PNAS RESEARCH ARTICLE | SOCIAL SCIENCES OPEN ACCESS

Gender-diverse teams produce more novel and higher-impact scientific ideas

Yang Yang[a,b,c], Tanya Y. Tian[d], Teresa K. Woodruff[e], Benjamin F. Jones[f,g], and Brian Uzzi[b,f,h,1]

Edited by Susan Fiske, Princeton University, Princeton, NJ; received January 16, 2022; accepted July 24, 2022

Science's changing demographics raise new questions about research team diversity and research outcomes. We study mixed-gender research teams, examining 6.6 million papers published across the medical sciences since 2000 and establishing several core findings. First, the fraction of publications by mixed-gender teams has grown rapidly, yet mixed-gender teams continue to be underrepresented compared to the expectations of a null model. Second, despite their underrepresentation, the publications of mixed-gender teams are substantially more novel and impactful than the publications of same-gender teams of equivalent size. Third, the greater the gender balance on a team, the better the team scores on these performance measures. Fourth, these patterns generalize across medical subfields. Finally, the novelty and impact advantages seen with mixed-gender teams persist when considering numerous controls and potential related features, including fixed effects for the individual researchers, team structures, and network positioning, suggesting that a team's gender balance is an underrecognized yet powerful correlate of novel and impactful scientific discoveries.

team science | gender inequality | innovation | computational social science

Significance

Science teams made up of men and women produce papers that are more novel and highly cited than those of all-men or all-women teams. These performance advantages increase the greater the team's gender balance and appear nearly universal. On average, they hold for small and large teams, the 45 subfields of medicine, and women- or men-led teams and generalize to published papers in all science fields over the last 20 y. Notwithstanding these benefits, gender-diverse teams remain underrepresented in science when compared to what is expected if the teams in the data had been formed without regard to gender. These findings reveal potentially new gender and teamwork synergies that correlate with scientific discoveries and inform diversity, equity, and inclusion (DEI) initiatives.

Medical research is undergoing two transformations that are potentially remaking practice and research impact. First is the increased participation of women in medical science (1–3). The last decade in medical science has seen women's participation rates exceed men's participation rates in graduate and postdoctoral research training in 60 to 40% and 54 to 46% ratios, respectively (4). Second is the shift from individual to team science (5). Rising teamwork levels are broadly documented across scientific fields and show that teams are associated with more novel combinations of prior work (6) and higher citation impact (5).

The rise in women's participation in research and the rise in teamwork suggest that research formats may be evolving toward gender-diverse collaborations, potentially opening new pathways to inclusivity in science and opening new opportunities for medical research. Clues regarding impact come from a handful of laboratory experiments that have found that teams mixing women and men are better at general problem-solving tasks than all-women or all-men teams with equivalent IQ levels (7). Nevertheless, in professional settings gender dynamics can produce a propensity for same-gender teams (8–12) that potentially reduces workgroup diversity and fairness (13–16). In addition, the degree to which laboratory studies reliably generalize to practice and policy in real-world settings is unclear (17–19). Given the potential yet unknown implications of the rise in teamwork and women's participation in medical research, we conducted a large-scale study of how changing gender demographics, team creation (20, 21), and team performance (22, 23) are reshaping medical science research practice and impact using an original dataset of over 6.6 million medical research papers published in more than 15,000 journals over the last 20 y. A full description of our data and methods appears in *Materials and Methods* and *SI Appendix*, section 1.1.

Gender-Diverse Teams Have Proliferated but Remain Underrepresented

Fig. 1*A* plots the upward trend of women's participation in medical science from 2000 to 2019, while Fig. 1*B* shows the share of different team sizes per year (Fig. 1 *B*, *Inset* shows the average team size over time) for our full sample of 6.6 million medical papers. The major change in team size is that large teams have supplanted small teams, which is consistent with existing literature (5, 24). In 2000, papers with one or two authors accounted for 15 and 16% of publications, respectively, but by 2015 their shares dropped to only 8 and 12%, while the large teams (more than six authors) increased their share from 25 to 46% of papers. Large teams (more than six authors) now dominate the production of knowledge in medicine and the trend continues upward.

As teams rise in share and size, there is more opportunity for researchers of different genders to collaborate (25). To quantify the changing incidence of gender-diverse teams,

Author contributions: Y.Y., T.Y.T., B.F.J., and B.U. designed research; Y.Y., T.Y.T., T.K.W., B.F.J., and B.U. performed research; Y.Y. analyzed data; and Y.Y., T.Y.T., B.F.J., and B.U. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: uzzi@northwestern.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2200841119/-/DCSupplemental.

Published August 29, 2022.

