Supplementary Information for

Gender-diverse Teams Produce More Novel and Higher Impact Scientific Ideas

Yang Yang, Yuan Tian, Teresa Woodruff, Benjamin F. Jones, and Brian Uzzi

Brian Uzzi.
E-mail: uzzi@kellogg.northwestern.edu

This PDF file includes:

Supplementary text
Figs. S1 to S21
Tables S1 to S14
References for SI reference citations

Supporting Information Text

1. Materials and Methods

1.1. Data. Our main data sources are introduced in detail as below:

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 come from U.S. News best global universities rankings (https://www.usnews.com/education/best-global-universities/rankings), 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. Name-to-Gender Inference based on First and Last Names. The total number of authors associated with the 6.6 million medical papers 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 estimate a binary gender designation based on both the author's first name and last name. Furthermore, we also tested whether potential issues with misclassification have significant impact on our main results (see Section 3.3 for details). For the remaining 21% of authors with first initials alone, they are not included into our main analysis. Importantly, we conducted robustness tests by simulating the gender designation based on their first name initials to verify whether these missing data can undermine our main conclusions (see Section 3.4 for details).

Among the different methods used for gender designation, algorithmic name inference can be applied to large-scale database that contains personal names (1). Carsenat (1) describes the NamSor API v2 method (https://namsor.app/) for name-to-gender inference. The NamSor API v2 is based on a Naive Bayes machine learning model, which is calibrated on 1.3 million unique names collected from different countries and regions. To increase the accuracy, the NamSor model takes linguistic and cultural information into consideration. Therefore, it can infer the gender from first and last name together (last name is considered for cultural information). It is one of the few name-to-gender inference tools (2) that can make use of first and last name together for gender inference. Another important reason we use this algorithm is because it supports multiple languages and alphabets, including names written in Chinese, Korean and Japanese. To be specific, the NamSor API we used is version 2.0.11 (released on 2020-10-31). Our analysis demonstrates high agreement across versions (see Section 3). The robustness tests of our gender detection results can be found in **section S3**.

1.2.2. Novelty Measure Based on Journal Pairings. Novelty is an essential feature of creative ideas. Several existing novelty metrics at the paper level have been constructed by using references information (3, 4). Following prior research (3), 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 pairings that appear within paper reference lists with a null model of the journal pairing distribution created by randomized citation networks. Journal pairings that appear more than expected by chance are conventional and journal pairings 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 (3) 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 (3) and Figure S1 (a).

For simplicity, we define a binary variable *novel paper* as below:

$$\textit{novel paper} = \begin{cases} 0, & \textit{z-score} > 0 \\ 1, & \textit{z-score} \leq 0 \end{cases} \qquad [1]$$

The variable *novel paper* 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 (3) 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.

$$\textit{novelty} = \begin{cases} -log_2(\textit{z-score}+1), & \textit{z-score} > 0 \\ log_2(-\textit{z-score}+1), & \textit{z-score} \leq 0 \end{cases} \qquad [2]$$

1.2.3. Novelty Measure Based on Subject Pairings. For robustness check, we also use a different measure to evaluate novelty of scientific papers. Instead of journals, this measure focused on subject categories associated with references within a paper's reference list (4). In the work of (4), 244 WoS subject categories are used for the computation. The measure considers papers with rare combinations of subject pairings to be novel because they create new combinations of knowledge that have been infrequently co-cited together in previous research.

Similarly, we use 291 tier 1 fields of study recorded in MAG as subject categories in our context. In MAG, each paper is tagged with several subjects (tier 1 fields of study). After enumerating all papers' reference lists, we can calculate how many times each pair of subjects has been co-cited in prior research. With the subject co-citation matrix, we can further convert it into subject similarity matrix, denoted as S. S_{ij} indicates the cosine similarity between subject i and subject j using the information from the subject co-citation matrix. The novelty score of a paper based on subject pairings is defined as the Sterling index of all subject pairings within its reference list:

$$\textit{novelty}_{\textit{subjects}} = 1 - \sum_{ij} S_{ij} p_i p_j \qquad [3]$$

where p_i is the proportion of papers in the reference list associated with subject i, p_j is the proportion of papers in the reference list associated with subject j, and S_{ij} indicates the cosine similarity between subject i and subject j.

For simplicity, we also define a binary variable *novel paper*$_{\textit{subjects}}$ as below:

$$\textit{novel paper}_{\textit{subjects}} = \begin{cases} 0, & \textit{novelty}_{\textit{subjects}} < \textit{median} \\ 1, & \textit{novelty}_{\textit{subjects}} \geq \textit{median} \end{cases} \qquad [4]$$

Details of the method can be found in the work of (4) and Figure S1 (b).

1.2.4. 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. In this way, 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 $\widehat{c_i}$.

Similarly, as for the novelty variable above, we use a binary variable *upper-tail paper* to measure a paper's impact.

$$\textit{upper-tail paper} = \begin{cases} 0, & \text{if } \widehat{c_i} < 95_{th} \text{ percentile} \\ 1, & \text{if } \widehat{c_i} \geq 95_{th} \text{ percentile} \end{cases} \qquad [5]$$

The variable *upper-tail paper* indicates whether it is a top 5% highly cited paper or not. This is one of the dependent variables used in our main results (see **Figure 2** in the manuscript).

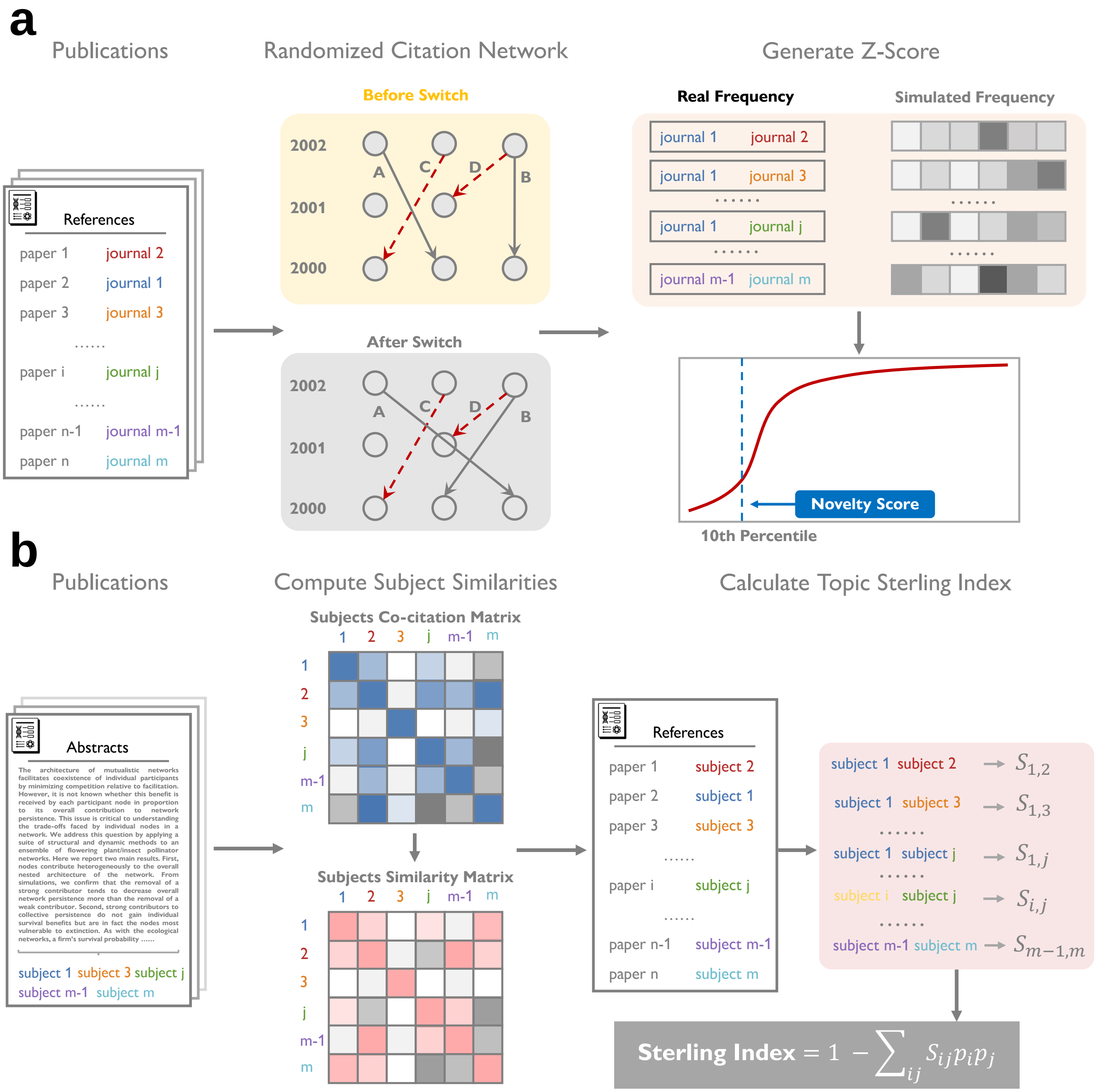
To further investigate whether our main results hold, we also run a similar regression by substituting the binary *upper-tail paper* variable with a continuous one. Given $\widehat{c_i}$ follows a heavy-tail distribution, we use a log transform of $\widehat{c_i}$ to measure a scientific paper's impact.

$$\textit{impact} = \log(\widehat{c_i} + 1) \qquad [6]$$

1.2.5. Team Gender Diversity. To measure the gender composition in a scientific team, we use a binary variable m (mixed-gender).

$$m_i = \begin{cases} 0, & \text{either all men or all women} \\ 1, & \text{the team has both men and women} \end{cases} \qquad [7]$$

This is a key independent variable used in our main results (see **Figure 2** in the manuscript). In summary, there are about 62% mixed-gender teams and 38% same-gender teams. Among same-gender teams, there are about 22% all women teams and 78% all men teams.



The architecture of mutualistic networks facilitates coexistence of individual participants by minimizing competition relative to facilitation. However, it is not known whether this benefit is received by each participant node in proportion to its overall contribution to network persistence. This issue is critical to understanding the trade-offs faced by individual nodes in a network. We address this question by applying a suite of structural and dynamic methods to an ensemble of flowering plant/insect pollinator networks. Here we report two main results. First, nodes contribute heterogeneously to the overall nested architecture of the network. From simulations, we confirm that the removal of a strong contributor tends to decrease overall network persistence more than the removal of a weak contributor. Second, strong contributors to collective persistence do not gain individual survival benefits but are in fact the nodes most vulnerable to extinction. As with the ecological networks, a firm's survival probability ......

$$\textbf{Sterling Index} = 1 - \sum_{ij} S_{ij} p_i p_j$$

Fig. S1. Novelty Measurements. (a) The novelty measure introduced in the work of (3) examines the combination of prior work referenced in a paper's bibliography using a z-score based metric. The measure considers papers with statistically atypical combinations of references to be novel because they create new combinations of knowledge that have not been joined, or rarely joined, in previous research. A z-score is used to compute a paper's novelty. The z-score is calculated from comparing the observed frequency of journal pairings that appear within a paper's reference list and the expected distribution of journal pairings created by randomized citation networks. Journal pairings that appear more than expected by chance are conventional and journal pairings that appear less than expected by chance are novel, with the z-score indicating the degree of novelty contained in the paper. For each paper's reference list, there is a distribution of z-scores. The tenth percentile value is defined as the novelty score of a paper. (b) The novelty measure introduced in the work of (4) examines the subjects combination of prior work referenced in a paper's bibliography using a Sterling index (5). Different from the work of (3), the measure considers papers with rare combinations of subjects to be novel. After enumerating all papers' references data, a subjects co-citation matrix can be calculated where each cell indicates how many times a pair of subjects is co-cited across all papers. The subjects co-citation matrix can be further converted to a subjects similarity matrix by computing each pair of subjects' cosine similarity score. Finally, for an individual paper, a novelty measure based on subjects can be calculated using the Sterling index (5) of all subjects within its reference list.

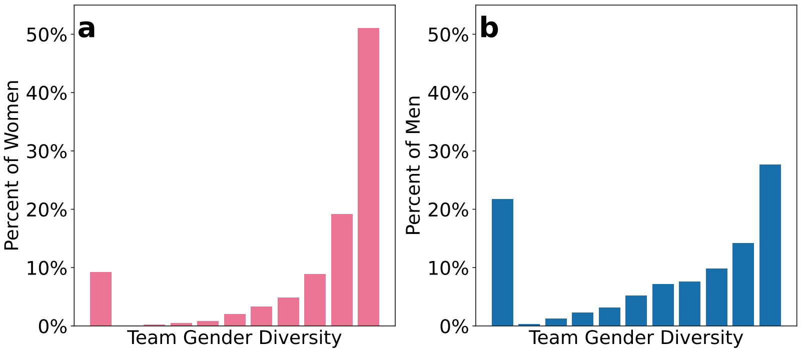


Fig. S2. Distribution of Women and Men across the Spectrum of Team Gender Diversity. (a) The distribution of women across gender diversity spectrum. (b) The distribution of men across gender diversity spectrum.

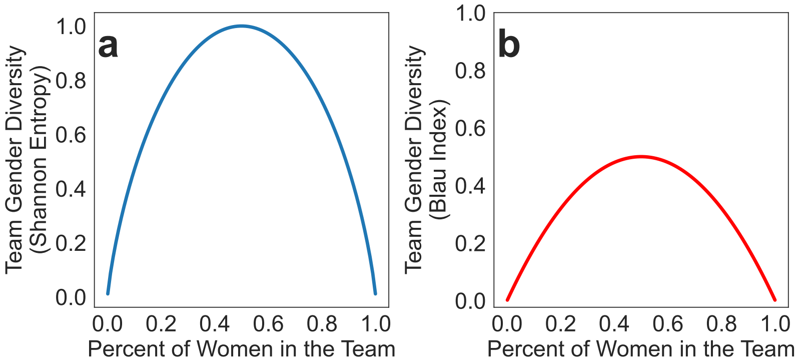


Fig. S3. Agreement between Blau and Shannon Entropy Indices in our data is (r = 1.0). (a) demonstrates the relationship between percent of women authors and Shannon Entropy. Shannon Entropy has the highest value when women and men have equivalent shares in the team. (b) demonstrates the relationship between percent of women authors and the Blau Index. Similarly, the Blau index also has the highest value when women and men have equivalent shares in the team.

