bioRxiv preprint doi: https://doi.org/10.1101/2021.06.21.449261; this version posted June 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

This manuscript was automatically generated on June 18, 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 The Gordon and Betty Moore Foundation (GBMF 4552)
- **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 journalism is a critical way in which the public can remain informed and benefit from new scientific findings. Such journalism also shapes the public's view of the current state of scientific findings and legitimizes experts. Those covering science can only cite and quote a limited number of sources. Sources may be identified by the journalist's research or by recommendations by other scientists. In both cases, biases may influence who is identified and ultimately included as an expert. We analyzed 22,001 non-research articles published by *Nature* to quantify possible disparities. Our analysis considered three possible sources of disparity: gender, name origin, and country affiliation. To explore these sources of disparity, we extracted cited authors' names and affiliations, as well as extracted names of quoted speakers. While citations and quotations within a piece do not reflect the entire information-gathering process, they can provide insight into the demographics of visible sources. We then used the extracted names to predict gender and name origin of the cited authors and speakers.

In order to appropriately quantify the level of difference, we must identify a suitable reference set for comparison. We chose first and last authors within primary research articles in *Nature* and a subset of *Springer Nature* articles in the same time period as our comparator. In our analysis, we found a skew towards male quotation in *Nature* journalism-related articles, but quotation is trending toward equal representation at a faster rate than first and last authorship in 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 predicted Celtic/English origin and under-representation of names with a predicted East Asian origin. This finding was observed both in extracted quotes and journal citations, but dampened in citations. Finally, we performed an analysis to identify how countries vary in the way that they're described in scientific journalism.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.21.449261; this version posted June 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

We focused on two groups of countries: countries that are often mentioned in articles, but do not often have affiliated authors cited, and countries that have affiliated authors that are often cited, but the country is not typically mentioned. We found that the articles in which the less cited countries occur tend to have more agricultural, extraction-related, and political 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

Science journalism is an indispensable part of scientific communication and provides an accessible way for everyone from researchers to the public to learn about new scientific findings and to consider their implications. However, it is important to identify the ways in which its coverage may skew towards particular demographics. Coverage of science shapes who is considered a scientist and field expert by both peers and the public. This indication of legitimacy can either help recognize people who are typically overlooked due to systemic biases or intensify biases. Journalistic biases in general-interest, online and printed news 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 subjects or sources, with independent studies finding that between 17-40% of total subjects were female across multiple general-interest printed news outlets between 1985 and 2015 [5,6,10]. One study found 27-35% of total subjects in international science and health related news were female between 1995 and 2015, and 46% in print, radio, and television in the United States in 2015 [10]. It should be noted that scientific news coverage is confounded by the existing differences in gender and racial demographics within the scientific field [11,12]. However, we are interested in quantifying disparities with respect to observed demographic differences in the scientific field, using academic authorship as an estimate for the existing demographics. This is similar to other studies that have quantified gender or racial disparities in science as observed in citation [13,14] and funding rates [15,16,17,18,19].

In researching a story, a journalist will typically interview multiple sources for their opinion, potentially asking for additional sources, thus allowing individual unconscious biases at any point along the interview chain 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 disparities in representation may go unnoticed in a single article, analyzing a large corpus of articles can identify and quantify these disparities and help guide institutional and individual self-reflection. In the same vein as previous media studies [5,6,7,8,9,10], we sought to quantify gender and regional differences of journalism beyond the existing demographic differences in the scientific field. Our study focused solely on scientific journalism, specifically content published by *Nature*. Since *Nature* also publishes primary research articles, we used these data to determine the demographics of the expected set of possible sources. For clarity, throughout this manuscript we will refer to journalistic articles as *news* and academic, primary research articles as *papers*. Furthermore, when we refer to “authors” we mean authors of academic papers, not journalists; this work did not scrape any journalists’ names, nor derive any insights about individual journalists. In our analysis, we identified quoted and cited people by analyzing the content and citations within all news articles from 2005 to 2020, and compared this demographic to the academic publishing

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.21.449261; this version posted June 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

demographic by analyzing first and last authorship statistics across all *Nature* papers during the same time period.

Through our analysis of 22,001 news articles, we were able to identify >100,000 quotes and >8,000 citations with sufficient speaker or author information. We also identified first and last authors of >13,000 *Nature* papers. We then identified possible gender or regional differences using the extracted names. The extracted names were used to generate three data-types: quoted, mentioned, and cited people. We used computational methods to predict gender and identified a trend towards quotes from people predicted male in news articles when compared to both the general population and predicted male authorship in papers. Within the period that we examined, the proportion of predicted male attributed quotes in news articles went from initially higher to currently lower than the proportion of male first and last authors in *Nature* papers. Furthermore, we found that the quote difference 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 predicted Celtic/English origin and under-representation of names with a predicted East Asian origin in both quotes and citations.

While we focused on news from *Nature*, our software can be repurposed to analyze other text. We hope that publishers will welcome systems to identify disparities and use them to improve representation in journalism. Furthermore, our approach is limited by the features we were able to extract, which only reflects a portion of the journalistic process. Journalists could additionally track all sources they contact to self-audit. 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 paper 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 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 Nature* API. The spider "article_author_crawl.py" scraped all articles designated "Article" or "Letters" from 2005 to 2020. We only scraped author names, author

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.21.449261; this version posted June 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

positions, and associated affiliations from research articles, which we refer to as *papers*. It should be noted that "News" article designations changed over time.

coreNLP

After the 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 Nature API

Springer Nature 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 Nature* API by obtaining author information for papers cited in news articles. We selected a random set of papers to generate the *Springer Nature* background set. These papers were the first 200 English language "Journal" papers returned by the *Springer Nature* API for each month, resulting in 2400 papers per year for 2005 through 2020. To obtain the author information for the cited papers, we queried the *Springer Nature* 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 predict that the gender of the speaker is the gender of the pronoun. We ignore all quotes with no name or partial names and no associated pronouns. A summary of processed gender predictions of quotes at each point of processing is provided in Table 1.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.21.449261; this version posted June 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Name Formatting for Gender Prediction of Authors

Because we separate first and last authors, we only considered papers with more than one author. As for quotes, we needed to extract the first name of the authors. We cast names to lowercase and processed them using the R package humaniformat [25]. humaniformat identifies if names are reversed (Lastname, Firstname), as well as identifies middle names. This processing was not required for quote prediction because names written in news articles did not appear to be reversed or abbreviated. 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 papers only contain first name initials (for example, “N. Davidson”), we remove any names less than four letters (length includes punctuation) and containing a “.” or “-”, then strip out all periods from the first name. This ensures that hyphenated names are not changed, e.g. Julia-Louise remains unchanged, but removes hyphenated initials, e.g. J-L. Finally, we only consider any remaining first names of more than two characters. This is to eliminate first and middle jointly-initialized names. For example, “NR Davidson” would be reduced to “Davidson” and then eliminated due to the lack of a first name. A summary of processed author gender predictions at each point of processing is provided in Tables 2 - 4.

Name Formatting for Name Origin Prediction

In contrast to the gender prediction, we require the entire name in all steps of name origin prediction. For names identified in the *Nature* news articles, we use the same process as described for the gender prediction; we again try to identify the full name. For author names, we process the names as previously described for the gender prediction of authors. For all names, we only consider them in our analyses if they consist of two distinct parts separated by a space. Additionally, if a full name is less than three characters, we were unable to consider it as the prediction model that we apply uses 3-mers. A summary of processed name origin predictions of quotes and citations at each point of processing is provided in Tables 1 - 4.

Gender Analysis

The quote extraction and attribution annotator from the coreNLP pipeline was employed to identify quotes and their associated speakers in the 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

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.21.449261; this version posted June 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 consider 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, Celtic/English, East Asian, European, Greek, Hispanic, Hebrew, Arabic/Turkish/Persian, Nordic, and South Asian. While a full description of Wiki-2019LSTM is outside the scope of this paper, we describe it here breifly. Wiki-2019LSTM is trained on name and nationality pairs, using 3-mers of the characters in a name to predict a nationality. To ensure robust predictions, nationalities were grouped together as described in NamePrism [29]. NamePrism chose to exclude the United States, Australia, and Canada from their country groupings and were therefore excluded during training of Wiki-2019LSTM. This choice was justified by NamePrism in stating that these countries had a high level of immigration. The treemap of country groupings defined in the NamePrism manuscript are found in figure 5 of the publication [29].

