NBER WORKING PAPER SERIES

CASSATTS IN THE ATTIC

Marlène Koffi
Matt Marx

Working Paper 31316
http://www.nber.org/papers/w31316

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2023

We thank Karin Hoisl, Fiona Murray, Xavier Jaravel, John Walsh, and participants at the African Econometric Society, the NBER Productivity Seminar, the Munich Summer Institute, the International Conference on Computational Social Science, the NBER Summer Institute, the Asia Innovation and Entrepreneurship Conference, and the Banff Empirical Microeconomics Workshop, seminars series of the University of Munich, Simon Fraser University and Dalhousie University for insightful feedback. We also thank WonJung Joey Ryu for excellent research assistance. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Cassatts in the Attic
Marlène Koffi and Matt Marx
NBER Working Paper No. 31316
June 2023
JEL No. J16,O31

ABSTRACT

We analyze more than 70 million scientific articles to characterize the gender dynamics of commercializing science. The double-digit gender gap we report is explained neither by the quality of the science nor its ex-ante commercial potential, and is widest among papers with female last authors (i.e., lab heads) when publishing high-quality science. Using Pitchbook database, we show that when authors self-commercialize scientific discoveries via new ventures, no gap appears, raising the question of whether incumbent firms are unaware of—or ignore—scientific contributions by women. A natural experiment based on the Obama administration's staggered introduction of open-access requirements for federally-funded research reveals that although easier access to scientific articles might facilitate commercialization, this benefit accrues primarily to male authors. Articles written with more "boastful" language are commercialized more often, and female scientists generally boast less, but even when they do their discoveries are commercialized no more often. We also observe gender homophily between scientific authors and commercializing inventors, the majority of whom are male. We conclude with the potential welfare effects of the gender gap: the disparity is more pronounced for higher-quality discoveries, as indicated by academic and patent citations or by predicted probabilities of commercialization derived from deep-learning algorithms.

Marlène Koffi
Department of Economics
University of Toronto
150 St. George Street
Toronto, ON M5S 3G7
and NBER
marlene.koffi@utoronto.ca

Matt Marx
Cornell University
137 Reservoir Avenue
Ithaca, NY 14853
and NBER
mmarx@cornell.edu

1 Introduction

Among the more remarkable transformations in science during the past half-century has been the increase in participation by women. As shown in Figure 1, whereas in 1980 barely one in five published papers included a female author, now more than 50% of papers include at least one woman on the authorship team. In fact, the percentage of papers in 2020 with a *majority* of female authors exceeds the percentage in 1980 with even a single woman. Broader gender diversity in scientific discovery can contribute to social welfare because the process of research and development itself is gendered (Koning, Samila, and Ferguson 2021). Indeed, prior work has shown that scientific teams combining male and female scientists can lead to discoveries that are more highly recognized within the academic community (Campbell et al. 2013; Yang et al. 2021).

Even if published, however, the full impact of a scientific discovery on society requires that it be brought to market or *commercialized*. As one example, the breakthrough drug erythropoietin (EPO), which treats anemia by stimulating the production of red blood cells, was purified in a lab at the University of Chicago in 1971 and subsequently tested in rats, with a flurry of academic papers published by Eugene Goldwasser and collaborators in the mid-1970s: (Goldwasser, Kung, and Eliason 1974; Miyake, Kung, and Goldwasser 1977; Sherwood and Goldwasser 1979). But commercialization did not occur until the late 1980s, when the biotech startup Applied Molecular Genetics (now Amgen) patented the production of a recombinant formulation later marketed as EPOGEN.[1]

Rivette and Kline 2000 coined the term *Rembrandts in the Attic*[2] to describe discoveries like EPO, which remain uncommercialized—for years, or even permanently. Given that in 2019 alone, the U.S. federal government invested $43.9 billion in basic research, there is a substantial societal interest in avoiding scientific discoveries from becoming trapped in the ivory tower. Nearly 200 academic articles have investigated the process of commercializing scientific discoveries (see Rothaermel, Agung, and Jiang 2007 for a review), not to mention countless task forces. In 2022, the National Science Foundation launched its first new directorate in 32 years: TIP (Technology, Innovation, and Partnerships), with the aim of "translation of research results to the market and society."[3]

The rise of women in the production of science, and the frequent failure of scientific discoveries to be commercialized, highlight and reinforce a concern raised frequently both in scholarly and policy circles: although women have made substantial strides in the production of science, they might nonetheless appear to be less involved in the commercialization of those same scientific discoveries. The implications of this potential gender gap in commercialization are troubling, including missed opportunities for career advancement and wealth creation.

1. https://news.uchicago.edu/story/eugene-goldwasser-biochemist-behind-blockbuster-anemia-drug-1922-2010.
2. Rivette and Kline 2000 use the phrase to refer to firms' failure to capitalize on technologies they developed but did not bring to market. Although they did not directly consider universities, we believe the notion of missed commercialization opportunities is also relevant to the academic sector.
3. https://beta.nsf.gov/tip/latest

Although many programs have arisen to boost women's participation in commercialization, such as REACH for Commercialization[4] and STEM to Market,[5] we lack large-scale data regarding the gender dynamics of commercialization. Most studies of commercialization have focused on patenting behavior and, moreover within a single field—most often, medicine or life sciences (Ding, Murray, and Stuart 2006; Fechner and Shapanka 2018; Thursby and Thursby 2005; Whittington and Smith-Doerr 2005). Alternatively, scholars have conducted surveys and interviews (Murray and Graham 2007) to characterize commercialization, which have tended to focus on choices by *individuals* whereas the role of teams in the production of science is ever-increasing (Wuchty, Jones, and Uzzi 2007). Noting that even the title "Rembrandts in the Attic" is gendered, the authors of the book having selected a male painter, our goal in this paper is to characterize the gender dynamics of scientific commercialization in the full canon of scientific inquiry.

We analyze commercialization among all published articles reported by the Microsoft Academic Graph (MAG) from 1800-2020. Automatically classifying the gender of the authors via forenames, and hand-coding gender for a substantial subsample, we explore the gender dynamics involved in the commercialization of these scientific discoveries. Commercialization is captured via a "patent-paper pair" (Ducor 2000) where (a) the patent cites the paper (b) the patent assignee is a firm (c) where the inventors on the patent overlap with the authors of the paper.

We find a double-digit gender gap in commercialization for papers with a woman in the last-author position on the article. Frequently, the last author is the "lab head" or Principal Investigator on the project. The head of the lab may play a critical role in commercialization; therefore, our analyses focus on the gender of the lab head. Because the gender gap in commercialization is larger among discoveries that are more highly cited and with higher commercial potential, we refer to these uncommercialized discoveries as *Cassatts in the Attic* after Pittsburgh-born painter and printmaker Mary Cassatt.

There are several potential explanations for this gap. First, we check that our results are not driven by the selection of less-commercializable projects by women. This is essential because, consistent with prior findings in specific fields (Yang et al. 2021), we find that in the full population of academic articles, those involving women are more highly cited by academic articles. Therefore, it might be that women focus on basic research as opposed to applied projects with greater commercial potential. We employ the method of Bikard 2020, identifying "twin" scientific articles that report the same scientific discovery (scaling his method up to the entirety of MAG) in order to control for unobserved commercial potential. Moreover, we show that this "twin" strategy is not dependent on adjacent self-citation. Indeed, we employ a similar methodology to that of Hill and Stein 2019 to identify pairs of papers with unique biological structures and sequences. Specifically, we retrieve all proteins archived in the Protein Data Bank (PDB) as of June 2022. We find that the gender gap in commercialization

4. https://ccts.osu.edu/content/reach-commercialization-inspiring-female-entrepreneurship
5. https://www.awis.org/stem-to-market/

remains salient when using this new approach.

We consider both supply-side and demand-side mechanisms underlying these gender dynamics. Given that prior work has suggested that lack of access to networks, as well as under-representation of women in science, may explain the reluctance to engage in venturing "beyond the lab" (Murray and Graham 2007), we explore the possible role of such factors in commercialization. We find some support for the notion that women are more likely to commercialize in scientific fields where a higher share of scientists are female. As well, we confirm that, consistent with past literature, women generally have limited commercial networks compared to men. However, we find no evidence to support the notion that the lack of women's commercial networks explains for the gender gap in commercialization.

Not finding strong evidence for supply-side factors, we separate our dependent variable into *cooperative* commercialization (i.e., involving incumbent firms) vs. *self*-commercialization via new ventures. To do so, we use the commercial dataset on startups provided by PitchBook. If supply-side factors played a major role, we would expect to see lower levels of self-commercialization for women, but we do not. Rather, the gender gap is exclusive to commercialization that occurs in cooperation with existing firms. We thus shift our focus to demand-side factors that might explain why firms appear less eager to commercialize scientific discoveries by women.

We begin our demand-side investigation by leveraging a natural experiment regarding Federal open-access (OA) mandates to ascertain whether the simple availability of science affects whether firms cooperatively commercialize scientific discoveries by women. The staggered-by-agency implementation of OA requirements for federally-funded research during the Obama administration facilitates a causal interpretation of having easier access to published scientific articles. Although we note that open access might drive commercialization generally for both men and women, we observe a noticeable deepening of the gender gap as articles become more accessible. Under the identification assumption, this is suggestive of a bias from the firm side.

Moreover, even when articles are available, firms' commercialization efforts are drawn to some over others due to simple word choice.[6] We extend the Lerchenmueller, Sorenson, and Jena 2019 study to the universe of academic publications showing that papers with women are less likely to use self-promoting language when publishing their scientific discoveries and that this is true even among twin articles that report the same scientific discovery. Although academics do not differentially cite papers that tout themselves as a "breakthrough" or "unprecedented", this sort of word choice appears to attract the attention of firms. Articles where a woman is the final author (often if not usually the Principal Investigator) use "boastful" language less often. Moreover, only teams with male lab heads appear to benefit from boasting when it comes to commercialization. We further show that homophily might be one of the sources of bias on the firms' side that could partly explain the gender gap in

6. See also Kolev, Fuentes-Medel, and Murray 2019 on the importance of words choice and gender in research grant proposals.

commercialization.

Our evidence concludes by underscoring the potential welfare effects of the gender gap in commercialization. We use academic citations and patent citations as indicators of higher-quality discoveries. In order to capture both quality and commercial potential beyond citations, we develop a machine learning algorithm that incorporates the latest advancements in natural language processing to predict the probability of commercialization. Our findings reveal that the gender gap in commercialization is more pronounced among higher-quality discoveries, as evidenced by academic citations, patent citations, and the predicted probability of commercialization. Therefore it truly appears that the gender gap leads to key scientific discoveries (i.e., "Cassatts") remaining in the ivory tower (i.e., in the "attic of science"). The gap, moreover, widened from the 1980s through the early 2000s but appears to have shrunk recently.

1.1 Related Literature

Our work is related to two broad literatures. First, nearly 200 articles characterizing the technology-transfer process whereby university scientific discoveries are licensed or commercialized have been written (see Rothaermel, Agung, and Jiang 2007 for an overview). Second, a smaller but growing literature observes gender bias in the production and recognition of science (Ding, Murray, and Stuart 2006; Fechner and Shapanka 2018; Murray and Graham 2007; Tartari and Salter 2015; Thursby and Thursby 2005; Whittington and Smith-Doerr 2005). Recent work has also branched out to consider the allocation of credit in the social sciences (Koffi 2021; Sarsons 2017).

One article deserves particular mention. Bikard and Fernandez-Mateo 2022 also use the twin-articles methodology to examine whether academic papers have an impact beyond collecting citations from other academic papers. There are several key differences between their work and ours. First is the dependent variable, where they use the patent-to-paper citation datasets of Marx and Fuegi (2020, 2021) to measure impact, whereas we examine only patent paper *pairs*—i.e., those patents citing papers where the inventors overlap with the authors and the patent assignee was a firm. Thus their work is more about cumulative innovation whereas we focus on commercialization. Second, whereas they focus on a subset of 295 academic publications and hand-code the gender of the authors in their sample, we explore the universe of academic research with 70,016,266 articles while relying both on forename-classification algorithms and hand-coded classification of last authors for 27,551 academic publications and all authors for 1,982.

The remainder of the paper is organized as follows. Section 2 describes the data. Section 3 presents the empirical strategy and the results, including robustness and placebo tests. Section 4 discusses potential mechanisms. Section 5 investigates the impact of the gender gap over time and according to the quality of the science. Finally, section 6 concludes.

2 Data and Variable Construction

We conduct our analysis using the Microsoft Academic Graph (MAG) (Sinha et al. 2015), which contains bibliometric metadata on nearly 190 million academic articles, conference proceedings, and working papers. Other sources of such metadata are available, such as the Clarivate Web of Science and Elsevier Scopus, we select MAG for three reasons. First, it is an open database, which permits replication and cumulative development of our results. (PubMed is also open but is limited to the life sciences whereas we seek to analyze the entire scientific canon.) Second, its coverage has been found to outstrip proprietary databases, particularly with respect to conference proceedings (Hug and Brändle 2017).[7] Third, and most importantly for our analysis, MAG seems to do particularly well at capturing *given* names (hereafter, "forenames" of authors. As we rely on the algorithmic determination of gender, forenames are critical to our analysis.

MAG captures metadata regarding the year of publication, the journal name, as well as the authors of the article and their affiliations. It also reports article-to-article citations. We use these fields to calculate our dependent variable *Citations from Scientific Articles*, which we limit to a fixed five-year window from the publication of the article. We also use these data to calculate the *Journal Impact Factor* (not included in MAG) as well as the number of citations for each author and institution on each paper; each of these is averaged to create a series of control variables: *Average citations per author*, *Average citations per institution.*

Finally, unlike the Clarivate Web of Science, MAG does not provide high-level categorizations of scientific fields. Instead, Microsoft automatically extracts more than 200,000 keywords from the abstracts and titles of the papers themselves. We mapped the MAG subjects to 6 OECD fields and 39 subfields.[8] Clarivate provides a public crosswalk between the OECD classifications and the 251 Web of Science scientific "categories", so we probabilistically mapped the MAG keywords to WoS categories. These probabilistic mappings are used as fixed effects in our cross-sectional models and for subsetting to examine field-specific findings.

Some of our analyses count citations to articles not from other articles but from patents. These are obtained from Marx and Fuegi 2020 for "front page" citations and Marx and Fuegi 2021 for "in-text" citations. Patent-to-article citations are also critical to identifying patent-paper pairs and, therefore, commercialization.

7. That said, one critique of MAG is that by collecting preprints and working papers from the web via the Bing web crawler, it includes many incomplete/duplicate drafts, inflating the count of papers. Therefore we restrict our analysis to entries in MAG that contain Digital Object Identifiers, reducing the number of academic articles from 190 million to approximately 95 million.

8. These are defined at http://www.oecd.org/science/inno/38235147.pdf.

2.1 Identifying the commercialization of scientific discoveries

What do we mean by "commercializing science", and how do we operationalize this construct? Must commercialization include that a product was actually brought to market, or does the attempt alone suffice, even if it fails? Is the simple referencing or acknowledgment of science by a commercial effort enough? And how can such activity be captured at scale?

One measure of commercialization could be lists of technology licenses issued by universities. Even if a product does not launch, a license might be written in anticipation of such an effort. The Association for University Technology Managers conducts surveys of university licensing activity, but these data are counts and cannot be mapped to the recipient of the license. Gathering licensing data from individual universities is another alternative and has been pursued by various authors (e.g., Lach and Schankerman 2008), but this approach has two limitations. But such data are difficult to collect at scale as no central repository of licensing data exists, and thus individual arrangements must be negotiated with each university.

An alternative to licensing data is using citations from patents to academic articles. Arguably, if a (commercial) patent references an academic article, it is relying upon the underlying science in a commercially-meaningful way. But citations can be made as background material or can be done strategically (Lampe 2012). Moreover, given the more than 40 million patent-to-paper citations, can all of these truly be said to be instances of commercializing science? It may be more useful to view citations from patents to articles as a superset of actual commercializations.

2.1.1 Capturing commercialization with patent-paper pairs

Our measure of commercialization relies on "patent-paper pairs" (PPPs). The concept of a PPP originates with Ducor 2000, who detected overlapping gene sequences between papers and patents as evidence of the same material being published and patented. Since Ducor, more than two dozen articles have employed the PPP method (see for example Murray 2002, Thompson, Ziedonis, and Mowery 2018, Ranaei et al. 2016, and Haeussler and Sauermann 2013). The intuition is that if we find a patent citing a paper *where the author(s) of the paper were also inventors on the patent*, we might reasonably assume that there was some act of commercialization involved.

We generalize the PPP algorithm of Marx and Hsu 2021 to the entire MAG. As we are focused not only on genetics, we cannot rely on genetic-sequence overlap; instead, we use techniques common to other PPP studies including overlap of authors and inventors as well as temporal similarity. Given the finding of Haeussler and Sauermann 2013 that papers often have "extra" authors who do not appear on a patent, and also being mindful of the critique of Ranaei et al. 2016 that misspellings and commonality can lead to both "lumping" and "splitting" problems in matching names, we adopt a probabilistic approach by taking into account prior probabilities of author and inventor names in the

respective corpora. For example, even if "John Smith" is an inventor on the patent and an author on the paper, we do not consider this a strong indicator of overlap unless there is also a match on the middle initial (stronger still if there is a match on the full middle name). Conversely, the less common the overlapping author/inventor name, the stronger the match.

Moreover, we deviate from prior approaches in the interest of avoiding false-positive PPPs by requiring that the patent cite the paper. Although it might be possible for an academic article to be patented without the involvement of the original scientist, in our estimation the false positives avoided by building PPP based on patent-to-paper citations likely outweighs possible false negatives. We also require that the patent be filed no later than five years after the publication of the article.

An example of a patent-paper pair linkage is illustrated in Panel A of Figure 2. The PNAS paper with "Jennifer A. Doudna" listed as the final author is cited by patent 9,260,752 in Panel B, where "Doudna, Jennifer A." is an inventor. Although Jennifer is a common forename, Doudna is an uncommon surname, composing 0.01625% of all scientific authors and 0.0339% of all patenting inventors. (The matching middle initial "A" on the paper and patent further raises our confidence that this is an overlapping author/inventor.) Indeed, patent 9,260,752 is assigned to Caribou Biosciences, of which Doudna is a founder and serves on the Scientific Advisory Board.

Figure 2 about here

2.1.2 "Transitive" patent-paper pairs

Because we are interested in the commercialization of science by firms, the PPP is not necessarily our end-point dependent variable. Rather, we want to know whether the original science is commercialized *by a firm* and therefore limit our scope to PPPs case where the patent assignee is a commercial entity. We recognize, however, that in some cases, a firm may commercialize a scientific discovery where the university has already claimed patent protection, often, by issuing a license to the patent.

As noted above, we cannot observe university-licensing data at scale. However, we take a step toward including such arrangements by also capturing what we refer to as "transitive" patent-paper pairs. In this scenario, a university patents a paper, and then a firm files a patent that cites that same PPP and with an overlap between the inventors on the firm-assigned patent and the university-assigned patent it cites (which is paired with the original scientific article). These "transitive" paper pairs account for approximately 14 percent of our commercialization instances; however, removing them does not affect the results.

