BMJ Open

Trends in gender of authors of original research in oncology among major medical journals: a retrospective bibliometric study

Shing Fung Lee,[1,2] Daniel Redondo Sánchez,[3,4,5] María-José Sánchez,[3,4,5,6] Bizu Gelaye,[7,8] Chi Leung Chiang,[1,2] Irene Oi Ling Wong,[9] Denise Shuk Ting Cheung,[10] Miguel Angel Luque Fernandez[3,4,5,7,11]

To cite: Lee SF, Redondo Sánchez D, Sánchez M-J, *et al*. Trends in gender of authors of original research in oncology among major medical journals: a retrospective bibliometric study. *BMJ Open* 2021;**11**:e046618. doi:10.1136/bmjopen-2020-046618

- Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2020-046618).

SFL and DRS are joint first authors.

M-JS and MALF are joint senior authors.

Received 04 November 2020
Accepted 19 September 2021

© Author(s) (or their employer(s)) 2021. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

For numbered affiliations see end of article.

Correspondence to
Dr Miguel Angel Luque Fernandez;
miguel-angel.luque@lshtm.ac.uk

ABSTRACT

Objective We evaluated the temporal trend in gender ratios of first and last authors in the field of oncological research published in major general medical and oncology journals and examined the gender pattern in coauthorship.

Design We conducted a retrospective study in PubMed using the R package RISmed. We retrieved original research articles published in four general medical journals and six oncology specialty journals. These journals were selected based on their impact factors and popularity among oncologists. We identified the names of first and last authors from 1 January 2002 to 31 December 2019. The gender of the authors was identified and validated using the Gender API database (https://gender-api.com/).

Primary and secondary outcome measures The percentages of first and last authors by gender and the gender ratios (male to female) and temporal trends in gender ratios of first and last authors were determined.

Results We identified 34 624 research articles, in which 32 452 had the gender of both first and last authors identified. Among these 11 650 (33.6%) had women as the first author and 7908 (22.8%) as the last author, respectively. The proportion of female first and last authors increased from 26.6% and 16.2% in 2002, to 32.9% and 27.5% in 2019, respectively. However, the gender ratio (male to female) of first and last authors decreased by 1.5% and 2.6% per year, respectively, which were statistically significant (first author: incidence rate ratio (IRR) 0.98, 95% CI 0.97 to 1.00; last author: IRR 0.97, 95% CI 0.96 to 0.99). Male first and last authorship was the most common combination. Male–female and female–female pairs increased by 2.0% and 5.0%, respectively (IRR 1.02, 95% CI 1.01 to 1.03 and IRR 1.05, 95% CI 1.04 to 1.06, respectively).

Conclusions The continued under-representation of women means that more efforts to address parity for advancement of women in academic oncology are needed.

Strengths and limitations of this study

- We have performed extensive literature search and described the gender gap in original oncology research publications in 10 major medical journals over the last 18 years.
- Gender inference algorithms currently only support gender binary classification systems.
- We assessed the gender composition of first and last authors on the assumption that these authorship positions are key positions in research activities and publication.
- Our analysis included only original research articles; other important contributions in research projects including conference abstracts for presentation, database management, coding and analysis could not be assessed.

INTRODUCTION

Gender disparity is being increasingly identified in many disciplines, including oncology. In 2017, the Association of American Medical Colleges reported that women accounted for 33.3% and 27.2% of active US physicians in haematology-oncology and radiation oncology, respectively.[1] Compared with men, women are also less likely to be the first authors in high-profile publications,[2] receive grant funding[3,4] and be invited as peer reviewers.[5]

Systematic biases in academia and research institutions may possibly hinder their academic development.[6] The possibility of unconscious bias against women in the peer review processes,[7] the way women tend to promote and present their research findings,[8] and the often under-representation of women in the editorial boards of medical journals may make the journals less likely to prioritise research topics more commonly studied by women.[9,10] However, publishing is important for building and maintaining a successful academic careers. It gives visibility to the author and provides a platform and opportunity for recognition as a researcher. It is not only the means through which research is communicated, but also a critical measure

BMJ Open: first published as 10.1136/bmjopen-2020-046618 on 18 October 2021. Downloaded from http://bmjopen.bmj.com/ on November 4, 2021 at Granada/Medicina/CC Salud PO Box 750. Protected by copyright.

of academic productivity and merit. The quantity and quality of publications are frequently used to assess job employment, tenure and promotion.[11]

Assessing the scientific publication trends in the field of oncology may highlight opportunities for improved awareness of the importance of gender parity, mentorship, representation and advancement for women. The existing data about gender trends in academic publication in oncology often examine only a small subset of journals[12] and include articles other than original research or oncology studies.[13 14] Therefore, we aimed to extend the scientific evidence by examining the global trends in women's representation as first and last authors of original oncology research in major oncology and general medical journals.

METHODS

Data and setting

This study focused on original research articles published from 2002 to 2019 in select high-impact English-language journals. Journals were selected after consideration of relevant literature,[15 16] impact factor, citation half-life and readership. Comments from faculty members regarding the long-term prestige and importance of the various journals were also collected. Four journals categorised by the Thomson Reuters International Scientific Indexing Journal Citation Index as general medical journals were included as follows: *The New England Journal of Medicine*, *The Journal of the American Medical Association (JAMA)*, *The Lancet* and *The BMJ*. Six journals categorised as oncology journals were also included: *Journal of Clinical Oncology*, *Journal of the National Cancer Institute*, *The Lancet Oncology*, *Cancer*, *International Journal of Radiation Oncology*, *Biology*, *Physics*, and *Radiotherapy and Oncology*. These 10 clinically orientated journals whose bibliographic citations include authors' full first name were included. We used the database from Gender API (https://gender-api.com/; Munich, Germany) to assign gender to first and last authors based on their first name. Their algorithm combined and verified data of multiple data sources, including publicly available governmental sources and social networks, to provide the best possible matches.

Search strategy

We developed a search query combining terms for journal names and cancer-related Medical Subject Heading terms. We focused on oncology-specific original research articles. Comments, editorials, review articles, retracted articles, errata biographies, personal narratives, portraits, introductory journal articles, practice guidelines, consensus development conferences, congresses and clinical conferences were excluded a priori using appropriate Medical Subject Heading terms for publication types. Guidelines, duplicate publications, legal cases, interviews and news articles were also excluded. We used the database from Gender API, which is a validated application programming interface for R to predict the probability of gender of an author's first name, to allow assignment of gender to each first and last author. At the time of the study, the database of this program contains 1 847 011 validated names from 177 different countries,[17] and prior study has demonstrated high reliability of this gender inference tool for all name origins, with the best features among the main gender inference services.[18] We used a threshold for accuracy of more than 60% to assign gender based on the author's first name, as has been implemented in previous works.[12 19 20] We also excluded the assignation of genders when the sample of Gender API was less than 10 cases.

Statistical analysis

We first calculated the numbers and percentages of first and last authors who were women and men, respectively. Then, the percentages within each authorship position were calculated out of the total number of women and men authors. To understand the pattern of collaboration of first and last authors, we paired the first and last authors by gender to create four categories. These four categories are (1) male first–male last, (2) male first–female last, (3) female first–male last, and (4) female last–female last. To assess the overall temporal trends, we fitted Poisson regression models with the count of articles in gender ratios as a dependent variable and the year of publication as an independent variable. We presented the rates of change as the interaction between gender ratios as first and last author and the calendar years of publications for the whole period and contrasted them by journals. Additionally, we assessed the temporal trends in percentages of publications when analysed by gender pairs of first and last authors.

As a sensitivity analysis, using data published by the Association of American Medical Colleges,[1] we compared the percentages of female first and last authors with those who are actively licensed female physicians and who worked in the haematology-oncology and radiation oncology practices in the USA. We also fitted separate logistic regression models with the gender of first and last authors of individual papers as a dependent variable and year of publication and the number of coauthors (per one increase, up to 10 authors) in each article as independent variables. Further, we analysed 47 journals in Q4 rank of oncology in 2020 and assessed the distribution of first and last author by gender. Finally, we analysed single-authored articles, computing the male to female ratio by year and for the period 2002–2019. All data analyses were completed using R V.3.6.1 (R Foundation for Statistical Computing, Vienna, Austria), the RISmed package (V.2.1.12) to search in PubMed and Stata V.16.1. We provide the program codes to search the databases and run the analysis in online supplemental file 1.

Patient and public involvement statement

No patients were involved.

BMJ Open: first published as 10.1136/bmjopen-2020-046618 on 18 October 2021. Downloaded from http://bmjopen.bmj.com/ on November 4, 2021 at Granada/Medicina/CC Salud PO Box 750. Protected by copyright.

RESULTS

Of the 34 624 articles analysed, we identified 9332 unique names of first and last authors. Of all the articles analysed, gender was identified for both the first and last authors in 32 452 (94%). Out of 34 624 articles, 11 650 (33.6%) had women as the first author, while 7908 (22.8%) had women as the last author (p value both <0.001) (see table 1). Gender information was not available for 3.6% of the first authors and 3.0% of the last authors. This is similar to the 3.0% non-classified names achieved by Gender API in the benchmark of different gender assignments tools.[18] As shown in table 2, the proportion of women first authors increased from 26.6% in 2002 to 32.9% in 2019. Similarly, the proportion of women last authors increased from 16.2% in 2002 to 27.5% in 2019 (p value for trend <0.001).

Figures 1 and 2 show changes in representation of women as first and last authors by year and journal, respectively. The average annual changes in the gender ratio of first and last authors were −1.5% and −2.6%, respectively (first author: incidence rate ratio (IRR) 0.98, 95% CI 0.97 to 1.00, p value for trend=0.12; last author: IRR 0.97, 95% CI 0.96 to 0.99, p value for trend=0.001).

The analyses for gender pairs of first and last authors are shown in figures 3 and 4. Although men as both first and last authors was the most common, over time male–male pairs decreased by 2.0% (IRR 0.98, 95% CI 0.98 to 0.99, p value for trend <0.001), and male–female and female–female pairs increased by 2.0% and 5.0%, respectively (IRR 1.02, 95% CI 1.01 to 1.03, p value for trend <0.001 and IRR 1.05, 95% CI 1.04 to 1.06, p value for trend <0.001, respectively).