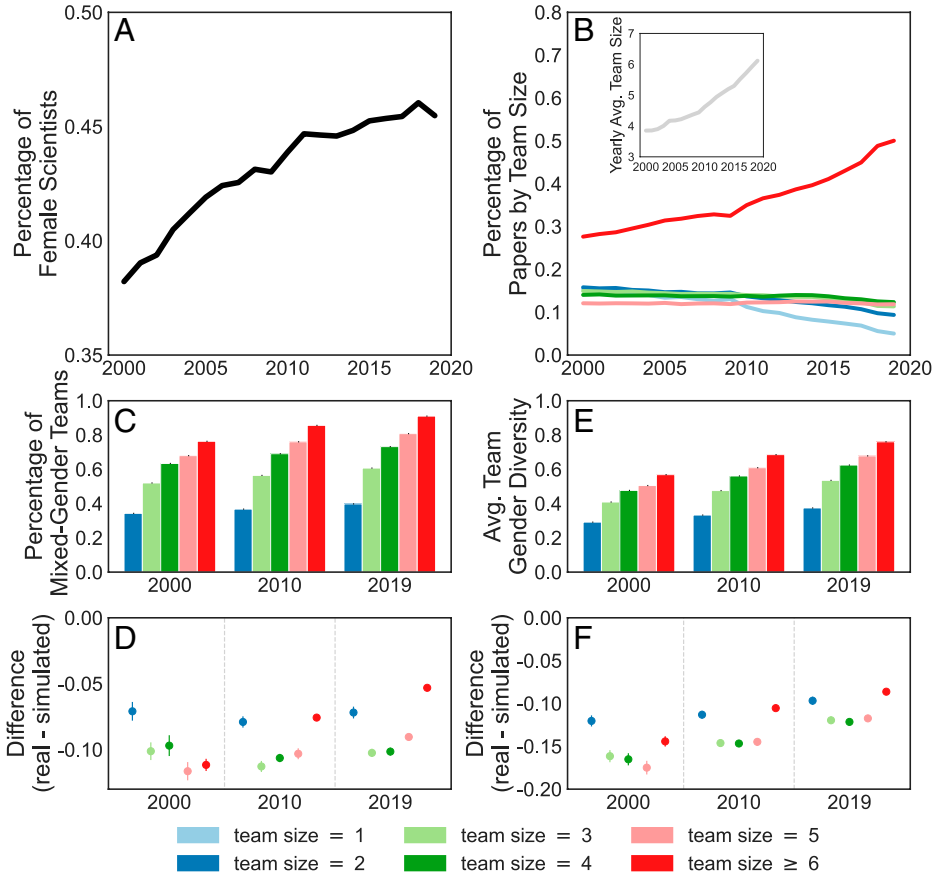

Fig. 1. The underrepresentation of gender-diverse teams. (*A*) The sharp upward trend of women's participation in medical research over the last 20 y. (*B*) The share of different team sizes by year over the same period. The major change is that large teams of six or more researchers per paper have supplanted teams of smaller sizes with the largest drop in solo authorship and two-person teams. *C* and *D* show that the share of publications from mixed-gender teams has steadily increased with time. Nonetheless, *E* and *F* indicate that mixed-gender teams are significantly underrepresented in medical science (no CIs cross the 0.00 difference line) by up to 17% depending on the team size when we measure team gender diversity using the Shannon entropy (*SI Appendix*, section 1.2.5).

we measured team gender composition in two ways. First, "mixed gender" is a binary variable equal to 0 if the team is all the same gender and 1 if the team has any combination of women and men. Second, "gender diversity" is a continuous variable that ranges from 0 when the team is made up of 100% women or men to 1 when the team includes equal percentages of women and men.

Following prior work (26–28), we use a validated name-to-gender inference method (29, 30) to compute a probability estimate of an author's gender based on the author's first name and last name. This algorithmic approach has the advantage of being comparable to most past measurements of gender in the literature (9, 28, 31). Also, it enables discipline-wide analyses of gender that would be unattainable with manual coding (32). In our case, the algorithm's estimate of the percentage of women and men authors in our study agrees with studies that use self-reported gender information. The 2019 and 2020 Association of American Medical Colleges (AAMC) census data (4) indicated that women comprised 42.7 and 43.2%, respectively, of faculty, which aligns with our estimate that women comprise 42.3% of authors in our data. Another study focused on the gender of 6,722 authors of 1,370 COVID-19–related articles found that that women comprised 34% of all authors (28). By comparison, 35.7% of 60,839 authors of 9,033 COVID-19–related papers in our data are estimated to be women. At the same time, our method has certain limitations. First, the algorithmic estimation approach is limited to binary gender classification (33). Second, although the method is widely used and validated (30) and the gender classification estimates of authors in our study agree with the fraction of scientists who self-report being a woman or a man, it is still possible that the gender of some individual authors is misclassified. *SI Appendix*, section 3 attempts to address the impact of these challenges on the reliability of results by estimating how different classification errors would alter our main findings.

Fig. 1*C* indicates the share of mixed-gender teams, by team size, and how this has evolved over time using our measures (*SI Appendix*, section 1.2.5). In 2000, about 60% of teams of size four included both men and women; by 2019, the percentage was about 70%. Fig. 1*D* similarly shows upward trends using a continuous rather than a binary measure of team gender composition (*SI Appendix*, section 1.2.5).

These results leave open the question of whether the increase in gender-diverse teams is more or less than expected if teams were assembled without regard to gender. To undertake this analysis, we designed a null model to estimate the expected rate of gender-diverse teams. The model holds constant the number of papers, distribution of team size, and number of men and women authors in each year. It then randomly interchanges women and men authors who have the same first year of publication, total publications, and country (see *SI Appendix*, section 6.1 for details).

Fig. 1 *E* and *F* shows the relative levels of underrepresentation of teams that included both women and men researchers according to our two measures of team gender diversity (*SI Appendix*, section 1.2.5). Fig. 1 *E* and *F* shows that gender-diverse teams are significantly underrepresented at every team size, with as much as 17% underrepresentation depending on the team's size (means with 95% CIs shown, all P values <0.01).

Further, the plots show that substantial underrepresentation remains persistent over the past 20 y despite the rise in teamwork, the rise in the participation of women in science, and policies of inclusiveness.

Gender-Diverse Teams Are More Novel and Impactful

We next examine novelty and impact outcomes, studying a scientific team's output from different perspectives. Novelty concerns the degree to which a paper combines past knowledge in a new way, while impact concerns the degree to which a paper influences future work. A paper's impact and novelty can be positively related but are also empirically distinct and can be driven by separate factors (6, 34), suggesting the necessity of studying both measures when characterizing a scientific team's output.

