a story, a journalist will typically interview multiple scientists for their opinion, potentially asking for additional sources, allowing individual unconscious biases to skew scientific coverage broadly. In addition, the repeated selection of a small set of field experts or the approach a journalist takes in establishing a new source may intensify existing biases [3,4,20].

While these biases may go unnoticed by an individual, analyzing a large corpus of articles can identify and quantify these biases and help guide institutional and individual self-reflection. In the same vein as previous media studies, we seek to quantify gender and regional biases of news coverage. Our study focuses solely on scientific news content, specifically news content published by *Nature*. Since *Nature* also publishes research articles, this provides a natural estimated background rate for comparison. Our goal is to identify quoted and cited scientists by analyzing the content and citations within all news articles from 2005 to 2020. We further analyze if the coverage is biased beyond the current state of academic publishing by analyzing the authorship statistics across all *Nature* research articles across the same period.

Through our analysis of 22,001 news-related articles, we were able to identify >100,000 quotes and >8,000 citations with sufficient speaker or author information within the news content. We then identified possible gender or regional biases using the extracted names. We used computational methods to predict gender and identified a bias towards male quotes in news articles. However, during the period that we examine the bias has decreased from being more extreme than in the research content of *Nature* to less extreme. Furthermore, we identified that the speaker bias was dependent on article type; the "Career-Feature" column achieved gender parity in quoted speakers. We also used computational methods to predict name origins and found a significant over-representation of names with Celtic/English origin and under-representation of names with an East Asian origin in both quotes and citations.

While we focused on scientific news coverage from *Nature*, our software can be repurposed to analyze other news text. We hope that news publishers will welcome bias-audting systems to help identify journalistic blind spots. However, auditing is only part of the solution; journalists and source recommenders must also change their source gathering patterns. To help change these patterns, there exist guides [20], databases [21], and affinity groups [20] that can help us all expand our vision of who can be a field expert.

Methods

Data Acquisition and Processing

Text Scraping

We scraped all text and metadata using the web-crawling framework Scrapy [23] (version 2.4.1). We created three independent scrapy web spiders to process the news text, news citations, and research article metadata. News articles were defined as all articles from 2005 to 2020 that were designated as "News", "News-Feature", "Career-Feature", "Technology-Feature", and "Toolbox". Using the spider "target_year_crawl.py", we scraped the title, author, and main text from all news articles. We character normalized the main text by mapping visually identical Unicode codepoints to a single Unicode codepoint and stripping all non-Unicode characters. Using an additional spider defined in "doi_crawl.py", we scaped all citations within news articles. For simplicity, we only considered citations with a DOI included in either text or a hyperlink in this spider. Other possible forms of citations, e.g., titles, were not included. The DOIs were then queried using the *Springer* API. The spider "article_author_crawl.py" scraped all articles designated "Article" or "Letters" from all possible research articles. We only scraped author names, author positions, and associated affiliations from research articles. It should be noted that news article designations changed over time.

coreNLP

After news articles were scraped and processed, the text was processed using the coreNLP pipeline [24] (version 4.2.0). The main purpose for using coreNLP was to identify named entities related to countries and quoted speakers. The full set of annotaters were: tokenize, ssplit, pos, lemma,ner, parse, coref, quote. We used the "statistical" algorithm to perform coreference resolution. All results were output to json format for further downstream processing.

Springer API

Springer was chosen over other publishers for multiple reasons: 1) it is a large publisher, second only to Elsevier; 2) it covers multiple subjects, in contrast to PubMed; 3) its API has a large daily query limit (5000/day); and 4) it provided more author affiliation information than found in Elsevier. We generated a comparative background set for supplemental analysis with the *Springer* API by obtaining author information for research articles cited in the news. We selected a random set of articles to generate the *Springer* background set. These articles were the first 200 English language "Journal" articles returned by the *Springer* API for each month, resulting in 2400 articles per year for 2005 through 2020. To get the author information for the cited articles, we queried the *Springer* API using the scraped DOI. For both API query types, the author names, positions, and affiliations for each publication were stored and are available in "all_author_country.tsv" and "all_author_fullname.tsv".