After running the pre-trained Wiki-2019LSTM model, we select the highest probability origin for each name as the resultant assignment. Similar to the gender analyses, quote proportions were again directly compared against publication rates. For citations, quotes, and mentions, we calculated the proportion for a given year for each name origin. This is shown in 3 to, for example, calculate the citation rate for last authors with a Greek name origin for a single year.

$$\text{Prop. Greek Last Author Cited} = \frac{|\text{Cited Last Authors w/Greek Name}|}{|\text{Cited Last Authors w/any Name}|} \quad (3)$$

$$\text{Prop. Greek Quotes} = \frac{|\text{Quotes w/Greek Named Speaker}|}{|\text{Quotes w/any Named Speaker}|} \quad (4)$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.21.449261; this version posted June 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

$$\text{Prop. Greek Names Mentioned} = \frac{|\text{Unique Greek Names Mentioned}|}{|\text{Unique Names w/any Origin Mentioned}|} \quad (5)$$

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 [30] 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 6 for calculating the mentioned rate for Mexico for a single year.

$$\text{Prop. Mexico Mentions} = \frac{|\text{Articles with} \geq 2 \text{ unique Mexico-related terms}|}{|\text{All News Articles}|} \quad (6)$$

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* papers (7) or the total number of papers cited by news articles for that year (8). Shown below are example calculations for Colombia for a single year.

$$\text{Prop. CO Affil. in Nature} = \frac{|\text{Articles with} \geq 1 \text{ CO affil. in Nature}|}{|\text{All Nature Research Articles}|} \quad (7)$$

$$\text{Prop. CO Affil. Citations} = \frac{|\text{Cited Articles in News with} \geq 1 \text{ CO affil.}|}{|\text{All Articles Cited in News}|} \quad (8)$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.21.449261; this version posted June 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 the set of C and M countries, we analyzed the word frequencies in all news articles where the set C or M country was mentioned but not cited. We believe this would provide insight into content differences between set C and M countries. Text from articles in 2020 were not considered due to an excess of SARS-CoV-2 related terms. Using the R package tidytext [31] we extracted tokens, removed stop words, and calculated the token frequencies across all articles. We only consider tokens in set C or 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. To avoid divide by zero errors, a pseudocount of 1 is added to both the numerator and denominator. We assert that the term must be observed at least once in each set.

Bootstrap Estimations

For all analyses related to equations 1 - 8, 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 - 8 were performed by sampling the denominator set. The mean, 5th, 95th quantiles across the estimates are reported as the estimated mean, lower, 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 mean, 5th, 95th quantiles across the estimates are reported as the estimated mean, lower, and upper bounds.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.21.449261; this version posted June 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Results

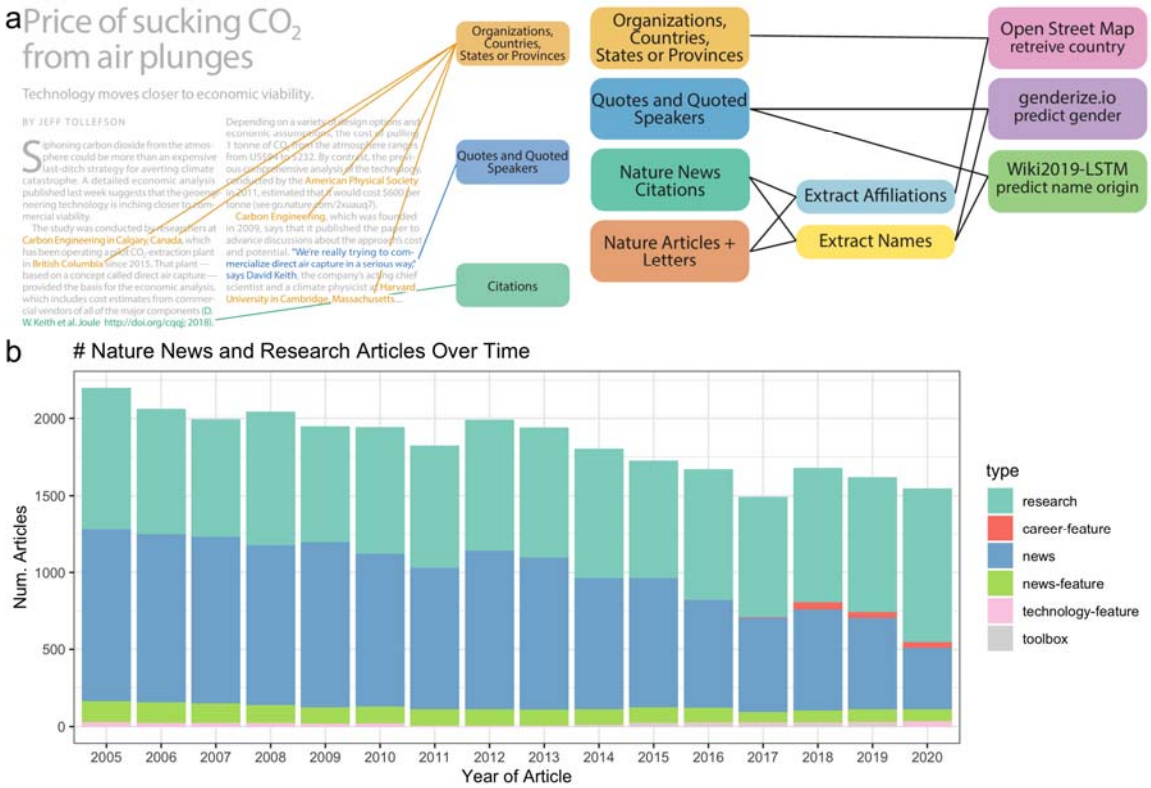


Figure 1: ***Data and Processing Pipeline Overview*** *Panel A, left, depicts an example news article and the type of data extracted from the text. Orange highlighted text depicts all named entities identified as either an organization, country, state, or province by the coreNLP pipeline. The coreNLP 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 news articles and papers 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 22,001 news-related articles hosted on "www.nature.com" that span 15 years from 2005 to 2020. Our primary focus is on 16,080 articles written by journalists which include the following five article types: "Career Feature", "News", "News Feature", "Technology Feature", and "Toolbox". "Career Feature" generally focuses on the career-related aspects of being a scientist. "News" and "News Feature" focuses on current events related to science as well as new scientific findings. It should be noted that the types of articles contained in "News" changed over time which may induce content shifts in a subset of the articles within our corpus. "Technology Feature" also covers current events and scientific findings, but additionally focuses on how science intersects with technology, such as apps, methodologies, tools, and practices. Lastly, "Toolbox" is similar to "Technology Feature", but is more centered on technology, especially the tools used to perform science. We also

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.21.449261; this version posted June 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

include one analysis of the scientist-written news articles, “Career Column” and “News and Views”, as an additional set of 5,921 articles. “Career Column” is similar to “Career Feature”, except it is not written by journalists, but individuals in the scientific field.
“News and Views” is similar to a review article, where a field expert writes an article relating to a recently written article within Nature.

The text and citations were then uniformly processed as depicted in Figure 1a to identify: 1) mentioned locations or organizations (light orange box), 2) quotes and quoted speakers (blue box), and 3) cited authors (green box). The extracted names from the text were used to generate three data types for downstream processing: quoted, mentioned, and cited people. A summary of frequencies for each data type at each point of processing is provided in Tables 1 - 4. We scraped the text using the web-crawling framework Scrapy [23], processed, and ran it through the coreNLP pipeline (Methods). 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 [30]. To identify quotes and speakers, we used the coreNLP quote extraction and attribution annotator. We performed multiple name formatting processes (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 Nature* API to attain all authors’ names, positions, and affiliations, however last authors were used as the primary comparator.

Next, we determined if the quoted speakers, mentioned countries, and cited authors in news articles have a similar demographic makeup as the scientists who publish their primary research in *Nature*. To make this determination, we used all authors’ names, positions, and affiliations of papers published by *Nature* over the same time period (Figure 1a, dark orange box). Again, last authors were used as the primary comparator. The author metadata of *Nature* papers from 2005 to 2020 totaled 13,414. To more broadly represent overall science authorship, we also separately analyzed 36,000 randomly selected *Springer Nature*-published papers from English-language journals over the same time. It should be noted that extracted quotes may come from multiple types of people, such as academic scientists, clinicians, the broader scientific community, politicians, and more. However, through anecdotal observation we believe that most sources come from either academic scientists or those actively involved in science. 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.