We also use a continuous variable to evaluate the gender composition of a scientific team. Similarly as in the work of (6), we use Shannon Entropy to measure the gender diversity of a team, which takes the form

$$g_i = -p_f log_2\left(p_f\right) - (1-p_f)log_2\left(1-p_f\right) \qquad [8]$$

where p_f indicates the portion of women 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.

In Figure S2, we provide the distribution of women and men across different team gender diversity. Figure S2(a) plots the distribution of women authors across the spectrum of our continuous measure of team gender diversity. Specifically, we divided the teams into 11 buckets of gender diversity with 0.1 increment in the continuous measure on the x-axis. Within each bucket, the bar and percent describe the proportion of women authors who have published in teams with the respective level of gender diversity. The percent numbers add up to 100%. For example, we can see that about 9% of women authors have published in scientific teams with team gender diversity = 0, which include all-women teams. By contrast, over 50% of the women authors have published in teams with gender diversity between 0.9 and 1. Similarly, Figure S2(b) plots the distribution of men authors across the team gender diversity spectrum. Generally speaking, women authors tend to publish in teams with high team gender diversity compared to men authors.

Blau Index Besides the Shannon Entropy measure described above, the Blau index is also widely used to measure diversity. The Blau index calculates the probability of two individuals chosen at random from the team being of different gender, which was proposed in the work of (7) devoted to sociological theory.

The comparison to the Shannon Entropy shows that they produce the same type of index variation. Specifically, Figure S3 below measures team gender diversity with both the Blau index and the Shannon Entropy. As the correlation between the Blau index and Shannon Entropy is 1, we present the main analyses using the Shannon Entropy.

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. We also conduct robustness tests to address several potential concerns: (1) whether our main findings still hold when using an alternative measure of novelty (see Section 2.3); (2) whether our main conclusions hold when articles from low-impact journals are excluded (see Section 2.5); (3) whether noises in gender designations by Namsor software (1) could significantly affect our conclusions (see Section 3.3); (4) whether missing data for authors with first initials are critical enough to significantly affect our existing conclusions (see Section 3.4; and (5) whether our main conclusions can generalize across different sub-fields in medicine (see Section 4).

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

$$\begin{aligned} y_i = \beta_m m_i + \sum_t \beta_t T_{ti} + \sum_t \beta_{mt} m_i T_{ti} + \beta_f f_i + \beta_l l_i + \sum_r \beta_r R_{ri} + \sum_d \beta_d D_{di} + \sum_e \beta_e E_{ei} \\ \sum_a \beta_a A_{ai} + \sum_h \beta_h H_{hi} + \sum_p \beta_p P_{pi} + \sum_q \beta_q Q_{qi} + \sum_s \beta_s S_{si} + \sum_j \beta_j J_{ji} + \in_i \end{aligned} \tag{9}$$

Dependent Variable. The dependent variable y_i measures whether a paper is novel or not, which is measured by the variable *novel paper* 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 *novel paper*. 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 7). Furthermore, a Shannon Entropy measure g_i is used to evaluate the detailed information of team gender composition (see definition in the equation 8). Similarly, our main results are based on the regression analysis using the binary variable m_i. 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 (2) 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 at the time of publication and author's career age. In model (1), we don't include the interaction term between m_i and team size. We add the interaction term between m_i and team size in model (2) to explore the relationsip between m_i and team size in predicting papers' novelty. In addition, when both dependent variable and key independent variable are measured as continuous variables, our observations in model (3) and model (4) 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. For example, the coefficients of f_i and l_i in model (1) are -0.0061(0.00066) and -0.0089(0.00072) respectively. This is consistent with observations made in the work of (8) that minority groups are more like to produce novel scientific papers.

Table S1. Relationship between team gender composition and paper novelty. In model (1) and (2), a binary variable m_i is used to indicate whether a scientific team is mixed-gender or not. The results in model (1) and (2) demonstrate that mixed-gender teams are more novel than same-gender teams across different team sizes. In model (3) and (4), a continuous variable g_i is used to measure the team gender balance. The results are consistent with the observations in model (1) and (2).

		Model (1)		Model (2)	Model (3)	Model (4)
Variable		*novel paper*		*novel paper*	*novelty*	*novelty*
Mixed-gender Team (m_i)		0.020*** (0.00067)		0.0051*** (0.0013)		
Team Size FE (T_{ti})	$t = 3$	0.016*** (0.00089)	$t = 3$	0.016*** (0.0012)		
	$t = 4$	0.031*** (0.00093)	$t = 4$	0.032*** (0.0013)		
	$t = 5$	0.040*** (0.00097)	$t = 5$	0.038*** (0.0015)		
	$t > 5$	0.062*** (0.00090)	$t > 5$	0.047*** (0.0013)		
$m_i \times T_{ti}$			$m_i = 1$ $t = 3$	0.0039* (0.0018)		
			$m_i = 1$ $t = 4$	0.0056** (0.0018)		
			$m_i = 1$ $t = 5$	0.011*** (0.0019)		
			$m_i = 1$ $t > 5$	0.025*** (0.0017)		
Team Gender Diversity (g_i)					0.41*** (0.0052)	0.44*** (0.017)
Team Size (t)					0.0025*** (0.00050)	0.0073*** (0.0029)
$g_i \times t$						-0.0053 (0.0031)
Controls†		Y		Y	Y	Y
Observations		4,822,724		4,822,724	4,822,724	4,822,724
R Squared		0.28		0.28	0.33	0.33

† The detail of control variables and fixed effects can be found in section 2.1.
*, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$.

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:

$$z_i = \beta_m m_i + \sum_t \beta_t T_{ti} + \sum_t \beta_{mt} m_i T_{ti} + \beta_f f_i + \beta_l l_i + \sum_r \beta_r R_{ri} + \sum_d \beta_d D_{di} + \sum_e \beta_e E_{ei} + \sum_a \beta_a A_{ai} + \sum_h \beta_h H_{hi} + \sum_p \beta_p P_{pi} + \sum_q \beta_q Q_{qi} + \sum_s \beta_s S_{si} + \sum_j \beta_j J_{ji} + \in_i \quad [10]$$

Dependent Variable. The dependent variable z_i measures whether a paper is an upper-tail paper being the top 5% papers gauged by citation, which is measured by the variable *upper-tail paper* defined in the equation 5. An alternative measure is a continuous variable *impact* defined in equation 6. Our main results are based on the analysis using the binary variable *upper-tail paper*.

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 using the binary variable m.

The results of regression analysis are presented in Table S2. The results in model (1) and (2) 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 at the time of publication and author's career age. We find that mixed-gender teams are more likely to produce highly cited scientific papers. We add the interaction term between m_i and team size in model (2) to explore the relationship between m_i and team size in predicting papers' impact. In addition, our observations in model (1) and model (2) remain valid when we measure both the dependent variable and the key independent variable as continuous variables (model (3) and model (4)). 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, the coefficients of f_i and l_i in Table S2 model (1) are 0.0036(0.00043) and 0.0036(0.00050) respectively when predicting a paper's citation impact. Results in both model (2), model (3) and (4) suggest that scientific teams with men as first or last author are more likely to be cited, which is consistent with existing literature (8).

Table S2. Relationship between team gender composition and paper impact. In model (1) and (2), a binary variable m_i is used to indicate whether a scientific team is mixed-gender or not. The results in model (1) and (2) demonstrate that mixed-gender teams are more likely to produce highly cited papers than same-gender teams across different team sizes. In model (3) and (4), a continuous variable g_i is used to measure the team gender balance. The results are consistent with the observations in model (1) and (2).

Variable		Model (1) *upper-tail paper*		Model (2) *upper-tail paper*	Model (3) *impact*	Model (4) *impact*
Mixed-gender Team (m_i)		0.0092*** (0.00038)		0.010*** (0.00072)		
Team Size FE (T_{ti})	$t = 3$	0.026*** (0.00047)	$t = 3$	0.027*** (0.00062)		
	$t = 4$	0.043*** (0.00050)	$t = 4$	0.046*** (0.00067)		
	$t = 5$	0.046*** (0.00053)	$t = 5$	0.051*** (0.00074)		
	$t > 5$	0.070*** (0.00053)	$t > 5$	0.067*** (0.00075)		
$m_i \times T_{ti}$			$m_i = 1$ $t = 3$	-0.0027** (0.00090)		
			$m_i = 1$ $t = 4$	-0.0049*** (0.00091)		
			$m_i = 1$ $t = 5$	-0.0067*** (0.00096)		
			$m_i = 1$ $t > 5$	0.0024*** (0.00092)		
Team Gender Diversity (g_i)					0.089*** (0.0012)	0.18*** (0.010)
Team Size (t)					0.0024*** (0.00025)	0.017*** (0.0020)
$g_i \times t$						-0.016*** (0.0020)
Controls†		Y		Y	Y	Y
Observations		4,822,724		4,822,724	4,822,724	4,822,724
R Squared		0.34		0.34	0.51	0.52

† The detail of control variables and fixed effects can be found in section 2.1.
*, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$.

2.3. Results of Novelty Measure Based on Subject Pairings.

Our main finding that mixed-gender teams are more novel than same-gender teams is based on the novelty measure using journal information (3) (see Section 1.2.2). The novelty measure based on journal pairings (3) considered papers with statistically atypical journal pairings to be novel because they create new combinations of knowledge that have not been joined, or rarely joined, in previous research.

In the work of (3), a journal is considered to be focused on several specialties. Therefore, journal pairings can be used to measure novel combination of knowledge in a granular level. We also use an alternative novelty measure based on subjects to evaluate novel combination of knowledge. In the work of (4), a Sterling index is used to evaluate a paper's tendency to

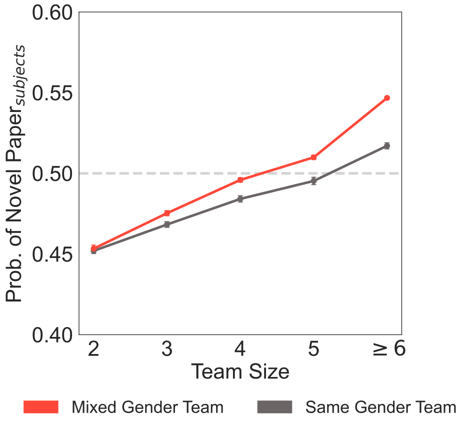


Fig. S4. Robustness Check for the Relationship Between Gender Balance and Novelty. We use novelty score based on subject pairings to measure novelty level of scientific papers. This figure shows that gender-diverse gender teams are more likely to produce novel papers than same-gender teams at all team sizes, and large gender-diverse teams (6+ authors) are 9.3% more likely to publish novel work than the base rate (0.5). The result is consistent with our observations in **Figure 2** of manuscript.

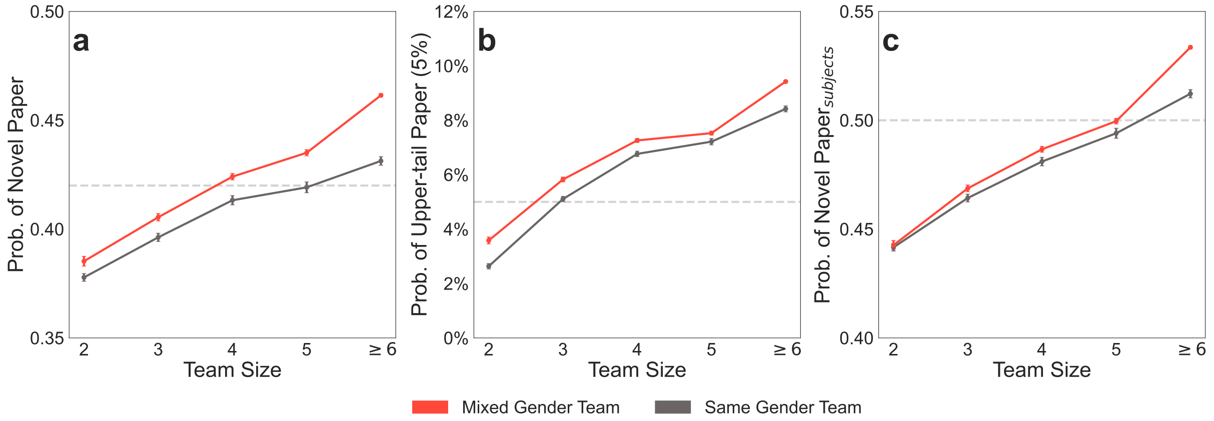


Fig. S5. Mixed-gender Teams are More Novel and highly cited Where Novelty and impact are Calculated Excluding Self-citations. (a) Mixed-gender teams are more likely to produce novel papers without self-citations than same-gender teams across all team sizes; for team sizes of 4 and above, mixed-gender teams are always more likely to produce a novel paper than the base rate (dashed line). (b) Similarly, mixed-gender teams are more likely to produce a highly cited paper without self-citations than same-gender teams of all team sizes. Furthermore, mixed-gender teams of size 3 and above are more likely to publish a highly cited paper than the base rate (dashed line). (c) Finally, mixed-gender teams are more novel in terms of novelty measure based on subject pairings without self-citations. Estimates shown are margins plots computed from fixed effect regressions.

reference unusual subject pairings. We argue that the novelty measure based on journal pairings and the novelty measure based on subject pairings are complementary. The novelty measure based on subjects ranges from 0 to 1, where higher value indicates higher novelty of scientific papers. Different from the novelty measure based on journal pairings which follows a heavy tailed distribution, the novelty measure based on subject pairings follows a normal distribution. In our sample of 6.6 million medical papers, the correlation between these two measures is as high as 0.43. When both measures are converted to binary variables, the agreement between the novelty measure based on journals (3) and the novelty measure based on subjects (4) is about 65%.