2.1.3 Self-commercialization via startups

Finally, we are interested in whether scientists *self*-commercialize their discoveries, launching new ventures instead of collaborating with existing companies in order to do so. Understanding this outcome

is critical because self-commercialization represents an avenue of greater autonomy. Even though academic entrepreneurs may seek funding from angels and venture capitalists, they may nonetheless be less beholden than otherwise to the willingness of large corporations. Moreover, those who self-commercialize scientific discoveries can take advantage of non-venture sources of funding; for example, in commercializing her doctoral research on pathogen detection, OmniVis founder/CEO Dr. Katharine Clayton was awarded a string of NIH and NSF SBIR grants which enabled her to build a medical device and market it largely without taking dilutive investments.

To classify PPPs as self-commercialized vs. cooperatively commercialized, we link PPPs to the PitchBook dataset of startups.[9] We fuzzy-match patent assignees to PitchBook company names through September 2021, Unlike PitchBook, patent assignees do not exclude suffixes such as "Corp", so these are removed in advance of matching in order to improve confidence scores. A low matching confidence threshold is employed to avoid false negatives, coupled with manual review of every match to eliminate false positives. Note: just because the patent assignee matches a company name in PitchBook does not mean that the PPP represents commercialization via a startup; for a PitchBook match to count as self-commercialization, the paper in the PPP must have been published no later than two years after the founding of the company (we allow a 2-year window to account for delays in the publishing process). The example in Figure 2 involving Jennifer Doudna and Caribou Biosciences is, therefore, an example of self-commercialization.

2.1.4 Characterization

The PPP method of detecting commercialization is not without limitations. PPPs might detect instances of sponsored research as commercialization (i.e., false positives), which we cannot algorithmically rule out. It might also yield false negatives by failing to detect instances of commercialization where, for example, a company was granted a license to a university technology, and the ensuing corporate patent(s) cited the original academic article, but because the company had different staff work on the project not including the original authors, our algorithm would not pick up that PPP. Employing a similar algorithm to detect instances of startup commercialization, Marx and Hsu 2021 hand-checked a random sample of 40 such PPPs and found that 39 were correct. This suggests that the false-positive rate is low, although we lack a "golden" dataset to characterize false negatives . It is not immediately obvious to us why the algorithm should be biased toward or away from detecting PPPs involving men vs. women, though if it were the case that, say, patenting or publishing women had much more rare (common) forenames than men, it would be easier (harder) to detect author/inventor overlap. In fact, Thursby and Thursby 2005 show that women are less likely to license their inventions than men. Appendix Table A1 shows the top 15 commercializers of scientific articles,

9. Unfortunately, no open-source dataset of startups is currently available and so this dependent variable is not publicly replicable.

and Figure 3 shows commercialization rates over time. Figure 4 illustrates commercialization rates by OECD scientific category.

Table A1, Figure 3, and Figure 4 about here

2.2 Gender classification

The Microsoft Academic Graph does not report the gender for authors. Given the volume of data, it is impractical to carry out a manual verification of every author's gender. Thus, we determine the gender of the authors from inference based on the first name (a common practice in big data analysis). We rely on Genderize.io API, a corpus-based dictionary that uses large databases of names collected from the US Census, international dictionaries, and social media and calculates the probability that a specific first name is associated with the male or female gender. In other words, for a name like "Anna" it computes the fraction of all individuals with that name that are women or girls and men or boys.

Further, we use a restrictive procedure applying a threshold of 90% to assign gender. Indeed, when comparing automatic classification vs. hand-classification in section 2.2.1, we found substantial errors for confidence scores lower than 90%. In addition, we do not assign a gender to authors without first names. Next, we identify the gender composition of each team based on the gender composition of the authors. In doing so, we primarily use a binary variable equal to one if the last author (usually the lab head or the principal investigator) is female and 0 if the last author is male. Given our focus on the lab head, we calculate the percentage of non-last authors who are female.

We also consider alternative definitions of the gender composition of a paper that captured more saturated structures such as a four-dummy model: first author male-last author male, first author female-last author male, first author male-last author female, first author female-last author female; or a three dummy-model: all-male teams, mixed gender teams with male last author, mixed gender teams with female last author, all-female teams; as well as less saturated gender composition (at least one female author, percentage of female authors on the team).

Finally, we present alternative specifications in which we discard papers where we cannot classify the gender of at least one-third (one-half, three-quarters, all) of the authors. For example, if there are two authors and only one can be classified, the paper is discarded. If there are three authors, two must be classified. However, the results are not overly sensitive to those restrictions, as can be seen in Appendix A2.

2.2.1 Manual classification

Given that most of our analyses focus on the gender of the final author (i.e., PI), we hand-coded the gender of every last author of a "twin" article, which did not receive a confidence score of 99% or

100% by the automatic classifier. We chose this cutoff based on our manual review of thousands of automatically-classified scientists. There were no errors at the 100% level and only one error out of thousands classified at the 99% confidence level. Therefore we hand-coded gender 9,748 PIs, looking up their faculty pages for visual or textual clues.

Finally, to increase confidence in the results, we hand-code the gender of *every* author for a subsample of of the "twin" articles. As described in Section 3.3, our primary identification strategy is to control for unobserved commercial potential by analyzing "twin" discoveries that yield multiple papers reporting the same science. There are 27,551 twin articles with more than 120,000 authors, a prohibitively large number to hand-code. However, our estimates are principally driven by the subset of 1,704 articles reporting twin discoveries where one article was commercialized and the other was not. (Indeed, our robustness checks using clogit instead of OLS necessarily omit twin discoveries where neither, or both, of the articles reporting the discoveries was commercialized.) We hand-coded 13,729 authors including all of those for the twin discoveries with variation in the outcome variable.

A team of research assistants (RAs) hand-checked the gender of all of these authors. RAs were instructed to find the author's website, LinkedIn page, or other individually identifying information. All but 527 of the authors were located, most of which had confidence scores of 99% or 100% from the algorithm. For the remaining authors, an RA checked for people with the same name to identify gender by proxy. Of the 7049 names that were automatically classified at 99% or 100% confidence, only one was determined to be incorrect. A total of 23,477 scientists were hand-coded. Descriptive statistics for the 70,016,266 papers analyzed are in Table 1.

Table 1 about here

3 Results

3.1 Empirical specification

Let us define by $COMM_{it}$ the commercialization status of a given scientific article i published at time t, to be the outcome of interest. $COMM_{it}$ is a dummy variable equal to 1 if the scientific article i has been commercialized and 0 otherwise. $Gender_i$ refers to the gender structure of the authors of the scientific article i. X_{it} captures a set of control variables such as the number of authors, the "prestige" of the authors, and their institutions as calculated by the average forward citations (in a five-year window) for each and the impact factor for the article's journal. Finally, we include a battery of fixed effects in the model with $Time_{FE}$ standing for the publication year fixed effects and $Field_{FE}$ for the article's field fixed effect where the field is defined by the WoS field category. We define the error term by ϵ_{it}. Then, we use the following specification to analyze the gender gap in commercialization:

$$COMM_{it} = \alpha_0 + \alpha_1 Gender_i + \beta X_{it} + Time_{FE} + Field_{FE} + \epsilon_{it} \quad (1)$$

We estimate equation 1 using linear probability models with robust standard errors. The coefficient of interest is α_1 which captures the gender gap in commercialization.

3.2 Cross-sectional Results

Table 2 examines the gender gap in commercialization. Although we will explore various ways of measuring the gender composition of the authors, in Table 2 we focus on the *last* author (often, or usually, the Principal Investigator). We observe higher commercialization rates for male-led scientific teams. In column (1), having a female last author is associated with a reduction of the likelihood of commercialization by 29 percent (–0.000957/0.0033). This gap persists in column (2)—although it is reduced in magnitude to about 17 percent—when we control for the percentage of female authors on the paper who are not the final author.[10]

Table 2 about here

Although Table 2 shows lower rates of commercialization among science conducted by women, this result could be mechanical if there is a quality difference that matches the different gender categories. To address these questions, in columns (3) and (4) we examine the scientific quality of articles by counting forward citations from other academic articles in a fixed, five-year window. When an article has a female final author, the citation count increases.[11] Therefore, lower commercialization rates among female scientists are not likely due to the underlying quality of the discovery.

The contrasting findings in Table 2 are striking. Although having a woman as the last author on a scientific team increases the scientific community citation, the commercialization pattern suggests a decrease in the likelihood of commercializing for teams led by women. However, there are potential threats to identifying the gender gap in commercialization in the cross-sectional data. Perhaps the most immediate interpretation is that women conduct more theoretical science (and are therefore appreciated by the academic community) but which is less applied and therefore ignored by commercial firms due to "latent commercializability" (Marx and Hsu 2021). Ruling this out is difficult because it is hard to know ex-ante the commercial potential of a scientific discovery, or to have a reliable proxy for paper quality. In the next section, we attempt to account for latent commercializability and quality by comparing commercialization outcomes for pairs of "twin" scientific discoveries with varying percentages of female authors.

3.3 "Twin" scientific discoveries

We attempt to control for latent commercializability by adopting the Bikard 2020 method of identifying "twin" discoveries based on co-citation. Although co-discoveries are uncommon in the social sciences,

10. For single-authored papers, this percentage is set to zero rather than a missing value.
11. Our findings are reminiscent of Yang et al. 2021 in Medicine, although we examine all fields of scientific inquiry.

they happen routinely in the hard sciences. Twin discoveries are identified as papers satisfying the following five conditions: (1) they are published no more than a year apart; (2) they have zero overlaps among authors; (3) they are cited at least 5 times; (4) they share at least 50% of forward citations; and (5) they are cited *adjacently* (i.e., within the same parenthesis) at least once. Bikard and Marx 2019 confirmed no false positives in a random sample of twin papers identified via the above algorithm. We combine the public twins made available by Bikard 2020 with those generated by Marx and Hsu 2021, crosswalking the Marx/Hsu twins from the Web of Science to MAG.[12] This exercise resulted in a set of 27,551 twins in MAG. Descriptive statistics for this subsample are in Table 3.

Table 3 about here

The empirical specification follows that of twin studies in the epidemiological literature (Carlin et al. 2005), introducing a fixed effect for each twin discovery (TwinDiscoveryFE). Year and field fixed effects are omitted, given the nature of the twin-matching exercise.

$$COMM_{it} = \alpha_0 + \alpha_1 Gender_i + \beta X_{it} + TwinDiscoveryFE + \epsilon_{it} \quad (2)$$

Table 4 refines the cross-sectional commercialization estimates of Table 2 while controlling for latent commercializability and quality. The effect of doing so becomes apparent in columns (1-2) of Table 4, where the estimated coefficient of the last author as female is associated with approximately a 1.52 percentage point decrease in commercialization. Given the base rate of commercialization among the twin papers of 4.5%, this implies a drop of nearly one-third. Therefore the gender gap in commercialization remains even when controlling for latent commercializability. This gap is materially unchanged when accounting for the percentage of women who are not last authors of the article in column (2). The somewhat weak significant coefficient of the percentage of women who are not the last authors suggests that both the presence and positionality of the female author matters.

Table 4 about here

Although our main objective is to verify that the findings in Table 2 are not driven by unobserved heterogeneity, we also confirm using the twins' method that the science published by women is not of lower quality. One might allege that, for some reasons, the results in our Table 2 or in (Yang et al. 2021) are due to women accruing artificially high numbers of citations from other academics. Disproving this is difficult in the cross-section, but when examining twin discoveries, one can test whether men vs. women receive more citations for the same scientific discovery. Columns (3-4) of Table 4 confirm that there is no gender bias in citation patterns from academic articles.

12. The majority of WoS-based twins from Marx and Hsu 2021 could be mapped directly via Digital Object Identifiers; the remainder were fuzzy-matched using author, title, year, journal, volume, and page(s).

In Table 5, we employ alternative measures for the gender composition of the scientific team. In column (1), we check whether there is at least one female author in any authorship position. The estimated coefficient is associated with a 20 percent reduction in the likelihood of commercialization, similar in sign to having a female last author but lower in magnitude.

In column (2), we instead enter the percentage of female authors in the article, again obtaining a negative and statistically significant estimated coefficient. There is only weak evidence in Column (3) that having a first author as female contributes to the gap, although one might suspect as much given the frequent "et al." citations of scientific articles with many authors. Even though the coefficient on first-or-last-author-female in column (4) achieves statistical significance at the 5 percent level, its magnitude is smaller than that of female-last-author: column (2) of Table 2 is repeated here in column (5) for convenience. Taken together, columns (1-5) of Table 5 show that positionality is a key factor in the commercialization gender gap.

A similar pattern holds in the final column of Table 5 when considering a two-dummy model with mixed-gender teams (teams with at least one man and at least one woman) and all-female teams. There is no commercialization gap for mixed-gender teams where the last author is male, only when the last author is female. (Although the estimated coefficient for all-female teams is positive, there is only a handful of commercializing all-female teams; therefore, the standard error exceeds the coefficient in magnitude by an order of magnitude.) Appendix Table **??** provides additional views into the gender composition of the authorship team.

Table 5 about here

The twins model shows that even for very similar academic papers, controlling for unobserved differences in quality and latent commercializability, papers with female lab heads experience a lower likelihood of commercializing. This result is at odds with the fact that those very same papers are also the ones with more academic recognition as measured by citations (column 3 of Table 2), suggesting that lower quality is not an explanation for the gender gap in commercialization.

3.4 Robustness

We test the robustness of this finding in several ways. First, we consider a relevant sample where the gender of all authors is hand-annotated. Then, we verify that the results for lab head are not systematically due to a patent citation gap. Next, we show that the results are robust to controlling for previous commercialization experience. Finally, we introduce another method of quality control and commercialization potential using the biological sequence database.

3.4.1 Hand-checked gender and conditional logistic regression

As described above, we hand-code gender for all last authors of articles. One may nonetheless worry that the control variable for the percentage of authors in non-final positions is subject to error. We, therefore, hand-coded all authors for the 1,704 'twin' articles, reporting a simultaneous discovery where one article was commercialized but not the other.[13] Columns (1-2) of Table 6, reveals that the main results are robust to this restriction.

A related concern may be that our inability to classify every author reliably leads to selection concerns. In particular, concerns may arise for Chinese authors who are likely to be excluded by the 90% confidence score. Again, we hand-code *every* last author for the twin articles and, for twin articles where one or the other is commercialized, we hand-code *all* authors as shown in columns (1-2) of Table 6. In models utilizing all twin articles, the covariate capturing the percentage of female non-last authors may be subject to inaccuracy. To show that our results are not sensitive to that issue, we vary the cutoff for the set of identified authors in a team as well as the cutoff of the probability of assigning the gender to an author.

In Appendix Table A2, we re-estimate both our twin and non-twin models, subsetting to continually stricter subsets of the dataset where a higher percentage of all authors on a paper can be reliably identified by gender. Panel A of Table A2 reveals that our twin results are robust when requiring that half of the authors of the article are reliably identified. When requiring that 75% of authors be identified, the results are statistically significant at the 5% level, still in the subset of about half of the twins. Requiring that all authors be identified against strengthens the precision of the estimate, albeit at the expense of sample size.

In Appendix Table A3, we perform a complementary analysis based on the confidence of the gender identification by the algorithm for each individual author. The columns represent increasingly stringent thresholds from 50-95% (the algorithm returns no predictions with confidence <50%). Interestingly, the results remain consistent even when we vary the threshold for assigning gender probabilities. However, it is important to note that when the thresholds become more lenient, there is an introduction of measurement error in the independent variable of interest. As a result, the classical attenuation bias due to measurement error becomes evident, as shown in columns (1) and (2) of Appendix Table A3. This same concern motivates our extensive hand-coding of gender for lab heads.

Columns (3-4) of Table 6 Panel A follow Beck 2020 by re-estimating our main twins model using conditional logit, which omits any twin discovery where neither of the twin articles reporting the discovery is commercialized. Observations counts are closer to those of the hand-coded only models, with similar estimated coefficients.

Table 6 about here

13. The number of observations in columns (1-2) of Table 6 exceeds that of column (3-4) somewhat because some authors of twins where one article was commercialized also appear in twins where neither was commercialized.

Furthermore, in Appendix Table A4 we exclude "transitive" PPPs, in Appendix Table **??** we present a more comprehensive table of alternative measures of gender composition, retaining the main result.

3.4.2 Placebo test: patent-to-article citations

Another concern may be that we are not capturing commercialization but merely citations by patents. Patents cite scientific articles for a variety of reasons having nothing to do with commercialization. The USPTO "duty of candor" requires patent applications to reference all prior art from whatever source, including but not limited to the scientific literature. Given that our strategy for detecting patent-paper pairs relies on the presence of a citation from a patent to the paper, it is possible that we are merely picking up patterns of citation.

In Table 7, instead of using commercialization as our dependent variable we instead use the count of citations from patents—whether or not that patent represented a patent-paper pair. Columns (1-2) consider only citations from the front page of patents, which are legally binding (Marx and Fuegi 2020); Columns (3-4) consider only citations from the body text of patents, which are not binding and are more likely to be added by the scientists themselves as opposed to patent attorneys (Bikard and Marx 2019; Marx and Fuegi 2021); Columns (5-6) consider both.

In none of these cases is the presence of a woman as the last author on the paper statistically significantly associated with the likelihood of the article being cited: either on the front page or in the body text. We do obtain a negative correlation between front-page citations (but not in-text) and a higher percentage of women who are not the last author on the paper. However, this does not seem to be economically significant (a variation of barely 2% from the mean).[14] Note that our main commercialization gap result exceeds 30%. In other words, this does not appear to explain the commercialization gap for female last authors.

Table 7 about here

3.4.3 Robustness: prior commercialization experience

Table 8 establishes the robustness of the results to including measures of authors' and institutions' prior experience with commercialization. The author-related variable captures whether any of the authors on a focal article had commercialized a paper published before the focal article. The institution-related variable is a logged count of all commercialized papers by any author at one of the institutions with which the authors of the focal paper is associated.

14. While most papers have 0 patent citations, we move from 0 (the 61st percentile) to 0.6931 (the 62nd percentile), from which a value of 0.0152 barely represents 2.1%. Similarly, when considering the probability of having at least one patent citation instead of the actual patent counts, the coefficient obtained is both statistically and economically insignificant. For instance, the coefficient for the variable indicating the presence of at least one patent citation (including both front-page and body citations) is -0.0037 (standard error: 0.00817). Moreover, the mean of this variable is 38.66%, further indicating its lack of substantial impact.

Across all columns, we estimate a negative and statistically significant coefficient for female last author, as in Table 4, despite controlling for the author's previous commercialization experience in columns (1-2) the institution's commercialization history in columns (3-4). Moreover, and consistent with Marx and Hsu (2021), commercialization experience matters. This is visible in the positive and precisely estimated coefficients on the author's previous commercialization experience (columns 1-2), and the institution's history of commercialization (columns 3-4). Although our results are robust to the inclusion of these variables, we do not include it in most models because prior commercialization (or lack thereof) may be explained by precisely the gender gap we measure and as such may be a "bad" control (Angrist and Pischke 2009).