The top three observed article frequencies are “Research” (including “Letters” and “Articles”), “News”, and “News Feature”. Since *Nature* merged “Letters” and “Research” papers in 2019, we combined them in our analysis. We observed substantial variability in the number of *Nature* news articles by type between 2005 and 2020 (Figure 1b). The changing classification of article types may explain temporal changes in news articles. Over time, the frequency of “News” articles decreased; however, more specific news-related article types increased, including the introduction of the new categories “Career Feature”, “Toolbox”, and “Career Column”.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.21.449261; this version posted June 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Quoted Speakers and Primary Research Authors in *Nature* are More Often Male

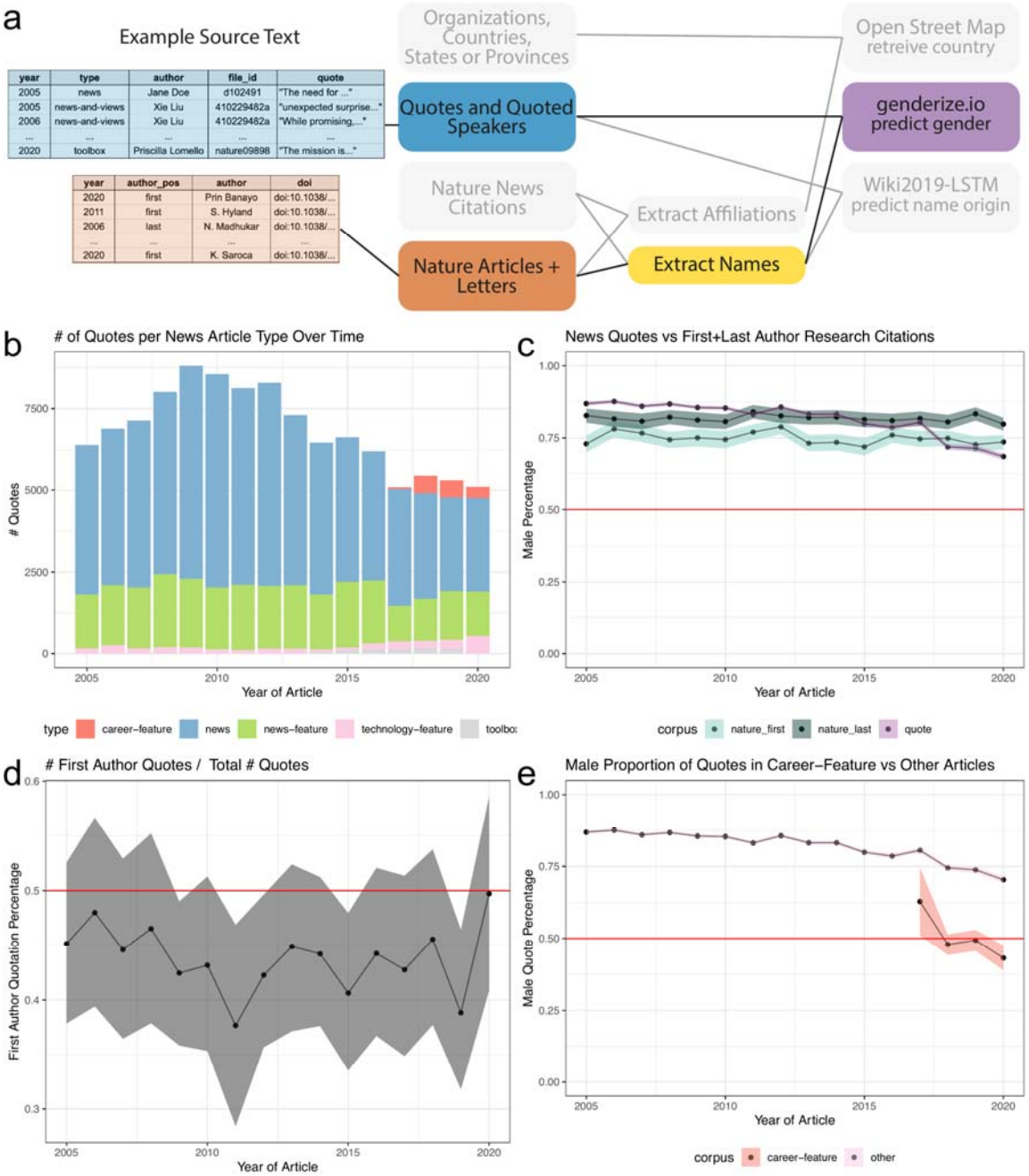


Figure 2: ***Predicted male speakers are overrepresented in quotes, but this depends on the article type.*** *Panel A, left, depicts an example of the names extracted from quoted speakers in news articles and authors in papers. Panel A, right, highlighted the data types and processes used to analyze the predicted 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 papers*

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.21.449261; this version posted June 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

from an estimated male author; Dark Blue: Proportion of predicted male last authors. 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 shift 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 5th and 95th bootstrap quantiles in all plots, and the point is the mean calculated from 5,000 bootstrap samples.

To quantify and compare the gender demographic of quoted people and authors, we analyzed their names. While we could have analyzed the proportion of unique male speakers, we were interested in measuring the overall participation rates by gender and analyzed the proportion of total quotes, e.g. a single speaker may have more than one quote in an article. Furthermore, we assume that a majority of quoted speakers are typically involved in scientific research and therefore primary research authors is a comparable demographic. Figure 2 shows an overview of the process and example input data for this analysis: 1) quotes and quoted speakers (blue box), 2) first and last authors' names of papers published by *Nature* (dark orange box). These analyses relied 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. Performance of binary prediction was evaluated on a benchmark data set of thirty randomly selected news articles, ten from each of the following years: 2005, 2010, 2015 (Figure Supplemental 1a).

We first examined the number of quotes identified within each type of news-related article (Figure 2b), totaling 119,998 quotes with 109,723 of them containing a gender prediction for the speaker. Quote frequencies vary by article type. We compared the number of quotes from predicted male people to the number of predicted male first and last authors published in *Nature*. The total number of authors with a gender prediction were 10,454 first authors and 10,488 last authors. As denoted by the red line, we found that the predicted genders of authors and source-quotes were far from gender parity (Figure 2c). Additionally, we observed a difference in the predicted genders between first and last authors, with the last authors more frequently predicted to be male.

To extend our analysis to primary research authors more broadly, we also examined a random selection of authors from English language journals published by *Springer Nature* (Figure Supplemental 2a). The predicted gender gap between first and last authors was larger in our selection of *Springer Nature* papers; however, both first and last authors were predicted to be closer to parity than for *Nature* authors. Overall, predicted male people were more frequently quoted than predicted female people in *Nature* news articles and first and last authors in *Nature* and *Springer Nature* papers over the same time period.

The gender proportions of authorship were relatively stable over time for both *Nature* and *Springer Nature* papers. In contrast, we found that the rate of quotes predicted to be from male

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.21.449261; this version posted June 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

people noticeably decreased over time. In 2005, the fraction of quotes predicted to be from male people was 86.87% (5,552/6,391) whereas in 2020 it was 68.5% (3,494/5,098). Indeed, the fraction of quotes from predicted male people was initially higher than the fraction of predicted male last authors, then slowly decreased until it was below the predicted male first and last authorship rates in 2020. We explored the possible reasons for this decrease. First, we looked at the authorship position of speakers who were quoted about their published paper (Figure 2d). We identified 8,064 quotes with an associated citation (3,382 first author and 4,682 last author quotes). We found that quotes trend slightly towards last authors from 2005 to 2020, but because the fraction of predicted male last authors remained stable over time both for *Nature* and the selection of *Springer Nature* papers, which likely does not explain the downward trend. We then analyzed the breakdown of gender predicted quotes by article type. Interestingly, one article type, “Career Feature”, achieved gender parity in its quotes (Figure 2e and Figure Supplemental 2b). In this article type, we identified a total of 1,454 quotes (759 predicted female and 695 predicted male quotes), which substantially pulled the overall quote gender ratio closer to parity from 2018 onward.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.21.449261; this version posted June 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Predicted Celtic English Name Origins are over-enriched in cited and quoted people, while predicted East Asian name origins are under-enriched

