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

This manuscript (permalink) was automatically generated from greenelab/nature_news_manuscript@9ba72ff on June 4, 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 Grant XXXXXXXX

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 22,000 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 22,094 news-related articles, we were able to identify >120,000 quotes and >10,000 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-and-Views", "News-Feature", "Career-Column", "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 Normalization

Names were cast to lowercase and processed using the R package humaniformat [25]. humaniformat identifies if names are reversed (Lastname, Firstname), as well as identifies middle names. Since many last or first authorships may be non-names, we additionally filtered out any identified names if they partially or fully match any of the following terms: "consortium", "group", "initiative", "team", "collab", "committee", "center", "program", "author", or "institute". Furthermore, since many articles only contain first and last name initials, we remove any names less than four letters with a "." or "-". Finally, we only consider any remaining names of more than two characters. We do this additional filtering to remove any web-scraping or coreNLP errors that only return partial names.

Since gender and origin predictions require the speaker's full name, when possible, a longer name with minimal edit (Levenshtein) distance mentioned within the same article replaces the coreNLP predicted speaker. 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. Furthermore, the identification of simultaneously cited and quoted speakers used this same name matching strategy.

Gender Analysis

The quote extraction and attribution annotator from the coreNLP pipeline was employed to identify quotes and their associated speakers in the news article text. In some cases, coreNLP could not identify an associated speaker's name but instead assigned a gendered pronoun. In these instances, we used the gender of the pronoun for the analysis. The R package genderizeR [26], a wrapper for the genderize.io API [27], predicted the gender of authors and speakers. We predicted a name as male using the first name with a minimum cutoff of 50%. To reduce the number of queries made to genderize.io, a previously cached gender prediction from [28] was also used and can be found in the file "genderize.tsv". All first name predictions from this analysis are in the file "genderize_update.tsv". To estimate the gender gap for the quote gender analyses, we used the proportion of total quotes, not quoted speakers. We used the proportion of quotes to measure speaker participation instead of only the diversity of speakers. The specific formulas for a single year are shown in equations 1 and 2. We did not consider any names where no prediction could be made or quotes where neither speaker nor gendered pronoun was associated.

$$\text{Prop. Male Quotes} = \frac{|\text{Male Speaker Quotes}|}{|\text{Male or Female Speaker Quotes}|} \quad (1)$$

$$\text{Prop. Male First Authors} = \frac{|\text{Male First Authors}|}{|\text{Male or Female First Authors}|} \quad (2)$$

Name Origin Analysis

We used the same quoted speakers as described in the previous section for the name origin analysis. In addition, we also take all authors cited in a *Nature* news article. In contrast to the gender prediction, we need to use the full name to predict name origin. We submitted all extracted full names to Wiki-2019LSTM [28] to predict one of ten possible name origins: African, CelticEnglish, EastAsian, European, Greek, Hispanic, Jewish, ArabTurkPers, Nordic, and SouthAsia. We set the highest probability origin as the resultant assignment. Similar to the gender analyses, quote proportions were again directly compared against publication rates. For citations, we calculated the proportion of overall news citations for a given year for each name origin, as exemplified in 3 to calculate the citation rate for first authors with a Greek name origin.

$$\text{Prop. Greek } 1^{\text{st}} \text{ Author Citations} = \frac{|\text{Cited } 1^{\text{st}} \text{ Authors w/Greek Name Orig}|}{|\text{Cited } 1^{\text{st}} \text{ Authors w/any Name Orig}|} \quad (3)$$

Country Mention Proportions

We estimated the prevalence of a country's mentions by including all identified organizations, countries, states, or provinces from coreNLP's named entity annotater. We queried the resultant terms using OpenStreetMap [29] to identify the associated country with the term. All terms that were identified in the text 25 or more times were visually inspected for correctness. Hand-edited entries are denoted in the OpenStreetMap cache file "osm_cache.tsv" by the column "hand_edited". Still, this only accounts for less than 5% of the total entries. Furthermore, country-associated terms identified by coreNLP may be ambiguous, causing OpenStreetMap to return incorrect locations. Therefore, we count country mentions only if we find at least two unique country-associated terms in an article. We calculate the mentioned rate as the proportion of country-specific mentions divided by the total articles for a particular year, as exemplified in 4 for calculating the mentioned rate for Mexico.

$$\text{Prop. Mexico Mentions} = \frac{|\text{Articles with} \geq 2 \text{ unique Mexico-related terms}|}{|\text{All News Articles}|} \quad (4)$$

Country Citation Proportions

To identify the citation rate of a particular country, we processed all authors' affiliations for a specific article. Since the affiliations could be in multiple formats, we again used OpenStreetMap to identify the country affiliation. Additionally, we considered all affiliations for a single author. We calculated a countries' citation rate as the number of citations for a country divided by either the number of *Nature* research articles (5) or the total number of papers cited by news articles for that year (6). Shown below are example calculations for Colombia for a single year.

$$\text{Prop. CO Affil. in Nature} = \frac{|\text{Articles with} \geq 1 \text{ CO affil. in Nature}|}{|\text{All Nature Research Articles}|} \quad (5)$$

$$\text{Prop. CO Affil. Citations} = \frac{|\text{Cited Articles in News with} \geq 1 \text{ CO affil.}|}{|\text{All Articles Cited in News}|} \quad (6)$$

Divergent Word Identification

After calculating the citation and mention proportion for each country, we identified countries outlying in their comparative citation or mention rate. Outlier detection was done by subtracting the citation and mention rates, then identifying which countries were in the top or bottom 5% from each year. We only considered countries identified as either high citation (Set C) or high mention (Set M) across all years. We did not consider any country that was in the top and bottom 5% in different years. Additionally, we only considered a country if cited or mentioned five times in a single year. Once we identified set C/M countries, we analyzed the word frequencies in all news articles where the set C/M country was mentioned but not cited. We believe this would provide insight into content differences between set C/M countries. Text from news articles in 2020 were not considered due to an excess of SARS-CoV-2 related terms. Using the R package tidytext [30] we extracted tokens, removed stop words, and calculated the token frequencies across all articles. We only consider tokens in set C/M articles if the token has been observed at least 100 times across all articles. We then identify tokens that have the most significant ratio of usage between the two sets. Since there are differences in the number of articles per country within each set, we calculated a token frequency within a set as the median frequency within each countries associated articles. We calculated the resultant token ratio as the country normalized citation frequency to the country normalized mention frequency. We assert that the term must be observed at least once in each set.

Bootstrap Estimations

For all analyses related to equations 1 - 6, we independently selected 5000 bootstrap samples for each year. We sampled with replacement of size equal to the cardinality of the complete set of interest. Bootstrap estimates for equations 1 - 6 were performed by sampling the denominator set. The 5th, 50th, and 95th quantiles across the estimates are reported as the lower, middle, and upper bounds. For the divergent word analysis, due to computational constraints, we only took 1000 bootstrap samples. The bootstrap estimates were taken by subsampling the news articles with replacement, each time recalculating the country-normalized token frequencies within each country set (C and M). After the normalized frequencies within each country set were calculated, we calculated the ratio between country sets for each subsample with a pseudocount of 1 in the numerator and denominator, (C+1)/(M+1). Again, the 5th, 50th, and 95th quantiles across the estimates are reported as the lower, middle, and upper bounds.

