From Quotas to Questions: How Shifting Gender Ratios Shape Medical Research

Fatima Aqeel*, Thea How Choon† and Yvette Zhang‡

August 2026

Abstract

How does a change in workforce gender composition influence research? We study a sudden rise in the share of women in medicine after a Pakistani medical school admissions reform. Using an IV difference-in-differences design and the universe of PubMed-indexed Pakistani medical research, we compare specialty-district units subjected to varying treatment intensity and show that a 1 percentage point (p.p.) increase in the female share raised female-relevant research by 0.5 p.p. This shift stemmed primarily from the inclusion of female subjects in research. Incumbent authors, especially female incumbents, became significantly more likely to participate in female-inclusive research. Our results suggest that gender diversity in the medical workforce can broaden the evidence base of research.

JEL Codes: O31, J16, I14

*Department of Economics, College of the Holy Cross. Email: faqeel@holycross.edu
†Department of Economics, St. Lawrence University. Email: thowchoon@stlawu.edu.
‡Department of Economics, Yale University. Email: yvette.zhang@yale.edu
The authors thank St. Lawrence University for generous funding through the William B. Bradbury, Jr. Faculty Award, and seminar participants at the Society of Labor Economists 2026 Meeting (SOLE), LacPal, St. Lawrence University, Boston University and Brown University for valuable feedback. Many thanks to Raquel Fernandez for thoughtful advice, and to Amelia Beaver for research assistance.

1 Introduction

In 2014, cardiac patient Joy O'Neal's heart failure was initially misdiagnosed as aging and emotional distress due to limited knowledge of women's heart failure symptoms (Mayo Clinic, 2022; Yale School of Medicine, 2023). In fact, women are 50% more likely to receive incorrect diagnoses of heart conditions than men (British Heart Foundation, 2016). One reason for these differences is that medical research has disproportionately been shaped by male-focused research, therefore providing less evidence on how health conditions and drugs affect women (Association of American Medical Colleges (AAMC), 2023).

In the medical profession, recent gender composition changes have been a near global phenomenon (Ramakrishnan et al., 2014; Penny et al., 2014; Shankar et al., 2012) and raise questions about implications for medical research. In the United States, for example, the share of women in medical schools rose from less than 6 percent in 1960 to 51 percent in 2020 (AAMC, 2023; Heiser, Stuart, 2017). But the causal effects of changing gender ratios are difficult to estimate because of the endogeneity of female entry: women typically enter medical fields that are already welcoming to them.

We examine how changes in the gender composition of the workforce affects medical research by leveraging a unique 1992 admissions reform in Pakistani medical schools that increased the share of women in medicine in a plausibly exogenous way. Prior to the reform, women's enrollment was capped at 20 percent.[1] A court struck down the cap after prospective female students who had scored above admitted men filed a case for discrimination, resulting in a rapid shift: the female share rose 10 percentage points (p.p.) in one year and from 28 percent pre-reform to 55 percent thereafter (Aqeel, 2024).

Crucially, the effects of the reform varied in intensity across specialty-district units: a lower pre-reform female share relative to similar units led to a larger gain in female share after the reform due to catch up effects. For instance, the female share of pediatricians in Hyderabad was 16.7 percent before the reform, which was lower than the female share

[1]While the policy was imperfectly enforced, we show that it was binding.

of doctors both in Hyderabad (28.4 percent) and in Pediatrics (20.7 percent). After the reform, the female share in that unit increased by 24.5 p.p., compared to 21.0 p.p. and 14.5 p.p. in Hyderabad and Pediatrics, respectively. We use these idiosyncrasies in pre-reform shares as a source of conditionally exogenous variation in the change in female shares after the reform.

Our empirical strategy is an instrumented difference-in-differences (DDIV) strategy, where the change in female share is instrumented by the pre-reform female share. First stage estimates show that a 1 p.p. higher pre-reform share predicted a 0.57 p.p. *lower* post-reform change in female share, conditional on specialty fixed effects and district fixed effects. While the instrument is pre-determined, we address concerns about persistent endogenous factors with specialty and district fixed effects, evidence supporting the parallel trends assumption, and alternative specifications that allow for trends in specialties and districts. Further, note that since more treated areas were *less* inclusive pre-reform, any persistent factors are likely to bias our results towards zero.

We collect data on the universe of PubMed-indexed medical research by Pakistani first-authors and calculate gender shares from doctor registration data. PubMed categorizes articles with "Female" and "Male" check tags based on the research content and/or participant characteristics, which provides an objective measure of female relevant research.

Our first result shows that a 1 p.p. increase in the female share raised the likelihood that articles include a "Female" tag by 0.50 p.p., or a 19% increase at the mean level of treatment. Notably, the exclusive presence of "Female" tags (indicating a focus on only female-relevant research) did not increase; instead, more "Female" tags were included in addition to "Male" tags. This suggests that the scope of research expanded to be more female-inclusive instead of crowding out male inclusiveness. We show descriptively that some of the greatest increases in female relevant research were in conditions ex-post known to affect women differently from men, such as heart disease.

Research primarily became more female-inclusive due to an increase in female subjects, as opposed to female-relevant topics. Since the "Female" tag does not distinguish between female participants and female-centric topics, we separate the two in the spirit of Koning et al. (2021) by using text analysis. We identify references in the abstract to the female anatomy and related health conditions, and to gender descriptors of participants. The presence of gender descriptor words increased by 0.71 p.p. for every 1 p.p. increase in treatment. Meanwhile, female anatomy/health condition words increased but not statistically significantly.

At least part of the overall increase in female-relevant research was driven by incumbent (pre-reform) authors. Pre-reform female authors became 0.76 p.p. more likely to work on female relevant research for every 1 p.p. higher treatment after the reform. The change is negative and under statistical significance for male incumbents, however differences in participation between male and female incumbents are also not significant. These results use within-author variation and suggest a treatment effect of the reform on existing researchers, rather than a pure compositional effect.

Turning to mechanisms, we find that female incumbents' collaborations with junior authors may be important for research inclusiveness. In more treated units, female incumbents became more likely to coauthor with junior researchers on work that also included female participant descriptor words. The corresponding increase is much smaller for female-incumbent research without junior coauthors. We also do not observe similar effects for male incumbent authors, which suggests these effects were not driven by overall trends in female patient participation. The timing of the effects also suggests that gender composition changes in the occupation, not in student populations alone, were the primary driver of research inclusiveness.

Finally, increases in female-inclusive research were not limited to PubMed. We additionally collected data on research published in non-PubMed indexed journals which are local and often do not undergo peer review. Using our lexicon of gender related words,

we find evidence that female-related words increased in the abstracts of research published in local journals as well.

We first contribute by providing causal evidence connecting author identity and research. Author identity can affect research based on factors such as researcher experience, preferences and opportunity (Einiö et al., 2019; Airoldi and Moser, 2024),[2] and in turn shape the direction of innovation (Acemoglu, 2002, 2023).[3] Female research teams are more likely to produce female-oriented research and patents (Koning et al., 2021, 2020; Koffi and Marx, 2026), and researchers are better able to recruit clinical trial participants who share their identity (Gupta, 2022; Gupta et al., 2025). However, little causal evidence exists on the relationship between author gender and research because women may sort into environments that are supportive of women's research. We leverage a plausibly exogenous policy to show that as the female share among doctors increased, research became more inclusive. Among incumbent researchers, female authors significantly increased participation in inclusive research, and especially in collaboration with junior authors. This paper's setting is also a rapidly expanding research sector in a developing country, which has previously received little attention.

Second, our paper is linked to the effects of sector-wide gender composition changes (Greenberg et al., 2024; Karpowitz et al., 2024). Truffa and Wong (2025) study gender composition changes after colleges transitioned to co-education and find a stark increase in gender-related research, driven mostly by male incumbents, and also a compositional effect through the hiring of female and male researchers interested in gender. Our work extends existing literature by studying gender composition changes in an entire occupation, and examines the effect of treated cohorts graduating and joining the workforce. This allows us to show that changes in workforce demographics can influence the research tra-

[2]Existing research also examines how socioeconomic status (Bell et al., 2019; Akcigit et al., 2017; Aghion et al., 2017), gender (Jensen et al., 2018) and race (Cook and Kongcharoen, 2010) affect differences in innovating and patenting.

[3]There is vast research on the direction of innovation which includes Acemoglu et al. (2012); Acemoglu and Restrepo (2022); Acemoglu (2023); Aghion et al. (2016), among others.

jectory of incumbent scholars. Further, while women crowded out men in medical school, female-relevant research did not significantly crowd out male relevant research.

Third, this paper helps to understand how gender gaps in medical research can be reduced. Historically, medical research based on men was generalized to diagnose women (Merone et al., 2022; Graves Jr et al., 2022). We now know that women experience conditions such as stroke and Alzheimer's differently (American Heart Association, 2024; The Harvard Gazette, 2024), and innovation differences can generate inequality in who benefits (Bryan and Williams, 2021). Scholars and policymakers are therefore calling for more inclusive research practices to improve the quality and usefulness of research (Gualtierotti, 2025; American Psychological Association, 2023; The Conference Forum, 2024). One way to reduce the medical research gap is through increasing women's representation in clinical trials (Michelman and Msall, 2021). The findings in this paper suggest that it could also be reduced through workforce diversity. In this context, topics do not change significantly, but evidence suggests that female participants were increasingly recruited into medical studies.

The remaining paper is structured as follows. In section 2 we describe the context, and research in Pakistan. In section 3 and 4 we detail our sources of data and identification strategy, respectively. Section 5 to section 7 present results, robustness and mechanisms, respectively, before concluding in 9.

2 Background

Background on the Pakistani Medical School Admissions: Students apply to Pakistani medical school immediately after completing the required pre-medical coursework in high school. Prior to 2000, admissions were based solely on academic performance in national high school and national middle school exams (Andrabi and Singh, 2015). Since then, admission criteria have included a standardized entrance exam (50 percent of the

composite score), high school grades (40 percent), and middle school grades (10 percent).

Medical education typically lasts five years, after which graduates can register to practice with an M.B.B.S. degree (Bachelor of Medicine, Bachelor of Surgery), which is equivalent to a general medical license. Alternatively, they may pursue a residency (referred to locally as a "house job") and specialization.

The Policy Change: In 1948, Pakistan introduced a gender quota in medical schools, reserving 20 percent of seats for women and 80 percent for men. While this was initially to guarantee seats for women, women and men did not then compete for the same seats. Of 858 seats in medical colleges, women could secure admissions in about 181 seats under the women's quota. The remaining 677 seats were awarded to men. However, female applications grew over time, competition among women increased and the quota became a constraint (Masood, 2019).