Novelty Measures. To measure a paper's novelty, we followed prior literature and denote novel papers as those that combine knowledge in a new way relative to existing combinations. Because novelty is a broad concept, we used two different measures of a paper's novelty from the literature (6, 34). The novelty measure used in our main results is based on ref. 6, which uses the journals referenced in a paper and examines whether given journal pairings are common or unusual. Specifically, the novelty measure quantifies the observed co-occurrence frequencies of all journal pairings in the literature prior to the publication year of the target paper. The observed co-occurrence of journal pairings in each paper is compared to a null model of what the pairing frequency would be if the journal pairings were combined by chance. Papers whose bibliographies contain journal pairings that have frequently occurred together in the past (more than expected by chance) indicate relatively conventional and familiar pairings of knowledge; by contrast, papers whose bibliographies contain journal pairings that happen less than expected by chance indicate novel combinations of knowledge (6).

Our second novelty measure is the Sterling index (34), which uses the subject categories of the papers cited in a paper's bibliography to compute a paper's novelty. In our study, we used 291 subject categories as defined by the Microsoft Academic Graph (MAG) (35). Papers whose bibliographies contain subject pairings that have been frequently cocited in the past represent conventional pairings of knowledge, while papers whose bibliographies contain subject pairings that have been rarely cocited represent novel combinations of knowledge. The measure ranges from 0 to 1, where high values indicate higher novelty. *Materials and Methods* and *SI Appendix*, sections 1.2.2, 1.2.3, and 2.3 compare the two novelty measures and show that the results are robust to either measure.

Impact Measures. To measure a paper's impact, we followed prior literature and denoted high-impact papers as those in the top 5% of citations for papers published in a given year (*SI Appendix*, section 1.2.4) (36). A continuous measure of impact, which is defined as a paper's citations normalized by the publication year average, produced confirmatory findings (*SI Appendix*, section 1.2.4).

Novelty and Impact Results. We ran fixed-effects regressions to investigate a paper's novelty and citation impact conditional on the team's gender diversity and team size. We further control for numerous author, journal, and institutional characteristics (see *Materials and Methods* and *SI Appendix*, section 2.1 for further information about regression methods). In *SI Appendix*, Tables S1 and S2 present regression details for team gender diversity when it is measured as a binary and a continuous variable.

Fig. 2*A* presents regression results for mixed-gender teams and indicates that mixed-gender teams publish significantly more novel papers than same-gender teams (two-sample t test, P value <0.001). For example, large (more than authors) mixed-gender teams are 9.1% more likely to publish a novel paper than the base rate [(48 to 44%)/44%]. Given that novelty positively correlates with team size (5, 6, 37), the substantial added explanatory power of mixed-gender teams vs. same-gender teams holding team size constant is striking. Proportionally, the increase in novelty for a mixed-gender team of six or more authors relative to same-gender teams is equivalent to the increase in novelty obtained by doubling a same-gender team size from two to four.

Fig. 2*B* shows that mixed-gender teams publish significantly more highly cited papers than same-gender teams. The impact advantage of mixed-gender teams appears at all team sizes. Focusing on large teams (more than six authors), mixed-gender teams are 14.6% [(10.2 to 8.9%)/8.9%) = 14.6%] more likely to publish an upper-tail paper than same-gender teams of equal size. Measuring team gender diversity as a continuous variable confirmed the above results and indicated that as gender-diverse teams approach a 50/50 split of women and men, the association between gender-diverse teams and novelty and impact increases (*SI Appendix*, Tables S1 and S2).

We further examined how team performance varies with the gender of the team's first and last author (38). Consistent with our main finding, mixed-gender teams, irrespective of the leader's gender, produced significantly more novel and highly cited research than same-gender teams ($P < 0.001$; *SI Appendix*, Fig. S16). Comparing women-led mixed-gender teams to men-led mixed-gender teams, we found that women-led mixed-gender teams have greater novelty but fewer citations than men-led mixed-gender teams (*SI Appendix*, section 7).

Generalizability across Subfields in Medicine. Subfields in medical research have important differences in the gender distribution of active researchers, research questions addressed, research funding, age and gender of team leaders and mentors, and other features (39–41). To examine how the findings might vary across subfields in medicine, we conducted several tests. We grouped papers in one of 45 MAG-designated primary medical subfields and ran a separate regression for each subfield using the same fixed-effects regression specification as above.

Fig. 2 *C* and *D* shows that the findings strongly generalize across subfields in medicine. The y axis shows the regression coefficient value with 95% CIs when novelty (Fig. 2*C*) and citation impact (Fig. 2*D*) are regressed on the team's gender diversity for 45 separate subfields in medical science (x axis). The 45 subfields are sorted from largest to smallest in terms of total number of publications (*SI Appendix*, Table S5 provides names of each subfield), which naturally leads to more precise coefficient estimates for the larger subfields to the left and less precise estimates for the smaller subfields to the right. Fig. 2 *C* and *D* demonstrates that the team's gender diversity significantly and positively predicts a team's novelty and impact for most subfields with the smaller subfields exhibiting noisy relationships (*SI Appendix*, section 4). In Fig. 2*C*, we can see that 43 of 45 subfields (representing 99% of papers in medicine) have positive coefficients when predicting novelty. The coefficients are both positive and significant in 29 subfields (representing 85% of papers in medicine). In Fig. 2*D*, we observe that 30 of 45 subfields (representing 83% of papers in medicine) have positive coefficients when predicting papers' impact. The coefficients are both positive and significant in