For robustness check, we regress novelty score based on subject pairings (see Section 1.2.3) on team gender diversity with a number of controls and fixed effects, which are consistent with our observations in Table S1.

Similar to equation 1, we define a paper as novel when its novelty score based on subject pairings is larger than median value (see equation 4).

Our result in **Figure 2(a)** in the manuscript is still valid when we use the novelty measure based on subjects to evaluate novelty of scientific papers. In Figure S4, we can see mixed-gender teams are more likely to produce novel papers in terms of novelty measure based on subjects than same-gender teams across different team sizes. The results are very consistent with our observations using novelty measure based on journals (**Figure 2(a)** in the manuscript). Large mixed-gender teams (6+) are 10.5% more likely to publish novel work than the base rate.

2.4. Robustness Test of Removing Self-citations. Previous research reveals that men tend to be more likely to engage in self-citation than women (9, 10). To test whether the gendered nature of self-citation behaviors has significant impact on our main conclusions that mixed-gender teams are more novel and impactful than same-gender teams, we analyzed whether our main findings are robust to the exclusion of self-citation behavior. To answer this question, we developed alternative measures of our dependent variables novelty and impact by removing self-citations in the calculation. In our sample of 6.6 million

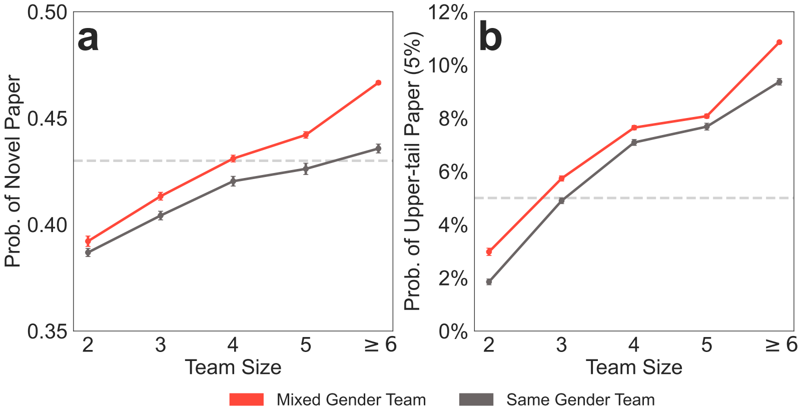


Fig. S6. Mixed-gender Teams are More Innovative (a) and highly cited (b) among ranked journals in the Scimago Journal Ranking. Mixed-gender teams are more likely to produce innovative papers than are same-gender teams of all team sizes; for teams of size 4 or more mixed-gender are always more likely 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.

medical papers published between 2000 and 2019, the proportion of self-citations per paper is about 11.7%, which is consistent with prior research that reported 13.2% (11). Figure S5 presents the robustness test where self-citations are excluded in the calculation of novelty and impact. Based on the comparison, we have several observations.

First, our main finding that mixed-gender teams are more novel than same-gender teams holds after removing self-citations in measurement. We excluded all self-citation references when calculating the novelty measure based on journal pairings. We replicated the analysis underlying **Figure 2(a)** in the manuscript by regressing the recalculated novelty measure on team gender diversity with the same set of controls and fixed effects (see see **section S2.1**). The result is presented in Figure S5(a), which is highly consistent with **Figure 2(a)** in the manuscript. Additionally, we also conducted a robustness test on novelty measured based on subject pairings. Figure S5(c) reports the result when we excluded all self-citations when computing the novelty measure based on subjects. The result in Figure S5(c) is consistent with Figure S4 where self-citations are included in the calculation of novelty measure based on subject pairings.

Second, our main finding that mixed-gender teams are more highly cited than same-gender teams also holds after excluding self-citations. We excluded all self-citations when calculating the paper impact (citation). Similarly, we reproduce our original main analysis underlying **Figure 2(b)** in the manuscript by regressing the recalculated impact variable on team gender diversity with the same set of controls and fixed effects. The finding presented in Figure S5(b) confirms that self-citation behavior does not have significant impact on our main finding in terms of impact.

In sum, Figure S5 provides results that address concerns related to the gendered nature of self-citation behaviors. Our main findings remain robust to the exclusion of self-citations.

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, $H\text{-}index = 5$ means that 5 papers published in the journal have at least 5 citations.

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 $H\text{-}index \geq 5$ can be found in MAG. And the matching ratio reaches 93.3% when we only consider Scimajo journals with $H\text{-}index \geq 10$. 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 about 5.6 million medical papers. In Figure S6, 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.

3. Robustness Test for Gender Inference Algorithm

In this section, we conduct robustness tests to address several potential concerns: (1) whether our gender inference results are consistent with existing literature (see Section 3.1); (2) whether noises in gender designations by Namsor software (1) could significantly affect our conclusions (see Section 3.3); and (3) whether missing data for authors with first initials are critical enough to significantly affect our existing conclusions (see Section 3.4).

3.1. Cross-check Gender Inference Results with Existing Literature. There are about 9.6 million scientists in the field of Medicine from 2000 to 2019. Among these authors, 79% have their full first names recorded in MAG (as opposed to the initials alone), which allows us to estimate a binary gender designation based on both the author's first name and last name. Based on the gender classification by Namsor API, about 42.3% of them are women in medicine. The remaining 21% of authors are not included in our main analysis.

Our gender inference results are found to be highly consistent with both AAMC (Association of American Medical Colleges) census data and existing literature (12) using a similar gender inference algorithm.

First, our gender detection results are consistent with AAMC (Association of American Medical Colleges) census data in 2019 and 2020, where women comprised 42.7% and 43.2% respectively (https://www.aamc.org/data-reports/faculty-institutions/report/faculty-roster-us-medical-school-faculty). This is highly consistent with our finding that women comprised around 42.3% of authors.

Second, in the work of (12), Pinho-Gomes et al. found that 34% of 6,722 authors in 1,370 COVID-19 related papers were women. To investigate our speculation and to further address the concern of gender disambiguation issues, we extracted 9,033 COVID-19 related journal articles in medicine published between Dec 2019 and Dec 2020, with a total of 60,839 authors. Using NamSor API, we found that women comprised 35.7% of all authors in this sample, which is very similar to Pinho-Gomes et al.'s finding (12) that 34% of 6,722 authors are women. This robustness test helps confirm that our gender detection results are consistent with existing literature, which also provides further evidence that the pandemic affects women disproportionately.

Thirdly, Nielsen et. al. (13) found that women comprised 40% of first authors, 27% of last authors and, on average, 35% of authors per paper using a sample of 1.5 million disease-specific medical papers published between 2008 and 2015. The analysis is done using the Gender API (https://gender-api.com/), which is a comparable gender inference algorithm with NamSor API (2). Using the sub-field information (see Table S5), we extracted a similar sample of 1.6 million disease-specific medical papers published between 2008 and 2015 in MAG. According to the gender designations by NamSor API, we find that women comprised 39% of first authors, 28% of last authors and, on average, 35% of authors per paper in the extracted sample. This further shows that our gender detection results are consistent with existing literature. This also implies that there is a high agreement between NamSor API and Gender API.

Table S3. Robustness tests for relationship between team gender composition and paper novelty. We run regressions on multiple subsets where scientists are included only when their gender confidence scores are larger than a threshold. The results of threshold equal to 60%, 70%, 80% and 90% are provided in model (1) to model (4) respectively. In model (5), we report the result of regression where we include authors with first initials with simulated gender labels. The results are consistent with model (1) in Table S1. Comparing to the coefficient of m_i in Table S1 model (1) (0.020), the difference ranges from 5% to 10%.

	Model (1)		Model (2)		Model (3)		Model (4)		Model (5)	
Variable	threshold=60%		threshold=70%		threshold=80%		threshold=90%		Include Authors with First Initials	
Mixed-gender Team (m_i)	0.021*** (0.00068)		0.021*** (0.00070)		0.021*** (0.00072)		0.021*** (0.00074)		0.018*** (0.00064)	
Team Size FE (T_{ti})	t = 3	0.016*** (0.00091)	t = 3	0.016*** (0.00093)	t = 3	0.016*** (0.00096)	t = 3	0.016*** (0.00098)	t = 3	0.015*** (0.00083)
	t = 4	0.031*** (0.00096)	t = 4	0.031*** (0.00097)	t = 4	0.031*** (0.0010)	t = 4	0.032*** (0.0010)	t = 4	0.031*** (0.00087)
	t = 5	0.040*** (0.0010)	t = 5	0.040*** (0.0010)	t = 5	0.040*** (0.0011)	t = 5	0.041*** (0.0011)	t = 5	0.040*** (0.00091)
	$t > 5$	0.062*** (0.00092)	$t > 5$	0.062*** (0.00094)	$t > 5$	0.063*** (0.00097)	$t > 5$	0.063*** (0.0010)	$t > 5$	0.061*** (0.00085)
Controls†	Y		Y		Y		Y		Y	
Observations	4,544,652		4,309,503		4,030,382		3,742,596		5,624,228	
R Squared	0.28		0.29		0.29		0.29		0.28	

† The detail of control variables and fixed effects can be found in section 2.1.
*, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$.

3.2. Cross-check Gender Inference Results with Alternative Algorithms. Gender API (https://gender-api.com/) is an alternative gender inference algorithm, which is demonstrated to have comparable performance with NamSor API (2). To further verify the level of agreement between these two software, we extract (1) top 5,000 frequent first names and (2) 1,000 randomly selected non-frequent first names in our sample as a test set. Gender API and NamSor API are applied to infer gender for those 6,000 names separately. We find that the agreement level is as high as 97.1%.

NamSor API v2.0.11 was used to infer the gender of authors in our analysis. The version was first released on October 31 2020. And the latest version of NamSor API is v2.0.18, which was released on Jan-16-2022. To test the level of consistency between NamSor API v2.0.11 and NamSor API is v2.0.18, we conduct similar analysis described above. We find that the agreement between v2.0.11 and v2.0.18 is about 97.9%.

3.3. 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.

Table S4. Robustness tests for relationship between team gender composition and paper impact. We run regressions on multiple subsets where scientists are included only when their gender confidence scores are larger than a threshold. The results of threshold equal to 60%, 70%, 80% and 90% are provided in model (1) to model (4) respectively. In model (5), we report the result of regression where we include authors with first initials with simulated gender labels. The results are consistent with model (1) in Table S2. Compared to the coefficient of m_i in Table S2 model (1) (0.0092), the difference ranges from 3.5% to 8.7%.

		Model (1)		Model (2)		Model (3)		Model (4)		Model (5)
Variable		threshold=60%		threshold=70%		threshold=80%		threshold=90%		Include Authors with First Initials
Mixed-gender Team (m_i)		0.0095*** (0.00039)		0.0098*** (0.00041)		0.0099*** (0.00043)		0.010*** (0.00045)		0.0086*** (0.00036)
Team Size FE (T_{ti})	t = 3	0.026*** (0.00049)	t = 3	0.026*** (0.00050)	t = 3	0.027*** (0.00051)	t = 3	0.027*** (0.00053)	t = 3	0.025*** (0.00043)
	t = 4	0.044*** (0.00051)	t = 4	0.044*** (0.00052)	t = 4	0.045*** (0.00055)	t = 4	0.045*** (0.00056)	t = 4	0.042*** (0.00046)
	t = 5	0.047*** (0.00055)	t = 5	0.048*** (0.00056)	t = 5	0.049*** (0.00058)	t = 5	0.049*** (0.00060)	t = 5	0.045*** (0.00049)
	$t > 5$	0.071*** (0.00055)	$t > 5$	0.072*** (0.00056)	$t > 5$	0.074*** (0.00058)	$t > 5$	0.074*** (0.00060)	$t > 5$	0.068*** (0.00049)
Controls†		Y		Y		Y		Y		Y
Observations		4,544,652		4,309,503		4,030,382		3,742,596		5,624,228
R Squared		0.35		0.35		0.35		0.35		0.34

† The detail of control variables and fixed effects can be found in section 2.1.
*, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$.

The output of Namsor API allows us to conduct such a robustness test. In addition to a binary gender label, NamSor API also provides a confidence score calibrated using a large-scale database of names collected from different countries and regions (1). 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.

To be specific, 86% of scientists in our data have gender confidence scores larger than 80%. And 81% of them have gender confidence scores larger than 90%. This suggests that most of the authors have confident gender designations.

To verify whether noises of gender inference 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%. And the results are consistent with our findings in Table S1. In Table S4, the coefficients of m_i in predicting a paper's impact are also 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.

3.4. 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.

- We calculate authors' career age by using their first publication years as below:

$$\textit{career age} = 2019 - \textit{first publication year} + 1 \qquad [11]$$

And we categorize the career age into bins: $[0, 5]$, $[6, 10]$, $[11, 15]$, $[16, 20]$, $[21, 25]$, $[26, 30]$, $[31, 35]$, $[36, 40]$, $[41, 45]$, $[46, 50]$, $[51, Inf]$.

- 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]".

- 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.

- For an individual author with "S." as first initial and who is in a given career age group, we randomly classify her/him as woman or man 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.

4. Generalizability across Sub-fields in Medicine

In this section, we investigate the generalizability of our findings across 45 sub-fields in Medicine, such as "anatomy" and "biomedical engineering". Those 45 sub-fields are listed in Table S5.

Table S5. Sub-fields 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

First, 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 *novel paper* and *upper-tail paper* as dependent variables in the regression. The definition of these two variables can be found in Section 1.2. In the second setting, we are using *novelty* and *impact* measures as our dependent variables (see definitions in equation 2 and equation 6). The detailed information can also be found in Section 1.2.

