Supplementary materials for

Longitudinal Analyses of Gender Differences in First Authorship Publications Related to COVID-19

Carolin Lerchenmüller, MD, Leo Schmallenbach, MSc,
Anupam B. Jena, MD, PhD, Marc J. Lerchenmüller, MPH, PhD

Table of Contents

Additional Information on Data

We merged several databases to analyze potential gender differences in first authorships of COVID publications relative to a set of control publications in the same journals and within the same time period one year earlier. First, we extracted all articles from the PubMed database for which the term "COVID" appeared in the title or abstract and obtained all available article characteristics including, among others, the names of all authors, country affiliation per author, the journal ISSN (International Standard Serial Number), and time of publication (months and year). The U.S. National Library of Medicine maintains the PubMed XML database and a detailed data inventory can be found online (https://www.nlm.nih.gov/databases/download/pubmed_medline.html). We obtained the journals' major scientific discipline from the Clarivate Journal Citation Report of 2018 via the unique journal ISSNs. We used journal names as a crosswalk to identify publications that appeared a year earlier in the exact same journals as the COVID articles.

An overview of the sample creation is provided in **Figure S1**. In service of estimation accuracy, we included only journals that are listed in Clarivate. By construction that excludes all COVID publications in journals that had no publication on record in PubMed for 2019. These journals likely only came into being in 2020. We restricted our search query to articles published between February 1st of 2020 and January 31st of 2021, since these months were the most productive in terms of COVID publishing and we sought to mitigate seasonal influences, like gender differences in teaching load at certain times of year.

We used the forenames recorded in PubMed to designate the gender of authors (PubMed started to systematically record forenames in 2002). We determined the probable gender of the authors through the Genderize database, an established approach that allows gender assignment for a large number of authors. At the time of initial submission, Genderize included 86,710 distinct forenames drawn from 74 countries and 81 languages. Recent tests of the accuracy and comprehensiveness of four gender assignment algorithms, using a control sample of gender-matched forenames from a US government office, found that Genderize provided the most accurate estimates of gender (*1*). Our underlying code for calling the Genderize database with a large set of forenames has been posted to Figshare (*2*). Genderize uses a variety of information, such as social media records, to assign a probability that an individual with a particular forename is a man or a woman. For example, Genderize designates the forename "Chris" as male with 93% probability based on 8,631 verified records in the database. We considered gender determined if Genderize assigned a probability of greater than 90%. Applying this threshold, we designated the gender for more than 72% of the authors in our dataset. However, there is variation across author origins (**Figure S2**). For example, we designated the gender for 84% of authors with an affiliation from North America and for 52% of authors with an affiliation from Asia. The lower accuracy for authors from Asia is a common challenge in name-based gender designation and a limitation to our analysis of authors from these countries. Yet, there is no difference in the accuracy of gender designation across men and women authors (**Figure S3**) or COVID and non-COVID articles (**Figure S4**). Hence, there is no reason to be concerned that the gender designation would systematically bias our results. Additionally, our main findings do not change when setting different gender designation thresholds.

Next, we compared the distribution of disciplines producing COVID research relative to the articles in the control sample (**Table S1**). Ranking the disciplines in terms of publication output, and testing a Spearman Rank correlation, we obtain a coefficient of greater 0.80. While this correlation would generally be considered strong (*3*) lending credence to our basic design, it does not consider the possibility that men and women may sort differently into these fields. However, our **Figure 2** in the main text documents that it is primarily fields where women tend to be well represented that produce COVID research.

To execute country-level analyses, we use regular expressions to extract the full country name or country codes from affiliation data for the first author. We also ranked countries by productivity for COVID-articles and control articles, obtaining a Spearman rank correlation of 0.94, again supporting our approach of using non-COVID articles in the prior year as a control group (**Table S2**). This also mitigates concerns that countries with larger gender gaps in general produce more COVID research.

Additional information on methods

Measurement

To assess the effect of the COVID pandemic on the gender gap in publishing, we reported unadjusted differences in the percent of women first authorships versus male first authorships for COVID and non-COVID publications. This straightforward metric provides a direct and easy to understand measure of how the COVID pandemic impacts women's versus men's publication productivity.

$$\begin{aligned}\Delta GenderGap &= \{FirstAuthor_{Female} - FirstAuthor_{Male} \mid COVID\} \\ &\quad - \{FirstAuthor_{Female} - FirstAuthor_{Male} \mid Non-COVID\}\end{aligned}$$

To conduct subgroup analysis for discipline and country, we calculated the change in the gender gap based on the percent of first authorships by men and women for the specific discipline and country.