The policy was overturned following the 1989 Shirin Munir case, in which female applicants challenged the quota for violating constitutional protections against sex-based discrimination (PLD 1990 SC 295) (LUMS University, 2024). The court stated:

"As the two Articles 25 and 22 (of the Constitution of Pakistan) stand at present, the general rule is that "sex" cannot be adopted as the sole criteria for discrimination except for advancing the cause of women and children as permitted by clause (3) of Article 25....When an educational institution is thrown open for co-education, then a further restriction of numbers on the ground of sex is permissible only as a protective measure of women and children but in no case so as to protect comparatively less meritorious boys to the prejudice and exclusion of women. This amounts to gross violation of the constitutional mandate" (Supreme Court of Pakistan, 1989).

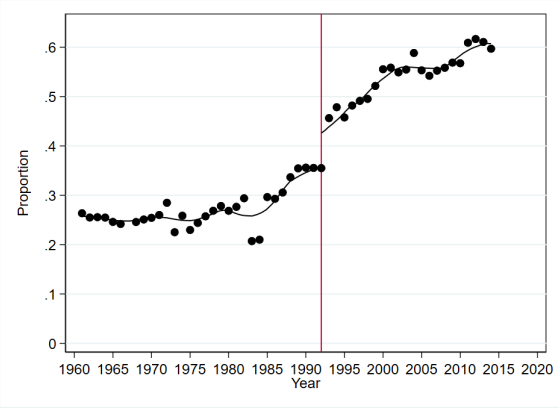
As a result, admissions became merit-based nationwide. The first gender-neutral class was admitted in 1992 and is expected to have graduated in 1997.

The female share among students enrolled in medical schools is depicted in figure 1 (which is reproduced from Aqeel (2024) and adapted to the current context). The figure

suggests imperfect compliance before the reform. This may be for two reasons. First, the data are aggregate and include 2 all-female schools that existed at the time, which could increase the female share observed.[4] Second, the 20 percent quota was enforced via a higher test score threshold for women, and this manner of enforcement could mean that the share of women was not always exactly 20 percent.

However, a distinct increase in the female share after 1992 suggests that the old policy was binding. The female share among medical students jumped by almost 10 p.p. in a single year. It then continued to increase thereafter, reaching about 60% in 2018.

Figure 1: Female Share Among Enrolled Medical Students



Note: The graph is reproduced from Aqeel (2024) but adapted to the current setting. It plots the mean share of female students enrolled in medical schools. The vertical line is at the year 1992, and data are from administrative data (provincial development reports). Data are from provincial development reports published by the Pakistan Bureau of Statistics.

Medical Research in Pakistan: The medical research sector in Pakistan, while small, is a rapidly growing one. For example, the number of PubMed articles increased 7.5 times between 2001 and 2011 (Ghaffar et al., 2013). Pakistan had the second highest growth rate in medical publications among 46 countries, publishing 41,889 new articles between 2001 and 2020, and up from 4,378 in the preceding 53 years (Hassan and Ara, 2021).

The majority of published research is non-clinical.[5] There were 508 registered clinical

[4] A back-of-the-envelope calculation, available upon request, suggests that the share was closer to 20% if we account for the pre-reform enrollment of these schools.
[5] Including case reports, reports from surveys and review articles.

trials from September 1992 to February 2019 in Pakistan, which was below the 6,700 average for the Eastern Mediterranean region between 1999-2005 (World Health Organization, 2026). These included interventional and observational trials (77.2 percent and 22.8 percent, respectively) (Siddique et al., 2022). Patents generated includes those for producing antioxidants, enzyme inhibitors, herbal actives, and drug compounds for cancer, diabetes, inflammation, Parkinson's disease (University of Karachi, Pakistan, 2023).

Who conducts medical research? Medical research is mostly conducted by doctors and by non-medical academics. While some medical doctors are solely clinical practitioners, those in academic medicine are required to publish research papers to be appointed or promoted to a position of Assistant Professor or higher. Only 6.6 percent of medical students report publishing research (Ahmad et al., 2022). Meanwhile, researchers who are not doctors are affiliated with university departments such as Biology, Chemistry, Genetics and Pharmacology.[6]

There are no widely available estimates of the share of doctors who participate in research. Related estimates suggest that this share may be low. For example, 8 percent of PhDs in Pakistan were also medical doctors in 2013 (Qureshi et al., 2013). Even in a research-centric country such as the United States, only 4 percent of respondents in the National Sample Survey of Physicians by the AMA (2015) held a medical degree and a PhD. Medical doctors holding PhDs were more likely to conduct research in the US: 14% of medical doctors and 33% of medical doctors holding a PhD conducted research in 2022 (Association of American Medical Colleges (AAMC), 2022).

The share of medical research produced by medical doctors is a related question. In our data, described below, registered medical specialties in Pakistan match with about 46.5% of PubMed articles first-authored by a Pakistani. While this may be for several

[6]Guidelines by the Pakistan Medical and Dental Council require hiring to the position of Assistant Professor with a minimum of 2 publications, promotion to Associate Professor with a minimum of 7 new publications, and to Professor with 6 additional publications. The Higher Education Commission of Pakistan (HEC) similarly governs promotion for non-medical academics.

reasons, we were able to code 60% of unmatched articles with known non-clinical subjects such as Statistics, Physics, Chemistry and Genetics. Further manual checks of unmatched articles suggest they are authored by academics affiliated with non-medical departments.

Publication process: PubMed-indexed medical research is typically peer reviewed, and in most scientific fields, the first author listed is the main author of the paper (Lapidow and Scudder, 2019; of Medical Journal Editors et al., 1997).

About 25 percent of Pakistan's medical schools published their own journals in 2013, and among them, only the Journal of Ayub Medical College in Abbottabad is indexed in PubMed (Ghaffar et al., 2013). Within specialties, the majority of medical research appears to be published in PubMed-indexed journals, however. For example, a study of skin cancer publications showed that 81% of the research analyzed was published in PubMed-indexed journals (Rehman et al., 2021), and in Pediatrics, 66% were published in PubMed (Abbas et al., 2024).

MeSH tags in PubMed: Each PubMed article contains "Medical Subject Heading" (MeSH) terms, which are unique identifiers assigned to articles using the MeSH vocabulary, which itself is a controlled vocabulary created by the National Library of Medicine (NLM) (National Library of Medicine, 2025a). These terms are used to index and catalog biomedical and health-related information, primarily in the MEDLINE/PubMed database. They are, in some ways, similar to JEL codes in Economics.

MeSH terms are organized in a hierarchical structure, with broader terms at the top and more specific terms branching out below. Most classification is conducted using the Medical Text Indexer (MTI) algorithm,[7] after which human indexers historically reviewed articles for accuracy.[8]

[7]The MTI algorithm uses natural language processing to classify medical articles.

[8]Since 2022, the process has become more automated, relying on abstracts, titles, and the MTI algorithm (National Library of Medicine, 2024). The MTI achieves a precision (how many predicted classifications were correct) of about 93 percent, while its recall (how often correct terms were identified) is lower at around 65 percent, though still superior to alternative indexing methods (Koning et al., 2021).

The ”Male” and ”Female” tags in MeSH terms are not core MeSH term subject headings, but are tags assigned by human indexers to the MeSH terms of articles. Indexers will assign a ”Female” check tag to the publication's MeSH record if the publication covers ”female organs, diseases psychological processes, genetics, etc.”, and analogously assign the ”Male” tag (Koning et al., 2021; National Library of Medicine, 2025b). Note that a study on general subjects that includes female clinical trial participants would satisfy the condition for a ”Female” tag (Virtual Health Library, 2024). The tag can be included for human and non-human subjects, and an additional tag of ”Human” indicates a human female or male.

Recruiting participants in studies: To our knowledge, there was no formal limit to enlisting female participants in Pakistan, unlike historically in the US (Michelman and Msall, 2021). Participants in medical studies are typically recruited through health care provider networks, nonprofit organizations, and social media platforms (NOVOTECH, 2021), with junior doctors often actively involved in recruiting (Vas et al., 2021). Existing research suggests that individuals from underrepresented groups may be more likely to participate in clinical trials when the Principal Investigator (PI) is from the same group (Alsan et al., 2025), and female PIs are more effective at recruiting women (Gupta, 2022).

3 Data

Our primary data source for medical publications is PubMed, an extensive and publicly available database maintained by the United States National Library of Medicine (NLM). We use the Entrez API provided by the National Center for Biotechnology Information (NCBI) to compile data at the article level from PubMed.

To identify articles first-authored by Pakistanis, we combine two additional data sets: the ”Author-ity” dataset (Torvik and Smalheiser, 2009, 2021), which matches PubMed authors across publications, and ”MapAffil” (Torvik, 2015), which tracks location data.

MapAffil uses the location of the first-listed author's affiliation as indexed by the NLM.[9] Cross-referencing MapAffil with author publication histories from Author-ity allows us to assemble a set of authors whose research has been linked to Pakistan, and we only include authors whose top 20 most frequent affiliations include a location in Pakistan. We then comb through the list to verify that researchers were likely to be based in Pakistan by checking if their name is a "Pakistani name" based on the proprietary software "Namsor" and medical registration data.

We also assign a likely gender to author names in articles. To assign gender, we use the Facebook names data set used in Chari and Goldsmith-Pinkham (2017), which contains names of Facebook users in New York City along with a probability of gender based on their demographic information on Facebook. We combine this with Pakistani doctor registration data which contains doctor names with their gender to increase the likelihood that we capture uniquely Pakistani names. Using both datasets, we assign a name a likely gender if the associated gender probability is at least 90 percent. Finally, research published on PubMed before 2002 often only provided the initials of authors instead of full first names. We hand-collected first names by searching for the author and article and locating the profile of the author (for example, their Google Scholar profile, Research Gate profile, or personal website). Overall, we were able to assign genders to about 75 percent of names in our entire data with this approach.

Registration data from the Pakistan Medical and Dental Council (PMDC) are used to calculate pre-reform female shares and the change in female share by medical specialty and district.[10] According to the PMDC, all doctors who practice in Pakistan must register in the district in which they practice. We compute the average female share before 1992

[9]For each paper, the NLM only provides the affiliation linked to the first author, starting from 1988. Affiliations became linked to authors, rather than papers, in 2014 (Tybaert, 2014), which is outside our time period of interest.

[10]Small specialties with observations less than 1.5 percent (1,500) of total observations are grouped with their parent or closest related specialty. This reduces our 53 initial specialties into 37 clinical specialties. We also exclude non-clinical specialties to reduce the chance that we capture non-doctors in our sample. From hereon, "specialties" refers to this group.

(pre-reform female share) and after 1997 (post-reform female share) by specialty and district. Years between 1992 and 1997 are omitted to avoid individuals who graduated early. The change in the female share is calculated as the pre-reform share subtracted from the pre-reform share.[11]

3.1 Summary Statistics

We present summary statistics based on our regression sample in table 1. Our regression sample includes 7,128 articles with Pakistani first authors and female shares in the registration data. The publications are restricted to those in clinical fields to limit the inclusion of non-medical doctors, and the time frame is restricted to at most 10 years before 1997 (1988) until the last year that we have data available (2018) to ensure a sufficient sample size.