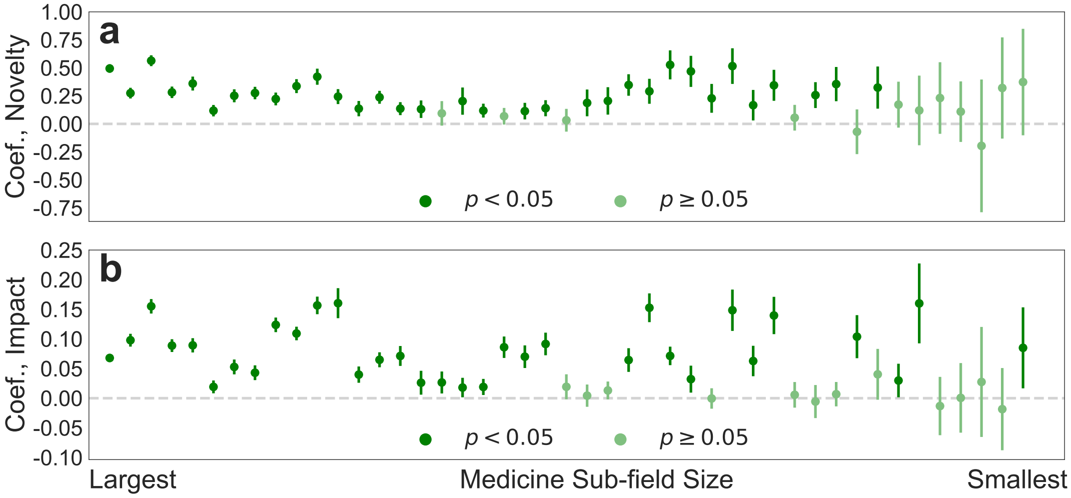


Fig. S7. Coefficients of Team Gender Diversity across 45 Sub-fields in Medicine (Dependent variables are also continuous). Each bar indicates the coefficient of team gender diversity with 95% confidence interval in a sub-field. We sort sub-fields from smallest to largest. The observations are consistent with **Figure 2(c)** and **2(d)** in manuscript. (a) Coefficients of team gender diversity in predicting papers' novelty. Dark green color indicates significant (p-value < 0.05) coefficients. Light green color indicates non-significant coefficients. We can see 43 out of 45 sub-fields (99.1% papers in medicine) have positive coefficients. In addition, 33 of them are significant (91.2% papers in medicine). This demonstrates the generalizability of our main finding that team gender diversity is predictive of papers' novelty. (b) Similar to sub figure (a), 41 out of 45 sub-fields (97.6% papers in medicine) have positive coefficients when predicting papers' impact. And the coefficients are both positive and significant in 33 sub-fields (90.5% papers in medicine).

In **Figure 2(c)** and **2(d)** of manuscript, we visualize the coefficients with 95% confidence intervals for 45 sub-fields when binary measures *novel paper* and *upper-tail 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 sub-fields. For example, when predicting novel papers, team gender diversity has consistent and positive sign in 43 out of 45 sub-fields (**Figure 2(c)**). When predicting upper-tail papers, the sign of team gender diversity is consistent and positive in 30 out 45 sub-fields.

Furthermore, we also verify the generalizability when dependent variables are continuous. In Figure S7, we observe that the sign of team gender diversity is consistent and positive in most sub-fields. For example, when predicting papers' novelty, team gender diversity has consistent and positive sign in 43 out of 45 sub-fields (Figure S7(a)). When predicting papers' impact, the sign of team gender diversity is consistent and positive in 41 out 45 sub-fields.

4.1. Regression Analysis Controlling for Sub-field Fixed Effect. To further examine how the main findings might vary across different sub-fields in medicine, we conducted several more tests.

We control for the sub-field fixed effect, such as "cardiology" and "immunology". And we find that our main findings are consistent with **Figure 2** in the manuscript. The results are presented in Figure S8. We can see that mixed-gender teams are

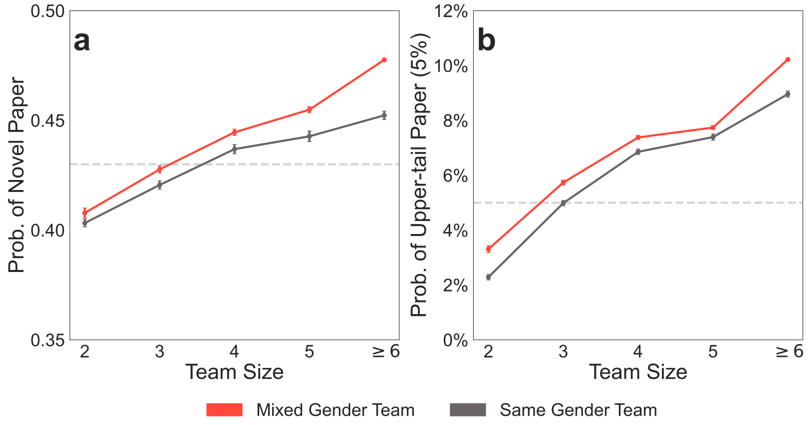


Fig. S8. Mixed-gender Teams are More Innovative (a) and Highly Cited (b) Where Sub-field Fixed Effect is Controlled. (a) Mixed-gender teams are more likely to produce novel papers than are same-gender teams of all team sizes; for teams of size 4 or more mixed-gender are always more likely to produce a novel paper than the base rate (dashed line). (b) Similarly, mixed-gender teams are more likely to produce a highly cited paper than same-gender teams of all team sizes. Estimates shown are margins plots computed from fixed effect regressions.

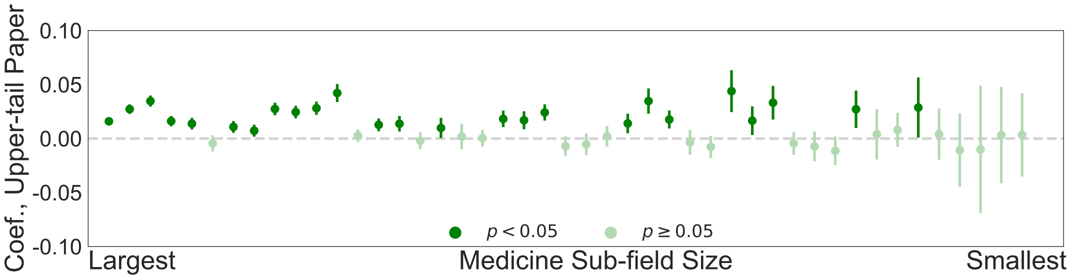


Fig. S9. Coefficients of Team Gender Diversity across 45 Sub-fields in Predicting Upper-tail Paper (the paper impact is normalized by sub-field year average). Each bar indicates the coefficient of team gender diversity with 95% confidence interval in a sub-field predicting papers' impact. The paper impact is normalized by sub-field year average. We sort sub-fields from largest (left) to smallest (right). Dark green color indicates significant (p-value < 0.05) coefficients. Light green color indicates insignificant coefficients. We observe that 34 out of 45 sub-fields (85.5% papers in medicine) have positive coefficients when predicting upper-tail paper. And the coefficients are both positive and significant in 25 sub-fields (75.5% papers in medicine).

more likely to produce novel and impactful scientific ideas than same-gender teams after controlling for sub-field fixed effect (Figure S8(a) and Figure S8(b)). In Figure S8(a), we find that mixed-gender teams are more likely to produce novel papers than are same-gender teams of all team sizes. And large mixed-gender teams (6+) are about 7.9% more likely to produce a novel paper than the base rate (dashed line). In Figure S8(b), mixed-gender teams are demonstrated to be more likely to produce a highly cited paper than same-gender teams of all team sizes. And large mixed-gender teams (6+) are about 14.5% more likely to produce a highly cited paper than same-gender teams of similar size.

4.2. Robustness Test for Paper Impact Normalized by Sub-field Year. In Section 1.2.4, the paper impact is normalized by year average. Because there are variations in citations across different sub-fields, we further measure the paper impact by normalizing citation by sub-field year average. Based on the normalized citation, we further measured high impact papers by the top 5% upper-tail papers in each sub-field and ran a separate regression for each sub-field. The results are presented in Figure S9, which further confirms that mixed-gender teams are more likely to produce highly cited papers despite potential citation variations across different sub-fields in medicine.

In Figure S9, the x-axis indicates the size of the sub-field, sorting the sub-fields from the largest on the left to the smallest on the right. We observe that 34 out of 45 sub-fields (85.5% papers in medicine) have positive coefficients when predicting top 5% upper-tail paper. The coefficients are both positive and significant in 25 sub-fields (75.5% papers in medicine). And most of the noisy relationships (non-significant) are exhibited in comparably small sub-fields, such as "andrology" and "traditional medicine".

In conclusion, based on the robustness tests described above, we confirm that our main findings are valid despite potential citation variations across different medical sub-fields.

5. Generalizability across Scientific Fields

Besides expanded analyses across 45 sub-fields in medicine, we also collected new data to investigate in which scientific fields other than medicine our main findings still hold. Including medicine, we detect the gender of about 33 million authors who wrote about 26 million journal articles between 2000 and 2019 in 19 scientific fields, such as Economics, Psychology and Sociology. According to the NamSor API, about 36.7% of 33 million authors are women.

Using the same regression specification described in Section 2, we further control for the field fixed effect, such as Economics, Psychology and Medicine. And we find that our main findings still hold. The results in Figure S10 indicate that the positive association between mixed-gender teams and novelty and impact hold for all sciences and display a similar pattern of effects. As shown in Figure S10(a), about 41% of scientific papers have novelty scores smaller than zero (see Section 1.2.2). Similar to **Figure 2** in the manuscript, we observe that large mixed-gender teams (6+ authors) are about 5% more novel than same-gender teams. And large mixed-gender teams (6+ authors) are also 10.7% more likely to produce highly cited papers. In conclusion, our main finding in Medicine that mixed-gender teams publish more innovative and impactful research than same-gender teams still hold for all sciences.

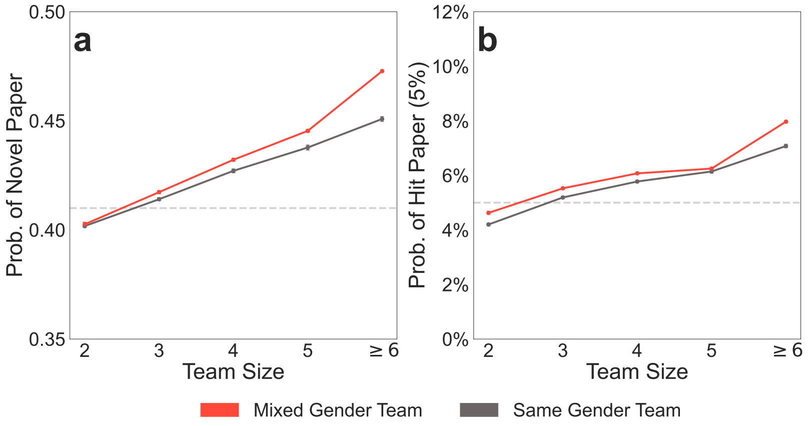


Fig. S10. Mixed-gender Teams are More Novel (a) and Highly Cited (b) where Field Fixed Effect is Controlled for. (a) Mixed-gender teams are more likely to produce novel papers than are same-gender teams of all team sizes; for teams of size 3 or more mixed-gender are always more likely to produce a novel paper than the base rate (dashed line). (b) Similarly, mixed-gender teams are more likely to produce a highly cited paper than same-gender teams of all team sizes. Estimates shown are margins computed from fixed effect regressions.

6. 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.

6.1. 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.

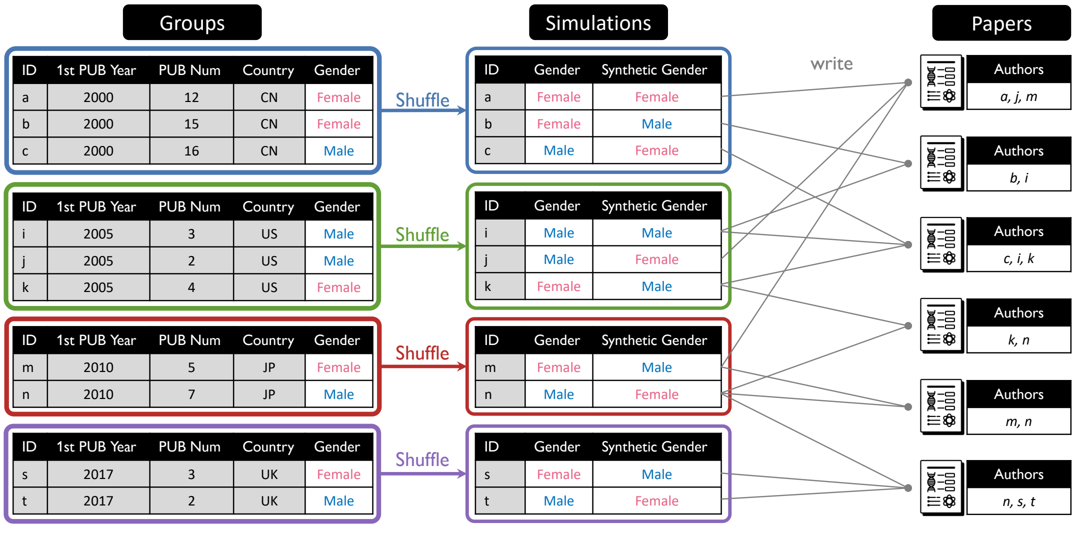
- For each author in our sample, we extract her first publication year (publication year of her first paper), total number of publications, and the country where her affiliated institution is located (i.e., CN, US, UK, JP and etc.).
- Second, we categorize the total number of publications into bins: $[0, 1]$, $[2, 4]$, $[5, 8]$, $[9, 16]$, $[17, 32]$, $[33, 64]$, $[65, 128]$, $[129, 256]$, $[257, 512]$, $[513, 1024]$, $[1025, Inf]$.
- 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; (iii) their affiliated institutions are in the same country; and (iv) their primary publications are in the same sub-field. In Figure S11, 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.
- 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 S11 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.
- 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 under-representation 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-gender 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:

$$z = \frac{(obs - exp)}{\sigma} \qquad [12]$$

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

Authors
a, j, m

Authors
b, i

Authors
c, i, k

Authors
k, n

Authors
m, n

Authors
n, s, t

Fig. S11. 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.