In a sensitivity analysis, we compared the ratios with those of practising physicians in the USA. Women represent 33.3% and 27.2% of actively licensed physicians in haematology-oncology and radiation oncology, respectively[1] (online supplemental table 1 and figure 1). Our results indicate that the representation of women as first and last authors was comparable with the overall representation of women as actively licensed physicians in oncology specialties in the USA. For every one increase in the number of coauthors, the odds of the first author being male are 1.04 (95% CI 1.03 to 1.05, p<0.001) and the odds of the last author being male are 1.07 (95% CI 1.06 to 1.08, p<0.001). The sensitivity analysis based on Q4 rank of oncology journal in PubMed in 2020 (13 of 47 journals were retrievable in PubMed) showed consistent distribution of first and last authors by gender and gender combinations of first and last authors as presented in the main analysis (online supplemental table 2).

Finally, single-authored articles represented 2.8% of all articles (977 single-authored articles of 34 624 articles with available genders for first and last authors). The overall male to female ratio for single-authored articles was 2:1, while a decreasing pattern was found when analysing these data by year (online supplemental table 3).

DISCUSSION

We found that women were first authors of 33.6% articles. We also found evidence for continuously increasing representation of women as senior authors and authors of single-authored articles. The trend has been largely stable in recent years. However, disparity still exists, especially in the first author position. Female–female author pairs showed the most obvious increase relative to other gender pairings.

Our study results are largely consistent with prior studies that investigated female authorship in journals with high-impact factors, with some differences likely attributable to the averaging of multiple journals in our analysis. Dalal *et al*[21] reported comparable increasing trends in female authorship in five major oncology journals between 1990 and 2017. Their analysis included editorials, reviews, letters, notes and proceedings, in addition to original research articles. For the year 2017, the percentage of female senior authorship was nearly 10% lower than that of first authorship, which corroborated with our findings.[21] Ahmed *et al*[15] examined trends in female authorship in the *International Journal of Radiation Oncology, Biology, Physics* between 1980 and 2012. For all original articles in 2012, they reported that 29.7% and 22.6% of first authors and last authors were women, respectively.[15] Although gender differences in medical research have been examined,[2–5 22] our study is unique as we described the gender gap in original oncology research publications in 10 major medical journals over the last 18 years.

The number of female first authors was consistently lower than that of male first authors. Nearly one-third of actively licensed oncology physicians in the USA are women,[1] and the annual increase in female oncologists is greater than that of male oncologists in many European countries.[23] However, female oncologists are less likely to have leadership roles and feel that their gender is adversely affecting their career.[23] The relative lack of female investigators publishing oncology topics is consistent with the trend in gender differences in other research areas, despite an increase in women entering scientific careers.[19 24] It is noteworthy that the proportion of US oncologists was used for comparison in the sensitivity analysis, but the gender findings were derived from literature by authors across the globe. Thus, this comparison should be at best considered as hypothesis-generating.

Simpson's paradox may bias the interpretation of data on gender disparity in publication, including our results. This paradox implies that an apparent association between two variables can actually be the result of a joint dependency on a third variable.[25] For example, a finding that female researchers succeeded less often in grant applications could be biased because women pursue funding more frequently in more competitive research fields, and the funded topics are under-represented in the analysed journals.[25] Our results could also be related to self-selection of career paths; for example, fewer women choose to publish papers due to fewer opportunities to submit papers and multiple demands on their

BMJ Open: first published as 10.1136/bmjopen-2020-046618 on 18 October 2021. Downloaded from http://bmjopen.bmj.com/ on November 4, 2021 at Granada/Medicina/CC Salud PO Box 750. Protected by copyright.

Table 1 Author gender percentages by journal (2002–2019)

Journal abbreviations*	Articles (n)	Women as first authors, n (%)	Men as first authors, n (%)	Undetermined gender of first authors, n (%)	P value†	Women as last authors, n (%)	Men as last authors, n (%)	Undetermined gender of last authors, n (%)	P value†
Total	34 624	11 650 (33.6)	21 723 (62.7)	1251 (3.6)	<0.001	7908 (22.8)	25 683 (74.2)	1033 (3.0)	<0.001
The BMJ	531	219 (41.2)	301 (56.7)	11 (2.1)		179 (33.7)	344 (64.8)	8 (1.5)	
Cancer	8971	3213 (35.8)	5392 (60.1)	366 (4.1)		2264 (25.2)	6409 (71.4)	298 (3.3)	
Int J Radiat Oncol Biol Phys	7870	2275 (28.9)	5270 (67.0)	325 (4.1)		1447 (18.4)	6146 (78.1)	277 (3.5)	
JAMA	705	275 (39.0)	414 (58.7)	16 (2.3)		216 (30.6)	464 (65.8)	25 (3.6)	
J Clin Oncol	7942	2680 (33.7)	5012 (63.1)	250 (3.2)		1882 (23.7)	5861 (73.8)	199 (2.5)	
J Natl Cancer Inst	2209	962 (43.6)	1154 (52.2)	93 (4.2)		660 (29.9)	1487 (67.3)	62 (2.8)	
Lancet	521	129 (24.8)	373 (71.6)	19 (3.7)		96 (18.4)	414 (79.5)	11 (2.1)	
Lancet Oncol	1442	388 (26.9)	1013 (70.3)	41 (2.8)		334 (23.2)	1073 (74.4)	35 (2.4)	
NEJM	903	236 (26.1)	646 (71.5)	21 (2.3)		188 (20.8)	701 (77.6)	14 (1.6)	
Radiother Oncol	3530	1273 (36.1)	2148 (60.9)	109 (3.1)		642 (18.2)	2784 (78.9)	104 (3.0)	

*Journals are arranged in alphabetical order.

†χ^2 p value.

Int J Radiat Oncol Biol Phys, *International Journal of Radiation Oncology, Biology, Physics*; *JAMA*, *The Journal of the American Medical Association*; *J Clin Oncol*, *Journal of Clinical Oncology*; *J Natl Cancer Inst*, *Journal of the National Cancer Institute*; *Lancet Oncol*, *The Lancet Oncology*; *NEJM*, *The New England Journal of Medicine*; *Radiother Oncol*, *Radiotherapy and Oncology*.

BMJ Open: first published as 10.1136/bmjopen-2020-046618 on 18 October 2021. Downloaded from http://bmjopen.bmj.com/ on November 4, 2021 at Granada/Medicina/CC Salud PO Box 750. Protected by copyright.

Table 2 Author gender percentages by year (2002–2019)

Years	Articles (n)	Women as first authors, n (%)	Men as first authors, n (%)	Undetermined gender of first authors, n (%)	P value*	Women as last authors, n (%)	Men as last authors, n (%)	Undetermined gender of last authors, n (%)	P value*
Total	34 624	11 650 (33.6)	21 723 (62.7)	1251 (3.6)	<0.001	7640 (22.1)	24 812 (71.7)	1033 (3.0)	<0.001
2002	1862	496 (26.6)	1259 (67.6)	107 (5.7)		302 (16.2)	1453 (78.0)	107 (5.7)	
2003	1856	545 (29.1)	1231 (66.3)	80 (4.3)		330 (17.8)	1446 (77.9)	80 (4.3)	
2004	1873	541 (28.9)	1216 (64.9)	116 (6.2)		342 (18.3)	1415 (75.5)	116 (6.2)	
2005	2452	764 (31.2)	1555 (63.4)	133 (5.4)		469 (19.1)	1850 (75.4)	133 (5.4)	
2006	2069	690 (33.3)	1269 (61.3)	110 (5.3)		399 (19.3)	1560 (75.3)	110 (5.3)	
2007	2167	686 (31.7)	1358 (62.3)	123 (5.7)		439 (20.3)	1605 (74.1)	123 (5.7)	
2008	2282	732 (32.1)	1399 (61.3)	151 (6.6)		484 (21.2)	1647 (72.2)	151 (6.6)	
2009	2372	763 (32.2)	1455 (61.3)	154 (6.5)		494 (20.8)	1724 (72.7)	154 (6.5)	
2010	2434	774 (31.8)	1487 (61.1)	173 (7.1)		516 (21.2)	1745 (71.7)	173 (7.1)	
2011	2412	822 (34.1)	1423 (59.0)	167 (6.9)		544 (22.6)	1701 (70.5)	167 (6.9)	
2012	1970	647 (32.8)	1191 (60.5)	132 (6.7)		423 (21.5)	1415 (71.8)	132 (6.7)	
2013	1731	616 (35.6)	1001 (57.8)	114 (6.6)		403 (23.3)	1214 (70.1)	114 (6.6)	
2014	1759	632 (35.9)	1019 (57.9)	108 (6.1)		448 (25.5)	1203 (68.4)	108 (6.1)	
2015	1804	627 (34.8)	1049 (58.1)	128 (7.1)		455 (25.2)	1221 (67.7)	128 (7.1)	
2016	1733	668 (38.5)	979 (56.5)	86 (5.0)		494 (28.5)	1153 (66.5)	86 (5.0)	
2017	1770	631 (35.6)	1010 (57.1)	129 (7.3)		501 (28.3)	1140 (64.4)	129 (7.3)	
2018	1577	576 (36.5)	881 (55.9)	120 (7.6)		459 (29.1)	998 (63.3)	120 (7.6)	
2019†	501	165 (32.9)	295 (58.9)	41 (8.2)		138 (27.5)	322 (64.3)	41 (8.2)	

*χ^2 for trend p value.

†Not information from all articles was available at the time of retrieval and analysis.

BMJ Open: first published as 10.1136/bmjopen-2020-046618 on 18 October 2021. Downloaded from http://bmjopen.bmj.com/ on November 4, 2021 at Granada/Medicina/CC Salud PO Box 750. Protected by copyright.