Results

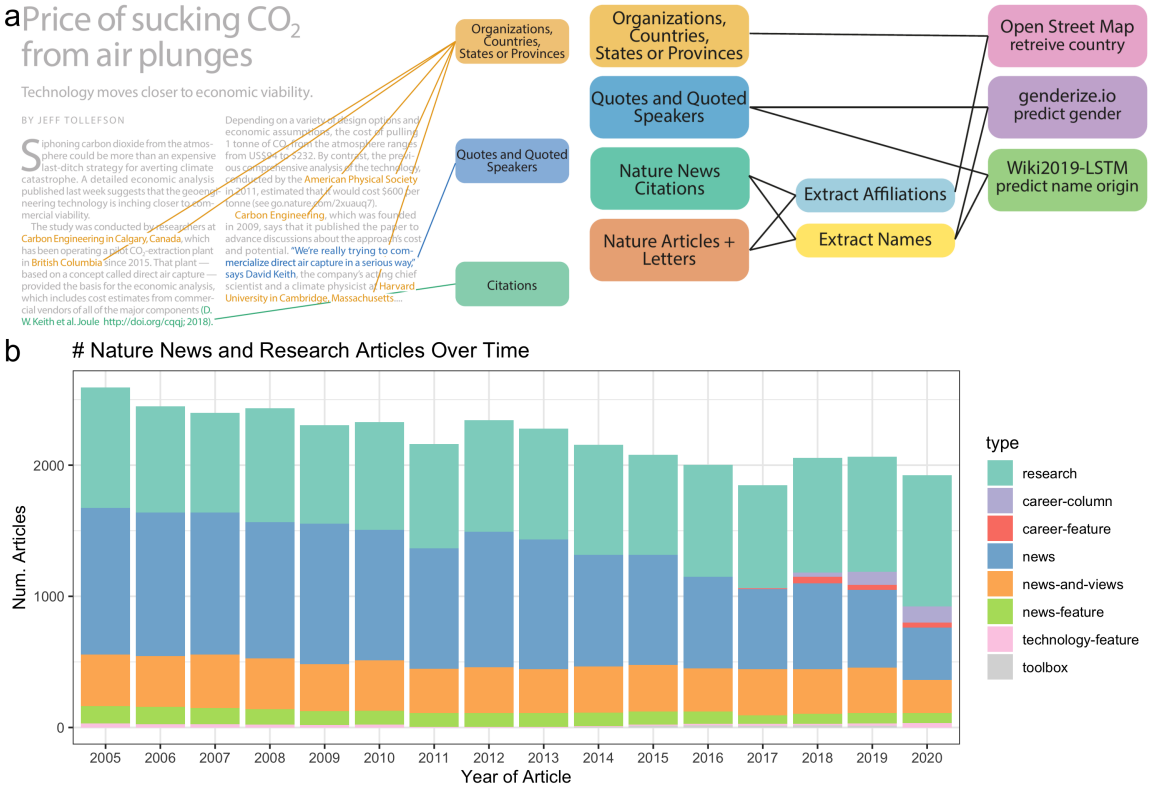

Figure 1: Data and Processing Pipeline Overview Panel A, left, depicts an example news article and the type of data extracted from the text. Orange highlighted text depicts all named entities identified as either an organization, country, state, or province by the coreNLP pipeline. The coreNLP pipeline also extracts all quotes and associated speakers. A custom script described in section Methods identifies all citations. Panel A, right, charts the analyses done on the extracted names and locations from the news, research articles, and letters published by *Nature*. Panel B shows the types and amounts of articles that we have used for analyses.

Creation of an Annotated News Dataset

We have analyzed the text and citations of 22,094 news-related articles hosted on "www.nature.com". The articles span 15 years from 2005 to 2020, and seven article types: Career Column, Career Feature, News, News and Views, News Feature, Technology Feature, and Toolbox. The text and citations were then uniformly processed as depicted in Figure 1a. We scraped the text using the web-crawling framework Scrapy [23], processed, and run through the coreNLP pipeline (Methods). From the coreNLP pipeline results, we wanted to identify two distinct data types to analyze, country mentions and quotes with their associated speakers. To identify country mentions, we used the following named entities as possible mentions: "organizations", "countries", "states or provinces". We then mapped the named entity to a country prediction using OpenStreetMap [29]. To identify quotes and speakers, we used the coreNLP quote extraction and attribution annotator. We performed name normalization (Methods) to identify the speaker's full name for gender and name origin prediction. We scraped the citations using an independent scraper to the text scraper. All identified DOI's were queried using the *Springer* API to attain author names, positions, and affiliations.

We used the author names, positions, and affiliations of Research articles and Letters published by *Nature* to perform comparative analyses. Since the news articles of current science, the quoted speakers, mentioned countries, and cited authors should have a similar demographic makeup to contemporaneous *Nature* authorship. The author metadata of research-related articles was analyzed over the same period as the news-related articles and totaled 13,414 articles. Since the demographic makeup of authors published in *Nature* may not represent overall science authorship, we also analyzed 36,000 randomly selected *Springer*-published articles from English language journals over the same time. The extracted author affiliations from both data sources were mapped to a country

using OpenStreetMap. Similarly, author names were uniformly processed and then used to predict both gender and name origin.

When considering all *Nature* articles, news and research, across all time, we find significant variability in the number of articles by type (Figure 1b). The top three observed article frequencies are research (including letters and research), news, and news-and-views. Since *Nature* merged letters and research articles in 2019, we combine them in our analysis. The changing classification of article types may also explain temporal changes in news articles. Over time, the frequency of news-related articles is decreasing; however, more specific article types are increasing. We observe this in the introduction of two new categories, “Career-column” and “Career-feature” in 2017. These articles were previously identified as “Features” articles before 2017 (https://www.nature.com/nature/articles?type=feature).