Estimation

In addition to the unadjusted differences, we also provided adjusted differences in first authorships from women and men obtained from linear probability models (**Table S3**), adjusting for the number of authors on a publication, the month of publication, the field of research and country. We run the same analysis for last authorships from women and men (**Table S4**). Both regression analyses support the descriptive evidence presented in Figure 1 of the main text. Logistic regression as an alternative estimation model has two disadvantages in our analysis. First, the large number of fixed effects when including countries and discipline dummies, for example, raises the possibility of incidental parameters bias and could prevent the convergence of some of our models. Second, logistic regressions can overestimate effect sizes as a result of the high leverage of marginal cases (i.e., identifying larger gender differences than reported in the main text), whereas linear probability models average across observations and produce more conservative results (see also **Table S5**).

We provided adjusted estimates in the supplement as one might be concerned, for example, that men are more numerous in fields that produce COVID research. This would also lower women's observed COVID productivity but not due to pandemic related constraints as hypothesized, but rather due to underlying structural differences in subspecialties. Of note, the descriptive data paint a different picture, such that women tend to be at least equal if not overrepresented in the most productive COVID disciplines.

We conducted four robustness checks to establish the reliability of our findings (**Table S6**). In the first two robustness checks, we vary the threshold applied to the accuracy of the gender designation. In Model 1, we consider all authors, for which gender was assigned with a probability higher than chance (>50%). In Model 2, we only consider authors, for which the gender designation accuracy was reported with 100%. Both models show very similar estimates for the decrease in women authorship on COVID publications (8.2%-points vs. 9.0%-points). Next, we excluded articles from the analysis, for which collective authorship was indicated in PubMed. This concerns roughly 8% of articles but excluding them does not alter the effect estimate. Last, we reran the analysis on the full sample, that is including COVID articles published in journals, which are not listed in Clarivate's journal citation report and for which the first author's gender could be designated. As we do not know the disciplines these journals fall into, we include journal instead of discipline fixed effects in this last model specification. Again, the results are consistent with our previous analysis. Accordingly, a descriptive comparison of the articles in- and excluded from the analysis shows that they are near identical with respect to the representation of women first and last authors (**Table S7**).

References

1. C. N. G. detection, (http://codingnews.info/post/genderdetection (accessed 11/15/17) (2015)).
2. Lerchenmueller. Marc, *Genderize_unlimited_API_request*. (2017).
3. M. G. Pagano, Kimberlee, *Principle of Biostatistics*. (Brooks/Cole, ed. Secon Edition, 2000).

Figure S1: Sample construction for COVID articles and non-COVID (control) articles