In our sample, 68% of articles have a "Female" tag and 19% have a "Female" tag but no "Male" tag in their MeSH terms. Among articles about humans ("Human"-tagged), about 19% of articles have "Female" tags only and 15% percent have "Male" tags only. This suggests that relatively few studies involve only female or male human participants. The share of incumbent female authors among articles with at least 1 incumbent author is 30%. The mean pre-reform female share in the matched sample is 19%, and mean change in female share is 24 percentage points.

The top 10 MeSH terms in our research sample indicate the disease categories that are most researched in this context (table A1). This list includes Diabetes, which is included in the MeSH terms of 1,721 articles, Carcinoma, in 1,632 articles, and Tuberculosis, in 1,339 articles.

[11] Using the full period 1952-1991 for the pre-reform female share allows us to address the fact that (i) in the years immediately before 1991, the female share started to change in an ad-hoc way and may not be representative of the average female share, and (ii) before the mid 1980's the female share was more stable and similar, suggesting a lack of significant trends.

Table 1: Summary Statistics

	Mean	S.D.	N
Female MeSH Tag	0.68	0.47	7,128
Female MeSH Tag, No Male MeSH Tag	0.19	0.39	7,128
Female (Human) MeSH Tag, No Male MeSH Tag	0.19	0.39	7,128
Male (Human) MeSH Tag, No Female MeSH Tag	0.15	0.36	7,128
Number of Authors	4.21	2.47	7,128
Incumbent Female Author Share	0.30	0.46	3,823
Δ Female Share	0.24	0.16	7,128
Pre-Reform Female Share	0.19	0.21	7,128

Note: The table presents summary statistics for the regression sample. Article-level variables are measured using PubMed articles published between 1988 and 2018, while female-share variables are constructed from medical registration data matched to the article data. The incumbent female author share is measured among articles with at least one incumbent author.

4 Empirical Strategy

Given our setting, a first-order identification strategy would be to estimate a difference-in-differences regression where the research outcome is regressed on an interaction of a continuous treatment (the change in female share) and a post-reform indicator. While informative as a first step, this approach is likely to suffer from endogeneity. For example, units that experienced larger increases in female share may also be open to female-focused research. Estimates may therefore reflect underlying unit characteristics, rather than the change in gender ratios.

To address endogeneity concerns, we incorporate an instrumental variable (IV) in our

difference-in-differences estimations. We show that the change in the female share of doctors was negatively related to the pre-reform female share of doctors, even after including specialty fixed effects and district fixed effects. Therefore, we instrument for the change in female share ("treatment") with the pre-reform female share (the instrument). Our strategy is similar in spirit to empirical designs where exposure to treatment varies due to pre-reform characteristics (Angrist and Imbens, 1995; Topalova and Khandelwal, 2011; De Loecker et al., 2016; Duflo, 2001).[12]

To see why the pre-reform female share and the change in the female share may be negatively related, consider the set of district-specialty units $I = \{1, 2, ..., N\}$, with pre-existing female shares $\{s_i^0\}_{i=1}^N$. If units have a turnover rate of $\lambda \in (0, 1)$,[13] and the female share among new researchers is μ, then the new female share in unit i is

$$s_i^1 = (1 - \lambda)s_i^0 + \lambda\mu$$

which gives the change in female share

$$\Delta s = s_i^1 - s_i^0 = \lambda(\mu - s_i^0)$$

Then, in units with a smaller pre-reform female share, because the difference in female share between the outgoing and incoming researchers is larger, the change in the female share is larger.

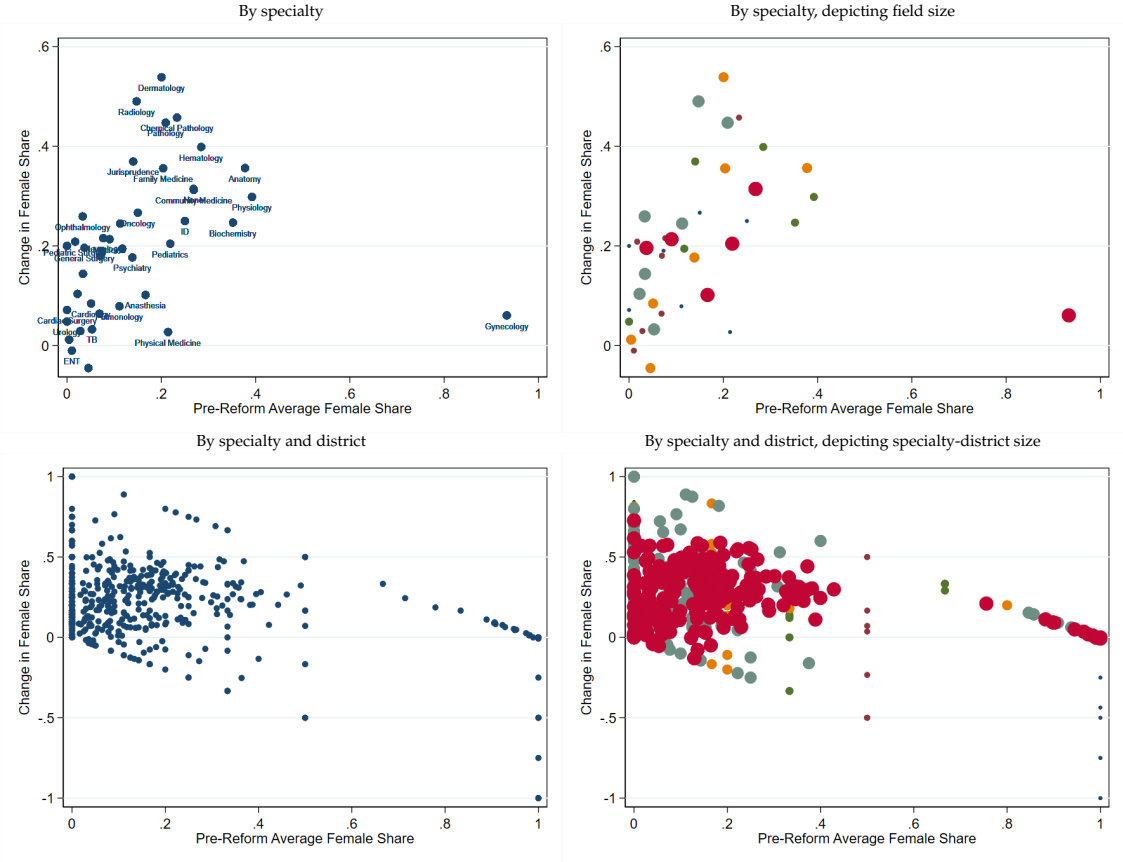
Figure 2 illustrates the relationship between the pre-reform female share and the post-reform change in female share among doctors. Panels (a) and (b) plot variation across specialties, with the size of points in panel (b) representing the pre-reform size of the field in terms of the number of doctors. While this relationship is positive for specialties

[12] Our approach is also related to shift-share (Bartik) instruments in that it exploits pre-reform characteristics to generate differential exposure to a policy shock (Goldsmith-Pinkham et al., 2020). However, unlike a standard Bartik design which involves exposure to multiple national shocks using pre-determined shares, our setting involves a single national reform that affects all specialty-districts simultaneously. Pre-reform female shares therefore determine the intensity of exposure to this common shock, rather than serving as a weight over multiple shocks.

[13] In figure A2 we provide support for a constant turnover rate by showing that turnover is proportional to the size of the unit in registration data.

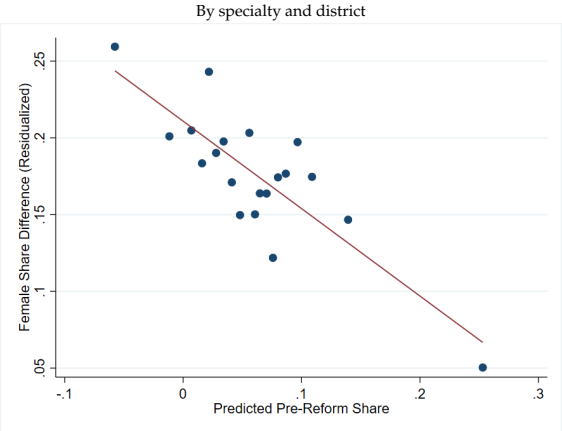
alone, it flips to being negatively sloped at the specialty-district level in panels (c) and (d). Figure 3 sharpens the relationship further by netting out district and specialty effects, and reveals a strong negative relationship between the pre-reform share and the change in female share.

Figure 2: Change in the Female Share of Specialties



Note: The graphs plot the relationship between the pre-reform average share of female doctors and the change in female share of doctors. Data are from the PMDC Medical Doctor Database between 1952 and 2018.

Figure 3: Change in Female Share against Pre-Reform Share, Partialling out Controls



Note: The graph plots the change in female share against the pre-reform female share after partialling out control variables. The scatter plot was obtained by regressing the change in female share on the pre-reform share, and controlling for specialty and district fixed effects. Data are from the PMDC Medical Doctor Database between 1952 and 2018.

Our structural equation is run on article data matched with specialty and district level shares from medical doctor registration data, and of the form:

$$Y_{isdt} = \beta_0 + \beta_1(Treatment_{sd} * Post_t) + \beta_s + \beta_d + \beta_t + \pi FemaleShare_{sd} + u_{isdt} \qquad (1)$$

where Y_{isdt} is an outcome for article *i* published in year *t* by 1st author in district *d* and medical specialty *s*, and $FemaleShare_{sd}$ is the pre-reform female share of doctors in specialty *s* and district *d*. The treatment is defined as the change in female share of doctors in the author's specialty and district. The main outcome at the article level is the female-inclusiveness of articles, measured primarily by the "Female" tag in the article's MeSH terms. β_s are specialty fixed effects, β_d are district fixed effects, β_t are year fixed effects, and u_{isdt} is an individual disturbance term. Standard errors are clustered by medical specialty and district.

We instrument for treatment intensity using the pre-reform female share in specialty s

and district d. The first stage is:

$$Treatment_{sd} = \alpha_0 + \alpha_1 FemaleShare_{sd} + \alpha_s + \alpha_d + \epsilon_{sd} \tag{2}$$

All fixed effects in equation 1 are included in the second stage, and standard errors are clustered by medical specialty and district.[14]

First stage estimates in table 2 show that a 1 p.p. higher pre-reform share in a specialty-district pair leads to a 0.57 percentage point *lower* change in the female share after the reform (column 1), conditional on specialty and district fixed effects and after removing outlier specialty-districts where the pre-reform female share exceeded the 99th percentile.[15]

Finally, we present reduced form results estimates based on:

$$Y_{isdt} = \delta_0 + \delta_1 FemaleShare_{sd} + \delta_s + \delta_d + \delta_t + \eta_{isdt} \tag{3}$$

In practice, our reduced form regressions use $(1-FemaleShare_{sd})$ instead of $FemaleShare_{sd}$ so that estimates are directionally consistent across reduced form, 2SLS estimates and event studies.