Quoted Speakers are More Often Male

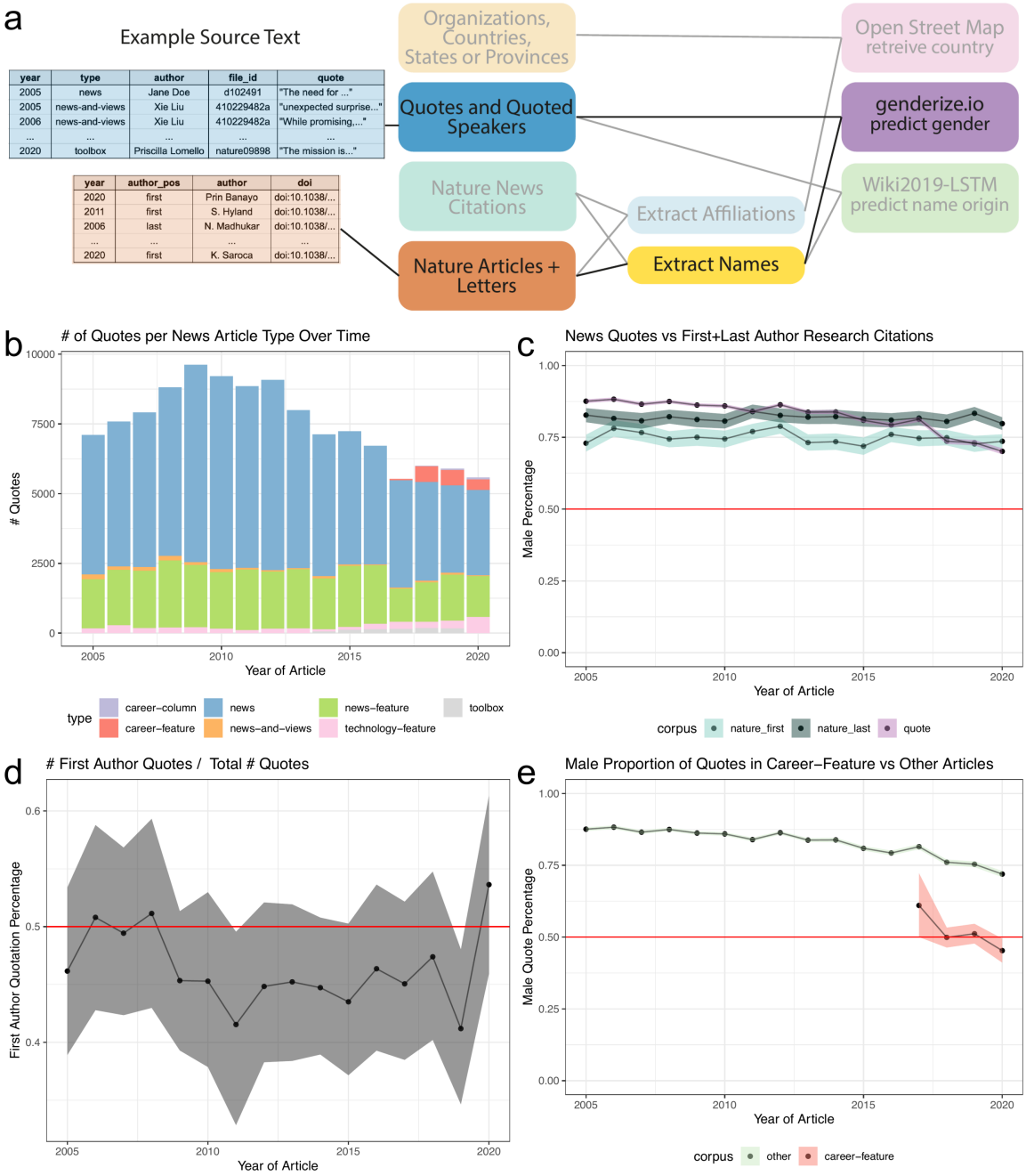

Figure 2: Male bias is observed in news quotes but is dependent on article type Panel A, left, depicts an example of the names extracted from quoted speakers in news articles and authors in research articles. Panel A, right, highlighted the data types and processes used to analyze the gender of extracted names. Panel B shows an overview of the number of quotes extracted for each article type. Panel C depicts three trend lines: Purple: Proportion of quotes for

an estimated male speaker; Light Blue: Proportion of first author articles from an estimated male author; Dark Blue: Proportion of last author articles from an estimated male author. We observe that the proportion of estimated male quotes is steadily decreasing, most notably from 2017 onward. This decreasing trend is not due to a change in quotes from the first or last authors, as observed in Panel D. Panel D shows a consistent but slight bias towards quoting the last author of a cited article than the first author. Instead, the observed downward trend of male quotes coincides with additional article types introduced in 2017. Panel E depicts the frequency of quote by article type highlighting an increase in quotes from "Career-Feature" articles. Panel E depicts that the quotes obtained in this article type have reached parity. The colored bands represent a 95% confidence interval in all plots, and the point is the median calculated from 5,000 bootstrap samples.

To quantify the amount of gender bias in quotes, we analyzed the names of quoted speakers and compared them against the first and last authors of research articles. While we could have analyzed the proportion of all quotes from a male speaker, we were interested in measuring the overall participation rates by gender. Figure 2 shows an overview of the process and example input data for this analysis. This analysis relies upon accurate gender prediction of both authors and speakers. To predict the gender of the speaker or author, we used the package genderizeR [26], an R package wrapper to access the genderize.io API [27] to get binary gender predictions for each identified first name. We, unfortunately, cannot identify non-binary gender expression with the tools we used. To identify our binary gender prediction error rate, we created a benchmark data set of thirty randomly selected news articles; ten from each of the following years: 2005, 2010, 2015. In these articles, we hand-annotated the quoted speakers and their assumed gender. We then compared our hand-annotated gender prediction against the gender prediction after the entire scraping, coreNLP quote annotator, and genderizeR pipeline. Using our hand-annotated prediction as ground truth, we correctly predicted a total of 42/55 female and 157/162 male names within the benchmark articles. We found that the gender prediction had an accuracy of 0.914, with a Kappa statistic of 0.767 (Figure Supplemental 1a). After establishing a reasonable accuracy for gender prediction, we determined the rate of gendered quotes within articles. We first examined the number of quotes identified within each news-related article (Figure 2b), totaling 121,684 quotes with 120,309 of them containing a gender prediction for the speaker. Quote frequencies vary by article type. The only article types in which there is a large discrepancy between the number of quotes and articles are "news-and-views" and "career-column." This discrepancy is expected in "news-and-views" since this article type typically is an overview of a specific scientific topic and rarely quotes authors. Similarly, "career-column" is a perspective piece from a scientist who also rarely quotes other persons.

We then compared the number of male quotes to the number of male first and last authors published in *Nature* (Figure 2c). We analyzed a total of 10,466 first authors and 10,494 last authors with a gender prediction. As denoted by the red line, we find that both authorship and quotes are far from gender parity. Additionally, we find a difference in author genders between first and last authors, with the last authors being more male-dominated. Since *Nature* authorship may not represent scientific publishing as a whole, we also compared against a random selection of authors from English language journals published by *Springer* (Figure Supplemental 2a). We observed a larger gender gap between first and last authors in our selection of *Springer* articles; however, both first and last authors are much closer to parity than *Nature* authors.

In contrast to the relatively stable gender proportions of authorship, we find the rate of male quotes significantly decreases over time. The male quote rate in 2005 is 87.6% (884/7103) and in 2020 is 70.1% (3911/5578). From 2005-2010, we find the male quote rate is higher than the male last authorship rate, then slowly decreasing until it matches the male first authorship rate in 2018 and continues to trend downwards. We then explored the possible reasons for this decrease. First, we looked at changes in the rate of authorship position by quoted speakers (Figure 2d). We were able to identify 10,167 quotes with an associated citation (4630 first author and 5537 last author quotes). We found that quotes are slightly biased towards the last authors. However, this rate does not change over time, and we do not believe this drives the downward trend. We then analyzed the breakdown of gendered quotes by article type. Interestingly, we found one article type, "career-feature" that has achieved gender parity in its quotes (Figure 2e and Figure Supplemental 2b). Since the introduction of

this article, it has identified a total of 1,550 quotes (781 female and 769 male quotes) and has substantially pulled the overall quote gender ratio much closer to parity from 2018 onward.

Celtic English Names are over enriched in cited and quoted persons, while East Asian Names are under enriched