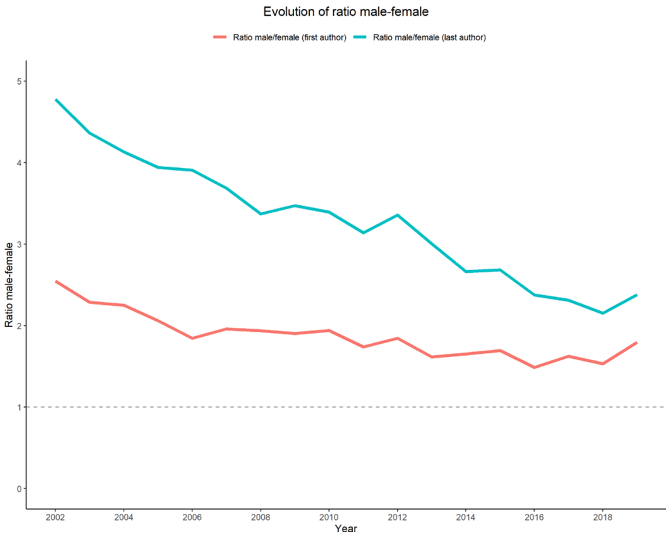


Figure 1 Gender ratios of first authors and last authors from 2002 to 2019, all journals.

time.[26] Unfortunately, we did not have information about the number of submitted papers or the distribution of the topics accepted and rejected by journals, respectively. Additionally, women often have greater caregiving responsibility, and work–family conflicts can be at odds with achieving academic outcomes.[25] Further, positive presentation of research findings is associated with higher rates of subsequent citations,[8] which could drive differential recognition of accomplishments in women and men. A complex interaction of the above gender biases and policies can contribute to the under-representation of women in publication and in leadership positions.[27 28] Finally, we were not able to examine this in our study but under-representation is much more pronounced for women from racial/ethnic minority groups. Future studies are warranted to fully understand the historical, structural, and institutional barriers that contribute to the chronic under-representation of these subgroups.

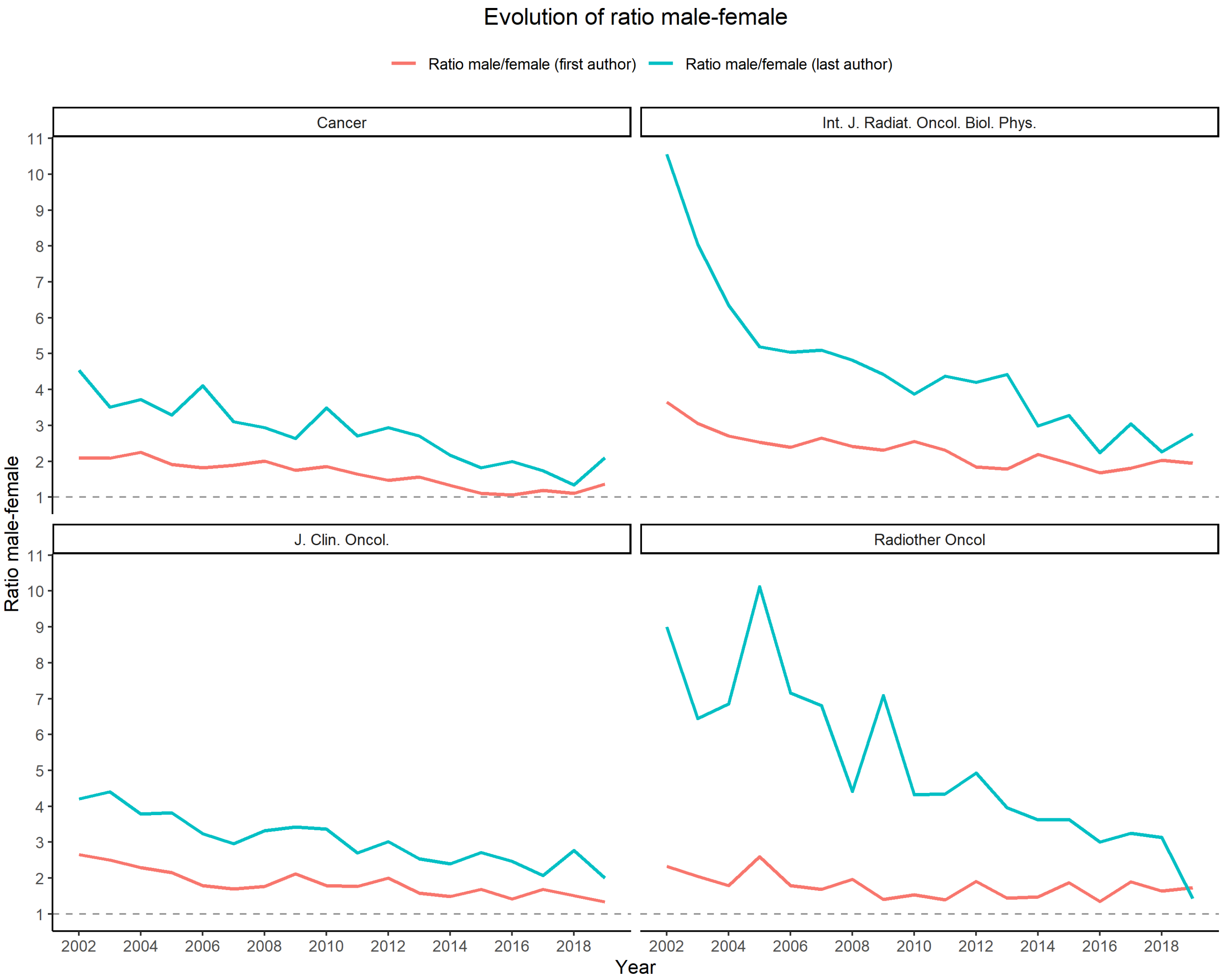


Figure 2 Gender ratios of first authors and last authors from 2002 to 2019. The top 4 journals (by number of articles) are shown because some of the other journals had zero cases for gender in some years. *Int J Radiat Oncol Biol Phys*, *International Journal of Radiation Oncology, Biology, Physics*; *J Clin Oncol*, *Journal of Clinical Oncology*; *Radiother Oncol*, *Radiotherapy and Oncology*.

BMJ Open: first published as 10.1136/bmjopen-2020-046618 on 18 October 2021. Downloaded from http://bmjopen.bmj.com/ on November 4, 2021 at Granada/Medicina/CC Salud PO Box 750. Protected by copyright.

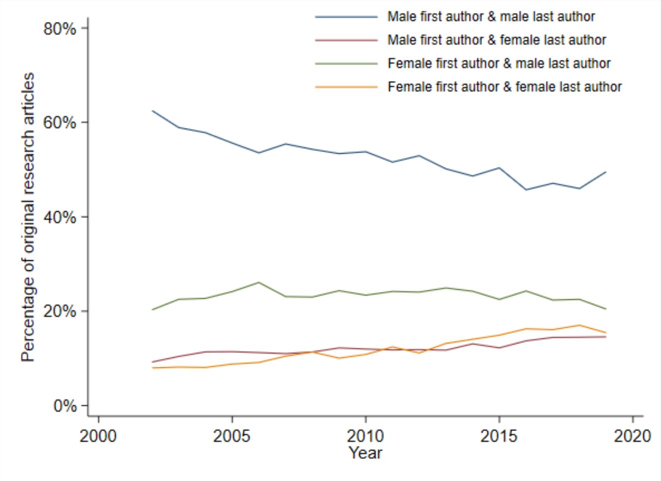


Figure 3 Temporal trends in percentage of original research articles published by female and male authors when paired by gender as first and last authors from 2002 to 2019, all journals.

Our study has some imitations. First, our use of Gender API and choice of the 60% accuracy threshold represent a trade-off between the accuracy and the comprehensiveness of the data analysed. Second, gender inference algorithms currently only support gender binary classification systems. Third, available information did not allow adjustment for some variables. For instance, H-index is not available for all authors; the rank of institutions could be difficult to adjust for because many articles were published by hospitals and laboratories, authors often have multiple affiliations and articles were sometimes written in non-standard formats. Fourth, we assumed that the first and last authors were oncologists, but the papers might actually have been written by other non-oncologist professionals. Finally, we assessed the gender composition of first and last authors on the assumption that these authorship positions are key positions in research activities and publication. Our analysis included only original research articles. Other important contributions in research projects including conference abstracts for presentation, database management, coding and analysis could not be assessed.

Overall, the variation between journals and time periods suggests the need and opportunity for continued efforts to support the advancement of women in oncology research. Strengthening existing initiatives and supporting new pipeline programmes are needed to address under-representation of female oncologists in leadership roles and promote equal access to career development opportunities. The proportion of single authorship by women in prestigious journals is smaller in contrast to men. We argue that initiatives to promote

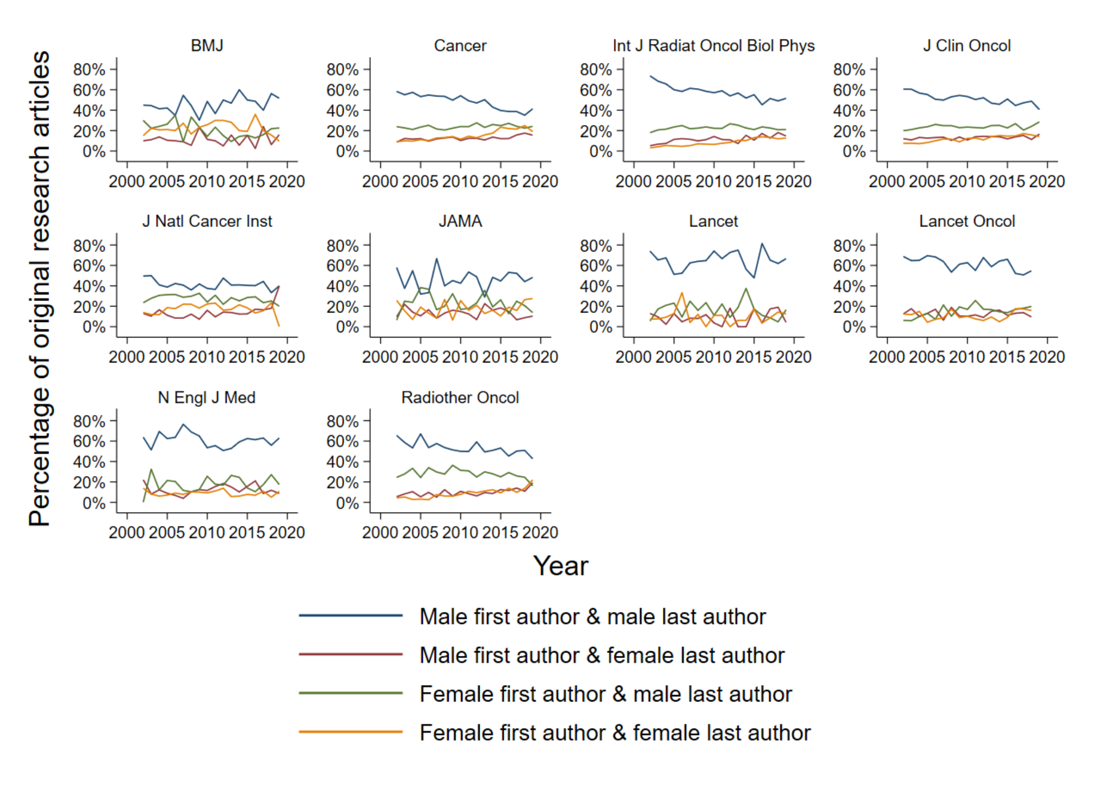


Figure 4 Temporal trends in percentage of original research articles published by female and male authors when paired by gender as first and last authors from 2002 to 2019, by individual journals. *Int J Radiat Oncol Biol Phys*, *International Journal of Radiation Oncology, Biology, Physics*; *JAMA*, *The Journal of the American Medical Association*; *J Clin Oncol*, *The Journal of Clinical Oncology*; *J Natl Cancer Inst*, *The Journal of the National Cancer Institute*; *Lancet Oncol*, *The Lancet Oncology*; *N Engl J Med*, *The New England Journal of Medicine*; *Radiother Oncol*, *Radiotherapy and Oncology*.