To support the parallel trends assumption for IV difference-in-differences, we estimate a reduced form event studies that replace $Post_t$ in equation 1 with two-year or one-year interval indicators interacted with the pre-reform female share of doctors:

$$Y_{isdt} = \gamma_0 + \sum_{t=-\underline{T}}^{\bar{T}} \gamma_1^t D_t^k * FemaleShare_{sd} + \gamma_s + \gamma_d + \gamma_t + \mu FemaleShare_{sd} + v_{isdt} \tag{4}$$

Importantly, the reform is likely to have induced changes in the selection of admitted students and into specialty and districts further down. To address concerns about

[14]Note that the pre-reform female share and change in the female share of doctors is constant at the specialty-district level, and so we do not include specialty-district fixed effects.
[15]Standard errors are clustered by specialty and district. The coefficient without excluding outliers is -0.82.

selection, we also analyze if incumbent authors began participating in female-relevant research. This exercise is based on a sample restricted to authors who were active prior to the reform and whose selection is less likely to change.

Table 2: First Stage

	Δ Female Share
	(1)
Pre-Reform Female Share	-0.570***
	(0.12)
Observations	991
R^2	0.44
Mean DV	0.18
F-statistic	19.5

Notes: This table presents results from first stage regressions. The regression excludes outliers where the pre-reform share exceeded the 99th percentile. All regressions include specialty and district fixed effects. Standard errors are clustered by specialty and district. Stars indicate significant differences: *p $<$ 0.10, **p $<$ 0.05, ***p $<$ 0.01.

Instrument Exogeneity and Other Identifying Assumptions The exclusion restriction (Angrist, 2002), in this context requires that the pre-reform share is only related to our outcomes through the change in female shares. In other words, if Y_i is the outcome for article *i*, D_{sd} is the change in female share and Z_{sd} the pre-reform female share for district *d* and specialty *s* then,

$$Y_{isdt}(d_{sd,t}, 0) = Y_{isdt}(d_{sd,t}, 1) \qquad \forall d = 0, 1 \tag{5}$$

Specialty and district time-invariant factors, as well as year-specific factors, can be accounted for with specialty, district and year fixed effects. Trends that are common across specialties or across districts will also cancel out in our strategy. Any remaining con-

founders would need to vary across specialty-districts by specialty, or across specialty-districts by district, which is a stricter requirement.

An important threat to the exclusion restriction is that persistent trends over time may be correlated with both the pre-reform female share and research outcomes. For example, high pre-reform female shares in units such as Pediatrics in Karachi may lead to cultural idiosyncrasies that affect female-relevant research, even in the absence of the reform. Event studies help to mitigate this concern by showing support for parallel trends before the reform. In Section 6, we also present a version of results which relaxes this assumption and includes specialty trends and district trends. Finally, note that, were this assumption violated, persistent cultural trends are likely to bias our estimates towards zero, since pre-reform female shares are likely to be positively correlated with trends in female inclusiveness, whereas our first stage is negative. In other words, units that had a high pre-reform female share, such as Pediatrics in Karachi, actually received a *lower* treatment.

Other main identifying assumptions include instrument exogeneity and monotonicity. Since the specialty-specific features are often the main determinant of pre-reform female shares, we assume that the instrument is exogenous conditional on specialty and district fixed effects. Regarding monotonicity, we assume that, everything else constant, there are no units that would have received a lower treatment if the pre-reform female share had been lower, or a higher treatment if the pre-reform female share had been larger. Finally, to the best of our knowledge, the reform was not anticipated by individuals and the case not heavily covered until a judgment was passed.[16]

Since the reform passed in a single year, our strategy does not involve a staggered design (Wing et al., 2024) but a discrete indicator of when the reform's effect is expected to begin. Further, our main independent variable is instrumented with the pre-reform

[16]The SUTVA requirement is that treatment varies at the specialty-district level. We assume that article-level outcomes depend only on the treatment in their own specialty–district. SUTVA is reasonable if research norms and collaboration patterns are primarily local within specialty–districts.

female share. Under the monotonicity assumption the instrument can only move treatment in one direction, so that there cannot be negative weights. Our strategy therefore does not correspond to the dose-response difference-in-differences literature (Callaway et al., 2021), but belongs in the average causal effects for models with varying exposure of treatment intensity (Angrist and Imbens, 1995).

Timing of Reform Effects The first cohort of medical students admitted purely on merit entered medical school in 1992 and are expected to have registered/graduated in 1997 after completing five years of training. However, we only observe the year in which an article is published, not when the research that produced it began. As a result, articles published in 1997 may still have been developed under pre-reform gender ratios.

To account for the gap between doctor registration and first publication, we matched first authors in PubMed to registration data to measure the time between doctor registration and first publication. Removing outliers below the 5th percentile, the minimum time between registration and first publication year is about 3 years (figure A1).[17] This suggests that articles published in 2000 and later were likely to be developed in the post-reform environment. In our main regressions, we treat 2000 as the first year in which publications reflect research developed after the reform.

5 Results

5.1 IV Results on "Female" Tags

Our first result is that an increase in the female share of doctors (treatment) increased the production of female-relevant research, but not the production of exclusively female-

[17] PubMed does not provide the unique registration ID of doctors, which limited the matching rate. We matched the PubMed medical researchers and medical doctor registration data sets using name, district of practice and medical specialty. With this relatively strict matching requirement, and after dropping any potential duplicate matches, about 1 percent of individual researchers in our PubMed sample (who are fewer than the number of articles) matched. The mean and median years between registration and first publication was approximately 7 years.

focused research. In table 3, the outcome is an indicator for articles having a ”Female” MeSH tag (column 1) or exclusively a ”Female” tag (column 2). Columns 3 and 4 repeat this analysis restricting the sample to articles with a ”Human” tag.

The DDIV estimate in column 1 of table 3 implies that a 1 p.p. increase in treatment raises the probability of a “Female” tag by 0.50 p.p. At the mean level of treatment, female inclusiveness increases by 13 p.p., which is a 19% increase over the mean of the dependent variable. Articles that are ”Human”-tagged became 0.51 p.p. more likely to be ”Female”-tagged for a 1 p.p. increase in treatment (column 3).[18] In columns 5 and 6, we allow for specialty-specific trends. Coefficients are positive and larger, but lose statistical significance. Estimations with additional district trends are presented in section 6.

In contrast, we find no change in the exclusive presence of the “Female” tag. The small and statistically insignificant results in columns 2 and 4 suggest that the reform increased the likelihood that articles include the “Female” tag alongside the “Male” tag, but not in substitution of it. In other words, the scope of research expanded to be more female-inclusive but did not crowd out male-inclusive research.

5.2 Reduced Form Results on ”Female”-Tags

Reduced form results in table 4 are consistent with 2SLS findings: an increase in the pre-reform female share of doctors decreases the likelihood that an article is ”Female”-tagged (columns 1 and 3) and does not change the exclusive presence of the ”Female”-tag (columns 2 and 4).[19] Coefficients remain significant with the inclusion of post-reform specialty trends (columns 5 and 6).

The naive difference-in-differences estimates in table A2 are smaller than the DDIV estimates, suggesting the IV indeed removed a source of endogeneity. Specifically, units with a higher pre-existing share were less exposed to treatment, but may also be pre-

[18] For the remainder of the paper, we focus on the “Female” tag indicator, as in column 1.

[19] Note that we use 1-(FemaleShare)$_{sd}$ in reduced form regressions instead of FemaleShare$_{sd}$ to keep the direction of the results consistent between reduced form and 2SLS results and event studies.

Table 3: Female Tagged Research - 2SLS

	Probability Female in					
	MeSH Term	MeSH Term^{+}	MeSH Term^{h}	MeSH Term$^{h,+}$	MeSH Term	MeSH Term^{h}
	(1)	(2)	(3)	(4)	(5)	(6)
Δ Female Share $\times$ Post	0.504**	0.175	0.510**	0.168	0.704	0.777
	(0.217)	(0.167)	(0.226)	(0.170)	(0.469)	(0.500)
Observations	7,128	7,128	7,128	7,128	7,128	7,128
Sanderson-Windmeijer F-stat	18.30	18.30	18.30	18.30	12.18	12.18
Kleibergen-Paap F-stat	11.92	11.92	11.92	11.92	5.31	5.31
SY 10% maximal IV size	7.03	7.03	7.03	7.03	7.03	7.03
A-R p-value	0.01	0.38	0.01	0.35	0.14	0.10
Mean DV	0.68	0.19	0.67	0.19	0.68	0.67
Specialty Trends					Yes	Yes

Notes: This table presents results from 2SLS regressions of the change in female share by specialty and district on articles' female relevance. In columns 1 and 3, the outcome is an indicator for a "Female" MeSH tag. In columns 2 and 4, the outcome is an indicator for a "Female" tag and *not* a "Male" tag in the MeSH term, indicated by $^{+}$. The "h" superscript in columns 3 and 4 indicates that the MeSH term includes a "Human" tag. Columns 5 and 6 include post-reform specialty trends in column 1 and column 3 regressions, respectively. SY maximal IV size refers to the Stock-Yogo maximal IV size for weak instruments (Stock and Yogo, 2002). All regressions include specialty, district and year fixed effects. Standard errors are clustered by specialty and district. Stars indicate significant differences: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

disposed to a higher acceptance of female inclusive research, which biases difference-in-differences estimates in the opposite direction.