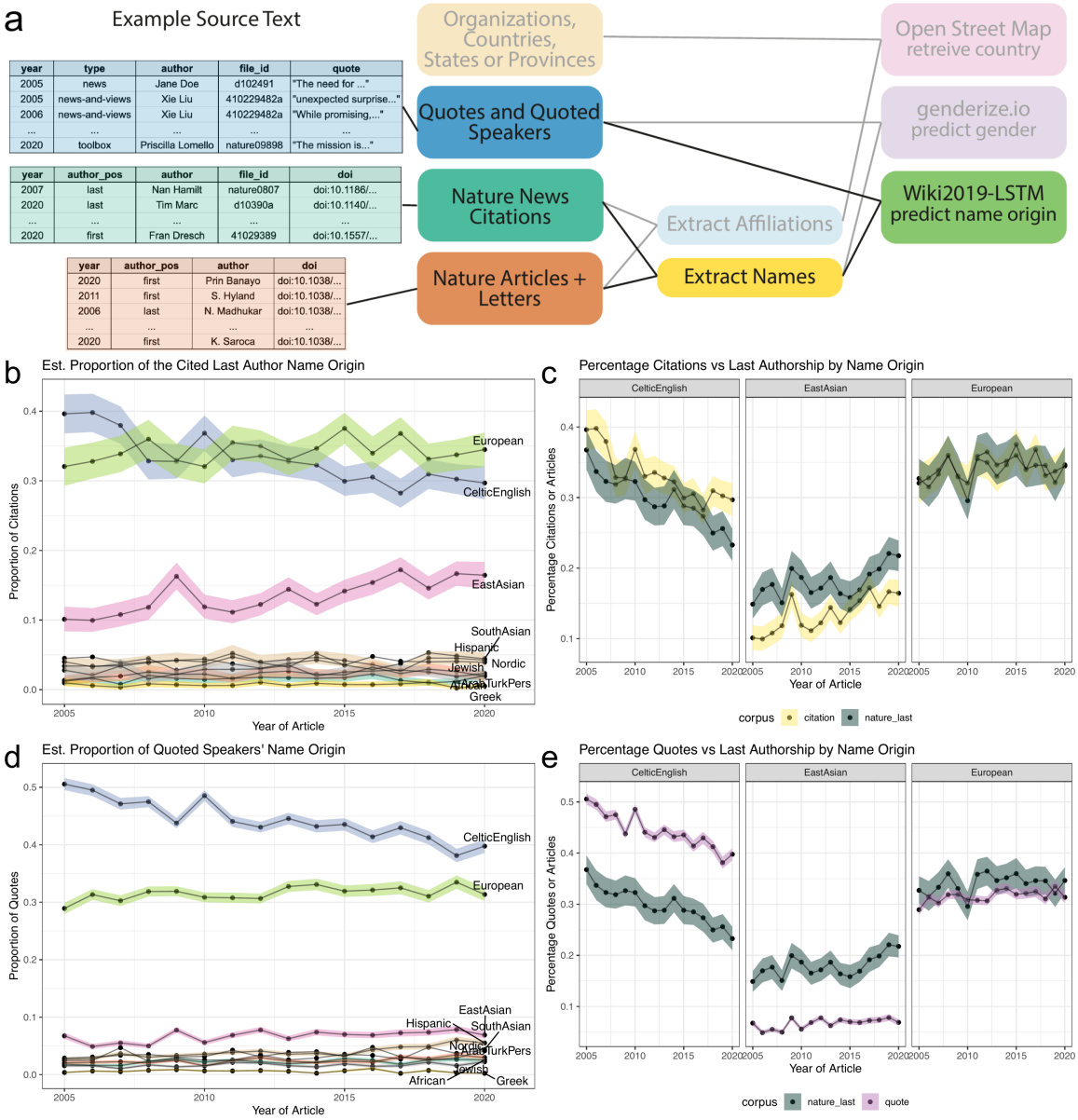

Figure 3: Analysis of Quotes and Citations found Overrepresentation of Celtic/English and underrepresentation of East Asian names Panel A, left, depicts an example of the names extracted from quoted speakers and citations found within news articles and authors in research articles. Panel A, right, highlights the data types and processes used to analyze the origin of extracted names. Panels B and D depicts nine trend lines, each representing a possible name origin. Panels C and E depict two possible trend lines, comparing either cited last authors or quoted speakers against last authors of *Nature* research articles. We observed that the overrepresentation of Celtic/English names is present both in citations and quotes. Furthermore, the overrepresentation of East Asian names is also present both in citations and quotes. We recapitulate this finding when comparing against an additional background of *Springer* published last authors (Figure Supplemental 3)

To identify possible bias in name origin, we again used the extracted quoted speakers and last authors published in *Nature*. In addition, we also identified the last authors of all articles cited by a news-related article. All processed names were then inputted into Wiki2019-LSTM and used to assign one of nine possible name origins (Methods). Figure 3a displays a schematic of the data processing and example names used in this analysis. We divided our analysis into two parts: firstly, looking at the number of cited last authors with a specific name origin overall cited last authors in a particular year.

Secondly, looking at the proportion of quotes from a speaker with a specific name origin overall quotes in a particular year. We used the proportion of last authors in *Nature* with a specific name origin overall *Nature* articles in a particular year for both analyses. Additionally, in our supplemental analyses, we compare against the last authorship in a random selection of *Springer* articles. We find that the number of quotes dramatically outnumbers the number of authors and citations (Figure Supplemental 3a). Still, we believe that the total number of observations per data type across all years is sufficient for our analysis. Minimum and median per data type: *Nature* articles, (682, 788); *Springer* articles, (1630, 1904); quotes, (4626, 6229); citations, (823, 1070).

In comparing the citation rate of last author name origins in news articles, we find most names are Celtic/English or European, both with a bootstrapped estimated citation rate between 26.2-42.5% (Figure 3b). East Asian names are the third highest proportion of cited names, with a bootstrapped estimated citation rate between 8.3-19.0%. All other name origins individually account for less than 7% of total cited authors. To compare against the composition of the last authors in *Nature*, we compare the top three most frequent name origins (Figure 3c). We find a slight over-enrichment for Celtic/English names and a small under-enrichment for East Asian names. Furthermore, we find no significant difference for European or other name origins (Figure 3c, Figure Supplemental 3b). We also observed the Celtic/English over-enrichment and East Asian under-enrichment when considering our subset of *Springer* articles (Figure Supplemental 3c). In contrast to *Nature*, in our *Springer* articles, we see a significant difference in European names, with a growing over-enrichment. Additionally, we see a significant difference in Arabic/Turkish/Persian, Hispanic, Jewish, and South Asian name frequencies between cited authors and *Springer* authors, however, the difference is lower than observed for Celtic/English and East Asian.

We then sought to see if quoted speaker replicated the cited authors' over- and under-enrichment patterns. We found a much stronger Celtic/English over-enrichment, with quotes from Celtic/English speakers at a much higher frequency than quotes from European speakers (Figure 3c). Additionally, we also found a much stronger under-enrichment of quotes from East Asian speakers (Figure 3c), with never more than 7.8% of quotes even though they constitute between 8.3-19.0% of cited authors. When we again compare against last authorship in *Nature*, we find the same patterns observed in the previously described citation analysis with all name origins except for East Asian and Celtic/English closely matching the background rate. Similarly, we find the same patterns in quoted speakers when comparing against the *Springer* background as we did in the previous citation analysis (Figure Supplemental 3e).