BMJ Open: first published as 10.1136/bmjopen-2020-046618 on 18 October 2021. Downloaded from http://bmjopen.bmj.com/ on November 4, 2021 at Granada/Medicina/CC Salud PO Box 750. Protected by copyright.

academic choices and careers of women at the student level more than focusing on journals or peer review must be developed. To objectively evaluate programmes aimed at addressing gender inequalities, collecting high-quality data will be critical.

Author affiliations
[1]Department of Clinical Oncology, University of Hong Kong, Hong Kong
[2]Department of Clinical Oncology, Tuen Mun Hospital, Hong Kong
[3]Department of Non-Communicable Disease and Cancer Epidemiology, Instituto de Investigacion Biosanitaria de Granada (ibs.GRANADA), University of Granada, Granada, Spain
[4]Andalusian School of Public Health, Granada, Spain
[5]Centro de Investigación Biomédica en Red de Epidemiología y Salud Pública (CIBERESP), Madrid, Spain
[6]Department of Preventive Medicine and Public Health, University of Granada, Granada, Spain
[7]Department of Epidemiology, Harvard University T H Chan School of Public Health, Boston, Massachusetts, USA
[8]Psychiatry, Harvard Medical School, Boston, Massachusetts, USA
[9]School of Public Health, University of Hong Kong, Hong Kong
[10]School of Nursing, Li Ka Shing Faculty of Medicine, University of Hong Kong, Hong Kong
[11]Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK

Twitter Shing Fung Lee @bestmic1 and Miguel Angel Luque Fernandez @WATZILEI

Contributors SFL and MALF had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Concept and design: SFL, DRS, MALF. Acquisition, analysis or interpretation of data: SFL, DRS, MALF. Drafting of the manuscript: SFL, DRS, M-JS, BG, CLC, IOLW, DSTC, MALF. Critical revision of the manuscript for important intellectual content: SFL, DRS, M-JS, BG, CLC, IOLW, DSTC, MALF. Obtained funding: MALF. Administrative, technical or material support: DRS, M-JS, BG, MALF. Supervision: MALF.

Funding MALF was supported by the Spanish National Health Institute Carlos III (Instituto de Salud Carlos III - ISCIII), Miguel Servet-I Investigator Grant/Award Number CP17/ 00206-EU-FEDER.

Disclaimer The funders had no role in the study design, data collection and analysis, decision to publish or preparation of the manuscript.

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval The study was exempted from institutional ethics review because only publicly available data were analysed.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available in a public, open access repository.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/.

ORCID iDs
Shing Fung Lee http://orcid.org/0000-0003-4696-9434
Miguel Angel Luque Fernandez http://orcid.org/0000-0001-6683-5164

REFERENCES

1 Association of American Medical Colleges. Physician specialty data report, 2018. Available: https://www.aamc.org/data-reports/workforce/report/physician-specialty-data-report [Accessed 22 Mar 2020].
2 Nature. Nature's under-representation of women. *Nature* 2018;558:344. doi:10.1038/d41586-018-05465-7
3 Witteman HO, Hendricks M, Straus S, *et al*. Are gender gaps due to evaluations of the applicant or the science? a natural experiment at a national funding agency. *Lancet* 2019;393:531–40.
4 Kaatz A, Gutierrez B, Carnes M. Threats to objectivity in peer review: the case of gender. *Trends Pharmacol Sci* 2014;35:371–3.
5 Lerback J, Hanson B. Journals invite too few women to referee. *Nature* 2017;541:455–7.
6 Silver JK. Medical journals must tackle gender bias. *BMJ* 2019;367:l5888.
7 Hengel E. Evidence from peer review that women are held to higher standards. Available: https://voxeu.org/article/evidence-peer-review-women-are-held-higher-standards
8 Lerchenmueller MJ, Sorenson O, Jena AB. Gender differences in how scientists present the importance of their research: observational study. *BMJ* 2019;367:l6573.
9 Jagsi R, Tarbell NJ, Henault LE, *et al*. The representation of women on the editorial boards of major medical journals: a 35-year perspective. *Arch Intern Med* 2008;168:544–8.
10 Erren TC, Groß JV, Shaw DM, *et al*. Representation of women as authors, reviewers, editors in chief, and editorial board members at 6 general medical journals in 2010 and 2011. *JAMA Intern Med* 2014;174:633–5.
11 Beasley BW, Simon SD, Wright SM. A time to be promoted. The prospective study of promotion in academia (prospective study of promotion in academia). *J Gen Intern Med* 2006;21:123–9.
12 Hart KL, Perlis RH. Trends in proportion of women as authors of medical Journal articles, 2008-2018. *JAMA Intern Med* 2019;179:1285–7.
13 Larivière V, Ni C, Gingras Y, *et al*. Bibliometrics: global gender disparities in science. *Nature* 2013;504:211–3.
14 Jagsi R, Silver JK. Gender differences in research reporting. *BMJ* 2019;367:l6692.
15 Ahmed AA, Egleston B, Holliday E, *et al*. Gender trends in radiation oncology in the United States: a 30-year analysis. *Int J Radiat Oncol Biol Phys* 2014;88:33–8.
16 Jagsi R, Sheets N, Jankovic A, *et al*. Frequency, nature, effects, and correlates of conflicts of interest in published clinical cancer research. *Cancer* 2009;115:2783–91.
17 Gender API, 2020. Available: https://gender-api.com/ [Accessed 11 Mar 2020].
18 Santamaría L, Mihaljević H. Comparison and benchmark of name-to-gender inference services. *PeerJ Comput Sci* 2018;4:e156.
19 Hart KL, Frangou S, Perlis RH. Gender trends in authorship in psychiatry journals from 2008 to 2018. *Biol Psychiatry* 2019;86:639–46.
20 Shen YA, Webster JM, Shoda Y. Persistent Underrepresentation of Women's Science in High Profile Journals. *bioRxiv* 2018. doi:10.1101/275362
21 Dalal NH, Chino F, Williamson H, *et al*. Mind the gap: gendered publication trends in oncology. *Cancer* 2020;126:2859–65.
22 Silver JK, Ghalib R, Poorman JA, *et al*. Analysis of gender equity in leadership of physician-focused medical specialty societies, 2008-2017. *JAMA Intern Med* 2019;179:433–5.
23 Banerjee S, Dafni U, Allen T, *et al*. Gender-Related challenges facing oncologists: the results of the ESMO women for oncology Committee survey. *ESMO Open* 2018;3:e000422.
24 Metheny WP, Jagadish M, Heidel RE. A 15-year study of trends in authorship by gender in two U.S. obstetrics and gynecology journals. *Obstet Gynecol* 2018;131:696–9.
25 Albers CJ, funding Dresearch. Dutch research funding, gender bias, and Simpson's paradox. *Proc Natl Acad Sci U S A* 2015;112:E6828–9.
26 Teele DL, Thelen K. Gender in the journals: publication patterns in political science. *PS: Political Science & Politics* 2017;50:433–47.
27 Beeler WH, Cortina LM, Jagsi R. Diving beneath the surface: addressing gender inequities among clinical Investigators. *J Clin Invest* 2019;129:3468–71. doi:10.1172/JCI130901
28 Gharzai LA, Jagsi R. Ongoing gender inequity in leadership positions of academic oncology programs: the broken pipeline. *JAMA Netw Open* 2020;3:e200691–e91. doi:10.1001/jamanetworkopen.2020.0691

BMJ Open: first published as 10.1136/bmjopen-2020-046618 on 18 October 2021. Downloaded from http://bmjopen.bmj.com/ on November 4, 2021 at Granada/Medicina/CC Salud PO Box 750. Protected by copyright.