Name Formatting

Name Formatting for Gender Prediction in Quotes or Mentions

To identify the gender of a quoted or mentioned person, we first attempt to identify the person's full name. Even though genderizeR only uses the first name to make the gender prediction, identifying the full name gives us greater confidence that we are using the first name. To identify the full name, we take the predicted speaker by coreNLP and match it to the longest matching name within the same article. We match names by finding the longest mentioned name in the article with minimal edit (Levenshtein) distance. The name with the smallest edit distance, where character deletions have zero cost, is defined as the matching name. Character deletion was assigned a zero cost because we would like exact substring matches. For example, the calculated cost, including a cost for character deletion, between John and John Steinberg is 10; without character deletion, it is 0. Compared with the distance between John and Jane Doe, with character deletion cost, it is 7; without it is 2. If we are still unable to find a full name, or if coreNLP cannot identify a speaker at all, we also determine whether or not coreNLP linked a gendered pronoun to the quote. If so, we assign the gender of the pronoun to the speaker.

1. **The enduring whiteness of the American media | Howard French**
   the Guardian
   (2016-05-25) http://www.theguardian.com/world/2016/may/25/enduring-whiteness-of-american-journalism

2. **I Analyzed a Year of My Reporting for Gender Bias and This Is What I Found**
   Adrienne LaFrance
   Medium (2013-09-30) https://medium.com/ladybits-on-medium/i-analyzed-a-year-of-my-reporting-for-gender-bias-and-this-is-what-i-found-a16c31e1cdf

3. **I Analyzed a Year of My Reporting for Gender Bias (Again)**
   Adrienne LaFrance
   The Atlantic (2016-02-17) https://www.theatlantic.com/technology/archive/2016/02/gender-diversity-journalism/463023/

4. **I Spent Two Years Trying to Fix the Gender Imbalance in My Stories**
   Ed Yong
   The Atlantic (2018-02-06) https://www.theatlantic.com/science/archive/2018/02/i-spent-two-years-trying-to-fix-the-gender-imbalance-in-my-stories/552404/

5. **A Paper Ceiling**
   Eran Shor, Arnout van de Rijt, Alex Miltsov, Vivek Kulkarni, Steven Skiena
   American Sociological Review (2015-09-30) https://doi.org/f7tzps
   DOI: 10.1177/0003122415596999

6. **Time Trends in Printed News Coverage of Female Subjects, 1880–2008**
Eran Shor, Arnout van de Rijt, Charles Ward, Aharon Blank-Gomel, Steven Skiena
Journalism Studies (2013-09-12) https://doi.org/gj3z8b
DOI: 10.1080/1461670x.2013.834149

7. **Women and news: A long and winding road**
Karen Ross, Cynthia Carter
Media, Culture & Society (2011-11-22) https://doi.org/ccxhvz
DOI: 10.1177/0163443711418272

8. **Women Are Seen More than Heard in Online Newspapers**
Sen Jia, Thomas Lansdall-Welfare, Saatviga Sudhahar, Cynthia Carter, Nello Cristianini
PLOS ONE (2016-02-03) https://doi.org/f8q47g
DOI: 10.1371/journal.pone.0148434 · PMID: 26840432 · PMCID: PMC4740422

9. **Lack of female sources in NY Times front-page stories highlights need for change**
Poynter
(2013-07-16) https://www.poynter.org/reporting-editing/2013/lack-of-female-sources-in-new-york-times-stories-spotlights-need-for-change/

10. **Who Makes the News | GMMP 2015 Reports** https://whomakesthenews.org/gmmp-2015-reports/

11. **Men Set Their Own Cites High: Gender and Self-citation across Fields and over Time**
Molly M. King, Carl T. Bergstrom, Shelley J. Correll, Jennifer Jacquet, Jevin D. West
Socius: Sociological Research for a Dynamic World (2017-12-08) https://doi.org/ddzq
DOI: 10.1177/2378023117738903