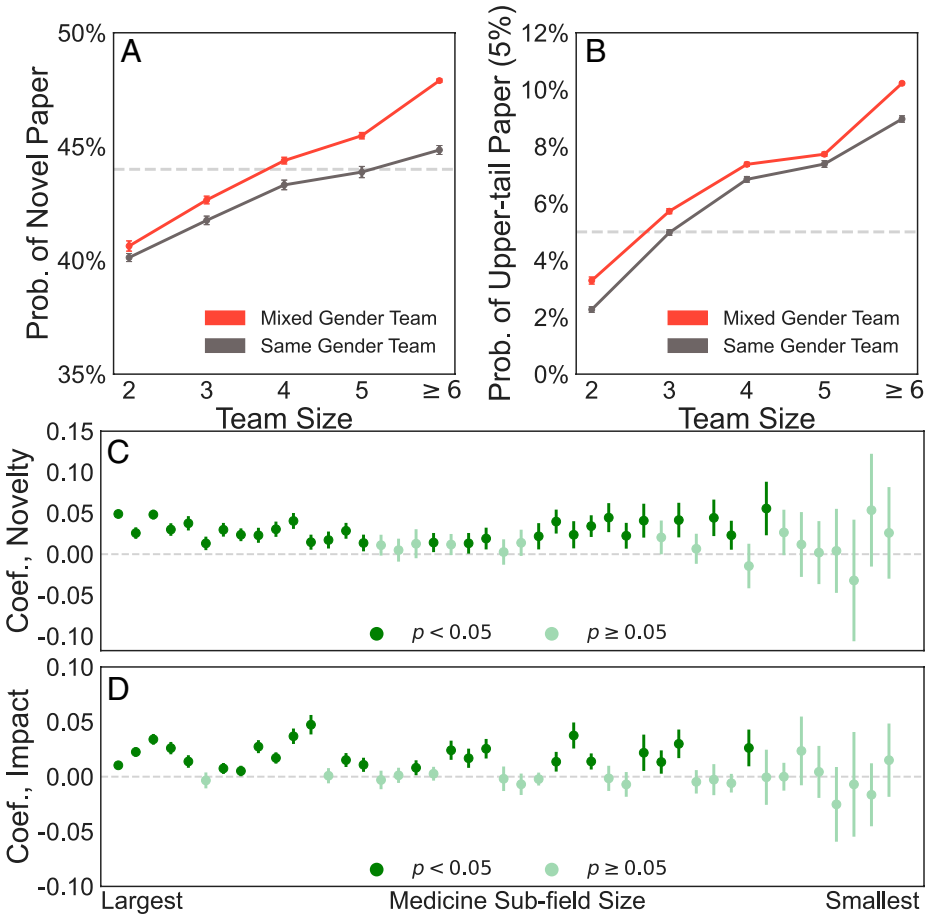

Fig. 2. Mixed-gender teams produce more novel and highly cited research publications. (*A*) Mixed-gender teams are more likely to produce novel papers than same-gender teams at all team sizes. Teams of six or more authors are 9.1% more likely to publish novel work than the base rate. (*B*) Mixed-gender teams are more likely to publish an upper-tail paper than same-gender teams by as much as 14.6%, depending on team sizes. (*C* and *D*) The performance benefits of gender diversity generalize across 45 medical subfields. Medical subfields are arranged from left to right according to the number of publications in the subfields (largest to smallest) with statistically significant relationships for novelty and impact becoming noisier for smaller medical subfields. Dark-green coloring indicates significant coefficients (*P* value <0.05), while light-green coloring indicates nonsignificant coefficients.

24 subfields (representing 75% of papers in medicine). Detailed results of generalizability across medical subfields can be found in *SI Appendix*, section 4.

Potential Alternative Explanatory Factors

Our final analysis investigates possible factors behind the new empirical regularities that we have identified. Prior work on team science and gender (18, 42–45) has explained team performance as a function of expertise (18, 25, 45–47), demographics (48, 49), and network characteristics (47, 50–52) that enhance a team's access to problem-solving and promotional resources (45, 53) (see *SI Appendix*, section 10 for a summary of this literature). To examine these possibilities, we further consider measures that quantify the diversity of expertise across team members; the network size, range, and density of team members; and the career age and international diversity of team members. The measurements of these factors follow past literature and are described in *Materials and Methods* and *SI Appendix*, section 10.

We first examined whether mixed-gender teams have different expertise, network, age diversity, and international diversity characteristics compared to same-gender teams. We found that mixed-gender teams are associated with significantly 1) higher topic-related expertise diversity (measured by unique expertise on the team and Shannon entropy of team expertise; *Materials and Methods*); 2) lower network density, larger network range, and larger network size; 3) higher career age diversity; and 4) higher geographic diversity and internationalism (*SI Appendix*, Table S11 and Fig. S19).

Given these differences, we further consider regression models that predict team performance based on the gender diversity of the team while also controlling for these potential related factors. Fig. 3 summarizes the regression results when accounting for these features separately or together. We find that the size of the coefficient for mixed-gender teams tends to decrease, yet the substantially and statistically significant main effects of mixed-gender teams on novelty and impact hold across all specifications. These findings taken together indicate that mixed-gender teams correlate with expertise, network, and demographic drivers of team success, which may inform the performance advantages seen among mixed-gender teams. Yet the performance advantages cannot be fully explained by those drivers, suggesting that a team's gender balance is an underrecognized yet powerful correlate of novel and impactful scientific discoveries that increases in magnitude with the gender balance of the team.

Finally, we examined whether the gender homophily in citation behavior may explain our findings. Research indicates that citations show homophily—men cite papers by men more than papers by women and vice versa (53, 54). Importantly, we find that mixed-gender teams receive more citations than same-gender teams regardless of the citing team—from all-women, all-men, and mixed-gender citing teams alike (*SI Appendix*, section 9).