Analysis Codes in R

```
# In order not to overload the E-utility servers, NCBI recommends that users post no
more
# than three URL requests per second and limit large jobs to either weekends or
between
# 9:00 PM and 5:00 AM Eastern time during weekdays. Failure to comply with this
policy may
# result in an IP address being blocked from accessing NCBI.

# ----- Packages -----

# devtools::install_github("skoval/RISmed")
library(RISmed)
library(dplyr)
library(ggplot2)
library(rjson)
library(httr)
library(reshape)

# ----- Get data from PubMed -----

# Define the terms of the query
terms_query <- '("Neoplasms"[Mesh])
  AND humans[MeSH Terms]
  AND (("N ENGL J MED"[JOURNAL])
      OR ("JAMA"[JOURNAL])
      OR ("Lancet"[JOURNAL])
      OR ("BMJ"[JOURNAL])
      OR ("J Clin Oncol"[JOURNAL])
      OR ("JAMA Oncol"[JOURNAL])
      OR ("Lancet Oncol"[JOURNAL])
      OR ("J Natl Cancer Inst"[JOURNAL])
      OR ("Cancer"[JOURNAL])
      OR ("ANN Oncol"[JOURNAL])
      OR ("Int J Radiat Oncol Biol Phys"[JOURNAL])
      OR ("Radiother Oncol"[JOURNAL]))
  AND Journal Article[ptyp]
```

```
    NOT (Letter[ptyp] OR Case Reports[ptyp] OR Comment[ptyp] OR Editorial[ptyp]
            OR Review[ptyp] OR News[ptyp] OR Congress[ptyp] OR Retracted
Publication[ptyp]
            OR Published Erratum[ptyp] OR Biography[ptyp] OR Personal
Narrative[ptyp]
            OR Book Illustrations[ptyp] OR Introductory Journal Article[ptyp]
            OR Guideline [ptyp] OR Practice guideline[ptyp]
            OR consensus development conferences[ptyp] OR Clinical Conference[ptyp]
            OR Address[ptyp] OR Duplicate Publication[ptyp] OR Interview[ptyp] OR
Legal case[ptyp])
    AND (2002/01/01:2019/12/31[Date - Publication]
            OR 2002[Date - Publication] OR 2003[Date - Publication]
            OR 2004[Date - Publication] OR 2005[Date - Publication]
            OR 2006[Date - Publication] OR 2007[Date - Publication]
            OR 2008[Date - Publication] OR 2009[Date - Publication]
            OR 2010[Date - Publication] OR 2011[Date - Publication]
            OR 2012[Date - Publication] OR 2013[Date - Publication]
            OR 2014[Date - Publication] OR 2015[Date - Publication]
            OR 2016[Date - Publication] OR 2017[Date - Publication]
            OR 2018[Date - Publication] OR 2019[Date - Publication]
            )
    NOT (1800/01/01:2001/12/31[Date - Publication])
    NOT (2020/01/01:3000[Date - Publication])'

# Get information
query <- EUtilsSummary(terms_query, mindate = 2002, maxdate = 2019, retmax =
50000)

# Number of results (23/03/2020 - 42066 articles)
QueryCount(query)

# Get the data from PubMed
# Downloaded 23/03/2020 - 42066 articles
#records <- EUtilsGet(query, type = "efetch", db = "pubmed")
#save(records, file = "records.RData")
load("records.RData")

# ----- Data preprocessing    -----
# Convert authors from list to dataframe
authors_list <- Author(records)

# Initialise authors
authors <- data.frame(first_forename = rep("", length(authors_list)),
                      last_forename = rep("", length(authors_list)),
                      number_authors = NA)

# Convert to character
class(authors[, 1]) <- "character"
class(authors[, 2]) <- "character"

# Extract names and number of authors
for(i in 1:length(authors_list)){
  # Counter
  if(i %% 500 == 0) cat(".")
  # Extract forenames from first and last authors
  authors$first_forename[i] <- authors_list[i][[1]][1,]$ForeName
  authors$last_forename[i] <-
authors_list[i][[1]][nrow(authors_list[i][[1]]),]$ForeName
  authors$number_authors[i] <- max(authors_list[i][[1]]$order)
}

# Create data.frame
pubmed_data <- data.frame(
  "Title" = ArticleTitle(records),
  "first_forename" = as.character(authors$first_forename),
  "last_forename" = as.character(authors$last_forename),
  "number_authors" = authors$number_authors,
  "PMID" = PMID(records),
  "Year" = YearPubmed(records),
  "Journal" = ISOAbbreviation(records)
)

# 14/04/2020 Remove Ann. Oncol and JAMA Oncol
pubmed_data <- pubmed_data %>%
  filter(! Journal %in% c("JAMA Oncol", "Ann. Oncol."))
```

```
# See the first rows
nrow(pubmed_data)
head(pubmed_data)

# To character
pubmed_data$first_forename <- as.character(pubmed_data$first_forename)
pubmed_data$last_forename <- as.character(pubmed_data$last_forename)

# Results by year
table(pubmed_data$Year)

# Results by journal
table(pubmed_data$Journal)

# Barplot by journal
table(pubmed_data$Journal) %>% as.data.frame() %>%
  ggplot() +
  geom_col(aes(x = Var1, y = Freq), fill = "darkblue") +
  ylab("Frequency of articles") +
  xlab("Journal") +
  theme_classic() +
  theme(axis.text = element_text(size = 7))

# ----- Text mining -----

# Names more frequent
table(pubmed_data$first_forename) %>% sort(decreasing = TRUE) %>% head(20)
table(pubmed_data$last_forename) %>% sort(decreasing = TRUE) %>% head(20)

# keep articles if first and last names are not NA
nrow(pubmed_data)
pubmed_data <- pubmed_data %>% filter(is.na(first_forename) == FALSE)
pubmed_data <- pubmed_data %>% filter(is.na(last_forename) == FALSE)
nrow(pubmed_data)

# Lisa M -> Lisa
# Michael M -> Michael
```

```
for(i in 1:nrow(pubmed_data)){
    if(i %% 500 == 0) cat(".")
    pubmed_data$first_forename[i] <- strsplit(pubmed_data$first_forename[i], split =
" ", fixed = TRUE)[[1]][1]
    pubmed_data$last_forename[i] <- strsplit(pubmed_data$last_forename[i], split =
" ", fixed = TRUE)[[1]][1]
}

# keep only names with length > 1
table(pubmed_data$Year)
nrow(pubmed_data)
pubmed_data <- pubmed_data %>% filter(nchar(first_forename)>1)
pubmed_data <- pubmed_data %>% filter(nchar(last_forename)>1)
nrow(pubmed_data)
table(pubmed_data$Year)

# 20 names more frequent
table(pubmed_data$first_forename) %>% sort(decreasing = TRUE) %>% head(20)
table(pubmed_data$last_forename) %>% sort(decreasing = TRUE) %>% head(20)

# ----- Table to get genders -----

# This code is to do the minimum number of querys to GenderAPI:
names <- unique(c(pubmed_data$first_forename, pubmed_data$last_forename))
print(paste0("Number of articles: ", nrow(pubmed_data), ". Number of names: ",
length(names), "."))

head(names)

# ----- GenderAPI - Creating files -----

# This code uses an API from GenderAPI to obtain a dataframe containing:
# - name
# - gender
# - samples (Number of samples used to obtain the gender)
# - accuracy (Percentage of accuracy of the gender)

## # API of GenderAPI
## api = "uNcRMbLdfhrTDPuAEs"

## # Get the genders
## for(i in 1:length(names)){
##     # Small sleep between names
##     Sys.sleep(0.1)
##     # Counter
##     if(i %% 50 == 0) print(i)
##     # Build URL
##     url <- paste0("https://gender-api.com/get?name=", names[i], "&key=", api)
##     # Get the gender from GenderAPI
##     content <- url %>% GET %>% content %>% data.frame
##     name_i_with_gender <- content %>% select(names = name, gender, accuracy)
##     if(i == 1){
##         # Create table
##         names_with_genders <- name_i_with_gender
##     } else {
##         # Append to the table
##         names_with_genders <- rbind(names_with_genders, name_i_with_gender)
##     }
##     # Save raw result in case it's needed
##     write.csv(content, file = paste0("genderapi/", names[i], ".csv"))
## }

# ----- GenderAPI - Joining files -----

## # Slight modifications of characters, only in the name of the CSV:
## # ?    ?     ar.csv -> ?     glar.csv
## # Do?     ng 羹 n.csv -> Dogang 羹 n.csv
## # Gra 髒 yna.csv -> Grazyna.csv
## # Ji?    ?.csv -> Jir 穩.csv
## # ?     kasz.csv -> Lukasz.csv
##
## # Changes reflected also in pubmed_data (using UNICODE characters)
## names_with_issues <- c("?    \U011Flar", paste0("Do", "\U011F", "ang\U00FCn"),
"Gra\U017Cyna", "Ji\U0159 穩", "\U0141ukasz")
## print(names_with_issues) # See the names with issues
## names_corrected <- c("?     glar", "Dogang 羹 n", "Grazyna", "Jir 穩", "Lukasz")
## for(i in 1:length(names_with_issues)){
##      pubmed_data$first_forename[pubmed_data$first_forename ==
names_with_issues[i]] <- names_corrected[i]
##      pubmed_data$last_forename[pubmed_data$last_forename ==
names_with_issues[i]] <- names_corrected[i]
## }
##
## # Join all files produced with GenderAPI
## setwd("genderapi")
## files <- list.files()
##
## for(i in 1:length(files)){
##      # Counter
##      if(i %% 500 == 0) cat(".")
##      file_i <- read.csv(files[i])
##      if(i == 1) names_with_genders <- file_i
##      else names_with_genders <- rbind(names_with_genders, file_i)
## }
##
## setwd("..")

## # Some names are duplicated: (e.g. Ay?      is processed through GenderAPI like
"Ayse", that was present already)
## # We remove duplicates
## names_with_genders <- unique(names_with_genders)
##
## head(names_with_genders)
## save(names_with_genders, file = "names_with_genders.RData")
load("names_with_genders.RData")

# ----- Processing genders -----

# Convert name to character
names_with_genders$name <- as.character(names_with_genders$name)
head(names_with_genders)

# Removing unknowns and empty gender:
table(names_with_genders$gender, useNA = "always")
```

```
nrow(names_with_genders)
names_with_genders <- names_with_genders %>% filter(gender != "unknown",
gender != "")
names_with_genders$gender <- factor(names_with_genders$gender)
nrow(names_with_genders)

# Establish a threshold for accuracy (in percentage)
nrow(names_with_genders)
threshold <- 60
# See names + genders that are going to be removed
names_with_genders %>% filter(accuracy < threshold) %>% head(10)
# Remove
names_with_genders <- names_with_genders %>% filter(accuracy >= threshold)
nrow(names_with_genders)

# Establish a threshold for samples
nrow(names_with_genders)
threshold <- 10 # For example
# See names + genders that are going to be removed
names_with_genders %>% filter(samples < threshold) %>% head(10)
# Remove
names_with_genders <- names_with_genders %>% filter(accuracy >= threshold)
nrow(names_with_genders)



# ----- Updating gender in pubmed_data -----

# Remove capitals from names
pubmed_data$first_forename <- tolower(pubmed_data$first_forename)
pubmed_data$last_forename <- tolower(pubmed_data$last_forename)

# Create columns
pubmed_data$first_gender <- NULL
pubmed_data$last_gender <- NULL

# Merge to get the gender
pubmed_data$first_gender <- left_join(x = pubmed_data, y = names_with_genders,
                                      by = c("first_forename" =
"name"))$gender %>% as.character
pubmed_data$last_gender <- left_join(x = pubmed_data, y = names_with_genders,
                                     by = c("last_forename" =
"name"))$gender %>% as.character

# Export pubmed_data
save(pubmed_data, file = "pubmed_data.RData")
write.csv(pubmed_data, file = "pubmed_data.csv", row.names = F)

# ----- Analysis of single authored articles -----
sa_articles <- pubmed_data %>%
  filter(number_authors == "1") %>%
  filter(!is.na(first_gender)) %>%
  select(Year, first_gender) %>%
  group_by(Year, first_gender) %>%
  summarise(n = n()) %>%
  reshape2::dcast(formula = Year ~ first_gender, value.var = "n") %>%
  mutate(ratio = round(male/female, 1))

sa_articles

ggplot(data = sa_articles, aes(x = Year, y = ratio, label = ratio)) +
  geom_col(color = "grey") +
  geom_label(color = "white", fill = NA, label.size = 0, nudge_y = -0.3) +
  scale_x_continuous(breaks = 2002:2019, minor_breaks = 2002:2019) +
  labs(x = "Year", y = "Male-female ratio for single-authored articles") +
  theme_minimal() +
  theme(axis.text = element_text(color = "black"))
ggsave(filename = "online_figure_2.png", width = 8, height = 5, dpi = 300)

sum(sa_articles$female) + sum(sa_articles$male) # 977 articles
977*100 / nrow(pubmed_data) # 2.8% of all articles with gender for first and last
author
977*100 / length(Author(records)) # 2.3% of all articles

# ----- Pubmed_data (each row is an article) -----

# Added 27/03/2020
```

```
first_and_last_combinations <- pubmed_data %>%
   filter(is.na(first_gender) == FALSE, is.na(last_gender) == FALSE) %>%
   mutate(first_and_last = paste0("first_", first_gender, "_last_", last_gender))

# Table
table(first_and_last_combinations$first_and_last)

# Save file
save(first_and_last_combinations, file =
"masterdata/first_and_last_combinations.RData")

# ----- Master data file -----

# Create
master <- pubmed_data %>%
   group_by(Year, Journal) %>%
   summarise() %>%
   mutate(first_male = 0,
              first_female = 0,
              last_male = 0,
              last_female = 0) %>%
   as.data.frame

# Auxiliar tables
pubmed_data_F <- pubmed_data %>%
   group_by(Year, Journal, first_gender) %>%
   summarise(n = n()) %>%
   as.data.frame
head(pubmed_data_F)

# Auxiliar tables
pubmed_data_L <- pubmed_data %>%
   group_by(Year, Journal, last_gender) %>%
   summarise(n = n()) %>%
   as.data.frame
head(pubmed_data_L)

# Update - first
```

```
for(i in 1:nrow(pubmed_data_F)){
    if(i %% 50 == 0) cat(".")
    for(j in 1:nrow(master)){
        if(pubmed_data_F[i, "Year"] == master[j, "Year"] & pubmed_data_F[i, "Journal"]
== master[j, "Journal"]){
            if(is.na(pubmed_data_F[i, "first_gender"]) == FALSE){
                if(pubmed_data_F[i, "first_gender"] == "male") master[j, "first_male"] =
pubmed_data_F[i, "n"]
                if(pubmed_data_F[i, "first_gender"] == "female") master[j, "first_female"]
= pubmed_data_F[i, "n"]
            }
            break()
        }
    }
}

# Update - last
for(i in 1:nrow(pubmed_data_L)){
    if(i %% 50 == 0) cat(".")
    for(j in 1:nrow(master)){
        if(pubmed_data_L[i, "Year"] == master[j, "Year"] & pubmed_data_L[i, "Journal"]
== master[j, "Journal"]){
            if(is.na(pubmed_data_L[i, "last_gender"]) == FALSE){
                if(pubmed_data_L[i, "last_gender"] == "male") master[j, "last_male"] =
pubmed_data_L[i, "n"]
                if(pubmed_data_L[i, "last_gender"] == "female") master[j, "last_female"] =
pubmed_data_L[i, "n"]
            }
            break()
        }
    }
}

head(master)

# Export
save(master, file = "master.RData")
write.csv(master, file = "master.csv", row.names = FALSE)
```

```