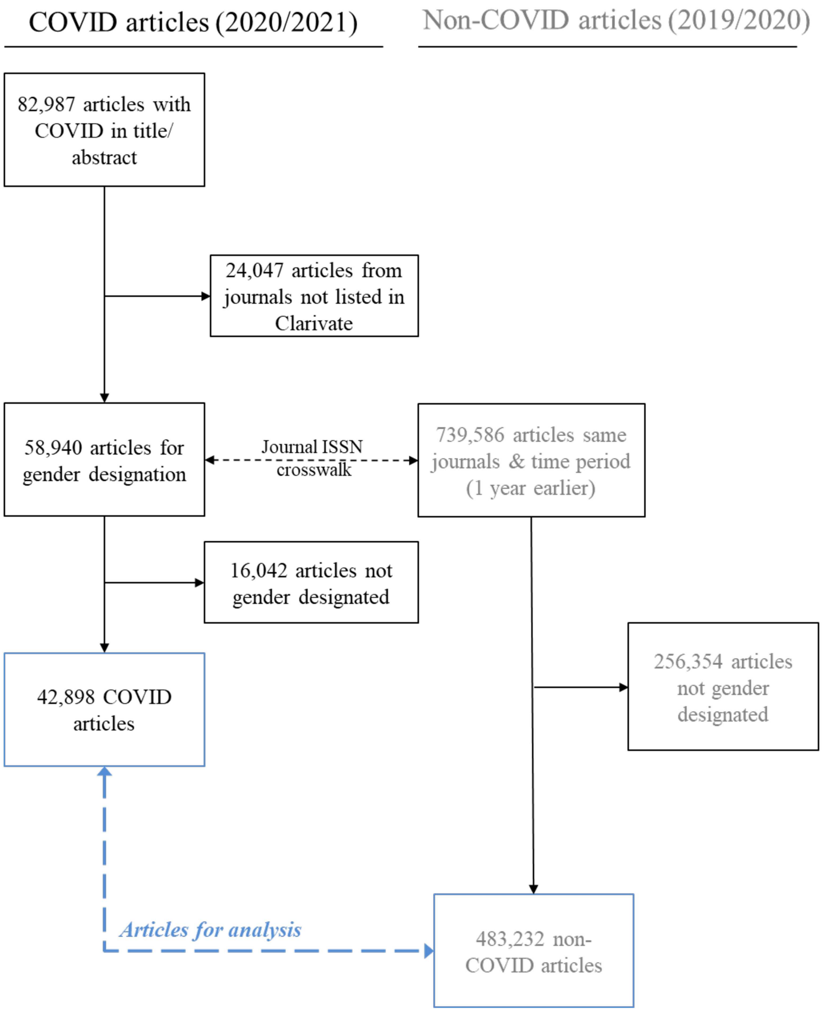

COVID articles (2020/2021)
Non-COVID articles (2019/2020)
82,987 articles with COVID in title/ abstract
24,047 articles from journals not listed in Clarivate
58,940 articles for gender designation
Journal ISSN crosswalk
739,586 articles same journals & time period (1 year earlier)
16,042 articles not gender designated