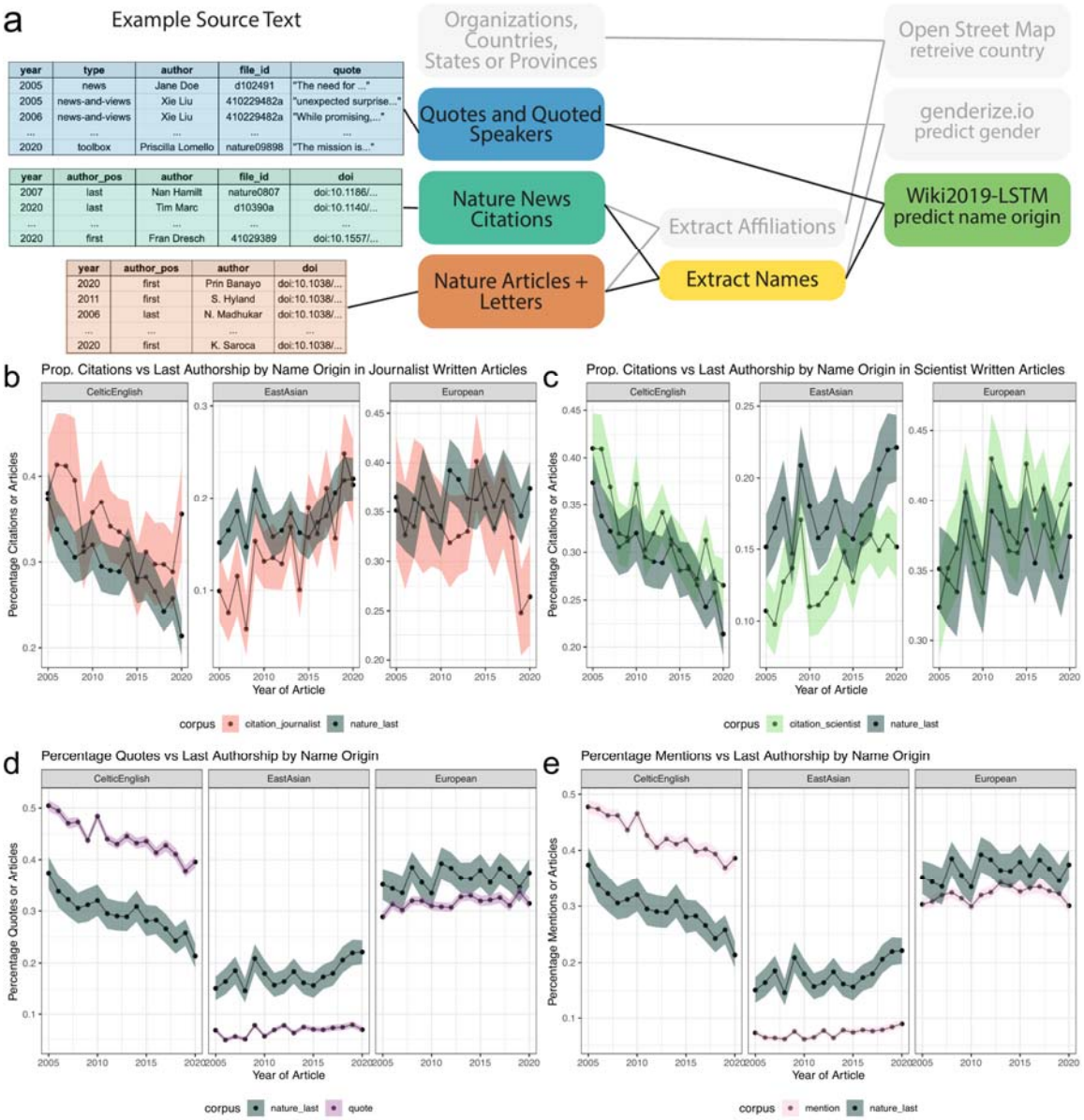


Figure 3: ***Analysis of Quotes and Citations found Over-representation of Celtic/English and under-representation of East Asian predicted name origins.*** *Panel A, left, depicts an example of the names extracted from quoted speakers and citations found within news articles and authors in papers. Panel A, right, highlights the data types and processes used to analyze the predicted origin of extracted names. Panels B and C depict a comparison between the predicted name origins of last authors in Nature and cited papers in the news. Panel B and C differ in the news article types. Panel B calculates the predicted name origin proportion using only journalist-written articles, whereas Panel C only uses scientist-written articles. Similarly,*

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.21.449261; this version posted June 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Panels D and E depict two possible trend lines, comparing predicted name origins of either quoted or mentioned people against name origins of last authors of Nature research papers. .

To identify possible disparities with respect to name origin, we again used the extracted names of quoted speakers and last authors published in *Nature*. In addition, we also identified the last authors of all papers cited by a news article. All processed names were then input into Wiki2019-LSTM and assigned one of ten possible name origins (Methods). Figure 3a shows an overview of the process and example input data for this analysis: 1) quotes and quoted speakers (blue box), 2) names of cited last authors in news articles (green) 3) last authors' names of papers published by *Nature* (dark orange box). We divided our analysis into three parts: firstly, quantifying the proportions of predicted name origins of last authors cited in *Nature* news articles. Secondly, calculating the proportion of quotes from speakers with a predicted name origin. Thirdly, calculating the proportion of unique names mentioned within an article with a predicted name origin. As a comparator set, we again used the last author names in *Nature* papers for all three analyses. Additionally, in our supplemental analyses, we compared against the last authorship in a random selection of *Springer Nature* papers. We found that the number of quotes and unique names mentioned dramatically outnumbered the number of cited authors in *Nature* news articles, as well as last authors within *Nature* papers (Figure Supplemental 3a). Still, since we have more than one hundred observations per time point for each data type, we believe this is sufficient for our analysis. Minimum and median per data type over all years: *Nature* papers, (565, 679); *Springer Nature* papers, (1298, 1684); quotes, (4577, 6194); mentions, (3634, 5002); citations in journalist-written article, (139, 267) citations in a scientist-written article, (503, 660).

In comparing the citation rate of last author name origins in news articles, we decided to additionally analyze scientist-written articles. Though fewer in number, scientist-written news articles have many citations, making the set sufficient for analysis and providing an opportunity to measure differences in citation patterns between journalists and scientists. In both journalist- and scientist-written articles, we found that most cited name origins were predicted Celtic/English or European, both with a bootstrapped estimated citation rate between 24.8-43.0% (Figure Supplemental 3b,c). East Asian predicted name origins are the third highest proportion of cited names, with a bootstrapped estimated citation rate between 5.7-24.8%. All other predicted name origins individually account for less than 9% of total cited authors.

We determined how these distributions compare to the composition of the last authors in *Nature*, by examining the top three most frequent predicted name origins (Figure 3b,c). We found a slight over-enrichment for predicted Celtic/English name origins and a small under-enrichment for predicted East Asian name origins in scientist-written and journalist-written news articles when compared to the composition of last authors in *Nature* (Figure 3b, c). Interestingly, the under-enrichment for predicted East Asian name origins in journalist-written articles was only from 2005 to 2009. Furthermore, we found no substantial difference for European or other predicted name origins (Figure Supplemental 4a). However, we did observe that papers in which the last author had European predicted name origins were more highly cited in news articles written by scientists than journalists (Figure Supplemental 44b,c). We also observed the predicted Celtic/English over-enrichment and East Asian under-representation when considering our subset of *Springer Nature* papers (Figure Supplemental 4b) for both journalist- and scientist-written news articles. In contrast to *Nature*, in the *Springer Nature* set, we see a

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.21.449261; this version posted June 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

difference in predicted European name origins, with a growing over-enrichment. Additionally, we see a difference in predicted Arabic/Turkish/Persian name origins frequencies between cited authors and *Springer Nature* authors, however the absolute difference is lower than observed for Celtic/English and East Asian predicted name origins.

We then sought to determine whether or not the quoted speaker demographic replicated the cited authors' over- and under-enrichment patterns. We found a much stronger Celtic/English over-enrichment, with quotes from those with Celtic/English name origins at a much higher frequency than quotes from those with European name origins (Figure Supplemental 3d). Additionally, we also found a much stronger depletion of quotes from people with predicted East Asian name origins (Figure Supplemental 3b), with never more than 7.9% of quotes even though they constitute between 5.7-24.8% of last authors cited in either journalist- or scientist-written news articles (Figure 3b,c). When we again compare against last authorship in *Nature*, we observe patterns consistent with the citation analysis with all predicted name origins, except for East Asian and Celtic/English closely matching the predicted name origin rate of last authors in *Nature* (Figure 3d).

Similarly, we find the same patterns in quoted speakers with East Asian, Celtic/English, and Arabic/Turkish/Persian predicted name origins when comparing against the *Springer Nature* set of last authors as we did in the previous citation analysis (Figure Supplemental 4d). In addition, we find an under-enrichment of predicted Hispanic, South Asian, and Hebrew name origins when comparing against the predicted name origin rate of last authors in our *Springer Nature* set.