Fig. 3. Mixed-gender teams and research outcomes controlling for numerous factors. (*A* and *B*) The regression coefficient and 95% CIs for mixed-gender teams in predicting novelty (*A*) and citation impact (*B*) while controlling for the features indicated in the panel headings. The leftmost panels indicate the coefficients of mixed-gender team in baseline regressions. The rightmost panels indicate the coefficients of mixed-gender team after controlling collectively for expertise diversity, network structure, career age diversity, and international diversity.

Discussion

Conducting an analysis of 6.6 million published papers from more than 15,000 different medical journals worldwide, we find that mixed-gender teams—teams combining women and men scientists—produce more novel and more highly cited papers than all-women or all-men teams. Mixed-gender teams publish papers that are up to 7% more novel and 14.6% more likely to be upper-tail papers than papers published by same-gender teams, results that are robust to numerous institutional, team, and individual controls and further generalize by subfield. Finally, in exploring gender in science through the lens of teamwork, the results point to a potentially transformative approach for thinking about and capturing the value of gender diversity in science (1, 3, 22).

Another key finding of this work is that mixed-gender teams are significantly underrepresented compared to what would be expected by chance. This underrepresentation is all the more striking given the findings that gender-diverse teams produce more novel and high-impact research and suggests that gender-diverse teams may have substantial untapped potential for medical research. Nevertheless, the underrepresentation of gender-diverse teams may reflect research showing that women receive less credit for their successes than do men teammates (14, 55), which in turn inhibits the formation of gender-diverse teams and women's success in receiving grants (56), prizes (3), and promotions (22).

We are conservative in speculating on the theoretical mechanisms at work given the noncausal nature of our study, but the richer descriptive findings in such a large-scale dataset are informative. We found that the association between gender diversity and team performance both supports and challenges current thinking. Examining causal and noncausal factors reported in the literature, we find that mixed-gender teams are correlated with expertise, network, career age, and international features of a team. Nevertheless, we found that the strong positive association between mixed-gender teams and team success can only partly be explained by expertise, network, career age, or international measures, suggesting that a rich set of factors, including factors as yet unveiled in the literature, may be at work in the research advantages of mixed-gender teams.

Thus, future research should examine the causal mechanisms that might explain why gender-diverse teams outperform same-gender teams (52, 57) and how those mechanisms translate into actionable practices and policy (24). Laboratory experiments suggest that women on a team improve information-sharing processes on teams, such as turn taking (7). It might also be that women provide a perspective on research questions that men do not possess and vice versa or it may be that when a team has both women and men teammates, there are synergies specific to gender-diverse teams that are more than the additivity of team processes and information typically associated with all-women and all-men teams (45).

Beyond experimental studies of the gender balance phenomena, future research should examine the phenomena in other areas of science (14, 58, 59). In this research we focused on medicine, a large branch of science and among the most funded with its wide array of downstream applications (3, 60). In addition, the advantages of gender-diverse teams in other branches of science were preliminarily studied. The results indicate that the novelty and impact benefits of gender-diverse teams generalize on average to published papers in all science fields over the last 20 y (details in *SI Appendix*, section 5). Future research should further examine the role of gender-diverse teams in STEM (science, technology, engineering, and mathematics) fields more broadly to sharpen insights on practices that promote innovation, inclusiveness, and equality.

Similarly, teamwork characterizes work practices within most for-profit and nonprofit organizations, which also stand to gain from a better understanding of the link between gender balance and team performance and diversity, equity, and inclusion (DEI) initiatives. For example, COVID-19 has fundamentally changed interpersonal interactions in organizations. In-person meetings have been substantially supplanted by virtual meetings, which are weaker at creative ideation (61). While it is beyond the scope of this study to examine how the phenomenon of team gender diversity relates to in-person vs. virtual teamwork, it would be valuable to investigate whether the creativity loss of virtual meetings relative to in-person meetings can be partially mitigated by gender balance on teams. One important upshot of these

future research questions suggests that team gender balance is a phenomenon at the center of a confluence of changes in the composition of the workforce, innovation, fairness, and inclusion. In this sense, our findings provide a different lens on potential gender and teamwork synergies that correlate with the rate of scientific discoveries and inform opportunities in DEI initiatives.

Materials and Methods

Data and methods used in our work are described below.

Data Samples. We conduct a large-scale systemic investigation of the performance of gender-diverse research teams in the medical sciences. Our field-wide dataset has over 6.6 million research publications by 3.2 million women and 4.4 million men scientists in more than 15,000 medical science journals from 2000 to 2019 as recorded in Microsoft Academic Graph (35). MAG is a scientific publication database that records journal articles' bibliographic information (title, journal, journal field, volume, issue, page, publication date), authorship (name), author affiliations (name, webpage, and wikipage), and citation links to other papers in the database.

Name-to-Gender Inference Based on First and Last Names. The main procedure for estimating a scientist's gender uses the Namsor software (29, 30), which has been developed and examined for multiple languages (e.g., Chinese, English, French, Spanish, etc.) (30). The algorithm was run on the first and last names of all authors as they appear in the MAG database. This procedure implements a binary gender system to remain consistent with most existing gender studies in science (28, 30), which is not designed to address the important issue of nonbinary gender distinctions in the data. The analysis suggests the method applied in our sample has face validity. As noted above, the algorithm's estimate of the gender distribution of women and men authors in our study is within a few percentage points of the percentages of women and men estimated from the self-reported gender information disclosed by women and men researchers and reported by the Association of American Medical Colleges (4) and an independent study of more than 6,000 researchers (28). Further, to empirically alleviate concerns of misclassification, we conduct several robustness checks to quantify how the results vary given different levels of assumed misclassification of gender (*SI Appendix*, Tables S3 and S4). Finally, our gender detection results are demonstrated to be consistent with the Gender API (application programming interface) (62) (another widely used name-to-gender method; *SI Appendix*, section 3.2). These results are explained in *SI Appendix*, section 3.