6.1.1. Same-gender Teams are Over-represented. Using the same null model explained above, we can also verify whether all-women teams and all-men teams are over-represented over time.

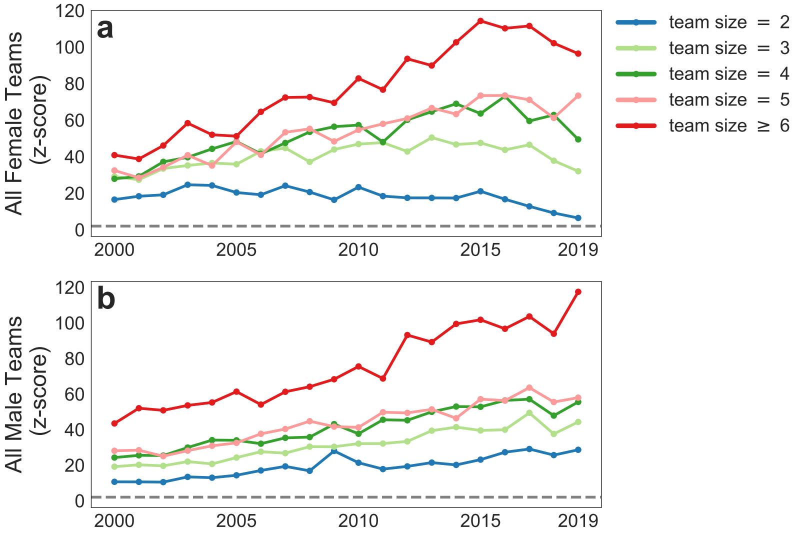


Fig. S12. Same-gender Teams Remain Over-represented in Medical Science. Using the same null model, we verify whether same-gender teams are over-represented over time. We can see both female teams and male teams are over-represented.

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 over-represented in medical science. This is supported by results in the FigureS12. In FigureS12, we can see both all-women teams and all-men teams are significantly over-represented despite the growing trend of mixed-gender teams.

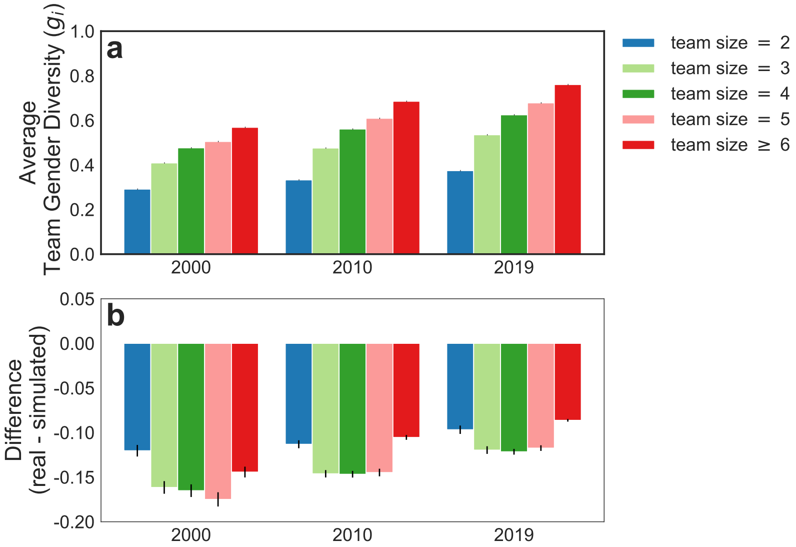


Fig. S13. Gender Balanced Teams Dominate Science. Instead of using a binary variable m_i, we use a Shannon Entropy measure g_i to evaluate gradations of gender diversity. (a) shows the average team gender diversity increased steadily over time. (b) shows the results of a null model that indicate that the observed team gender diversity remains underrepresented in medical science.

6.2. Robustness Test for Figure 1c and 1e Using a Shannon Entropy Measure. In this section, we demonstrate that our observations in **Figure 1(c)** and **1(e)** 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 8 of Section 1.2.5.

In the manuscript, we use a binary variable m_i (see equation 7) 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 1(c)**, we are now measuring the average g_i for each team size category.

In Figure S13, we find the results using team gender diversity g_i are consistent with what we observe in **Figure 1(c)** and **1(e)** in the manuscript.

6.3. 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 1.2.2).

6.3.1. Robustness Test for Figure 1(a) 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 S14, 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).

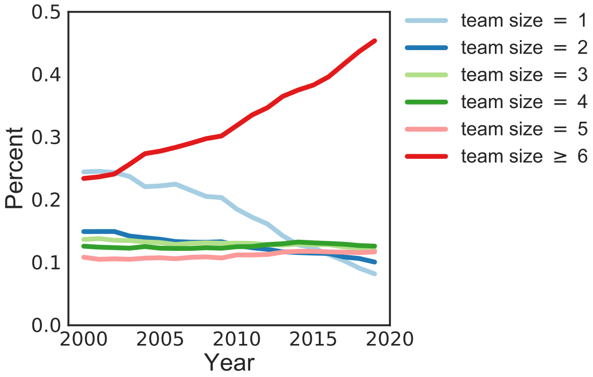


Fig. S14. Big Teams Dominate Medical Science. This figure plots the share of publications (y-axis) by team size and year from 2000-2019 (x-axis). Over time, large teams have replaced small teams. For example, in 2000 solo and two-person teams each had more than 15% of the share of publications but by 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%.

6.3.2. Robustness Test for Figure 1(c) and 1(e) by Including Papers without References. In **Figure 1(c)** and **1(e)** of the manuscript, we have demonstrated that mixed-gender teams steadily increased over time. Similarly, we also replicate the results of **Figure 1(c)** and **1(e)** in the manuscript using the sample of 12 million medical journal articles, which is presented in Figure S15. The results are consistent with our observations in the manuscript.

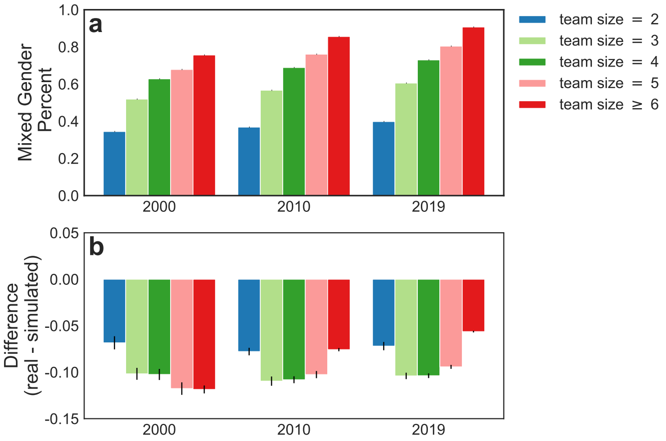


Fig. S15. Mixed-gender Teams Dominate Science. Using all 12 million papers, we repeat the analysis in **Figure 1(c)** and **1(e)** and find similar results. (a) shows the share of publications from mixed- vs same-gender teams steadily increased with time and that the increase is proportionally greater 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.

7. Mixed-gender Teams Led by Women and Men

In this section, we analyze whether gender of leaders have different effects on our findings about mixed-gender teams. Here, we estimate leadership by using author's order. Both first author and last author are considered as leaders in a scientific team.

In this way, we can classify mixed-gender teams into the following categories:

- mixed-gender team led by woman first author
- mixed-gender team led by man first author
- mixed-gender team led by woman last author
- mixed-gender team led by man last author

We further define two variables

$$mf_i = \begin{cases} 0, & m_i = 1 \textit{ and } f_i = 0 \\ 1, & m_i = 1 \textit{ and } f_i = 1 \\ 2, & m_i = 0 \end{cases} \qquad [13]$$

$$ml_i = \begin{cases} 0, & m_i = 1 \textit{ and } l_i = 0 \\ 1, & m_i = 1 \textit{ and } l_i = 1 \\ 2, & m_i = 0 \end{cases} \qquad [14]$$

where the definitions of m_i, f_i, and l_i can be found in section 2.1.

7.1. Gender of First Author and Novel Paper. First, we examine whether woman first author led mixed-gender teams and man first author led mixed-gender teams have different levels of novelty using a fixed-effect ordinary least squares regression as below:

$$y_i = \beta_m mf_i + \sum_t \beta_t T_{ti} + \beta_l l_i + \sum_r \beta_r R_{ri} + \sum_d \beta_d D_{di} + \sum_e \beta_e E_{ei}$$
$$\sum_a \beta_a A_{ai} + \sum_h \beta_h H_{hi} + \sum_p \beta_p P_{pi} + \sum_q \beta_q Q_{qi} + \sum_s \beta_s S_{si} + \sum_j \beta_j J_{ji} + \in_i \qquad [15]$$

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

Predictors of Interest. We use a variable mf_i to indicate whether a scientific team is woman first author led mixed-gender team, man first author led mixed-gender team or same-gender team (see definition in the equation 13).

Control Variables. We also include several other explanatory variables to control for other possible predictors. Details can be found in section 2.1.

In Figure S16(a) left panel, we consider the gender of leadership by the first authors and examine the dependent variable of novelty. We find that women-led mixed-gender teams are more likely to produce novel outputs than men-led mixed-gender teams. Despite the difference by the gender of leadership, we note that both of them are significantly more novel than same-gender teams. However, the difference between women led mixed-gender teams and men led mixed-gender teams is much smaller than the difference between mixed-gender teams and same-gender teams. For example, the difference between men led mixed-gender teams and same-gender teams is about 3.5 times of its difference with women led mixed-gender teams. Similarly, Figure S16(a) right panel considers the gender of leadership by the last authors and presents consistent results.

7.2. Gender of Last Author and Novel Paper. Second, we consider the gender of leadership by the last authors and examine the dependent variable of novelty using a fixed-effect ordinary least squares regression as below:

$$y_i = \beta_m ml_i + \sum_t \beta_t T_{ti} + \beta_l l_i + \sum_r \beta_r R_{ri} + \sum_d \beta_d D_{di} + \sum_e \beta_e E_{ei}$$
$$\sum_a \beta_a A_{ai} + \sum_h \beta_h H_{hi} + \sum_p \beta_p P_{pi} + \sum_q \beta_q Q_{qi} + \sum_s \beta_s S_{si} + \sum_j \beta_j J_{ji} + \in_i \qquad [16]$$

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

Predictors of Interest. We use a variable ml_i to indicate whether a scientific team is woman last author led mixed-gender team, man last author led mixed-gender team or same-gender team (see definition in the equation 14).

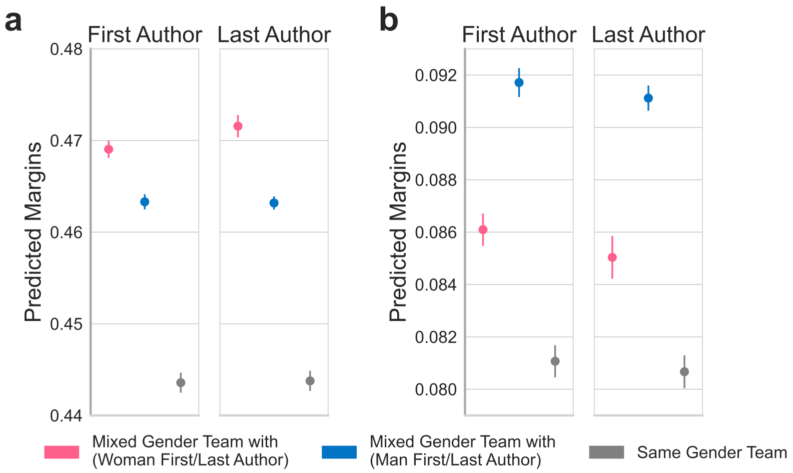


Fig. S16. Gender of Leadership in Predicting Novelty and Impact. In panel (a), we find that women first/last author led mixed-gender teams are more novel than men first/last author led mixed-gender teams and same-gender teams. In panel (b), we observe that men first/last author led mixed-gender teams are more impactful than women first/last author led mixed-gender teams and same-gender teams. Importantly, the difference between mixed-gender teams and same-gender teams remains much larger than that among mixed-gender teams by the gender of leadership.

Control Variables. We also include several other explanatory variables to control for other possible predictors. Details can be found in section 2.1.

Consistent with our findings in Figure S16(a) left panel, we find that woman last author led mixed-gender teams are more likely to produce novel papers than man last author led mixed-gender teams. And both of them are significantly more novel than same-gender teams.

7.3. Gender of First Author and Upper-tail Paper. Thirdly, we examine whether woman first author led mixed-gender teams and man first author led mixed-gender teams have different levels of paper impact using a fixed-effect ordinary least squares regression as below:

$$y_i = \beta_m mf_i + \sum_t \beta_t T_{ti} + \beta_l l_i + \sum_r \beta_r R_{ri} + \sum_d \beta_d D_{di} + \sum_e \beta_e E_{ei}$$
$$\sum_a \beta_a A_{ai} + \sum_h \beta_h H_{hi} + \sum_p \beta_p P_{pi} + \sum_q \beta_q Q_{qi} + \sum_s \beta_s S_{si} + \sum_j \beta_j J_{ji} + \in_i \qquad [17]$$

Dependent Variable. The dependent variable y_i measures whether a paper is upper-tail paper or not, which is measured by the variable *upper-tail paper* defined in equation 5.

Predictors of Interest. We use a variable mf_i to indicate whether a scientific team is woman first author led mixed-gender team, man first author led mixed-gender team or same-gender team (see definition in the equation 13).

Control Variables. We also include several other explanatory variables to control for other possible predictors. Details can be found in section 2.1.

In a similar vein, Figure S16(b) examines the dependent variable of impact, wherein the left panel considers the gender of leadership by the first authors and the right panel considers the gender of leadership by the last authors. We find that men-led mixed-gender teams receive more citations than women-led mixed-gender teams. Importantly, both of them are much more impactful than same-gender teams. For example, women (first author) led mixed-gender teams are about 6.1% less impactful than men (first author) led mixed-gender teams, while same-gender teams are about 12% less impactful.