Since many journalists use additional sources that are not directly quoted, we also also analyzed likely paraphrased speakers, e.g. a case in which the person was a source and mentioned in the story but not directly quoted. To do this, we identified all unique names that appeared in an article, which we term *mentions*. We found the same pattern of over-enrichment for predicted Celtic/English name origins and under-enrichment for East Asian name origins when comparing against both *Nature* and *Springer Nature* last authorships (Figure 3e, Figure Supplemental 3d,e, Figure Supplemental 4e,f).

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.21.449261; this version posted June 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Content of Science Coverage Differs between Countries

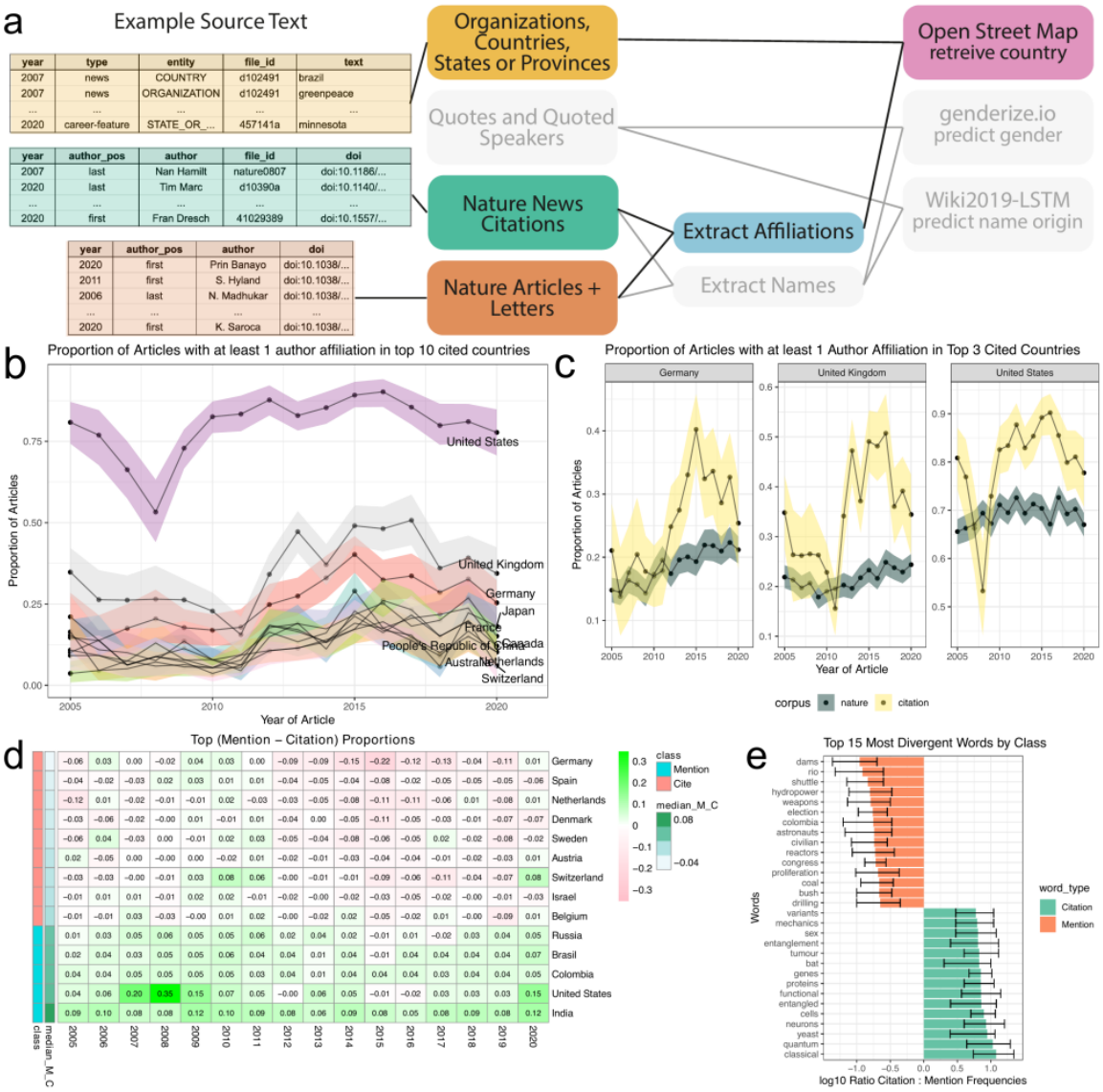


Figure 4: ***Type of country representation in news articles differ across countries*** *Panel A, left, depicts an example of the country mentions extracted from news article text, citations found within news articles, and author affiliations in papers. Panel A, right, highlights the data types and processes 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 within Nature, 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 large 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*

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.21.449261; this version posted June 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 largest proportional frequency between the two defined country sets. The width of the bar depicts the log10(Frequency in Mentions +1 / Frequency of Citations +1).

After finding name origin differences between cited and quoted people in comparison to last authorship rates, we wanted to determine if news articles 1) represent countries at different rates, or 2) vary in the language used to describe scientific content related to each country. To perform this analysis, we used three sources of information: 1) country-related entities mentioned in the news article text (light orange), 2) country affiliations of cited authors in news articles (green), 3) country affiliation of authors in *Nature* and *Springer Nature* (dark orange). 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 paper if any author, not only first or last, has affiliation with a specific country. Therefore a single paper may have multiple country affiliations. It was not possible to only identify country affiliations for a specific author position due to limitations in the *Springer Nature* API. Affiliation query results from the *Springer Nature* API return all country affiliations for a specific paper and are not linked to one particular author.

After post-processing, we analyzed a total of 1,989 papers with a citation accessible through the *Springer Nature* API. We considered all authors, not only first or last, within the article and their affiliations for this analysis. We found that most cited papers 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 of countries included in NamePrisms' grouping of countries with East Asian name origins (Figure 4c, Figure Supplemental 5a).

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 needed to identify in an article when a country is mentioned and an affiliated author from that country is cited. Our assumption is that if a country is not cited, but it is talked about, then the topic of the article is related to something happening within that country. Similarly, if a country is not mentioned, but has an affiliated author that is cited, then the science output from that country is likely to be the subject of the article. We quantified this by counting all the journalist-written news articles in which a country, region within a country, or organization affiliated with a country was mentioned, which we term a country's "mention rate". 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 found that 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)

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.21.449261; this version posted June 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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, Spain, Netherlands, Denmark, Sweden, Austria, Switzerland, Israel, and Belgium. Those with a low relative citation to mention rate were: Russia, Brazil, Colombia, the United States, and India. We removed all countries that were in both the top or bottom 5% in different years, which excluded Australia, Canada, the United Kingdom, China, France, and Japan 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 mostly related to environmental, extractive or political topics. The top 5 terms were “dams”, “rio”, “shuttle”, “hydropower”, and “weapons” (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 “classical”, “quantum”, “yeast”, “neurons”, and “cells”.

Discussion

Scientific journalism is the critical conduit between the academic and public spheres, and consequently shapes the public’s view of science and scientists. However, as observed in other forms of recognition in science, biases may shift coverage away from the known demographics within science [28]. Ideally, scientific journalism is representative of academic papers. Though it would be best for news coverage to promote equitable representation, at a minimum quotes and citations would ideally match the regional and gender demographics of scientific academia. To examine this last point, we analyzed over 22,000 news articles published by *Nature* to identify quoted, mentioned, and cited people. We then compared this to the authorship statistics from *Nature*’s papers and a subset of *Springer Nature*’s English language papers.

We first looked at possible gender differences in quotes and found a large, but decreasing, gender gap when compared to the broader population in all but one article type. We found that the decreasing trend was largely driven by the recent introduction of a single column, “Career Feature”. This column has an equal number of quotes from both genders, showing that gender parity is possible in science journalism. However, we do recognize that different journalistic columns have different purposes or may represent different demographics and be inherently more difficult to reach parity. In order to draw these conclusions, we analyzed the proportion of all identified quotes that were from a speaker predicted to be male compared to the proportion of first and last authors in *Nature* predicted to be male, which similarly is a measure of scientific participation. Using computational methods, we performed quote association and gender prediction. We observed a strong skew towards predicted male participation across both quotes

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.21.449261; this version posted June 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

within news articles and authorship within *Nature* and *Springer Nature* papers. We also identify a gender differences between first and last authors, as previously shown [32,33,34].