Table 8 about here

3.4.4 Robustness: twins based on identical biological sequence and structure

One critique of Bikard's algorithm for detecting twin papers is that articles may be cited jointly for reasons other than them having duplicate content. We, therefore, re-establish our results using a second method for determining twins, albeit one that applies only to the life-sciences subset of our data. Following Hill and Stein 2019, we construct pairs of papers in structural biology, but with key refinements. We begin with all proteins deposited as of June 2022 to the Protein Data Bank (PDB), which is a repository used in structural biology. PDB clusters proteins with similar underlying structural entities using its MMseqs2 sequence-clustering algorithm.[15] MMseqs2 can be implemented at varying levels of similarity; for example, Hill and Stein 2019 employ a 50% similarity match. To ensure exact matching on biological structure, we employ a 100% similarity match.

Even so, a 100% similarity structural match in PDB can be misleading because this match can be, as Hill and Stein describe, for one of many structural entities between proteins. We, therefore, employ the Uniprot database to ensure a unique *sequence* as well for proteins that share even identical portions of their substructure. Whereas PDB focuses on structures, Uniprot focuses on sequences. We submit the PDB identifiers from the previous step to Uniprot, which returns a Uniprot identifier for each PDB identifier. Because a unique protein substructure may be employed by multiple proteins with different sequences, Uniprot may map a single Uniprot identifier to multiple PDB identifiers. We, therefore, retain only the unique PDB-Uniprot mapping to obtain a list of proteins that are unique in both structure *and* sequence.

Uniprot also returns a list of published articles for each of its unique identifiers. Some Uniprot identifiers are associated with many articles because a single protein can be employed for a variety of studies and may not represent the same scientific discovery. Therefore, in columns (1-2) of Table 9, we examine only proteins with a unique structure and sequence that are reflected in exactly two published

15. MMseqs2 is an upgraded version of the BLAST algorithm employed by Hill and Stein 2019.

articles. This exercise yields 5974 articles from 2987 proteins with identical structure and sequence. Because we define twins according to biological structure and sequence, we employ fixed effects for each PDB identifier. Column (2) of Table 9, suggests that among these "biological twins" there is a commercialization gap for scientific teams with female last authors, but not given the percentage of women in other authorship positions. Here, the estimated gender gap exceeds 50 percent. Columns (3-4) demonstrate that estimation becomes less precise when allowing any number of twins with shared structure and sequence (even sextuplets and septuplets), which would seem less likely to be replications of the same scientific discovery.

Table 9 about here

4 Mechanisms

Although the twins regressions in Table 4 control for unobserved commercial potential and show lower rates of commercialization among science conducted by women-led teams, it is unclear what mechanisms drive these dynamics.

On the supply side, one might observe a gender gap in commercialization if female scientists have different preferences, i.e., are not interested in commercializing and/or patent less (Whittington and Smith-Doerr 2005). In this case, the PPP-based result will be a mechanical derivation from the argument of women patenting less. On the other hand, conditional on being interested in the patenting process, women may lack networks, connections, and other forms of support to commercialize (Murray and Graham 2007).

On the demand side, firms might be biased against female scientists and discriminate against women in searching for commercializable discoveries, as has been found with venture investors and female entrepreneurs (Brooks et al. 2014). Alternatively, it could be that the lack of self-promotion of their work attracts less attention from firms compared to similar male-authored works. We investigate both supply-side and demand-side factors in the remainder of this section.

4.1 Supply side: commercialization and gender representation

We begin with a possible mechanism suggested by prior literature: that gender disparities in commercial activity by scientists may be attributable to a lack of representation. For example, Tartari and Salter 2015 suggest that women are more likely to engage in commercialization in fields with higher representation of women. In Table 10 we explore whether we can find evidence for this mechanism at scale.

For each of the 251 fields defined by the Clarivate Web of Science and crosswalked to keywords in MAG, we calculate the percentage of scientists publishing in that field each year who are female. This variable is entered into Equation 2 as a covariate in column (1) of Table 10 and is then interacted

with the presence of a female last author in column (2). (These models are then repeated in columns 3-4 with controls for the percentage of non-last female authors.)

The estimated coefficient on the interaction of female last author and within-field female representation in (2) is positive, consistent with the finding of Tartari and Salter 2015, but only weakly statistically significant. We take this as suggestive evidence at best that higher gender representation in a scientific field promotes commercialization.

Table 10 about here

4.2 Supply side: commercialization and professional networks

Next, we explore the extent to which boundary-spanning networks play a role in commercialization. For example, Murray and Graham 2007 report that female scientists who wish to commercialize their work are excluded due to lack of access to professional networks at firms. We proxy for professional networks by counting, for each author on a focal paper, their unique prior coauthors who were affiliated with a commercial firm at the time of the prior paper's publication. For each author, this reflects the first-degree commercial reach of their coauthorship networks. This figure is summed among the authors on a focal paper, with (unreported) alternative calculations of the mean and maximum yielding similar results.

Table 11 first utilizes this paper-level measure of professional access as a dependent variable. In column (1), the last author on papers with female last authors have fewer prior coauthors at firms. This is true also when excluding all articles that have even a single author from industry, in column (2), and in unreported results for the mean and maximum for the authorship team as well as for the number of prior coauthors belonging only to the last author.

Table 11 about here

Given that scientific projects led by women have fewer prior commercial coauthors, in columns (3-4) of Table 11 we investigate the association between this and the likelihood that the focal paper is commercialized. Prior work would suggest that stronger connections to industry would predict commercialization, all else equal. Controlling for latent commercializability in Table 11 as well as various aspects of the authorship team—including its size, aggregate citations, and institutional affiliation prestige—we obtain a positive sign on the estimated coefficient, but with little precision. This does not necessarily contradict prior findings, which relied on finer-grained data from interviews with practicing scientists; indeed, it may be that useful professional contacts extend beyond coauthorships on published articles. Another possible explanation is that professional contacts could be endogenously determined by other covariates, such as the prominence of the authors. In this case, conditional on these covariates, the expected importance of professional outreach diminishes.[16]

16. Moreover, we note that the (logged) count of prior coauthors at firms is highly collinear with the controls for team

Column (3) reveals that the disparity in professional networks accounts for only a minimal fraction of the gender gap in commercialization: Over 80% of the gap remains unexplained. Additionally, even when women possess a commercial network, there is no significant increase in their likelihood of engaging in commercialization, as evidenced by Column (4). Even if we were to assume that the interaction coefficient was significant, our findings suggest that scientific teams led by women would require a substantial increase in "ln prior coauthors at firms" to narrow just half of the gap. To be precise, they would need approximately ten times the natural logarithm of prior coauthors at firms, where the value of ten corresponds to thousands of coauthors working at firms.

4.3 Supply side vs. demand side: commercialization "mode"

Given that we fail to find strong evidence for supply-side mechanisms, we proceed by decomposing our dependent variable into two forms of commercialization. Our definition thus far is a PPP where the assignee is a commercial firm. However, different types of firms may indicate different commercialization mechanisms, which can provide insight into the role of supply vs. demand factors. One mode of commercialization assumes cooperation between an extant firm and one or more publishing scientists. But in another mode, one may choose to *self*-commercialize by founding a startup company instead of cooperating with an existing firm. We predict that demand-side pressures will matter more in the case of cooperative commercialization, whereas supply-side factors may have greater influence for self-commercialization.

Table 12 splits our dependent variable into cooperative (columns 1-2) vs. self (3-4) commercialization, the former with existing firms and the latter via startups. As described above, for the latter dependent variable we link patent-paper pairs to PitchBook. Of all patent-paper pair twins, 3.7% are cooperatively commercialized and 0.5% are self-commercialized, a ratio of about 8:1.

Table 12 about here

In Table 12, we are able to precisely estimate a gender gap only for cooperative commercialization with existing firms. The estimated coefficient on having a female last author in column (1) is similar in sign and magnitude to that of column (1) in Table 4, and is even somewhat more precisely estimated. However, no such result is found in columns (3-4) where self-commercialization is the dependent variable. The estimated coefficient on female last author is imprecisely estimated and positive; the positive sign on the estimated coefficient for female-last-author in columns (3-4) indicates that the lack of a result for self-commercialization is not simply a power problem. This moreover appears consistent with our failure to find strong evidence for the supply-side mechanisms of within-field representation

size (.71) and aggregate forward citations (.79), which may make it difficult to obtain a precise estimate. In unreported models where these controls are omitted, the estimated coefficient on prior coauthors at firms is positive and significant at the 1% level. The alternative measures of average or maximum past coauthors are also collinear with the controls.

and professional networks. We conclude that although firms appear to avoid commercializing science discovered by women, there are not substantial differences in *self*-commercialization rates.

On the one hand, this result might be anticipated because founding a company is a more autonomous act than finding an established company to license one's discovery or with whom to collaboratively cooperate to commercialize. On the other hand, this may seem surprising given the well-documented gender gap in entrepreneurship more generally (Guzman and Kacperczyk 2019). One possible explanation for the lack of a gender gap in self-commercialization could be a substitution effect from (failed) cooperative commercialization to self-commercialization. Another is that it is not the case that every commercialized discovery requires venture capital, at least not to launch the company. For example, when Katharine Clayton founded OmniVis to commercialize her Ph.D. research on pathogen detection from Purdue University, she elected to fund the company initially via SBIR grants.

The results of Table 12 also help to allay the concern that our findings are driven simply by differences in patenting rates between men and women. Azoulay, Ding, and Stuart 2007 and others have documented lower patenting rates and disclosures (Bercovitz and Feldman 2011) among women, despite similar baseline research productivity. Similar (or higher) quality of research in the face of lower commercialization output could be explained simply by lower patenting rates, given that we measure commercialization via patent-paper pairs. That we do not find a gender gap in self-commercialization provides some reassurance that our findings may not be mechanical and likely not due to women's lack of interest in commercial activities.

4.4 Demand side: accessibility of articles

Why are scientific discoveries led by women less likely to be commercialized in cooperation with existing firms? We begin our investigation of demand-side factors by assessing the impact of awareness. Collaboration between a researcher and a firm can occur in different ways: published papers and reports, public conferences and meetings, informal information exchange, and consulting, geographic hubs (Cohen, Nelson, and Walsh 2002, Markman, Siegel, and Wright 2008, Bikard and Marx 2019). While it is challenging to quantify the relative share of each method that can lead to commercialization, it is widely acknowledged that access to scientific publications promotes scientific collaboration (Gowers and Nielsen 2009, Friesike et al. 2015, McKiernan et al. 2016). Thus, it seems plausible that knowledge about one research could lead to a collaboration with a firm.

Therefore the gender gap in commercialization could also be salient in an environment where access to information on scientific articles is not perfectly distributed and gender-specific.[17] In particular, if scientific articles with women are less visible than male-authored scientific articles, this could prevent

17. Indeed, dissemination of academic research via social media, for example, has been shown to increase the visibility and the likelihood of citation (Eysenbach 2011 and Klar et al. 2020).

companies from accessing the former's publications and therefore reduce their probability of commercializing relative to the latter. If such a hypothesis turns out to be accurate, then a shock that would increase awareness and access to scientific research should contribute to reducing the gender gap in commercialization.

To test this, we use a natural experiment provided by the open-access mandates for federally-funded research. In many scientific fields, most articles and working papers are not freely available (Bjork, Roos, and Lauri 2009, Khabsa and Giles 2014, Ware and Mabe 2015). At the same time, one of the most common rationales behind the evolution of scientific discovery is to expand the frontier of knowledge by building upon previously available research. In fact, many authors have shown that limited awareness (limited access or openness constraints) about scientific production can limit the use of science (Furman and Stern 2011, Williams 2013, Murray et al. 2016, Staudt 2020, Bryan and Ozcan 2021). This channel could be more important in the commercialization of academic research as firms may need to explicitly collaborate with a researcher from an academic institution.

In 2008, the National Institutes of Health (NIH) leveraged an initiative to make freely available the academic research they funded such that any article accepted for publication after April 7, 2008, must be archived in the open-access PubMed Central (PMC) database within 12 months of publication.[18] In 2013, the White House Office of Science Technology Policy mandated agencies with an R&D budget of $100M in order to develop plans to make the results of the federally funded research freely available. This gives rise to a staggered implementation of the "Public Access policy" (PAP) with, for example, the Department of Energy (DOE) implementing this policy in 2014 and the National Science Foundation (NSF) in 2016.

Our empirical model takes advantage of the gradual implementation of the PAP by constructing an event study where the event date is the starting year of the PAP for one agency. Therefore, an article in the database is considered to be "treated" in a given year if a federal agency financed this paper, and during that year, this agency started to implement the PAP. In this setup, we are particularly interested in the triple difference that captures the effect on the commercialization of federally-funded publications written by women after the implementation of the PAP relative to those written by men. Therefore, assessing the effect of the PAP on narrowing the gender gap in commercialization. We further add the control variables similar to our baseline and include the journal, year, and field fixed effects. Our identifying assumption is that there are no shocks correlated with the introduction of the PAP that differentially affect scientific teams with men/women commercialization likelihood. To address concerns regarding heterogeneous treatment effects, we use a robust staggered difference and difference approach by Sun and Abraham 2021. Other procedures to solve this issue have been proposed by Goodman-Bacon 2021, Callaway and Sant'Anna 2021. (Baker, Larcker, and Wang 2022

18. Most of the paper in the literature of open access on academic citation finds a non-negative effect. In particular, Bryan and Ozcan 2021 show that after 2008, NIH-funded researches were 12 to 27% more likely to get cited, while Staudt 2020 finds a positive but more moderate effect.

shows an interesting equivalence between those different procedures.)

Figure 5 shows the result of this estimation, with Appendix Table A5 providing the corresponding numerical values. We use one year before the introduction of the PAP as the reference year. Panel A of Figure 5 presents the difference-in-difference estimate for papers with a woman as the last author (i.e., lab manager). Although the pre-trend for the sample difference and difference is not non-significant, impeding the interpretation of the simple difference, we clearly see that both genders are moving in an almost perfect one-to-one mapping. There seems to be a sharp jump in the commercialization of science following the advent of Open Access mandates, but we do not see a material convergence of the gender gap. Rather, the gap widens starting in year 1, diverging further in years 2-4.

Panel B of Figure 5 plots the result of the triple difference exercise. There is no statistically significant pre-trend and no effect after the PAP. This means that the PAP has not shrunk the gender gap in commercialization, as we would expect if a differential in access to information about articles were responsible for the gap.

Figure 5 about here

We conclude that contrary to priors that increased information might help to close the gender gap, the introduction of open access mandates in fact exacerbated the gender gap in commercialization for scientific projects led by women. By contrast, for teams where women are the last authors, the open access policy causes a widening of the commercialization gap. This is suggestive of bias on the part of firms and reinforces our focus on demand-side factors.

4.5 Demand side: firms' attention to word choice

Previously we supplied causal evidence that increase ease of access to scientific literature only exacerbates the gender gap in commercialization. The question remains why firms would be less likely to pay attention to science led by women. One possibility explored in the literature is that women use less self-promoting language when writing scientific articles (Lerchenmueller, Sorenson, and Jena 2019 hereafter, "LSJ" and also). If so, it might be that firms are more enticed by scientific articles by men if they are more prone to describe their research findings using words such as "breakthrough" or "unprecedented."

We collected the titles and abstracts for *all* articles in the Microsoft Academic Graph and checked whether they contained any of the "boastful" words used by LSJ, with one exception. The most commonly found boastful word in LSJ's analysis was novel. Worrying that novel is often used not for boasting but rather to identify novelty as in *novel coronavirus*, we excluded such bigrams. With these refinements, in column (1) of Table 13, we extend LSJ analysis to the universe of academic publications and confirm LSJ's finding that overall, papers with female last authors tend to use boastful words less often. Note that this is also true for papers with a higher percentage of women who are not the last

authors. When the sample is restricted to twins articles, the association is dominated by the percentage of female authors involved in the article, with an attenuation of the effect of the last author.

Next, we add an indicator for whether "boastful" words occur in the title or abstract of the twin articles.[19] In column (3) we analyze whether boasting is associated with the likelihood of being cited by academic papers. We see no correlation between word choice and citation counts. In other words, academic scientists are not swayed by boasting. Firms, however, appear to pay more attention to articles that describe themselves as "breakthroughs" and the like, even when controlling for content via twins. In column (5) of Table 13, articles with boastful words are commercialized at a higher rate.

Finally, in column (6) we interact the indicator for presence of a boastful word with the presence of a female last author. Although the estimated coefficient is negative, it is rather imprecise. We conclude that to the extent boasting draws more attention from firms, it appears to happen primarily for articles with male last authors.

Table 13 about here

4.6 Demand side: commercialization and gender homophily

In this section, we present evidence of gender homophily in the commercialization process. In an ideal experiment, we would randomly seed commercialization partners with heterogenous gender composition and assess the likelihood of commercializing scientific articles of otherwise identical quality but heterogeneous composition of the scientific teams. The analytic approach used so far (i.e. Equation 2) cannot accomplish this because the setup is at the level of the academic paper and thus cannot compare the gender composition of the paper with that of the patent.

Thus, we switch the level of analysis from the paper to a patent-paper dyad that *potentially* forms a patent-paper pair. As before, we account for the quality and nature of the paper with "twin" discoveries. To approximate the random seeding of patents that could form a patent-paper pair, we adopt a case-control setup. We reduce our set of twin papers to those where one or the other indeed formed a pair with some patent. A dyad is formed both for the patent and the paper with which it *is* paired as well as for the patent and the twin of the (actually paired) paper, with which the paper was *not* paired but should be about as likely to have been paired. This unrealized patent-paper pairing forms a counterfactual for our case-control analysis.

For the patent, we calculate the percentage of male vs. female inventors on the patent using USPTO inventor-level classification (therefore this step is limited to USPTO-issued patents only). To avoid biasing this measure in favor of the paper that was actually cited, in calculating the gender composition of the patent we omit the inventor(s) who were matched to authors on the paper in the

19. Columns (2-6) replicate with a count of boastful words instead of an indicator. We did not generate a count variable for the entirety of MAG in column (1).

(realized) patent-paper pair. We then estimate the following equation, which deviates from Equation 2 by including a) the gender composition of patent j b) fixed effects for patent j.

$$\begin{aligned} COMM_{ijt} = \alpha_0 + \alpha_1 PaperGender_i + \alpha_2 PatentGender_j + \\ \alpha_3 PaperGender_i X PatentGender_j + \\ \beta X_{it} + TwinDiscoveryFE + Patent_j FE + \epsilon_{ijt} \end{aligned} \tag{3}$$

Column (1) of Table 14 replicates our main (i.e., non-homophily) finding, using a dummy for a female last author. In column (2), the indicator for having a female last author is interacted with the percentage of male inventors on the focal (i.e., citing) patent in the patent-paper pair. (The base coefficient is not estimated due to the patent fixed effects.) The negative and statistically-significant estimated coefficient on the interaction of female last author and percentage-of-male-inventors indicates that articles with a female author are less likely to be commercialized by a patent with more male inventors than its counterpart "twin" paper without any female authors. Columns (3-4) have similar results when controlling for the percentage of non-last female authors.

