Author Mentions in U.S. Science News Reveal Widespread Disparities across Name-inferred Ethnicities
(Supplementary Information)

(Dated: September 27, 2022)

I. SUPPLEMENTARY TEXT

A. Detailed Dataset Description

1. News Stories Mentioning Research Papers

The dataset of news stories mentioning scientific papers was collected from *Altmetric.com* (accessed on Oct 8, 2019), which tracks a variety of sources for mentions of research papers, including coverage from over 2,000 news outlets around the world. To control for differences in the frequency of scientific reporting and potential confounds from variations in journalistic practices across different countries, the list of news outlets was curated to 423 U.S.-based news media outlets, with each having at least 1,000 mentions in the Altmetric database. Location data for each outlet is provided by Altmetric. This exclusion criterion ensures that the dataset has sufficient volume to estimate outlet-level disparities, while still retaining sufficient diversity in outlet types, stories, and the scientific articles they cover. This initial dataset consists of 2.4M mentions of 521K papers by 1.7M news articles before 2019-10-06. Each mention in the Altmetric data has associated metadata that allows us to retrieve the original citing news story as well as the DOI for the paper itself.

2. Scraping News Content and Identifying Journalists

Due to access and permission limitations when retrieving the content of news stories, 135 outlets were excluded due to insufficient volume (27 outlets denied our access entirely; 65 outlets had less than 100 urls crawled; 43 outlets had at least 100 urls crawled, but only with non-news content such as subscription ads). For the remaining 288 outlets, 44.1% of the stories were successfully retrieved after cleaning, including dropping duplicated htmls and removing all html tags and unrelated content such as advertisements. Stories with less than 100 words were removed (less than 1%) as a manual inspection showed that the vast majority of these do not contain the complete content of the story. This process resulted in 520,061 downloaded news stories mentioning 275,403 papers from the 288 outlets.

In order to control for the effects of journalists' ethnicity and gender, we first used the *newspaper* Python package (`https://github.com/codelucas/newspaper`) to extract the journalists' names from the retrieved html news content. Since not all stories in each outlet contain the journalist information and the *newspaper* package does not work perfectly for every story that has journalist information, we focused on the top 100 outlets (ranked by the story count). With manual inspection, we verified that this package can consistently and reliably identify journalists' names for 41 of the top 100 outlets. We excluded extracted names with words signaling institutions and organizations (such as "University", "Hospital", "World", "Arxiv", "Team", "Staff", and "Editors"). We also cleaned names by removing prefix words, such as "PhD.", "M.D.", and "Dr.". We eventually obtained the journalist's name in 100,163 news stories (18.1% of all cleaned stories) for 41 outlets. Note that we did not drop any data where the journalist's name is missing. When coding journalists' gender and ethnicity, we assigned "Unknown" to those missing names.

3. News Outlets Categorization

To estimate differences across outlets, we grouped 288 news outlets into three categories based on their news production mechanisms (Table S8). The three categories are (1) Press Releases, (2) Science & Technology, (3) General News. The categorization is based on manual inspections of five random stories per outlet.

The Press Releases category is unique since many outlets in this group commonly—if not exclusively—republish university press-releases as stories, making them reasonable proxies for estimating disparity in universities' own press office. The Science & Technology category consists of magazines that focus on reporting science, such as "MIT Technology Review" and "Scientific American." These outlets typically construct a large scientific narrative referencing several papers in their stories. The General News category includes mainstream news media such as "The New York Times" and "CNN.com" that publish stories in a wide variety of topics. They have well-trained editorial staff and science journalists who are focused on accurately reporting science. Table S4 shows the number of (story, paper, author) triplets by outlet types. The average number of words per story for each outlet type is shown in Fig. S2.

4. Retrieving Paper Metadata

The Altmetric database does not contain detailed author information and therefore an additional dataset is needed to identify the authors of mentioned papers. We used the Microsoft Academic Graph (MAG) data [1] (accessed on June 01, 2019) to retrieve information for each paper based on its Document Object Identifier (DOI). Not all papers with a DOI in the Altmetric database are indexed in the MAG. We were ultimately able to retrieve 251,630 papers (all have author names) from MAG based on DOIs (matching based on lower-cased strings), which were mentioned by 472,762 stories from 288 outlets. MAG also provides rich metadata for papers, including author names, author rank, author affiliations, affiliation rank, publication year, publication venue, the paper abstract, and paper topical keywords. As all of this information will be used in our regression models, we excluded papers with missing metadata and story-paper-author triplets from rare ethnicity groups, leaving us with 100,486 papers in the final dataset.

5. Cleaning Author Names

Author names in the MAG have varying amounts of completeness. While most have the first name and surname, special care was taken for three cases: (1) If the name has a single word (e.g., Curie), the ethnicity and the gender were both set to *Unknown*, as *Ethnea* requires at least an initial. Single-word name cases occurred for 208 authorships in the final dataset. (2) If the name has an initial and surname (e.g., M. Curie), we directly fed it into the API, which provides an ethnicity inference but returns *Unknown* for gender due to the inherent ambiguity. (3) If the name has three or more words, we took the first word as the given name and the last word as the surname. However, if the first word is an initial and the second word is not an initial, we took the second word as the given name (e.g., M. Salomea Curie would be Salomea Curie) to improve prediction accuracy and retrieve a gender inference.

6. Story-Paper-Author Triplets and Corresponding Authors

We further used the Web of Science database (2019 version) to retrieve the corresponding authors for 86.0% papers in the final dataset based on the DOI. The remaining papers are mainly from disciplines such as computer science that do not have the norm to designate corresponding authors.

We focused on several authors whom journalists are likely to mention by name when covering a paper in a news story, including the first author, the last author, and any middle author who is designated as the corresponding author (note that the first author and the last author can be corresponding as well). It is possible that some papers could have equal-contributing first authors, however, our data does not have this information. We estimate that such cases are rare. For solo-author papers, we included the single author in the analyses. Papers in a few research fields that commonly use the alphabetic-based authorship ordering are also included as journalists may be unfamiliar with this norm. To examine whether a specific author is mentioned, we treated each (*story, paper, author*) triplet as an observation in the regression.

7. Final Dataset and Statistics

The final dataset consists of 223,587 news stories referencing 100,486 research papers. As some stories mentioned more than one paper and some papers were mentioned in more than one story, we have 276,202 (story, paper) mention pairs. Since multiple authors are likely to be mentioned per paper, we have 524,052 (story, paper, author) triplets in total to test whether an author is mentioned in a story.

The distribution of the number of papers and news stories over time and attention per paper are shown in Fig. S1. News story data is left censored and primarily includes stories written after 2010, as *Altmetric.com* was only launched in 2012, which limits the collection of earlier news. As shown in Fig. S1, news stories can mention papers that were published several decades before, highlighting the potential lasting value of scientific work. However, the majority of papers are mentioned within the same year or just a few years after publication. Table S2 shows the the number of authorships and triplets for authors in each broad ethnicity group, and Table S3 shows the number of triplets by journalists' inferred ethnicities.