42,898 COVID articles
256,354 articles not gender designated
Articles for analysis
483,232 non-COVID articles

Figure S2: Gender designation accuracy for first authors from North America, Latin America, Europe, Asia, Oceania, and Africa separately

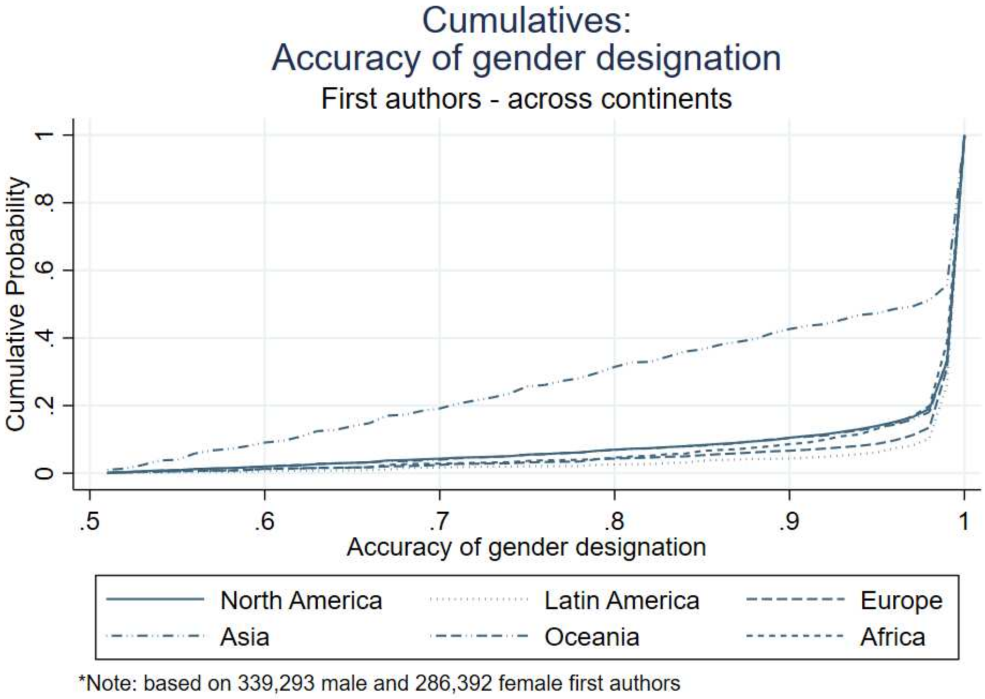
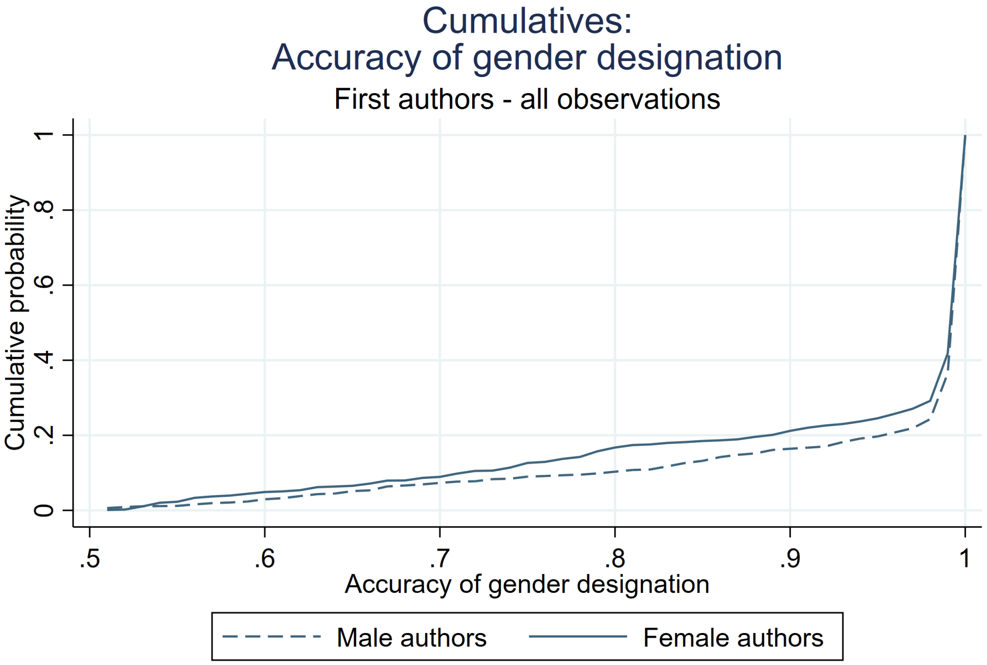


