Gender Diverse Teams Produce More Innovative and Influential Ideas in Medical Research

Yang Yang, PhD[1,2], Teresa Woodruff, PhD[3], Yuan Tian, PhD[4], Benjamin F. Jones, PhD[2,5,6], Brian Uzzi, PhD[2,5,7*]

Syracuse University[1]
Northwestern Institute on Complex Systems[2]
Michigan State University[3]
New York University[4]
Kellogg School of Management, Northwestern University[5]
National Bureau of Economic Research[6]
McCormick School of Engineering, Northwestern University[7]

*corresponding author: Brian Uzzi, email: uzzi@northwestern.edu

Words Count: 2,208

Key Points

Question: Is a research team's gender composition related to its innovativeness and impact?

Findings: Gender diverse teams are significantly more innovative and impactful than same gender teams. The advantages of gender diverse teams generalize to medical science subfields, hold whether the gender of the diverse team's leader is a man or women, and increase as the team's gender mix becomes more balanced. Nevertheless, gender diverse teams are still significantly underrepresented in medical science, potentially inhibiting the advancement of science and women in science.

Meaning: Gender diverse teams play an unrecognized key role in medical science's innovative and high impact research.

Abstract

Importance: This work addresses questions about how concurrent increases in teamwork[1] and women in science[2,3] have changed how the gender composition of a team positively coincides with its research innovativeness, citation impact, and potential for recognizing women's new contributions to science[4-11].

Objectives: Statistically examine the link between gender diverse research teams and team performance relative to same gender teams.

Design: Systemic study of team science using a massive, longitudinal scientometric data covering all 6.6 million publications from 15,033 medical science journals and published by 3.2 million female and 4.4 million male authors, 2000-2019.

Setting: International sample of 7.6 million medical researchers.

Participants: All authors of academic medical research published from 2000-2019 in 15,033 medical journals worldwide as curated in Microsoft Academic Graph (MAG)[12].

Design: We regressed each publication's innovativeness and normalized citation impact on the gender composition of the team using fixed-effects, matching, and null models to control for confounds due to institutional, journal, time, and team and individual performance variables.

Main outcomes and measures: Research performance is measured in terms of innovativeness and impact and the representativeness of gender diverse teams.

Results: Gender diverse teams publish papers that are up to 7% more innovative than same gender teams, controlling for confounds due to the authors' past performance, prestige, subfield, and journal. Similarly, gender diverse teams can be twice as likely to publish hit papers (top 5% of citations) relative to the base rate and 16.7% more likely than same-gender teams. The results strongly generalize to the subfields of medical science, whether the diverse team's leader is a woman or a man, and further strengthen as gender diverse teams become more gender balanced.

Conclusions: Gender diverse teams outperform same gender teams in the creation of innovative and high impact papers, an advantage that generalizes to medical science subfields, holds whether the gender of the diverse team's leader is a man or women, and increases as the team's gender mix becomes more balanced. Despite their advantages, gender diverse teams are significantly underrepresented in medical science, potentially constraining scientific advances and under-utilizing human capital.

Introduction

In the past 20 years, medicine has undergone two transformations that may potentially be remaking medical research outcomes. One change involves a shift from individual to team science[1]. Rising teamwork levels are broadly documented in science. Teams produce more highly cited research and influence stratification[13], yet the implications of team science for medical research has yet to be fully investigated[14,15]. The second change is the increased participation of women in medical science[3,4,16]. Perhaps more than any other branch of science, the last decade has seen women's participation rates exceed the participation rates of men in graduate and post-doctoral training in 60%-to-40% and 54%-to-46% ratios, respectively[17].

These systemic changes create a need to study how a research team's gender composition[18,19] is related to research performance[2,20] and acknowledgement of credit[11,21-24]. Is a medical science team's performance correlated with gender diversity? Are the implied gains of gender diverse teams measurable? Does team performance vary with a leader's gender?

New experiments suggest that gender diverse teams are potentially consequential to the innovativeness and impact of research because they may combine the benefits of teams and diversity. Lab experiments indicate that mixed-gender teams exchange more diverse information than same gender teams and can do better at general problem-solving tasks than all-male or all-female teams with equivalent IQ levels[8]. Yet, when gender stereotypes and inequities exist in professional settings like medicine, same gender relationships are preferred[25-27], reducing workgroup diversity and equitable representation[28]. These theoretical and practical repercussions merit new empirical analysis of the incidence and impact of gender diverse teams on medical science research and practice.

Methods

Sample

We conduct the first large-scale systemic investigation of the performance of gender diverse research teams in the medical sciences. Our original field-wide dataset has over 6.6 million research publications by 3,225,279 female and 4,386,020 male scientists in 15,033 medical science journals from 2000-2019 as recorded in Microsoft Academic Graph[12]. Microsoft Academic Graph (MAG) is a scientific publication database that records journal article's bibliographic information (title, journal, journal field, volume, issue, page, publication date), authorship (name), author affiliations (name, webpage, and wikipage), and citation links to other papers in the database.

Variables

Our main independent variable is team gender diversity, and our two outcome variables are research innovativeness and research impact. All variables were constructed with MAG data.

Team Gender Diversity: To measure the gender composition in a scientific team, we use a binary variable "mixed-gender". A mixed-gender team has both men and women. Otherwise, it is a same-gender team. We also use a continuous variable to evaluate the gender composition of a scientific team in finer detail[26] that takes the form

$$g_i = -p_f log_2(p_f) - (1 - p_f) log_2(1 - p_f) \quad (1)$$

where p_f indicates the portion of female authors in a team i. The value of g_i ranges from 0 to 1. When the value of g_i is low, either women or men are majority of a team. When $g_i = 0$, the team is either an all-women team or an all-men team. By contrast, when the value of g_i is high, women and men have roughly equivalent presence in the team. When $g_i = 1$, the team has 50% women and 50% men. A given author's gender was estimated using an accepted and validated algorithm[29]. Further supplementary validity tests demonstrated that algorithm's results are robust to measurement error (see **SM S1.2** and **Table S3** & **S4**).

Innovativeness of Scientific Papers: We use a popular and validated innovation measure that uses a paper's listed references as an indicator of its mixture of knowledge[30,31]. The measure considers papers with statistically atypical combinations of references to be innovative because they create new combinations of knowledge that have not been joined, or rarely joined, in previous research. By contrast, papers that have statistically common combinations of references represent conventional, familiar groupings of knowledge. To compute a paper's innovativeness in terms of novelty vs conventionality, we use a z-score. The z-score is computed from the observed frequency of journal pairs that appear within a paper's reference list and the expected distribution of journal pairings created by randomized citation networks. When a z-score is less than zero, the combination of prior work is considered innovative (i.e., a novel pairing of ideas) and when the z-score is above zero, the combination of prior work is considered conventional (i.e., a common pairing of ideas). A paper is defined as innovative if its novelty z-score is smaller than zero (see **SM S1.2.2** for detail).

Impact of Scientific Papers: The MAG database tracks a paper's annual citations. We define high impact papers as those in the top 5% of citation for papers published in the same year. To provide a fair and comparable measure, we normalize paper i's final citations by the publication year (see **SM S1.2.3** for detail).

Analysis

To quantify the link between team diversity, leadership, and performance, we used fixed effects regressions. The regressions control for confounds due to the authorship team's size, leadership, and institutional prestige rank, year, journal quality, prior citation impact, average team age, and individual fixed effects[1,13,30,32,33] and were run for the full data. To control for subfield differences, we also run a separate regression for 45 medical subfields (see **SM S2** for detail). Alternative measurements and null models confirm the results (see **SM S1** and **S3** for detail).

Results

Figure 1a shows team size trends in medical science. Large teams (6+ authors) are now the modal authorship form. Today almost 50% of all papers are by large teams, and the shares of papers at all other team sizes are concomitantly declining[1,34]. The share of publications by gender diverse teams is also changing[35]. We measure gender diversity two ways. First, "mixed-gender" is a binary variable equal to 0 if the team is all the same gender and 1 if the team has both female and male authors. Second, "gender diversity" is a continuous variable that rises from 0 when the team is all the same gender to 1 when the team includes equal numbers of women and men. **Supplemental Materials S1.2.4** describes the measures in detail. **Figures 1b** and **1d** indicate that the share of papers by mixed-gender and gender-balanced teams respectively is growing with time.

However, **Figure 1c** and **1e** show that gender diverse are significantly underrepresented (means with 95% CI shown, all p-values < 0.01). The expected level of gender diverse teams was computed using a null model. The model holds constant the yearly observed share of female authors in medical science (i.e., ~45% in 2019), the coauthorship network, team size, and number of publications and randomly interchanges on a yearly basis female and male authors who are matched on having the same first year of publication, total publications, and country into simulated groups (see **SM S3.1** for details).

Gender Diverse Teams and Innovation and Impact

Despite their underrepresentation, gender diverse teams are more innovative and impactful than same gender teams. Innovativeness is measured by whether a paper combines past knowledge in a novel way relative to past combinations and is updated yearly to account for the evolving corpus of research and how it has been combined[30]. Overall, 44% of papers in our data are defined as innovative. High impact papers are in the top 5% of citation for papers normalized by yearly average citations (see **SM S1.2**).

Figures 2a and **2b** show the predicted margins of a paper's novelty and citation impact conditional on the authorship team's gender diversity and team size, while controlling for institutional prestige, the authors' prior citation impact (the mean for the team members and the prior citation impact of the first author and the last author separately), the authors' career age (the mean for the team members and that of the first author and last author separately), the gender of the first and last authors, journal-year fixed effects, and individual fixed effects (see **SM S2** for detail).

Figure 2a demonstrates that mixed-gender teams are significantly more likely to publish more novel papers than same-gender teams (two sample t-test, p-value < 0.001). For example, large (6+) mixed-gender team are 9.1% more likely to publish a novel article than the base rate [(0.48-0.44)/0.44]. Given that novelty positively correlates with team size[1,30,36], the substantial added explanatory power of mixed- vs. same-gender teams seen in **Figure 2** is striking. Proportionally, the increase in novelty for mixed-gender teams of 6+ relative to same-gender teams is equivalent to the increase in novelty obtained by doubling a same-gender team size from 2 to 4.

Figure 2b shows that mixed-gender teams are significantly more likely to publish a citation hit than same gender teams. Comparing increases in publishing a hit to the baseline rate, large (6+) mixed-gender teams are 16.7% [(10.5%-9.0%)/9.0%) =16.7%] more likely to publish a hit paper than same gender teams of equal size (two sample t-test, p-value < 0.001).

The results replicate when we examine the link between the continuous measure of team gender balance and performance. We find that as team composition moves from same gender to gender balanced, the positive impact on a paper's innovativeness and impact significantly increases irrespective of team size. **SM Tables S1** and **S2** present regression details and exact significance levels while controlling for confounds.

Lastly, we examined the generalizability of the link between team gender diversity and performance. First, we tested whether the gender of the team's first or last author matters[37]. We found that gender diverse teams led by men or women alike produce significantly more novel and higher impact research than same gender teams (two sample t-test, $p < 0.001$, see **SM S4**). Second, we tested whether the benefits of gender diverse teams generalize across medical research subfields. We grouped papers by their MAG designated primary subfield and ran a separate regression for each subfield using the same specification as described in **Figure 2**.

Figure 3 shows that the findings strongly generalize across subfields. The y-axis shows the regression coefficient value with 95% CIs when innovativeness (**Figure 3a**) and citation impact (**Figure 3b**) are regressed on the team's "mixed-gender" variable for 45 separate subfields (x-axis). **Figure 3a** and **3b** demonstrate that the team's gender diversity significantly and positively predicts a team's novelty and impact for a significant majority of subfields (binomial test $p < 0.03$ for mixed-gender and $p < 0.001$ for gender balance) with the smallest subfields exhibiting a noisy relationship. The continuous measure of gender balance replicates the findings of the mixed-gender measure (see **SM S2.6**).

Discussion

This research presents striking performance advantages of gender diverse teams relative to same-gender teams that generalizes broadly across subfields and whether the diverse team's leader is a woman or a man. The large body of work documenting the differences in the careers and advancement of women scientists has raised awareness of gender inequities that inhibit science, reduce workplace fairness, and require new policy[2-4]. This work reveals new, team-level patterns of gender relations in science. Teams that combine the efforts and talents of men and women scientists do better than either all men or all women teams. Gender diverse teams publish papers that are up to 7% more innovative and 16.7% more likely to be citation hits than same-gender teams, controlling for confounds due to the authors' past performance, prestige, subfield, and journal. These results cannot be explained by the increased frequency of teamwork, team size, or the surge in women's participation in medical science. Thus, this work provides a new perspective on the potentially transformative benefits of gender diversity in science.

While the research finds that gender diverse research teams have robust performance advantages over teams of all women or all men, our analysis also reveals that gender diverse teams remain significantly underrepresented in medical science. This suggests that medical science may have the potential to speed breakthroughs by breaking down barriers to the formation of gender diverse teams.