Table 14 about here

One implication of Table 14 is that women might commercialize more given a more gender-balanced profile of inventors at potential cooperation partners in industry. Although this result tentatively points toward a homophilic pattern, it is worth acknowledging that teams' compositions for collaboration may be endogenous. In fact, it could rely on (1) the firms wanting to reduce communication costs by having a less diverse gendered group given the gender of the author with whom the collaboration will occur (Reagans and Zuckerman 2001); or (2) this could also be an implicit or explicit preference of the author with whom the collaboration may occur.[20]

4.7 Additional discussion: The case of University patents in PPPs

Throughout the analysis, we have excluded PPPs composed of university patents that are not transitive patent-paper pairs (although, PPPs where the patent is assigned both to a university and to a corporation are included). Typically, these patents do not reflect commercialization activities unless they are accompanied by a license. However, due to the absence of detailed license data, a comprehensive analysis of licenses is not feasible, as explained in Section 2.1. In this section, we utilize university patents as a placebo group to strengthen our arguments and emphasize the distinctiveness of commercialization as a cooperative activity.

Column (1) of Appendix Table A7 shows that within the sample of twin articles, we do not observe a significant gender gap in the likelihood of converting science into a university patent. Furthermore,

20. Note that we do *not* see an interaction effect in Appendix Table A6 when simply measuring citations from patents to papers, which do not require cooperation between the inventors and authors.

in terms of magnitude, the gender gap is four times smaller than the gap observed in our baseline model of commercialization activity.

Additionally, column (4) reveals that using more self-promoting language does not increase the likelihood of obtaining a university patent. In other words, the impact of language choice appears to be more significant when collaborating with individuals outside academia. Particularly, firms seem to be sensitive to self-promotion, a phenomenon also observed in a broader industrial context (see Bolino et al. 2008 for insights on how self-promotion affects interviews, job prospects, and career outcomes).

Furthermore, the proportion of women does not influence the likelihood of obtaining a university patent (column 3). The reverse observation is only evident within our commercialization setting. This is in contrast to a role-model story where the percentage of females in the field should have increased the likelihood of undertaking patenting activities. It rather suggests that as a collaborative pattern, commercialization may be influenced by stereotypical behavior (Bordalo et al. 2019), wherein women in fields dominated by women are perceived to have greater potential for commercialization of their work.

Lastly, there is another potential mechanism that some may consider, which could be referred to as the commercialization "races". In fact, under this channel, women may be slower in terms of commercialization compared to men. However, it is important to note that this mechanism is not in line with most of our findings. Specifically, it sharply contrasts with our observations regarding open access and the potential influence of homophily in team formation, and also to some extent the representation of women in the fields. Although we cannot entirely dismiss this possibility, we interpret these results as evidence that the gender gap is less likely to be attributed to women "losing out" in potential commercialization races.

We conclude this section by presenting a fully-saturated model in Table A8. Interestingly, none of the additional controls accounted for the gender gap observed in the commercialization of academic research.

5 Implications

The gender gap we find in commercialization, as well as the associated mechanisms, are troubling on multiple levels. In this section, we assess the implications of our findings in three ways. First, we examine the relative importance of discoveries by women that remain uncommercialized. Second, we identify the fields of science with larger vs. smaller gaps. Third, we chart time trends to reveal whether the gender gap in commercialization is widening vs. shrinking.

5.1 Gender gap in commercialization by quality of scientific discovery

One might ask whether our results suggest any loss in welfare. In other words, are there truly any "Cassatts" in the "attic of science"? Or are the scientific discoveries by women that remain uncommercialized not of great importance?[21] The results in columns (3-4) of Table 2 indicate that science conducted by women is more highly cited, and columns (3-4) of Table 4 indicate no difference in citation rates to academic papers for twin papers with male vs. female last authors. Therefore, we think it unlikely that the gender gap represents no loss as far as a failure to commercialize valuable scientific discoveries. Nonetheless, in Table 16, we analyze commercialization rates for scientific articles according to their quality.

5.1.1 Methodology

We assess the quality of an article through four distinct proxies. The first proxy relies on the number of citations from academic articles, while the second is based on the number of patent citations. The third and fourth proxies utilize predictive algorithms to evaluate the article's quality in terms of commercial relevance.

We consider two types of data in our analysis: structured and unstructured. The structured data encompass various variables we have utilized thus far, including the number of authors, publication year, field, average authors' prominence, average affiliation prominence, journal impact factors, funding status, use of novel words, and the coauthorship network. Notably, we intentionally exclude citation measures to capture additional information about the potential for commercialization not already accounted for by citations. Additionally, we augment this dataset with unstructured data derived from the text content of the research papers, especially the abstracts and the titles.

Articles abstracts are highly appropriate for conducting this type of analysis. They offer a succinct overview of the main outcomes, methodology, conclusions, and significant contributions of the study, along with its distinguishing characteristics that may indicate its potential quality. As a result, abstracts are less prone to digressions and exhibit a more focused and structured format. This feature proves valuable in minimizing noise within the prediction task. Thus, on top of the computational constraints, this explains the extensive utilization of abstracts when analyzing academic research papers.

To incorporate the unstructured data, we leverage the advanced deep learning technique known as Bidirectional Encoder Representations from Transformers (BERT). BERT is considered a state-of-the-art approach in natural language processing (Devlin et al. 2018). It has been trained on extensive datasets such as books corpus and English Wikipedia, enabling it to excel in capturing intricate

21. In the case of twin-article analysis, as long as one of the twin articles is commercialized, one might argue that there is no loss to society as the underlying scientific discovery was brought to market. (One might nonetheless ask whether it is troubling that twin articles with female last authors are systemically commercialized less often.)

and nuanced text structures. The key feature of BERT lies in its enhanced ability to capture word context comprehensively. Unlike previous models trained in a unidirectional manner, BERT takes into account both the left and right context of a word simultaneously. This bidirectional training enables BERT to effectively capture the subtle contextual nuances and dependencies among words within a sentence, resulting in improved language comprehension. However, it is worth noting that the BERT algorithm can be computationally intensive. To address this issue, we adopt a pre-trained BERT model and randomly sample approximately 2.5 million observations, representing roughly 5% of the available data. Subsequently, the model without the language model will be applied to the entire dataset of MAG papers. In contrast, the model with the language model will be applied to a random 5% subsample. This approach allows us to balance computational considerations while still incorporating the benefits of the language model in our analysis.[22]

We train and fine-tune our models for the 56 distinct groups, employing stochastic gradient descent learning models and tree-based methods. After identifying the optimal model for each category, we observe that the average area under the curve (AUC) for both models, with and without the text data, is approximately 87%. However, the model with the text data and the language model demonstrates a 25% higher recall and a 20% higher F1 score compared to the model without the text data and the language model. This indicates that including text data is crucial for attaining improved accuracy levels.

5.1.2 Results

First, we begin by establishing the significance of employing our four metrics to evaluate quality, particularly in terms of commercial relevance. Table 15 showcases the relationship between the commercialization variable and the proxies, utilizing a randomly selected hold-out sample that was not used in the predictive models (as using the entire sample would introduce bias into the algorithm's actual performance). Article citations account for approximately 18% of the variation in commercialization, patent citations for 61%, the model without the language model for 27%, and the model with the language model for 33%. Although patent citations explain a substantially large variation of the commercialization variable, each of these proxies contains valuable additional information, as evidenced by columns (5) and (6). Hence, utilizing any of these proxies is a reasonable attempt to capture the paper's quality in terms of its potential for commercialization.[23]

Table 15 about here

In Panel A of Table 16, the sample is divided into deciles based on the number of forward citations

22. More details can be found in Appendix 6.

23. Furthermore, we conducted additional evaluations considering twin pairs, where one twin is commercialized, and the other is not. For approximately 75% of the uncommercialized twins, our predictive algorithms indicate a probability of commercialization over 50%, i.e., the uncommercialized twin could have as well gotten commercialized. This once again ensures the accuracy of our metrics.

from academic articles. Column (1) presents estimates for the commercialization of articles within the first five deciles, representing articles below the median, as 50% of the articles have no citations. The remaining columns progressively focus on higher deciles at and above the median. With a significant number of observations yielding statistically significant coefficient estimates in each column, our attention is directed toward the magnitude of the findings. Columns (1)-(6) in Panel A exhibit a consistent upward trend in the commercialization gap for papers with female last authors as the scientific quality improves.

Table 16 about here

Panel B demonstrates a comparable trend when examining patent citation deciles, suggesting the underlying invention's potential commercial relevance. Both of these trends are illustrated in Figure 6, with the academic-citations trend in Panel A and the patent-citation trend in Panel B.[24]

Moving on to Panel C of Table 16, the sample is divided into three subsamples based on the algorithm's predicted probability of commercialization. The first three columns correspond to predictions without the language model, while the last three pertain to predictions with the language model. Similar to Panel A and B, columns (1)-(3) and (4)-(6) of Panel C exhibit a consistent upward trend in the commercialization gap for papers with female last authors as the commercial potential of the scientific research increases.

We conclude that the gender gap in commercialization does not reflect the loss of "unimportant" scientific discoveries. If anything, the gap tends to widen as the scientific quality and commercial relevance of scientific discoveries increase. In other words, there are truly valuable contributions by female researchers in science that are not commercialized, i.e., there are genuinely "Cassatts" in the "attic" of science.[25]

Figure 6 about here

5.2 Gender gap in commercialization by scientific field

Appendix Tables A10 and A11 show where in the attic the "Cassatts" are, segmenting the scientific space into six top-level OECD categories: Natural Sciences, Engineering, Medical and Health Sciences, Agricultural Sciences Social Sciences, and Humanities. Each panel shows first the cross-section, and then the subset of twin papers for that category.

24. These trends for academic citations hold moreover in the subset of twin articles, and somewhat less robustly for patent citations (Appendix Table A9).

25. Because patent citations explain a significant variation in our commercialization variable, one could argue that commercialization may still occur through firms utilizing the paper via patent citations. To rule out this hypothesis and show that we might still have some important losses, we examine papers with less than one patent citation. We conduct the same analysis while incrementally increasing the values for the other two indicators. This is further supported by the findings in Table 15, which confirm that each of our indicators provides valuable information not captured by the others. Appendix Table A13 illustrates that when considering papers with minimal patent citations, the gender gap in commercialization is still more pronounced for papers with higher quality and commercial potential, as measured by the other indicators.

Although in the cross-section a gender gap appears in every category when applying the twins' methodology, this is limited to Natural Sciences and even then only achieves weak statistical significance. Of course, the twin estimates are considerably smaller subsets of the 27,551. We, therefore, examine the magnitude of the gap in the cross-section (columns (1) and (3) of each panel. The gender gap appears largest in the Natural Sciences and Engineering (Panels A and B of Table A10) as well as Agricultural Sciences (Panel A of Table A11). The gender gap is smaller in magnitude for Medical and Health Sciences (Panel C of Table A10), and smaller still for the Social Sciences and Humanities (Panels B and C of Table A11).

Overall, these results reveal that not all fields are guided by the same level of gender commercialization gap. In particular, the coefficient of social sciences could also testify of a non-systematic lesser likelihood to commercialize for women teams across all the fields (but note that the sample size for this field is meager in the twins' sample).

Tables A10 and A11 about here

5.3 Gender gap in commercialization over time

As more women participate in science, one would hope that any bias against women would begin to dissipate, particularly with the myriad of examples of women doing high-quality research. Thus, we could expect a reduction in the commercialization gap as time passes.

In Figure 7, we plot coefficients from Appendix Table A12, which re-estimates the baseline model but interacts a series of decade dummies with the indicator for a female last author as well as the percentage of female non-last authors. Panel A analyzes the entire Microsoft Academic Graph, and Panel B focuses on the subset of twin articles. In each panel, the predicted coefficients for female last author are in the left-hand graph and for the percent of non-last female author are on the right.

Figure 7 and Table A12 about here

Beginning with Panel A of Figure 7, which examines all articles, it appears that the gender gap increased from the 1980s through the first decade of the 2000s. This is true both for papers with female last authors and with a higher percentage of non-last female authors. Also in both cases, the gap narrows significantly—but has not entirely closed–in the past decade.

Panel B is restricted to twin articles. Because estimation is restricted to twin articles, interacting these with decade dummies results in less precisely estimated coefficients than in Panel A. Nonetheless, something of a trend is discernible. For papers with women as last author, the gap was somewhat imprecisely estimated in the 1980s (probably due to few twin papers from that decade). Estimates tightened in the two successive decades, showing a gap in the 1990s and early 2000s. During the last decade, it is difficult to establish any gap in commercialization when a woman is the last author.

The pattern for the percentage of non-last female authors suggests more of an enduring gap. Overall, Figure 7 is suggestive that the gender gap in commercialization has narrowed in the past decade.

6 Conclusion

We provide the first large-scale characterization of the gender dynamics underlying the commercialization of science. Analyzing more than 70 million articles from the Microsoft Academic Graph, we find that scientific teams with women—and especially women as last authors (i.e. lab managers)—suffer a commercialization penalty relative to all-male teams. This result is not explained either by the underlying quality of the science or by its commercial potential, which we establish by using "twin" discoveries. We hand-code the gender for every last author in our twins sample to increase confidence in our estimates. The gender gap is not a mechanical result of commercialized articles being cited more often by patents, nor is it explained by levels of commercialization activity at the authors' institutions or by the authors' prior commercialization. Finally, our results are robust not only to twin discoveries assembled via adjacent co-citation but also to identical biological sequences and structure.

What mechanisms explain this gap? We explore both supply-side and demand-side factors. We find only weak support for supply-side factors, including the representation of women in a particular scientific field, and also that women have fewer coauthorship ties to industry. Moreover, we establish that the gap does not exist for entrepreneurial self-commercialization via new ventures but only for cooperative commercialization in collaboration with existing firms. One question is whether firms simply pay less attention to women-led science. We leverage a natural experiment that shifted open access to scientific articles. Although open access seems to spur commercialization generally, easier access to scientific articles only exacerbates the gender gap—suggesting possible bias on the part of firms. One possibility is that firms pay less attention to women-led science because, as has been shown previously, women are less likely to use self-promoting language in their articles. We find that articles that "boast" are more likely to be commercialized, but this benefit is captured only when the lab head is male. Thus it does not appear to be simply due to word choice. It may be that the percentage of male inventors at potential commercialization partners may help to explain the gap. Papers with women as the last authors are less likely to be commercialized in a (potential) patent-paper pair with a higher percentage of male inventors on the patent, suggesting gender homophily in the construction of commercialization teams.

We lastly explore the implications of our findings. The gender gap is the largest among the highest-quality papers and those with the most commercial potential, suggesting negative welfare implications. In other words, the uncommercialized discoveries are not unimportant but may indeed represent "Cassatts" in the "attic" of science. The gap appears most pronounced in the natural sciences, and although the gap has narrowed in the last decade, it grew substantially from the 1980s through the

early 2000s.

To summarize, our findings tend to point toward a bias from the firms' side, not systematic across all fields, but with some stereotypical and homophilic features. The results of this analysis are keen in informing the public debate about the economic and welfare losses of such discriminatory behavior against scientific publications by women.

References

Angrist, Joshua D, and Jörn-Steffen Pischke. 2009. *Mostly harmless econometrics: An empiricist's companion.* Princeton university press.

Azoulay, Pierre, Waverly Ding, and Toby Stuart. 2007. "The determinants of faculty patenting behavior: Demographics or opportunities?" *Journal of Economic Behavior & Organization* 63 (4): 599–623.

Baker, Andrew C, David F Larcker, and Charles CY Wang. 2022. "How much should we trust staggered difference-in-differences estimates?" *Journal of Financial Economics* 144 (2): 370–395.

Beck, Nathaniel. 2020. "Estimating grouped data models with a binary-dependent variable and fixed effects via a logit versus a linear probability model: The impact of dropped units." *Political Analysis* 28 (1): 139–145.

Bercovitz, Janet, and Maryann Feldman. 2011. "The mechanisms of collaboration in inventive teams: Composition, social networks, and geography." *Research Policy* 40 (1): 81–93.

Bikard, Michaël. 2020. "Idea twins: Simultaneous discoveries as a research tool." *Strategic Management Journal* 41 (8): 1528–1543.

Bikard, Michaël, and Isabel Fernandez-Mateo. 2022. "Standing on the Shoulders of (Male) Giants: Gender Inequality and the Technological Impact of Scientific Ideas." *Available at SSRN 4059813.*

Bikard, Michaël, and Matt Marx. 2019. "Bridging academia and industry: How geographic hubs connect university science and corporate technology." *Management Science.*

Bjork, Bo-Christer, Annikki Roos, and Mari Lauri. 2009. "Scientific journal publishing: yearly volume and open access availability." *Information Research: An International Electronic Journal* 14 (1).

Bolino, Mark C, K Michele Kacmar, William H Turnley, and J Bruce Gilstrap. 2008. "A multi-level review of impression management motives and behaviors." *Journal of Management* 34 (6): 1080–1109.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. 2019. "Beliefs about gender." *American Economic Review* 109 (3): 739–73.

Brooks, Alison Wood, Laura Huang, Sarah Wood Kearney, and Fiona E Murray. 2014. "Investors prefer entrepreneurial ventures pitched by attractive men." *Proceedings of the National Academy of Sciences* 111 (12): 4427–4431.

Bryan, Kevin A, and Yasin Ozcan. 2021. "The impact of open access mandates on invention." *Review of Economics and Statistics* 103 (5): 954–967.

Callaway, Brantly, and Pedro HC Sant'Anna. 2021. "Difference-in-differences with multiple time periods." *Journal of Econometrics* 225 (2): 200–230.

Campbell, Lesley G, Siya Mehtani, Mary E Dozier, and Janice Rinehart. 2013. "Gender-heterogeneous working groups produce higher quality science." *PloS one* 8 (10): e79147.

Carlin, John B, Lyle C Gurrin, Jonathan AC Sterne, Ruth Morley, and Terry Dwyer. 2005. "Regression models for twin studies: a critical review." *International Journal of Epidemiology* 34 (5): 1089–1099.

Cohen, Wesley M, Richard R Nelson, and John P Walsh. 2002. "Links and impacts: the influence of public research on industrial R&D." *Management science* 48 (1): 1–23.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805.*

Ding, Waverly W, Fiona Murray, and Toby E Stuart. 2006. "Gender differences in patenting in the academic life sciences." *science* 313 (5787): 665–667.

Ducor, Philippe. 2000. "Coauthorship and coinventorship." *Science* 289 (5481): 873–875.

Eysenbach, Gunther. 2011. "Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact." *Journal of medical Internet research* 13 (4): e2012.