B. Detect Author Attributions in Science News

1. Identifying Author Name Mentions

We normalized both the news content and the author names to ensure that this approach works for names with diacritics. For each story-paper-author triplet, the author's last name was searched for using a regular expression with word boundaries around the name, requiring that the name's initial letter be capitalized. While the chance exists that this process may introduce false positives for authors with common words as last names (e.g., "White"), such cases are rare because (i) few authors in our dataset have common English words as their last names, and (ii) these words rarely appear at the beginning of a sentence in the story when they would be capitalized. However, a particular exception is for two common Chinese last names "He" and "She," which can appear as third person pronouns at the start of sentences. We thus imposed additional constraints for these two names such that they must be immediately preceded with one of the following titles to be considered as a name mention: "Professor", "Prof.", "Doctor", "Dr.", "Mr.", "Miss", "Ms.", 'Mrs.". Occasionally, the author name can occur within a reference to the paper at the end of the story, which should not be counted as a name mention. As authors are typically mentioned at the beginning or in the middle of the news story, we removed the last 10% of the story content when checking name mentions (note that we obtained similar results without this filtering). Ultimately, author names were found in 41.2% of all (story, paper, author) triplets.

2. Author-Quote Detection

Authors can be mentioned by name in different forms, including quotation (e.g., “We are getting close to the truth.” said Dr. Xu.), paraphrasing (e.g., Timnit says she is confident, however, that the process will soon be perfected.), and simple passing (e.g., A recent research conducted by Dr. Jha found that drinking coffee has no harmful effects on mental health.).

We used a rule based matching method to detect explicit quotes for each (story, paper, author) triplet. We first parsed our news corpus using *spacy* (`https://spacy.io/`). We identified 18 verbs that were commonly used to integrate quoted materials in news stories, from the most 50 frequently used verbs in our news corpus, including “describe”, “explain”, “say”, “tell”, “note”, “add”, “acknowledge”, “offer”, “point”, “caution”, “advise”, “emphasize”, “see”, “suggest”, “comment”, “continue”, “confirm”, “accord”. A sentence is determined to contain a quote from the author if the following two conditions are met: (i) both the quotation mark and the author’s last name appear in the sentence, and (ii) any of the 18 quote-signaling verbs (or their verb tenses) appears within five tokens before or after the author’s last name. A manual inspection of 100 extracted quotes revealed no false quote attributes. This conservative method only gives an underestimation of the quote rate, as it may not be able to detect every quote due to unusual writing styles or article formatting. So the benefit of British-origin named scholars in getting a quote (Fig. **??**) may be even higher.

3. Detecting Institution Mentions

We checked institution mentions based on exact string matching with authors’ listed institution names in the MAG, i.e., for each (story, paper, author) triplet, we examined whether any of the author’s full institution name appears in the news story. Similar to quote detection, this method may not be able to identify every instance of institution mentions due to noise in the MAG or the story using slightly different nomenclature such as an institution’s abbreviation. However, a full list of alternative names for each institution is not available to us, we thus used this conservative method. For this reason, minority scholars’ trend in being substituted by institutions is likely underestimated.

C. Associations of Control Variables with Author Mentions

Although our focus is on ethnicity and gender, we find that many control variables are strongly associated with author mention rates. Examining the influence of these factors can lead to a better understanding of the mechanisms at play in science reporting. Below we interpret their effects based on Model 5 (Table **??**) along three themes: (1) prestige related inequality, (2) impact of co-authorship, and (3) story content effects.

Not surprisingly, being designated as the corresponding author is positively associated with name mentions. Scholars who have a high professional rank or are affiliated with prestigious institutions receive outsized name mentions in science news when their research is covered. Popular authors whose research received many press coverage are more likely to be mentioned by name. This result suggests that the benefits of status, the so-called “Matthew Effect” [2], persist even after publication.

Having more co-authors on a paper has a negative effect on the author being mentioned. Compared to the last author position, the first author is more likely to be mentioned by name, whereas the middle author is less likely to be named. The observed first position effect might due to the fact that, among papers (excluding solo-author papers) that have the corresponding author information, 59.9% have the first author as corresponding and only 36.1% have the last author as corresponding. Solo-authored papers have been decreasing over time and are associated with lower impact on average [3, 4]. However, our results highlight an underappreciated benefit—conditional on a paper being referenced in the news, a solo author is significantly more likely to be mentioned compared to authors of a multi-author paper. Although seemingly counter to previous studies, it has a natural explanation—there is only one person to mention if need be.

The coefficients for story features point to the multifaceted nature of science reporting. Although the volume of science reporting is increasing over time (Fig. S1a), journalists tend to mention authors less frequently in later years. At the same time, while older papers are still discussed in the media (Fig. S1c), journalists are less likely to mention authors of these studies as often. When more papers are referenced in a story, their authors are less likely to be mentioned. We hypothesize that such stories are often citing multiple scientific papers to construct a large narrative and thus those papers are only mentioned in passing. Longer stories are more likely to mention author names as they have more space to engage the authors.

D. Does It Matter Who Is Reporting?

Understanding whether ethnic disparities are related to journalists' own identities may help uncover the mechanisms producing them. First, journalists of different ethnicities may differ in their overall tendencies to mention authors. If so, disparities may be driven by the composition of journalists. Our fullest model controls for journalists' name-inferred ethnicity, and shows that journalists with minority-identity associated names are not more or less likely to mention authors compared with journalists with Male or British-origin names (main text, Table 2, Model 5). We also note that, when dropping controls for outlets (main text, Table 2, Models 3-4), journalists' ethnicities become significant, suggesting that journalists' differential behavior might be explained by variations at the outlet level, *i.e.* certain news outlets mention authors more or less often and certain groups of journalists are under- or over-represented in those outlets.

Second, there might exist interactive relationships between authors' and journalists' ethnic identities. One intuitive hypothesis, which we call "ethnic hierarchy," is that all journalists, regardless of their perceived ethnicity, prefer to mention British-origin named scholars over others. On the other hand, journalists may prefer to mention authors of the same ethnicity, which we call "ethnic homophily". Evidence for demographic homophily is pervasive[5]. For example, concordance of gender identities between actors has been found to predict outcomes in domains such as healthcare[6]. However, the relatively small number of cases of identified journalists (Table S3) prevents us from including the full interactions between author's and journalist's ethnicities in the model. The present study thus lacks the evidence to suggest either ethnic hierarchy or homophily hypotheses. However, this is an important avenue for future research.