Figure S3: Gender designation accuracy for all articles in the sample (COVID articles and non-COVID articles) for women and men first authors



*Note: based on 349,489 male 296,937 female first authors

Figure S4: Gender designation accuracy for women first authors for COVID articles and non-COVID (control) articles separately

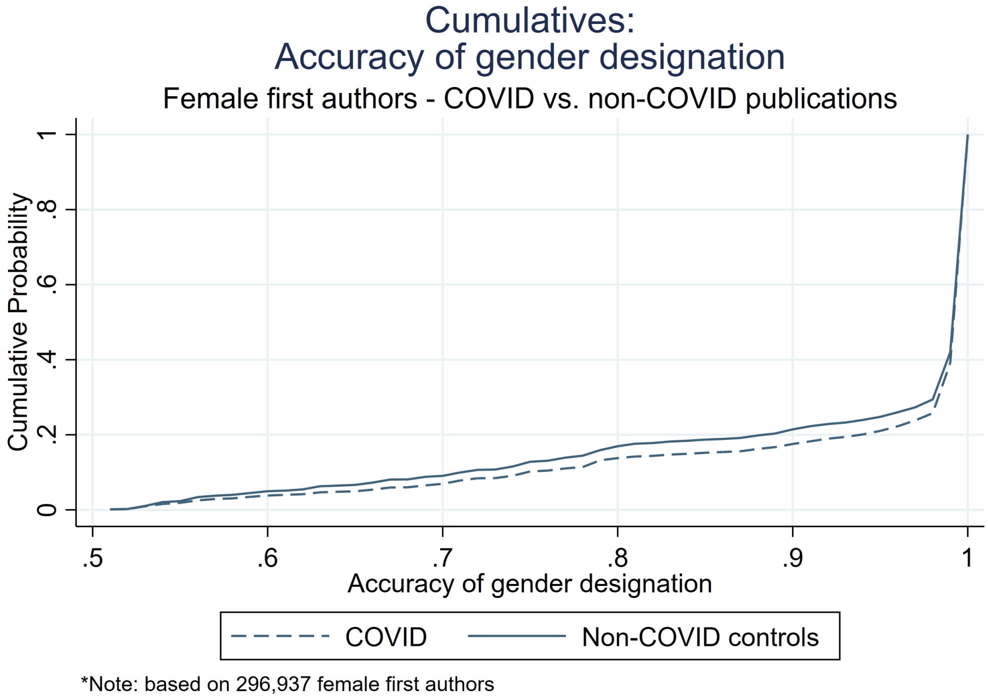


Table S1: Major disciplines for COVID articles and non-COVID (control) articles and concordance statistic