Relatedly, the underrepresentation of gender-diverse teams highlights earlier work that has documented the inequities women experience in relation to accurate perceptions of credit[11]. The reasons for underrepresentation may relate to gender inequities in grants[5], prizes[4], leaky pipelines[2], and credit allocation in teams[3,24] that have been identified elsewhere but bear further study and evaluation from the lens of team gender diversity. Bias in team formation may be related to these other challenges and point to mechanisms and options for instituting practices that can advance gender balance[11,38]. For example, adopting practices for listing each author's contribution in publications can further transparency, accountability, and fairness[11] that can otherwise inhibit team formation.[38] In lab experiments, it has been shown that misperceptions can be ameliorated by providing feedback about individual team member performance[23]. Similarly, research examining causal tests of gender diversity in teams and mechanisms that inhibit gender-diverse team assembly is welcomed and could clarify actionable practices and support new policy[34].

Conclusions

Gender diverse teams are significantly more likely to publish research that innovatively combines existing ideas in new ways and is more influential than research by equivalent same gender teams. The advantages of mixed-gender teams generalize to medical science subfields, hold whether the gender of the diverse team's leader is a man or women, and increase as the team's gender mix becomes more balanced. Nevertheless, gender diverse teams are significantly underrepresented, potentially constraining scientific advances. More generally, this work expands science of science studies from analyses of gender differences to analyses of gender complementarities, recognizing distinctive advantages in the scientific outcomes from gender-diverse teams.

References

1. Wuchty S, Jones BF, Uzzi B. The increasing dominance of teams in production of knowledge. Science 2007;316:1036-9.
2. Etzkowitz H, Kemelgor C, Neuschatz M, Uzzi B, Alonzo J. The paradox of critical mass for women in science. Science 1994;266:51-3.
3. Etzkowitz H, Kemelgor C, Uzzi B. Athena Unbound: The Advancement of Women in Science and Technology. Cambridge: Cambridge University Press; 2000.
4. Ma Y, Oliveira DFM, Woodruff TK, Uzzi B. Women who win prizes get less money and prestige. Nature Publishing Group; 2019.
5. Oliveira DFM, Ma Y, Woodruff TK, Uzzi B. Comparison of National Institutes of Health Grant Amounts to First-Time Male and Female Principal Investigators. JAMA 2019;321:898-900.
6. Ma Y, Mukherjee S, Uzzi B. Mentorship and protégé success in STEM fields. Proceedings of the National Academy of Sciences 2020.
7. Woolley AW, Gerbasi ME, Chabris CF, Kosslyn SM, Hackman JR. Bringing in the experts - How team composition and collaborative planning jointly shape analytic effectiveness. Small Group Research 2008;39:352-71.
8. Woolley AW, Chabris CF, Pentland A, Hashmi N, Malone TW. Evidence for a collective intelligence factor in the performance of groups. Science 2010;330:686-8.
9. Woolley AW, Bear JB, Chang JW, DeCostanza AH. The effects of team strategic orientation on team process and information search. Organizational Behavior and Human Decision Processes 2013;122:114–26.
10. Nielsen MW, Alegria S, Börjeson L, et al. Opinion: Gender diversity leads to better science. Proceedings of the National Academy of Sciences 2017;114:1740-2.
11. Sugimoto C. Is science built on the shoulders of women? a study of gender differences in contributorship. Acad Med 2016;91.
12. Wang K, Shen Z, Huang C-Y, et al. A Review of Microsoft Academic Services for Science of Science Studies. Frontiers in Big Data 2019;2:45.
13. Jones BF, Wuchty S, Uzzi B. Multi-University Research Teams: Shifting Impact, Geography, and Stratification in Science. Science 2008;322:1259-62.
14. Fontanarosa P, Bauchner H, Flanagin A. Authorship and team science. Jama 2017;318:2433-7.
15. Falk-Krzesinski HJ, Börner K, Contractor N, et al. Advancing the science of team science. Clinical and translational science 2010;3:263-6.
16. Ataman LM, Ma Y, Duncan FE, Uzzi B, Woodruff TK. Quantifying the growth of oncofertility. Biology of reproduction 2018;99:263-5.
17. U.S. Medical School Faculty Trends: Percentages. 2020. at https://www.aamc.org/data-reports/faculty-institutions/interactive-data/us-medical-school-faculty-trends-percentages.)
18. Long JS. Measures of sex differences in scientific productivity. Social Forces 1992;71:159-78.
19. Bear JB, Woolley AW. The role of gender in team collaboration and performance. Interdisciplinary science reviews 2011;36:146-53.
20. Campbell LG, Mehtani S, Dozier ME, Rinehart J. Gender-heterogeneous working groups produce higher quality science. PloS one 2013;8:e79147.
21. Jagsi RG, Elizabeth A; Worobey, Cynthia Coooper; Henault, Lori E; Chang, Yuchiao; Starr, Rebecca; Tarbell, Nancy J; Hylek, Elaine M. The gender gap in authorship of academic medical literaturea 35-year perspective. New England Journal of Medicine 2006;355:281-7.

22. Patton EWG, Kent A; Jones, Rochelle D; Stewart, Abigail; Ubel, Peter A; Jagsi, Reshma. Differences in mentor-mentee sponsorship in male vs female recipients of national institutes of health grants. JAMA Internal Medicine 2017;177:580-2.

23. Heilman ME, Haynes MC. No credit where credit is due: attributional rationalization of women's success in male-female teams. Journal of applied Psychology 2005;90:905.

24. Sarsons H. Recognition for group work: Gender differences in academia. American Economic Review 2017;107:141-45.

25. Gorman EH. Gender stereotypes, same-gender preferences, and organizational variation in the hiring of women: Evidence from law firms. American Sociological Review 2005;70:702-28.

26. Yang Y, Chawla NV, Uzzi B. A network's gender composition and communication pattern predict women's leadership success. Proceedings of the National Academy of Sciences 2019;116:2033-8.

27. DeCastro R, Sambuco D, Ubel PA, Stewart A, Jagsi R. Mentor networks in academic medicine: moving beyond a dyadic conception of mentoring for junior faculty researchers. Academic medicine: journal of the Association of American Medical Colleges 2013;88:488.

28. Rivera LA. Hiring as cultural matching: The case of elite professional service firms. American Sociological Review 2012;77:999-1022.

29. Namsor. Available from: https://github.com/namsor/namsor-api. Retrieved October 30, 2020; 2020.

30. Uzzi B, Mukherjee S, Stringer M, Jones B. Atypical combinations and scientific impact. Science 2013;342:468-72.

31. Leahey E, Beckman CM, Stanko TL. Prominent but less productive: The impact of interdisciplinarity on scientists' research. Administrative Science Quarterly 2017;62:105-39.

32. Börner K, Contractor N, Falk-Krzesinski HJ, et al. A Multi-Level Systems Perspective for the Science of Team Science. Science Translational Medicine 2010;2:49cm24.

33. Jones B, Reedy EJ, Weinberg BA. Age and Scientific Genius. In: Simonton DK, ed. The Wiley Handbook of Genius: Wiley-Blackwell; 2014.

34. Fortunato S, Bergstrom CT, Börner K, et al. Science of science. Science 2018;359.

35. Ma Y, Uzzi B. Scientific prize network predicts who pushes the boundaries of science. Proceedings of the National Academy of Sciences 2018;115:12608-15.

36. Boudreau KJ, Guinan EC, Lakhani KR, Riedl C. Looking across and looking beyond the knowledge frontier: Intellectual distance, novelty, and resource allocation in science. Management Science 2016;62:2765-83.

37. Chowdhary M, Chowdhary A, Royce TJ, et al. Women's Representation in Leadership Positions in Academic Medical Oncology, Radiation Oncology, and Surgical Oncology Programs. JAMA Network Open 2020;3:e200708-e.

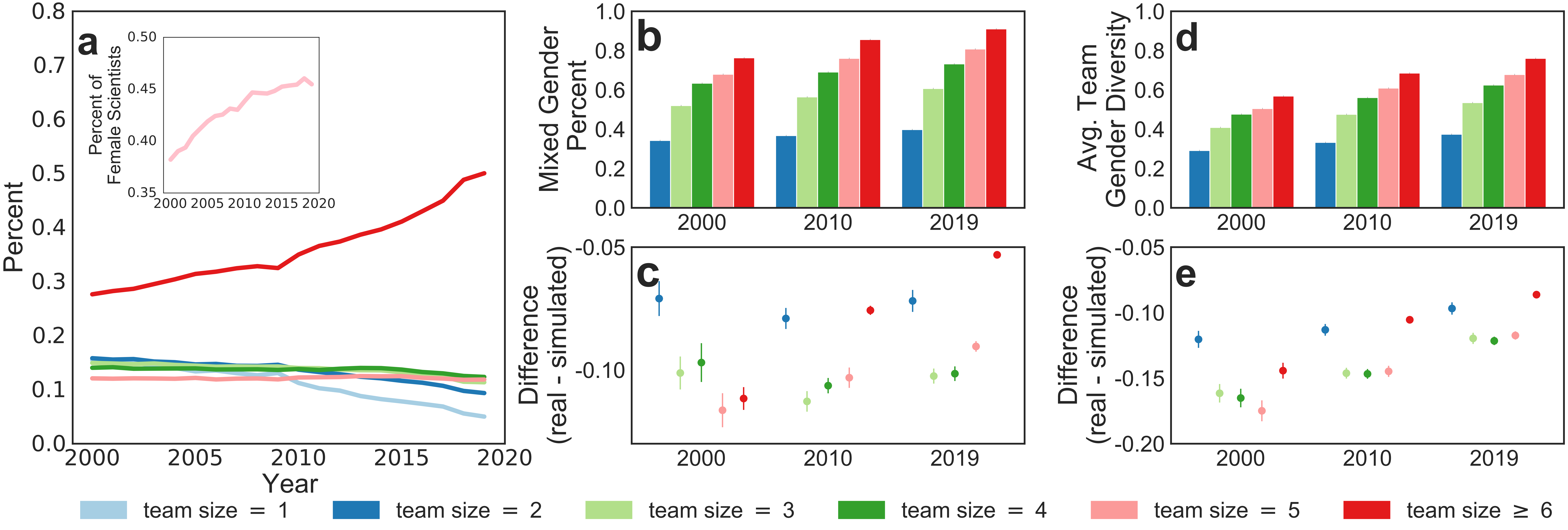
38. Ahmadpoor M, Jones BF. Decoding team and individual impact in science and invention. Proceedings of the National Academy of Sciences 2019;116:13885-90.