II. SUPPLEMENTARY TABLES

Broad Ethnic Category	**Individual Ethnicity**
African	*African*
British-origin	*English*
Chinese	*Chinese*
non-Chinese East Asian	*Indonesian*, *Japanese*, *Korean*, *Mongolian*, *Thai*, *Vietnamese*
Eastern European	*Hungarian*, *Romanian*, *Slav*
Indian	*Indian*
Middle Eastern	*Arab*, *Israeli*, *Turkish*
Southern European	*Hispanic*, *Italian*, *Greek*
Western & Northern European	*Baltic*, *Dutch*, *French*, *German*, *Nordic*
Caribbean	*Caribbean*
Polynesian	*Polynesian*
Unknown	Note: names are unrecognized by *Ethnea*.

TABLE S1. 26 individual ethnicities were grouped into 11 broad ethnic categories. The last two groups, Caribbean and Polynesian, were excluded due to less than 100 observations.

Authors Broad Ethnic Category	**# Paper Authorships**	**# Triplets**
British-origin	81,226	234,510
Western & Northern European	39,007	106,331
Southern European	19,109	51,134
Chinese	16,054	43,039
Middle Eastern	9,185	26,082
Indian	7,505	21,314
non-Chinese East Asian	7,816	19,068
Eastern European	6,315	17,251
African	1,079	2,774
Unknown Ethnicity	898	2,549
Total	188,194	524,052

TABLE S2. The number of paper authorships and the total number of (story, paper, author) triplets for the 9 high-level ethnic groups. Note that there are 100,486 unique papers, with some counted twice or more for authorships. For example, if a paper has 3 authors and gets covered by 2 news stories, it contributes 3 (paper, author) pairs, and 6 (story, paper, author) triplets.

Journalists Broad Ethnic Category	**# Triplets**
British-origin	68,652
Western & Northern European	13,790
Southern European	10,594
Middle Eastern	3,494
Eastern European	2,924
Chinese	2,449
Indian	2,409
non-Chinese East Asian	910
African	643
Unknown Ethnicity	418,187
Total	524,052

TABLE S3. The number of (story, paper, author) triplets in our regression data by journalists' ethnicity.

Outlet Type	**# Outlets**	**Example Outlet**	**# Triplets**	**Perc. Aut. Ment.**
Press Releases	21	EurekAlert!	165,343	63.5%
Science & Technology	86	MIT Technology Rev.	137,851	41.9%
General News	181	The New York Times	220,858	24.2%

TABLE S4. The number of outlets, the number of (story, paper, author) triplets, and the percentage of triplets that have mentioned the author, for three outlet types. The full list of 288 outlets are available in Appendix Table S8.

Author Name	***Ethnea***	**U.S. Census**	**Wikipedia**
Alana Lelo	African	White	Romance Language
Samuel Lawn	African	White	British-origin
Saka S Ajibola	African	Black	East Asian
Mosi Adesina Ifatunji	African	Black	African
Sebastian Giwa	African	White	African
Olabisi Oduwole	African	White	African
Chidi N. Obasi	African	White	African
Habauka M. Kwaambwa	African	Asian	African
Esther E Omaiye	African	White	African
Aurel T. Tankeu	African	White	British-origin

TABLE S5. A random sample of 10 African-named authors predicted by *Ethnea* (out of 908 in total in our data) and their ethnicity or race categories based on the Wikipedia data or the U.S. census data.

Author Name	**U.S. Census**	***Ethnea***	**Wikipedia**
E. Robinson	Black	British-origin	British-origin
Momar Ndao	Black	Romance Language	African
Angela F Harris	Black	British-origin	British-origin
Daddy Mata-Mbemba	Black	Romance Language	African
A Bolu Ajiboye	Black	African	African
Lasana T. Harris	Black	British-origin	British-origin
John M. Harris	Black	British-origin	British-origin
Edwin S Robinson	Black	British-origin	British-origin
Eric A. Coleman	Black	British-origin	British-origin
Mp Coleman	Black	British-origin	British-origin

TABLE S6. A random sample of 10 Black authors predicted based on the U.S. census data (out of 892 in total in our data) and their ethnicity categories based on *Ethnea* or the Wikipedia data.

III. SUPPLEMENTARY FIGURES

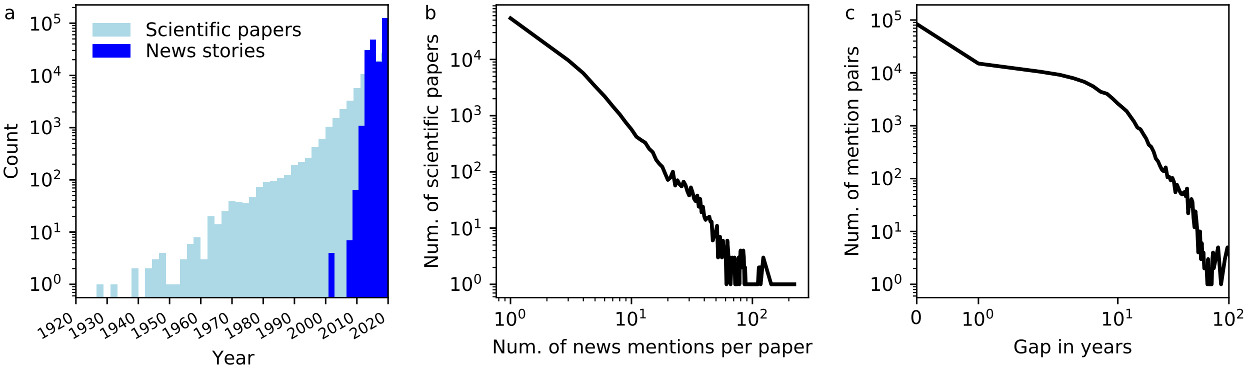

FIG. S1. **a,** The number of news stories and research papers in our mention date over time. **b,** The distribution of the number of news mentions per paper. **c,** The distribution of the *year gap* between paper publication date and news story mention date for all 276,202 story-paper mention pairs in the final dataset.

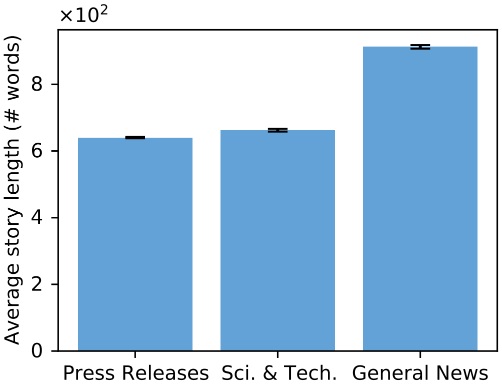

FIG. S2. The average story length for three types of outlets. Error bars show 95% confidence intervals.