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

Team gender diversity. To measure the gender composition in a scientific team, we use a binary variable, mixed-gender team, in the main analysis. A mixed-gender team has both men and women. Otherwise, it is a same-gender team. Additionally, we use a continuous variable to evaluate the gender composition of a scientific team (9) that takes the form

$$g_i = -p_f log_2\,(p_f) - (1-p_f) log_2\,(1-p_f)\,, \qquad [1]$$

where p_f indicates the portion of women authors in a team i. The value of g_i ranges from 0 to 1. When the value of g_i is low, either women or men are the 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.

Novelty measures. We use two measures of novelty to characterize the mixture of knowledge in a given paper (6, 34). The main 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. To compute a paper's novelty, we use a z score. The z score is computed from 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. When a z score is less than zero, the combination of prior work is considered novel (i.e., a novel pairing of ideas), and when the z score is above zero, the combination of prior work is considered conventional (i.e., a common pairing of ideas). A journal pairing is defined as novel if its novelty z score is smaller than zero.

Our second measure of the novelty of scientific papers (34) focuses on subject (a.k.a. topic) pairings cited within a paper's reference list. The measure considers papers with rare combinations of subjects to be novel because they create new combinations of knowledge in terms of subjects that have been infrequently cocited together in previous research. The 244 Web of Science (WoS) subject categories are used in the work of ref. 34. Similarly, we use the 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 cocited in prior research. We denote by S_{ij} the cosine similarity between subject i and subject j in the subject cocitation matrix. Formally, 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 (further details are provided in *SI Appendix*, sections 1.2.2 and 1.2.3):

$$novelty_{subjects} = 1 - \sum_{ij} S_{ij} p_i p_j. \qquad [2]$$

Impact measures. The MAG database tracks a paper's annual citations received. We define high-impact papers as those in the top 5% of citations for papers published in the same year. The variable is denoted as an upper-tail paper if it is a top 5% highly cited paper. We alternatively measure impact as a ratio to normalizing a paper i's final citations by the publication year average, which is denoted as $\widehat{c_i}$ (see *SI Appendix*, section 1.2.3 for details). For robustness tests, we run a similar regression by substituting the binary upper-tail paper variable with this continuous one. Given that $\widehat{c_i}$ follows a heavy-tail distribution, we use a log transform of $\widehat{c_i}$ to measure a scientific paper's impact:

$$impact = \log(\widehat{c_i} + 1). \qquad [3]$$

Regression Analysis. To quantify the link between team gender diversity and performance, we used fixed-effects regressions. The regressions control for confounds due to the authorship, team size, leadership, institutional prestige rank, year, journal quality, prior citation impact at the time of publication, average team career age, and individual fixed effects (5, 6, 18, 36, 49) (see *SI Appendix*, section 2 for details). Alternative measurements and null models confirm the results (see *SI Appendix*, sections 2 and 3 for details), and to test the generalizability of our findings, we ran separate regressions for each of the 45 medical subfields (see *SI Appendix*, section 4 and Table S5 for details).

We further 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. Accordingly, we investigate the following measures:

Topic-related expertise. 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, unique expertise on the team is defined as

$$unique\ expertise\ on\ the\ team = log|\bigcup \mu(\tau_{a_i})|. \qquad [4]$$

After we quantified the unique expertise of individuals, 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 topics in the team. If the weighted portion of topic k is denoted as p_k, the team Shannon entropy of team expertise takes the following form:

$$Shannon\ entropy\ of\ team\ expertise = -\sum_{k=1}^{K} p_k log(p_k). \qquad [5]$$

Detailed information can be found in *SI Appendix*, section 10.1.

Network characteristics. 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

$$network\ density = \frac{\sum\sum e_{ij}}{n(n-1)}. \qquad [6]$$

Further, 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

$$network\ range = log|\bigcup \phi(A_i)|. \quad [7]$$

We also measure the team network size as

$$network\ size = log|\bigcup A_i|. \quad [8]$$

An illustrative example of these variables is presented in *SI Appendix*, section 10.2.

Career age diversity. 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 refs. 52 and 63, we measure the team age diversity in the form of

$$age\ diversity = \log \frac{\sum\sum |\gamma_i - \gamma_j|}{n(n-1)},\ i \neq j. \quad [9]$$

Internationalism. 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 $C = c_1, c_2, \ldots, c_m$. In this way, internationalism is measured by the logged number of countries associated with a team:

$$internationalism = log(|C|). \quad [10]$$

Data, Materials, and Software Availability. Previously published data were used for this work (https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/).

ACKNOWLEDGMENTS. This study is supported by the Air Force Office of Scientific Research under Minerva Award FA9550-19-1-0354, the Northwestern Alumnae Grant, and the Northwestern University Institute on Complex Systems and Data Science. The access to NamSor API is funded by a research budget grant when Y.Y. was at Syracuse University. We wish to thank the participants at the Science of Science Funding National Bureau of Economic Research Summer Institute meeting, the 2022 International Conference on the Science of Science & Innovation, and the two anonymous reviewers and Susan Fiske for their excellent comments.