12. **Bibliometrics: Global gender disparities in science**
Vincent Larivière, Chaoqun Ni, Yves Gingras, Blaise Cronin, Cassidy R. Sugimoto
Nature (2013-12-11) https://doi.org/qgf
DOI: 10.1038/504211a · PMID: 24350369

13. **Fund Black scientists**
Kelly R. Stevens, Kristyn S. Masters, P. I. Imoukhuede, Karmella A. Haynes, Lori A. Setton, Elizabeth Cosgriff-Hernandez, Muyinatu A. Lediju Bell, Padmini Rangamani, Shelly E. Sakiyama-Elbert, Stacey D. Finley, … Omolola Eniola-Adefeso
Cell (2021-02) https://doi.org/ghvqv5
DOI: 10.1016/j.cell.2021.01.011 · PMID: 33503447

14. **NIH peer review: Criterion scores completely account for racial disparities in overall impact scores**
Elena A. Erosheva, Sheridan Grant, Mei-Ching Chen, Mark D. Lindner, Richard K. Nakamura, Carole J. Lee
Science Advances (2020-06-03) https://doi.org/gjnjbz
DOI: 10.1126/sciadv.aaz4868 · PMID: 32537494 · PMCID: PMC7269672

15. **Topic choice contributes to the lower rate of NIH awards to African-American/black scientists**
Travis A. Hoppe, Aviva Litovitz, Kristine A. Willis, Rebecca A. Meseroll, Matthew J. Perkins, B. Ian Hutchins, Alison F. Davis, Michael S. Lauer, Hannah A. Valantine, James M. Anderson, George M. Santangelo

Science Advances (2019-10-09) https://doi.org/gghp8t
DOI: 10.1126/sciadv.aaw7238 · PMID: 31633016 · PMCID: PMC6785250

16. **SOCIOLOGY: The Gender Gap in NIH Grant Applications**
T. J. Ley, B. H. Hamilton
Science (2008-12-05) https://doi.org/frdj6k
DOI: 10.1126/science.1165878 · PMID: 19056961

17. **Disparities in publication patterns by gender, race and ethnicity based on a survey of a random sample of authors**
Allison L. Hopkins, James W. Jawitz, Christopher McCarty, Alex Goldman, Nandita B. Basu
Scientometrics (2012-11-10) https://doi.org/gffmpv
DOI: 10.1007/s11192-012-0893-4

18. **The Diversity–Innovation Paradox in Science**
Bas Hofstra, Vivek V. Kulkarni, Sebastian Munoz-Najar Galvez, Bryan He, Dan Jurafsky, Daniel A. McFarland
Proceedings of the National Academy of Sciences (2020-04-28) https://doi.org/ggskr7
DOI: 10.1073/pnas.1915378117 · PMID: 32291335 · PMCID: PMC7196824

19. **Last Place? The Intersection of Ethnicity, Gender, and Race in Biomedical Authorship**
Gerald Marschke, Allison Nunez, Bruce A. Weinberg, Huifeng Yu
AEA Papers and Proceedings (2018-05-01) https://doi.org/gjg9k8
DOI: 10.1257/pandp.20181111 · PMID: 30197432 · PMCID: PMC6124503

20. **Including Diverse Voices in Science Stories**
Christina Selby
The Open Notebook (2016-08-23) https://www.theopennotebook.com/2016/08/23/including-diverse-voices-in-science-stories/

21. **gage. Discover Brilliance** https://gage.500womenscientists.org/

22. **WMC SheSource - Women's Media Center** https://www.womensmediacenter.com/shesource

23. **Scrapy | A Fast and Powerful Scraping and Web Crawling Framework** https://scrapy.org/

24. **The Stanford CoreNLP Natural Language Processing Toolkit**
Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, David McClosky
Association for Computational Linguistics (ACL) (2014) https://doi.org/gf3xhp
DOI: 10.3115/v1/p14-5010