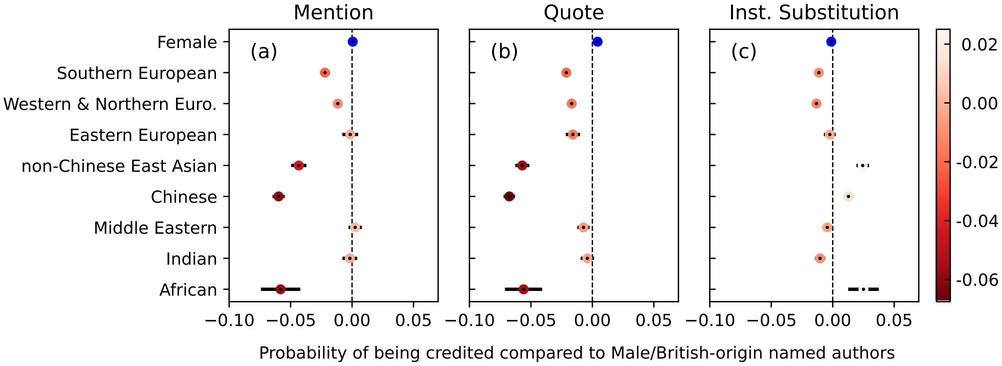

FIG. S3. The average marginal effects of ethnicity estimated based on 524,052 observations in the full data. Authors with minority-ethnicity names are less likely to be mentioned by name (**left**) or quoted (**middle**), and are more likely to be substituted by their institution (**right**). A negative (positive) marginal effect indicates a decrease (increase) in probability compared to authors with Male (for gender) or British-origin (for ethnicity) names. The colors are proportional to the absolute probability changes. *Female* is colored as blue to reflect its difference from ethnicity identities. The error bars indicate 95% bootstrapped confidence intervals.

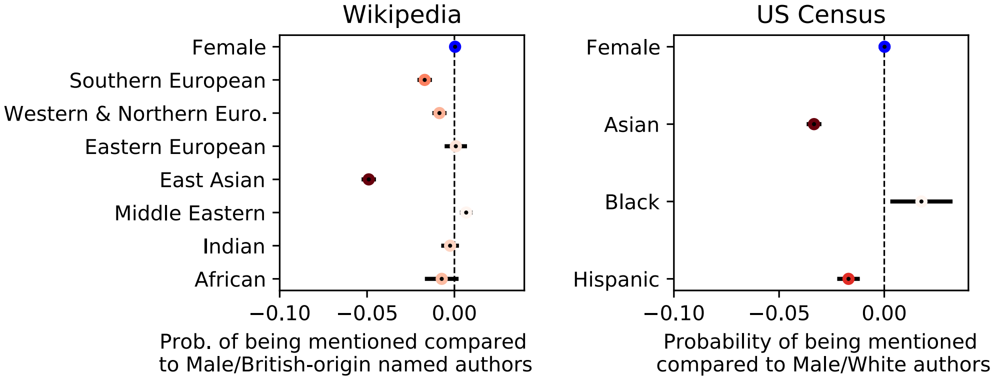

FIG. S4. The average marginal effects in mention probability for author names' demographic associations, using Wikipedia data for coding ethnicity (**Left**) or U.S. Census data for coding race (**Right**) based on author (or journalist) names. Note that gender is still inferred using *Ethnea*.

[1] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-j. P. Hsu, and K. Wang, in *WWW* (2015).

[2] R. K. Merton, Science **159**, 56 (1968).

[3] M. Greene, Nature **450**, 1165 (2007).

[4] S. Milojević, Proceedings of the National Academy of Sciences **111**, 3984 (2014).

[5] M. McPherson, L. Smith-Lovin, and J. M. Cook, Annual Review of Sociology **27**, 415 (2001).

[6] B. N. Greenwood, S. Carnahan, and L. Huang, Proceedings of the National Academy of Sciences **115**, 8569 (2018).

Appendix A: Tables

TABLE S7: A random sample of 10 names for each of the 24 individual ethnicities and the "Unknown" category. All 6 MONGOLIAN names in our data are shown here.

Ethnicity	**Name Example**	**Gender**
AFRICAN	Dora Wynchank	F
	Benjamin D. Charlton	M
	J. Nwando Olayiwola	unknown
	Ayodeji Olayemi	M
	Elizabeth Gathoni Kibaru	F
	Christopher Changwe Nshimbi	M
	Naganna Chetty	unknown
	Benjamin Y. Ofori	M
	Khadijah Essackjee	F
	Jeanine L. Marnewick	F
	Habtamu Fekadu Gemede	M
ARAB	Zaid M. Abdelsattar	M
	Alireza Dirafzoon	M
	Ahmad Nasiri	M
	Saleh Aldasouqi	M
	Ibrahim A. Arif	M
	Sameer Ahmed	M
	A Elgalib	unknown
	Taha Adnan Jan	M
	Mohsen Taghizadeh	M
	Behnam Nabet	M
BALTIC	Skirmantas Kriaucionis	M
	Airidas Korolkovas	M
	Egle Cekanaviciute	F
	Arunas L. Radzvilavicius	M
	Ieva Tolmane	F
	Alberts B	M
	Gediminas Gaigalas	M
	Armandas Balcytis	unknown
	Ruta Ganceviciene	F
	Andrius Pašukonis	M
CHINESE	Chin Hong Tan	unknown
	Li Yuan	unknown
	Yalin Li	unknown
	Xian Adiconis	unknown
	Philip Sung-En Wang	M
	Xiaohui Ni	unknown
	Minghua Li	unknown
	Fang Fang Zhang	F
	Li-Qiang Qin	M
	Jian Tan	unknown
DUTCH	Pieter A. Cohen	M
	I. Vandersmissen	unknown
	Marleen Temmerman	F
	Gerard 't Hooft	M
	A. Yool	unknown
	G. A W Rook	unknown
	Fatima Foflonker	F
	Mirjam Lukasse	F
	Sander Kooijman	M
	Izaak D. Neveln	M
ENGLISH	Isabel Hilton	F
	Gavin J. D. Smith	M
	Katherine A. Morse	F