Fechner, Holly, and Matthew S Shapanka. 2018. "Closing diversity gaps in innovation: Gender, race, and income disparities in patenting and commercialization of inventions." *Technology & Innovation* 19 (4): 727–734.

Friesike, Sascha, Bastian Widenmayer, Oliver Gassmann, and Thomas Schildhauer. 2015. "Opening science: towards an agenda of open science in academia and industry." *The journal of technology transfer* 40:581–601.

Furman, Jeffrey L, and Scott Stern. 2011. "Climbing atop the shoulders of giants: The impact of institutions on cumulative research." *American Economic Review* 101 (5): 1933–1963.

Goldwasser, Eugene, Charles K-H Kung, and James Eliason. 1974. "On the mechanism of erythropoietin-induced differentiation: XIII. The role of sialic acid in erythropoietin action." *Journal of Biological Chemistry* 249 (13): 4202–4206.

Goodman-Bacon, Andrew. 2021. "Difference-in-differences with variation in treatment timing." *Journal of Econometrics* 225 (2): 254–277.

Gowers, Timothy, and Michael Nielsen. 2009. "Massively collaborative mathematics." *Nature* 461 (7266): 879–881.

Guzman, Jorge, and Aleksandra Olenka Kacperczyk. 2019. "Gender gap in entrepreneurship." *Research Policy* 48 (7): 1666–1680.

Haeussler, Carolin, and Henry Sauermann. 2013. "Credit where credit is due? The impact of project contributions and social factors on authorship and inventorship." *Research Policy* 42 (3): 688–703.

Hill, Ryan, and Carolyn Stein. 2019. "Scooped! Estimating rewards for priority in science." *Job Market Paper.*

Hug, Sven E, and Martin P Brändle. 2017. "The coverage of Microsoft Academic: Analyzing the publication output of a university." *Scientometrics* 113 (3): 1551–1571.

Khabsa, Madian, and C Lee Giles. 2014. "The number of scholarly documents on the public web." *PloS one* 9 (5): e93949.

Klar, Samara, Yanna Krupnikov, John Barry Ryan, Kathleen Searles, and Yotam Shmargad. 2020. "Using social media to promote academic research: Identifying the benefits of twitter for sharing academic work." *PloS one* 15 (4): e0229446.

Koffi, Marlène. 2021. "Gendered citations at top economic journals." In *AEA Papers and Proceedings,* 111:60–64.

Kolev, Julian, Yuly Fuentes-Medel, and Fiona Murray. 2019. *Is blinded review enough? How gendered outcomes arise even under anonymous evaluation.* Technical report. National Bureau of Economic Research.

Koning, Rembrand, Sampsa Samila, and John-Paul Ferguson. 2021. "Who do we invent for? Patents by women focus more on women's health, but few women get to invent." *Science* 372 (6548): 1345–1348.

Lach, Saul, and Mark Schankerman. 2008. "Incentives and invention in universities." *The RAND Journal of Economics* 39 (2): 403–433.

Lampe, Ryan. 2012. "Strategic citation." *Review of Economics and Statistics* 94 (1): 320–333.

Lerchenmueller, Marc J, Olav Sorenson, and Anupam B Jena. 2019. "Gender differences in how scientists present the importance of their research: observational study." *bmj* 367.

Markman, Gideon D, Donald S Siegel, and Mike Wright. 2008. "Research and technology commercialization." *Journal of Management Studies* 45 (8): 1401–1423.

Marx, Matt, and Aaron Fuegi. 2020. "Reliance on science: Worldwide front-page patent citations to scientific articles." *Strategic Management Journal.*

———. 2021. *Reliance on science by inventors: Hybrid extraction of in-text patent-to-article citations.* Technical report. National Bureau of Economic Research.

Marx, Matt, and David H Hsu. 2021. "Revisiting the Entrepreneurial Commercialization of Academic Science: Evidence from "Twin" Discoveries." *Management Science.*

McKiernan, Erin C, Philip E Bourne, C Titus Brown, Stuart Buck, Amye Kenall, Jennifer Lin, Damon McDougall, Brian A Nosek, Karthik Ram, Courtney K Soderberg, et al. 2016. "How open science helps researchers succeed." *elife* 5:e16800.

Miyake, Takaji, Charles K Kung, and Eugene Goldwasser. 1977. "Purification of human erythropoietin." *Journal of Biological Chemistry* 252 (15): 5558–5564.

Murray, Fiona. 2002. "Innovation as co-evolution of scientific and technological networks: exploring tissue engineering." *Research Policy* 31 (8-9): 1389–1403.

Murray, Fiona, Philippe Aghion, Mathias Dewatripont, Julian Kolev, and Scott Stern. 2016. "Of mice and academics: Examining the effect of openness on innovation." *American Economic Journal: Economic Policy* 8 (1): 212–252.

Murray, Fiona, and Leigh Graham. 2007. "Buying science and selling science: Gender differences in the market for commercial science." *Industrial and Corporate Change* 16 (4): 657–689.

Ranaei, Samira, Antti Knutas, Juho Salminen, and Arash Hajikhani. 2016. "Cloud-based Patent and Paper Analysis Tool for Comparative Analysis of Research." In *CompSysTech,* 315–322.

Reagans, Ray, and Ezra W Zuckerman. 2001. "Networks, diversity, and productivity: The social capital of corporate R&D teams." *Organization science* 12 (4): 502–517.

Reimers, Nils, and Iryna Gurevych. 2019. "Sentence-bert: Sentence embeddings using siamese bert-networks." *arXiv preprint arXiv:1908.10084.*

Rivette, Kevin G, and David Kline. 2000. *Rembrandts in the attic: Unlocking the hidden value of patents.* Harvard Business Press.

Rothaermel, Frank T, Shanti D Agung, and Lin Jiang. 2007. "University entrepreneurship: a taxonomy of the literature." *Industrial and Corporate Change* 16 (4): 691–791.

Sarsons, Heather. 2017. "Recognition for group work: Gender differences in academia." *American Economic Review* 107 (5): 141–45.

Sherwood, Judith B, and Eugene Goldwasser. 1979. "A radioimmunoassay for erythropoietin."

Sinha, Arnab, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Hsu, and Kuansan Wang. 2015. "An overview of microsoft academic service (mas) and applications." In *Proceedings of the 24th international conference on world wide web,* 243–246.

Staudt, Joseph. 2020. "Mandating access: assessing the NIH's public access policy." *Economic policy* 35 (102): 269–304.

Sun, Liyang, and Sarah Abraham. 2021. "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects." *Journal of Econometrics* 225 (2): 175–199.

Tartari, Valentina, and Ammon Salter. 2015. "The engagement gap:: Exploring gender differences in University–Industry collaboration activities." *Research Policy* 44 (6): 1176–1191.

Thompson, Neil C, Arvids A Ziedonis, and David C Mowery. 2018. "University licensing and the flow of scientific knowledge." *Research Policy* 47 (6): 1060–1069.

Thursby, Jerry G, and Marie C Thursby. 2005. "Gender patterns of research and licensing activity of science and engineering faculty." *The Journal of Technology Transfer* 30 (4): 343–353.

Ware, Mark, and Michael Mabe. 2015. "The STM report: An overview of scientific and scholarly journal publishing."

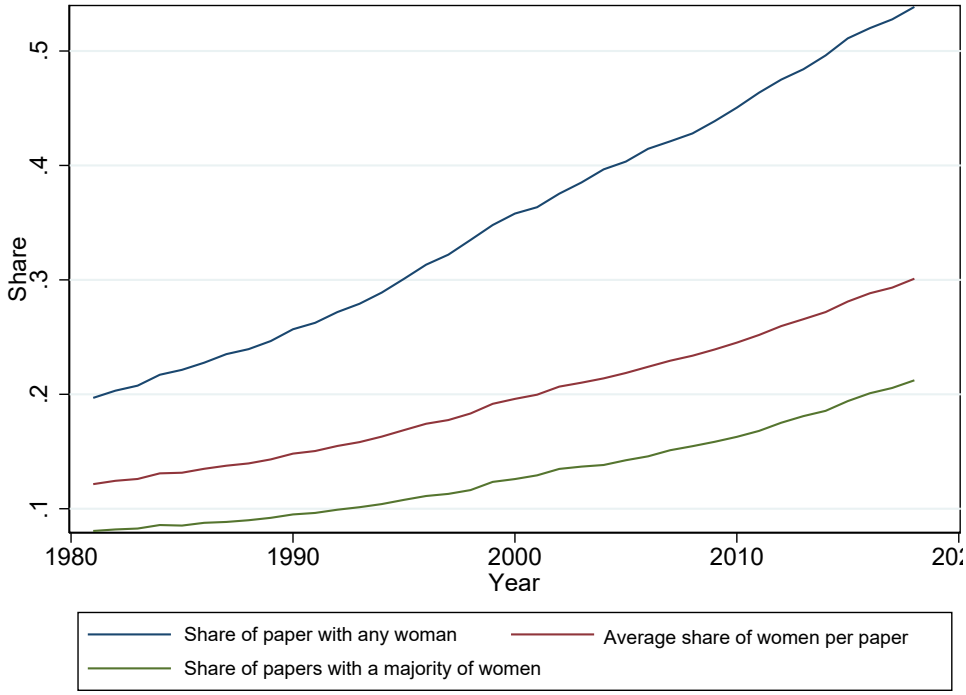
Whittington, Kjersten Bunker, and Laurel Smith-Doerr. 2005. "Gender and commercial science: Women's patenting in the life sciences." *The Journal of Technology Transfer* 30 (4): 355–370.

Williams, Heidi L. 2013. "Intellectual property rights and innovation: Evidence from the human genome." *Journal of Political Economy* 121 (1): 1–27.

Wuchty, Stefan, Benjamin F Jones, and Brian Uzzi. 2007. "The increasing dominance of teams in production of knowledge." *Science* 316 (5827): 1036–1039.

Yang, Yang, Teresa Woodruff, Yuan Tian, Benjamin F Jones, and Brian Uzzi. 2021. "Gender Diverse Teams Produce More Innovative and Influential 2 Ideas in Medical Research 3."

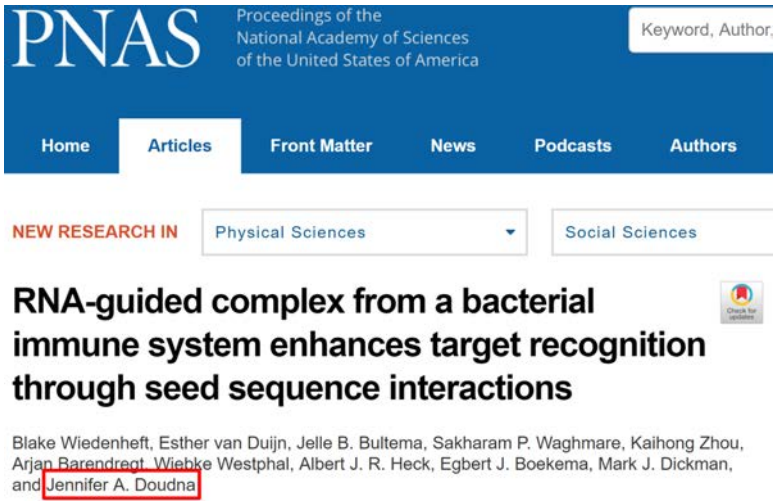
Figure 1: Representation of women in authorship of scientific articles, 1980-2020



Notes: Data are counts of published articles captured by the Microsoft Academic Graph, limited to those with Digital Object Identifiers. Author gender is determined via algorithm as described in Section 2.2.

Figure 2: Patent-paper pair example

Panel A: PNAS Article in the Patent-Paper Pair



Panel B: U.S. Patent 9,260,752 in the Patent-Paper Pair

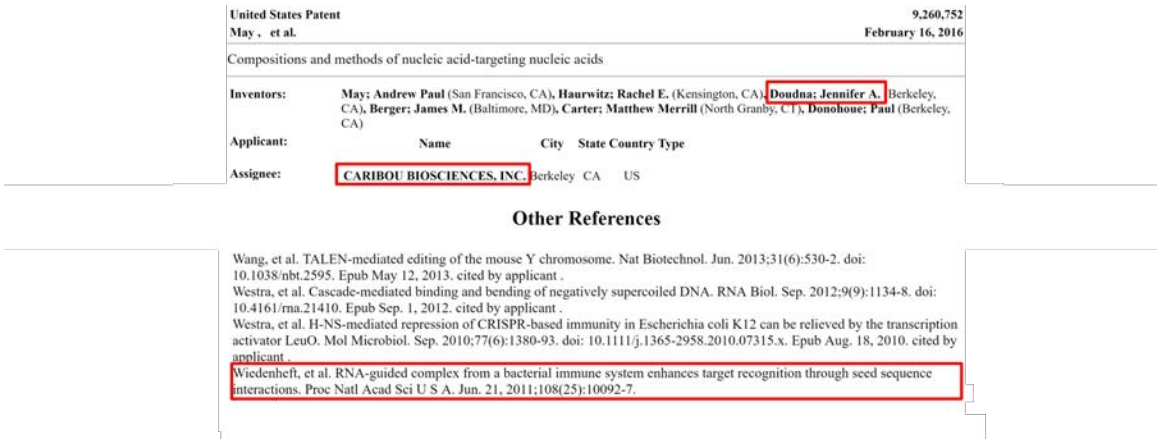
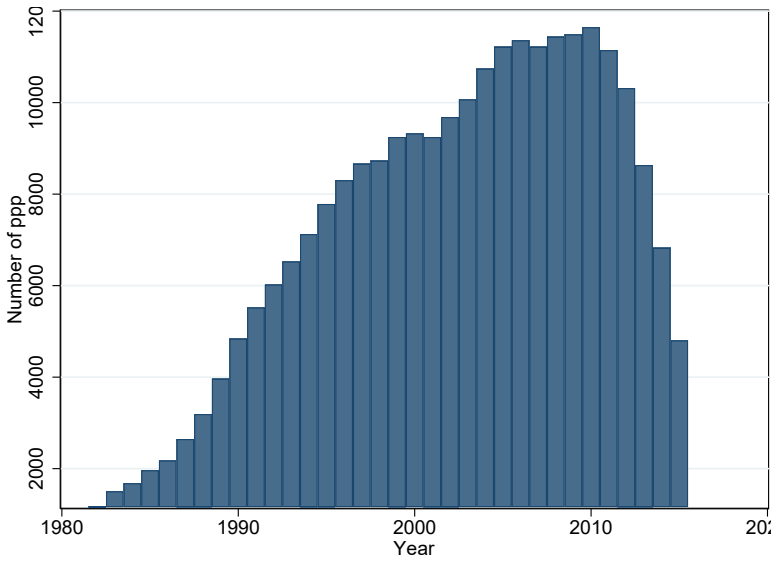
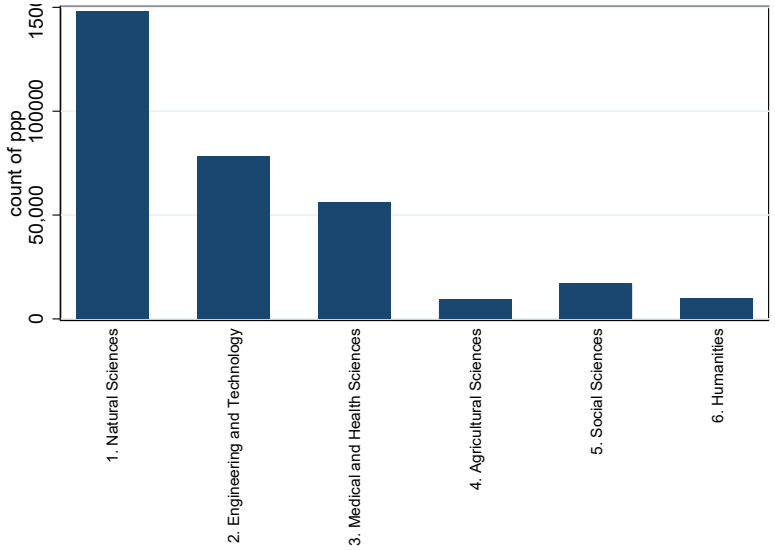


Figure 3: Commercialization rates over time



Notes: Count of commercialized papers per year as per our PPP methodology in Section 2.1.

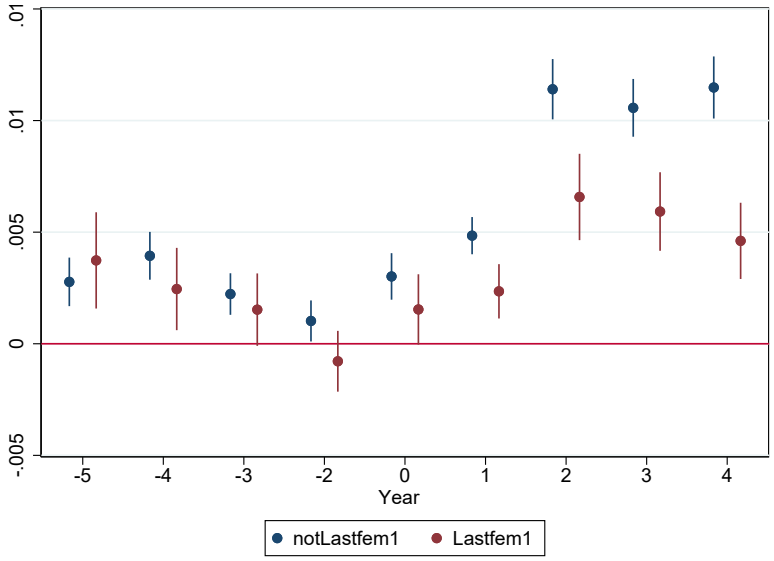
Figure 4: Commercialization rates by OECD category



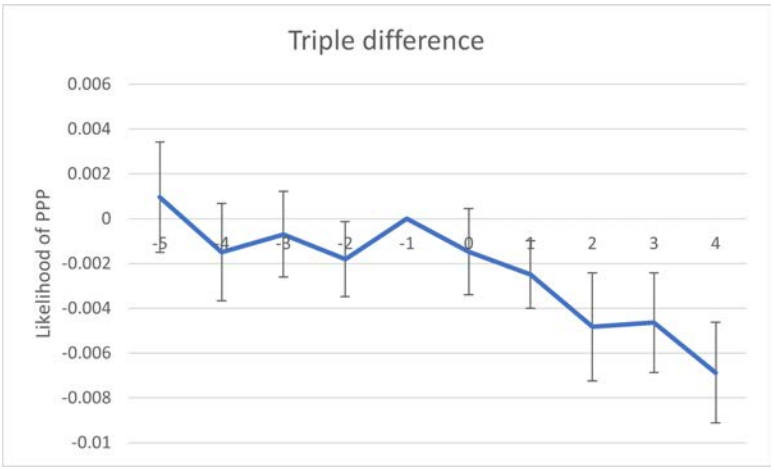
Notes: Count of commercialized papers as per our PPP methodology in Section 2.1, segmented by OECD top-level categories defined at http://www.oecd.org/science/inno/38235147.pdf.