To further our analysis of possible coverage disparities, we looked to differences in predicted 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 [35] and under-recognition [28] of East Asian people. Interestingly, we found under-citation of people with predicted East Asian name origins to be much less pronounced than under-quotation. We do not believe that the under-quotation is driven by paraphrasing sources, which may occur more frequently with non-native English speakers. This is because our findings of under-enrichment of predicted East Asian name origins was recapitulated when we additionally looked at unique names mentioned within news articles. Furthermore, we find that scientist-written news articles tend to under-cite people with predicted East Asian name origins more than journalist-written articles. Our finding of under-quotation of people with predicted East Asian name origin was also recapitulated when we additionally looked at unique names mentioned within news articles. Overall, we find that most quotes, mentions, and citations are from people with predicted Celtic/English or European name origins, followed by East Asian, with the remaining origins individually making up less than 10% of both citations or quotes. Except Celtic/English (over-representation) and East Asian (under-representation), all predicted name origins roughly match the expected background rate estimated by *Nature* last authorship. We also found this same pattern in our *Springer Nature* data set.

After observing name origin differences, we determined if there was a difference in the frequency or content of coverage across countries. We first looked at possible citation disparities for cited authors with specific country affiliations, and found that most papers cited by *Nature* news articles have at least one author affiliated with the United States, United Kingdom, or 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 paper citations with authors having affiliations in China is increasing at the same rate as *Springer Nature* 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, Spain, and the Netherlands. The top "Mention" countries were India, the United States, and Colombia. We then found that these two sets of countries were discussed differently. The resultant words for "Mention" countries were most related to extraction, agriculture and politics, 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. We hypothesize that 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. This hypothesis assumes that no country has a specialization in a scientific topic, which is likely not true. This does, however, give us an indication that countries differ in their scientific journalism.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.21.449261; this version posted June 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 disparity, the rate of female representation in scientific news is increasing and outpacing first and last authorships on scientific papers. Furthermore, we identified a significant depletion of quotes from scientists with a predicted East Asian name origin when compared to paper authorship, and a significant but smaller depletion of cited authors with a predicted East Asian name origin in news content. Finally, we showed that coverage of specific countries differ in 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].

It should also be mentioned that we were only able to analyze the data provided through scraping "www.nature.com". This is a major limitation, because the only measures that we have of demographics of sources are people who have their name mentioned or research cited within the article. Journalists do not quote or mention all of the sources that they interviewed or cite all of the papers that they read when researching an article. For example, a person may not be mentioned or quoted in the article because of length limitations, because they do not want to be named, or if they provide information that is not directly quotable but that still shapes the content of the article. A more accurate reflection of journalists' sources would be a self-maintained record of people they interview. Our work examines disparities with respect to recognition within articles, which can be measured by mentions, quotes, or citations of people.

Furthermore, many journalists are limited by who responds to their requests for an interview or recommendations from prominent scientists. Scientists fielding reporter inquiries can also audit themselves to examine the extent to which there are disparities in the sets of experts they recommend. Journalists and the scientists they interview 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 help to reduce disparities in the outputs of science-related journalism.

Data and Resource availability

This manuscript was written using Manubot [36] 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. To ensure reproducability without violating copyright, we provide the word frequencies for each news article and the coreNLP output. Furthermore, we provide a docker image that can re-run the analysis pipeline using intermediate, pre-processed data and produce all the main

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.21.449261; this version posted June 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

and supplemental figures. To re-run the entire pipeline (including scraping), the docker image contains all necessary packages and code. The shell scripts to re-run the entire analysis are provided in the README file in the github repository.

Acknowledgements

We would like to thank Jeffrey Perkel for asking thoughtful questions that spurred this line of research, and providing feedback and insight into the news-gathering process during the course of this project.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.21.449261; this version posted June 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplemental Figures

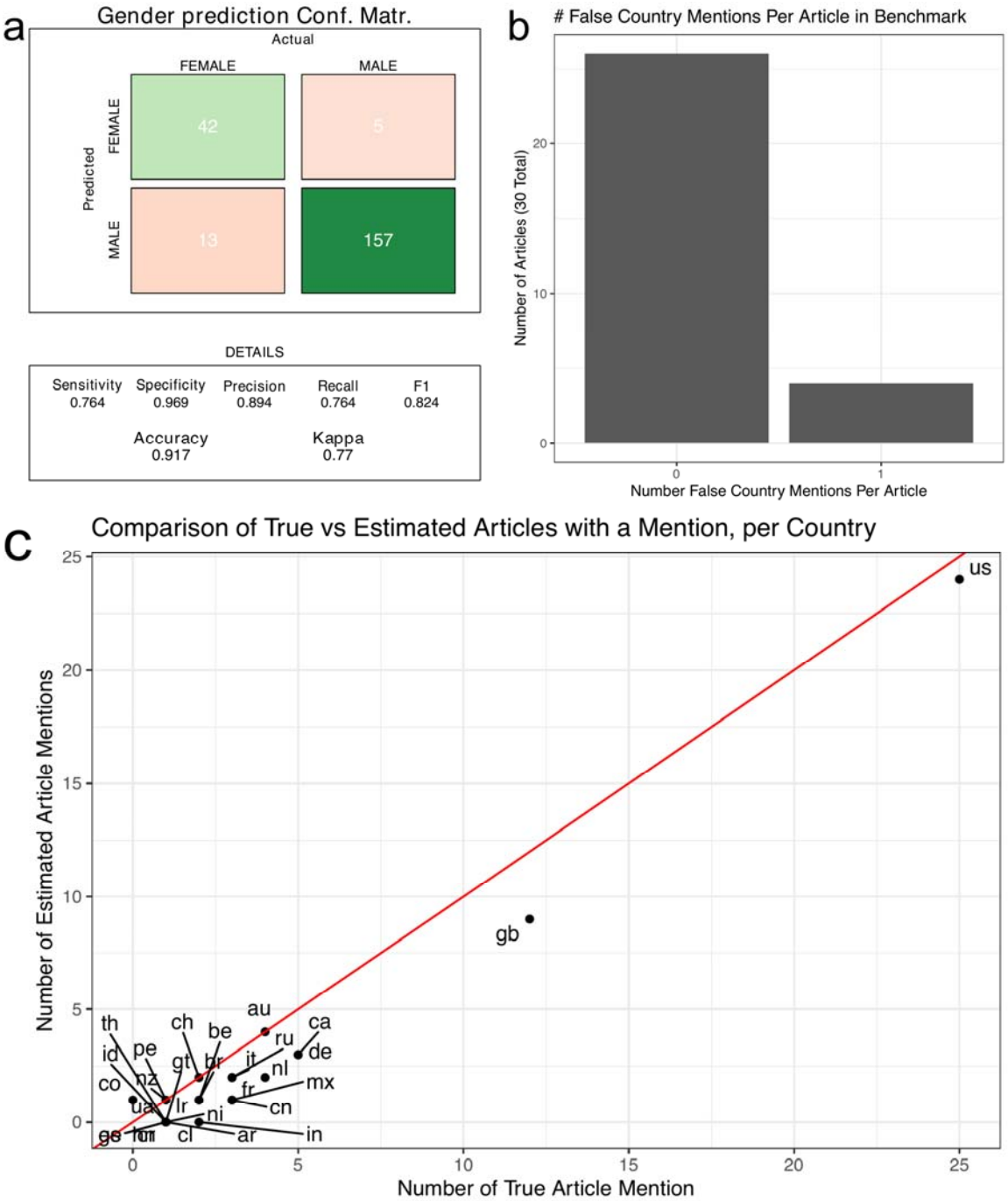


Figure Supplemental 1: ***Benchmark Data*** *Panel A, depicts the performance of gender prediction for pipeline-identified quoted speakers. Panel B is a histogram of the number of articles that were falsely identified to mention a country by our processing pipeline. Panels C shows the estimated versus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

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.21.449261; this version posted June 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