	Andrew S. Bowman	M
	T. M. L. Wigley	unknown
	Francis Markham	M
	Neil T. Roach	M
	Brooke Catherine Aldrich	F
	Vaughn I. Rickert	M
	Kellie Morrissey	F
FRENCH	Lucas V. Joel	M
	Daniel Clery	M
	Pierre Jacquemot	M
	Scott Le Vine	M
	Nathalie Dereuddre-Bosquet	F
	Stéphane Colliac	unknown
	Adelaide Haas	F
	Julie M. D. Paye	F
	Justine Lebeau	F
	Arnaud Chiolero	M
GERMAN	Laure Schnabel	F
	Jeff M. Kretschmar	M
	E. Homeyer	unknown
	Maren N. Vitousek	F
	D. Wild	unknown
	Hany K. M. Dweck	M
	E. M. Fischer	unknown
	Paul Marek	M
	Hans-Jörg Rheinberger	M
	Daniel James Cziczo	M
GREEK	Mary J. Scourboutakos	F
	Anita P Courcoulas	F
	Elgidius B. Ichumbaki	unknown
	Stavros G. Drakos	M
	Nikolaos Konstantinides	M
	Constantine Sedikides	M
	Maria A. Spyrou	F
	Panos Athanasopoulos	M
	Aristeidis Theotokis	M
	Amy H. Mezulis	F
HISPANIC	Mirela Donato Gianeti	F
	Julio Cesar de Souza	M
	Paulina Gomez-Rubio	F
	José A. Pons	M
	Arnau Domenech	M
	Nicole Martinez-Martin	F
	Mauricio Arcos-Burgos	M
	Raquel Muñoz-Miralles	F
	Annmarie Cano	F
	Merika Treants Koday	F
HUNGARIAN	Andrea Tabi	F
	Róbert Erdélyi	M
	Gabor G. Kovacs	M
	Xenia Gonda	F
	Erzsébet Bukodi	unknown
	Julianna M. Nemeth	F
	Ian K. Toth	M
	Zoltan Arany	M
	Cory A. Toth	M
	Ashley N. Bucsek	unknown
INDIAN	Sachin M. Shinde	M
	Govindsamy Vediyappan	M
	Ashish K. Jha	M
	Tamir Chandra	M

	Hariharan K. Iyer	M
	Chanpreet Singh	unknown
	Ravi Chinta	M
	Madhukar Pai	M
	Lalitha Nayak	F
	Ravi Dhingra	M
INDONESIAN	Dewi Candraningrum	unknown
	Richard Tjahjono	M
	T. A. Hartanto	unknown
	Johny Setiawan	M
	Truly Santika	unknown
	Chairul A. Nidom	unknown
	Christine Tedijanto	F
	Alberto Purwada	M
	Ardian S. Wibowo	M
	Anna I Corwin	F
ISRAELI	Ron Lifshitz	M
	Martin H. Teicher	M
	Ruth H Zadik	F
	Gil Yosipovitch	M
	Mor N. Lurie-Weinberger	unknown
	J. Tarchitzky	unknown
	Ilana N. Ackerman	F
	B. Trakhtenbrot	unknown
	Yoram Barak	M
	Mendel Friedman	M
ITALIAN	Tiziana Moriconi	F
	Marco Gobbi	M
	Marco De Cecco	M
	F. Govoni	unknown
	Theodore L. Caputi	M
	Mark A Bellis	M
	Fernando Migliaccio	M
	Julien Granata	M
	Jennifer M. Poti	F
	Brendan Curti	M
JAPANESE	Takuji Yoshimura	M
	Maki Inoue-Choi	F
	Masaaki Sadakiyo	M
	Moeko Noguchi-Shinohara	F
	Naoto Muraoka	M
	Shigeki Kawai	M
	Koji Mikami	M
	Masayoshi Tokita	M
	Naohiko Kuno	M
	Saba W. Masho	F
KOREAN	Jih-Un Kim	M
	Hanseon Cho	unknown
	Hyung-Soo Kim	M
	Yun-Hee Youm	F
	Yoon-Mi Lee	unknown
	Soo Bin Park	F
	Yungi Kim	unknown
	Woo Jae Myung	unknown
	Kunwoo Lee	unknown
	Sandra Soo-Jin Lee	F
MONGOLIAN	C. Jamsranjav	unknown
	Jigjidsurengiin Batbaatar	unknown
	Khishigjav Tsogtbaatar	unknown
	Migeddorj Batchimeg	unknown
	Tsolmon Baatarzorig	unknown

NORDIC	Steven G. Rogelberg	M
	Kirsten K. Hanson	F
	Jan L. Lyche	M
	Morten Hesse	M
	Karolina A. Aberg	F
	Britt Reuter Morthorst	F
	Kirsten F. Thompson	F
	Shelly J. Lundberg	F
	G Marckmann	unknown
	David Hägg	M
ROMANIAN	Afrodita Marcu	F
	Iulia T. Simion	F
	Liviu Giosan	M
	Alina Sorescu	F
	Liviu Giosan	M
	Mircea Ivan	M
	Dana Dabelea	F
	Constantin Rezlescu	M
	Christine A. Conelea	F
	R. A. Popescu	unknown
SLAV	Noémi Koczka	F
	Mikhail G Kolonin	M
	Richard Karban	M
	Branislav Dragović	M
	H Illnerová	unknown
	Marte Bjørk	F
	Jacek Niesterowicz	M
	Justin R. Grubich	M
	Mikhail Salama Hend	M
	Snejana Grozeva	F
THAI	Piyamas Kanokwongnuwut	unknown
	Clifton Makate	M
	Noppol Kobmoo	unknown
	Kabkaew L. Sukontason	unknown
	Aroonsiri Sangarlangkarn	unknown
	Yossawan Boriboonthana	unknown
	Ekalak Sitthipornvorakul	unknown
	Tony Rianprakaisang	M
	Apiradee Honglawan	F
	Wonngarm Kittanamongkolchai	unknown
TURKISH	Iris Z. Uras	F
	Metin Gurcan	unknown
	Mustafa Sahmaran	M
	Pinar Akman	F
	Joshua Aslan	M
	Selin Kesebir	F
	Tan Yigitcanlar	unknown
	Thembela Kepe	unknown
	Ulrich Rosar	M
	Selvi C. Ersoy	F
VIETNAMESE	Huong T. T. Ha	unknown
	Vu Van Dung	M
	H ChuongKim	unknown
	Daniel W. Giang	M
	Nhung Thi Nguyen	unknown
	V. Phan	unknown
	Oanh Kieu Nguyen	F
	Phuc T. Ha	M
	Bich Tran	unknown
	Oanh Kieu Nguyen	F
Unknown	Gene Y. Fridman	M

	Judith Glück	F
	Noor Edi Widya Sukoco	unknown
	Charlene Laino	F
	Benoît Bérard	unknown
	David Zünd	M
	Katarzyna Adamala	F
	K.A. Godfrin	unknown
	Shadd Maruna	M
	Mariette DiChristina	F

TABLE S8: The 288 U.S.-based outlets are grouped into 3 categories based on their topics of reports. Note that other 135 U.S.-based outlets, which are not shown in this table, are excluded in our analyses due to technical limitations in accessing sufficient volumes of their content (e.g., view-limited paywalls or anti-crawling mechanisms).