# ----- Results - Overall male-female ratio -----

table(pubmed_data$first_gender, useNA = "always")
table(pubmed_data$last_gender, useNA = "always")

# ----- Figures - Ratio male/female -----

# First author
df <- table(pubmed_data$first_gender, pubmed_data$Year)
df <- rbind(df, df["male",]/df["female",])
rownames(df)[3] <- "ratio_male_female"
df <- melt(df) %>% filter(X1 == "ratio_male_female")
names(df) <- c("ratio", "year", "value")
df

ggplot(df) +
    geom_line(aes(x = year, y = value), size = 0.75) +
    #dgeom_col(aes(x = year, y = value), size = 0.75, col = "black", fill = "black") +
    geom_hline(yintercept = 1, lty = 2, col = "gray60", size = 0.5) +
    ggtitle("Evolution of ratio male-female (First author)") +
    ylab("Ratio male-female") +
    ylim(c(0, 5)) +
    scale_x_continuous("Year", breaks = seq(2002, 2020, 2)) +
    theme_classic() +
    theme(plot.title = element_text(hjust = 0.5))
ggsave(filename = "plots/r_first.png", width = 200, height = 200, units = "mm")

# Save graph to combine
graph_first <- ggplot(df) +
    geom_line(aes(x = year, y = value), size = 0.75) +
    #dgeom_col(aes(x = year, y = value), size = 0.75, col = "black", fill = "black") +
    geom_hline(yintercept = 1, lty = 2, col = "gray60", size = 0.5) +
    ggtitle("Evolution of ratio male-female (First author)") +
    ylab("Ratio male-female") +
    ylim(c(0, 5)) +
    scale_x_continuous("Year", breaks = seq(2002, 2020, 2)) +
    theme_classic() +
    theme(plot.title = element_text(hjust = 0.5))

# Last author
df <- table(pubmed_data$last_gender, pubmed_data$Year)
df <- rbind(df, df["male",]/df["female",])
rownames(df)[3] <- "ratio_male_female"
df <- melt(df) %>% filter(X1 == "ratio_male_female")
names(df) <- c("ratio", "year", "value")
df

ggplot(df) +
    geom_line(aes(x = year, y = value), size = 0.75) +
    #geom_col(aes(x = year, y = value), size = 0.75, col = "black", fill = "black") +
    geom_hline(yintercept = 1, lty = 2, col = "gray60", size = 0.5) +
    ggtitle("Evolution of ratio male-female (Last author)") +
    ylab("Ratio male-female") +
    ylim(c(0, 5)) +
    scale_x_continuous("Year", breaks = seq(2002, 2020, 2)) +
    theme_classic() +
    theme(plot.title = element_text(hjust = 0.5))

ggsave(filename = "plots/r_last.png", width = 200, height = 200, units = "mm")

# Save graph to combine
graph_last <- ggplot(df) +
    geom_line(aes(x = year, y = value), size = 0.75) +
    #geom_col(aes(x = year, y = value), size = 0.75, col = "black", fill = "black") +
    geom_hline(yintercept = 1, lty = 2, col = "gray60", size = 0.5) +
    ggtitle("Evolution of ratio male-female (Last author)") +
    ylab("Ratio male-female") +
    ylim(c(0, 5)) +
    scale_x_continuous("Year", breaks = seq(2002, 2020, 2)) +
    theme_classic() +
    theme(plot.title = element_text(hjust = 0.5))

# Combine figures
library(ggpubr)
ggarrange(graph_first, graph_last)
```

```
ggsave(filename = "plots/r_first_last_combined.png", width = 400, height = 200,
units = "mm")

# ----- 21/09/2020 NEW figures - Ratio male/female -----

# First author
df <- table(pubmed_data$first_gender, pubmed_data$Year)
df <- rbind(df, df["male",]/df["female",])
rownames(df)[3] <- "ratio_male_female"
df <- melt(df) %>% filter(X1 == "ratio_male_female")
names(df) <- c("ratio", "year", "value")
df

# Last author
df2 <- table(pubmed_data$last_gender, pubmed_data$Year)
df2 <- rbind(df2, df2["male",]/df2["female",])
rownames(df2)[3] <- "ratio_male_female"
df2 <- melt(df2) %>% filter(X1 == "ratio_male_female")
names(df2) <- c("ratio", "year", "value")
df2

# cbind first and last
names(df) <- c("ratio", "year", "value_first")

df$year == df2$year

df <- cbind(df, "value_last" = df2$value)
head(df)

ggplot(df) +
    geom_line(aes(x = year, y = value_first, col = "Ratio male/female (first author)"),
size = 1.5) +
    geom_line(aes(x = year, y = value_last,    col = "Ratio male/female (last author)"),
size = 1.5) +
    geom_hline(yintercept = 1, lty = 2, col = "gray60", size = 0.5) +
    ggtitle("Evolution of ratio male-female") +
    ylab("Ratio male-female") +
    ylim(c(0, 5)) +
  scale_x_continuous("Year", breaks = seq(2002, 2020, 2)) +
  theme_classic() +
  theme(text = element_text(color = "black"),
        legend.title = element_blank(),
        legend.position="top") +
  theme(plot.title = element_text(hjust = 0.5))

ggsave(filename = "plots/20200921_r_first_and_last.png", width = 250, height = 200,
units = "mm")

# ----- Figures - Ratio male/female by journal -----

# Only 4 journals with more articles
table(pubmed_data$Journal) %>% sort(decreasing = TRUE)
journals <- table(pubmed_data$Journal) %>% sort(decreasing = TRUE) %>% head(4)
%>% names()

# journals <- c(journals, "N. Engl. J. Med.")

# First author
df <- pubmed_data %>%
  filter(is.na(first_gender) == FALSE) %>%
  filter(is.na(Journal) == FALSE) %>%
  filter(Journal %in% journals) %>%
  mutate(first_gender = as.character(first_gender)) %>%
  mutate(Journal = as.character(Journal)) %>%
  group_by(Year, Journal, first_gender) %>%
  summarise(n = n()) %>%
  mutate(n = as.double(n)) %>%
  as.data.frame()

# Calculate ratio
for(i in 2002:2019){
  for(j in 1:length(journals)){
    ratio <- df[df$Year == i & df$Journal == journals[j] & df$first_gender == "male",
"n"] /
      df[df$Year == i & df$Journal == journals[j] & df$first_gender == "female", "n"]
    aux <- data.frame(i, journals[j], "ratio", ratio)
```

```
        names(aux) <- names(df)
        df <- rbind(df, aux)
    }
}

df <- df %>% filter(first_gender == "ratio")
names(df)[4] <- "ratio"
head(df)

ggplot(df) +
    geom_line(aes(x = Year, y = ratio), size = 0.75) +
    geom_hline(yintercept = 1, lty = 2, col = "gray60", size = 0.5) +
    ggtitle("Evolution of ratio male-female (First author)") +
    scale_x_continuous("Year", breaks = seq(2002, 2020, 2)) +
    facet_wrap(vars(Journal)) +
    scale_y_continuous("Ratio male-female", breaks = 1:11) +
    theme_classic() +
    theme(plot.title = element_text(hjust = 0.5))

ggsave(filename = "plots/r_first_by_journal.png", width = 200, height = 200, units = "mm")

# Save graph to combine figures
graph_first_by_journal <- ggplot(df) +
    geom_line(aes(x = Year, y = ratio), size = 0.75) +
    geom_hline(yintercept = 1, lty = 2, col = "gray60", size = 0.5) +
    ggtitle("Evolution of ratio male-female (First author)") +
    scale_x_continuous("Year", breaks = seq(2002, 2020, 2)) +
    facet_wrap(vars(Journal)) +
    scale_y_continuous("Ratio male-female", breaks = 1:11) +
    theme_classic() +
    theme(plot.title = element_text(hjust = 0.5))

# Last author
df <- pubmed_data %>%
    filter(is.na(last_gender) == FALSE) %>%
    filter(is.na(Journal) == FALSE) %>%
    filter(Journal %in% journals) %>%
  mutate(last_gender = as.character(last_gender)) %>%
  mutate(Journal = as.character(Journal)) %>%
  group_by(Year, Journal, last_gender) %>%
  summarise(n = n()) %>%
  mutate(n = as.double(n)) %>%
  as.data.frame()

# Calculate ratio
for(i in 2002:2019){
  for(j in 1:length(journals)){
    ratio <- df[df$Year == i & df$Journal == journals[j] & df$last_gender == "male",
"n"] /
      df[df$Year == i & df$Journal == journals[j] & df$last_gender == "female", "n"]
    aux <- data.frame(i, journals[j], "ratio", ratio)
    names(aux) <- names(df)
    df <- rbind(df, aux)
  }
}

df <- df %>% filter(last_gender == "ratio")
names(df)[4] <- "ratio"
head(df)

ggplot(df) +
  geom_line(aes(x = Year, y = ratio), size = 0.75) +
  geom_hline(yintercept = 1, lty = 2, col = "gray60", size = 0.5) +
  ggtitle("Evolution of ratio male-female (Last author)") +
  scale_x_continuous("Year", breaks = seq(2002, 2020, 2)) +
  scale_y_continuous("Ratio male-female", breaks = 1:11) +
  facet_wrap(vars(Journal)) +
  theme_classic() +
  theme(plot.title = element_text(hjust = 0.5))

ggsave(filename = "plots/r_last_by_journal.png", width = 200, height = 200, units =
"mm")

# Save graph to combine figures
graph_last_by_journal <- ggplot(df) +
    geom_line(aes(x = Year, y = ratio), size = 0.75) +
    geom_hline(yintercept = 1, lty = 2, col = "gray60", size = 0.5) +
    ggtitle("Evolution of ratio male-female (Last author)") +
    scale_x_continuous("Year", breaks = seq(2002, 2020, 2)) +
    scale_y_continuous("Ratio male-female", breaks = 1:11) +
    facet_wrap(vars(Journal)) +
    theme_classic() +
    theme(plot.title = element_text(hjust = 0.5))

# Combine figures
library(ggpubr)
ggarrange(graph_first_by_journal, graph_last_by_journal)
ggsave(filename = "plots/r_first_last_combined_by_journal.png", width = 400, height
= 200, units = "mm")

# ----- 21/09/2020 NEW figures - Ratio male/female by journal -----

# Only 4 journals with more articles
table(pubmed_data$Journal) %>% sort(decreasing = TRUE)
journals <- table(pubmed_data$Journal) %>% sort(decreasing = TRUE) %>% head(4)
%>% names()

# journals <- c(journals, "N. Engl. J. Med.")

# First author
df <- pubmed_data %>%
    filter(is.na(first_gender) == FALSE) %>%
    filter(is.na(Journal) == FALSE) %>%
    filter(Journal %in% journals) %>%
    mutate(first_gender = as.character(first_gender)) %>%
    mutate(Journal = as.character(Journal)) %>%
    group_by(Year, Journal, first_gender) %>%
    summarise(n = n()) %>%
    mutate(n = as.double(n)) %>%
    as.data.frame()

# Calculate ratio
for(i in 2002:2019){
    for(j in 1:length(journals)){
        ratio <- df[df$Year == i & df$Journal == journals[j] & df$first_gender == "male",
"n"] /
            df[df$Year == i & df$Journal == journals[j] & df$first_gender == "female", "n"]
        aux <- data.frame(i, journals[j], "ratio", ratio)
        names(aux) <- names(df)
        df <- rbind(df, aux)
    }
}

df <- df %>% filter(first_gender == "ratio")
names(df)[4] <- "ratio_first"
head(df)

# Last author
df2 <- pubmed_data %>%
    filter(is.na(last_gender) == FALSE) %>%
    filter(is.na(Journal) == FALSE) %>%
    filter(Journal %in% journals) %>%
    mutate(last_gender = as.character(last_gender)) %>%
    mutate(Journal = as.character(Journal)) %>%
    group_by(Year, Journal, last_gender) %>%
    summarise(n = n()) %>%
    mutate(n = as.double(n)) %>%
    as.data.frame()

# Calculate ratio
for(i in 2002:2019){
    for(j in 1:length(journals)){
        ratio <- df2[df2$Year == i & df2$Journal == journals[j] & df2$last_gender ==
"male", "n"] /
            df2[df2$Year == i & df2$Journal == journals[j] & df2$last_gender == "female",
"n"]
        aux <- data.frame(i, journals[j], "ratio", ratio)
        names(aux) <- names(df2)
        df2 <- rbind(df2, aux)
    }
}
```

```
df2 <- df2 %>% filter(last_gender == "ratio")
names(df2)[4] <- "ratio_last"
head(df2)

# cbind first and last
df$Year == df2$Year
df$Journal == df2$Journal
df <- cbind(df, "ratio_last" = df2$ratio_last)
head(df)

ggplot(df) +
    geom_line(aes(x = Year, y = ratio_first, col = "Ratio male/female (first author)"),
size = 1) +
    geom_line(aes(x = Year, y = ratio_last, col = "Ratio male/female (last author)"), size
= 1) +
    geom_hline(yintercept = 1, lty = 2, col = "gray60", size = 0.5) +
    ggtitle("Evolution of ratio male-female") +
    scale_x_continuous("Year", breaks = seq(2002, 2020, 2)) +
    scale_y_continuous("Ratio male-female", breaks = 1:11) +
    facet_wrap(vars(Journal)) +
    theme_classic() +
    theme(text = element_text(color = "black"),
                legend.title = element_blank(),
                legend.position="top") +
    theme(plot.title = element_text(hjust = 0.5))

ggsave(filename = "plots/20200921_r_first_and_last_by_journal.png", width = 250,
height = 200, units = "mm")

# ----- Figures - Number -----

# First author
df <- table(pubmed_data$first_gender, pubmed_data$Year) %>% melt
names(df) <- c("gender", "year", "value")
df

ggplot(df) +
    geom_line(aes(x = year, y = value, color = gender), size = 0.75) +
    geom_hline(yintercept = 1, lty = 2, col = "gray60", size = 0.5) +
    ggtitle("Evolution of number of articles by gender (First author)") +
    ylab("Number of articles") +
    scale_x_continuous("Year", breaks = seq(2002, 2020, 2)) +
    theme_classic() +
    theme(plot.title = element_text(hjust = 0.5))
ggsave(filename = "plots/n_first.png", width = 200, height = 200, units = "mm")


# Last author
df <- table(pubmed_data$last_gender, pubmed_data$Year) %>% melt
names(df) <- c("gender", "year", "value")
df

ggplot(df) +
    geom_line(aes(x = year, y = value, color = gender), size = 0.75) +
    geom_hline(yintercept = 1, lty = 2, col = "gray60", size = 0.5) +
    ggtitle("Evolution of number of articles by gender (Last author)") +
    ylab("Number of articles") +
    scale_x_continuous("Year", breaks = seq(2002, 2020, 2)) +
    theme_classic() +
    theme(plot.title = element_text(hjust = 0.5))
ggsave(filename = "plots/n_last.png", width = 200, height = 200, units = "mm")

# Trends Ratio male/female over time (accouting for overdispersion)
library(sandwich)
load("First.RData")
load("Last.RData")

# First author
trendF <- glm(n ~ I(first_gender) + Year + first_gender:Year,
family=poisson(link="log"), data=pubmed_data_F)
summary(trendF)
confint(trendF)
# Robust Standard Errors (accounting for overdispersion)
cov.F <- vcovHC(trendF, type="HC0")
stdf.err <- sqrt(diag(cov.F))
```

```
rf.est <- cbind(Estimate= coef(trendF), "Robust SE" = stdf.err,
                    "Pr(>|z|)" = 2 * pnorm(abs(coef(trendF)/stdf.err),
lower.tail=FALSE),
                    LL = coef(trendF) - 1.96 * stdf.err,
                    UL = coef(trendF) + 1.96 * stdf.err)
rf.est

# Last author
trendL <- glm(n ~ I(last_gender) + Year + last_gender:Year, family=poisson(link="log"),
data=pubmed_data_L)
summary(trendL)
confint(trendL)

# Robust Standard Errors (accounting for overdispersion)
cov.L <- vcovHC(trendF, type="HC0")
stdl.err <- sqrt(diag(cov.L))
rl.est <- cbind(Estimate= coef(trendL), "Robust SE" = stdl.err,
                     "Pr(>|z|)" = 2 * pnorm(abs(coef(trendL)/stdl.err),
lower.tail=FALSE),
                     LL = coef(trendL) - 1.96 * stdl.err,
                     UL = coef(trendL) + 1.96 * stdl.err)
rl.est
```