7.4. Gender of Last Author and Upper-tail Paper. Thirdly, we examine whether woman last author led mixed-gender teams and man last author led mixed-gender teams have different levels of paper impact using a fixed-effect ordinary least squares regression as below:

$$y_i = \beta_m ml_i + \sum_t \beta_t T_{ti} + \beta_l l_i + \sum_r \beta_r R_{ri} + \sum_d \beta_d D_{di} + \sum_e \beta_e E_{ei}$$
$$\sum_a \beta_a A_{ai} + \sum_h \beta_h H_{hi} + \sum_p \beta_p P_{pi} + \sum_q \beta_q Q_{qi} + \sum_s \beta_s S_{si} + \sum_j \beta_j J_{ji} + \in_i \qquad [18]$$

Dependent Variable. The dependent variable y_i measures whether a paper is upper-tail paper or not, which is measured by the variable *upper-tail paper* defined in equation 5.

Predictors of Interest. We use a variable ml_i to indicate whether a scientific team is woman last author led mixed-gender team, man last author led mixed-gender team or same-gender team (see definition in the equation 14).

Control Variables. We also include several other explanatory variables to control for other possible predictors. Details can be found in section 2.1.

In Figure S16(b) right panel, we find that man last author led mixed-gender teams are more likely to produce impactful papers than woman last author led mixed-gender teams. Importantly, both of them are much more impactful than same-gender teams.

8. Direction of Team Gender Diversity

The gender balance in this study is measuring to what degree women and men are evenly distributed because we are trying to emphasize on the relationship between gender balance and team performance.

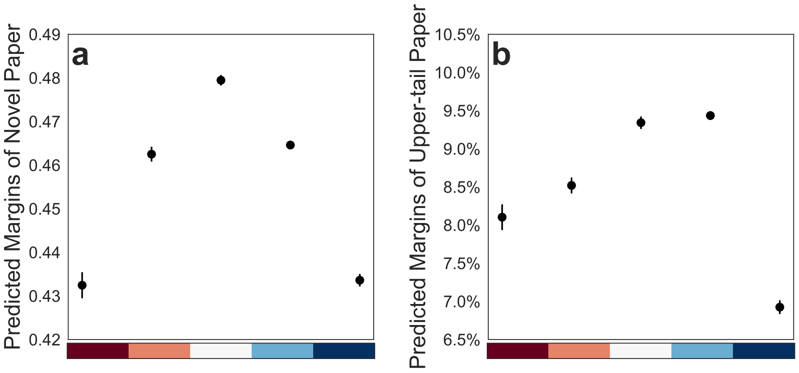


Fig. S17. Margins of Five Categories of Teams in Predicting Novel Paper and Upper-tail Paper. Dark red, light red, white, light blue and dark blue represent all women teams, teams with more women than men, teams with roughly equivalent numbers of women and men, teams with more men than women, and all men teams respectively. (a) demonstrates margins of five categories of teams in predicting novel paper. (b) demonstrates margins of five categories of teams in predicting upper-tail paper.

Here, we engaged in more analyses to further investigate the gender balance in both directions. Specifically, we categorized the teams into five buckets: all-women teams, teams with more women than men, teams with roughly equivalent numbers of women and men, teams with more men than women, and all-men teams. In Figure S18 below, dark red, light red, white, light blue and dark blue represent all-women teams, teams with more women than men, teams with roughly equivalent numbers of women and men, teams with more men than women, and all-men teams respectively.

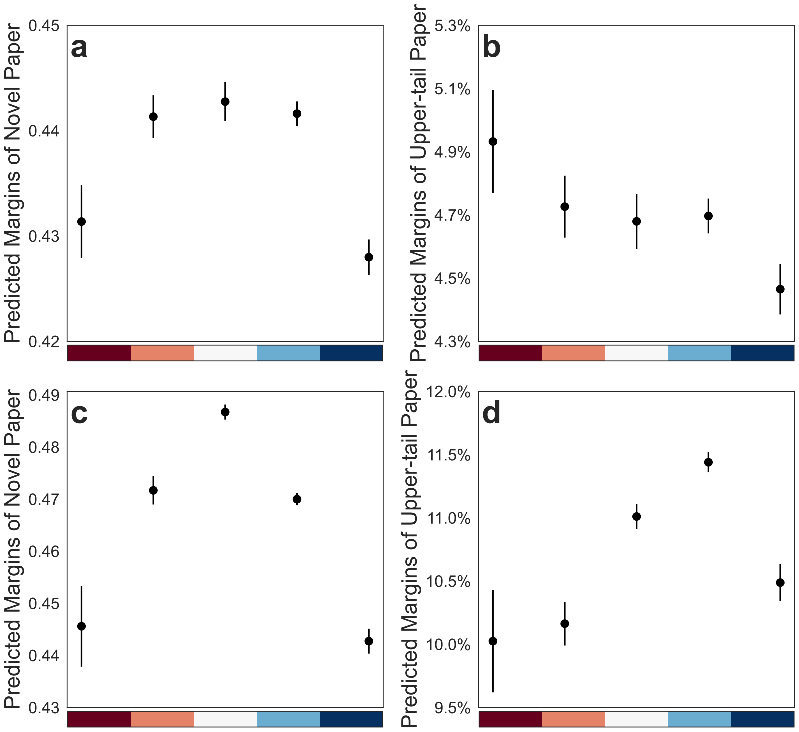


Fig. S18. Margins of Five Categories of Teams in Predicting Novel Paper and Upper-tail Paper. Dark red, light red, white, light blue and dark blue represent all women teams, teams with more women than men, teams with roughly equivalent numbers of women and men, teams with more men than women, and all men teams respectively. Top panel (a & b) indicates results of teams with 2-5 authors. Bottom panel (c & d) represents results of teams with more than 5 authors. (a & c) demonstrates margins of five categories of teams in predicting novel paper when $2 \leq$ team size ≤ 5 and team size ≥ 6 respectively. (b & d) demonstrates margins of five categories of teams in predicting upper-tail paper when $2 \leq$ team size ≤ 5 and team size ≥ 6 respectively.

Figure S17(a) plots the relation between different types of gender composition and novelty. We observe that the effect of gender balance is symmetric when predicting novelty. Teams with roughly equivalent numbers of women and men are significantly more novel than teams with more women than men and teams with more men than women. All-women teams and all-men teams are at similar novelty level. Figure S17(b) plots the relation between different types of gender composition and impact. We can see that all-women teams receive significantly more citations than all-men teams. Among mixed-gender teams, teams that have more men than women receive more citations than teams with more women than men. Consistent with our main results, both Figure S17 demonstrates that mixed-gender teams outperform same-gender teams in novelty and citation.

Furthermore, we divided teams into $2 \leq$ team size ≤ 5 and team size ≥ 6. We have similar observations. First, we observe that the effect of gender balance is symmetric when predicting novelty (see Figure S18 (a & c)). When $2 \leq$ team size ≤ 5, teams with roughly equivalent numbers of women and men are more novel than teams with more women than men and teams with more men than women, but the difference is not statistically significant. All-women teams and all-men teams are at similar novelty level. When team size ≥ 6, teams with roughly equivalent numbers of women and men are significantly more novel than teams with more women than men and teams with more men than women. In conclusion, we observe that having more women or more men in a team does not significantly affect a paper's novelty level.

Second, we investigated the effect of gender balance when predicting impact (see Figure S18 (b & d)). The effect of gender balance is not symmetric when predicting upper-tail paper. In Figure S18(b), we find that all-women teams receive significantly more citations than all-men teams when $2 \leq$ team size ≤ 5. But they do not display significant advantages over mixed-gender teams (teams with more women than men, teams with roughly equivalent number of women and men and teams with more men than women). By contrast, when team size ≥ 6, all-women teams and all-men teams have similar citations, and both receive fewer citations than mixed-gender teams. Among mixed-gender teams, we observe that teams with more men are more likely to receive citations. Given that these analyses support the main results rather than fundamentally change their interpretation

Related to the nuance that we observe symmetric effect of gender balance in novelty but not in citations, we also included a paragraph in the manuscript to describe the connection between our two important dependent variables – novelty and impact. Existing literature (3, 4) confirms that a paper's citations is related to its novelty. However, there are significant conceptual and empirical distinctions. Novelty measure focuses on what are combined, while citation is looking at impact in the community and in science. While novel papers may generate higher citations, citations are driven by things other than just novelty. We also know that novelty poses challenges for acceptance (3, 4). They are not necessarily completely correlated. Given the divergence between novelty and citation, both novelty and citations are independently important, which motivated us to use both novelty and impact to measure a scientific team's performance. What we find in the present study is that they are in the same direction based on mixed-gender teams. Mixed-gender teams are better than same-gender teams whether we evaluate team performance by novelty or citations. These analyses support the main results rather than fundamentally change their interpretation. The nuance observed in Figure S18 bears further analyses to understand the underlying mechanisms (8).

9. Gendered Nature of Citations

In this section, we tried to answer the following questions:

- is the relationship between gender balance and novelty simply due to the increasing citation of women in these collaborations?
- is the link between gender balance and impact explained by gender-based homophily in citations?

To sum up, we find that (1) mixed-gender teams are more novel after controlling for the gender composition of their papers' references; and (2) the citation advantage of mixed-gender teams is not due particularly to citations from women; mixed-gender teams receive more citations from all gendered categories of teams.

Table S6. Relationship between mixed-gender teams and novelty based on journal pairings (model 1-3) /subject pairings (model 4-6) when controlling for the gender composition of references.

	Model (1)	Model (2)	Model (3)	Model (4)	Model (5)	Model (6)
Variable	*novel paper*	*novel paper*	*novel paper*	*novel paper*$_{subjects}$	*novel paper*$_{subjects}$	*novel paper*$_{subjects}$
Mixed-gender Team (m_i)	0.019*** (0.00067)	0.014*** (0.00067)	0.017*** (0.00067)	0.016*** (0.00064)	0.012*** (0.00064)	0.013*** (0.00064)
All Women Ref.	0.021*** (0.0046)			0.18** (0.0042)		
Ref. Entropy	0.19*** (0.0015)			0.21*** (0.0014)		
Women First Author Ref.		0.068*** (0.0022)			0.040*** (0.0020)	
First Author Ref. Entropy		0.16*** (0.0015)			0.14*** (0.0014)	
Women Last Author Ref.			0.025*** (0.0026)			0.060*** (0.0023)
Last Author Ref. Entropy			0.10** (0.0015)			0.12** (0.0015)
Reference Num.	0.052*** (0.00049)	0.056*** (0.00048)	0.063*** (0.00048)	0.11*** (0.00046)	0.12*** (0.00045)	0.12*** (0.00045)
Team Size FE (T_{ti})	Y	Y	Y	Y	Y	Y
Controls†	Y	Y	Y	Y	Y	Y
Observations	4,797,634	4,797,634	4,797,634	4,797,634	4,797,634	4,797,634
R Squared	0.29	0.29	0.29	0.37	0.37	0.37

† The detail of control variables and fixed effects can be found in Section 2.1.
*, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$.

9.1. Gender Composition of References and Novelty. A study of the citation data (14) suggested the existence of gender-based homophily in citations. Such gender-based homophily in citations has also been detected in the work of (15, 16). As both metrics of novelty (Equation 1 and Equation 3) are derived based on the reference list of papers, it is important to interrogate whether the gender composition of references and increased citation towards women could be a factor underlying the higher novelty level of papers written by mixed-gender teams. We took several steps to conduct additional analyses to address this important issue.

To begin with, we detected the gender of 13.9 million authors who are associated with the papers cited by our sample of 6.6 million medical papers, which enabled us to measure the gender composition of the papers backward cited by our analysis sample following the same procedure described in Section 1.2. Next, we defined several variables to measure the gender composition of these papers in the reference list:

- *All Women Ref.* indicates the proportion of papers in the bibliography written by all-women teams.
- *Ref. Entropy* indicates the Shannon Entropy of papers in the bibliography written by all-men, all-women, and mixed-gender teams.
- *Women First Author Ref.* indicates the proportion of papers in the bibliography written by teams with woman first author.
- *First Author Ref. Entropy* indicates the Shannon Entropy of papers in bibliography written by teams with woman first author and man first author.

- *Women Last Author Ref.* indicates the proportion of papers in the bibliography written by teams with woman last author.
- *Last Author Ref. Entropy* indicates the Shannon Entropy of papers in bibliography written by teams with woman last author and man last author.
- *Reference Num.* indicates the total number of references in a paper.

As the results in Table S6 show, we find that the gender composition of references does predict paper novelty (based on journal pairings, see Equation 1). In model (1), *all women ref.*, *ref. entropy* and *reference num.* are all significantly predictive of novel paper. We have similar observations in model (2) and model (3), where *women first author ref.*, *first author ref. entropy*, *women last author ref.* and *last author ref. entropy* are all significantly predictive of novel paper. However, our main finding on the relationship between mixed-gender teams and novelty still hold. Mixed-gender teams are more novel than same-gender teams after controlling for a battery of proxies for the gender composition of references (see Table S6 model (1), (2) and (3)).

Similar conclusions can also be drawn when novelty is measured by the metric based on subject pairings (see Equation 3). Results in model (4), (5) and (6) of Table S6 are consistent with our findings above. This implies that the gender composition of references is unlikely to alter our finding that mixed-gender teams are more novel when measured by novelty based on subject pairings.

9.2. Gender Composition of Citation and Impact. Consistent with the notion of gendered nature of citation, a study of citation data (14) revealed the existence of gender homophily in citations, which has also been detected in the works of (15, 16). Here, we investigate whether gender based homophily in citations can account for our main finding that mixed-gender teams are more impactful than same-gender teams.