Content of Science Coverage Differs between Countries

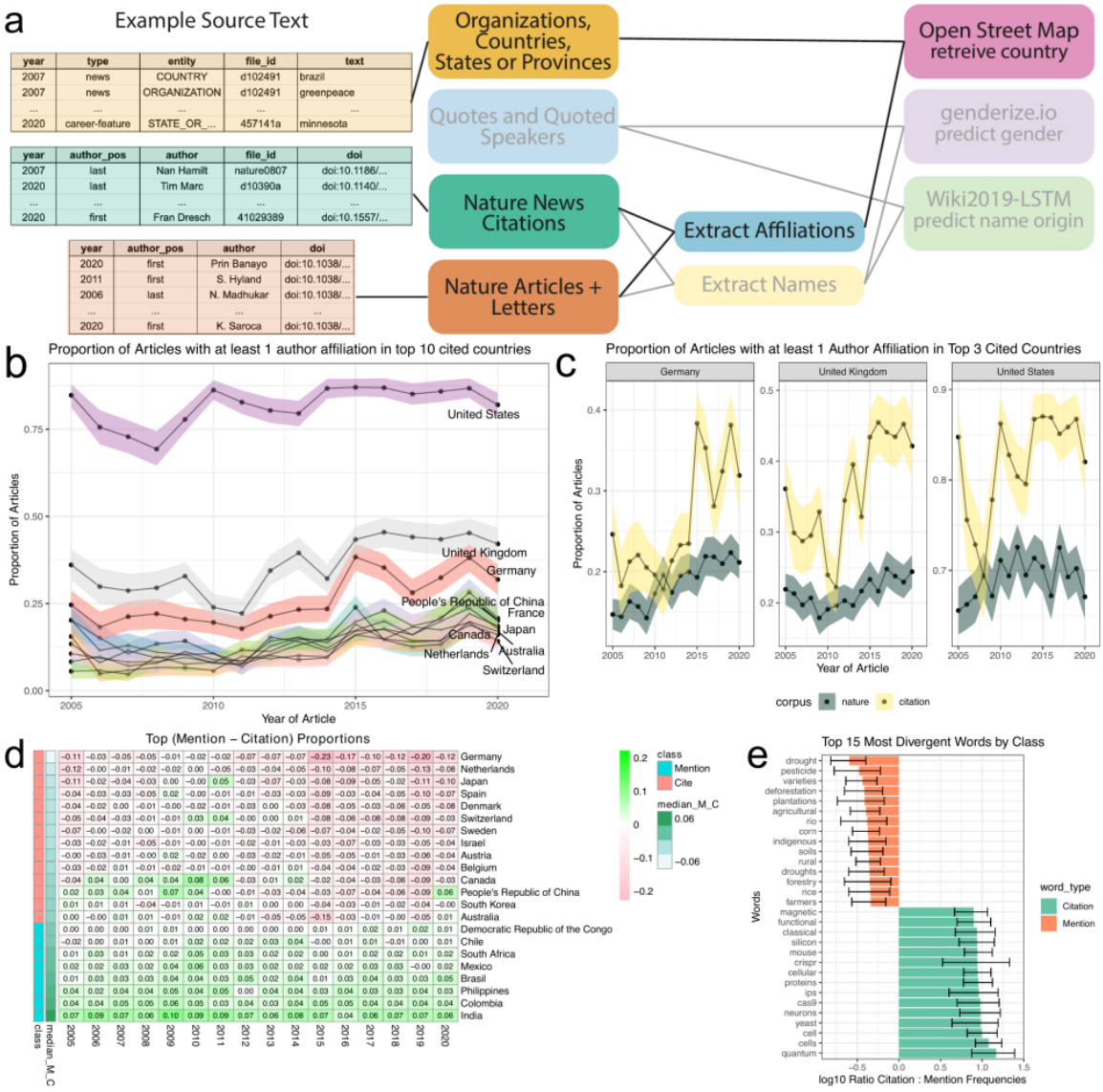

Figure 4: Type of country representation in news articles differs depending on whether the or not country itself is the subject Panel A, left, depicts an example of the country mentions extracted from the news article text and citations found within news articles and author affiliations in research articles. Panel A, right, highlights the data types and processes used to analyze the countries cited or mentioned. Panel B depicts the citation rate of the top ten most-cited countries over time. Panel C depicts the citation rate of the top three most-cited countries (yellow) compared to that countries citation rate, as measured by author affiliation (grey). Panel D is a heatmap depicting the yearly difference in citation and mention rate for a specific country. We only depict countries with a consistent and significant difference across all years. Each cell contains the difference between citation and mention rates, with red denoting the lower difference between mention and citation and green a more considerable difference. The left annotation bar titled "median_M_C" is the median difference across all observed years. The "Class" annotation column denotes the binarized set definition of each country, either "Cited" or "Mentioned". Panel E shows the top 15 words extracted from articles mentioning the countries depicted in Panel D, with the most significant proportional frequency between the two defined country sets. The width of the bar depicts the log10(Frequency in Mentions / Frequency of Citations).

After finding name origin differences between cited and quoted persons, we wanted to determine if news articles 1) represent countries at different rates 2) vary in the language used to describe scientific content related to each country. To do this analysis, we used three sources of information 1) country-related entities mentioned in the news text, 2) country affiliations of cited authors in news articles, 3) country affiliation of authors in *Nature* and Springer. Figure 4a shows example input data and a schematic of the analysis. We provide further processing details in Methods. First, we interrogated the country affiliations of cited authors. We assigned an affiliation to a manuscript if any author has affiliation from a specific country. Therefore a single manuscript may have multiple country affiliations. We analyzed any possible country affiliation for an article due to limits imposed by the *Springer* API. Affiliation query results from the *Springer* API return all country affiliations for a specific manuscript and are not linked to one particular author.

After post-processing, we analyzed a total of 12,853 articles and considered all authors within the article and their affiliations for this analysis. We found that most cited articles have at least one author with an affiliation within the United States, followed by the United Kingdom, Germany, and France (Figure 4b). Interestingly, we found a strong citation over-enrichment of many top-cited countries, but we found no evidence of under-enrichment by East Asian countries (Figure 4c, Figure Supplemental 4a). Next, we examined content differences between countries or groups of countries. For example, we wanted to determine the extent to which a country was the subject (i.e., their scientific policies, environment, pollution) or the research being performed within that country was the subject. To do this, we need to calculate the mention rate of each country. To identify if a country was mentioned in an article, we started with all organizations, countries, states, or provinces identified by coreNLP's named entity tagger. We then linked all the identified region-related named entities to countries with OpenStreetMap. Since there may be errors in both coreNLP and OpenStreetMap, we only assumed a country was mentioned when at least two unique entities mapped to the same country in a single article. On a benchmark set, we identify 4 country identifications from a total of 59 country predictions were incorrect (Figure Supplemental 1b,c). When aggregating over articles, we find that 4/30 articles contain exactly one incorrect country mention (Figure Supplemental 1b)

Once we calculated the mention rate and used the previously described citation rate, we identified countries with a consistent skew towards either a higher or lower mention to citation rate (Figure 4d and Figure Supplemental 4b). This is defined as countries where the difference between citation and mention rates is in the top or bottom 5% per year. This outlier description allowed us to identify two sets of countries based on their citation and mention rates. Those with a high relative citation to mention rate were: Germany, Netherlands, Japan, Spain, Denmark, Israel, Sweden, Switzerland, Belgium, Austria, China, South Korea, and Australia. Those with a low relative citation to mention rate: Canada, Democratic Republic of the Congo, Chile, Indonesia, Russia, South Africa, Mexico, Brasil, Philippines, Colombia, and India. We removed all countries that were in both the top or bottom 5% in different years, which excluded the United States, France, and the United Kingdom from consideration.

We then identified content differences between these two sets of countries by analyzing all of the main text from articles that mentioned and did not cite an author affiliated with each of the specified countries. After properly identifying high-frequency words across the entire corpus, we identified the top 15 most discriminative terms of each country type (Methods). Interestingly, we identified that the words most linked with mentioned countries were related to environmental or extractive topics. The top 5 terms were br, pesticide, dams, spores, and herd (Figure 4e, Figure Supplemental 4c,d). In contrast, we find that the words most related to countries with a higher citation than mention rate were science or research-related ones. The top five terms were ips, entanglement, graphene, classical, and epigenetic.

Discussion

Scientific news coverage, unlike academic articles, should represent not only the current state of research-related disciplines but also the diversity of its readership. News coverage is the conduit between the academic and public spheres, and consequently shapes the public's view of science and scientists. Though it would be best for news coverage to promote equitable representation, at a minimum news coverage should match the current rates of regional and gender participation in academia. To test whether this last point was true, we analyzed over 10,000 news articles published by *Nature* to identify quoted and cited persons. We then compared this to the authorship statistics from *Nature*'s research articles and a subset of *Springer*'s English language articles.

We first looked at possible gender biases in quotes and found a large, but decreasing, gender gap in all but one article type. We found that the decreasing trend was largely driven by a single column,

career-feature. This column has an equal number of quotes from both genders showing that gender parity is possible in science-related news coverage. In order to draw these conclusions, we performed an analysis examining the proportion of all identified quotes that were from a male speaker. As a comparator, we analyzed the proportion of male authors, which similarly is a measure of scientific participation. Using computational methods, we performed quote association and gender prediction. We observe a strong bias towards male participation across both authorship and quotes. We also identify a gender bias between first and last authors, previously shown in [31,32,33]. Interestingly, we found that while the rate of male quotes and male authorship are decreasing over time, quotation rate is decreasing faster than authorship rates. We then tried to identify what drove the observed trend. First or last authorship changes, which we anticipated to affect gendered quotation rates, did not drive the downward trend. Despite first authors more commonly being female than last authors, first and last authors were quoted at comparable rates over time. Instead, we found that column types were driving the trend, with career-feature reaching gender parity.