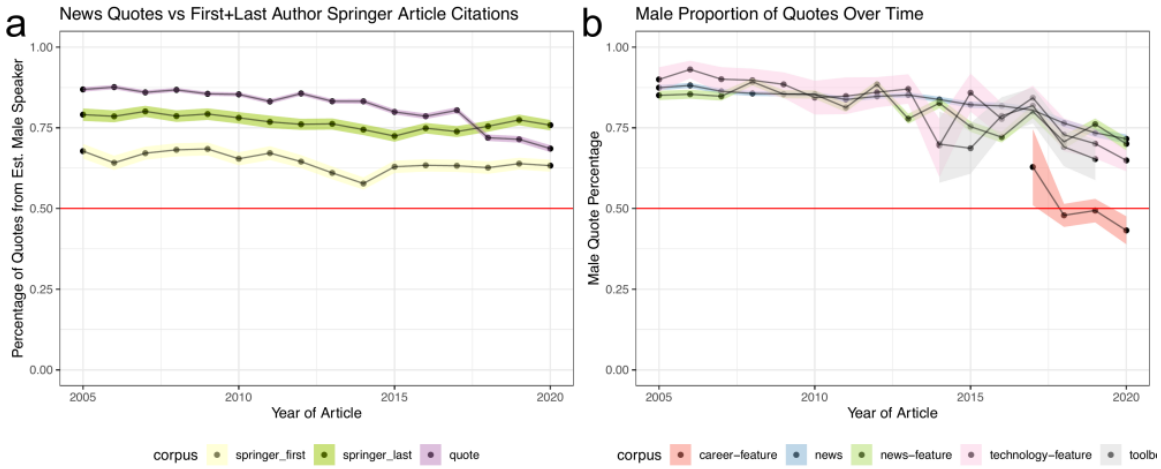


Figure Supplemental 2: ***Predicted male speakers are overrepresented in news quotes when compared against Springer Nature authorship*** *Panel A depicts three trend lines: Purple: Proportion of quotes for an estimated male speaker; Yellow: Proportion of first author articles from an estimated male author in Springer Nature; Dark Mustard: Proportion of last author articles from an estimated male author in Springer Nature. We observe a larger gender difference between first and last authors in Springer Nature articles, however the proportion of predicted male speakers 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 the 5th and 95th bootstrap quantiles and the point is the mean calculated from 5,000 bootstrap samples.*

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.21.449261; this version posted June 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

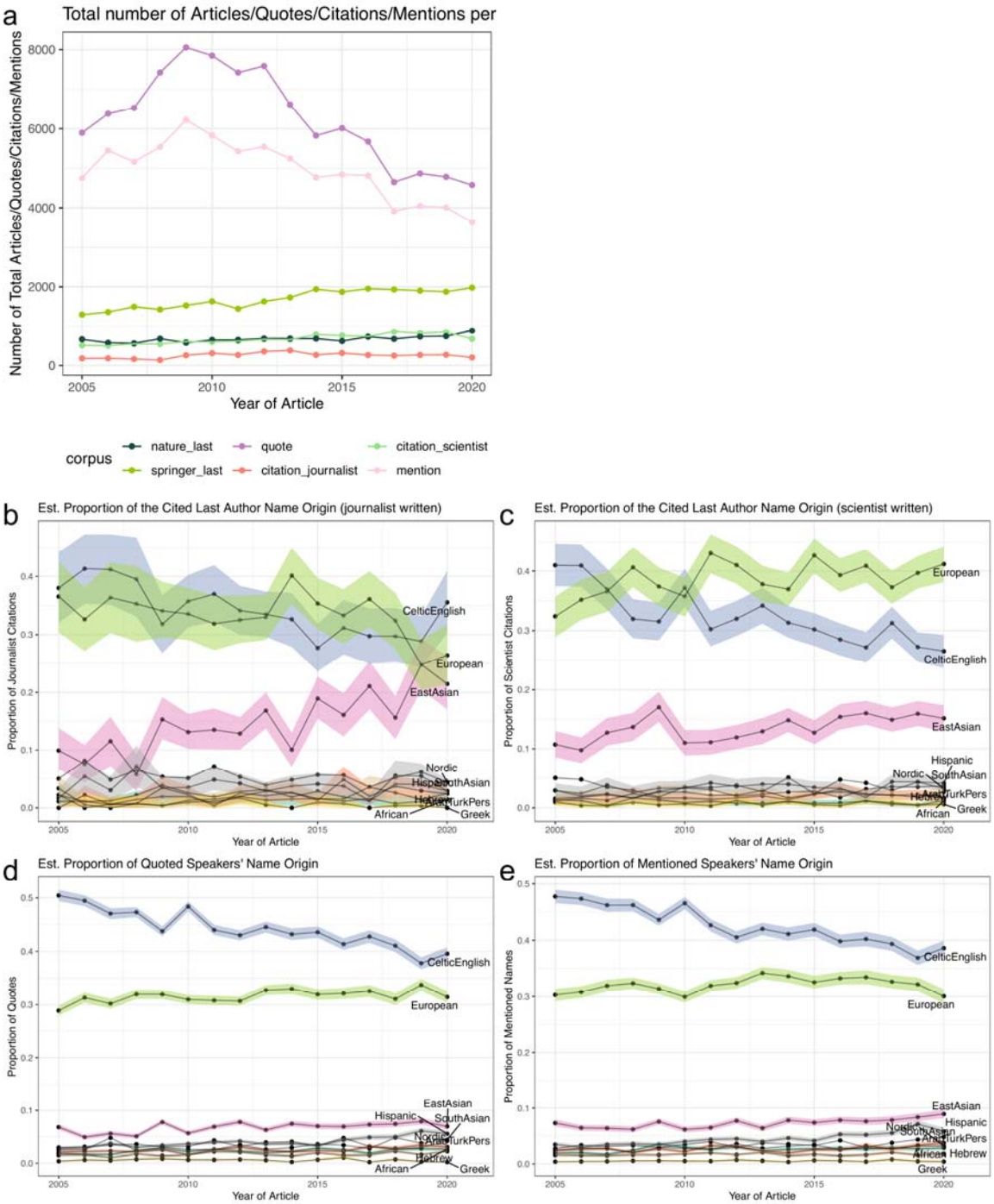


Figure Supplemental 3: ***Predicted Celtic/English, and European name origins are the highest cited, quoted, and mentioned*** *Panel A, depicts the number of quotes, mentions, citations, or research articles considered in the name origin analysis. Panels B-E depicts the proportion of a name origin in a given dataset, citations in articles written by journalists or writers, quoted speakers or mentions. In all plots the colored bands represent the 5th and 95th bootstrap quantiles and the point is the mean calculated from 5,000 bootstrap samples.*

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.21.449261; this version posted June 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