Table 4: Female Tagged Research - Reduced Form

	Probability Female in					
	MeSH Term	MeSH Term^{+}	MeSH Term^{h}	MeSH Term$^{h,+}$	MeSH Term	MeSH Term^{h}
	(1)	(2)	(3)	(4)	(5)	(6)
FemaleShare $\times$ Post	0.127***	0.041	0.127***	0.039	0.163*	0.179**
	(0.044)	(0.040)	(0.046)	(0.041)	(0.085)	(0.089)
Observations	7,128	7,128	7,128	7,128	7,128	7,128
Mean DV	0.68	0.19	0.67	0.19	0.68	0.67
Specialty Trends					Yes	Yes

Notes: This table presents results from reduced form regressions of the pre-reform female share by specialty and district on articles' female relevance. In columns 1 and 3, the outcome is an indicator for a "Female" MeSH tag. In columns 2 and 4, the outcome is an indicator for a "Female" tag and *not* a "Male" tag in the MeSH term, indicated by $^{+}$. The "h" superscript in columns 3 and 4 indicates that the MeSH term includes a "Human" tag. Columns 5 and 6 include specialty trends for the outcomes in columns 1 and 3, respectively. All regressions include specialty, district and year fixed effects. Standard errors are clustered by specialty and district. Stars indicate significant differences: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

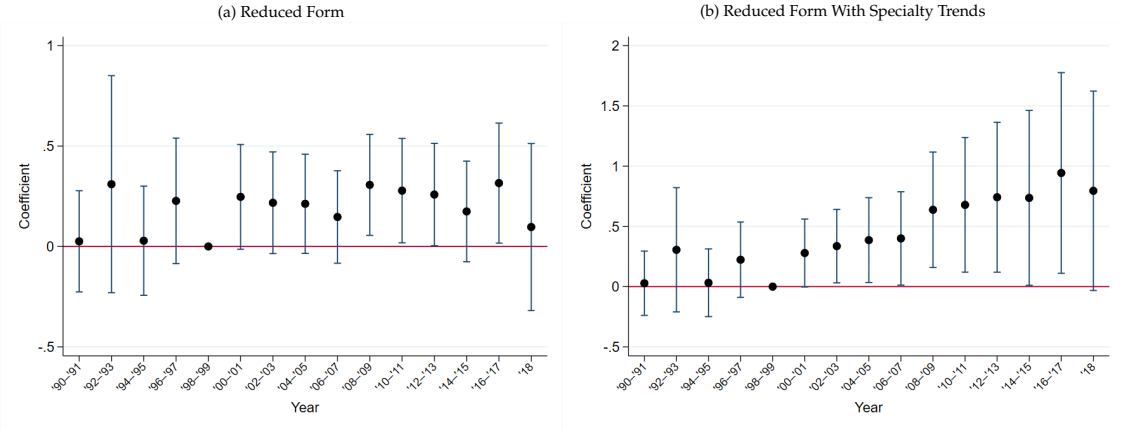
5.3 Event Studies on "Female"-Tags

Figure 4 shows our main reduced form event studies where the outcome is the probability that an article has a "Female" tag, and the main independent variable is the pre-reform female share interacted with time period indicators. In panel (b), we further include specialty trends for the post-period which attenuate the cyclical quality of the event study in panel (a). Overall, the coefficients before the reform are not statistically different from 0, while the coefficients starting in 1999-2000 suggest an increase in the likelihood of the "Female" tag in articles. Event studies with 1 year intervals are presented in Appendix figure A3 and are similar.

The increase in female-relevant research could also be driven by the salience of the court case and subsequent reform, or the higher female shares among students, which could create awareness of gender issues or increase the number of female research assistants. We provide an alternative event study centered around 1992 (figure 5), when medical school admissions became merit-based. The event study includes vertical lines that divide event-time into 3 components: pre-1992,[20] between 1992-1997, and after 1997. Event study coefficients suggests that female-inclusiveness did not increase until 2000, when individuals who graduated in 1997 are expected to have started publishing research.

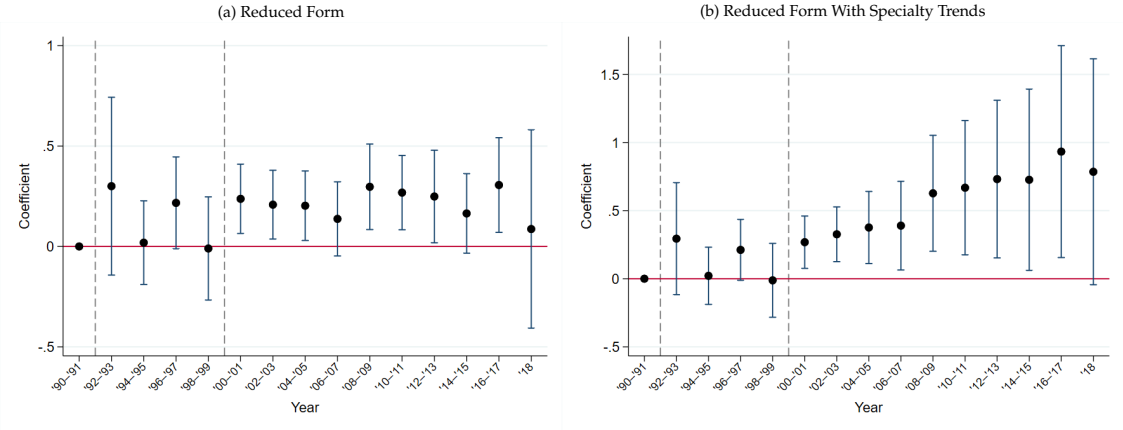
[20] Insufficient observations before 1990 lead to the coefficient for 1988-1989 being dropped.

Figure 4: Event Study of "Female" Tags in Article



Note: The graph plots event studies for the likelihood of a "Female" tag in article based on equation 3 and PubMed data. The event study ranges from 1988 to 2018, and includes fixed effects for specialty, year and district. Standard errors are clustered by specialty and district. Round markers represent point estimates and vertical lines represent confidence intervals. The base period is 1998-1999.

Figure 5: Event Study of "Female"-Tags Around the 1992 Reform



Note: The graph plots event studies for the likelihood of a "Female" tag in article based on equation 3 and PubMed data. The left event study is around the original reform year of 1992; the right event study includes specialty trends. The event studies range from 1988 to 2019, and include fixed effects for specialty, district and year. Standard errors are clustered by specialty and district. Round markers represent point estimates and vertical lines represent confidence intervals. The base period is 1990-1991, one interval before the interval of the reform.

5.4 Descriptive Results on "Female"-Tags in Research on Specific Conditions

Did female-inclusiveness extend to conditions which we know ex-post to affect women differently from men? Women and men are now known to experience conditions such as heart conditions and dementia differently (WebMD, 2024). These specific conditions can be identified from article MeSH tags, but unfortunately regression sample sizes shrink substantially for sub-groups of diseases. Instead we plot the 30 conditions with the greatest change in their share of "Female"-tagged research after the reform (figure A4). Research which includes MeSH tags such as "Cardiac Imaging Techniques" or "Aorta" increase by 17 p.p. and 15 p.p., respectively. Another example is of "Aberrant motor behavior in dementia", which became 9 p.p. more female-inclusive. While the graph does not provide causal evidence, these conditions are studied in specialties that experienced an increase in their female share of doctors, such as 11 p.p. in Cardiology and 31 p.p. in Neurology. Additionally, conditions such as "deficiency diseases" which are more prevalent in developing than developed countries (Ritchie and Moser, 2023),[21] and diarrhea, one of the top MeSH terms in Pakistan (table A1), also became more inclusive, providing examples of research pertinent to developing countries also changing.

5.5 Do "Female"-Tag Increases Reflect Topics or Study Participants?

While our primary measure of female inclusiveness is through the gender tags with article MeSH terms, notably, these do not separate between article topics, or participants. The observed increase in "Female" tags could be due to more research on female-centric topics, or through general topics with more female participants, or both.

We use text analysis of article abstracts to separate the likely reason that an article received the "Female" tag by constructing a lexicon of words that are most associated with

[21] An example is anemia; 45% of pregnant women in South Asia have anemia, compared to 10% in North America.

the tag. Our algorithm is based on Koning et al. (2021) and Monroe et al. (2008) and calculates the log-odds ratio for how likely a word ("token") is to occur in "Female" tagged articles, for all words in article abstracts.[22] The resulting log-odds ratio is converted to a z-score that measures how disproportionately a token appears in "Female"-tagged articles. Consistently with Koning et al. (2021), words with high z-scores (3.891 or more, i.e., $p < 0.0001$) are retained in the lexicon. About 76 percent of articles which have any word in the lexicon also have "Female" tags, so that our word list is largely but not exactly a subset of "Female" tagged articles, consistent with PubMed assigning "Female" tags based on a broader analysis of articles. Details on the method are described in the Appendix.

Alternatively, we also use the lexicon provided in Koning et al. (2021) of about 200 words that are most frequently associated with the "Female" tag in articles and in patents.[23] These words are based on the universe of PubMed articles and patent files, and allow us to consider female words that may be new to Pakistani research.

Each methodology creates a lexicon which includes both medical terms linked to the anatomy or related health conditions (e.g., uterine, lactating and kidney) and gender descriptor words such as "women" or "females". Manually parsing through abstracts suggests that the latter almost always refers to the gender of study participants.[24] The complement of this subset of participant nouns in the full lexicon are words that depict anatomy and/or health conditions.[25] We run both the full lexicon and each subset across abstracts of all articles to classify the possible source, or sources, of the "Female" tag.

One concern is that our coding may miss mentions of participants in the article text that are not included in the abstract. To attenuate this concern, we combine our measure

[22] In other words, it compares abstracts with a "Female" tag to those without a "Female" tag; for each word it asks "Is this word unusually more common in female-tagged abstracts than in non-female-tagged abstracts".

[23] We use all the words provided, which includes the 100 most common words in articles and patents each, for a total of 200 words.

[24] Examples of abstracts with female participants include a gastronomical study which presented the "case of a 58-year-old woman with a GCF (Gastrocolic fistulas)", or a study in Pediatrics which "infratentorial brain abscesses in a ten month old girl".

[25] We provide a few examples of these articles in the Appendix A.2.

of gender descriptor words with data from Feldman et al. (2019), who developed an algorithm to extract data from summary statistics tables in over 25,000 papers on PubMed with registered clinical trials.[26]

We find that "Female" tag increases after the reform in more treated fields primarily coincide with increases in gender descriptor words, which signal a rise in the inclusion of female subjects. A 1 percentage point increase in the female share led to a 0.71 p.p. increase in the mention of female gender descriptors (column 3 of table 5). In column 6, we use the words provided by Koning et al. (2021), and in column 7, additionally include the participant measure from Feldman et al. (2019). Both columns show increases similar to those in column 3. The Kleibergen–Paap rk Wald F statistic is 9.76, exceeding the Stock–Yogo critical value of 7.03 for 10% maximal IV size.

Coefficients for female words overall and for anatomy/conditions are positive but smaller and statistically insignificant. This may be because certain topics are limited in how exclusively female-focused they can be, or may require a greater change in research agendas. For example, several female-focused topics explicitly involve studies of the female anatomy, and may not be easily combined with studies of a general interest topic such as diabetes. Instead, a less costly margin of creating female-inclusiveness is to include female subjects in general-interest research topics.

5.6 Female-Inclusiveness of Articles by Author Gender

As new individuals entered medical school after the reform, the selection of medical doctors is likely to have changed, and selection differences could affect the production of research. To mitigate selection effects, we restrict our analysis to articles with at least 1 researcher who was active before 1997 (incumbents).[27] This holds pre-reform authors constant and examines the projects that these authors participated in as gender ratios

[26] This data is based on the entire PubMed archive limited to records that are related to clinical trials and further subset to those for which Semantic Scholar had full PDFs.