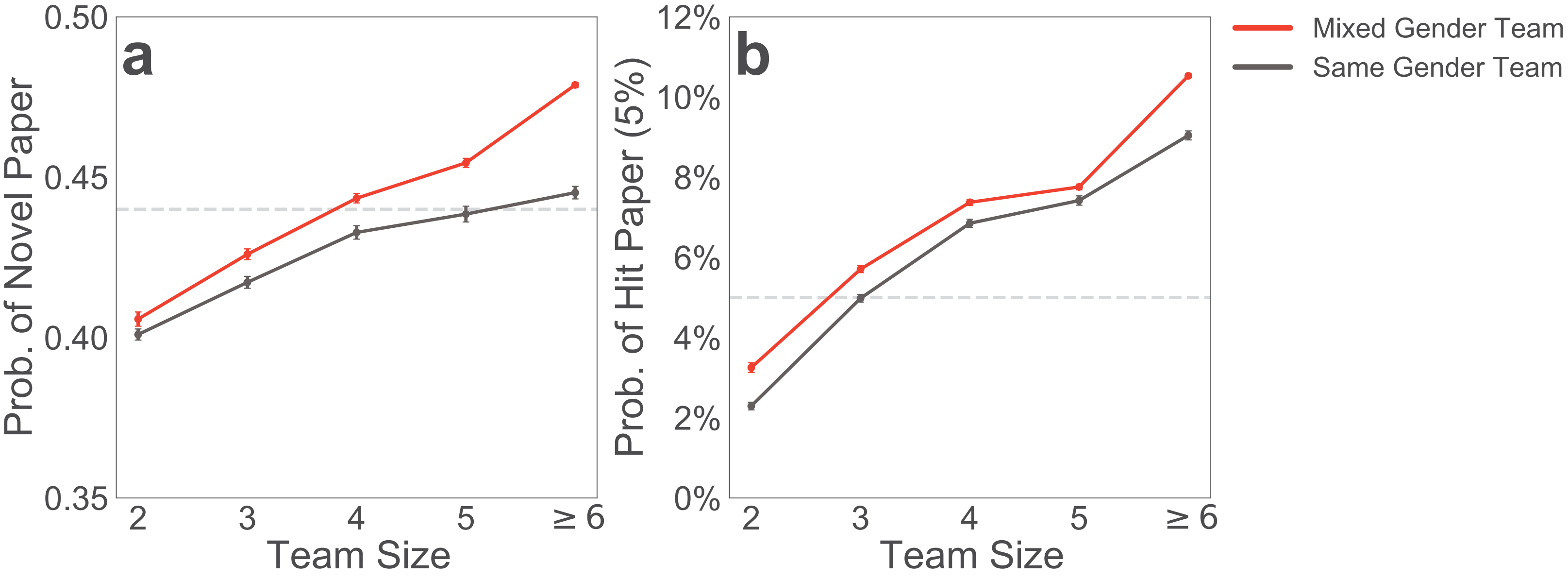
Figures

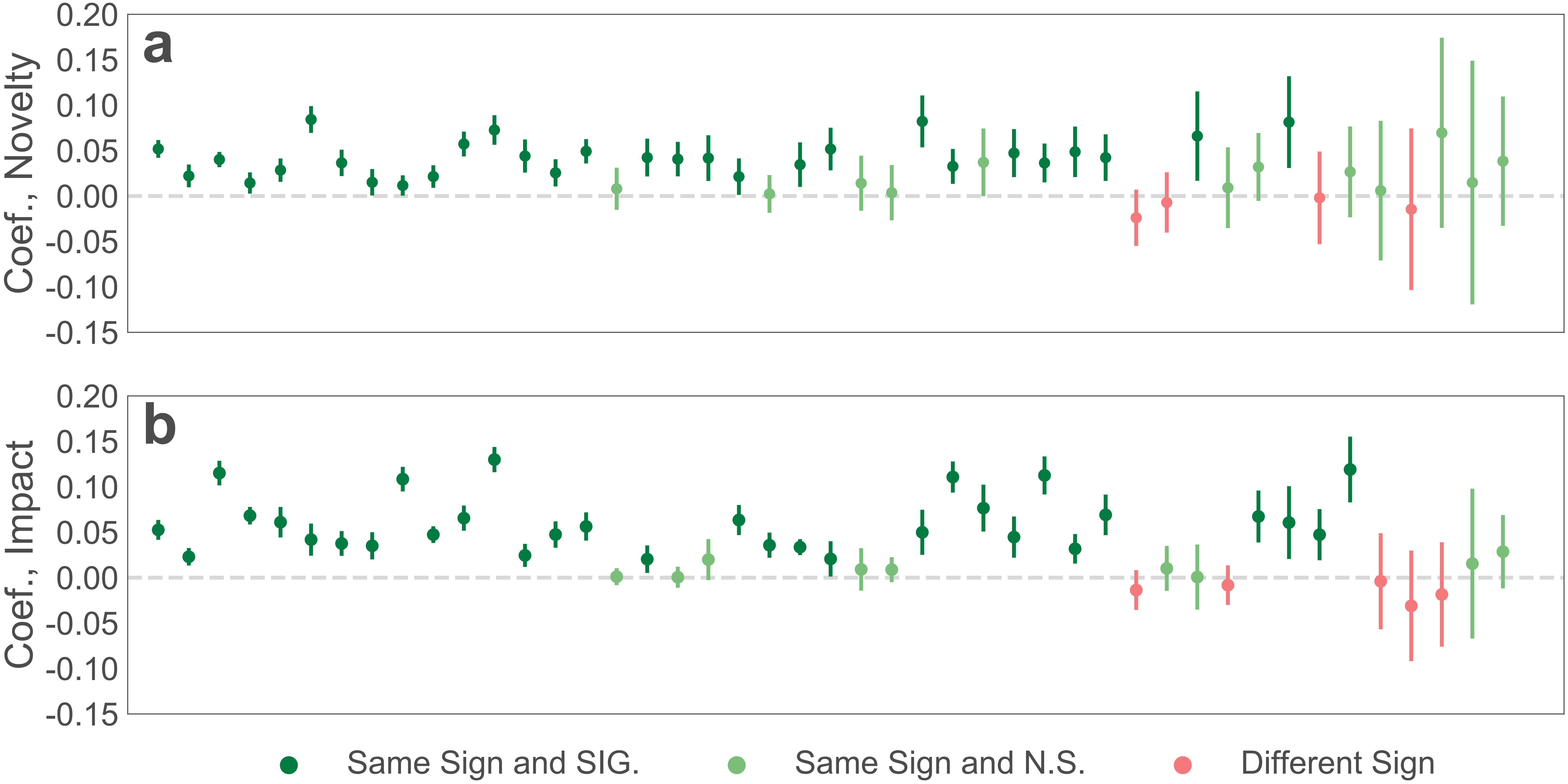
Figure 1: Gender Diverse Teams Remain Underrepresented in Medical Science. Figure 1a plots the share of publications (y-axis) by team size and year from 2000-2019 (x-axis) and figure 1a inset demonstrates the sharp upward trend of women's participation in medical science over the same time period. Large teams have replaced small teams. In 2000, solo and two-person teams accounted for 15% and 16% of publications, respectively, but by 2015 their shares dropped to 8% and 12% while the largest teams increased their market share dominance from 25% to 46%. Figure 1b and 1d shows that the share of publications from gender diverse teams steadily increased with time. Nonetheless, Figure 1c and 1e indicate that mixed-gender and gender balanced teams are significantly underrepresented in medical science by up to 17% depending on the team size.

Figure 2: Gender Diverse Teams Produce the Most Innovative and Highly Cited Research Publications. Mixed-gender teams are more likely to produce innovative papers than same-gender teams at all team sizes; for teams of size 4 or more, mixed-gender teams are also more likely to produce a novel paper than the base rate (dashed line). Mixed-gender teams of size 6+ (about 1/2 of all 2019 publications) are 9.1% more likely to publish novel work than the base rate. Figure 2b shows that the same performance patterns are largely mirrored for whether a paper is a citation hit (top 5% of papers). Large (6+) mixed-gendered teams can be twice as likely to publish hit papers than the base rate and 16.7% more likely than large (6+) same-gender teams [(10.5%-9.0%)/9.0%) = 16.7%]. The results are reproduced when team gender diversity is measured in terms of Shannon Entropy and shows that as gender balance of a team increases, innovativeness and impact significantly increase (see **SM Tables S1** and **S2**).

Figure 3: The Benefits of Gender Diverse Teams Generalize Across Medical Science Subfields. The plots demonstrate that the performance benefits of team gender diversity generalize across medical subfields. Figure a and b show the regression coefficients of 45 separate regressions. Each regression estimates the relationship between either novelty (figure a) or citation impact (figure b) and the team's gender diversity for the papers in one of 45 separate medical subfields. The regressions control for confounds due to the authors' experience, prestige, subfield, individual success, journal of publication and the gender of the team's leadership as was done in Figure 2, which pooled the data for subfields in medical science. Subfields are arranged from left to right according to the number of papers in the subfield (largest to smallest). The larger subfields show positive and statistically significant relationships with novelty and impact while the relationships become noisier with smaller fields with team gender diversity being significantly predictive of team performance for the majority of subfields, a statistical regularity that is not explained by chance (binomial test $p < 0.03$ for mixed-gender and $p < 0.001$ for gender balance).


a
Percent
Year
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
2000
2005
2010
2015
2020
Percent of Female Scientists
0.35
0.40
0.45
0.50
2000 2005 2010 2015 2020
b
Mixed Gender Percent
0.0
0.2
0.4
0.6
0.8
1.0
2000
2010
2019
c
Difference (real - simulated)
-0.05
-0.10
2000
2010
2019
d
Avg. Team Gender Diversity
0.0
0.2
0.4
0.6
0.8
1.0
2000
2010
2019
e
Difference (real - simulated)
-0.05
-0.10
-0.15
-0.20
2000
2010
2019
team size = 1
team size = 2
team size = 3
team size = 4
team size = 5
team size ≥ 6


a
Prob. of Novel Paper
0.50
0.45
0.40
0.35
2
3
4
5
≥ 6
Team Size
b
Prob. of Hit Paper (5%)
12%
10%
8%
6%
4%
2%
0%
2
3
4
5
≥ 6
Team Size
Mixed Gender Team
Same Gender Team


a
b
Coef., Novelty
Coef., Impact
0.20
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
● Same Sign and SIG.
● Same Sign and N.S.
● Different Sign

1 Materials and Methods

1.1 Data

1.1.1 Microsoft Academic Graph

Our analysis is based on Microsoft Academic Graph (MAG), a scientific publication database. MAG records journal article's bibliographic information (title, journal, journal field, volume, issue, page, publication date), authorship (name), author affiliations (name, official page, and wikipage), and citation links to other papers in the database. We focused on medical journal articles published from 2000 to 2019.

Microsoft Academic Graph (MAG) is publicly available (https://docs.microsoft.com/en-us/academic-services/).

1.1.2 U.S. News University Ranking Data

Our institutional ranking data comes from U.S. News best global universities rankings (https://www.usnews.com/education/best-global-universities/rankings), which was accessed on June 27, 2020. There are about 1,200 U.S. News recognized universities across 80 different countries. The rankings were calculated using 13 weighted indicators that U.S. News chose to measure a university's global research performance. These 13 indicators include global research reputation (12.5%), regional research reputation (12.5%), publications (10%), total citations (12.5%), number of publications that are among hit papers (12.5%) etc. The U.S. News ranking data were used in our analysis related to institutional rank. For institutions that are not listed in U.S. News University ranking, we classify them into a category called "unranked" in regression analyses. These data are publicly available on the U.S. News website.

1.1.3 Scimago Journal Rank

MAG covers a wide range of journals of varying quality ratings. To control for journal quality, we match the subset of MAG medical journals with the Scimago Journal Ranking. Scimago Journal Ranking (https://www.scimagojr.com) ranks 10,368 medical journals from 1999 to 2020. Using journal name or journal ISSN, we matched journals listed in the Scimago Journal Ranking and journals recorded in MAG. 7,808 out of 10,368 (75.3%) Scimago journals can be matched to MAG; 87% of journals with H-index ≥ 5 can be matched; and the number rises to 93.3% when we consider journals with H-index ≥ 10. The match ratio between Scimago and MAG is relatively high for high impact journals.

There are about 12 million medical journal articles published between 2000 and 2019 (MAG). The total number of authors associated with those 12 million medical journal articles is about 16 million. However, 5.4 million medical journal articles among them do not have information about references. Those 5.4 million medical articles without references can be divided into three categories: (1) non research articles, such as comments, where references are not required; (2) articles in foreign languages where references cannot be mapped back to their English formats; (3) articles published in low-impact journals. Our analysis shows that over 90% of the 5.4 million medical articles without references have zero citations. This implies that almost all medical papers without references are most likely non-research papers or from low-impact journals, which can be excluded from our sample. Therefore, our main observations are based on 6.6 million medical journal articles with references information.

1.2 Methods

1.2.1 Gender Detection of Scientists Based on Names

The total number of authors associated with the 6.6 million medical journal articles is 9.6 million. Among these authors, 79% have their full first names recorded in MAG (as opposed to the initials alone), which allows us to algorithmically estimate a binary gender designation based on both the author's first name and last name. For the remaining 21% of authors, we conducted robustness tests by simulating the gender designation based on their first name initials (see Section **S2.4** for details). We use the popular Namsor software (*1*), which handles multiple languages (e.g., Chinese, English, French, Spanish, etc.). Another advantage of this algorithm is its ability to classify binary gender for Asian names (*1*).

1.2.2 Novelty of Scientific Papers

Novelty is an essential feature of creative ideas. Several existing novelty metrics at the paper level have been constructed by using references information (*2*, *3*). Following prior research (*2*), we measure novelty at the paper level by examining the combination of prior work referenced in a paper's bibliography using a z-score based metric. To compute novelty, we compare the observed frequency of journal pairs that appear within paper reference lists with a null model of the journal pair distribution created by randomized citation networks. Reference pairs that appear more than expected by chance are conventional and reference pairs that appear less than expected by chance are novel, with the z-score indicating the degree of novelty contained in the paper. Formally, the novelty measure in the prior work (*2*) is a z-score, where lower values indicate higher novelty. Details of the method can be found in pages 3 – 5 in the supplementary information of work (*2*).

For simplicity, we define a binary variable $innovative$ as below.

$$innovative = \begin{cases} 0, & novelty > 0 \\ 1, & novelty \leq 0 \end{cases} \tag{1}$$

The variable $innovative$ is used to indicate whether a paper is novel or not in our main results (see **Figure 2** in the manuscript).

To test the robustness of our analyses, we further investigate whether our main results hold when novelty is measured by a continuous variable. The original measure introduced in the work of (*2*) follows a heavy-tail distribution. Therefore, we use a log transformation to convert the z-score to the form below. The new measure also improves readability, such that a higher score indicates greater novelty.

$$novelty = \begin{cases} -log_2(\text{z-score} + 1), & \text{z-score} > 0 \\ log_2(-\text{z-score} + 1), & \text{z-score} \leq 0 \end{cases} \tag{2}$$

1.2.3 Impact of Scientific Papers

The MAG database keeps track of paper reference information, where the tuple $\{j, i,\ t\}$ lists a paper j that cites paper i at time t. We can calculate the total number of citations to paper i. We denote the final number of citations for a paper i as c_i. To provide a fair and comparable measure, we further normalize a paper i's final citations by the corresponding year average, which is denoted as $\widehat{c_\iota}$.

Similarly, as for the $innovative$ variable above, we use a binary variable $hit\ paper$ to measure a paper's impact.

$$hit\ paper = \begin{cases} 0, & if\ \widehat{c_i} < 95_{th}\ percentile \\ 1, & if\ \widehat{c_i} \geq 95_{th}\ percentile \end{cases} \quad (3)$$

The variable $hit\ paper$ indicates whether it is a top 5% home run paper or not. This is one of the dependent variables used in our main results (see **Figure 2** in the manuscript).