Figure 5: Impact of Open Access mandates on commercialization

Panel A

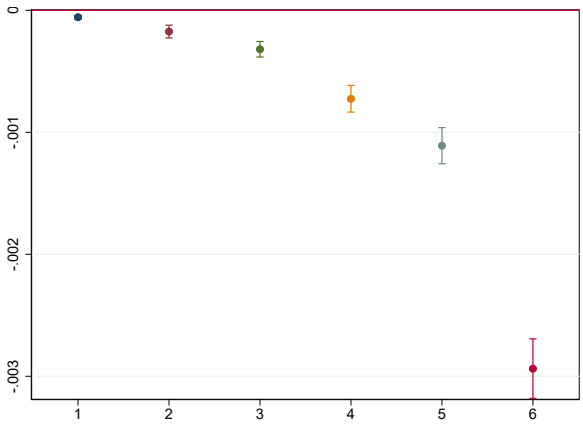


Panel B

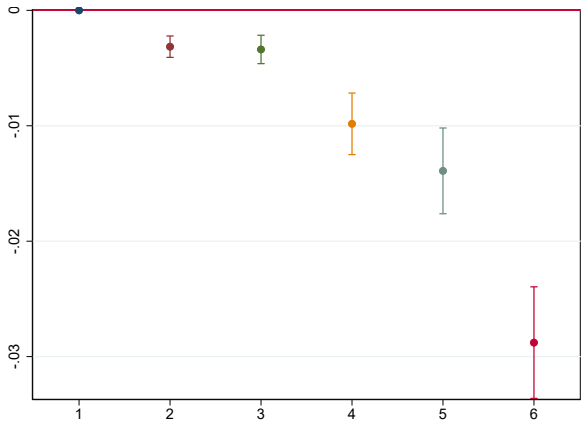


Notes: Figure 5 shows the estimation results of the staggered triple difference to assess the effect of Open Access on the gender gap in commercialization. Panel A shows the staggered difference-in-difference separately for male-authored and female-authored papers (papers with any woman). Panel B plots the triple difference results, therefore assessing the effect of the Open Access policy on the gender gap. The unit of observation is the academic article. The dependent variable is the commercialization measured by the patent-paper-pair whose assignee is a firm. All estimates include controls for the number of authors, the authors' average prominence and institutions, the fields, and years dummies. The coefficient for event time -1 is omitted to normalize the gender commercialization gap to zero in the year prior to the policy.

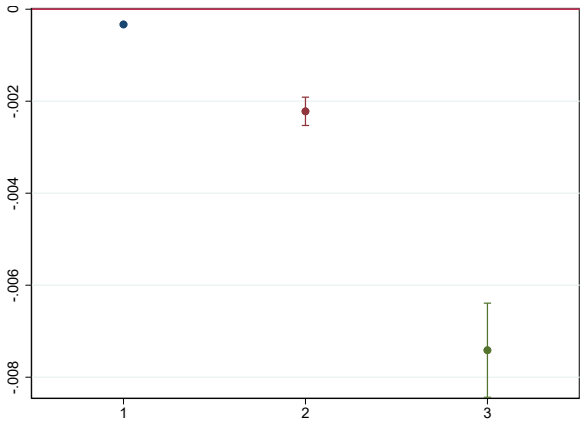
Figure 6: Assessing impact of the gender gap by paper quality



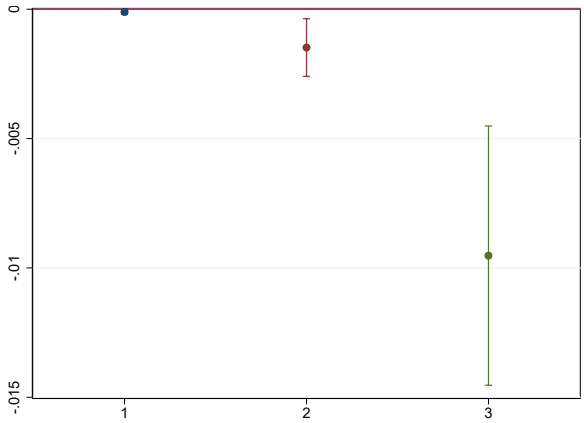
(a) Academic papers citations



(b) Patent citations



(c) Probability of commercialization (no language model)

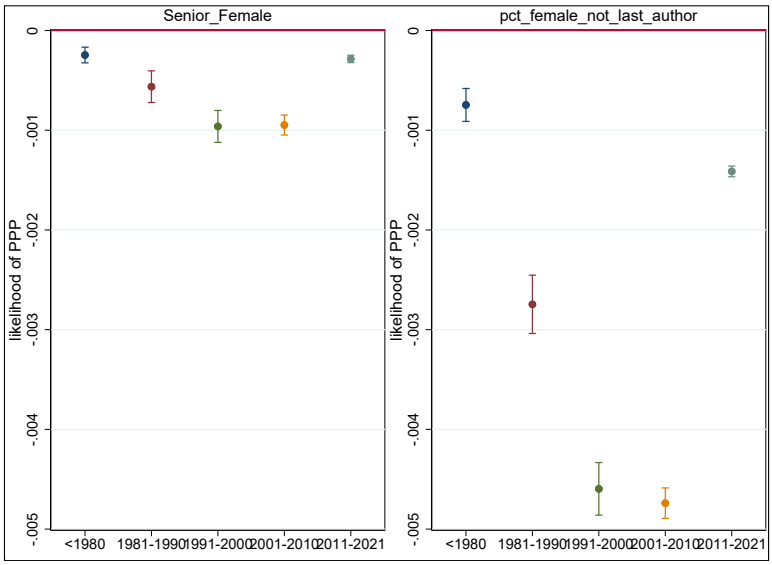


(d) Probability of commercialization (with language model)

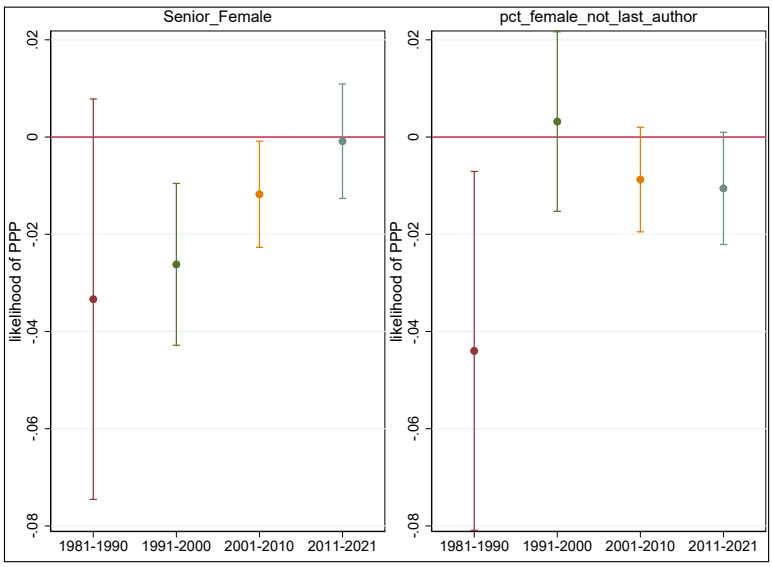
Notes: Figure 6 shows the estimation results of the gender dynamic in the commercialization of academic research splitting the sample into deciles according to the number of forward citations from other academic articles (plots (a) and (b)) and splitting the sample based on the probability of commercialization (plots (c) and (d)). The dependent variable is the patent-paper pairs while the independent variable of interest is having a female last author (PI). In plots (a) and (b), the blue bar represents the first five deciles (i.e., below the median as 50% of the paper does not have any citation), the red bar is the 60th decile, the green bar the 70th decile, the orange bar the 80th decile, the light green the 90th decile and the pink above the 90th of forward citations. Plot (a) shows the plot using forward citations, and Plot (b) uses patent citations. In plots (c) and (d), the blue bar, the red bar, and the green bar represent, respectively, articles with an estimated probability of commercialization below 33%, between 33% and 66% and above 66%. Plot (c) shows the plot using the predictive model with no text data and language model, and Plot (d) shows the plot using the predictive model with text data and language model.

Figure 7: Gender gap in commercialization over time

Panel A



Panel B



Notes: Figure 7 shows the estimation results of the gender dynamic in the commercialization of academic research over time. The dependent variable is the patent-paper pairs, while the independent variable of interest is having a female last author (plot on the left) and the percentage of female authors excluding the last author (plot on the right). Panel A shows the plot in the cross-sectional sample and Panel B in the twins' sample.

Table 1: Descriptive statistics for 70,016,266 articles

Variable	Obs	Mean	Std. Dev.	Min	Max
Commercialized	70016266	.0033	.0575	0	1
Cooperative Commercialization (i.e., w/established firm)	70016266	.0032	.0568	0	1
Self-Commercialization (i.e., via startup)	70016266	.0001	.009	0	1
Ln scientific citations (5-years, forward)	70016266	1.0963	1.239	0	11.4892
Ln patent citations (forward)	70016266	.0781	.3902	0	10.3432
Ln patent citations (forward, in-text)	70016266	.0317	.2357	0	9.9398
Ln patent citations (forward, front-page)	70016266	.0633	.336	0	9.2406
At least one female	70016266	.4147	.4927	0	1
\% female	70015225	.2446	.349	0	1
First author is female	59404819	.2636	.4406	0	1
First or last author is female	67854103	.3082	.4618	0	1
Female last author	60037015	.2173	.4124	0	1
\% female authors (not last)	60036555	.1721	.3083	-1	2
Pct female in field-year	60523991	.2332	.1065	0	1
Ln total prior coauthors at firms	70015446	2.4665	2.1861	0	11.4922
Paper has author at a firm	46687037	.0516	.2212	0	1
Paper has boastful words	70016266	.0873	.2823	0	1
Author previously commercialized	70016266	.1844	.3878	0	1
Ln num commercializations at institution(s)	70016266	.2689	.9251	0	8.997
Ln authors	70016266	1.3103	.5362	.6931	4.5951
Ln average citations per author	70016266	4.6333	2.9354	0	12.5171
Ln average citations per institution	70016266	9.0132	9.6503	0	20.9692
Ln Journal Impact Factor	70016266	.4833	.6099	0	6.7052

Notes: All articles reported in the Microsoft Academic Graph are included as long as the article contains a DOI. Count of observations for female first author, and female last author, is lower because many of these are not automatically classifiable at the 90 confidence level.

Table 2: Cross-sectional commercialization vs. citation

	Commercialized		Ln academic citations	
	(1)	(2)	(3)	(4)
Female last author=1	-0.000957***	-0.000568***	0.0328***	0.0255***
	(0.0000172)	(0.0000178)	(0.000313)	(0.000321)
% female authors (not last)		-0.00248***		0.0467***
		(0.0000280)		(0.000468)
Observations	51818941	51818473	51818941	51818473
year FE	y	y	y	y
article-field FE	y	y	y	y

Notes: This table shows the estimation results of the gender gap in commercialization on the cross-sectional data of MAG academic research papers. In columns (1)-(2), the left-hand side variable is the commercialization of a given paper as measured by the patent-paper-pair described in section 2.1. In columns (3)-(4), the left-hand side variable is the logarithm of academic citations over five years after the publication of the paper. All models include controls for the number of authors, the prestige of authors, the prestige of authors' institutions, and the journal impact factor. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 3: Descriptive statistics for twin articles

Variable	Obs	Mean	Std. Dev.	Min	Max
Commercialized	27551	.0451	.2075	0	1
Cooperative Commercialization (i.e., w/established firm)	27551	.0379	.191	0	1
Self-Commercialization (i.e., via startup)	27551	.0072	.0843	0	1
Ln scientific citations (5-years, forward)	27551	3.7074	1.2196	0	8.8364
Ln patent citations (forward)	27551	.8435	1.3138	0	9.2109
Ln patent citations (forward, in-text)	27551	.5501	1.0348	0	9.2014
Ln patent citations (forward, front-page)	27551	.6525	1.1066	0	7.1115
At least one female	27551	.5927	.4913	0	1
\% female	26164	.2582	.2674	0	1
First author is female	21033	.2933	.4553	0	1
First or last author is female	25743	.2848	.4513	0	1
Female last author	24224	.1635	.3698	0	1
\% female authors (not last)	23889	.276	.3135	-1	2
Pct female in field-year	27528	.247	.0768	.0098	.5783
Ln total prior coauthors at firms	26707	4.3042	1.8028	0	10.0858
Paper has author at a firm	26445	.045	.2073	0	1
Paper has boastful words	27551	.1479	.355	0	1
Author previously commercialized	27551	.5477	.4977	0	1
Ln num commercializations at institution(s)	27551	.6245	1.4455	0	8.4489

"Twin" articles are defined via adjacent co-citation as per Bikard 2020, including the public dataset from that article as well as from Marx and Hsu 2021. Count of observations for female first author is lower because many of these are not automatically classifiable at the 90 confidence level. Gender of last authors is classified by hand.

Table 4: Twins commercialization vs. citation

	Commercialized		Ln academic citations	
	(1)	(2)	(3)	(4)
Female last author=1	-0.0152**	-0.0146**	0.0193	0.0155
	(0.00512)	(0.00528)	(0.0161)	(0.0163)
% female authors (not last)		-0.0105+		0.0303
		(0.00579)		(0.0210)
Observations	21488	20806	21488	20806
twin FE	y	y	y	y

Notes: This table shows the estimation results of the gender gap in commercialization in the twins' sample defined in Section 3.3. The estimations include fixed effects for the twin scientific discovery. In columns (1)-(2), the left-hand side variable is the commercialization of a given paper as measured by the patent-paper-pair described in section 2.1. In columns (3)-(4), the left-hand side variable is the logarithm of academic citations over five years after the publication of the paper. All models include controls for the number of authors, the prestige of authors, the prestige of authors' institutions, and the journal impact factor. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 5: Robustness: alternative measures of gender composition

	Commercialized					
	(1)	(2)	(3)	(4)	(5)	(6)
At least one female	-0.00892*					
	(0.00366)					
% female		-0.0188**				
		(0.00583)				
First author is female=1			-0.00984+			
			(0.00584)			
First or last author is female=1				-0.00810*		
				(0.00398)		
Female last author=1					-0.0146**	
					(0.00528)	
% female authors (not last)					-0.0111+	
					(0.00579)	
Mixed-gender teams with male last author						-0.00584
						(0.00459)
Mixed-gender teams with female last author						-0.0222***
						(0.00648)
All Female						0.000634
						(0.00491)
Observations	26734	24956	16540	24114	20808	23014
twin FE	y	y	y	y	y	y

Notes: This table shows the estimation results of the gender gap in commercialization in the twins' sample defined in Section 3.3 for alternative definitions of the gender structure of the authors on a given article. The estimations include fixed effects for the twin scientific discovery. Column (1) measures the gender structure using a binary variable equal to 1 if at least one female author is on the team. Column (2) measures the gender structure using the percentage of female authors on the team. Column (3) measures the gender structure using a binary variable equal to 1 if the first author is female. Column (4) measures the gender structure using a binary variable equal to 1 if the first or the last author is female. Column (5) replicates the result from Table 4- Column (2) for convenience. Column (6) measures the gender structure using a four-dummy model with labels: all-male teams, mixed-gender teams with a male last author, mixed-gender teams with a female last author, and all-female teams. The reference category is all-male teams. All models include controls for the number of authors, the prestige of authors, the prestige of authors' institutions, and the journal impact factor. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 6: Robustness: by-hand gender classification and conditional logit

	Commercialized			
	Every author hand-coded		Conditional logit	
	(1)	(2)	(3)	(4)
main				
Female last author=1	-0.190***	-0.185***	-0.470**	-0.460**
	(0.0556)	(0.0557)	(0.152)	(0.152)
% female authors (not last)		-0.0975		-0.292
		(0.0785)		(0.209)
Constant	-1.217***	-1.164***		
	(0.198)	(0.204)		
Observations	1982	1982	1853	1853
twin FE	y	y	y	y

Notes: This table shows the estimation results of the gender gap in commercialization in the twins' sample defined in Section 3.3 for a subset of hand-collected data (columns (1)-(2)) and the conditional logistic regression (columns (3)-(4)). Every author was hand-coded for the subset of papers in columns (1-2) where one or the other twin in the simultaneous discovery was commercialized (twin discoveries where neither paper was commercialized are excluded). Conditional logit necessarily excludes these. The estimations include fixed effects for the twin scientific discovery. In columns (1)-(2), the left-hand side variable is the commercialization of a given paper as measured by the patent-paper-pair described in section 2.1. In columns (3)-(4), the left-hand side variable is the logarithm of academic citations over five years after the publication of the paper. All models include controls for the number of authors, the prestige of authors, the prestige of authors' institutions, and the journal impact factor. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 7: Placebo: patent citations to papers

	Ln citations from patents					
	Front-page		In-text		Front-page or in-text	
	(1)	(2)	(3)	(4)	(5)	(6)
Female last author=1	-0.0110	-0.00614	-0.0131	-0.0117	-0.0153	-0.0110
	(0.0140)	(0.0144)	(0.0111)	(0.0114)	(0.0151)	(0.0155)
% female authors (not last)		-0.0540**		-0.0197		-0.0477*
		(0.0175)		(0.0151)		(0.0196)
Observations	21490	20808	21490	20808	21490	20808
twin FE	y	y	y	y	y	y

Notes: This table shows the estimation results of the gender dynamic of patent citations in the twins' sample defined in Section 3.3. The estimations include fixed effects for the twin scientific discovery. In columns (1)-(2), the left-hand side variable is the front-page patent citations (likely to be legally binding). In columns (3)-(4), the left-hand side variable is the in-text patent citations (less likely to be legally binding and more likely to be added by the scientists). In columns (5)-(6), the left-hand side variable is the sum of front-page citations and in-text patent citations. Citations to front-page articles are from Marx and Fuegi 2020 and in-text citations are from Marx and Fuegi 2021. All models include controls for the number of authors, the prestige of authors, the prestige of authors' institutions, and the journal impact factor. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 8: Robustness: prior commercialization

	Commercialized					
	(1)	(2)	(3)	(4)	(5)	(6)
Female last author=1	-0.0138**	-0.0133*	-0.0152**	-0.0147**	-0.0138**	-0.0134*
	(0.00512)	(0.00531)	(0.00512)	(0.00531)	(0.00512)	(0.00531)
Author previously commercialized	0.0243***	0.0244***			0.0240***	0.0241***
	(0.00435)	(0.00448)			(0.00435)	(0.00449)
% female authors (not last)		-0.00896		-0.0102+		-0.00876
		(0.00587)		(0.00586)		(0.00587)
Ln num commercializations at institution(s)			0.00279+	0.00274	0.00257	0.00253
			(0.00169)	(0.00171)	(0.00169)	(0.00171)
Constant	-0.0523***	-0.0560***	-0.0677***	-0.0717***	-0.0493**	-0.0529**
	(0.0157)	(0.0170)	(0.0153)	(0.0166)	(0.0157)	(0.0171)
Observations	21488	20832	21488	20832	21488	20832
twin FE	y	y	y	y	y	y