Analysis Codes in Stata

```
gen first_female =0
gen first_male =0
replace first_male=1 if first_gender=="male"
gen last_female =0
replace last_female=1 if last_gender=="female"
replace first_female=1 if first_gender=="female"
gen last_male =0
replace last_male=1 if last_gender=="male"

//to drop these two journals because their articles are problematic
drop if journal =="Ann. Oncol."
drop if journal =="JAMA Oncol"



// percentages of female first authors and female last authors
gen percent_ff = first_female/(first_female+first_male)
bys journal: sum percent_ff
bys year: sum percent_ff
bys year: sum percent_ff if last_gender!="NA" & first_gender!="NA"
sum percent_ff if last_gender!="NA" & first_gender!="NA"

gen percent_fs = last_female/(last_female+last_male)
bys journal: sum percent_fs
bys year: sum percent_fs
bys year: sum percent_fs if last_gender!="NA" & first_gender!="NA"
sum percent_fs

gen percent_mf = first_male/(first_female+first_male)
bys journal: sum percent_mf
bys year: sum percent_mf
bys year: sum percent_mf if last_gender!="NA" & first_gender!="NA"
sum percent_mf if last_gender!="NA" & first_gender!="NA"

gen percent_ms = last_male/(last_female+last_male)
bys journal: sum percent_ms
bys year: sum percent_ms
bys year: sum percent_ms if last_gender!="NA" & first_gender!="NA"
sum percent_ms


tab last_female if last_gender!="NA" & first_gender!="NA"
by year: tab last_female if last_gender!="NA" & first_gender!="NA"
tab last_female if last_gender!="NA" & first_gender!="NA"
tab last_male if last_gender!="NA" & first_gender!="NA"
tab last_male
tab last_female if last_gender!="NA" & first_gender!="NA"
tab last_male if last_gender!="NA" & first_gender!="NA"
by year: tab last_male if last_gender!="NA" & first_gender!="NA"
by year: tab last_male
by year: tab last_male if last_gender!="NA" & first_gender!="NA"
by year: tab last_male if last_gender!="NA" & first_gender!="NA"
by year: tab last_male if last_gender!="NA" & first_gender!="NA"
by year: tab last_male if last_gender!="NA" & first_gender!="NA"
tab last_male if last_gender!="NA" & first_gender!="NA"
by year: tab last_male if last_gender!="NA" & first_gender!="NA"
by year: tab last_female if last_gender!="NA" & first_gender!="NA"


//Chi Sq test in Table 1
gen new_first_gen = .
replace new_first_gen =1 if first_gender=="male"
replace new_first_gen =0 if first_gender=="female"
replace new_first_gen = 2 if new_first_gen == .
tab journal new_first_gen , chi

gen new_last_gen = .
replace new_last_gen =1 if last_gender=="male"
replace new_last_gen =0 if last_gender=="female"
replace new_last_gen = 2 if new_last_gen == .
tab journal new_last_gen , chi

//Chi Sq test for trend in Table 2    (Patrick Royston's ptrend command in Stata)
gen new_first_male = 1 if    new_first_gen ==1
gen new_first_female = 1 if    new_first_gen ==0
gen new_last_male = 1 if    new_last_gen ==1
gen new_last_female = 1 if    new_last_gen ==0
replace new_first_male = 0 if new_first_male ==.
replace new_first_female = 0 if new_first_female ==.
replace new_last_male = 0 if new_last_male ==.
replace new_last_female = 0 if new_last_female ==.
ptrend new_first_male new_first_female year
ptrend new_last_male new_last_female year



//gender ratios
gen ratio_ff = first_male/first_female
bys journal: sum ratio_ff
bys year: sum ratio_ff
sum ratio_ff

gen ratio_fs = last_male/last_female
bys journal: sum ratio_fs
bys year: sum ratio_fs
sum ratio_fs

poisson ratio_ff year, irr
poisson ratio_fs year, irr

bys journal: sum rmff      //rmff = male/female for 1st author
bys journal: sum rmfs      //rmfs = male/female for last author

mean rmff rmfs
        estimates store mean
        coefplot (matrix(C[,1]), ci((C[,2]))) (mean)

graph bar (mean)    rmff, over(year) by(journal)    scheme(emma)
graph bar (mean)    rmfs, over(year) by(journal)    scheme(emma)

graph bar (mean)    rmff, over(year)    scheme(emma)
graph bar (mean)    rmfs, over(year)    scheme(emma)
```