Discpline	**Non-COVID**			**COVID**			**Total**		
	Total	%	Rank	Total	%	Rank	Total	%	Rank
MEDICINE GENERAL & INTERNAL	23,163	4.79%	5	4,229	9.88%	1	27,392	5.21%	4
PUBLIC ENVIRONMENTAL & OCCUPATIONAL HEALTH	17,860	3.70%	8	3,530	8.24%	2	21,390	4.07%	6
SURGERY	26,457	5.48%	3	2,254	5.26%	3	28,711	5.46%	3
IMMUNOLOGY	11,151	2.31%	12	1,925	4.50%	4	13,076	2.49%	11
CARDIAC & CARDIOVASCULAR SYSTEMS	16,196	3.35%	9	1,479	3.45%	5	17,675	3.36%	9
PHARMACOLOGY & PHARMACY	18,408	3.81%	7	1,439	3.36%	6	19,847	3.77%	8
MULTIDISCIPLINARY SCIENCES	38,524	7.97%	1	1,435	3.35%	7	39,959	7.60%	1
PSYCHIATRY	9,890	2.05%	14	1,333	3.11%	8	11,223	2.13%	14
ENVIRONMENTAL SCIENCES	19,607	4.06%	6	1,278	2.98%	9	20,885	3.97%	7
ONCOLOGY	23,207	4.80%	4	1,159	2.71%	10	24,366	4.63%	5
BIOCHEMISTRY & MOLECULAR BIOLOGY	30,224	6.25%	2	1,118	2.61%	11	31,342	5.96%	2
PEDIATRICS	10,280	2.13%	13	1,105	2.58%	12	11,385	2.16%	13
INFECTIOUS DISEASES	3,143	0.65%	41	1,001	2.34%	13	4,144	0.79%	37
CLINICAL NEUROLOGY	9,614	1.99%	15	922	2.15%	14	10,536	2.00%	15
HEALTH CARE SCIENCES & SERVICES	5,209	1.08%	29	921	2.15%	15	6,130	1.17%	27
MEDICINE RESEARCH & EXPERIMENTAL	5,645	1.17%	25	887	2.07%	16	6,532	1.24%	25
NEUROSCIENCES	14,611	3.02%	10	835	1.95%	17	15,446	2.94%	10
DERMATOLOGY	5,452	1.13%	26	777	1.81%	18	6,229	1.18%	26
VIROLOGY	2,198	0.45%	47	771	1.80%	19	2,969	0.56%	43
RESPIRATORY SYSTEM	4,074	0.84%	35	732	1.71%	20	4,806	0.91%	32
GASTROENTEROLOGY & HEPATOLOGY	6,367	1.32%	23	719	1.68%	21	7,086	1.35%	22
RADIOLOGY NUCLEAR MEDICINE & MEDICAL IMAGING	7,092	1.47%	20	715	1.67%	22	7,807	1.48%	19
ENDOCRINOLOGY & METABOLISM	6,468	1.34%	22	628	1.47%	23	7,096	1.35%	21
HEMATOLOGY	4,449	0.92%	31	618	1.44%	24	5,067	0.96%	31
UROLOGY & NEPHROLOGY	7,369	1.52%	18	581	1.36%	25	7,950	1.51%	18
ANESTHESIOLOGY	2,100	0.43%	48	519	1.21%	26	2,619	0.50%	44
MICROBIOLOGY	6,086	1.26%	24	499	1.17%	27	6,585	1.25%	24
NURSING	4,602	0.95%	30	484	1.13%	28	5,086	0.97%	30
EMERGENCY MEDICINE	1,958	0.41%	52	459	1.07%	29	2,417	0.46%	48
RHEUMATOLOGY	3,146	0.65%	40	455	1.06%	30	3,601	0.68%	40
PSYCHOLOGY MULTIDISCIPLINARY	3,271	0.68%	38	413	0.96%	31	3,684	0.70%	39
OPHTHALMOLOGY	5,303	1.10%	27	412	0.96%	32	5,715	1.09%	28
GERIATRICS & GERONTOLOGY	1,818	0.38%	55	392	0.92%	33	2,210	0.42%	52
OBSTETRICS & GYNECOLOGY	4,388	0.91%	32	353	0.82%	34	4,741	0.90%	33
CELL BIOLOGY	8,490	1.76%	16	352	0.82%	35	8,842	1.68%	16
OTORHINOLARYNGOLOGY	2,098	0.43%	49	303	0.71%	36	2,401	0.46%	49
CRITICAL CARE MEDICINE	2,018	0.42%	50	282	0.66%	37	2,300	0.44%	50
DENTISTRY ORAL SURGERY & MEDICINE	3,212	0.66%	39	275	0.64%	38	3,487	0.66%	41
ECONOMICS	495	0.10%	81	274	0.64%	39	769	0.15%	74
BIOTECHNOLOGY & APPLIED MICROBIOLOGY	4,021	0.83%	36	246	0.57%	40	4,267	0.81%	34
NUTRITION & DIETETICS	5,241	1.08%	28	229	0.53%	41	5,470	1.04%	29
PSYCHOLOGY CLINICAL	2,392	0.50%	43	210	0.49%	42	2,602	0.49%	45
ETHICS	674	0.14%	75	195	0.46%	43	869	0.17%	71
MEDICAL LABORATORY TECHNOLOGY	1,217	0.25%	63	185	0.43%	44	1,402	0.27%	61
SPORT SCIENCES	3,065	0.63%	42	182	0.43%	45	3,247	0.62%	42
PATHOLOGY	2,011	0.42%	51	181	0.42%	46	2,192	0.42%	53
GENETICS & HEREDITY	7,184	1.49%	19	178	0.42%	47	7,362	1.40%	20
ORTHOPEDICS	2,345	0.49%	44	168	0.39%	48	2,513	0.48%	46
PERIPHERAL VASCULAR DISEASE	1,700	0.35%	57	133	0.31%	49	1,833	0.35%	56
MATHEMATICS INTERDISCIPLINARY APPLICATIONS	70	0.01%	120	126	0.29%	50	196	0.04%	104