To further investigate whether our main results hold, we also run a similar regression by substituting the binary $hit\ paper$ variable with a continuous one. Given $\widehat{c_i}$ follows a heavy-tail distribution, we use a log transform of $\widehat{c_i}$ to measure a scientific paper's impact.

$$impact = \log\left(\widehat{c_i} + 1\right) \quad (4)$$

1.2.4 Team Gender Diversity

To measure the gender composition in a scientific team, we use a binary variable m (mixed gender).

$$m_i = \begin{cases} 1, & the\ team\ has\ both\ men\ and\ women \\ 0, & either\ all\ men\ or\ all\ women \end{cases} \quad (5)$$

This is a key independent variable used in our main results (see **Figure 2** in the manuscript).

We also use a continuous variable to evaluate the gender composition of a scientific team. Similarly as in the work of (*4*), we use Shannon Entropy to measure the gender diversity of a team, which takes the form

$$g_i = -p_f log_2\left(p_f\right) - \left(1 - p_f\right) log_2\left(1 - p_f\right) \quad (6)$$

where p_f indicates the portion of female scientists in a team i. The value of g_i ranges from 0 to 1. When the value of g_i is low, either women or men are majority of a team. When $g_i = 0$, the team is either an all-women team or an all-men team. By contrast, when the value of g_i is high, women and men have roughly equivalent presence in the team. When $g_i = 1$, the team has 50% women and 50% men (see **Figure S1** below).

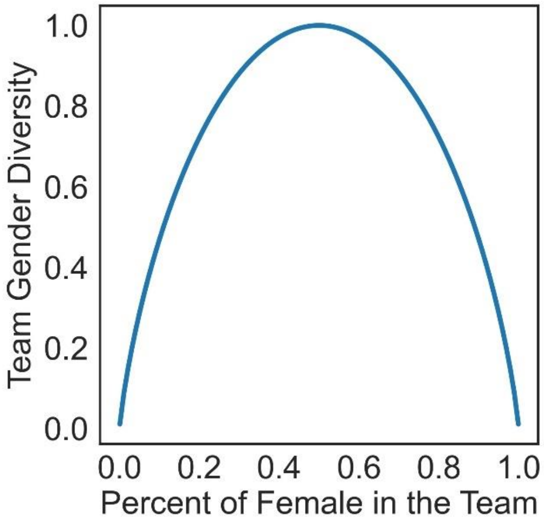


Figure S1 The Relationship between Percent of Female and Team Gender Diversity.

2 Regression Analysis

The results in **Figure 2** of manuscript are based on fixed-effect ordinary least squares regressions. In this section, we discuss the details of our regression analyses.

In addition, we also conduct robustness tests to address several potential concerns: (1) whether noise in gender designations by Namsor software could significantly affect our conclusions; (2) whether missing data for authors with first initials are critical enough to significantly affect our existing conclusions; and (3) whether our main conclusions hold when articles from low-impact journals are excluded.

2.1 Regression of Figure 2A

Our results in **Figure 2A** are based on a fixed-effect ordinary least squares regression as below.

$$\begin{aligned} y_i = {} & \beta_m m_i + \sum_t \beta_t T_{ti} + \sum_t \beta_{mt} m_i T_{ti} + \beta_f f_i + \beta_l l_i + \sum_r \beta_r R_{ri} + \sum_d \beta_d D_{di} + \sum_e \beta_e E_{ei} \\ & + \sum_a \beta_a A_{ai} + \sum_h \beta_h H_{hi} + \sum_p \beta_p P_{pi} + \sum_q \beta_q Q_{qi} + \sum_s \beta_s S_{si} + \sum_j \beta_j J_{ji} \\ & + \epsilon_i \end{aligned} \tag{7}$$

Dependent Variable: The dependent variable y_i measures whether a paper is novel or not, which is measured by the variable $innovative$ defined in equation (1). An alternative measure is a continuous variable $novelty$ defined in equation (2). Our main results are based on the analysis using the binary variable $innovative$. The continuous variable $novelty$ is also considered as a robustness check.

Predictors of Interest: We use a binary variable m_i to indicate whether a scientific team is mixed gender or not (see definition in the equation (5)). Furthermore, a Shannon Entropy measure g_i is used to evaluate the detailed information of team gender composition (see definition in the equation (6)). Similarly, our main results are based on the regression analysis using the binary variable m. The continuous variable g_i is used for robustness check.

Control Variables: We also include several other explanatory variables to control for other possible predictors of paper impact.

- T_{ti}: T_{ti} indicates fixed effects that account for the size of a scientific team. We categorize a scientific team into 6 bins: $t = 1$, $t = 2$, $t = 3$, $t = 4$, $t = 5$, and $t \geq 6$. $T_{ti} = 1$ if the team size of a paper i is in bin t and $T_{ti} = 0$ otherwise.
- f_i: f_i measures the gender of first author. $f_i = 1$ if the first author is male and $f_i = 0$ otherwise.
- l_i: l_i measures the gender of last author. $l_i = 1$ if the first author is male and $l_i = 0$ otherwise.
- R_{ri}: R_{ri} indicates fixed effects that account for the highest institution rank affiliated with a paper i. We categorize institution rank into 9 bins: [1, 10], [11, 20], [21, 40], [41, 80], [81, 160], [161, 320], [321, 640], [641, 1250] and no rank (no rank means that the institution is not recognized in U.S. News Ranking Database). $R_{ri} = 1$ if the highest institution rank affiliated with a paper i is in bin r and $R_{ri} = 0$ otherwise.

- D_{di}: D_{di} indicates fixed effects that account for the career age of first author when a paper i was published. We categorize the career age of first author into bins: [0, 5], [6, 10], [11, 15], [16, 20], [21, 25], [26, 30], [31, 35], [36, 40], [41, 45], [46, 50], [51, Inf]. $D_{di} = 1$ if the career age of first author is in a bin d and $D_{di} = 0$ otherwise.
- E_{Ei}: E_{ei} indicates fixed effects that account for the career age of last author when a paper i was published, using the same bin definitions as for first authors above. $E_{ei} = 1$ if the career age of last author is in a bin e and $E_{ei} = 0$ otherwise.
- A_{ai}: A_{ai} indicates fixed effects that account for the average career age of a team, using the same bin definitions as for first authors above. $A_{ai} = 1$ if the average career age is in a bin a and $A_{ai} = 0$ otherwise.
- H_{hi}: H_{hi} indicates fixed effects that account for the impact of first author when a paper i was published. We categorize the impact of first author into 21 exponential bins. $H_{hi} = 1$ if the impact of first author is in an exponential bin h and $H_{hi} = 0$ otherwise.
- P_{pi}: P_{pi} indicates fixed effects that account for the impact of last author when a paper i was published. It has a similar setting as H_{hi}.
- Q_{qi}: Q_{qi} indicates fixed effects that account for the average impact of authors when a paper i was published. It has a similar setting as H_{hi}.
- S_{si}: S_{si} indicates fixed effects that account for an individual scientist. $S_{si} = 1$ if a paper i is written by scientist s and $S_{si} = 0$ otherwise.
- J_{ji}: J_{ji} indicates fixed effects that account for the journal-year. For example, 'Science' and '2020' is one journal-year pair indicating all papers published by Science in the year of 2020. Therefore, $J_{ji} = 1$ if a paper i belongs to the journal-year pair j and $J_{ji} = 0$ otherwise.

The results of regression analysis are presented in **Table S1**. First, the results in Models (1) and (3) confirm that there is a strong connection between a paper's novelty and its team gender composition. After controlling for a number of other explanatory variables including institution rank, author's prior impact and author's career age, we find that mixed gender teams are more likely to produce novel scientific papers (model (2)). In addition, when both dependent variable and key independent variable are measured as continuous variables, our observations in model (2) remain valid. We have several interesting observations. First, we can see large teams are more likely to produce novel papers. Second, scientific teams with women as first author or last author are more likely to produce novel scientific papers. This is consistent with observations made in the work of (5) that minority groups are more like to produce novel scientific papers.

2.2 Regression of Figure 2B

Similarly, when we examine a paper's impact, we specify a regression model that examines the relationship between team gender diversity and paper impact as follows.

$$z_i = \beta_m m_i + \sum_t \beta_t T_{ti} + \sum_t \beta_{mt} m_i T_{ti} + \beta_f f_i + \beta_l l_i + \sum_r \beta_r R_{ri} + \sum_d \beta_d D_{di} + \sum_e \beta_e E_{ei} + \sum_a \beta_a A_{ai} + \sum_h \beta_h H_{hi} + \sum_p \beta_p P_{pi} + \sum_q \beta_q Q_{qi} + \sum_s \beta_s S_{si} + \sum_j \beta_j J_{ji} + \in_i \quad (8)$$

Table S1 Relationship between team gender composition and paper novelty.

	Model (1)		Model (2)		Model (3)	Model (4)
Variable	DV = $innovative$		DV = $innovative$		DV = $novelty$	DV = $novelty$
Mixed Gender Team (m_i)	0.083*** (0.00082)		0.031*** (0.0011)			
Team Size FE (T_{ti})	t = 2	0.023*** (0.0010)	t = 2	0.017*** (0.0012)		
	t = 3	0.037*** (0.0010)	t = 3	0.033*** (0.0013)		
	t = 4	0.047*** (0.0010)	t = 4	0.049*** (0.0014)		
	t = 5	0.050*** (0.0012)	t = 5	0.055*** (0.0016)		
	t > 5	0.038*** (0.0011)	t > 5	0.065*** (0.0014)		
$m_i \times T_{ti}$	m_i = 1 t = 2	-0.053*** (0.0014)	m_i = 1 t = 2	-0.026*** (0.0017)		
	m_i = 1 t = 3	-0.043*** (0.0013)	m_i = 1 t = 3	-0.022*** (0.0016)		
	m_i = 1 t = 4	-0.035*** (0.0013)	m_i = 1 t = 4	-0.020*** (0.0016)		
	m_i = 1 t = 5	-0.025*** (0.0015)	m_i = 1 t = 5	-0.015*** (0.0017)		
Team Gender Diversity (g_i)					0.89*** (0.0046)	0.49*** (0.015)
Team Size (t)					0.026*** (0.00076)	0.011*** (0.0029)
$g_i \times t$					-0.016*** (0.00089)	-0.0094** (0.0030)
First Author Gender			-0.0062*** (0.00065)			-0.023*** (0.0044)
Last Author Gender			-0.0089*** (0.00072)			-0.056*** (0.0047)
Institution FE			Y			Y
First Author Age FE			Y			Y
Last Author Age FE			Y			Y
Mean Author Age FE			Y			Y
First Author Impact FE			Y			Y
Last Author Impact FE			Y			Y
Mean Author Impact FE			Y			Y
Observations	6,606,294		5,276,515		6,606,294	5,276,515
R Squared	0.0075		0.28		0.012	0.33

*, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$.

Dependent Variable: The dependent variable z_i measures whether a paper is a hit paper being the top 5% home run papers gauged by citation, which is measured by the variable $hit\ paper$ defined in the equation (3). An alternative measure is a continuous variable impact defined in equation (4). Our main results are based on the analysis using the binary variable $hit\ paper$. The continuous variable $impact$ is used in the robustness check.

Predictors of Interest: We use a binary variable m_i to indicate whether a scientific team is a mixed gender team. Furthermore, a Shannon Entropy measure g_i is used to evaluate team gender composition. Similarly, our main results are based on the regression analysis using the binary variable m. The continuous variable g_i is used for the robustness checking purpose.

Control Variables are the same as those defined in Section **S2.1.**

Table S2 Relationship between team gender composition and paper impact.

	Model (1)		Model (2)		Model (3)	Model (4)
Variable	DV = $hit\ paper$		DV = $hit\ paper$		DV = $impact$	DV = $impact$
Mixed Gender Team (m_i)	0.037*** (0.00036)		0.013*** (0.00062)			
Team Size FE (T_{ti})	t = 2	0.012*** (0.00040)	t = 2	0.039*** (0.00062)		
	t = 3	0.011*** (0.00042)	t = 3	0.066*** (0.00069)		
	t = 4	0.0091*** (0.00045)	t = 4	0.085*** (0.00075)		
	t = 5	0.011*** (0.00051)	t = 5	0.090*** (0.00082)		
	t > 5	0.023*** (0.00042)	t > 5	0.11*** (0.00083)		
$m_i \times T_{ti}$	m_i = 1 t = 2	-0.030*** (0.00063)	m_i = 1 t = 2	-0.0034*** (0.00091)		
	m_i = 1 t = 3	-0.028*** (0.00058)	m_i = 1 t = 3	-0.0057*** (0.00083)		
	m_i = 1 t = 4	-0.025*** (0.00060)	m_i = 1 t = 4	-0.0077*** (0.00081)		
	m_i = 1 t = 5	-0.024*** (0.00063)	m_i = 1 t = 5	-0.0096*** (0.00085)		
Team Gender Diversity (g_i)					0.28*** (0.00081)	0.22*** (0.010)
Team Size (t)					0.038*** (0.00013)	0.020*** (0.0020)
$g_i \times t$					-0.024*** (0.00016)	-0.019*** (0.0021)
First Author Gender			0.0036*** (0.00043)			0.011*** (0.0011)
Last Author Gender			0.0035*** (0.00049)			0.008*** (0.0011)
Institutional FE			Y			Y
First Author Age FE			Y			Y
Last Author Age FE			Y			Y
Mean Author Age FE			Y			Y
First Author Impact FE			Y			Y
Last Author Impact FE			Y			Y
Mean Author Impact FE			Y			Y
Observations	6,606,294		5,276,515		6,606,294	5,276,515
R Squared	0.01		0.34		0.06	0.51

*, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$.