[27] There were an insufficient number of articles with only incumbent authors.

Table 5: Text Analysis of Article Abstracts

	Female Word	Anatomy, Condition	Participant	Female^k Word	Anatomy ^{k} Condition	Participant^k Word	Participant
	DDIV	DDIV	DDIV	DDIV	DDIV	DDIV	DDIV
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Δ Female Share $\times$ Post	0.214	0.193	0.714***	0.104	-0.114	0.688**	0.716***
	(0.192)	(0.168)	(0.231)	(0.248)	(0.300)	(0.293)	(0.236)
Observations	7,517	7,517	7,517	7,517	7,517	7,517	7,517
Sanderson-Windmeijer F-stat	33.01	33.01	33.01	33.01	33.01	33.01	33.01
Kleibergen-Paap F-stat	9.76	9.76	9.76	9.76	9.76	9.76	9.76
SY 10% maximal IV size	7.03	7.03	7.03	7.03	7.03	7.03	7.03
A-R p-value	0.53	0.37	0.00	0.28	0.18	0.01	0.00
Mean DV	0.26	0.19	0.13	0.14	0.10	0.13	0.14

Notes: This table presents results from DDIV regressions based on text analysis of article abstracts. The superscript ^{k} indicates that the lexicon provided by Koning et al. (2021) is used. In columns 1 and 4, the outcome is an indicator for a match with any female word (including both words pertaining to anatomy and related conditions, and gender descriptor words), in columns 2 and 5, it is an indicator for a match with a word describing the female anatomy or related health conditions, and in columns 3 and 6, a match with a gender descriptor word. In column 7, the outcome additionally includes the measure of participants used in column 3 and also any papers identified to include female participants in the data provided by Feldman et al. (2019). SY maximal IV size refers to the Stock-Yogo maximal IV size for weak instruments (Stock and Yogo, 2002). All regressions include district, specialty and year fixed effects. Standard errors are clustered by specialty and district. Stars indicate significant differences: *p $<$ 0.10, **p $<$ 0.05, ***p $<$ 0.01.

changed.

Incumbent authors in more treated units became increasingly likely to participate in female-inclusive research. A 1 p.p. higher treatment raises the likelihood that incumbents' projects are female-relevant by 0.96 p.p. (table 6, column 1).[28] While we cannot disentangle whether pre- or post-reform researchers drive the overall increase in female-relevant research, this result indicates the presence of a treatment effect on incumbents' research.

We can also use this analysis to examine if female or male incumbent authors participated in female-relevant research. The increase for female incumbents was strong and statistically significant: a 1 p.p increase in treatment led pre-reform female authors to become 0.76 p.p. more likely to participate in female-relevant research after the reform (column 2). The estimated effect for male researchers is negative and statistically insignif-

[28]The number of observations grows because we match authors to articles, and the same article may be matched to different authors.

icant (column 3), however, the difference between estimates for men and women are not statistically significant (column 4).[29]

Together, these results imply two additional features of research production under gender composition changes. First, incumbent authors who were already producing research and whose selection is less likely to change start to participate in inclusive research. Second, while we cannot rule out differences between male and female incumbents, the increase in participation in inclusive projects is especially strong for female incumbents.

Table 6: Articles by Authors Active Before 1997

	Probability Female in MeSH Term			
	All Incumbents	Female Incumbents	Male Incumbents	Triple Difference
	DDIV	DDIV	DDIV	DDIV
	(1)	(2)	(3)	(4)
Δ Female Share $\times$ Post	0.961*	0.763**	-8.655	-4.634
	(0.492)	(0.371)	(49.399)	(22.118)
Δ Female Share $\times$ Post $\times$ Female				5.505
				(22.511)
Observations	4,279	1,146	2,677	3,823
Sanderson-Windmeijer F-stat	18.39	18.39	0.05	0.16
Kleibergen-Paap F-stat	2.10	9.40	0.01	0.01
SY 10% maximal IV size	7.03	7.03	7.03	7.03
A-R p-value	0.00	0.00	0.12	0.01
Mean DV	0.70	0.73	0.69	0.70
Author FE	✓	✓	✓	✓

Notes: This table presents results, at the author-article level, from DDIV regressions of the probability that an article's MeSH term includes the "Female" tag on the change in female share of doctors. In column 1, the sample is of articles with a matched pre-reform researcher. In column 2, we restrict the sample to articles by pre-reform female authors, and in column 3, articles by pre-reform male authors. Column 4 represents a triple difference where the main interaction of the change in female share and post is further interacted with pre-reform author gender. SY maximal IV size refers to the Stock-Yogo maximal IV size for weak instruments (Stock and Yogo, 2002). All regressions include author and year fixed effects in the second stage and specialty and district fixed effects in the first stage. Standard errors are clustered by specialty and district. Stars indicate significant differences: *p $<$ 0.10, **p $<$ 0.05, ***p $<$ 0.01.

[29] Our triple difference estimates in column 4 may be underpowered from dropping observations where no author gender could be assigned.

6 Robustness

Non-doctor researchers as a placebo group: Our results could be due to overall trends in the popularity of female-inclusive research or changes in the way articles are indexed by indexers. Another factor may be the founding of the Higher Education Commission in 2002, which centralized and increased resources for research.[30]

To address these concerns, we leverage non-medical doctor researchers as a placebo group. Non-medical researchers were exposed to similar trends as medical doctors, but should not be affected by the medical school reform. We randomly assign the change in female share by district to articles with affiliations in non-medical subjects.[31] The results suggest that research produced by non-medical doctors did not become more female-inclusive after the reform (panel A, table A3).

Placebo Year: We also test the possibility that our effects are affected by time trends or other omitted factors that remain uncontrolled for, by using a placebo year as our "treatment year". Instead of the year 2000, we use the year 1990 as the year that articles affected by the "reform" would be affected in the placebo exercise.[32] Coefficients on the interaction term with this exercise are positive but statistically insignificant (panel B,table A3), which suggests that time trends may not be the main driving force of our results.

Trends in Districts and Specialty: In this robustness check, we relax the assumption that districts and specialties are on the same trend. We control for a district trend and specialty trend term. The results in table A3 indicate that while coefficients are slightly smaller, they remain significant.

[30] https://www.iiu.edu.pk/quality-assurance/quality-assurance/
To the best of our knowledge, this funding did not target female-relevant research, nor units with a low pre-reform female share of doctors.

[31] We were able to identify subject fields for 60% of all articles that do not match with medical specialties, including Chemistry, Physics, Genetics and Statistics.

[32] We change the sample to articles between the years 1975 to 1999 for these regressions. Note that the sample size is smaller before 1987.

Small Clusters and Outliers: To mitigate the concern that estimates are influenced by small clusters, we restrict regression clusters to have a minimum size of 5 observations. The results in panel C (table A3) show that coefficients remain close to the main coefficient size of 0.5 p.p. change for a 1 p.p. treatment increase.

Further, since some units have a zero pre-reform female share, we examine if a few units with zero women before the reform drive the effect. We now additionally include the interaction between an indicator for the presence of any women in the unit before the reform with an indicator for post-reform periods. The estimates in table A4 indicate that the relationship persists among units with some female representation before the reform. The main results therefore do not appear to be driven solely by units that initially had no female doctors.

7 Mechanisms

We now examine mechanisms for the increase in female inclusiveness. First, since research inclusiveness increased after about 2000 (figure 5), the gender composition of the student body alone does not seem to drive female inclusive research. Instead, we see an effect *after* these changes permeate downstream into the workforce. Second, while we find an increase in female authors after the reform, including junior female authors (figure A6), the increase is similar across treatment intensity (figure A5). This may indicate barriers to research for junior female doctors in units with lower pre-existing female shares.[33]

However, we find that incumbent *collaborations* with junior authors may be an important channel for research becoming inclusive. Similar to Truffa and Wong (2025), we measure interactions that produce female-relevant research as the probability that an in-

[33]More treated units historically had a smaller share of female doctors. Therefore, consistent with literature on the importance of mentorship in academia (Blau et al., 2010; Lundberg and Stearns, 2019), there may be slow catchup in the pipeline of female researchers.

cumbent researcher publishes a female-inclusive study *and* that it is coauthored with at least one junior author.[34] As mentioned earlier, junior authors often play a significant role in recruiting participants for studies (Vas et al., 2021). Following section 5, we utilize female participant words as our main measure of female inclusiveness in this section, and also examine "Female" tags in Appendix A.

In table 7, we compare incumbent research coauthored with at least 1 junior researcher (columns 1-3) as opposed to no junior researchers (columns 4-6). For female incumbents, the DDIV estimate in column 1 is more than twice as large as in column 4 (1.34 p.p. vs. 0.62 p.p. for a 1 p.p increase in treatment after the reform). Together, the coefficients suggest that interactions between female incumbents and junior researchers may have played a role in producing female inclusive research.

In contrast, for men, the DDIV estimates in columns 2 and 5 are nearly identical. Consistently, in columns 3 and 6 the triple difference, although not statistically significant, indicates a larger effect for female incumbents when coauthoring with junior authors, and a smaller effect without, implying a different pattern of results for female incumbents. If, instead, our results were due to an increase in Pakistani women's participation in healthcare, or their seeking out doctors to participate in research studies, we may have expected to see similar results for male and female incumbents.

Finally, in table A5, we see a consistent pattern of coefficients for producing a "Female"-tagged paper and coauthoring with junior researchers, even though our estimates are noisier. Overall, these results suggest that collaborations between female incumbents and junior authors may play a role in increasing research female-relevance.

[34]Junior authors are defined as authors who have not published any research in previous years.

Table 7: Incumbent and Junior Author Collaborations

	Female Participant Word					
	With Junior Coauthor			No Junior Coauthor		
	Female Incumbents	Male Incumbents	Triple Difference	Female Incumbents	Male Incumbents	Triple Difference
	DDIV	DDIV	DDIV	DDIV	DDIV	DDIV
	(1)	(2)	(3)	(4)	(5)	(6)
Δ Female Share $\times$ Post	1.342***	1.862	1.115	0.622*	1.838	2.428
	(0.321)	(2.161)	(1.534)	(0.372)	(1.504)	(3.071)
Δ Female Share $\times$ Post $\times$ Female			0.406			-2.059
			(1.613)			(3.420)
Observations	1,110	2,627	3,737	1,110	2,627	3,737
Sanderson-Windmeijer F-stat	20.34	1.05	0.25	20.34	1.05	0.25
Kleibergen-Paap F-stat	10.89	0.62	0.06	10.89	0.62	0.06
SY 10% maximal IV size	7.03	7.03	7.03	7.03	7.03	7.03
A-R p-value	0.00	0.16	0.00	0.04	0.02	0.06
Mean DV	0.04	0.02	0.03	0.04	0.02	0.03
Author FE	✓	✓	✓	✓	✓	✓

Notes: This table presents results from DDIV regressions of the probability that an article's MeSH term includes a Female gender descriptor and was coauthored with at least 1 junior author. In columns 1 and 4, the sample is restricted to articles with a matched incumbent female authors, and in columns 2 and 5, to matched incumbent male authors. Columns 3 and 6 represent IV triple differences models. SY maximal 10% IV size refers to the Stock-Yogo maximal IV size for weak instruments (Stock and Yogo, 2002). All regressions include author, year, specialty and district fixed effects. Standard errors are clustered by specialty and district. Stars indicate significant differences: *p $<$ 0.10, **p $<$ 0.05, ***p $<$ 0.01.