Outlet	**Type**
OnMedica	Sci. & Tech.
Huffington Post	General News
KiiiTV 3	General News
Carbon Brief	Sci. & Tech.
PR Newswire	Press Releases
Nutra Ingredients USA	Sci. & Tech.
The Bellingham Herald	General News
CNN News	General News
Health Medicinet	Press Releases
Herald Sun	General News
EurekAlert!	Press Releases
AJMC	Press Releases
The University Herald	General News
Lincoln Journal Star	General News
Cardiovascular Business	Sci. & Tech.
MinnPost	General News
CNET	Sci. & Tech.
Infection Control Today	Sci. & Tech.
Science 2.0	Sci. & Tech.
Lexington Herald Leader	General News
Statesman.com	General News
Nanowerk	Press Releases
The San Diego Union-Tribune	General News
The Daily Beast	General News
Lab Manager	Press Releases
SDPB Radio	General News
New Hampshire Public Radio	General News
Health Day	Press Releases
Rocket News	General News
KPBS	General News
Technology.org	Press Releases
UPI.com	General News
WUWM	General News
Central Coast Public Radio	General News
The Hill	General News
The Epoch Times	General News
Biospace	Sci. & Tech.
Minyanville: Finance	General News
Nature World News	Sci. & Tech.
New York Post	General News
Action News Now	General News
WUNC	General News
Futurity	Press Releases
Reason	General News
azfamily.com	General News
Idaho Statements	General News
Google News	General News
Tri States Public Radio	General News
American Physical Society - Physics	Press Releases
KTEP El Paso	General News
LiveScience	Sci. & Tech.
KUNC	General News
The Daily Meal	Sci. & Tech.
AOL	General News
Women's Health	Sci. & Tech.

Prevention	Sci. & Tech.
ECN	Sci. & Tech.
Iowa Public Radio	General News
Becker's Hospital Review	Sci. & Tech.
7th Space Family Portal	Press Releases
Springfield News Sun	General News
Environmental News Network	Press Releases
Sky Nightly	Sci. & Tech.
Quartz	Sci. & Tech.
Benzinga	General News
Headlines & Global News	General News
The Denver Post	General News
Science Daily	Press Releases
The Advocate	General News
ABC News	General News
Newswise	Press Releases
hellogiggles.com	General News
WLRN	General News
EarthSky	Sci. & Tech.
Becker's Spine Review	Sci. & Tech.
MIT News	Press Releases
MarketWatch	General News
Arstechnica	Sci. & Tech.
Journalist's Resource	Sci. & Tech.
Northern Public Radio	General News
Everyday Health	Sci. & Tech.
Star Tribune	General News
TCTMD	Sci. & Tech.
The Verge	General News
She Knows	General News
SeedQuest	Sci. & Tech.
Tech Times	Sci. & Tech.
Witchita's Public Radio	General News
Oncology Nurse Advisor	Sci. & Tech.
Delmarva Public Radio	General News
Medical Daily	Sci. & Tech.
Homeland Security News Wire	General News
Discover Magazine	Sci. & Tech.
Washington Post	General News
MSN	General News
Hawaii News Now	General News
The Daily Caller	General News
News Tribune	General News
The Fresno Bee	General News
King 5	General News
Star-Telegram	General News
CNBC	General News
Salon	General News
WJCT	General News
WVPE	General News
KTEN	General News
Wired.com	General News
Daily Kos	General News
USA Today	General News
Men's Health	Sci. & Tech.
Boise State Public Radio	General News
Voice of America	General News
PR Web	Press Releases
Georgia Public Radio	General News
FiveThirtyEight	General News
Public Radio International	General News

Harvard Business Review	General News
Inverse	General News
Doctors Lounge	Sci. & Tech.
North East Public Radio	General News
The Charlotte Observer	General News
National Geographic	Sci. & Tech.
Pharmacy Times	Sci. & Tech.
Popular Science	Sci. & Tech.
ABC Action News WFTS Tampa Bay	General News
News Channel	General News
The University of New Orleans Public Radio	General News
Mic	General News
Health Canal	Sci. & Tech.
KOSU	General News
Raleigh News and Observer	General News
The Atlantic	General News
newsmax.com	General News
Yahoo! Finance USA	General News
Government Executive	General News
International Business Times	General News
Emaxhealth.com	Press Releases
Newsweek	General News
FOX News	General News
The New York Observer	General News
Sign of the Times	General News
The Inquisitr	General News
ABC News 15 Arizona	General News
Parent Herald	General News
The ASCO Post	Sci. & Tech.
Clinical Advisor	Sci. & Tech.
Slate Magazine	General News
NPR	General News
Health	Sci. & Tech.
Dayton Daily News	General News
Guardian Liberty Voice	General News
Belleville News-Democrat	General News
Yahoo! News	General News
WCBE	General News
Buzzfeed	General News
Sci-News	Sci. & Tech.
The Seattle Times	General News
Philly.com	General News
Renal & Urology News	Sci. & Tech.
Arizona Public Radio	General News
Interlochen Public Radio	General News
12 News KBMT	General News
New York Magazine	General News
Medium US	General News
KPCC : Southern California Public Radio	General News
2 Minute Medicine	Sci. & Tech.
Pediatric News	Sci. & Tech.
redOrbit	Sci. & Tech.
Insurance News Net	General News
Drug Discovery and Development	Sci. & Tech.
USNews.com	General News
Yahoo!	General News
The Body	Sci. & Tech.
GEN	Sci. & Tech.
Pacific Standard	General News
Northwest Indiana Times	General News
Psychology Today	Sci. & Tech.

Oregon Public Broadcasting	General News
Mother Nature Network	Sci. & Tech.
Pressfrom	General News
Physician’s Weekly	Sci. & Tech.
Pettinga: Stock Market	General News
Winona Daily News	General News
Runner’s World	Sci. & Tech.
Bio-Medicine.org	Press Releases
Alternet	General News
Mother Jones	General News
The Wichita Eagle	General News
Cornell Chronicle	Press Releases
Politico Magazine	General News
Equities.com	General News
WBUR	General News
ABC 7 WKBW Buffalo	General News
Billings Gazette	General News
My Science	Sci. & Tech.
The Week	General News
BioTech Gate	Sci. & Tech.
Kansas City Star	General News
The Deseret News	General News
PBS	General News
Space.com	Sci. & Tech.
Astrobiology Magazine	Sci. & Tech.
Outside	General News
Value Walk	General News
WYPR	General News
Bustle	General News
Science World Report	Sci. & Tech.
Inside Science	Sci. & Tech.
Science Alert	Sci. & Tech.
Breitbart News Network	General News
St. Louis Post-Dispatch	General News
HowStuffWorks	General News
Wyoming Public Radio	General News
UBM Medica	Sci. & Tech.
Fight Aging!	Sci. & Tech.
MIT Technology Review	Sci. & Tech.
WVXU	General News
The Ecologist	Sci. & Tech.
Alaska Despatch News	General News
Health Imaging	Sci. & Tech.
Kansas City University Radio	General News
Christian Science Monitor	General News
Medicinenet	Sci. & Tech.
WTOP	General News
Business Insider	General News
Real Clear Science	Sci. & Tech.
Counsel & Heal	Sci. & Tech.
The Raw Story	General News
Medcity News	Sci. & Tech.
Drugs.com	Sci. & Tech.
Relief Web	Press Releases
SPIE Newsroom	Sci. & Tech.
New York Daily News	General News
Newser	General News
The Sacramento Bee	General News
Vice	General News
R&D	Sci. & Tech.
KCENG12	Sci. & Tech.