The results of regression analysis are presented in **Table S2**. The results in model (1) and (3) demonstrate that there is a strong connection between team gender composition and a paper's impact. After controlling for several other explanatory variables, such as institution rank, author's prior impact and author's career age, we find that mixed gender teams are more likely to produce highly cited scientific papers (model (2)). In addition, our observations in model (2) remain valid when we measure both the dependent variable and the key independent variable as continuous variables. We have several observations that are very similar to **Table S1**. First, we can see large teams are more likely to produce papers of high impact. In **Table S1**, we find that scientific teams with women as first author or last author are more likely to produce novel scientific papers. However, it is opposite when predicting a paper's citation. Results in both model (2) and (4) suggest that scientific teams with men as first or last author are more likely to be cited.

2.3 Robustness Test for Disambiguated Gender Classification

Our work classifies scientists' gender based on their first and last names. To the extent that some cases may be misclassified when scientists have unisex first names, we further test whether potential issues with misclassification have significant impact on our main results.

The output of Namsor allows us to conduct such a robustness test. In addition to a gender label, the machine learning model also provides a confidence score calibrated from big data. For example, the name "Anderson Cooper" is labeled as male with 98.3% confidence score. The confidence score allows us to examine whether the uncertainty levels are large enough to nullify the findings. Specifically, there are about 9.6 million scientists in the field of Medicine from 2000 to 2019. Based on the gender classification by Namsor, about 42.3% of them are women. In addition, 86% of scientists in Medicine have gender confidence scores larger than 80%. And 81% of them have gender confidence scores larger than 90%.

To verify whether these noises are large enough to nullify our findings, we run regressions on multiple subsets where scientists are included only when their gender confidence scores are larger than a threshold. In this work, we use four different thresholds: 60%, 70%, 80% and 90%.

The robustness test results can be found in **Table S3** and **Table S4**. We can see that our main conclusions are valid across different confidence thresholds. In **Table S3**, we can see the variation in the coefficients of m_i in predicting a paper's novelty is no more than 5% when we increase the threshold from 60% to 90%. In **Table S4**, the coefficients of m_i in predicting a paper's impact are consistent with our observations in **Table S2** when we increase the threshold from 60% to 90%. Overall, our main finding that mixed gender teams are more innovative remains consistent.

2.4 Robustness Test for Authors with First Initials

In our main results, we do not include authors who only have first initials in our analysis. Therefore, we are interested in whether the resulting missing data might undermine our main findings. To test this, we simulate gender labels for authors with first initials using detected gender designations. To illustrate these procedures, consider the following example.

1. We calculate authors' career age by using their first publication years as below.
$$career\ age = 2019 - first\ publication\ year + 1 \tag{9}$$
And we categorize the career age into bins: [0, 5], [6, 10], [11, 15], [16, 20], [21, 25], [26, 30], [31, 35], [36, 40], [41, 45], [46, 50], [51, Inf].
2. We classify authors with detected gender designations into different groups based on (i) whether they have the same first initial and (ii) whether their career ages are in the same career age bin described above. For example, *Alexander Jones* who has worked in the field of

Medicine for 8 years and *Aaron Smith* who has worked in the field of Medicine for 10 years are classified into the same group “A. [6, 10]”.

3. We calculate the portion of female scientists in each group where full names are available. For example, the group “A. [0, 5]” has about 48% female scientists. And 23% of the group “S. [41,45]” are female scientists.
4. For a single author with S. as first initial and who is in a given career age group, we randomly classify her/him as female or male based on the computed gender ratio in the that career age group.

With the procedures described above, we can assign gender to authors who only have first initials recorded. To verify whether those missing data nullifies our observations, we run a regression on the data where authors with initials are randomly assigned gender labels in this way.

The result can be found in **Table S3** and **S4**. When we include authors with initials into our analysis, the coefficient of m_i in predicting novelty decreases about 10%. And the coefficient of m_i in predicting paper impact does not change. In conclusion, the missing data does not nullify our main results even as we add this data which has higher noise.

Table S3 Robustness tests for relationship between team gender composition and paper novelty.

Variable	Model (1)	Model (2)	Model (3)	Model (4)	Model (5)
	Threshold = 60%	Threshold = 70%	Threshold = 80%	Threshold = 90%	Include Authors with First Initials
Mixed Gender Team (m_i)	0.031*** (0.0011)	0.031*** (0.0010)	0.031*** (0.0011)	0.030*** (0.0011)	0.028*** (0.0011)
Team Size FE (T_{ti})	Y	Y	Y	Y	Y
$m_i \times T_{ti}$	Y	Y	Y	Y	Y
First Author Gender	-0.0065*** (0.00068)	-0.0065*** (0.00069)	-0.0073*** (0.00073)	-0.0075*** (0.00076)	-0.0062*** (0.00060)
Last Author Gender	-0.0094*** (0.00075)	-0.0095*** (0.00077)	-0.0098*** (0.00081)	-0.011*** (0.00084)	-0.0075*** (0.00066)
Institutional FE	Y	Y	Y	Y	Y
First Author Age FE	Y	Y	Y	Y	Y
Last Author Age FE	Y	Y	Y	Y	Y
Mean Author Age FE	Y	Y	Y	Y	Y
First Author Impact FE	Y	Y	Y	Y	Y
Last Author Impact FE	Y	Y	Y	Y	Y
Mean Author Impact FE	Y	Y	Y	Y	Y
Observations	4,991,066	4,749,982	4,461,380	4,161,852	6,110,387
R Squared	0.28	0.28	0.29	0.29	0.28

*, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$.

2.5 Robustness Test for Scimago Journal Ranking Subsample

Our data sample includes 6.6 million medical journal articles, which are published in 15,033 journals with heterogeneous levels of impact and quality. To address potential issues of low-quality journals, we rely on the Scimago Journal Ranking (https://www.scimagojr.com). There are 10,368 medical journals from 1999 to 2020 recorded by Scimago. The ranking also calculates H-index for each journal which allows us to proxy for quality of the journal. For example, H-index = 5 means that 5 papers published in the journal have at least 5 citations.

Table S4 Robustness tests for relationship between team gender composition and paper impact.

Variable	Model (1)	Model (2)	Model (3)	Model (4)	Model (5)
	Threshold = 60%	Threshold = 70%	Threshold = 80%	Threshold = 90%	Include Authors with First Initials
Mixed Gender Team (m_i)	0.014*** (0.00063)	0.014*** (0.00065)	0.014*** (0.00067)	0.015*** (0.00070)	0.013*** (0.00061)
Team Size FE (T_{ti})	Y	Y	Y	Y	Y
$m_i \times T_{ti}$	Y	Y	Y	Y	Y
First Author Gender	0.0037*** (0.00046)	0.0039*** (0.00048)	0.0042*** (0.00050)	0.0044*** (0.00052)	0.0035*** (0.00040)
Last Author Gender	0.0037*** (0.00052)	0.0036*** (0.00054)	0.0039*** (0.00057)	0.0039*** (0.00058)	0.0031*** (0.00045)
Institutional FE	Y	Y	Y	Y	Y
First Author Age FE	Y	Y	Y	Y	Y
Last Author Age FE	Y	Y	Y	Y	Y
Mean Author Age FE	Y	Y	Y	Y	Y
First Author Impact FE	Y	Y	Y	Y	Y
Last Author Impact FE	Y	Y	Y	Y	Y
Mean Author Impact FE	Y	Y	Y	Y	Y
Observations	4,991,066	4,749,982	4,461,380	4,161,852	6,110,387
R Squared	0.34	0.34	0.34	0.35	0.34

*, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$.

Using journal name or journal ISSN, we match journals listed in Scimago Journal Ranking and journals in MAG. We find that 7,808 out of the 10,368 (75.3%) Scimago journals can be found in MAG. Furthermore, 87% of Scimago journals with H-index ≥ 5 can be found in MAG. And the matching ratio reaches 93.3% when we only consider Scimajo journals with H-index ≥ 10. This implies that the match ratio between Scimago and MAG is higher for high-quality journals.

We extract a subsample from the 6.6 million medical journal articles according to the Scimago journal ranking, which includes 5.6 million medical papers. In **Figure S2**, we replicate our results of **Figure 2** in the manuscript by using 5.6 million medical papers published in journals that are listed in the Scimago Journal Ranking. We can see that our findings remain valid when considering this subsample.

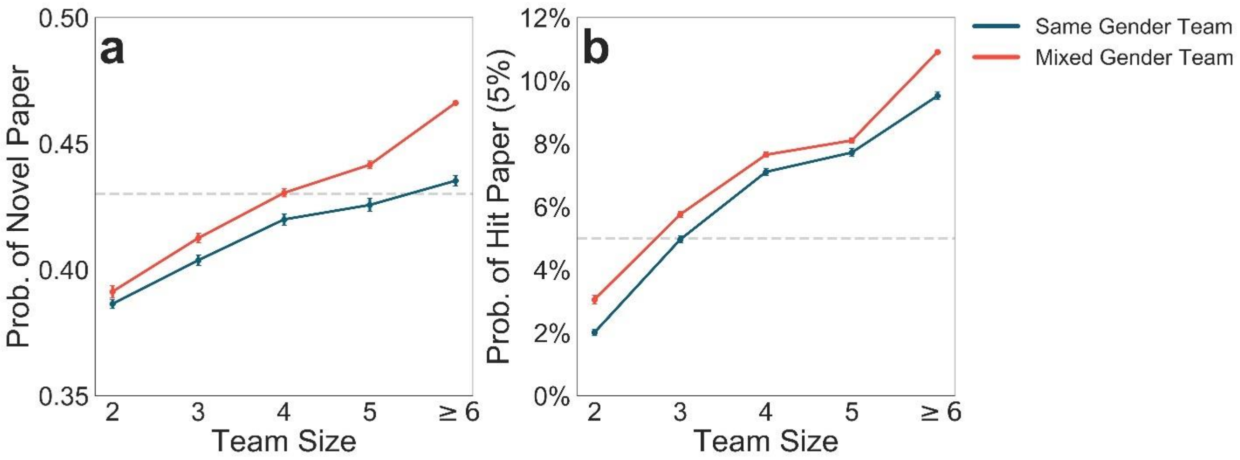


Figure S2 Mixed Gender Teams are More Innovative (A) and highly cited (B) among ranked journals in the Scimago Journal Ranking. Mixed gender teams are more likely to produce innovative papers than are same-gender teams of all team sizes; for teams of size 4 or more mixed-gender are always more likely to produce a novel paper than the base rate (dashed line). Mixed gender teams of size 5 or 6 are more likely to publish novel work than the base rate. Estimates shown are margins plots computed from fixed effect regressions.

Similarly, we also consider again the analyses in **Table S1** and **Table S2** based on this subsample of 5.6 million medical papers. Results are presented in **Table S5** and **Table S6.** respectively. The results remain consistent.

Table S5 Relationship between team gender composition and paper novelty (Scimago Journal Ranking subsample).

	Model (1)		Model (2)		Model (3)	Model (4)
Variable	DV = $innovative$		DV = $innovative$		DV = $novelty$	DV = $novelty$
Mixed Gender Team (m_i)	0.077*** (0.00088)		0.031*** (0.0011)			
Team Size FE (T_{ti})	t = 2	0.019*** (0.00098)	t = 2	0.016*** (0.0012)		
	t = 3	0.033*** (0.0011)	t = 3	0.034*** (0.0014)		
	t = 4	0.043*** (0.0011)	t = 4	0.050*** (0.0015)		
	t = 5	0.048*** (0.0012)	t = 5	0.055*** (0.0017)		
	t > 5	0.037*** (0.0011)	t > 5	0.065*** (0.0015)		
$m_i \times T_{ti}$	m_i = 1 t = 2	-0.047*** (0.0015)	m_i = 1 t = 2	-0.026*** (0.0018)		
	m_i = 1 t = 3	-0.037*** (0.0014)	m_i = 1 t = 3	-0.022*** (0.0017)		
	m_i = 1 t = 4	-0.030*** (0.0015)	m_i = 1 t = 4	0.020*** (0.0017)		
	m_i = 1 t = 5	-0.021*** (0.0016)	m_i = 1 t = 5	0.015*** (0.0018)		
Team Gender Diversity (g_i)					0.88*** (0.0050)	0.52*** (0.016)
Team Size (t)					0.024*** (0.00081)	0.011*** (0.0029)
$g_i \times t$					-0.015*** (0.00095)	-0.0090** (0.0031)
First Author Gender			-0.0068*** (0.00070)			-0.027*** (0.0048)
Last Author Gender			-0.010*** (0.00078)			-0.063*** (0.0052)
Institutional FE			Y			Y
First Author Age FE			Y			Y
Last Author Age FE			Y			Y
Mean Author Age FE			Y			Y
First Author Impact FE			Y			Y
Last Author Impact FE			Y			Y
Mean Author Impact FE			Y			Y
Observations	5,619,077		4,530,400		5,619,077	4,530,400
R Squared	0.007		0.27		0.011	0.31

*, p < 0.05; **, p < 0.01; ***, p < 0.001.