Author affiliations: [a]Mendoza College of Business, University of Notre Dame, Notre Dame, IN 46556; [b]Northwestern Institute on Complex Systems and Data Science, Northwestern University, Evanston, IL 60208; [c]Lucy Family Institute for Data and Society, University of Notre Dame, Notre Dame, IN 46556; [d]New York University Shanghai, New York University, Shanghai, China; [e]Department of Obstetrics, Gynecology, and Reproductive Biology, Michigan State University, East Lansing, MN 48824; [f]Kellogg School of Management, Northwestern University, Evanston, IL 60208; [g]National Bureau of Economic Research, Cambridge, MA 02138; and [h]The McCormick School of Engineering, Northwestern University, Evanston, IL 60208

1. H. Etkowitz, C. Kemelgor, B. Uzzi, *Athena Unbound: The Advancement of Women in Science and Technology* (Cambridge University Press, Cambridge, UK, 2000).
2. L. M. Ataman, Y. Ma, F. E. Duncan, B. Uzzi, T. K. Woodruff, Quantifying the growth of oncofertility. *Biol. Reprod.* **99**, 263–265 (2018).
3. Y. Ma, D. F. Oliveira, T. K. Woodruff, B. Uzzi, Women who win prizes get less money and prestige. *Nature* **565**, 287–288 (2019).
4. Association of American Medical Colleges, U.S. medical school faculty trends: Percentages. https://www.aamc.org/data-reports/faculty-institutions/interactive-data/2020-us-medical-school-faculty. Accessed 1 January 2022.
5. S. Wuchty, B. F. Jones, B. Uzzi, The increasing dominance of teams in production of knowledge. *Science* **316**, 1036–1039 (2007).
6. B. Uzzi, S. Mukherjee, M. Stringer, B. Jones, Atypical combinations and scientific impact. *Science* **342**, 468–472 (2013).
7. A. W. Woolley, C. F. Chabris, A. Pentland, N. Hashmi, T. W. Malone, Evidence for a collective intelligence factor in the performance of human groups. *Science* **330**, 686–688 (2010).
8. E. H. Gorman, Gender stereotypes, same-gender preferences, and organizational variation in the hiring of women: Evidence from law firms. *Am. Sociol. Rev.* **70**, 702–728 (2005).
9. Y. Yang, N. V. Chawla, B. Uzzi, A network's gender composition and communication pattern predict women's leadership success. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 2033–2038 (2019).
10. R. DeCastro, D. Sambuco, P. A. Ubel, A. Stewart, R. Jagsi, Mentor networks in academic medicine: Moving beyond a dyadic conception of mentoring for junior faculty researchers. *Acad. Med.* **88**, 488–496 (2013).
11. P. Fontanarosa, H. Bauchner, A. Flanagin, Authorship and team science. *JAMA* **318**, 2433–2437 (2017).
12. E. H. Chang, E. L. Kirgios, R. K. Smith, Large-scale field experiment shows null effects of team demographic diversity on outsiders' willingness to support the team. *J. Exp. Soc. Psychol.* **94**, 104099 (2021).
13. L. A. Rivera, Hiring as cultural matching: The case of elite professional service firms. *Am. Sociol. Rev.* **77**, 999–1022 (2012).
14. B. Macaluso, V. Larivière, T. Sugimoto, C. R. Sugimoto, Is science built on the shoulders of women? A study of gender differences in contributorship. *Acad. Med.* **91**, 1136–1142 (2016).
15. H. Sarsons, Recognition for group work: Gender differences in academia. *Am. Econ. Rev.* **107**, 141–145 (2017).
16. B. K. AlShebli, T. Rahwan, W. L. Woon, The preeminence of ethnic diversity in scientific collaboration. *Nat. Commun.* **9**, 5163 (2018).
17. M. W. Nielsen *et al.*, Opinion: Gender diversity leads to better science. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 1740–1742 (2017).
18. K Börner *et al.*, A multi-level systems perspective for the science of team science. *Sci. Transl. Med.* **2**, 49cm24 (2010).
19. A. W. Woolley, R. M. Chow, A. T. Mayo, C. Riedl, J. W. Chang, Collective attention and collective intelligence: The role of hierarchy and team gender composition. *Organ. Sci.*, 10.1287/orsc.2022.1602 (2022).
20. J. S. Long, Measures of sex differences in scientific productivity. *Soc. Forces* **71**, 159–178 (1992).
21. J. B. Bear, A. W. Woolley, The role of gender in team collaboration and performance. *Interdiscip. Sci. Rev.* **36**, 146–153 (2011).
22. H. Etzkowitz, C. Kemelgor, M. Neuschatz, B. Uzzi, J. Alonzo, The paradox of critical mass for women in science. *Science* **266**, 51–54 (1994).
23. L. G. Campbell, S. Mehtani, M. E. Dozier, J. Rinehart, Gender-heterogeneous working groups produce higher quality science. *PLoS One* **8**, e79147 (2013).
24. S. Fortunato *et al.*, Science of science. *Science* **359**, eaao0185 (2018).
25. R. Guimerà, B. Uzzi, J. Spiro, L. A. N. Amaral, Team assembly mechanisms determine collaboration network structure and team performance. *Science* **308**, 697–702 (2005).
26. J. D. West, J. Jacquet, M. M. King, S. J. Correll, C. T. Bergstrom, The role of gender in scholarly authorship. *PLoS One* **8**, e66212 (2013).
27. V. Larivière, C. Ni, Y. Gingras, B. Cronin, C. R. Sugimoto, Bibliometrics: Global gender disparities in science. *Nature* **504**, 211–213 (2013).
28. A. C. Pinho-Gomes *et al.*, Where are the women? Gender inequalities in COVID-19 research authorship. *BMJ Glob. Health* **5**, e002922 (2020).
29. NamSor API, Namsor: Name checker for gender, origin and ethnicity determination. https://github.com/namsor/namsor-api. Accessed 30 October 2020.
30. L. Santamaría, H. Mihaljević, Comparison and benchmark of name-to-gender inference services. *PeerJ Comput. Sci.* **4**, e156 (2018).
31. G. Shannon *et al.*, Gender equality in science, medicine, and global health: Where are we at and why does it matter? *Lancet* **393**, 560–569 (2019).
32. N. Caplar, S. Tacchella, S. Birrer, Quantitative evaluation of gender bias in astronomical publications from citation counts. *Nat. Astron.* **1**, 1–5 (2017).
33. C. Richards *et al.*, Non-binary or genderqueer genders. *Int. Rev. Psychiatry* **28**, 95–102 (2016).
34. E. Leahey, C. M. Beckman, T. L. Stanko, Prominent but less productive: The impact of interdisciplinarity on scientists' research. *Adm. Sci. Q.* **62**, 105–139 (2017).
35. K. Wang *et al.*, A review of Microsoft academic services for science of science studies. *Front Big Data* **2**, 45 (2019).
36. B. F. Jones, S. Wuchty, B. Uzzi, Multi-university research teams: Shifting impact, geography, and stratification in science. *Science* **322**, 1259–1262 (2008).
37. K. J. Boudreau, E. C. Guinan, K. R. Lakhani, C. Riedl, Looking across and looking beyond the knowledge frontier: Intellectual distance, novelty, and resource allocation in science. *Manage. Sci.* **62**, 2765–2783 (2016).
38. M. Chowdhary *et al.*, Women's representation in leadership positions in academic medical oncology, radiation oncology, and surgical oncology programs. *JAMA Netw. Open* **3**, e200708–e200708 (2020).
39. S. Mukherjee, D. M. Romero, B. Jones, B. Uzzi, The nearly universal link between the age of past knowledge and tomorrow's breakthroughs in science and technology: The hotspot. *Sci. Adv.* **3**, e1601315 (2017).
40. B. F. Jones, The burden of knowledge and the "death of the renaissance man": Is innovation getting harder? *Rev. Econ. Stud.* **76**, 283–317 (2009).
41. Y. Ma, S. Mukherjee, B. Uzzi, Mentorship and protégé success in STEM fields. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 14077–14083 (2020).
42. S. K. Horwitz, I. B. Horwitz, The effects of team diversity on team outcomes: A meta-analytic review of team demography. *J. Manage.* **33**, 987–1015 (2007).
43. B. Stvilia *et al.*, Composition of scientific teams and publication productivity at a national science lab. *J. Am. Soc. Inf. Sci. Technol.* **62**, 270–283 (2011).
44. A. W. Woolley, M. E. Gerbasi, C. F. Chabris, S. M. Kosslyn, J. R. Hackman, Bringing in the experts: How team composition and collaborative planning jointly shape analytic effectiveness. *Small Group Res.* **39**, 352–371 (2008).
45. S. E. Page, *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies* (Princeton University Press, 2007).
46. National Research Council, *Enhancing the Effectiveness of Team Science* (National Academies Press, 2015).
47. B. Uzzi, A social network's changing statistical properties and the quality of human innovation. *J. Phys. A Math. Theor.* **41**, 224023 (2008).
48. R. Reagans, E. Zuckerman, B. McEvily, How to make the team: Social networks vs. demography as criteria for designing effective teams. *Adm. Sci. Q.* **49**, 101–133 (2004).
49. B. Jones, E. Reedy, B. A. Weinberg, Age and scientific genius. https://www.nber.org/papers/w19866. Accessed 5 August 2022.
50. B. Uzzi, J. Spiro, Collaboration and creativity: The small world problem. *Am. J. Sociol.* **111**, 447–504 (2005).
51. M. A. Marks, L. A. DeChurch, J. E. Mathieu, F. J. Panzer, A. Alonso, Teamwork in multiteam systems. *J. Appl. Psychol.* **90**, 964–971 (2005).
52. R. Reagans, E. W. Zuckerman, Networks, diversity, and productivity: The social capital of corporate R&D teams. *Organ. Sci.* **12**, 502–517 (2001).
53. M. L. Dion, J. L. Sumner, S. M. Mitchell, Gendered citation patterns across political science and social science methodology fields. *Polit. Anal.* **26**, 312–327 (2018).