Inc.	General News
Science/AAAS	Sci. & Tech.
The Atlanta Journal Constitution	General News
Brookings	General News
Common Dreams	General News
Physician's Briefing	Press Releases
KERA News	General News
Space Daily	Sci. & Tech.
Tech Xplore	Sci. & Tech.
US News Health	Sci. & Tech.
KUOW	General News
WRKF	General News
TIME Magazine	General News
Smithsonian Magazine	Sci. & Tech.
Herald Tribune	General News
Lifehacker	General News
Fast Company	General News
Kansas Public Radio	General News
Omaha Public Radio	General News
New York Times	General News
Technology Networks	Sci. & Tech.
Elite Daily	General News
Centre for Disease Research and Policy	Sci. & Tech.
Business Wire	General News
KUNM	General News
CBS News	General News
Scientific American	Sci. & Tech.
NBC News	General News
Sun Herald	General News
KRWG TV/FM	General News
TODAY	General News
Radio Acadie	General News
The Columbian	General News
Houston Chronicle	General News
WABE	General News
The Modesto Bee	General News
American Council on Science and Health	Sci. & Tech.
WKAR	General News
Psych Central	Sci. & Tech.
WebMD News	Sci. & Tech.
Green Car Congress	Sci. & Tech.
ABC News WMUR 9	General News
Healthline	Sci. & Tech.
Mongabay	Sci. & Tech.
Vox.com	General News
WPTV 5 West Palm Beach	General News
Popular Mechanics	Sci. & Tech.
PM 360	Sci. & Tech.
SFGate	General News
Seed Daily	Sci. & Tech.

TABLE S9: The coefficients of all variables (including 199 keywords) in Model 5 in predicting whether the author is mentioned by name in a news story referencing a research paper. Random effects for 288 outlets and 8,261 publication venues are also included in the model.

	Dependent variable:	
	Is author mentioned	
Author ethnicity African	−0.366	p = 0.000
Author ethnicity Chinese	−0.376	p = 0.000
Author ethnicity non-Chinese East Asian	−0.272	p = 0.000
Author ethnicity Eastern European	−0.009	p = 0.653
Author ethnicity Indian	−0.011	p = 0.560
Author ethnicity Middle Eastern	0.016	p = 0.366
Author ethnicity Southern European	−0.138	p = 0.000
Author ethnicity Western & Northern Euro.	−0.072	p = 0.000
Author ethnicity Unknown	−0.227	p = 0.00002
Author gender Female	0.003	p = 0.695
Author gender Unknown	−0.113	p = 0.000
Reporter ethnicity Asian	−0.051	p = 0.176
Reporter ethnicity European	−0.033	p = 0.095
Reporter ethnicity Other Unknown	0.054	p = 0.047
Reporter gender Female	−0.015	p = 0.405
Reporter gender Unknown	0.015	p = 0.560
Last name length	−0.010	p = 0.000
Last name frequency	0.004	p = 0.028
First author position	0.397	p = 0.000
Middle author position	−0.814	p = 0.000
Is the paper solo authored	0.683	p = 0.000
Author rank	−0.0001	p = 0.000
Not a top author	−0.090	p = 0.004
Not a corresponding author	−1.448	p = 0.000
Corresponding status unknown	−0.506	p = 0.000
Affiliation rank	−0.00004	p = 0.000
Affiliation international (location)	−0.307	p = 0.000
Affiliation unknown (location)	0.056	p = 0.571
Number of authors in the paper	−0.007	p = 0.000
Year of news story (mention year)	−0.051	p = 0.000
Year gap between story and paper	−0.145	p = 0.000
News story length	0.0002	p = 0.000
Num. of papers mentioned in a story	−0.120	p = 0.000
Flesch-Kincaid score	−0.001	p = 0.000
Sentences per paragraph	0.008	p = 0.00002
Type-Token ratio	0.300	p = 0.00000
Cell biology	0.301	p = 0.00000
Genetics	0.001	p = 0.980
Biology	0.032	p = 0.701
Body mass index	−0.329	p = 0.00001
Health care	−0.183	p = 0.0005
Disease	−0.103	p = 0.018
Gerontology	−0.607	p = 0.000
Population	−0.103	p = 0.00003
Public health	−0.165	p = 0.004
Medicine	−0.361	p = 0.00001
Materials science	0.352	p = 0.001
Composite material	0.162	p = 0.188
Nanotechnology	0.255	p = 0.007
Cohort study	−0.009	p = 0.861
Social psychology	−0.154	p = 0.006
Cohort	0.069	p = 0.155
Psychological intervention	0.009	p = 0.879

Young adult	−0.309	p = 0.00000
Family medicine	−0.306	p = 0.00001
Cancer	−0.097	p = 0.038
Surgery	−0.019	p = 0.779
Randomized controlled trial	−0.095	p = 0.062
Placebo	0.019	p = 0.790
Clinical trial	−0.105	p = 0.190
Nursing	−0.288	p = 0.002
Applied psychology	−0.425	p = 0.011
Human factors and ergonomics	−0.220	p = 0.061
Injury prevention	0.335	p = 0.002
Suicide prevention	0.003	p = 0.978
Psychiatry	−0.362	p = 0.000
Occupational safety and health	−0.471	p = 0.00002
Intensive care medicine	−0.286	p = 0.001
Pediatrics	−0.241	p = 0.0003
Hazard ratio	0.266	p = 0.00001
Confidence interval	−0.147	p = 0.020
Retrospective cohort study	0.148	p = 0.039
Vaccination	0.059	p = 0.493
Psychology	0.078	p = 0.384
Perception	0.185	p = 0.021
Cognition	−0.117	p = 0.034
Environmental health	−0.347	p = 0.00000
Obesity	−0.203	p = 0.003
Risk factor	0.236	p = 0.001
Quality of life	−0.035	p = 0.702
Physical therapy	−0.094	p = 0.095
Weight loss	−0.357	p = 0.0001
Anatomy	0.625	p = 0.000
Mental health	0.140	p = 0.030
Psychosocial	0.271	p = 0.011
Anxiety	−0.334	p = 0.00000
Distress	0.269	p = 0.012
Business	−0.660	p = 0.00001
Public relations	−0.244	p = 0.023
Marketing	0.168	p = 0.295
Immunology	−0.164	p = 0.007
Global warming	−0.100	p = 0.178
Economics	−0.040	p = 0.741
Climatology	−0.254	p = 0.003
Climate change	−0.461	p = 0.000
General surgery	0.008	p = 0.960
Endocrinology	−0.154	p = 0.007
Internal medicine	0.341	p = 0.000
Receptor	−0.160	p = 0.055
Inflammation	0.199	p = 0.019
Stimulus physiology	0.091	p = 0.390
Immune system	0.132	p = 0.050
Meta analysis	−0.696	p = 0.000
Sociology	0.371	p = 0.008
Gene	−0.131	p = 0.031
Cancer research	−0.025	p = 0.705
Breast cancer	0.075	p = 0.230
Cell	0.385	p = 0.00001
Diabetes mellitus	−0.062	p = 0.159
Blood pressure	−0.127	p = 0.177
Oncology	−0.172	p = 0.049
Gynecology	−0.338	p = 0.006
Communication	0.319	p = 0.006
Cognitive psychology	0.002	p = 0.983