Table S6 Relationship between team gender composition and paper impact (Scimago Journal Ranking subsample).

	Model (1)		Model (2)		Model (3)	Model (4)
Variable	DV = $hit\ paper$		DV = $hit\ paper$		DV = $impact$	DV = $impact$
Mixed Gender Team (m_i)	0.038*** (0.00039)		0.014*** (0.00067)			
Team Size FE (T_{ti})	t = 2	0.011*** (0.00044)	t = 2	0.042*** (0.00069)		
	t = 3	0.010*** (0.00047)	t = 3	0.071*** (0.00077)		
	t = 4	0.0087*** (0.00051)	t = 4	0.092*** (0.00082)		
	t = 5	0.010*** (0.00058)	t = 5	0.099*** (0.00090)		
	t > 5	0.023*** (0.00047)	t > 5	0.12*** (0.00091)		
$m_i \times T_{ti}$	m_i = 1 t = 2	-0.031*** (0.00070)	m_i = 1 t = 2	-0.0035*** (0.0010)		
	m_i = 1 t = 3	-0.029*** (0.00065)	m_i = 1 t = 3	-0.0060*** (0.00092)		
	m_i = 1 t = 4	-0.026*** (0.00066)	m_i = 1 t = 4	-0.0085*** (0.00090)		
	m_i = 1 t = 5	-0.025*** (0.00071)	m_i = 1 t = 5	-0.010*** (0.00093)		
Team Gender Diversity (g_i)					0.29*** (0.00088)	0.23*** (0.011)
Team Size (t)					0.039*** (0.00014)	0.021*** (0.0022)
$g_i \times t$					-0.025*** (0.00017)	-0.020*** (0.0022)
First Author Gender			0.0036*** (0.00047)			0.013*** (0.0012)
Last Author Gender			0.0038*** (0.00054)			0.0098*** (0.0012)
Institutional FE			Y			Y
First Author Age FE			Y			Y
Last Author Age FE			Y			Y
Mean Author Age FE			Y			Y
First Author Impact FE			Y			Y
Last Author Impact FE			Y			Y
Mean Author Impact FE			Y			Y
Observations	5,619,077		4,530,400		5,619,077	4,530,400
R Squared	0.01		0.34		0.06	0.49

*, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$.

2.6 Generalizability across Subfields in Medicine

Finally, we also investigate the generalizability of our findings across 45 subfields in Medicine, such as anatomy and biomedical engineering. Those 45 subfields are listed in **Table S7**.

Table S7 Subfields in Medicine.

Anatomy	Andrology	Anesthesia	Audiology	Biomedical engineering
Cancer research	Cardiology	Clinical psychology	Dentistry	Dermatology
Emergency medicine	Endocrinology	Environmental health	Family medicine	Gastroenterology
General surgery	Gerontology	Gynecology	Immunology	Intensive care medicine
Internal medicine	Medical education	Medical emergency	Medical physics	Nuclear medicine
Nursing	Obstetrics	Oncology	Ophthalmology	Optometry
Orthodontics	Pathology	Pediatrics	Pharmacology	Physical medicine and rehabilitation
Physical therapy	Physiology	Psychiatry	Radiology	Surgery
Traditional medicine	Urology	Veterinary medicine	Virology	Miscellaneous

We examined the coefficients of team gender diversity g_i in two different regression settings. In the first case, the dependent variables are binary. Namely, we are using $innovative$ and $hit\ paper$ as dependent variables in the regression. The definition of these two variables can be found in Section **S1.2**. In the second setting, we are using $novelty$ and $impact$ measures as our dependent variables. The detailed information can also be found in Section **S1.2**.

In **Figure 3**, we visualize the coefficients with 95% confidence intervals for 45 subfields when binary measures $innovative$ and $hit\ paper$ are used as dependent variables. We have several important observations. First, we observe that the sign of team gender diversity is consistent and positive in most of subfields. For example, when predicting innovative papers, team gender diversity has consistent and positive sign in 41 out of 45 subfields (**Figure 3** (a)). When predicting hit papers, the sign of team gender diversity is consistent and positive in 40 out 45 subfields.

Furthermore, we also verify the generalizability when dependent variables are continuous. In **Figure S3**, we observe that the sign of team gender diversity is consistent and positive in most of subfields. For example, when predicting papers' novelty, team gender diversity has consistent and positive sign in 40 out of 45 subfields (**Figure S3** (a)). When predicting papers' impact, the sign of team gender diversity is consistent and positive in 42 out 45 subfields.

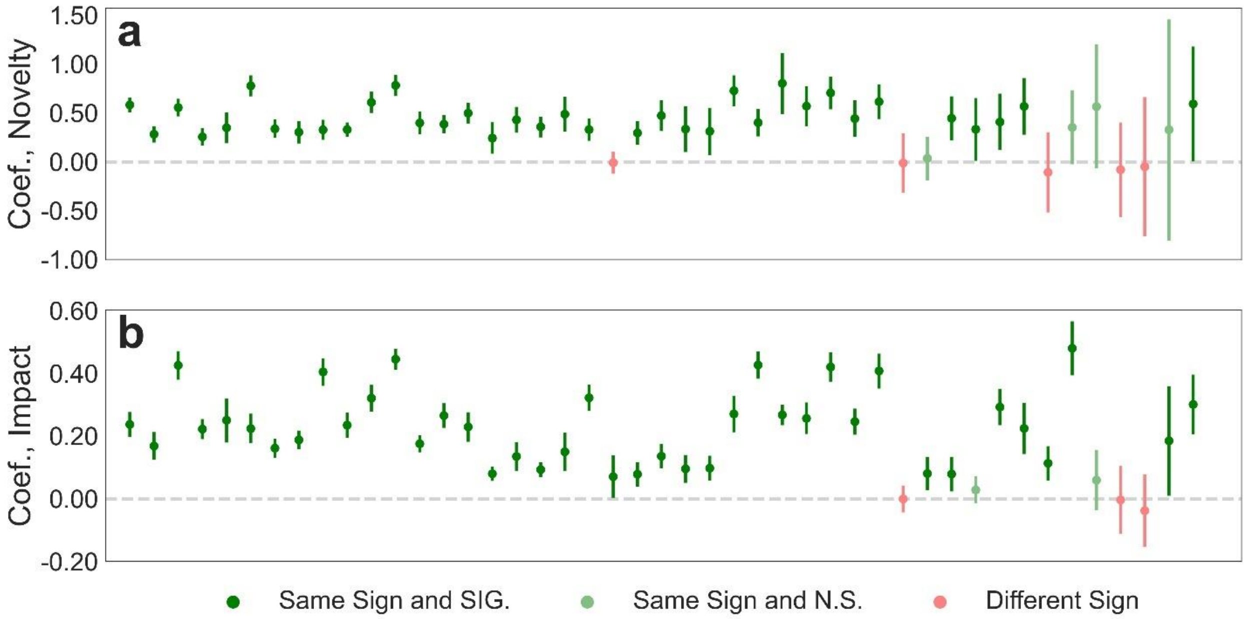


Figure S3 Coefficients of Team Gender Diversity across 45 Subfields in Medicine (Dependent variables are also continuous). Each bar indicates the coefficient of team gender diversity with 95% confidence interval in a subfield. We sort subfields from smallest to largest. The observations are consistent with **Figure 3** in manuscript. (a) Coefficients of team gender diversity in predicting papers' novelty. Dark green color indicates positive and significant (p-value < 0.05) coefficients. Light green color indicates positive but non-significant coefficients. Red color indicates negative coefficients. We can see 40 out of 45 subfields have positive coefficients. In addition, 36 of them are significant. This demonstrates the generalizability of our main finding that team gender diversity is predictive of papers' novelty. (b) Similar to subfigure (a), 42 out of 45 subfields have positive coefficients when predicting papers' impact. And the coefficients are both positive and significant in 40 subfields.

3 Gender Diverse Teams over Time

Figure 1 in the manuscript shows the increasing dominance of teamwork in medical science over the last 20 years. In addition, the share of papers written by mixed-gender teams at all team sizes has increased annually with growth concentrated in larger teams. In this section, we consider the rise in mixed gender teams compared to a null model, together with robustness tests.

3.1 Null Model

To understand the increase of mixed gender teams in light of the increased presence of female scientists, we design a null model. To illustrate the null model, consider the following steps and example.

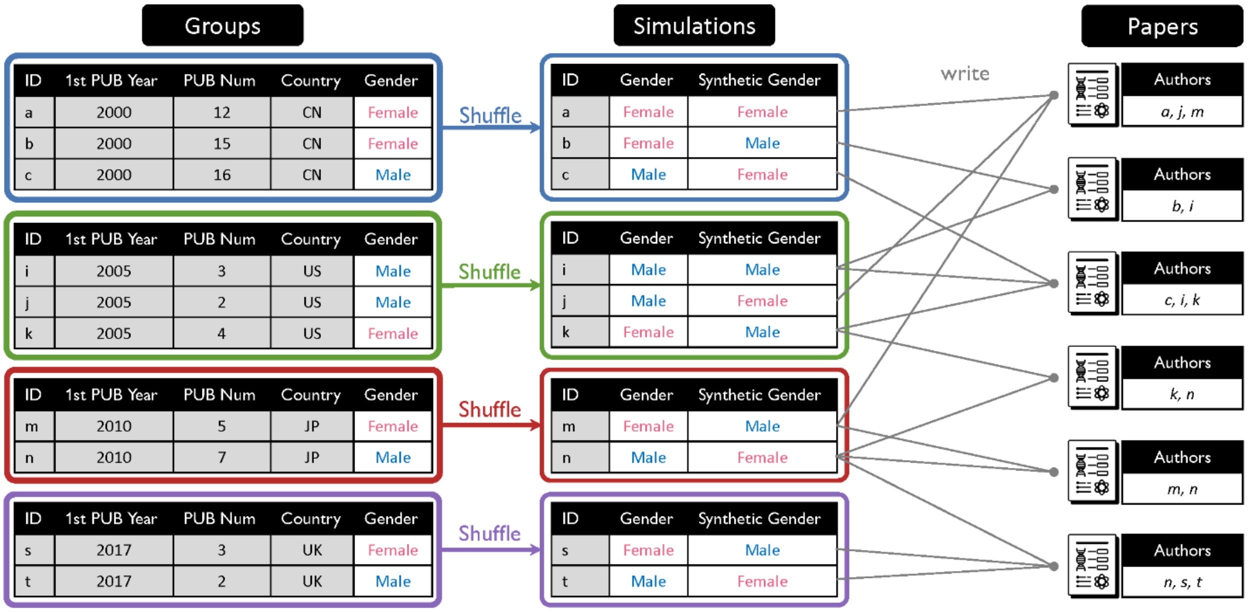
1. For each author in our sample, we extract her first publication year (publication year of her first paper), total number of publications, and the country where her affiliated institution is located (i.e., CN, US, UK, JP and etc.).
2. Second, we categorize the total number of publications into bins: [0, 1], [2, 4], [5, 8], [9, 16], [17, 32], [33, 64], [65, 128], [129, 256], [257, 512], [513, 1024], [1025, Inf].
3. With that information, we classify scientists into different groups if (i) they have the same first publication year; (ii) their total numbers of publications are in the same bin; and (iii) their affiliated institutions are in the same country. In **Figure S4**, we provide several examples. For example, scientists a, b and c are classified into the same group because their first publication year = 2005, their total numbers of publications are in the bin of [9,16], and they are all in China. Similarly, scientists i, j and k are categorized into the same group because they wrote their first papers in the year of 2005, their total numbers of publications are in the bin of [2, 4], and they are in United States.
4. In each round of our simulations, we randomly shuffle scientists' gender designations within each group. In this way, the gender ratio in each group is preserved. Take the blue group in **Figure S4** as an example, scientists b and c exchange their gender labels. But the gender ratio of blue group is still two women versus one man.
5. With this gender shuffling within groups, we then turn to the actual papers written and consider the resulting gender distributions that emerge among the papers.

In this way, the null model can provide randomness while keeping several factors intact, for example the portion of women among newcomers with similar productivity in the same country-year.