Notes: This table shows the estimation results of the gender gap in commercialization in the twins' sample defined in Section 3.3. The estimations include fixed effects for the twin scientific discovery. The left-hand side variable is the commercialization of a given paper as measured by the patent-paper-pair described in section 2.1. Columns (1), (2), and (5) control for the history of commercialization of the authors in the paper. In particular, it is added a binary variable equal to 1 if the paper's authors have commercialized previously. Columns (3), (4), and (6) control for the history of commercialization of the authors' institution in the paper, which is measured by the logarithm of the total count of commercialization of a given institution. All models include controls for the number of authors, the prestige of authors, the prestige of authors' institutions, and the journal impact factor. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 9: Robustness: twins based on biological sequence and structure

	Commercialized			
	only 2 twins		all possible twins	
	(1)	(2)	(3)	(4)
Female last author=1	-0.0323**	-0.0317**	-0.00809*	-0.00664+
	(0.0113)	(0.0113)	(0.00393)	(0.00396)
% female authors (not last)		-0.0196		-0.0165**
		(0.0169)		(0.00515)
Constant	-0.00804	0.000735	-0.0794***	-0.0718***
	(0.0311)	(0.0320)	(0.0126)	(0.0128)
Observations	4304	4304	24724	24723
twin FE	y	y	y	y

Notes: This table shows the estimation results of the gender dynamic in commercialization for the subset of "twin" articles based on identical biological sequence and structure as defined in Section 3.4.4. The estimations include fixed effects for the twin scientific discovery. The left-hand side variable is the commercialization of a given paper as measured by the patent-paper-pair described in section 2.1. All models include controls for the number of authors, the prestige of authors, the prestige of authors' institutions, and the journal impact factor. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 10: Female scientists: commercialization and representation in scientific fields

	Commercialized			
	(1)	(2)	(3)	(4)
Female last author=1	-0.0154**	-0.0421*	-0.0148**	-0.0430*
	(0.00514)	(0.0174)	(0.00531)	(0.0187)
Pct female in field-year	0.0536	0.0317	0.0661+	0.0435
	(0.0377)	(0.0411)	(0.0375)	(0.0409)
Female last author=1 × Pct female in field-year		0.0993+		0.104+
		(0.0572)		(0.0614)
% female authors (not last)			-0.0113+	-0.0114*
			(0.00580)	(0.00580)
Observations	21460	21460	20778	20778
twin FE	y	y	y	y

Notes: This table shows the estimation results of the gender gap in commercialization in the twins' sample defined in Section 3.3, focusing on the effect of female representation in scientific fields. The estimations include fixed effects for the twin scientific discovery. All models include controls for the number of authors, the prestige of authors, the prestige of authors' institutions, and the journal impact factor. "Pct female in field-year" is defined as the share of authors publishing in the same field in that same year. Fields are determined by probabilistically crosswalking Microsoft Academic Graph keywords to 251 Web of Science categories. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 11: Commercialization and Networks

	Ln total prior coauthors at firms		Commercialized	
	(1)	(2)	(3)	(4)
Female last author=1	-0.118***	-0.138***	-0.0121*	-0.0155
	(0.0242)	(0.0260)	(0.00544)	(0.0129)
% female authors (not last)	-0.130***	-0.144***	-0.0127*	-0.0127*
	(0.0316)	(0.0337)	(0.00597)	(0.00597)
Ln total prior coauthors at firms			0.00145	0.00131
			(0.00236)	(0.00242)
Female last author=1 × Ln total prior coauthors at firms				0.000794
				(0.00295)
Observations	19954	17267	19954	19954
Twin FE	y	y	y	y
Focal paper has an author at a firm	y	n	y	y

Notes: This table shows the estimation results of the gender gap in commercialization in the twins' sample defined in Section 3.3, focusing on the industry network of the authors. The estimations include fixed effects for the twin scientific discovery. All models include controls for the number of authors, the prestige of authors, the prestige of authors' institutions, and the journal impact factor. "Ln last-author coauthors at firms" is the logarithm of the count of coauthors of the last author *not on the focal paper* with industrial affiliations. In columns (1)-(2), the left-hand side variable is the "Ln last-author coauthors at firms". In columns (3)-(4), the left-hand side variable is the commercialization of a given paper as measured by the patent-paper-pair described in section 2.1. Column (2) omits articles that have an industry author. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 12: Female scientists and commercialization "mode"

	Cooperatively Commercialized with Existing Firms		*Self*-Commercialized via Startups	
	(1)	(2)	(3)	(4)
Female last author=1	-0.0163***	-0.0157**	0.00109	0.00111
	(0.00489)	(0.00504)	(0.00183)	(0.00193)
% female authors (not last)		-0.0118*		0.000724
		(0.00556)		(0.00216)
Observations	21490	20808	21490	20808
twin FE	y	y	y	y

Notes: This table shows the estimation results of the gender gap in commercialization in the twins' sample defined in Section 3.3, focusing on the commercialization mode. The estimations include fixed effects for the twin scientific discovery. All models include controls for the number of authors, the prestige of authors, the prestige of authors' institutions, and the journal impact factor. "Cooperatively Commercialized" indicates that the article was part of a Patent-Paper Pair where the assignee was an existing/incumbent firm. "Self-commercialized" indicates that the assignee in the PPP was a new venture, as determined by merging assignees to PitchBook and ensuring that the article was published before the startup was founded (and no more than five years earlier). + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 13: Commercialization, gender, and attention

	Paper has boastful words		Ln academic citations		Commercialized	
	(1)	(2)	(3)	(4)	(5)	(6)
Female last author=1	-0.00274***	-0.00194	0.0156	0.00686	-0.0146**	-0.0134*
	(0.0000980)	(0.00934)	(0.0163)	(0.0179)	(0.00528)	(0.00565)
% female authors (not last)	-0.00972***	-0.0239*	0.0302	0.0303	-0.0108+	-0.0108+
	(0.000146)	(0.0111)	(0.0210)	(0.0210)	(0.00579)	(0.00579)
Paper has boastful words			-0.00709		0.0141*	
			(0.0170)		(0.00640)	
Paper has boastful words=1				-0.0162		0.0153*
				(0.0184)		(0.00708)
Paper has boastful words=1 × Female last author=1				0.0557		-0.00738
				(0.0420)		(0.0146)
Observations	51818473	20808	20808	20808	20808	20808
Twin FE	n	y	y	y	y	y
year FE	y	n	n	n	n	n
article-field FE	y	n	n	n	n	n

Notes: This table shows the estimation results of the gender gap in commercialization in the twins' sample defined in Section 3.3 (except column (1), which is using all the publications in MAG), focusing on self-promotion. Self-promotion is measured by the use of "boastful words". "Paper has boastful words" indicates that the title or abstract uses one or more words such as "breakthrough" which are defined by Lerchenmueller, Sorenson, and Jena 2019 as boasting, with the exception that "novel" is not treated as a boasting word when it appears in a bigram with "coronavirus." In columns (1)-(2), the left-hand side variable is "Paper has boastful words". In columns (3)-(4), the left-hand side variable is the logarithm of academic citations over five years after the publication of the paper. In columns (5)-(6), the left-hand side variable is the commercialization of a given paper as measured by the patent-paper-pair described in section 2.1. The estimations include fixed effects for the twin scientific discovery. All models include controls for the number of authors, the prestige of authors, the prestige of authors' institutions, and the journal impact factor. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 14: Commercialization and gender homophily (Counterfactual paper-patent twin dyads)

	Commercialized			
	(1)	(2)	(3)	(4)
Female last author=1	-0.195***	0.511***	-0.145**	0.559***
	(0.0440)	(0.153)	(0.0464)	(0.155)
Female last author=1 × Pct male inventors		-0.887***		-0.886***
		(0.180)		(0.180)
% female authors (not last)			-0.268***	-0.268***
			(0.0677)	(0.0686)
Constant	-1.587***	-1.489***	-1.392***	-1.293***
	(0.128)	(0.132)	(0.135)	(0.138)
Observations	2849	2849	2834	2834
twin-paper FE	y	y	y	y
citing patent FE	y	y	y	y

Notes: This table shows the estimation results of the gender gap in commercialization in the twins' sample defined in Section 3.3, focusing on homophily in team composition. In particular, the table estimates Equation 3, which instead of paper-level analysis, performs patent-paper level analysis of *possible* PPPs. The sample is limited to twin discoveries where one or the other twin is commercialized. The commercializing patent in the realized PPP is then artificially paired with the uncommercialized article in the twin to create a counterfactual PPP, given the intuition that the uncommercialized article in the twin discovery might well have been paired with the patent that commercialized the other article in the twin. The percentage of male inventors on the focal patent is calculated using USPTO's inventor-gender file, but excludes inventors who are also on the paper in order to avoid biasing toward realized PPPs. All models include controls for the number of authors, the prestige of authors, the prestige of authors' institutions, and the journal impact factor. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 15: Commercialization and proxies for quality and commercialization relevance

	Commercialized					
	(1)	(2)	(3)	(4)	(5)	(6)
Ln scientific citations (5-years, forward)	0.122***				0.003***	0.005***
	(0.001)				(0.000)	(0.000)
Ln patent citations (forward)		0.293***			0.266***	0.255***
		(0.001)			(0.001)	(0.001)
Probability of commercialization without language model			0.799***		0.191***	
			(0.003)		(0.003)	
Probability of commercialization with language model				0.910***		0.241***
				(0.003)		(0.003)
N	264374	264374	264374	264374	264374	264374
AdjR-sqr	0.178	0.611	0.269	0.329	0.623	0.627

Notes: This table shows the correlation between the probability of commercialization and the different proxies of paper quality and commercialization relevance. The proxy in Column (1) is academic article citations. The proxy in Column (2) is academic patent citations. The proxy in Column (3) is the probability of commercialization found by the predictive algorithm without the text data and the language model. The proxy in Column (4) is the probability of commercialization found by the predictive algorithm with the text data and the language model. Columns (5) and (6) put together all the proxies. The sample is limited to a hold-out sample, a randomly chosen sample unseen by the different algorithms in the estimation stage.

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

Table 16: Welfare?

Panel A: Commercialization gap by increasing academic citation percentile (full)

	Commercialization					
Academic citation decile	*0-49*	*50-59*	*60-69*	*70-79*	*80-89*	*90-100*
	(1)	(2)	(3)	(4)	(5)	(6)
Female last author=1	-0.0000691***	-0.000216***	-0.000383***	-0.000734***	-0.00118***	-0.00302***
	(0.00000766)	(0.0000269)	(0.0000324)	(0.0000555)	(0.0000741)	(0.000121)
% female authors (not last)	-0.000332***	-0.000642***	-0.00105***	-0.00170***	-0.00225***	-0.00600***
	(0.0000162)	(0.0000424)	(0.0000445)	(0.0000725)	(0.0000956)	(0.000152)
Observations	20706226	5544756	8352129	5558386	5685858	5971082
year FE	y	y	y	y	y	y
article-field	y	y	y	y	y	y

Panel B: Commercialization gap by increasing patent citation percentile (full)

	Commercialized					
Patent citation decile	*0-49*	*50-59*	*60-69*	*70-79*	*80-89*	*90-100*
	(1)	(2)	(3)	(4)	(5)	(6)
Female last author=1	0.000000121	-0.00329***	-0.00398***	-0.00994***	-0.0142***	-0.0295***
	(0.000000329)	(0.000443)	(0.000585)	(0.00128)	(0.00178)	(0.00232)
% female authors (not last)	-0.000000568	-0.00752***	-0.0131***	-0.0153***	-0.0188***	-0.0386***
	(0.000000467)	(0.000540)	(0.000672)	(0.00149)	(0.00208)	(0.00272)
Observations	48798988	958096	1117796	404857	283410	255209
year FE	y	y	y	y	y	y
article-field	y	y	y	y	y	y

Panel C: Commercialization gap by increasing probability of commercialization

	Commercialized					
	Prediction without language model			Prediction with language model		
Probability of commercialization	*0-0.33*	*0.33-0.66*	*0.66-1*	*0-0.33*	*0.33-0.66*	*0.66-1*
	(1)	(2)	(3)	(4)	(5)	(6)
Female last author=1	-0.000331***	-0.00222***	-0.00741***	-0.000105*	-0.00148**	-0.00953***
	(0.0000121)	(0.000157)	(0.000522)	(0.0000510)	(0.000568)	(0.00256)
% female authors (not last)	-0.000874***	-0.00437***	-0.00977***	-0.000569***	-0.00310***	-0.00974**
	(0.0000198)	(0.000190)	(0.000731)	(0.0000696)	(0.000708)	(0.00328)
Observations	45696200	4844813	1277462	1780702	250524	51005
year FE	y	y	y	y	y	y
article-field	y	y	y	y	y	y

Notes: This table analyzes the possible missed innovation due to the gender gap in commercialization. The left-hand side variable is the commercialization of a given paper as measured by the patent-paper-pair described in section 2.1. Each column of panel A and B estimates a subset of the 70,016,266 based on citation-count decile, with column (1) collapsing below-median deciles as most of these papers are never cited. Panel C shows a similar estimation based on groups of increasing probability of commercialization. All models include controls for the number of authors, the prestige of authors, the prestige of authors' institutions, and the journal impact factor.+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A1: Top 15 Commercializing Firms

1. International Business Machines Corporation
2. Semiconductor Energy Laboratory Co., Ltd
3. Microsoft Corporation
4. Ignis Innovation Inc.
5. Genentech, Inc.
6. Immunomedics, Inc.
7. Intel Corporation
8. Abbot Diabetes Care Inc.
9. Yeda Research and Development Co. Ltd.
10. Bristol-Myers Squibb Company
11. Immunex Corporation
12. Google Inc.
13. Hewlett-Packard
14. Lucent Technologies Inc.
15. Schlumberger Technology

Table A2: Robustness: Alternative percentage of known authors

Panel A: Twins

	Percentage of known authors is greater than				
	0%	33%	50%	75%	100%
	(1)	(2)	(3)	(4)	(5)
Female last author=1	-0.0143**	-0.0150**	-0.0156**	-0.0216*	-0.0493**
	(0.00533)	(0.00563)	(0.00602)	(0.00866)	(0.0172)
% female authors (not last)	-0.0111+	-0.0126*	-0.0146*	-0.0231*	-0.0394*
	(0.00586)	(0.00636)	(0.00702)	(0.0110)	(0.0183)
Observations	20842	19696	18284	12779	6513
twin FE	y	y	y	y	y

Panel B: All papers

	Percentage of known authors is greater than				
	0%	33%	50%	75%	100%
	(1)	(2)	(3)	(4)	(5)
Female last author=1	-0.000547***	-0.000568***	-0.000604***	-0.000713***	-0.000587***
	(0.0000178)	(0.0000178)	(0.0000179)	(0.0000183)	(0.0000167)
% female authors (not last)	-0.00234***	-0.00248***	-0.00262***	-0.00290***	-0.00225***
	(0.0000276)	(0.0000280)	(0.0000285)	(0.0000318)	(0.0000317)
Observations	52658916	51817669	50247100	42991426	37245527
twin FE	y	y	y	y	y

Notes: all models include controls for number of authors, prestige of authors, prestige of authors' institutions, and journal impact factor. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A3: Robustness: Alternative threshold for auto-generated gender

Panel A: twins%

	Auto-gender threshold			
	50%	75%	90%	95%
	(1)	(2)	(3)	(4)
Female last author=1	-0.00777+	-0.00873+	-0.0121*	-0.0122*
	(0.00471)	(0.00477)	(0.00536)	(0.00575)
% female authors (not last)	-0.00628	-0.00796	-0.00753	-0.00814
	(0.00621)	(0.00633)	(0.00699)	(0.00743)
Observations	21378	20630	17388	15958
twin FE	y	y	y	y

Panel B: cross-section%

	Auto-gender threshold			
	50%	75%	90%	95%
	(1)	(2)	(3)	(4)
Female last author=1	-0.000610***	-0.000631***	-0.000676***	-0.000714***
	(0.0000172)	(0.0000173)	(0.0000175)	(0.0000180)
% female authors (not last)	-0.00233***	-0.00235***	-0.00240***	-0.00245***
	(0.0000264)	(0.0000267)	(0.0000281)	(0.0000290)
Observations	55800878	55022560	51807745	49083539
twin FE	n	n	n	n

Notes: all models include controls for number of authors, prestige of authors, prestige of authors' institutions, and journal impact factor. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A4: Exclude transitive PPP

	Commercialization	
	(1)	(2)
Female last author=1	-0.0131**	-0.0125*
	(0.00496)	(0.00512)
% female authors (not last)		-0.00984+
		(0.00557)
Observations	21494	20812
Twin FE	y	y

Notes: all models include controls for number of authors, prestige of authors, prestige of authors' institutions, and journal impact factor * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A5: Accessibility of articles and commercialization

	Commercialized		
	Female last author	*Male last author*	*Triple difference*
	(1)	(2)	(3)
-5	0.00347997	0.00277673	0.00070324
	(0.00108522)	(0.00055488)	(0.0012188)
-4	0.00245346	0.00394614	-0.00149268
	(0.00094052)	(0.00054633)	(0.0010876)
-3	0.00152928	0.00221253	-0.00068325
	(0.00082808)	(0.00047442)	(0.0009543)
-2	-0.00078295	0.00100215	-0.0017851
	(0.00069476)	(0.00046824)	(0.00083781)
-1	0	0	0
	0	0	0
0	0.00153915	0.00299787	-0.00145872
	(0.00080394)	(0.00053096)	(0.0009634)
1	0.00235096	0.0048119	-0.00246094
	(0.00062102)	(0.0004248)	(0.0007524)
2	0.00650631	0.01138847	-0.00488216
	(0.00098411)	(0.00068883)	(0.0012012)
3	0.00592582	0.01053731	-0.00461149
	(0.00089778)	(0.00065969)	(0.00111409)
4	0.00461124	0.0114688	-0.00685756
	(0.00087099)	(0.00070895)	(0.0011230)

Notes: Table estimates likelihood of commercialization depending on open-access status of the article using the natural experiment described in Section 4.4. All models include controls for number of authors, prestige of authors, prestige of authors' institutions, and journal impact factor. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A6: Commercialization and gender homophily (Paper-patent citation dyads)

	Commercialized			
	(1)	(2)	(3)	(4)
Female last author=1	0.261***	0.262***	0.292***	0.297***
	(0.0152)	(0.0112)	(0.0154)	(0.0113)
Female last author=1 × Pct male inventors	0.0148	0.0132	0.0254	0.0192
	(0.0175)	(0.0125)	(0.0177)	(0.0126)
% female authors not last adjusted			-0.199***	-0.199***
			(0.00815)	(0.00705)
Constant	-1.228***	-1.255***	-1.143***	-1.191***
	(0.121)	(0.0170)	(0.121)	(0.0176)
Observations	262710	266674	261464	265781
twin-paper FE	y	y	y	y
citing patent FE	y	y	y	y