To further our analysis of possible coverage biases, we looked to biases in name origins of quoted and cited last authors across all the processed news articles. Our findings provide additional support for previous studies that identified under-citation [34] and under-recognition [28] of East Asian persons. Interestingly, we found that under-citation of East Asian persons to be less pronounced than under-quotation. Overall, we find that most quotes and citations are from persons with Celtic/English or European name origins, followed by East Asian, with the remaining individual origins individually making up less than 10% of both citations or quotes. Except Celtic/English (over-representation) and East Asian (under-representation), all name origins roughly match the expected academic background rate estimated by *Nature* last authorship. We also found this same pattern in our *Springer* data set.

After observing name origin biases, we sought to identify a bias in the frequency or content of coverage across countries. We first looked at possible citation biases for authors with specific country affiliations, and found that most manuscripts cited by *Nature* news have at least one author affiliated with the United States, United Kingdom, and Germany. In contrast to the name origins results, the citation rate of Chinese affiliated authors was not significantly depleted. Interestingly, we find the number of citations to articles with authors having affiliations in China is increasing at the same rate as *Springer* and *Nature* authorships. Furthermore, the increased citation and last authorship rates of Chinese affiliated authors is most pronounced in comparison to all other countries within the top ten most cited.

We then focused on identifying whether the news content about a country focused on the scientific output from that country or the country itself as the scientific subject. We postulated that a difference in citation and mention rates could indicate the difference in a news article's subject matter. To achieve this, we identified two sets of countries with a large and consistent difference in their citation and mention rates. The top "Citation" countries were Germany, Netherlands, and Japan. The top "Mention" countries were India, Colombia, and the Philippines. We then found that these two sets of countries were discussed differently. The resultant words for "Mention" countries were most related to agriculture; suggesting that the country was likely the article's subject. In contrast, the representative words for "Citation" countries were more diverse in topic, relating to biological, medical, and physics terms. The difference in discriminative terms between the two country sets is evidence that the news content may focus more on research of a country as a subject than science that comes out of it.

Through our comprehensive analysis, we were able to identify how news coverage varies by country, name origin, and gender, and compare it to scientific publishing background rates. While we found a significant gender bias, the rate of female representation in scientific news is increasing and outpacing authorships on scientific manuscripts. Furthermore, we identified a significant depletion of quotes from scientists with an East Asian name origin and a significant but smaller depletion of authors with an East Asian name origin among the citations of news content. Finally, we showed that

coverage of specific countries might be biased by content, with the country's scientific output being put in a more significant focus for some countries than the environmental aspects of other countries. Previous anecdotal studies from journalists have shown that awareness of their bias can help them to reduce it [2,3,4]. Once a bias is identified an individual can seek resources to help them find and retain diverse sources, such as utilizing international expert databases like gage [21] and SheSource [22]. Additional tips for journalists to achieve and maintain a diverse source pool is described by Christina Selby in the Open Notebook [20].

While removing biases from their coverage should be a focus for editors and journalists, many journalists are limited by the persons who can respond to their requests for an interview or leads from prominent scientists. Scientists themselves can also recommend junior researchers from diverse backgrounds, in contrast to referring journalists to well-established researchers that may already receive a lot of representation. Furthermore, since news coverage is accountable to their readership, it presents the opportunity to represent scientific perspectives that are more diverse than observed in academic publishing. We have shown that in at least one aspect that this is possible, as observed by the gender parity in quotes from the "career-features" column. News outlets and referee scientists have a unique opportunity to shape the public and their peers' perspectives on who is a scientific expert. Their choice of coverage topics and interviewees could reflect a more diverse view of science and scientists and help to diminish existing biases in scientific research.

Data and Resource availability

This manuscript was written using Manubot [35] and is available on github: manuscript repository link. All code and metadata is also available on github, full analysis repository link, under a BSD 3-Clause License. The code to generate all main and supplemental figures are available as R markdown documents within our main analysis github, in the following subfolder: notebooks. Due to copyright, we are unable to provide the scraped data used in this analysis. However, scraping code is available on our main analysis github, in the following subfolder: scraper.

Acknowledgements

Supplemental Figures

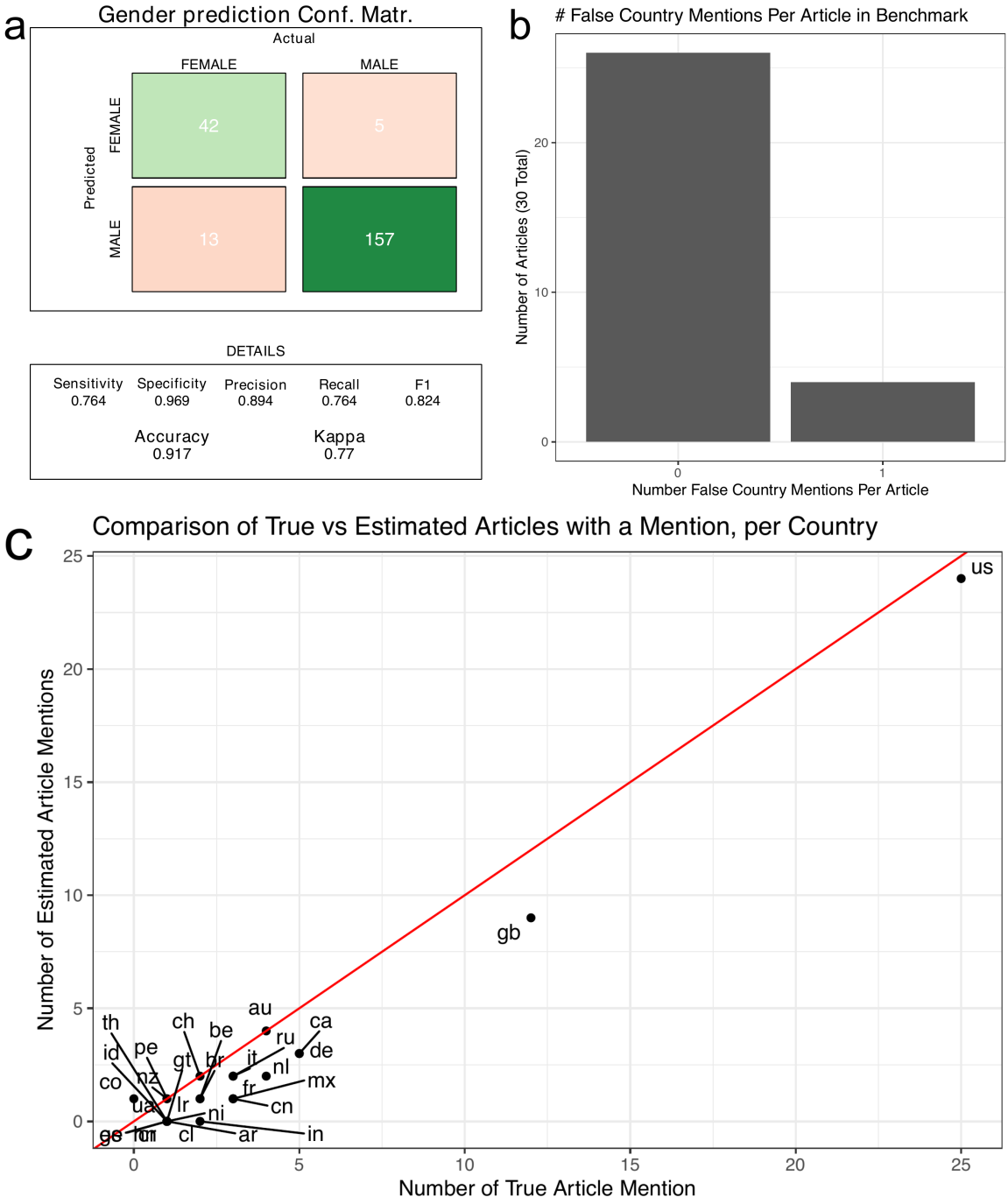

Figure Supplemental 1: Benchmark Data Panel A, depicts the performance of gender prediction for pipeline-identified quoted speakers. Panel B is a histogram of the number of articles that were falsely identified to mention a country by our processing pipeline. Panels C shows the estimated versus true frequency of country mentions within our benchmark dataset. The red line denotes the x = y line.