This null model allows us to verify whether the increase of mixed gender teams or women's underrepresentation can be explained by women's increased attendance in science. Take the portion of mixed gender teams as an example. First, we can calculate the real portion of mixed gender teams in each year. Then, we run the null model 100 rounds and get 100 sets of synthetic data of gender designations. For each set of synthetic data, we recalculate the portion of mixed teams in each year. Finally, we have 100 simulated values. We can then evaluate the z-score for each observed portion of mixed gender teams relative to what is expected by the null model:

$$z = \frac{(obs - exp)}{\sigma} \tag{10}$$

where *obs* is the observed portion of mixed gender team while *exp* is the mean and σ is the standard deviation of the simulated portions using the null model.



ID	1st PUB Year	PUB Num	Country	Gender
a	2000	12	CN	Female
b	2000	15	CN	Female
c	2000	16	CN	Male

ID	1st PUB Year	PUB Num	Country	Gender
i	2005	3	US	Male
j	2005	2	US	Male
k	2005	4	US	Female

ID	1st PUB Year	PUB Num	Country	Gender
m	2010	5	JP	Female
n	2010	7	JP	Male

ID	1st PUB Year	PUB Num	Country	Gender
s	2017	3	UK	Female
t	2017	2	UK	Male

ID	Gender	Synthetic Gender
a	Female	Female
b	Female	Male
c	Male	Female

ID	Gender	Synthetic Gender
i	Male	Male
j	Male	Female
k	Female	Male

ID	Gender	Synthetic Gender
m	Female	Male
n	Male	Female

ID	Gender	Synthetic Gender
s	Female	Male
t	Male	Female

Figure S4 Illustrative Example of the Null Model.

Similarly, following the same procedure, we can also evaluate whether women are underrepresented in the positions of first author and last author.

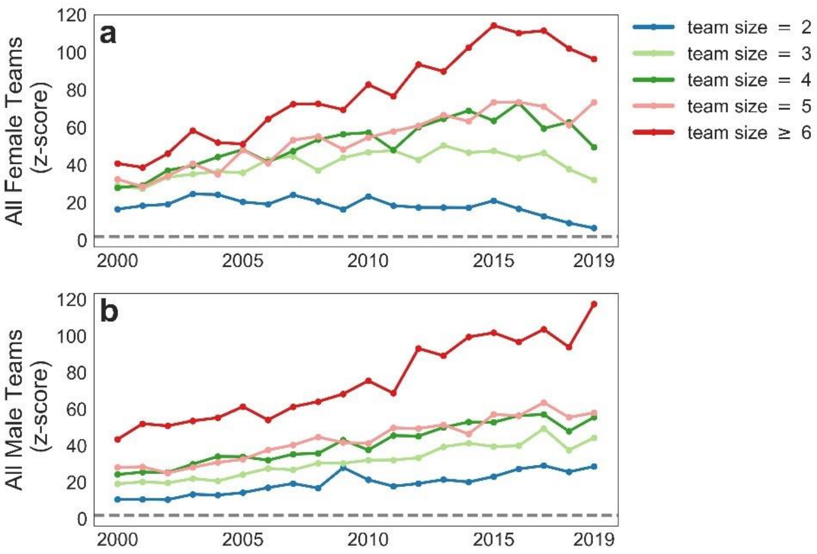


Figure S5 Same Gender Teams Remain Overrepresented in Medical Science. Using the same null model, we verify whether same gender teams are overrepresented over time. We can see both female teams and male teams are over-represented.

3.1.1 Same Gender Teams are Overrepresented

Using the same null model explained above, we can also verify whether all-women teams and all-men teams are overrepresented over time.

Based on the results presented in **Figure 1**, we conclude that the increase in mixed gender teams is not explained simply by the increase of women in science and that same gender teams remain overrepresented in medical science. This is supported by results in the **Figure S5**. In **Figure S5**, we can see both all-women teams and all-men teams are significantly over-represented despite the growing trend of mixed gender teams.

3.2 Robustness Test for Figure 1B and 1C Using a Shannon Entropy Measure

In this section, we demonstrate that our observations in **Figure 1B and 1C** of the manuscript are valid when we use a Shannon Entropy measure g_i to evaluate gradations of gender diversity. The measure is defined in equation (6) of section **S1.2.4**.

In the manuscript, we use a binary variable m_i (see equation (5)) to measure team gender composition. Here, we switch to a continuous variable g_i, which ranges from 0 to 1. When $g_i = 0$, the team is a same gender team. In contrast, when $g_i = 1$, 50% of members are women and the 50% are men. In contrast to **Figure 1B**, we are now measuring the average g_i for each team size category.

In **Figure S6**, we find the results using team gender diversity g_i are consistent with what we observe in **Figure 1B** and **1C** in the manuscript.

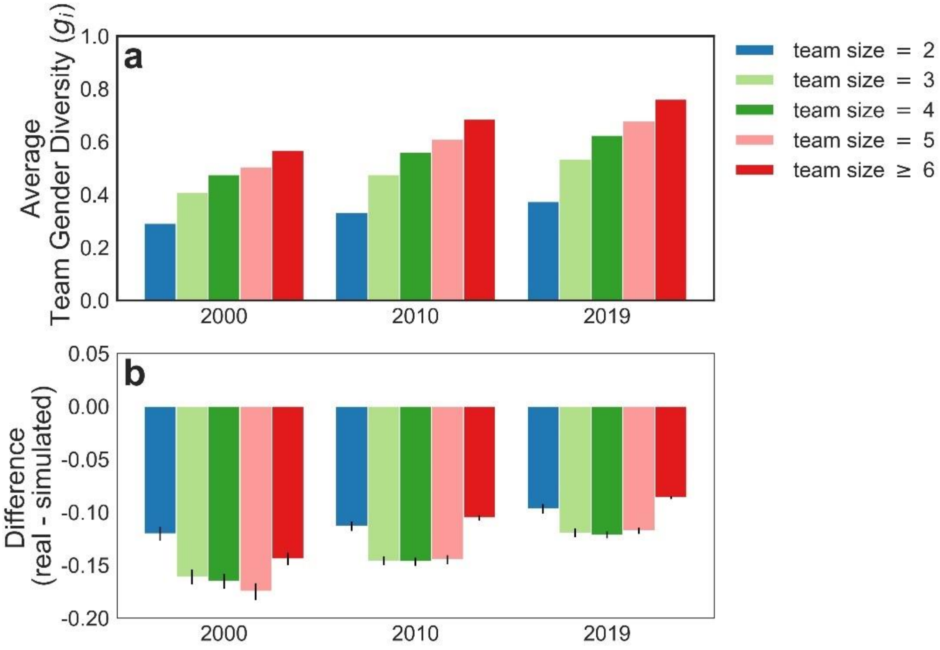


Figure S6 Gender Balanced Teams Dominate Science. Instead of using a binary variable m_i, we use a Shannon Entropy measure g_i to evaluate gradations of gender diversity. (a) shows the average team gender diversity increased steadily over time. (b) shows the results of a null model that indicate that the observed team gender diversity remains underrepresented in medical science.

3.3 Robustness Test for Medical Papers without References

As discussed in **Section 1.1.3**, our main analysis is based on 6.6 million medical journal articles with references information. We now conduct a robustness test to check whether the results presented in **Figure 1** and **Figure 2** remain consistent when we include the medical journal articles without references. The analysis sample uses 12 million papers that include 6.6 million medical journal articles with references and 5.4 million articles without references.

We cannot carry out the same robustness test for **Figure 2** because we need reference information to measure paper novelty (see section **S1.2.2**).

3.3.1 Robustness Test for Figure 1A by Including Papers without References

The results in **Figure 1** of the manuscript are based on 6.6 million medical journal articles. Here, we test whether the results hold when we include journal articles without references.

In **Figure S7**, the gap between small and large teams becomes smaller when using the 12 million papers. This implies that the trend of large team is even stronger in comparably high-impact journal articles (because ~90% of 5.4 million papers without references have zero citations).

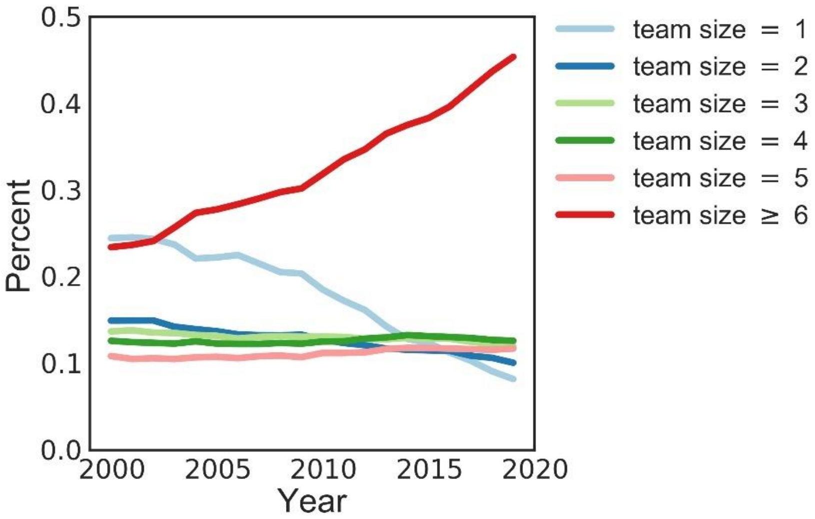


Figure S7 Big Teams Dominate Medical Science. This figure plots the share of publications (y-axis) by team size and year from 2000-2019 (x-axis). Over time, large teams have replaced small teams. For example, in 2000 solo and two-person teams each had more than 15% of the share of publications but by 2015 their shares dropped to 12% and 11% respectively. In contrast, large teams with more than 5 persons increased their market share dominance from 23% to 45%.

3.3.2 Robustness Test for Figure 1B and 1C by Including Papers without References

In **Figure 1B and 1C** of the manuscript, we have demonstrated that mixed-gender teams steadily increased over time. Similarly, we also replicate the results of **Figure 1B and 1C** in the manuscript using the sample of 12 million medical journal articles, which is presented in **Figure S8**. The results are consistent with our observations in the manuscript.

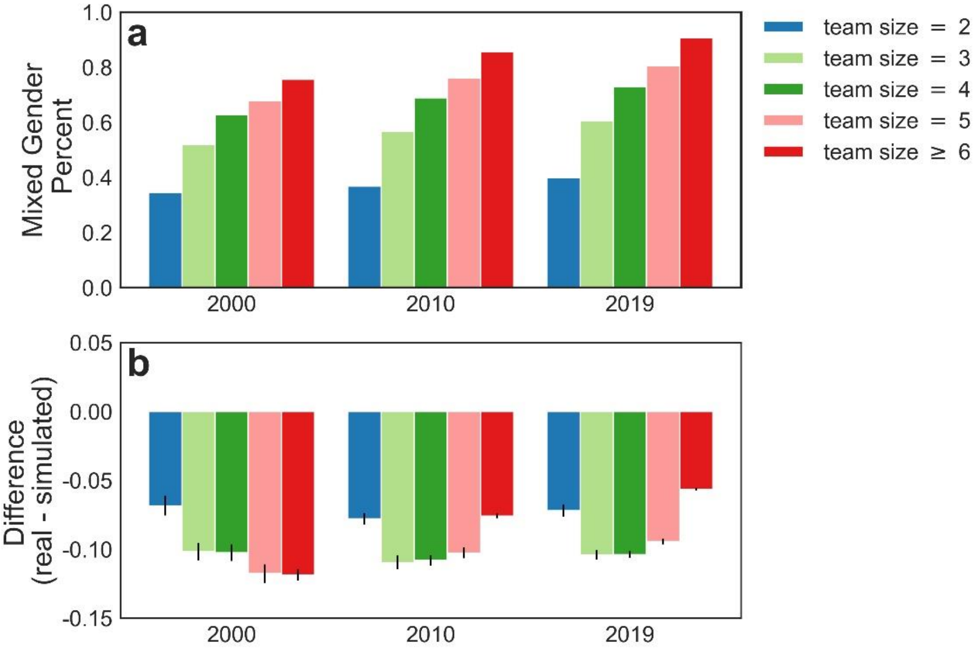


Figure S8 Mixed Gender Teams Dominate Science. Using all 12 million papers, we repeat the analysis in **Figure 1B and 1C** and find similar results. (a) shows the share of publications from mixed vs same gender teams steadily increased with time and that the increase is proportionally greater the larger the team size. (b) shows the results of a null model that indicate that mixed-gender teams remain underrepresented in medical science by up to 10% depending on the team size.

4 Generalizability across Teams Led by Women and Men

Here, we analyze the interaction among team gender composition, team size, and leadership. We further consider five binary variables as below:

$$mff_i = \begin{cases} 1, & m_i = 1 \; and \; f_i = 0 \\ 0, & otherwise \end{cases} \tag{11}$$

$$mfm_i = \begin{cases} 1, & m_i = 1 \; and \; f_i = 1 \\ 0, & otherwise \end{cases} \tag{12}$$

$$sgt_i = \begin{cases} 1, & m_i = 0 \\ 0, & m_i = 1 \end{cases} \tag{13}$$

$$mlf_i = \begin{cases} 1, & m_i = 1 \; and \; l_i = 0 \\ 0, & otherwise \end{cases} \tag{14}$$

$$mlm_i = \begin{cases} 1, & m_i = 1 \; and \; l_i = 1 \\ 0, & otherwise \end{cases} \tag{15}$$

where the definitions of m_i, f_i, and l_i can be found in section **S2.1** and these new binary variables are described as follows:

1. The binary variable mff_i indicates whether a paper i is written by a mixed gender team led by female first author.
2. The binary variable mfm_i indicates whether a paper i is written by a mixed gender team led by male first author.
3. The binary variable sgt_i indicates whether a paper i is written by a same gender team.
4. The binary variable mlf_i indicates whether a paper i is written by a mixed gender team led by female last author.
5. The binary variable mlm_i indicates whether a paper i is written by a mixed gender team led by male last author.