Notes: all models include controls for number of authors, prestige of authors, prestige of authors' institutions, and journal impact factor * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A7: Patent-paper pairs where the patent is assigned only to a university

	Commercialization		
	(1)	(2)	(3)
Female last author=1	-0.00326	-0.00322	-0.00327
	(0.00391)	(0.00392)	(0.00391)
% female authors (not last)	-0.00305	-0.00302	-0.00327
	(0.00441)	(0.00442)	(0.00441)
Pct female in field-year		-0.00511	
		(0.0268)	
paperhasnovelwords			-0.00935+
			(0.00495)
Observations	20806	20776	20806
Twin FE	y	n	y

Notes: all models include controls for number of authors, prestige of authors, prestige of authors' institutions, and journal impact factor. Column (1) includes twins fixed effects. Column (2) does not include twins fixed effects. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A8: Commercialization and gender: Fully saturated model

	Commercialized		
	(1)	(2)	(3)
Female last author=1	-0.0146**	-0.0136*	-0.0506+
	(0.00528)	(0.00557)	(0.0267)
% female authors (not last)	-0.0111+	-0.0104+	-0.0108+
	(0.00579)	(0.00600)	(0.00603)
Author previously commercialized		0.0202***	0.0226***
		(0.00455)	(0.00495)
Ln num commercializations at institution(s)		0.00244	0.00299
		(0.00176)	(0.00191)
Ln prior coauthors of last author at firms		-0.00268	-0.00279
		(0.00168)	(0.00184)
Paper has boastful words		0.0137*	0.0153*
		(0.00649)	(0.00718)
Pct female in field-year		0.0619	0.0384
		(0.0386)	(0.0422)
Female last author=1 × Author previously commercialized			-0.0146
			(0.0120)
Female last author=1 × Ln num commercializations at institution(s)			-0.00348
			(0.00420)
Female last author=1 × Ln prior coauthors of last author at firms			0.000570
			(0.00352)
Female last author=1 × Paper has boastful words			-0.00902
			(0.0150)
Female last author=1 × Pct female in field-year			0.0979
			(0.0632)
Female last author=1 × Ln authors			0.0111
			(0.0135)
Constant	-0.0733***	-0.0782***	-0.0713***
	(0.0164)	(0.0206)	(0.0211)
Observations	20808	19914	19914
Twin FE	y	y	y
year FE	n	n	n
article-field FE	n	n	n

Notes: This table shows the estimation results of the gender gap in commercialization in the twins' sample defined in Section 3.3. The estimations include fixed effects for the twin scientific discovery. All models include controls for the number of authors, the prestige of authors, the prestige of authors' institutions, and the journal impact factor. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A9: Welfare

Panel A: Commercialization gap by increasing academic citation percentile (Twins)

	Commercialization			
	Below-median citations		*Above-median citations*	
	(1)	(2)	(3)	(4)
Female last author=1	-0.00264	-0.00184	-0.0299*	-0.0301*
	(0.00404)	(0.00427)	(0.0118)	(0.0119)
% female authors (not last)		-0.00989+		0.00232
		(0.00508)		(0.0128)
Constant	-0.0162	-0.0150	-0.198***	-0.217***
	(0.0115)	(0.0127)	(0.0520)	(0.0534)
Observations	8684	8250	9289	9145
twin FE	y	y	y	y

Panel B: Commercialization gap by increasing patent citation percentile (twins)

	Commercialization			
	Below-median citations		*Above-median citations*	
	(1)	(2)	(3)	(4)
Female last author=1	0.00129	0.00158	-0.0365*	-0.0340*
	(0.000946)	(0.00106)	(0.0167)	(0.0171)
% female authors (not last)		-0.00304*		-0.0267
		(0.00147)		(0.0200)
Constant	-0.00389	-0.00356	-0.262***	-0.290***
	(0.00382)	(0.00406)	(0.0673)	(0.0728)
Observations	10578	10132	6869	6733
twin FE	y	y	y	y

Notes: all models include controls for number of authors, prestige of authors, prestige of authors' institutions, and journal impact factor. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A10: Field

Panel A: Natural Sciences

	Commercialization			
	(1)	(2)	(3)	(4)
Female last author=1	-0.00158***	-0.0200*	-0.00112***	-0.0203*
	(0.0000403)	(0.00974)	(0.0000418)	(0.0101)
% female authors (not last)			-0.00280***	-0.0118
			(0.0000585)	(0.0105)
Observations	17650196	8325	17649968	8081
twin FE	n	y	n	y
year FE	y	n	y	n

Panel B: Engineering and Technology

	Commercialization			
	(1)	(2)	(3)	(4)
Female last author=1	-0.00152***	-0.0169	-0.000954***	-0.0137
	(0.0000679)	(0.0434)	(0.0000709)	(0.0438)
% female authors (not last)			-0.00307***	-0.0436
			(0.0000950)	(0.0501)
Observations	7275551	938	7275521	922
twin FE	n	y	n	y
year FE	y	n	y	n

Panel C: Medical and Health Sciences

	Commercialization			
	(1)	(2)	(3)	(4)
Female last author=1	-0.000629***	-0.0101	-0.000318***	-0.00966
	(0.0000273)	(0.00871)	(0.0000283)	(0.00912)
% female authors (not last)			-0.00202***	-0.0143
			(0.0000457)	(0.0103)
Observations	13226409	3952	13226277	3814
twin FE	n	y	n	y
year FE	y	n	y	n

Notes: all models include controls for number of authors, prestige of authors, prestige of authors' institutions, and journal impact factor. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A11: Field

Panel A: Agricultural Sciences

	Commercialization			
	(1)	(2)	(3)	(4)
Female last author=1	-0.00113***	-0.0725	-0.000889***	-0.0785
	(0.000142)	(0.0524)	(0.000149)	(0.0550)
% female authors (not last)			-0.00136***	0.0558
			(0.000200)	(0.0766)
Observations	1185879	472	1185848	446
twin FE	n	y	n	y
year FE	y	n	y	n

Panel B: Social Sciences

	Commercialization			
	(1)	(2)	(3)	(4)
Female last author=1	-0.000339***	0.000550	-0.000186***	0.00318
	(0.0000181)	(0.0204)	(0.0000185)	(0.0214)
% female authors (not last)			-0.00109***	-0.00832
			(0.0000362)	(0.0229)
Observations	9699613	830	9699580	808
twin FE	n	y	n	y
year FE	y	n	y	n

Panel C: Humanities

	Commercialization			
	(1)	(2)	(3)	(4)
Female last author=1	-0.000435***	-0.0417	-0.000216***	-0.0579
	(0.0000410)	(0.0870)	(0.0000413)	(0.110)
% female authors (not last)			-0.00193***	-0.238
			(0.000101)	(0.345)
Observations	2781289	62	2781275	56
twin FE	n	y	n	y
year FE	y	n	y	n

Notes: all models include controls for number of authors, prestige of authors, prestige of authors' institutions, and journal impact factor. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A12: Gender gap in commercialization over time

	Commercialization			
	(1)	(2)	(3)	(4)
Female last author=1 × year19801989=1	-0.000846***	-0.00134***	-0.0539	-0.0543
	(0.0000806)	(0.0000785)	(0.133)	(0.0352)
Female last author=1 × year19901999=1	-0.00319***	-0.00368***	-0.0414	-0.0418**
	(0.0000855)	(0.0000836)	(0.129)	(0.0160)
Female last author=1 × year20002009=1	-0.00244***	-0.00293***	-0.00915	-0.00949
	(0.0000544)	(0.0000513)	(0.128)	(0.0131)
Female last author=1 × year2010plus=1	0.000492***		0.000339	
	(0.0000303)		(0.128)	
year19801989=1 × % female authors (not last)	0.00306***	0.00263***	-0.0764	-0.0727*
	(0.000149)	(0.000139)	(0.129)	(0.0365)
year19901999=1 × % female authors (not last)	0.00659***	0.00616***	0.0103	0.0140
	(0.000142)	(0.000131)	(0.126)	(0.0189)
year20002009=1 × % female authors (not last)	0.00234***	0.00191***	-0.00734	-0.00366
	(0.0000886)	(0.0000703)	(0.125)	(0.0169)
year2010plus=1 × % female authors (not last)	0.000426***		-0.00367	
	(0.0000614)		(0.125)	
Female last author=1 × year1979pre=1		-0.000492***		-0.000339
		(0.0000303)		(0.128)
year1979pre=1 × % female authors (not last)		-0.000426***		0.00367
		(0.0000614)		(0.125)
Observations	51818473	51818473	20834	20834
Twin FE	n	n	y	y
Article-field FE	y	y	n	n

Notes: Decade dummies and female last-author base variables are not shown. All models include controls for number of authors, prestige of authors, prestige of authors' institutions, and journal impact factor + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A13: Welfare given at most one patent citation

Panel A: Commercialization gap by increasing academic citation percentile (full)

	Commercialization					
Academic citation decile	*0-49*	*50-59*	*60-69*	*70-79*	*80-89*	*90-100*
	(1)	(2)	(3)	(4)	(5)	(6)
Female last author=1	-0.0000169***	-0.0000545***	-0.0000856***	-0.000134***	-0.000181***	-0.000189***
	(0.00000438)	(0.0000147)	(0.0000161)	(0.0000249)	(0.0000302)	(0.0000394)
% female authors (not last)	-0.000109***	-0.000164***	-0.000292***	-0.000291***	-0.000330***	-0.000479***
	(0.00000909)	(0.0000224)	(0.0000221)	(0.0000330)	(0.0000393)	(0.0000489)
Observations	20852950	5548874	8280883	5421406	5397483	5053360
year FE	y	y	y	y	y	y
article-field	y	y	y	y	y	y

Panel B: Commercialization gap by increasing probability of commercialization

	Commercialized		
	Prediction without language model		
Probability of commercialization	*0-0.33*	*0.33-0.66*	*0.66-1*
	(1)	(2)	(3)
Female last author=1	-0.0000609***	-0.000227***	-0.000361*
	(0.00000525)	(0.0000630)	(0.000183)
% female authors (not last)	-0.000169***	-0.000688***	-0.000642*
	(0.00000869)	(0.0000758)	(0.000253)
Observations	44699009	4141217	916870
year FE	y	y	y
article-field	y	y	y

Notes: This table analyzes the possible missed innovation due to the gender gap in commercialization, focusing on papers with at most one patent citation. The left-hand side variable is the commercialization of a given paper as measured by the patent-paper-pair described in section 2.1. Each column of panel A estimates a subset of the 70,016,266 based on citation-count decile, with column (1) collapsing below-median deciles as most of these papers are never cited. Panel B shows a similar estimation based on groups of increasing probability of commercialization. All models include controls for the number of authors, the prestige of authors, the prestige of authors' institutions, and the journal impact factor.+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Predictive model

Language model

BERT utilizes a bidirectional language model, allowing it to consider both each word's left and right context. To achieve high accuracy, BERT undergoes pretraining on a massive corpus consisting of billions of words (BooksCorpus and English Wikipedia).

By leveraging the contextual associations from those sources, BERT can generate 768-dimensional vector representations for words within a text block. These representations consider the surrounding words to capture the contextual meaning of each word.

BERT is designed to accomplish two essential tasks. First, it learns to predict masked words within a sentence, with approximately 15% of the words being masked. The model then predicts the masked words based on the context provided by the surrounding words. This masked language modeling task aims to minimize the cross-entropy loss between the predicted probabilities and the actual masked tokens.

Second, BERT is trained to understand connections between sentences through a task called next sentence prediction. Pairs of sentences are used, and the model is trained to classify whether the second sentence follows the first sentence (labeled as "IsNext") or if it is a randomly chosen sentence (labeled as "NotNext").

The main objective of the BERT model is to minimize the combined loss function, which consists of the cross-entropy loss from the masked token task and the binary loss from the next sentence prediction task.

To further enhance sentence-level understanding, an extension of BERT called Sentence-BERT (SBERT) is utilized. SBERT focuses on generating fixed-length vector representations (embeddings) specifically for sentences or short texts. Unlike BERT, which primarily focuses on word-level tasks, SBERT aims to capture entire sentences' meaning and semantic similarity. SBERT has demonstrated improved speed compared to BERT while maintaining the accuracy achieved by BERT (Reimers and Gurevych 2019).

Tuning parameters

Table A14: Tuning parameters

	Field	Tuning without Language model	Tuning with Language model
1	{'Acoustics'}	(5, 'hinge', 'l2')	(0.0003, 'log', 'l1')
2	{'Agricultural Engineering'}	(0.0007, 'log', 'l1')	(0.001, 'log', 'l1')
3	{'Allergy'}	(0.007, 'log', 'l2')	(0.0003, 'log', 'l1')
4	{'Andrology'}	(0.0005, 'hinge', 'l1')	(5, 'log', 'l2')
5	{'Biochemical Research Methods'}	(5, 'hinge', 'l2')	(1, 'hinge', 'l2')
6	{'Biochemistry & Molecular Biology'}	(0.03, 'hinge', 'l2')	(0.08, 'hinge', 'l2')
7	{'Biotechnology & Applied Microbiology'}	(0.07, 'hinge', 'l2')	(0.5, 'hinge', 'l2')
8	{'Cell & Tissue Engineering'}	(0.3, 'hinge', 'l2')	(0.0005, 'log', 'l1')
9	{'Cell Biology'}	(0.0005, 'log', 'l1')	(0.0007, 'hinge', 'l1')
10	{'Chemistry, Analytical'}	(0.03, 'hinge', 'l2')	(0.0003, 'log', 'l1')
11	{'Chemistry, Applied'}	(0.005, 'log', 'l1')	(0.0007, 'hinge', 'l1')
12	{'Chemistry, Inorganic & Nuclear'}	(0.0001, 'hinge', 'l1')	(0.0007, 'log', 'l1')
13	{'Chemistry, Medicinal'}	(0.05, 'hinge', 'l2')	(0.0007, 'hinge', 'l1')
14	{'Chemistry, Multidisciplinary'}	(0.007, 'log', 'l2')	(0.0007, 'log', 'l1')
15	{'Chemistry, Organic'}	(0.0003, 'log', 'l2')	(0.0005, 'hinge', 'l1')
16	{'Chemistry, Physical'}	(20, 'log', 'l2')	(0.0003, 'log', 'l1')
17	{'Computer Science, Artificial Intelligence'}	(0.0003, 'log', 'l1')	(0.0001, 'hinge', 'l1')
18	{'Computer Science, Cybernetics'}	(0.0005, 'hinge', 'l1')	(0.001, 'hinge', 'l1')
19	{'Computer Science, Hardware & Architecture'}	(0.05, 'hinge', 'l2')	(0.0003, 'log', 'l1')
20	{'Computer Science, Information Systems'}	(0.0007, 'log', 'l1')	(0.0001, 'hinge', 'l1')
21	{'Computer Science, Software Engineering'}	(0.0005, 'log', 'l1')	(0.0007, 'log', 'l1')
22	{'Computer Science, Theory & Methods'}	(10, 'hinge', 'l2')	(0.0003, 'log', 'l1')
23	{'Developmental Biology'}	(10, 'log', 'l2')	(0.001, 'log', 'l1')
24	{'Electrochemistry'}	(0.0003, 'log', 'l2')	(0.001, 'log', 'l1')
25	{'Engineering, Biomedical'}	(0.05, 'hinge', 'l2')	(0.0007, 'hinge', 'l1')
26	{'Engineering, Chemical'}	(0.0001, 'hinge', 'l1')	(0.0005, 'hinge', 'l1')
27	{'Engineering, Electrical & Electronic'}	(0.005, 'log', 'l2')	(0.03, 'log', 'l2')
28	{'Genetics & Heredity'}	(0.0007, 'hinge', 'l1')	(0.0003, 'hinge', 'l1')
29	{'Hematology'}	(0.007, 'log', 'l2')	(0.0007, 'log', 'l1')
30	{'Imaging Science & Photographic Technology'}	(0.01, 'log', 'l1')	(0.007, 'log', 'l1')
31	{'Immunology'}	(0.03, 'hinge', 'l2')	(0.05, 'log', 'l2')
32	{'Instruments & Instrumentation'}	(0.0005, 'hinge', 'l1')	(0.001, 'hinge', 'l1')
33	{'Materials Science, Ceramics'}	(0.005, 'hinge', 'l1')	(0.001, 'log', 'l1')
34	{'Materials Science, Coatings & Films'}	(0.001, 'hinge', 'l2')	(0.1, 'log', 'l2')
35	{'Materials Science, Multidisciplinary'}	(0.0007, 'hinge', 'l2')	(0.0001, 'hinge', 'l1')
36	{'Medicine, Research & Experimental'}	(0.0003, 'log', 'l1')	(0.1, 'log', 'l2')
37	{'Microbiology'}	(5, 'log', 'l2')	(0.0007, 'log', 'l1')
38	{'Multidisciplinary Sciences'}	(0.0007, 'hinge', 'l1')	(0.001, 'log', 'l1')
39	{'Nanoscience & Nanotechnology'}	(0.03, 'hinge', 'l1')	(0.7, 'log', 'l2')
40	{'Neurosciences'}	(0.005, 'log', 'l2')	(0.0001, 'log', 'l1')
41	{'Oncology'}	(0.0007, 'hinge', 'l1')	(0.0001, 'hinge', 'l1')
42	{'Ophthalmology'}	(0.0001, 'hinge', 'l1')	(0.0003, 'log', 'l1')
43	{'Optics'}	(0.005, 'log', 'l2')	(0.0003, 'hinge', 'l1')
44	{'Pharmacology & Pharmacy'}	(0.0007, 'hinge', 'l1')	(0.0003, 'hinge', 'l1')
45	{'Physics, Applied'}	(0.01, 'log', 'l2')	(0.0003, 'hinge', 'l1')
46	{'Physics, Mathematical'}	(0.007, 'hinge', 'l1')	(0.01, 'hinge', 'l1')
47	{'Plant Sciences'}	(0.005, 'log', 'l2')	(0.0005, 'log', 'l1')
48	{'Polymer Science'}	(0.01, 'log', 'l2')	(0.0005, 'hinge', 'l1')
49	{'Radiology, Nuclear Medicine & Medical Imaging'}	(0.001, 'log', 'l1')	(0.0003, 'hinge', 'l1')
50	{'Reproductive Biology'}	(0.0005, 'hinge', 'l1')	(0.0007, 'log', 'l1')
51	{'Rheumatology'}	(0.001, 'log', 'l1')	(0.0005, 'log', 'l1')
52	{'Robotics'}	(0.0007, 'log', 'l1')	(0.005, 'hinge', 'l1')
53	{'Spectroscopy'}	(0.007, 'hinge', 'l1')	(0.03, 'hinge', 'l1')
54	{'Toxicology'}	(0.005, 'hinge', 'l1')	(0.01, 'hinge', 'l1')
55	{'Virology'}	(0.3, 'hinge', 'l2')	(0.3, 'hinge', 'l2')
56	{All the others}	(0.0007, 'log', 'l1')	(0.0003, 'hinge', 'l1')

Notes: Table shows the tuning parameters for the different groups. The models were built to optimize the f1-score.