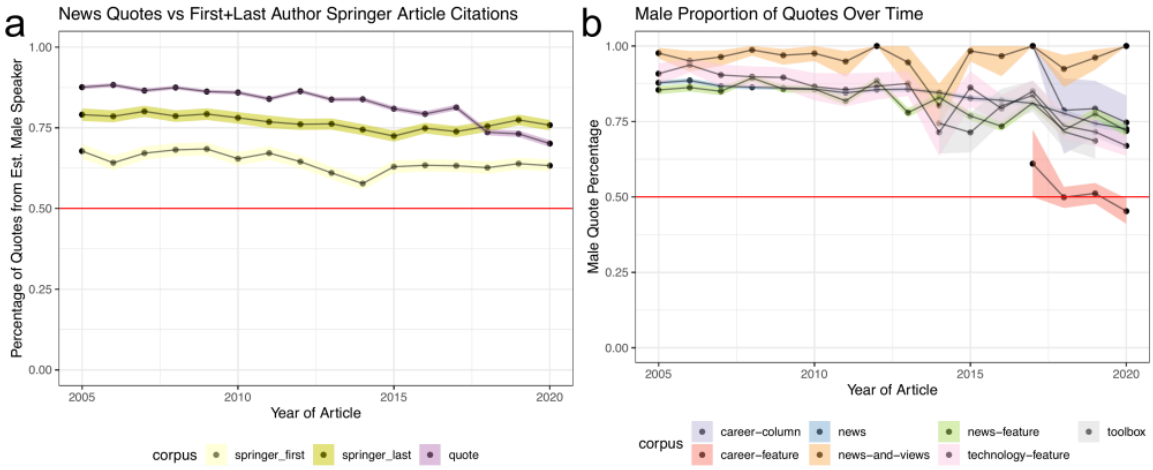

Figure Supplemental 2: Male bias is also observed in news quotes when compared against *Springer* authorship Panel A depicts three trend lines: Purple: Proportion of quotes for an estimated male speaker; Yellow: Proportion of first author articles from an estimated male author in *Springer*; Dark Mustard: Proportion of last author articles from an estimated male author in *Springer*. We observe a larger gender difference between first and last authors in *Springer* articles, however the over male bias is less than observed in *Nature* research articles. Panel B depicts the proportion of male quotes broken down by article type. In all plots the colored bands represent a 95% confidence interval and the point is the median calculated from 5,000 bootstrap samples.

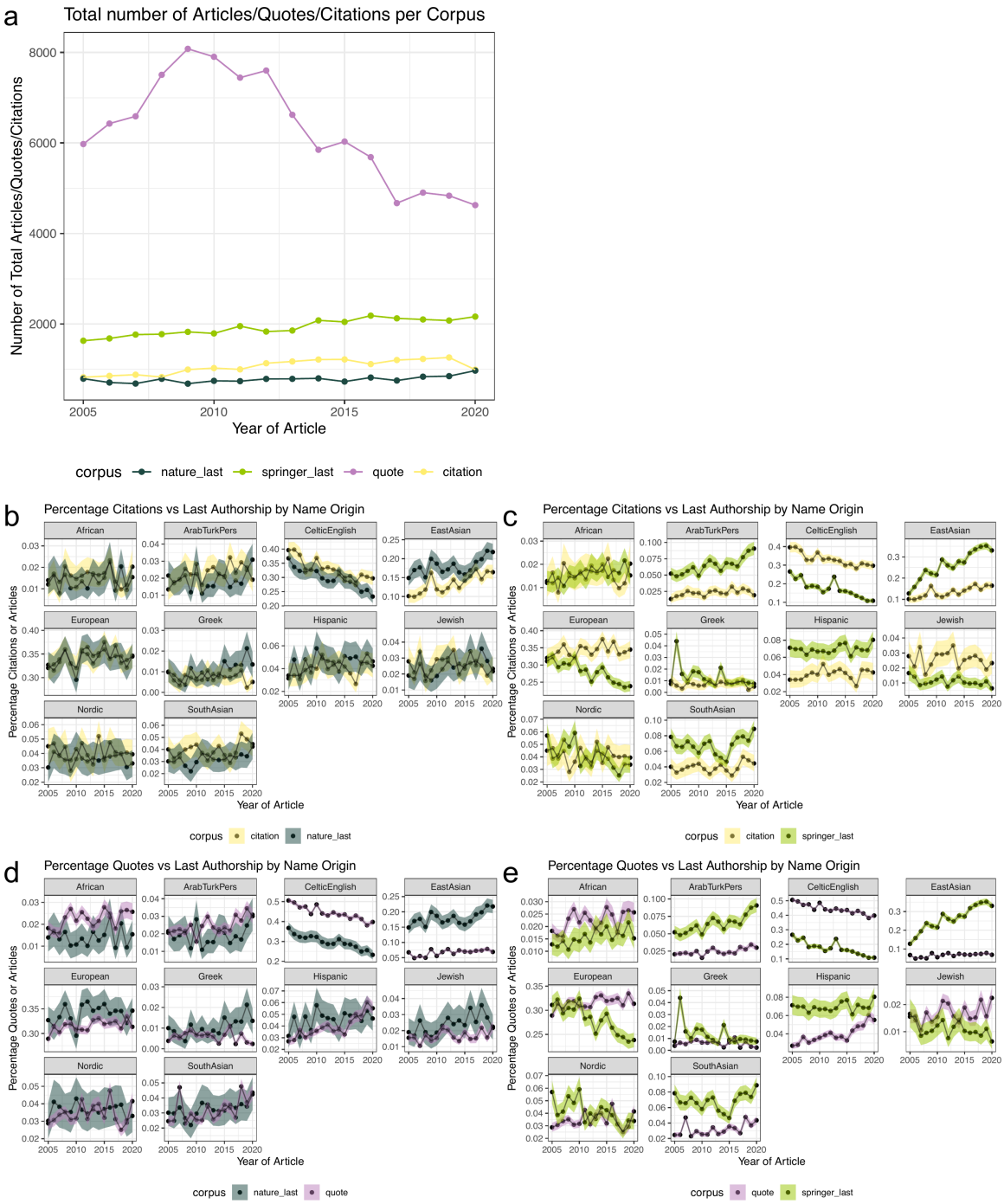

Figure Supplemental 3: Overrepresentation of Celtic/English names and underrepresentation of East Asian names is also found in *Springer* articles Panel A, depicts the number of quotes, citations, or reasearch articles considered in the name origin analysis. Panels B-D depicts ten plots, each for a possible name origin comparison against a background set. Panel B and C compare the citation rate against *Nature* (b), or *Springer* (c) last author name origins. Panel D and E compare the quote rate against *Nature* (d), or *Springer* (e) last author name origins.