54. 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).
55. M. B. Ross *et al.*, Women are credited less in science than are men. *Nature* **608**, 135–145 (2022).
56. D. F. M. Oliveira, Y. Ma, T. K. Woodruff, B. Uzzi, Comparison of national institutes of health grant amounts to first-time male and female principal investigators. *JAMA* **321**, 898–900 (2019).
57. S. Harvey, A different perspective: The multiple effects of deep level diversity on group creativity. *J. Exp. Soc. Psychol.* **49**, 822–832 (2013).
58. M. E. Heilman, M. C. Haynes, No credit where credit is due: Attributional rationalization of women's success in male-female teams. *J. Appl. Psychol.* **90**, 905–916 (2005).
59. M. Ahmadpoor, B. F. Jones, Decoding team and individual impact in science and invention. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 13885–13890 (2019).
60. Y. Yin, Y. Dong, K. Wang, D. Wang, B. Jones, Public use and public funding of science. *Nat. Hum. Behav.*, 10.1038/s41562-022-01397-5 (2022).
61. E. Á. Horvát, B. Uzzi, Virtual collaboration hinders a key component of creativity. *Nature* **605**, 38–39 (2022).
62. Gender API, Gender API - Determines the gender of a first name. https://gender-api.com. Accessed 30 April 2022.
63. M. G. Kendall, A. Stuart, J. K. Ord, *Kendall's Advanced Theory of Statistics* (Oxford University Press, Inc., 1987).