To answer this question, we further detected the gender of 16.1 million authors who cited our main analysis sample of 6.6 million medical papers. This step of gender detection enabled us to calculate the number of citations from all men teams, all women teams and mixed-gender teams. Following the same procedure described in Section 1.2, we further normalize a paper i's citations by the corresponding year average to measure a paper's impact. These variables are defined as below:

- $impact_{all\ women}$ indicates normalized citations from all women teams.
- $impact_{all\ men}$ indicates normalized citations from all men teams.
- $impact_{mixed\text{-}gender}$ indicates normalized citations from mixed-gender teams.

We regress these variables on mixed-gender team m_i with a number of controls and fixed-effects (see Section 2.1). The results are presented in Table S7. We observe that when using normalized citations measure, mixed-gender teams are significantly more likely to be cited by all women teams, all men teams and mixed-gender teams. This implies that gender homophily in citations is unlikely to undermine our main finding.

Table S7. Relationship between mixed-gender teams and paper impact within sub-samples of all women citing teams, all men citing teams and mixed-gender citing teams. Mixed-gender teams are significantly more likely to be cited by all women teams, all men teams and mixed-gender teams.

Variable	Model (1) *impact*$impact_{all\ women}$	Model (2) $impact_{all\ men}$	Model (3) $impact_{mixed\text{-}gender}$
Mixed-gender Team (m_i)	0.012*** (0.00077)	0.0021* (0.00083)	0.039*** (0.00071)
Team Size FE (T_{ti})	Y	Y	Y
Controls†	Y	Y	Y
Observations	4,822,724	4,822,724	4,822,724
R Squared	0.41	0.44	0.46

† The detail of control variables and fixed effects can be found in section 2.1.
*, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$.

To test the robustness of this analysis, we also compute the number of citations from teams with women first authors, teams with men first authors, teams with women last authors and teams with men last authors. Similarly, we also classify citations based on the gender of the first authors:

- $impact_{women\ first}$ indicates normalized citations from teams with women first authors.
- $impact_{men\ first}$ indicates normalized citations from teams with men first authors.

As Table S8 shows, we found that mixed-gender teams are more likely to receive citations from both teams with women first authors and teams with men first authors.

Lastly, we also classify citations based on the gender of last authors:

Table S8. Relationship between mixed-gender teams and paper impact within sub-samples of citing teams with women first authors and men first authors. Mixed-gender teams are significantly more likely to be cited by both teams with women first authors and teams with men first authors.

Variable	Model (1) $impact_{women\ first}$	Model (2) $impact_{men\ first}$
Mixed-gender Team (m_i)	0.035*** (0.00069)	0.027*** (0.00075)
Team Size FE (T_{ti})	Y	Y
Controls†	Y	Y
Observations	4,822,724	4,822,724
R Squared	0.47	0.47

† The detail of control variables and fixed effects can be found in section 2.1.
*, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$.

- $impact_{women\ last}$ indicates normalized citations from teams with women last authors.
- $impact_{men\ last}$ indicates normalized citations from teams with men last authors.

Table S9 presents consistent results that mixed-gender teams are more likely to receive citations from both teams with women last authors and teams with men last authors.

Table S9. Relationship between mixed-gender teams and paper impact within subsamples of citing teams with women last authors and men last authors. Mixed-gender teams are significantly more likely to be cited by both teams with women last authors and teams with men last authors.

Variable	Model (1) $impact_{women\ last}$	Model (2) $impact_{men\ last}$
Mixed-gender Team (m_i)	0.031*** (0.00069)	0.030*** (0.00073)
Team Size FE (T_{ti})	Y	Y
Controls†	Y	Y
Observations	4,822,724	4,822,724
R Squared	0.46	0.47

† The detail of control variables and fixed effects can be found in section 2.1.
*, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$.

In conclusion, our analyses suggest that the gendered nature of citations is unlikely to undermine our main findings on the relationship between mixed-gender teams and novelty and impact.

10. Potential Alternative Explanatory Factors

In this section, we engage in a series of analyses that examine several potential pathways through which mixed-gender teams may outperform same-gender teams when it comes to novelty and impact.

Specifically, we investigate the following speculations as to (a) whether mixed-gender teams have access to greater diversity in terms of topic-related expertise; (b) whether mixed-gender teams have favorable network characteristics (i.e., density and size); (c) whether mixed-gender teams benefit from diversity in terms of career age; (d) and whether mixed-gender teams tap into advantages from diversity in terms of geographic diversity. In what follows we describe how we measure these mechanisms in detail.

Table S10. Relationship between team gender composition and mechanism variables. From left to right, we find that mixed-gender team (m_i) is highly correlated with a team's unique expertise, Shannon entropy of team expertise, network density, network range, network size, age diversity and internationalism. The predicted margins are visualized in Figure S19.

Variable	Model (1) Unique Expertise on the Team	Model (2) Shannon Entropy of Team Expertise	Model (3) *Network Density*	Model (4) *Network Range*	Model (5) *Network Size*	Model (6) *Age Diversity*	Model (7) *Internationalism*
Mixed-gender Team (m_i)	0.061*** (0.00065)	0.051*** (0.00057)	-0.015*** (0.00037)	0.090*** (0.00082)	0.091** (0.00077)	0.032** (0.00043)	0.014*** (0.00038)
Team Size FE (T_{ti})	Y	Y	Y	Y	Y	Y	Y
Controls†	Y	Y	Y	Y	Y	Y	Y
Observations	4,822,724	4,822,724	4,822,724	4,822,724	4,822,724	4,822,724	4,822,724
R Squared	0.83	0.85	0.43	0.87	0.88	0.72	0.53

† The detail of control variables and fixed effects can be found in section 2.1.
*, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$.

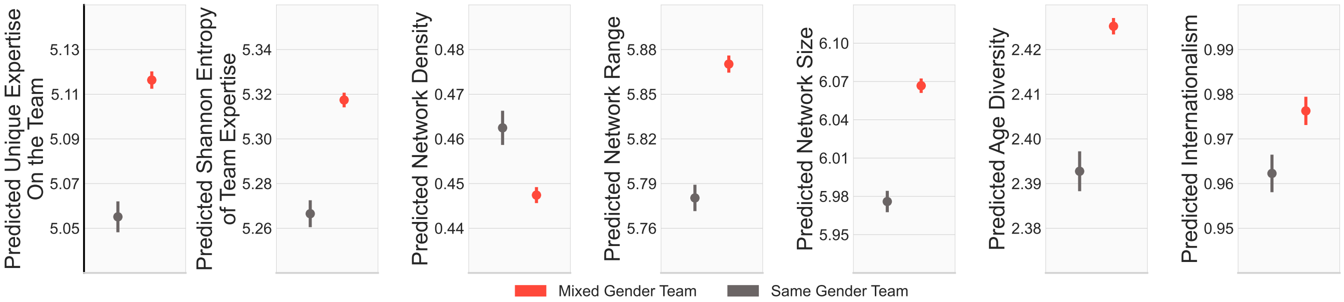


Fig. S19. Margins of Mixed-gender Teams and Same-gender Teams in Predicting Different Mechanism Variables. From left to right, we have unique expertise on the team, Shannon entropy of team expertise, network density, network range, network size, career age diversity and internationalism as dependent variables.

10.1. Team's Exposure to Topic-related Expertise. Here, we examined whether the exposure to different topic-related expertise (17–19) can explain the relationship between gender balance and novelty. To answer this question, we constructed two variables to measure a team's exposure to different topic-related expertise. The topics used in our analysis are tier 2 fields of study recorded in MAG.

For a paper with n authors (a_1, a_2, ..., a_n) published at time t, we can extract an author a_i's publication record and identify the set of topics a_i worked on before time t. Such a set of topics is denoted as τ_{a_i}. Let there be a function μ such that $\tau \in \mu(\tau_{a_i})$ and $\tau \notin \mu(\tau_{a_j})$ for all $j \neq i$, then $\mu(\tau_{a_i})$ is defined as unique topics of a_i. In this way, the number of *unique expertise on the team* is defined as (see Figure S20(a)):

$$unique\ expertise\ on\ the\ team = log|\bigcup \mu(\tau_{a_i})| \qquad [19]$$

Besides access to authors that bring unique expertise individually, we also measured whether the team is exposed to a diverse set of expertise overall, which is measured using the Shannon Entropy. After enumerating τ_{a_i} of all authors, we can calculate the weighted distribution of expertise in the team. If the weighted portion of topic-related expertise k is denoted as p_k, *Shannon entropy of team expertise* takes the form as below (see Figure S20(b)):

$$Shannon\ entropy\ of\ team\ expertise = -\sum_{k=1}^{K} p_k log(p_k) \qquad [20]$$

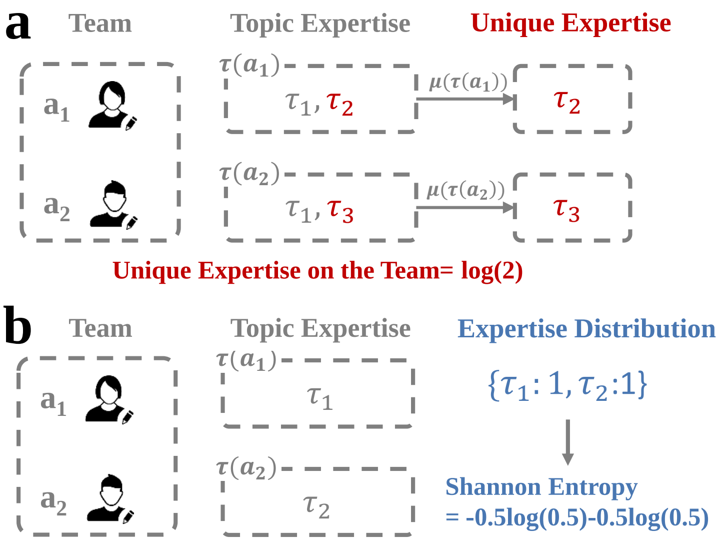


Fig. S20. Illustrative Examples of the Variables *unique expertise on the team* and *Shannon entropy of team expertise.* (a) demonstrates how we calculated the number of unique expertise of a team with two authors a_1 and a_2. For example, a_1 works on topics τ_1 and τ_2 where τ_2 is considered unique topic-related expertise because other authors in the team didn't work on them. Similarly, we know that τ_3 is an unique topic-related expertise brought by a_2. In total, there are two unique topic-related expertise. (b) demonstrates how we computed Shannon entropy of team expertise with two authors a_1 and a_2. We observe that topic τ_1's frequency is 1 and topic τ_2's frequency is also 1. In this way, Shannon entropy of team expertise is $-0.5log(0.5) - 0.5log(0.5)$.

First, we find that mixed-gender teams tend to be exposed to a set of both unique and diverse topic-related expertise (model (1) and (2) in Table S10). As shown in Table S11, we find that the number of unique expertise on the team and Shannon entropy of team expertise are significantly and positively predictive of novel paper (model (1) and (2)). Importantly, our key independent variable – mixed-gender team remains significantly predictive of novel paper across different regression settings. The coefficients of m_i drop about 10% comparing to model (1) in Table S1 where unique expertise on the team and Shannon entropy of team expertise are not included in regression.

Table S11. Relationship between team gender composition and paper novelty/impact when controlling for *unique expertise on the team* and *Shannon entropy of team expertise*. Definitions of *unique expertise on the team* and *Shannon entropy of team expertise* can be found in Equation 19 and Equation 20.

Variable	Model (1) *novel paper*	Model (2) *novel paper*	Model (3) *upper-tail paper*	Model (4) *upper-tail paper*
Mixed-gender Team (m_i)	0.018*** (0.00067)	0.018*** (0.00067)	0.0080*** (0.00038)	0.0082*** (0.00037)
Unique Expertise on the Team	0.0036*** (0.00052)		0.020*** (0.00035)	
Shannon Entropy of Team Expertise		0.0038*** (0.00053)		0.018*** (0.00035)
Team Size FE (T_{ti})	Y	Y	Y	Y
Controls†	Y	Y	Y	Y
Observations	4,822,724	4,822,724	4,822,724	4,822,724
R Squared	0.29	0.29	0.35	0.35

† The detail of control variables and fixed effects can be found in section 2.1. *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$.

Our findings are consistent when predicting upper-tail paper (model (3) and (4) in Table S11). Unique expertise on the team and Shannon entropy of team expertise are also significantly and positively indicative of upper-tail paper (model (3) and (4)). Importantly, mixed-gender team remains significant after controlling for both variables. Similarly, we also observe that the coefficients of m_i drop 11% to 13% compared to model (1) in Table S2 where unique expertise on the team and Shannon entropy of team expertise are not included in regression.

In conclusion, we have several important observations. First, we find that mixed-gender teams are more likely to be exposed to more unique and diverse topic-related expertise. Second, after controlling for unique expertise on the team and Shannon

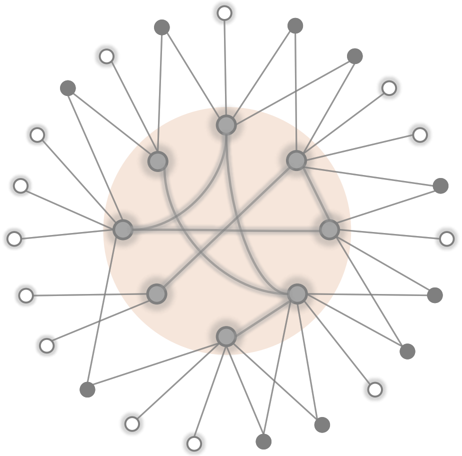


Fig. S21. Scientific Team's Network Density and Network Range. Each node in the graph indicates an individual scientist. The orange area represents a team of scientists, where two scientists are connected if they have collaborated before. The nodes outside the orange area indicate collaborators outside the team. Among them, dark gray nodes indicate scientists who have collaborated with at least two team members, while white nodes indicate scientists who have only collaborated with one team member. In this example, we can see the network density is $\frac{3}{14}$. The network range is $log(13)$. And the network size is $log(23)$.

entropy of team expertise, our main findings persist. Furthermore, the coefficients of mixed-gender team m_i drop slightly (10% to 13%) after controlling for these two variables. The finding implies that a team's exposure to topic-related expertise appears as an orthogonal factor for the relationship between gender balance and team performance.