Adverse effect	−0.092	p = 0.208
Clinical endpoint	−0.626	p = 0.000
Pharmacology	−0.392	p = 0.0001
Virology	−0.330	p = 0.0001
Risk assessment	0.250	p = 0.021
Transcription factor	0.383	p = 0.0001
Political science	−0.280	p = 0.054
Ecology	0.062	p = 0.270
Geography	0.018	p = 0.864
Cross sectional study	−0.024	p = 0.792
Odds ratio	−0.114	p = 0.040
Comorbidity	−0.136	p = 0.209
Environmental engineering	−0.452	p = 0.005
Chemistry	0.097	p = 0.320
Medical emergency	−0.711	p = 0.000
Physics	0.131	p = 0.214
Social science	0.448	p = 0.008
Ethnic group	0.018	p = 0.848
Labour economics	0.380	p = 0.015
Antibody	0.274	p = 0.008
Geomorphology	−0.160	p = 0.102
Geophysics	0.081	p = 0.461
Geology	−0.312	p = 0.002
Ranging	−0.113	p = 0.215
Stroke	−0.003	p = 0.974
Environmental resource management	−0.132	p = 0.203
Type 2 diabetes	0.169	p = 0.053
Cardiology	0.066	p = 0.502
Molecular biology	0.169	p = 0.007
Developmental psychology	−0.043	p = 0.499
Agriculture	−0.393	p = 0.00002
Signal transduction	−0.188	p = 0.053
Optoelectronics	−0.047	p = 0.651
Psychotherapist	−0.413	p = 0.004
Affect psychology	−0.319	p = 0.003
Clinical psychology	−0.036	p = 0.622
Anesthesia	−0.311	p = 0.001
Atmospheric sciences	−0.029	p = 0.774
In vivo	−0.117	p = 0.192
Biochemistry	0.0001	p = 0.999
Analytical chemistry	−0.078	p = 0.553
Neuroscience	0.310	p = 0.00001
Botany	−0.292	p = 0.015
Gene expression	0.242	p = 0.017
Politics	0.170	p = 0.070
Demography	0.339	p = 0.000
Socioeconomic status	−0.345	p = 0.00004
Mortality rate	−0.225	p = 0.002
Virus	0.066	p = 0.494
Optics	0.411	p = 0.0004
Condensed matter physics	−0.591	p = 0.000
Bioinformatics	−0.510	p = 0.00001
Law	−0.111	p = 0.494
Physical medicine and rehabilitation	−0.086	p = 0.583
Stem cell	−0.056	p = 0.496
Biodiversity	−0.167	p = 0.022
Astrophysics	−1.033	p = 0.000
Astronomy	−0.203	p = 0.041
Radiology	−0.400	p = 0.007
Pathology	−0.014	p = 0.858
Proportional hazards model	−0.137	p = 0.108

Chemotherapy	−0.662	p = 0.00000
Predation	−0.196	p = 0.029
Food science	−0.300	p = 0.034
Artificial intelligence	1.100	p = 0.00002
Overweight	−0.049	p = 0.571
Antibiotics	−0.043	p = 0.710
Microbiology	0.143	p = 0.173
Zoology	0.280	p = 0.002
Paleontology	0.200	p = 0.016
Habitat	0.546	p = 0.000
Public administration	0.924	p = 0.00001
Ecosystem	−0.062	p = 0.424
Economic growth	0.095	p = 0.450
Organic chemistry	0.254	p = 0.100
Government	−0.135	p = 0.199
Autism	−0.140	p = 0.133
Transplantation	0.250	p = 0.003
Gastroenterology	−0.297	p = 0.022
Insulin	0.018	p = 0.849
Engineering	−0.268	p = 0.133
Computer science	0.072	p = 0.529
Observational study	−0.154	p = 0.111
Heart disease	0.021	p = 0.836
Epidemiology	−0.106	p = 0.104
Obstetrics	0.158	p = 0.133
Pregnancy	−0.140	p = 0.040
Fishery	0.026	p = 0.839
Alternative medicine	−0.243	p = 0.041
Logistic regression	0.385	p = 0.00003
Offspring	0.196	p = 0.031
Mood	−0.287	p = 0.002
Bacteria	0.127	p = 0.248
Prostate cancer	−0.400	p = 0.00004
Evolutionary biology	0.130	p = 0.114
Phenomenon	0.022	p = 0.821
Longitudinal study	0.027	p = 0.758
Genome	0.088	p = 0.191
Mutation	0.204	p = 0.012
Pedagogy	−0.283	p = 0.101
Dementia	−0.186	p = 0.046
Relative risk	0.121	p = 0.109
Microeconomics	0.536	p = 0.003
Odds	0.004	p = 0.968
Feeling	0.451	p = 0.00004
Oceanography	−0.095	p = 0.376
Emergency medicine	0.029	p = 0.759
Personality	−0.023	p = 0.804
Prospective cohort study	−0.212	p = 0.0003
Hippocampus	−0.046	p = 0.650
Greenhouse gas	0.006	p = 0.948
Biomarker medicine	0.409	p = 0.00002
Myocardial infarction	−0.135	p = 0.140
Socioeconomics	0.297	p = 0.015
Drug	0.290	p = 0.004
Environmental science	−0.368	p = 0.0003
Epigenetics	−0.382	p = 0.0002
Inorganic chemistry	−0.233	p = 0.020
Emergency department	−0.205	p = 0.028
Medical prescription	0.270	p = 0.002
Phenotype	0.076	p = 0.450
Constant	0.968	p = 0.000

Observations	524,052
Log Likelihood	-255,530.5
Akaike Inf. Crit.	511,537.0
Bayesian Inf. Crit.	514,195.3