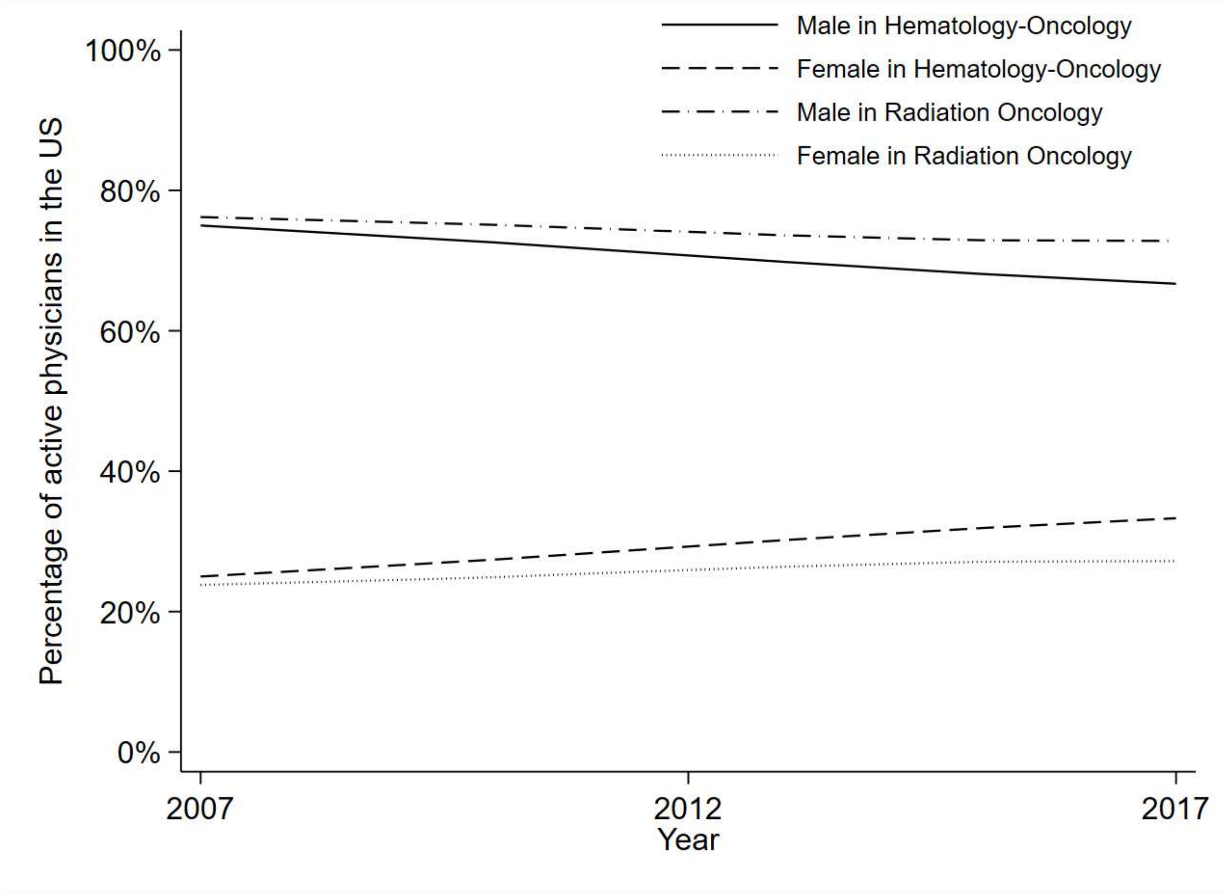
Supplementary Online Content

Trends in Gender of Authors of Original Researches in Oncology Among Major Medical Journals

Online Supplemental Table 1. Number and Percentage of Active Physicians by Sex and Specialty in the US, 2007 to 2017

Years	Hematology-Oncology		Radiation Oncology	
	Male, No. (%)	Female, No. (%)	Male, No. (%)	Female, No. (%)
2007	8 837 (75)	2 952 (25)	3 205 (76)	1 003 (24)
2010	9 241 (73)	3 483 (27)	4 456 (75)	1 108 (25)
2013	9 600 (70)	4 155 (30)	3 443 (74)	1 237 (26)
2015	9 846 (68)	4 611 (32)	3 533 (73)	1 312 (27)
2017	10 268 (67)	5 122 (33)	3 661 (73)	1 365 (27)

Online Supplemental Figure 1. The temporal trend for the percentages of active physicians in the US from years 2007 to 2017.



Online Supplemental Table 2. Distribution of First and Last Author by Genders, Q4 rank oncology journals in Scimago Journal Rank for 2020.

	First author, n (%)	Last author, n (%)
Female	103 (35.5)	67 (23.1)
Male	163 (56.2)	203 (70.0)
Unknown	24 (8.3)	20 (6.9)
Total	290 (100.0)	290 (100.0)

Note: In 250 articles of the 290 original articles, genders were identified for **both** the first and last author with the following distribution: first male, last male: 121 (49.0%), first female, last male: 66 (26.7%), first female, last female: 34 (13.8%), first male, last female: 29 (10.5%). The gender combinations of first and last authors in 2019 in the main analysis are: first male, last male: 228 (46.9%), first female, last male: 94 (20.4%), first female, last female: 71 (15.4%), first male, last female: 67 (14.6%)

Online Supplemental Table 3. Distribution of gender in single-authored articles by year.

Year	Number of single-authored articles by males	Number of single-authored articles by females	Male/female authorship ratio for single-authored articles
2002	49	8	6.1:1
2003	43	15	2.9:1
2004	44	6	7.3:1
2005	39	9	4.3:1
2006	36	14	2.6:1
2007	44	8	5.5:1
2008	54	19	2.8:1
2009	37	14	2.6:1
2010	37	24	1.5:1
2011	44	17	2.6:1
2012	44	20	2.2:1
2013	22	15	1.5:1
2014	25	25	1.0:1
2015	50	28	1.8:1
2016	46	41	1.1:1
2017	29	27	1.1:1

2018	17	16	1.1:1
2019	5	6	0.8:1
2002-2019	**665**	**312**	**2.1**:1