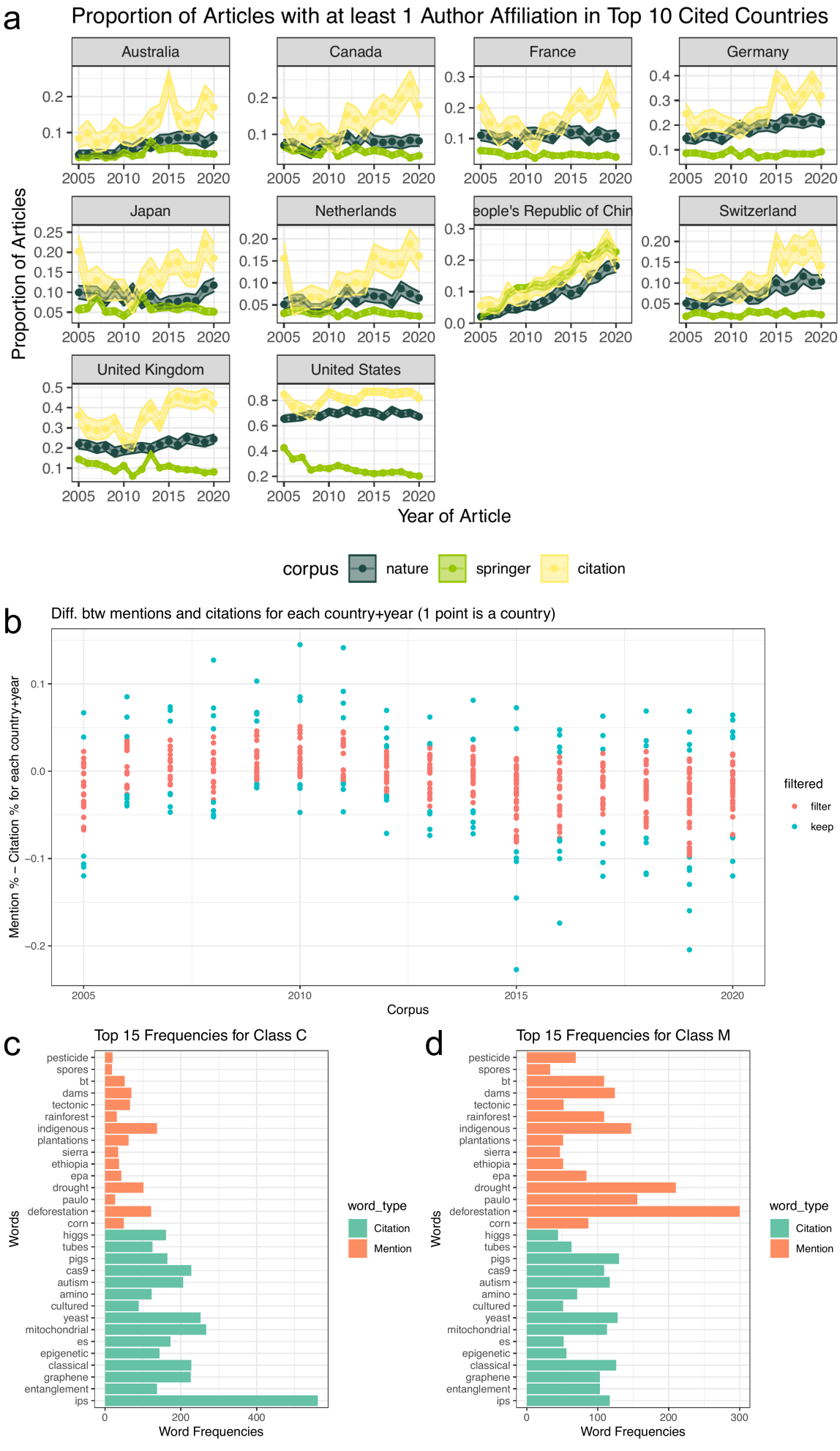
a
Proportion of Articles with at least 1 Author Affiliation in Top 10 Cited Countries
Australia
Canada
France
Germany
Japan
Netherlands
eople's Republic of Chin
Switzerland
United Kingdom
United States
Proportion of Articles
Year of Article
corpus
nature
springer
citation
b
Diff. btw mentions and citations for each country+year (1 point is a country)
Mention % − Citation % for each country+year
Corpus
filtered
filter
keep
c
Top 15 Frequencies for Class C
Words
Word Frequencies
word_type
Citation
Mention
d
Top 15 Frequencies for Class M
Words
Word Frequencies
word_type
Citation
Mention
pesticide
spores
bt
dams
tectonic
rainforest
indigenous
plantations
sierra
ethiopia
epa
drought
paulo
deforestation
corn
higgs
tubes
pigs
cas9
autism
amino
cultured
yeast
mitochondrial
es
epigenetic
classical
graphene
entanglement
ips

Figure Supplemental 4: Analysis of Country representation Panel A, depicts the citation rate for the top ten most cited articles by *Nature* news. Each plot is a comparison between the citation rate (yellow), *Nature* author affiliation (grey), and *Springer* author affiliations (dark mustard). Panel B depicts the top and bottom 5% of (mention rate - citation rate). Each point represents a country - year pair. Blue points are a country that is further considered to be a "Citation" or "Mention" country. Panel C and D show the overall word frequencies of the 15 words with the largest ratio of frequencies between "Citation" (panel C) and "Mention" (panel D) countries.

References

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

25. **humaniformat: A Parser for Human Names** https://CRAN.R-project.org/package=humaniformat

26. **Gender Prediction Methods Based on First Names with genderizeR**
Kamil Wais
The R Journal (2016) https://doi.org/gf4zqx
DOI: 10.32614/rj-2016-002

27. **Genderize.io | Determine the gender of a name** https://genderize.io/

28. **Analysis of ISCB honorees and keynotes reveals disparities**
Trang T. Le, Daniel S. Himmelstein, Ariel A. Hippen Anderson, Matthew R. Gazzara, Casey S. Greene
Cold Spring Harbor Laboratory (2020-09-22) https://doi.org/ggr64p
DOI: 10.1101/2020.04.14.927251

29. **OpenStreetMap**
OpenStreetMap
https://www.openstreetmap.org/

30. **Text Mining using "dplyr", "ggplot2", and Other Tidy Tools [R package tidytext version 0.3.1]** (2021-04-10) https://CRAN.R-project.org/package=tidytext

31. **Time's up for journal gender bias.**
Nina Schwalbe, Jennifer Fearon
Lancet (London, England) (2018-06-30) https://www.ncbi.nlm.nih.gov/pubmed/30070216
DOI: 10.1016/s0140-6736(18)31140-1 · PMID: 30070216

32. **Does academic authorship reflect gender bias in pediatric surgery? An analysis of the Journal of Pediatric Surgery, 2007–2017**
Alexandra F. Marrone, Loren Berman, Mary L. Brandt, David H. Rothstein
Journal of Pediatric Surgery (2020-10) https://doi.org/gj7mc9
DOI: 10.1016/j.jpedsurg.2020.05.020 · PMID: 32563536

33. **Gender Trends in Authorship in Psychiatry Journals From 2008 to 2018**
Kamber L. Hart, Sophia Frangou, Roy H. Perlis
Biological Psychiatry (2019-10) https://doi.org/gjtjzz
DOI: 10.1016/j.biopsych.2019.02.010 · PMID: 30935668 · PMCID: PMC6699930

34. **Racial and ethnic imbalance in neuroscience reference lists and intersections with gender**
Maxwell A. Bertolero, Jordan D. Dworkin, Sophia U. David, Claudia López Lloreda, Pragya Srivastava, Jennifer Stiso, Dale Zhou, Kafui Dzirasa, Damien A. Fair, Antonia N. Kaczkurkin, … Danielle S. Bassett
Cold Spring Harbor Laboratory (2020-10-12) https://doi.org/gj7mdc
DOI: 10.1101/2020.10.12.336230

35. **Open collaborative writing with Manubot**
Daniel S. Himmelstein, Vincent Rubinetti, David R. Slochower, Dongbo Hu, Venkat S. Malladi, Casey S. Greene, Anthony Gitter
PLOS Computational Biology (2019-06-24) https://doi.org/c7np
DOI: 10.1371/journal.pcbi.1007128 · PMID: 31233491 · PMCID: PMC6611653