8 Non-PubMed Journal Articles

Since not all Pakistani research is published in PubMed, we collected data on articles published in non-PubMed indexed journals from PakMediNet, similar to Research Gate, between 1993 and 2019.[35] While these articles do not have the "Female" tag or MeSH terms, we use article abstracts to quantify female inclusiveness using female words from our earlier analysis based on PubMed articles.

Female words increased in non-PubMed articles by 1 p.p. for a 1 p.p. increase in treatment. The result suggests that female inclusiveness increases were not restricted to highly competitive journals. Results for participant words and words on conditions/anatomy are more noisily estimated, potentially because the list of words used is based on PubMed articles.

9 Conclusion

This paper provides causal evidence on how gender composition in professional fields shapes the production of scientific knowledge. Exploiting the removal of a gender-based restriction in Pakistani medical school admissions, we show that a plausibly exogenous increase in the share of women in medicine led to research becoming more female-inclusive, and without crowding out male inclusiveness.

Our evidence suggests that increases in female-inclusiveness were driven more strongly by the inclusion of female participants in studies than by explicitly addressing female-focused questions. Pre-reform authors also significantly increased participation in female-relevant research, providing evidence of a treatment effect. Furthermore, the rise of female-inclusive research in which incumbents coauthored with a junior researcher im-

[35] We collected all articles on the website that are published in non-PubMed indexed journals. While there is an incentive for academics to promote their research, journal articles are voluntarily submitted to PakMediNet, so that some articles could be missing on the website. Examining the data that we extracted from the website does not indicate that articles are missing systematically from a given year or journal. The first year that data was available for was 1993.

Table 8: Non-PubMed Articles

	Female Word	Anatomy & Disease	Participant
	(1)	(2)	(3)
Δ Female Share $\times$ Post	1.029**	0.122	-0.270
	(0.518)	(0.253)	(0.415)
Observations	10,503	10,503	10,503
Sanderson-Windmeijer F-stat	15.53	15.53	15.53
Kleibergen-Paap F-stat	6.35	6.35	6.35
SY 10% maximal IV size	7.03	7.03	7.03
A-R p-value	0.00	0.77	0.18
Mean DV	0.59	0.30	0.48

Notes: This table presents results for articles published in journals that are not indexed by PubMed. In columns 1, 2, and 3, the outcome is the probability that an article's abstract contains a female word, a word related to female anatomy and conditions, and a potential participant word, respectively. SY maximal IV size refers to the Stock-Yogo maximal IV size for weak instruments (Stock and Yogo, 2002). All regressions include specialty, district and year fixed effects. Standard errors are clustered by specialty and district. Stars indicate significant differences: *p $<$ 0.10, **p $<$ 0.05, ***p $<$ 0.01.

plies that peer interactions with junior post-reform authors may have been a mechanism.

Taken together, our findings suggest that gender ratios in scientific production influence the knowledge that is generated. Policies that expand research access for different groups can thus enhance not only equity but also the evidence base for medical research. Meanwhile, continuing to study how research is affected by the demographics of the workforce is essential for evaluating the long-term effects of diversity initiatives in science and higher education.

References

AAMC (2023). A Brief timeline of women in medicine. https://www.aamc.org/news/brief-timeline-women-medicine.

Abbas, A., Q. Abbas, N. U. R. Siddiqui, S. Ishaque, S. Muneer, H. Jurair, A. R. Ahmed, A. Haque, Z. Cheema, and A. Afzal (2024). Publications in pediatric critical care medicine: Trends from pakistan. *Journal of Critical Care 81*, 154739.

Acemoglu, D. (2002). Directed technical change. *The review of economic studies 69*(4), 781–809.

Acemoglu, D. (2023). Distorted innovation: does the market get the direction of technology right? In *AEA Papers and Proceedings*, Volume 113, pp. 1–28. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

Acemoglu, D., P. Aghion, L. Bursztyn, and D. Hemous (2012). The environment and directed technical change. *American economic review 102*(1), 131–166.

Acemoglu, D. and P. Restrepo (2022). Demographics and automation. *The Review of Economic Studies 89*(1), 1–44.

Aghion, P., U. Akcigit, A. Hyytinen, and O. Toivanen (2017). The social origins of inventors. Technical report, National Bureau of Economic Research.

Aghion, P., A. Dechezleprêtre, D. Hemous, R. Martin, and J. Van Reenen (2016). Carbon taxes, path dependency, and directed technical change: Evidence from the auto industry. *Journal of political economy 124*(1), 1–51.

Ahmad, S., S. Ahmad, U. Ahmad, H. A. Cheema, N. Iqbal, A. Shahid, B. Malik, A. Siddiqui, H. Jaffar, U. Ghani, et al. (2022). An assessment of publishing practices and barriers faced by medical students to conduct research: A cross-sectional study from pakistan. *Health Science Reports 5*(5), e831.

Airoldi, A. and P. Moser (2024). Inequality in science: Who becomes a star? Technical report, National Bureau of Economic Research.

Akcigit, U., J. Grigsby, and T. Nicholas (2017). Immigration and the rise of american ingenuity. *American Economic Review 107*(5), 327–331.

Alsan, M., R. A. Campbell, L. Leister, and A. Ojo (2025). Investigator racial diversity and clinical trial participation. *Journal of Health Economics 100*, 102968.

AMA (2015). How medical specialties vary by gender. Specialty profiles, American Medical Association (AMA).

American Heart Association (2024). 7 Things to Know about How Stroke is Different for Women. https://www.heart.org/en/news/2024/05/22/7-things-to-know-about-how-stroke-is-different-for-women.

American Psychological Association (2023). The push for more equitable research is changing the field. https://www.apa.org/monitor/2023/01/trends-inclusivity-psychological-research.

Andrabi, T. and N. Singh (2015). Almost breaking the glass ceiling: open merit admissions in medical education in pakistan. *Mimeo*.

Angrist, J. (2002). How do sex ratios affect marriage and labor markets? evidence from america's second generation. *The Quarterly Journal of Economics 117(3)*, 997–1038.

Angrist, J. D. and G. W. Imbens (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American statistical Association 90*(430), 431–442.

Aqeel, F. (2024). Removing barriers to entry in medicine: Evidence from pakistan. *The World Bank Economic Review*, lhad039.

Association of American Medical Colleges (AAMC) (2022). Demographic Characteristics and Work Experiences of Physician Scientists in the U.S. https://www.aamc.org/data-reports/report/demographic-characteristics-and-work-experiences-physician-scientists-us.

Association of American Medical Colleges (AAMC) (2023). Why we know so little about women's health. https://www.aamc.org/news/why-we-know-so-little-about-women-s-health.

Bell, A., R. Chetty, X. Jaravel, N. Petkova, and J. Van Reenen (2019). Who becomes an inventor in america? the importance of exposure to innovation. *The Quarterly Journal of Economics 647*, 713.

Blau, F. D., J. M. Currie, R. T. A. Croson, and D. K. Ginther (2010). Can mentoring help female assistant professors? interim results from a randomized trial. *American Economic Review 100*(2), 348–352.

British Heart Foundation (2016). Misdiagnosis of Heart Attacks in Women. https://www.bhf.org.uk/what-we-do/news-from-the-bhf/news-archive/2016/august/women-are-50-per-cent-more-likely-than-men-to-be-given-incorrect-diagnosis-following-a-heart-attack.

Bryan, K. and H. Williams (2021). Innovation: market failures and public policies. *NBER Working Paper* (w29173).

Callaway, B., A. Goodman-Bacon, and P. H. Sant'Anna (2021). Difference-in-differences with a continuous treatment. *arXiv preprint arXiv:2107.02637*.

Chari, A. and P. Goldsmith-Pinkham (2017). Gender representation in economics across topics and time: Evidence from the nber summer institute. Technical report, National Bureau of Economic Research.

Cook, L. D. and C. Kongcharoen (2010). The idea gap in pink and black. Technical report, National Bureau of Economic Research.

De Loecker, J., P. K. Goldberg, A. K. Khandelwal, and N. Pavcnik (2016). Prices, markups, and trade reform. *Econometrica 84*(2), 445–510.

Duflo, E. (2001). Schooling and labor market consequences of school construction in indonesia: Evidence from an unusual policy experiment. *American Economic Review 91, no. 4*, 795–813.

Einiö, E., J. Feng, and X. Jaravel (2019). Social push and the direction of innovation. *Available at SSRN 3383703*.

Feldman, S., W. Ammar, K. Lo, E. Trepman, M. van Zuylen, and O. Etzioni (2019). Quantifying sex bias in clinical studies at scale with automated data extraction. *JAMA network open 2*(7), e196700–e196700.

Ghaffar, A., S. Zaidi, H. Qureshi, and A. Hafeez (2013). Medical education and research in pakistan. *The Lancet 381*(9885), 2234–2236.

Goldsmith-Pinkham, P., I. Sorkin, and H. Swift (2020). Bartik instruments: What, when, why, and how. *American Economic Review 110*(8), 2586–2624.

Graves Jr, J. L., M. Kearney, G. Barabino, and S. Malcom (2022). Inequality in science and the case for a new agenda. *Proceedings of the National Academy of Sciences 119*(10), e2117831119.

Greenberg, K., M. Wasserman, and E. A. Weber (2024). The effects of gender integration on men: Evidence from the us military. Technical report, National Bureau of Economic Research.

Gualtierotti, R. (2025). Bridging the gap: Time to integrate sex and gender differences into

research and clinical practice for improved health outcomes. *European Journal of Internal Medicine*.

Gupta, H. (2022). Do female researchers increase female enrollment in clinical trials? Technical report, Working paper.

Gupta, H., A. B. Jena, and E. C. Sun (2025). Gender diversity of research teams and clinical trial enrollment. *JAMA Network Open 8*(10), e2537667.

Hassan, W. and A. Ara (2021). Pakistan has the 2nd highest growth rate in medical sciences research publications for the year 2019-2020: Comparison with 46 countries: Bibliometric analysis of medical sciences research in pakistan for the year 2019-2020. *Proceedings of the Pakistan Academy of Sciences: B. Life and Environmental Sciences 58*(3), 43–54.