Spearman Rank Correlation - all disciplines	
coefficient (rs)	0.807
N	148
T statistic	16.537
DF	146
p-value	0.000

Spearman Rank Correlation - top 50 disciplines	
coefficient (rs)	0.738
N	50
T statistic	7.588
DF	48
p-value	0.000

Table S2: Major countries for COVID articles and non-COVID (control) articles and concordance statistics

Country (First Author)	**Non-COVID**			**COVID**			**Total**		
	Total	%	Rank	Total	%	Rank	Total	%	Rank
United States	120,478	26.65%	1	11,066	26.12%	1	131,544	26.75%	1
Italy	22,670	5.01%	5	4,309	10.15%	2	26,979	5.49%	4
United Kingdom	27,994	6.19%	3	3,157	8.13%	3	31,151	6.33%	2
India	11,388	2.52%	12	1,778	7.43%	4	13,166	2.68%	12
Spain	15,273	3.38%	9	1,672	4.19%	5	16,945	3.45%	9
China	29,078	6.43%	2	1,377	3.77%	6	30,455	6.19%	3
Canada	16,654	3.68%	7	1,324	3.11%	7	17,978	3.66%	7
France	13,837	3.06%	10	1,313	3.02%	8	15,150	3.08%	10
Germany	24,413	5.40%	4	1,161	2.64%	9	25,574	5.20%	5
Brasil	13,219	2.92%	11	1,160	2.64%	10	14,379	2.92%	11
Australia	16,132	3.57%	8	1,052	2.50%	11	17,184	3.49%	8
Iran	6,757	1.49%	15	878	2.07%	12	7,635	1.55%	14
Turkey	6,134	1.36%	18	851	1.97%	13	6,985	1.42%	16
Japan	18,952	4.19%	6	499	1.30%	14	19,451	3.96%	6
Netherlands	9,702	2.15%	13	425	1.24%	15	10,127	2.06%	13
Switzerland	6,455	1.43%	16	422	1.00%	16	6,877	1.40%	17
Singapore	1,823	0.40%	36	405	0.99%	17	2,228	0.45%	32
Israel	3,813	0.84%	22	378	0.96%	18	4,191	0.85%	22
Saudi Arabia	2,020	0.45%	33	341	0.96%	19	2,361	0.48%	30
Greece	2,605	0.58%	27	323	0.86%	20	2,928	0.60%	26
Belgium	4,132	0.91%	21	314	0.82%	21	4,446	0.90%	21
Pakistan	1,769	0.39%	38	286	0.80%	22	2,055	0.42%	36
Mexico	3,186	0.70%	24	281	0.74%	23	3,467	0.71%	24
Egypt	2,541	0.56%	28	254	0.71%	24	2,795	0.57%	28
Poland	6,318	1.40%	17	253	0.66%	25	6,571	1.34%	18
Hong Kong	1,285	0.28%	41	248	0.63%	26	1,533	0.31%	41
Ireland	2,100	0.46%	31	230	0.59%	27	2,330	0.47%	31
Austria	3,313	0.73%	23	203	0.58%	28	3,516	0.71%	23
South Korea	7,400	1.64%	14	199	0.52%	29	7,599	1.55%	15
Sweden	5,841	1.29%	19	192	0.45%	30	6,033	1.23%	19
Bangladesh	318	0.07%	60	154	0.45%	31	472	0.10%	55
Portugal	3,007	0.67%	25	151	0.36%	32	3,158	0.64%	25
Denmark	4,599	1.02%	20	143	0.35%	33	4,742	0.96%	20
South Africa	1,771	0.39%	37	142	0.35%	34	1,913	0.39%	38
United Arab Emirates	532	0.12%	51	129	0.34%	35	661	0.13%	50
Colombia	869	0.19%	45	120	0.33%	36	989	0.20%	45
Chile	1,520	0.34%	40	114	0.29%	37	1,634	0.33%	40
Taiwan	2,101	0.46%	30	112	0.27%	38	2,213	0.45%	33
Norway	2,812	0.62%	26	106	0.26%	39	2,918	0.59%	27
Malaysia	1,224	0.27%	42	94	0.25%	40	1,318	0.27%	42
Peru	346	0.08%	58	93	0.22%	41	439	0.09%	56
Argentina	1,739	0.38%	39	89	0.20%	42	1,828	0.37%	39
Romania	1,142	0.25%	43	84	0.20%	43	1,226	0.25%	43
Russia	2,083	0.46%	32	83	0.20%	44	2,166	0.44%	34
Lebanon	679	0.15%	49	82	0.20%	45	761	0.15%	49
New Zealand	2,011	0.44%	34	80	0.19%	46	2,091	0.43%	35
Nigeria	501	0.11%	52	75	0.19%	47	576	0.12%	52
Indonesia	289	0.06%	64	73	0.17%	48	362	0.07%	60
Jordan	345	0.08%	59	68	0.17%	49	413	0.08%	58
Morocco	304	0.07%	63	66	0.17%	50	370	0.08%	59

Spearman Rank Correlation - all countries	
coefficient (rs)	0.93
N	167
T statistic	32.32
DF	165
p-value	0.000

Spearman Rank Correlation - top 50 countries	
coefficient (rs)	0.85
N	50
T statistic	10.97
DF	48
p-value	0.000

Table S3: Hierarchical linear probability model for the likelihood of women first authorship for COVID articles versus non-COVID (control) articles

Dependent variable: First Author Female	(1)	(2)	(3)	(4)	(5)
COVID	**-0.074***** **(0.00)**	**-0.074***** **(0.00)**	**-0.075***** **(0.00)**	**-0.086***** **(0.00)**	**-0.089***** **(0.00)**
number of authors		0.002*** (0.00)	0.002*** (0.00)	0.002*** (0.00)	0.002*** (0.00)
publication month fixed effects (12)			Included	Included	Included
discipline fixed effects (148)				Included	Included
country fixed effects (167)					Included
constant	0.451*** (0.00)	0.442*** (0.00)	0.454*** (0.00)	0.269*** (0.02)	0.036 (0.10)
R-squared	0.002	0.002	0.002	0.043	0.060
Adjusted R-squared	0.002	0.002	0.002	0.043	0.059
Observations	526,130	526,130	526,130	526,130	491,912