10.2. Network, Gender Balance and Team Performance. In the work of (20), Reagans and Zuckerman proposed a causal model for diversity-performance relationship in network terms. The model presents pathways that link demographic diversity, the network variables, and team performance. As depicted by the model, increasing demographic diversity reduces internal network density but raises external network range. Both internal network density and external network range are expected to have positive effects on team performance.

To investigate the relationship among gender diversity, networks and team performance in our context, we constructed two variables of network characteristics: *network density* and *network range*.

For a team with n authors (a_1, a_2, ..., a_n) working on a paper at time t, let $e_{ij} = 1$ if a_i and a_j collaborated before time t and let $e_{ij} = 0$ otherwise. In this way, a team's *network density* is defined as below:

$$network\ density = \frac{\sum\sum e_{ij}}{n(n-1)} \qquad [21]$$

We denote a set of scientists who collaborated with a_i before time t as A_i. Let there be a function ϕ such that $\alpha \in \phi(A_i)$ and $\alpha \notin \phi(A_j)$ for all $j \neq i$, then $\phi(A_i)$ is defined as unique collaborators of a_i. In this way, a team's *network range* is defined as:

$$network\ range = log|\bigcup \phi(A_i)| \qquad [22]$$

Meanwhile, we also measure the team *network size* as:

$$network\ size = log|\bigcup A_i| \qquad [23]$$

An illustrative example of these variables is presented in Figure S21. In Figure S21, the orange area represents a team of 8 scientists, where two scientists are connected if they have collaborated before. The nodes outside the orange area indicate collaborators outside the team. Among them, dark gray nodes indicate scientists who have collaborated with at least two team members, while white nodes indicate scientists who have only collaborated with one team member.

Based on the definition of *network density*, there are 28 potential links among 8 scientists where 6 of them are real. Therefore, the *network density* is $\frac{3}{14}$. In total, 8 scientists have 23 collaborators outside the team where 13 of them are unique contacts based on the definition of *network range*. Therefore, *network range* = 13 in our example. Finally, based on the definition, *network size* = 23.

In Table S10, we find that the mixed-gender teams have lower network density (model (3)), which is consistent with existing research (20, 21). In contrast, mixed-gender teams tend to have higher network range (model (4)), which is also well observed in previous studies (20, 21). Finally, in model (5) we observe that mixed-gender teams tend to have larger number of collaborators than same-gender teams.

Table S12. Relationship between team gender composition and paper novelty/impact when controlling for network variables. Definitions of network variables can be found in Equation 21, Equation 22 and Equation 23.

	Model (1)	Model (2)	Model (3)	Model (4)	Model (5)	Model (6)
Variable	*novel paper*	*novel paper*	*novel paper*	*upper-tail paper*	*upper-tail paper*	*upper-tail paper*
Mixed-gender Team (m_i)	0.019*** (0.00067)	0.019*** (0.00067)	0.019*** (0.00067)	0.0086*** (0.00038)	0.0069*** (0.00038)	0.0067*** (0.00038)
Network Density	-0.053*** (0.00095)			-0.034*** (0.00065)		
Network Range		0.015*** (0.00053)			0.025*** (0.00037)	
Network Size			0.013*** (0.00060)			0.027*** (0.00042)
Team Size FE (T_{ti})	Y	Y	Y	Y	Y	Y
Controls†	Y	Y	Y	Y	Y	Y
Observations	4,822,724	4,822,724	4,822,724	4,822,724	4,822,724	4,822,724
R Squared	0.28	0.28	0.28	0.35	0.35	0.35

† The detail of control variables and fixed effects can be found in section 2.1.
*, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$.

To answer the question that whether network characteristics can explain the relationship between gender balance and team performance in our context, we control for *network density*, *network range* and *network size*. We find that mixed-gender team (m_i) is still significantly predictive of novel paper and upper-tail paper. The results are presented in Table S12.

In model (1), (2) and (3) of Table S12, we find that mixed-gender teams are more novel than same-gender teams after controlling for network variables, such as network density, network range and network size. And an interesting observation is that the network range and network size are significantly and positively indicative of novel paper while the network density is negatively indicative of novel paper. This is consistent with existing literature (20, 21). The coefficients of m_i drop about 5% after controlling for network density, network range and network size in regression.

In model (4), (5) and (6) of Table S12, we observe that the network density is also negatively correlated with paper impact. In contrast, network range and network size strongly and positively predict upper-tail paper. Most importantly, we observe that coefficients of m_i in predicting upper-tail paper drop about 6.5%, 25% and 27% after controlling for network density, network range and network size respectively.

In conclusion, while mixed-gender teams and same-gender teams are significantly different in terms of network characteristics, network characteristics are unlikely to fully explain the connection between gender balance and team performance.

10.3. Career Age Diversity and Team Performance. Career age diversity is another important factor that may be relevant to affect team performance in terms of novelty and impact. It is possible that mixed-gender teams are more likely to tap into the potential benefits from career age diversity. Despite increases in women's participation in science in recent years, the overall distribution of career stages may differ by gender as a greater proportion of women are in early careers. Insofar as scientists with diverse career age tend to have access to different information, the diversity of information enriches decision-making within the team and thus enhances performance (22–24). Therefore, team career age diversity is postulated to have a positive effect on team performance.

For a team with n authors (a_1, a_2, ..., a_n) working on a paper at time t, we denote the career age of a_i as γ_i. Following the work of (20, 25), we measure the team age diversity in the form of:

$$age\ diversity = \ log\frac{\sum\sum|\gamma_i\ -\ \gamma_j|}{n(n-1)},\ i\ \neq j \qquad [24]$$

First, based on the results of model (6) in Table S10, we find that mixed-gender teams tend to have higher career age diversity. To investigate whether *age diversity* can explain the relationship between gender balance and team performance, we control for *age diversity* when predicting novel paper and upper-tail paper.

In Table S13, we have several observations. First, we find that age diversity is significantly and positively predictive of novel paper and upper-tail paper. This is consistent with existing research that tenure diversity has a positive effect on team performance (22–24). Second, our main findings are still valid after controlling for tenure diversity. The coefficients of m_i in predicting *novel paper* and *upper-tail paper* drop about 0% and 2.2% respectively after controlling for *age diversity*. This implies that age diversity is an orthogonal factor in the relationship between gender balance and team performance.

10.4. Internationalism of scientific teams. Existing literature suggests that internationally diverse teams are more innovative in different scenarios (26–29). In the context of science, it has been demonstrated that there is a significant citation advantage of

Table S13. Relationship between team gender composition and paper novelty/impact when controlling for age diversity (Equation 24).

Variable	Model (1) *novel paper*	Model (2) *upper-tail paper*
Mixed-gender Team (m_i)	0.020*** (0.00067)	0.0090*** (0.00038)
Age Diversity	0.015*** (0.00062)	0.0040*** (0.00039
Team Size FE (T_{ti})	Y	Y
Controls†	Y	Y
Observations	4,822,724	4,822,724
R Squared	0.28	0.34

† The detail of control variables and fixed effects can be found in section 2.1.
*, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$.

Table S14. Relationship between team gender composition and paper novelty/impact when controlling for internationalism (Equation 25).

Variable	Model (1) *novel paper*	Model (2) *upper-tail paper*
Mixed-gender Team (m_i)	0.020*** (0.00067)	0.0081*** (0.00038)
Internationalism	-0.0024 (0.0017)	0.079*** (0.0012
Team Size FE (T_{ti})	Y	Y
Controls†	Y	Y
Observations	4,822,724	4,822,724
R Squared	0.28	0.35

† The detail of control variables and fixed effects can be found in section 2.1.
*, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$.

having an international presence in research teams (30). To disentangle the link between international diversity and gender diversity, we control for the number of countries involved in the team.

For a paper with n authors (a_1, a_2, ..., and a_n) published at time t, we can extract all affiliated countries with authors a_1, a_2, ..., and a_n, which is denoted as $C = c_1, c_2, ..., c_m$. In this way, internationalism is measured by the logged number of countries associated with a team:

$$\mathit{internationalism} = \mathit{log}(|C|) \qquad [25]$$

First, we find that mixed-gender teams tend to have higher internationalism in Table S10 (model (7)). To disentangle the link between internationalism and gender diversity, we control for the logged number of countries associated with a team. And we find that our main findings still hold after controlling for the number of countries associated with a team (Table S14).

In the Table S14, we find that mixed-gender teams are significantly more likely to produce novel and impactful papers after controlling for *internationalism* (model (1) and (2)). The effect size of our key independent variable mixed-gender Team (m_i) is almost the same as our observations in Table S1 and Table S2 where internationalism is not included as a control variable. Furthermore, we find that *internationalism* is not significantly predictive of a paper's novelty. But *internationalism* is significantly indicative of a paper's impact, which is consistent with findings in the work of (30).

In conclusion, this implies that the relationship between gender balance and team performance cannot be fully explained by higher level of internationalism of mixed-gender teams.

References

1. E Carsenat, Inferring gender from names in any region, language, or alphabet. *Unpublished* **10** (2019).
2. L Santamaría, H Mihaljević, Comparison and benchmark of name-to-gender inference services. *PeerJ Comput. Sci.* **4**, e156 (2018).
3. B Uzzi, S Mukherjee, M Stringer, B Jones, Atypical combinations and scientific impact. *Science* **342**, 468–472 (2013).
4. E Leahey, CM Beckman, TL Stanko, Prominent but less productive: The impact of interdisciplinarity on scientists' research. *Adm. Sci. Q.* **62**, 105–139 (2017).
5. A Porter, A Cohen, J David Roessner, M Perreault, Measuring researcher interdisciplinarity. *Scientometrics* **72**, 117–147 (2007).
6. Y Yang, NV Chawla, B Uzzi, A network's gender composition and communication pattern predict women's leadership success. *Proc. Natl. Acad. Sci.* **116**, 2033–2038 (2019).
7. PM Blau, *Inequality and heterogeneity: A primitive theory of social structure.* (Free Press New York) Vol. 7, (1977).
8. B Hofstra, et al., The diversity–innovation paradox in science. *Proc. Natl. Acad. Sci.* **117**, 9284–9291 (2020).
9. MM King, CT Bergstrom, SJ Correll, J Jacquet, JD West, Men set their own cites high: Gender and self-citation across fields and over time. *Socius* **3**, 2378023117738903 (2017).
10. DS Chawla, Men cite themselves more than women do. *Nature* **535**, 212 (2016).
11. S Mishra, BD Fegley, J Diesner, VI Torvik, Self-citation is the hallmark of productive authors, of any gender. *PloS one* **13**, e0195773 (2018).
12. AC Pinho-Gomes, et al., Where are the women? gender inequalities in covid-19 research authorship. *BMJ Glob. Heal.* **5**, e002922 (2020).
13. MW Nielsen, JP Andersen, L Schiebinger, JW Schneider, One and a half million medical papers reveal a link between author gender and attention to gender and sex analysis. *Nat. human behaviour* **1**, 791–796 (2017).
14. G Ghiasi, P Mongeon, C Sugimoto, V Larivière, Gender homophily in citations in *STI 2018 Conference Proceedings.* (Centre for Science and Technology Studies (CWTS)), pp. 1519–1525 (2018).
15. S Knobloch-Westerwick, CJ Glynn, The matilda effect—role congruity effects on scholarly communication: A citation analysis of communication research and journal of communication articles. *Commun. Res.* **40**, 3–26 (2013).
16. M Potthoff, F Zimmermann, Is there a gender-based fragmentation of communication science? an investigation of the reasons for the apparent gender homophily in citations. *Scientometrics* **112**, 1047–1063 (2017).
17. A Taylor, HR Greve, Superman or the fantastic four? knowledge combination and experience in innovative teams. *Acad. management journal* **49**, 723–740 (2006).
18. S Aral, M Van Alstyne, The diversity-bandwidth trade-off. *Am. journal sociology* **117**, 90–171 (2011).
19. NW Kohn, SM Smith, Collaborative fixation: Effects of others' ideas on brainstorming. *Appl. Cogn. Psychol.* **25**, 359–371 (2011).
20. R Reagans, EW Zuckerman, Networks, diversity, and productivity: The social capital of corporate r&d teams. *Organ. science* **12**, 502–517 (2001).
21. H Ibarra, Race, opportunity, and diversity of social circles in managerial networks. *Acad. management journal* **38**, 673–703 (1995).
22. DG Ancona, DF Caldwell, Demography and design: Predictors of new product team performance. *Organ. science* **3**, 321–341 (1992).
23. DC Hambrick, TS Cho, MJ Chen, The influence of top management team heterogeneity on firms' competitive moves. *Adm. science quarterly* **41**, 659–684 (1996).
24. HA Ndofor, DG Sirmon, X He, Firm resources, competitive actions and performance: investigating a mediated model with evidence from the in-vitro diagnostics industry. *Strateg. Manag. J.* **32**, 640–657 (2011).
25. MG Kendall, A Stuart, JK Ord, *Kendall's advanced theory of statistics.* (Oxford University Press, Inc.), (1987).
26. J Hunt, M Gauthier-Loiselle, How much does immigration boost innovation? *Am. Econ. Journal: Macroecon.* **2**, 31–56 (2010).
27. M Nathan, N Lee, Cultural diversity, innovation, and entrepreneurship: Firm-level evidence from l ondon. *Econ. geography* **89**, 367–394 (2013).
28. DM Hart, What do foreign-born founders bring to entrepreneurial teams? an exploration in the us high-tech sector. *An Explor. US High-Tech Sect. (July 28, 2010). GMU Sch. Public Policy Res. Pap.* **6** (2010).
29. JE Perry-Smith, CE Shalley, A social composition view of team creativity: The role of member nationality-heterogeneous ties outside of the team. *Organ. Sci.* **25**, 1434–1452 (2014).
30. D Hsiehchen, M Espinoza, A Hsieh, Multinational teams and diseconomies of scale in collaborative research. *Sci. advances* **1**, e1500211 (2015).