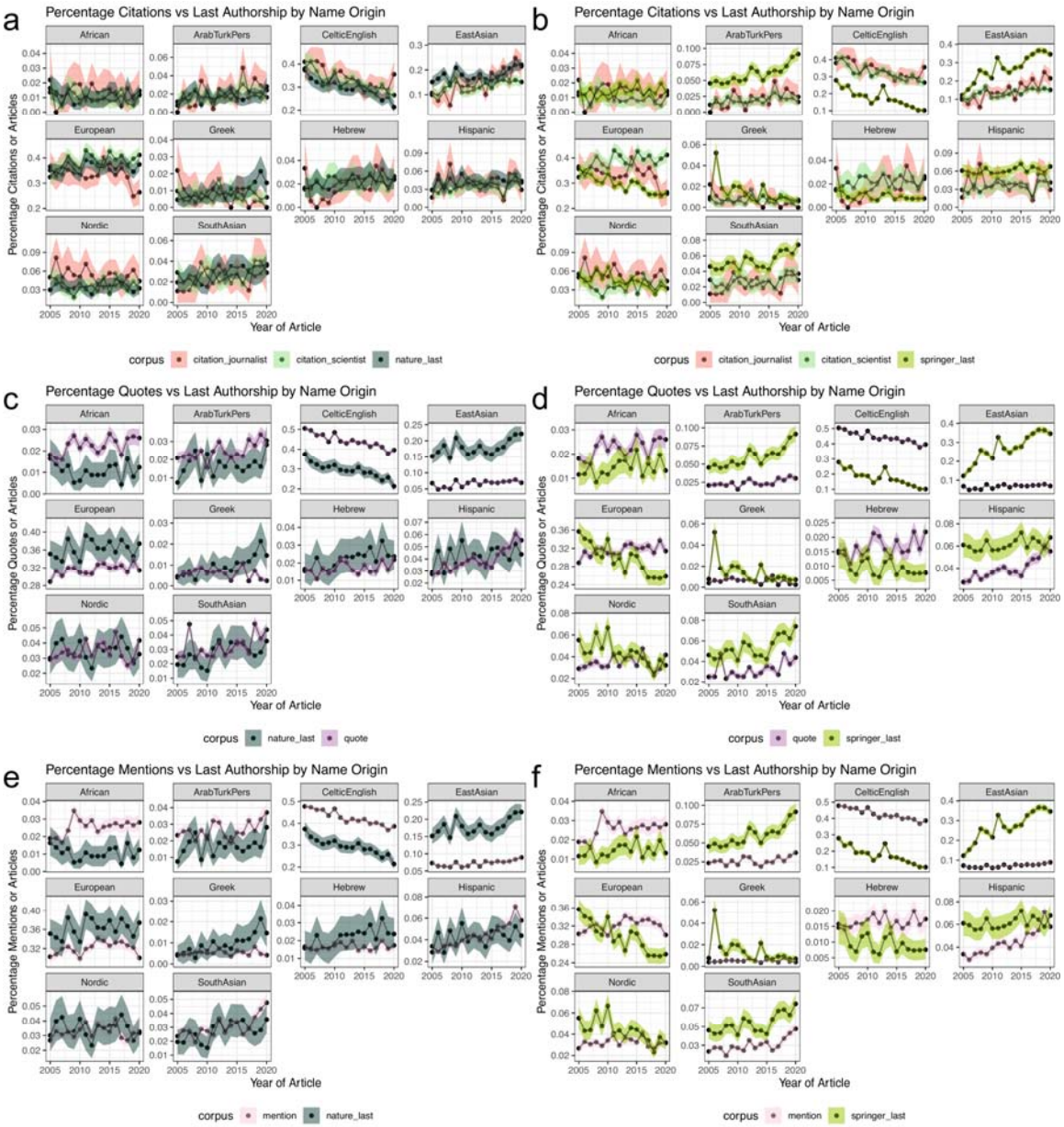
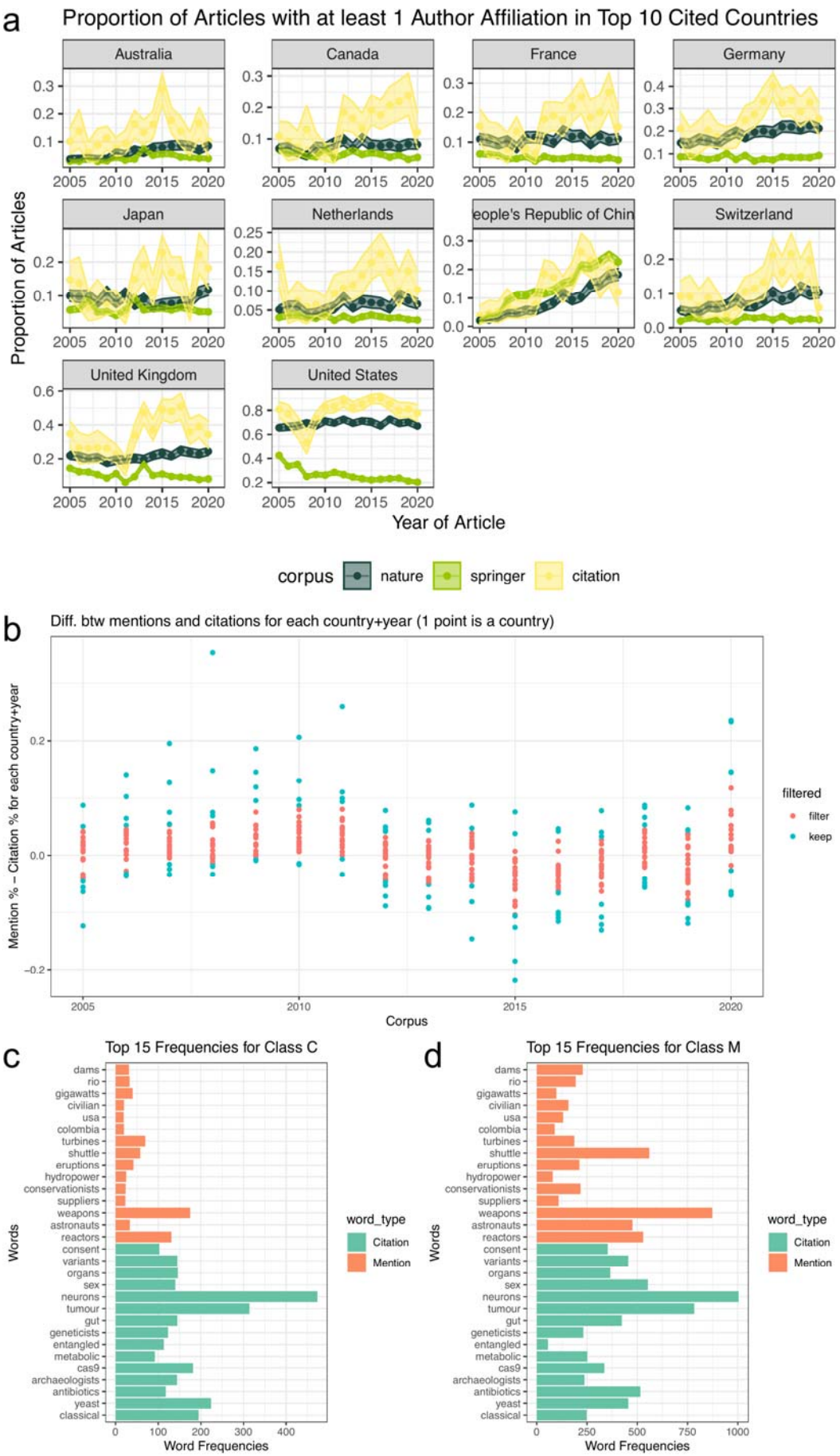


Figure Supplemental 4: ***Over-representation of predicted Celtic/English and under-representation of East Asian name origins is also found in Nature and Springer Nature articles*** *Panels A-F depicts ten plots, each for a possible name origin comparison against a background set. Panel A, C, and E compare the citation (a), quote (c), or mention (e) rate against Nature last author name origins. Panel B, D, and F compare the citation (a), quote (c), or mention (e) rate against Springer Nature last author name origins. Panels A and B additionally partition the citation rates calculated into two sets, journalist-written articles (salmon) and scientist-written articles (mint green). For C-F, only journalist written articles are considered.*

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.21.449261; this version posted June 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/2021.06.21.449261; this version posted June 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure Supplemental 5: ***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 Nature 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.*

Table 1: Breakdown of quotes at major processing steps

Processing Step	Frequency
Total Quotes	119998
Quotes with a full name or pronoun associated	110035
Quotes with a gender prediction	109723
Quote with a full name	100529
Quotes with a name origin prediction	100528

Table 2: Breakdown of citations at major processing steps

Writer of Article	Total citations	Total Springer Nature citations	First author citations with a full name	Last author citations with a full name	First author citations with a name origin prediciton	Last author citations with a name origin prediciton
Journalist	15713	5736	4405	4423	4402	4406
Scientist	40707	14597	11151	11083	11151	11065

Table 3: Breakdown of all Springer Nature articles at major processing steps

Processing Step	Frequency
# Springer Nature Articles	38400
# First + last authors with a full name in Springer Nature Articles	54509
# First + last authors with a gender prediction in Springer Nature Articles	50877
# First + last authors with a name origin prediction in Springer Nature Articles	54358

Table 4: Breakdown of all Nature articles at major processing steps

Processing Step	Frequency
# Nature Articles	13414
# First + last authors with a full name in Nature Articles	21765
# First + last authors with a gender prediction in Nature Articles	20942
# First + last authors with a name origin prediction in Nature Articles	21765

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.21.449261; this version posted June 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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/

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.21.449261; this version posted June 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

11. **Women, Minorities, and Persons with Disabilities in Science and Engineering: 2021 | NSF - National Science Foundation** https://ncses.nsf.gov/pubs/nsf21321/

12. **Why we need to increase diversity in the immunology research community**
Akiko Iwasaki
Nature Immunology (2019-08-19) https://doi.org/gkmwwv
DOI: 10.1038/s41590-019-0470-6 · PMID: 31427777

13. **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

14. **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

15. **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

16. **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

17. **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

18. **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

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.21.449261; this version posted June 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

19. **Race, Ethnicity, and NIH Research Awards**
D. K. Ginther, W. T. Schaffer, J. Schnell, B. Masimore, F. Liu, L. L. Haak, R. Kington
Science (2011-08-18) https://doi.org/csf8j8
DOI: 10.1126/science.1196783 · PMID: 21852498 · PMCID: PMC3412416

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. **Nationality Classification Using Name Embeddings**
Junting Ye, Shuchu Han, Yifan Hu, Baris Coskun, Meizhu Liu, Hong Qin, Steven Skiena
Association for Computing Machinery (ACM) (2017-11-06) https://doi.org/ggjc78
DOI: 10.1145/3132847.3133008

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.21.449261; this version posted June 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

31. **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

32. **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

33. **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

34. **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

35. **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

36. **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