Note: standard errors in brackets, * p < 0.05, ** p< 0.01, *** p < 0.001

Table S4: Hierarchical linear probability model for the likelihood of women last authorship for COVID articles versus non-COVID (control) articles

Dependent variable: Last Author Female	(1)	(2)	(3)	(4)	(5)
COVID	**-0.014*** (0.00)**	**-0.014*** (0.00)**	**-0.015*** (0.00)**	**-0.033*** (0.00)**	**-0.037*** (0.00)**
number of authors		-0.002*** (0.00)	-0.002*** (0.00)	-0.001*** (0.00)	-0.001*** (0.00)
publication month fixed effects (12)			Included	Included	Included
discipline fixed effects (148)				Included	Included
country fixed effects (167)					Included
constant	0.319*** (0.00)	0.332*** (0.00)	0.342*** (0.00)	0.232*** (0.02)	0.081 (0.09)
R-squared	0.000	0.001	0.001	0.043	0.060
Adjusted R-squared	0.000	0.001	0.001	0.043	0.059
Observations	539,103	539,103	539,103	539,103	504,148

Note: standard errors in brackets, * p < 0.05, ** p< 0.01, *** p < 0.001

Table S5: Hierarchical logit regression for the likelihood of women first authorship for COVID articles versus non-COVID (control) articles

Dependent variable: First Author Female	(1)	(2)	(3)	(4)	(5)
COVID	**0.738*** (0.01)**	**0.737*** (0.01)**	**0.733*** (0.01)**	**0.690*** (0.01)**	**0.678*** (0.01)**
number of authors		1.006*** (0.00)	1.006*** (0.00)	1.009*** (0.00)	1.009*** (0.00)
publication month fixed effects (12)			Included	Included	Included
discipline fixed effects (148)				Included	Included
country fixed effects (167)					Included
constant	0.821*** (0.00)	0.790*** (0.00)	0.829*** (0.01)	0.365*** (0.03)	0.117*** (0.06)
observations	526,130	526,130	526,130	526,112	491,837

Note: Coefficients reported as odds ratios, standard errors in brackets, * p < 0.05, ** p < 0.01, *** p < 0.001

Table S6: Robustness checks

	Accuracy of gender designation > 50%	Accuracy of gender designation = 100%	Exluding collective authorships	Full sample
Dependent variable: First Author Female	(1)	(2)	(3)	(4)
COVID	**-0.082*****	**-0.090*****	**-0.089*****	**-0.079*****
	(0.00)	**(0.00)**	**(0.00)**	**(0.00)**
number of authors	0.002***	0.002***	0.003***	0.002***
	(0.00)	(0.00)	(0.00)	(0.00)
publication month fixed effects (12)	Included	Included	Included	Included
discipline fixed effects (148)	Included	Included	Included	Included
country fixed effects (167)	Included	Included	Included	
journal fixed effects (5,101)				Included
constant	0.095	0.040	0.031	0.215*
	(0.10)	(0.10)	(0.10)	(0.10)
R-squared	0.049	0.067	0.060	0.091
Adjusted R-squared	0.048	0.066	0.059	0.082
Observations	607,598	443,711	483,308	507,653

Note: standard errors in brackets, * p < 0.05, ** p< 0.01, *** p < 0.001

Table S7: Descriptive statistics of articles included in the analysis versus articles not included in the analysis

	Included in analysis		Excluded from analysis		t-test	
Variable	**Mean**	**Std. Dev.**	**Mean**	**Std. Dev.**	**Difference**	**t-statistic**
First Author Female	0.38	0.48	0.38	0.49	0.00	0.45
Last Author Female	0.32	0.47	0.33	0.47	0.01	2.27
Publication Month	7.35	3.15	7.51	3.25	0.15	5.35
Number of Authors	6.42	8.56	5.75	6.25	-0.68	-10.76
North America	0.29	0.45	0.24	0.43	-0.05	-12.70
Europe	0.35	0.48	0.28	0.45	0.07	-16.15
Asia	0.19	0.39	0.23	0.42	0.04	9.69
Latin America	0.05	0.21	0.04	0.20	0.00	-1.44
Oceania	0.03	0.16	0.02	0.14	-0.01	-4.22
Africa	0.02	0.14	0.03	0.17	0.01	6.20
Observations	**42,898**		**17,445**		**60,343**	