4.1 Generalizability of Predicting Novelty across First Author's Gender

First, we examine the generalizability in predicting novelty across first author's gender using a fixed-effect ordinary least squares regression as below.

$$\begin{aligned} y_i = {} & \beta_{mff} mff_i + \beta_{mfm} mfm_i + \beta_{sgt} sgt_i + \sum_t \beta_t T_{ti} + \sum_t \beta_{mfft} mff_i T_{ti} \\ & + \sum_t \beta_{mfmt} mfm_i T_{ti} + \sum_t \beta_{sgtt} sgt_i T_{ti} + \beta_f f_i + \beta_l l_i + \sum_r \beta_r R_{ri} \\ & + \sum_d \beta_d D_{di} + \sum_e \beta_e E_{ei} + \sum_a \beta_a A_{ai} + \sum_h \beta_h H_{hi} + \sum_p \beta_p P_{pi} + \sum_q \beta_q Q_{qi} \\ & + \sum_s \beta_s S_{si} + \sum_j \beta_j J_{ji} + \in_i \end{aligned} \tag{16}$$

Dependent Variable: The dependent variable y_i measures whether a paper is novel or not, which is measured by the variable $innovative$ defined in equation (1).

Predictors of Interest: The key independent variables are three binary variables: mff_i, mfm_i, and sgt_i.

Control Variables: We also include several other explanatory variables to control for other possible predictors of paper impact. Details can be found in section **S2**.

4.2 Generalizability of Predicting Novelty across Last Author's Gender

Second, we examine the generalizability in predicting novelty across last author's gender using a fixed-effect ordinary least squares regression as below.

$$\begin{aligned} y_i = {} & \beta_{mlf} mlf_i + \beta_{mlm} mlm_i + \beta_{sgt} sgt_i + \sum_t \beta_t T_{ti} + \sum_t \beta_{mlft} mlf_i T_{ti} + \sum_t \beta_{mlmt} mlm_i T_{ti} \\ & + \sum_t \beta_{sgtt} sgt_i T_{ti} + \beta_f f_i + \beta_l l_i + \sum_r \beta_r R_{ri} + \sum_d \beta_d D_{di} + \sum_e \beta_e E_{ei} \\ & + \sum_a \beta_a A_{ai} + \sum_h \beta_h H_{hi} + \sum_p \beta_p P_{pi} + \sum_q \beta_q Q_{qi} + \sum_s \beta_s S_{si} + \sum_j \beta_j J_{ji} \\ & + \in_i \end{aligned} \tag{17}$$

Dependent Variable: The dependent variable y_i measures whether a paper is novel or not, which is measured by the variable $innovative$ defined in the equation (1).

Predictors of Interest: The key independent variables are three binary variables: mlf_i, mlm_i, and sgt_i.

Control Variables: We also include several other explanatory variables to control for other possible predictors of paper impact. Details can be found in section **S2**.

4.3 Generalizability of Predicting Impact across First Author's Gender

Thirdly, we examine the generalizability in predicting impact across first author's gender using a fixed-effect ordinary least squares regression as below.

$$\begin{aligned} z_i = {} & \beta_{mff} mff_i + \beta_{mfm} mfm_i + \beta_{sgt} sgt_i + \sum_t \beta_t T_{ti} + \sum_t \beta_{mfft} mff_i T_{ti} \\ & + \sum_t \beta_{mfmt} mfm_i T_{ti} + \sum_t \beta_{sgtt} sgt_i T_{ti} + \beta_f f_i + \beta_l l_i + \sum_r \beta_r R_{ri} \\ & + \sum_d \beta_d D_{di} + \sum_e \beta_e E_{ei} + \sum_a \beta_a A_{ai} + \sum_h \beta_h H_{hi} + \sum_p \beta_p P_{pi} + \sum_q \beta_q Q_{qi} \\ & + \sum_s \beta_s S_{si} + \sum_j \beta_j J_{ji} + \in_i \end{aligned} \tag{18}$$

Dependent Variable: The dependent variable z_i measures whether a paper is a top 5% home run paper.

Predictors of Interest: The key independent variables are three binary variables: mff_i, mfm_i, and sgt_i.

Control Variables: We also include several other explanatory variables to control for other possible predictors of paper impact. Details can be found in section **S2**.

4.4 Generalizability of Predicting Impact across Last Author's Gender

Thirdly, we examine the generalizability in predicting impact across last author's gender using a fixed-effect ordinary least squares regression as below.

$$
\begin{aligned}
z_i = {} & \beta_{mlf} mlf_i + \beta_{mlm} mlm_i + \beta_{sgt} sgt_i + \sum_t \beta_t T_{ti} + \sum_t \beta_{mlft} mlf_i T_{ti} + \sum_t \beta_{mlmt} mlm_i T_{ti} \\
& + \sum_t \beta_{sgtt} sgt_i T_{ti} + \beta_f f_i + \beta_l l_i + \sum_r \beta_r R_{ri} + \sum_d \beta_d D_{di} + \sum_e \beta_e E_{ei} \\
& + \sum_a \beta_a A_{ai} + \sum_h \beta_h H_{hi} + \sum_p \beta_p P_{pi} + \sum_q \beta_q Q_{qi} + \sum_s \beta_s S_{si} + \sum_j \beta_j J_{ji} \\
& + \in_i
\end{aligned}
\tag{19}
$$

Dependent Variable: The dependent variable z_i measures whether a paper is a top 5% home run paper.

Predictors of Interest: The key independent variables are three binary variables: mlf_i, mlm_i, and sgt_i.

Control Variables: We also include several other explanatory variables to control for other possible predictors of paper impact. Details can be found in section **S2**.

4.5 Regression Results

In **Table S8 and S9**, we present the results of the four regressions mentioned above.

Table S8 Interaction among team gender composition, team size, and first author gender in predicting paper novelty and impact.

	Model (1)		Model (2)	
Variable	DV = $innovative$		DV = $hit\ paper$	
Mixed Gender team led by female first author (mff_i)	0.022*** (0.0021)		-0.0069*** (0.0015)	
Mixed Gender team led by male first author (mfm_i)	0.014*** (0.0020)		0.0031* (0.0014)	
Same Gender Team (sgt_i)	-0.015*** (0.0022)		-0.013*** (0.0015)	
Team Size FE (T_{ti})	t = 2	0.017*** (0.0011)	t = 2	0.039*** (0.00061)
	t = 3	0.018*** (0.0056)	t = 3	0.057*** (0.0030)
	t = 4	0.033*** (0.0045)	t = 4	0.077*** (0.0024)
	t = 5	0.051*** (0.0042)	t = 5	0.075*** (0.0022)
	t > 5	0.063*** (0.0013)	t > 5	0.11*** (0.00081)
$mff_i \times T_{ti}$	mff_i = 1 t = 2	-0.028*** (0.0019)	mff_i = 1 t = 2	0.0082*** (0.0011)
	mff_i = 1 t = 3	-0.0095 (0.0056)	mff_i = 1 t = 3	0.014*** (0.0030)
	mff_i = 1 t = 4	-0.0083 (0.0044)	mff_i = 1 t = 4	0.0095*** (0.0023)
	mff_i = 1 t = 5	-0.015*** (0.0041)	mff_i = 1 t = 5	0.013*** (0.0022)
$mfm_i \times T_{ti}$	mfm_i = 1 t = 2	-0.025*** (0.0021)	mfm_i = 1 t = 2	-0.017*** (0.0012)
	mfm_i = 1 t = 3	-0.0067 (0.0056)	mfm_i = 1 t = 3	-0.0070* (0.0029)
	mfm_i = 1 t = 4	-0.0036 (0.0044)	mfm_i = 1 t = 4	-0.0065** (0.0024)
	mfm_i = 1 t = 5	-0.0096* (0.0041)	mfm_i = 1 t = 5	-0.00024 (0.0022)
$sgt_i \times T_{ti}$	sgt_i = 1 t = 2	0.014* (0.0056)	sgt_i = 1 t = 2	0.0093** (0.0030)
	sgt_i = 1 t = 3	0.014** (0.0045)	sgt_i = 1 t = 3	0.0081** (0.0023)
	sgt_i = 1 t = 4	0.0026 (0.0042)	sgt_i = 1 t = 4	0.015*** (0.0022)
First Author Gender	-		-	
Last Author Gender	-0.0091*** (0.00072)		0.0029*** (0.00050)	
Institution FE	Y		Y	
First Author Age FE	Y		Y	
Last Author Age FE	Y		Y	
Mean Author Age FE	Y		Y	
First Author Impact FE	Y		Y	
Last Author Impact FE	Y		Y	
Mean Author Impact FE	Y		Y	
Observations	5,545,641		5,545,641	
R Squared	0.28		0.34	

*, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$.

Table S9 Interaction among team gender composition, team size, and last author gender in predicting paper novelty and impact.

	Model (1)		Model (2)	
Variable	DV = $innovative$		DV = $hit\ paper$	
Mixed Gender team led by female last author (mlf_i)	0.020*** (0.0021)		-0.0066*** (0.0015)	
Mixed Gender team led by male last author (mlm_i)	0.0095*** (0.0020)		0.00055 (0.0015)	
Same Gender Team (sgt_i)	-0.020*** (0.0021)		-0.014*** (0.0015)	
Team Size FE (T_{ti})	t = 2	0.017*** (0.0011)	t = 2	0.039*** (0.00061)
	t = 3	0.014** (0.0050)	t = 3	0.060*** (0.0027)
	t = 4	0.036*** (0.0041)	t = 4	0.071*** (0.0022)
	t = 5	0.043*** (0.0038)	t = 5	0.075*** (0.0021)
	t > 5	0.063*** (0.0013)	t > 5	0.11*** (0.00080)
$mlf_i \times T_{ti}$	mlf_i = 1 t = 2	-0.027*** (0.0022)	mlf_i = 1 t = 2	-0.011*** (0.0012)
	mlf_i = 1 t = 3	-0.0070 (0.0051)	mlf_i = 1 t = 3	-0.0039 (0.0028)
	mlf_i = 1 t = 4	-0.012** (0.0041)	mlf_i = 1 t = 4	0.011*** (0.0022)
	mlf_i = 1 t = 5	-0.011** (0.0039)	mlf_i = 1 t = 5	0.0096*** (0.0021)
$mlm_i \times T_{ti}$	mlm_i = 1 t = 2	-0.028*** (0.0019)	mlm_i = 1 t = 2	0.0034** (0.0010)
	mlm_i = 1 t = 3	-0.0035 (0.0050)	mlm_i = 1 t = 3	-0.00096 (0.0027)
	mlm_i = 1 t = 4	-0.0070 (0.0040)	mlm_i = 1 t = 4	0.0052* (0.0021)
	mlm_i = 1 t = 5	-0.0036 (0.0037)	mlm_i = 1 t = 5	0.0048* (0.0020)
$sgt_i \times T_{ti}$	sgt_i = 1 t = 2	0.019*** (0.0050)	sgt_i = 1 t = 2	0.0063* (0.0027)
	sgt_i = 1 t = 3	0.012** (0.0041)	sgt_i = 1 t = 3	0.014*** (0.0022)
	sgt_i = 1 t = 4	0.0097* (0.0038)	sgt_i = 1 t = 4	0.015*** (0.0021)
First Author Gender	-0.0068*** (0.00065)		0.0043*** (0.00043)	
Last Author Gender	-		-	
Institution FE	Y		Y	
First Author Age FE	Y		Y	
Last Author Age FE	Y		Y	
Mean Author Age FE	Y		Y	
First Author Impact FE	Y		Y	
Last Author Impact FE	Y		Y	
Mean Author Impact FE	Y		Y	
Observations	5,585,550		5,585,550	
R Squared	0.28		0.34	

*, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$.

References

1. L. Santamaría, H. Mihaljević, Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science* **4**, e156 (2018).
2. B. Uzzi, S. Mukherjee, M. Stringer, B. Jones, Atypical combinations and scientific impact. *Science* **342**, 468-472 (2013).
3. E. Leahey, C. M. Beckman, T. L. Stanko, Prominent but less productive: The impact of interdisciplinarity on scientists' research. *Administrative Science Quarterly* **62**, 105-139 (2017).
4. Y. Yang, N. V. Chawla, B. Uzzi, A network's gender composition and communication pattern predict women's leadership success. *Proceedings of the National Academy of Sciences* **116**, 2033-2038 (2019).
5. B. Hofstra *et al.*, The Diversity–Innovation Paradox in Science. *Proceedings of the National Academy of Sciences* **117**, 9284-9291 (2020).