Heiser, Stuart (2017). More Women Than Men Enrolled in U.S. Medical Schools in 2017. https://www.aamc.org/news/press-releases.

Jensen, K., B. Kovács, and O. Sorenson (2018). Gender differences in obtaining and maintaining patent rights. *Nature biotechnology 36*(4), 307–309.

Karpowitz, C. F., S. D. O'Connell, J. Preece, and O. Stoddard (2024). Strength in numbers? gender composition, leadership, and women's influence in teams. *Journal of Political Economy 132*(9), 3077–3114.

Koffi, M. and M. Marx (2026). Cassatts in the attic: Is there a gender gap in the commercialization of science? *American Economic Journal: Applied Economics*.

Koning, R., S. Samila, and J.-P. Ferguson (2020). Inventor gender and the direction of invention. In *AEA Papers and Proceedings*, Volume 110, pp. 250–254. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

Koning, R., S. Samila, and J.-P. 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.

Lapidow, A. and P. Scudder (2019). Shared first authorship. *Journal of the Medical Library Association: JMLA 107*(4), 618.

LUMS University (2024). Shirin Munir v Government of Punjab [PLD 1990 SC 295]. https://sahsol.lums.edu.pk/node/17087.

Lundberg, S. and J. Stearns (2019). Women in economics: Stalled progress. *Journal of Economic Perspectives 33*(1), 3–22.

Masood, A. (2019). The influence of marriage on women's participation in medicine: The case of doctor brides of pakistan. *Sex Roles 80, no. 1*, 1–18.

Mayo Clinic (2022). Heart disease in women: Unsuspected? Overlooked? Ignored? https://mcpress.mayoclinic.org/women-health/part-1-heart-disease-in-women-unsuspected-overlooked-ignored/.

Merone, L., K. Tsey, D. Russell, and C. Nagle (2022). Sex inequalities in medical research: a systematic scoping review of the literature. *Women's Health Reports 3*(1), 49–59.

Michelman, V. and L. Msall (2021). Sex, drugs, and r&d: missing innovation from regulating female enrollment in clinical trials. Technical report, Working paper.

Monroe, B. L., M. P. Colaresi, and K. M. Quinn (2008). Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis 16*(4), 372–403.

National Library of Medicine (2024). Exploring MeSH for Expert Searching. https://www.nlm.nih.gov/oet/ed/pubmed/mesh/mod03/04-100.html.

National Library of Medicine (2025a). MeSH Record Types. https://www.nlm.nih.gov/mesh.

National Library of Medicine (2025b). Use of MeSH in Cataloging for Metadata Management Program (MMP). https://www.nlm.nih.gov/tsd/cataloging/MeSHCatPractices.html.

NOVOTECH (2021). How to Recruit Participants and How Participants can Take Part. https://novotech-cro.com/faq/how-recruit-participants-and-how-participants-can-take-part.

of Medical Journal Editors, I. C. et al. (1997). Uniform requirements for manuscripts submitted to biomedical journals. *Pathology 29*(4), 441–447.

Penny, M., R. Jeffries, J. Grant, and S. C. Davies (2014). Women and academic medicine: a review of the evidence on female representation. *Journal of the Royal Society of Medicine 107*(7), 259–263.

Qureshi, M. A., M. S. Qureshi, and M. R. Khanani (2013). Health and research in pakistan. *Lancet 382*(1245), 62091–7.

Ramakrishnan, A., D. Sambuco, and R. Jagsi (2014). Women's participation in the medical profession: insights from experiences in japan, scandinavia, russia, and eastern europe. *Journal of women's health 23*(11), 927–934.

Rehman, S. H., B. Majid, S. A. Syed, and M. A. Qureshi (2021). Quantitative analyses of skin cancer research in pakistan. *Pakistan Journal of Medical Sciences 37*(2), 582.

Ritchie, H. and M. Moser (2023). Micronutrient Deficiency.

Sarica, S. and J. Luo (2021). Stopwords in technical language processing. *Plos one 16*(8), e0254937.

Shankar, P., K. Singh, and S. Singh (2012). Possible impact of increase in female medical student admissions in nepal: Findings from a qualitative study among medical undergraduates. *The Australasian Medical Journal 5*(3), 184.

Siddique, A., A. Zafar, and T. Irfan (2022). Comprehensive analysis of clinical trials in pakistan from september 1992 to february 2019. *Journal of Ayub Medical College Abbottabad 34*(3), 468–473.

Stock, J. H. and M. Yogo (2002). Testing for weak instruments in linear iv regression.

Supreme Court of Pakistan (1989). Shirin Munir v Government of Punjab [PLD 1990 SC 295]. https://www.scribd.com/document/818358123/Shirin-Munir-v-Government-of-Pakistan-P-L-D-1990-Supreme-Court-295.

The Conference Forum (2024). Making Clinical Trials More Inclusive for Sex- and Gender-Diverse People. https://theconferenceforum.org/editorial/making-clinical-trials-more-inclusive-for-sex-and-gender-diverse-people.

The Harvard Gazette (2024). Why are Women Twice as Likely to Develop Alzheimer's as Men. https://news.harvard.edu/gazette/story/2025/07/why-are-women-twice-as-likely-to-develop-alzheimers-as-me.

Topalova, P. and A. Khandelwal (2011). Trade liberalization and firm productivity: The case of india. *Review of economics and statistics 93*(3), 995–1009.

Torvik, V. I. (2015). Mapaffil: A bibliographic tool for mapping author affiliation strings to cities and their geocodes worldwide. *D-Lib Magazine: The Magazine of the Digital Library Forum 21*(11-12).

Torvik, V. I. and N. Smalheiser (2021, apr). Author-ity 2018 - pubmed author name disambiguated dataset. University of Illinois Urbana-Champaign.

Torvik, V. I. and N. R. Smalheiser (2009). Author name disambiguation in medline. *ACM Transactions on Knowledge Discovery from Data 3*(3), 11.

Truffa, F. and A. Wong (2025). Undergraduate gender diversity and the direction of scientific research. *American Economic Review 115*(7), 2414–2448.

Tybaert, S. (2014, Nov-Dec). Medline data changes — 2015. *NLM Tech Bull* (401), e5.

University of Karachi, Pakistan (2023). ICCBS Patents. https://www.iccs.edu/page-patent.

Vas, A., P. D'sa, S. Gokhale, T. Agarwal, G. L. Roberts, K. Mohanty, and G. Roberts (2021). Trainee principal investigator could improve recruitment in trauma trials: review of literature and experience from a trauma center. *Cureus 13*(10).

Virtual Health Library (2024). MeSH Class 3 descriptors: "check tags". https://decs.bvsalud.org/en/class-3-descriptors-check-tags.

WebMD (2024). Conditions That Affect Men and Women Differently. https://www.webmd.com/a-to-z-guides/ss/slideshow-conditions-affect-men-women-differently.

Wing, C., M. Yozwiak, A. Hollingsworth, S. Freedman, and K. Simon (2024). Designing difference-in-difference studies with staggered treatment adoption: key concepts and practical guidelines. *Annual Review of Public Health 45*.

World Health Organization (2026). Number of clinical trials by year, country, who region and income group (1999-2025).

Yale School of Medicine (2023). How is Heart Disease Different in Women. https://medicine.yale.edu/news-article/how-is-heart-disease-different-in-women/.

A Appendix

Examples of abstracts from "Female"-tagged PubMed articles:

Article 1: *Diabetes mellitus constitutes one of the most important problems in developing and non-developing countries. The purpose of the study to estimate the concentrations of Chromium and Manganese in diabetic and non-diabetic population of Hazara division. The cross sectional comparative study was carried out on one hundred blood samples of Type 2 Diabetic patients collected non-randomly from Ayub Teaching Hospital and one hundred normal healthy controls from Women Medical College Abbottabad from September 2014 to April 2015.*

Article 2: *To assess the causes of low tetanus toxoid (TT) vaccination coverage in pregnant women a mixture of quantitative and qualitative methods were adopted at the community, primary health care delivery and management levels in Lahore district, Pakistan. Out of a random sample of 362 women who had delivered during the previous 3 months, 87 percent recalled receiving 2 doses of TT. The main reasons for non-vaccination were poor knowledge about the importance of TT (32 percent of women) or the place and time to get vaccinated (18 percent). According to the managers and primary health care medical officers, the main reasons for low coverage were lack of awareness about the importance of vaccination among the public and misconceptions about TT vaccination (e.g. that it was a contraceptive).*

Article 3: *Reported are the results of a community-based prospective study in four urban squatter settlements in Karachi that was carried out to assess the incidence of and risk factors for intrauterine growth retardation. The incidence of term intrauterine growth retardation was 24.4 percent among 738 singleton births. The socioeconomic and biological risk factors that were found to be statistically significant in a bivariate analysis were included in a logistic regression model to assess their independent effects. The major risk factors were low level of maternal education, paternal unemployment, consanguinity, short birth-to-conception intervals, short maternal stature, and low maternal weight. The population risk estimates suggest the desirability of public health interven-*

tions to improve maternal weight and birth spacing and of improvements in socioeconomic conditions, especially maternal education. Public education programs to discourage consanguineous marriages should also be considered.

A.1 Constructing a Lexicon for Abstract Text Parsing

To identify words associated with the "Female" tag, we use the method in Monroe et al. (2008) and Koning et al. (2021). The method uses a log-odds ratio to identify words that are most associated with "Female"-tagged articles and less likely to be associated with non-"Female" tagged articles. Since the log-odds ratio alone could over-value rare words, a small noise factor is added to make these rare words less influential (known as Dirichlet prior smoothing). The formal smooth log-odds ratio is calculated as:

$$\delta_w = \left[\log \frac{f_w + \alpha}{(N - f_w) + \alpha_0}\right] - \left[\log \frac{f'_w + \alpha}{(N' - f'_w) + \alpha_0}\right].$$

where f_w is a count of word w in "Female"-tagged patents, f'_w is a count of word w in the comparison group (non-"Female" tagged patents), N, N' are total token counts in each group, α is a small, artificial count (called a Dirichlet pseudo-count) added to the real count of word w, and α_0 is a total prior mass (e.g., sum of pseudo-counts over the vocabulary). The latter two help to shrink the log-odds ratio of rare words toward 0.

The variance of the log-odds difference is given by:

$$\sigma^2(\delta_w) = \frac{1}{f_w + \alpha} + \frac{1}{(N - f_w) + \alpha_0} + \frac{1}{f'_w + \alpha} + \frac{1}{(N' - f'_w) + \alpha_0}$$

Finally, we construct the standardized Z-score for each word as as:

$$z_w = \frac{\delta_w}{\sqrt{\sigma^2(\delta_w)}}$$

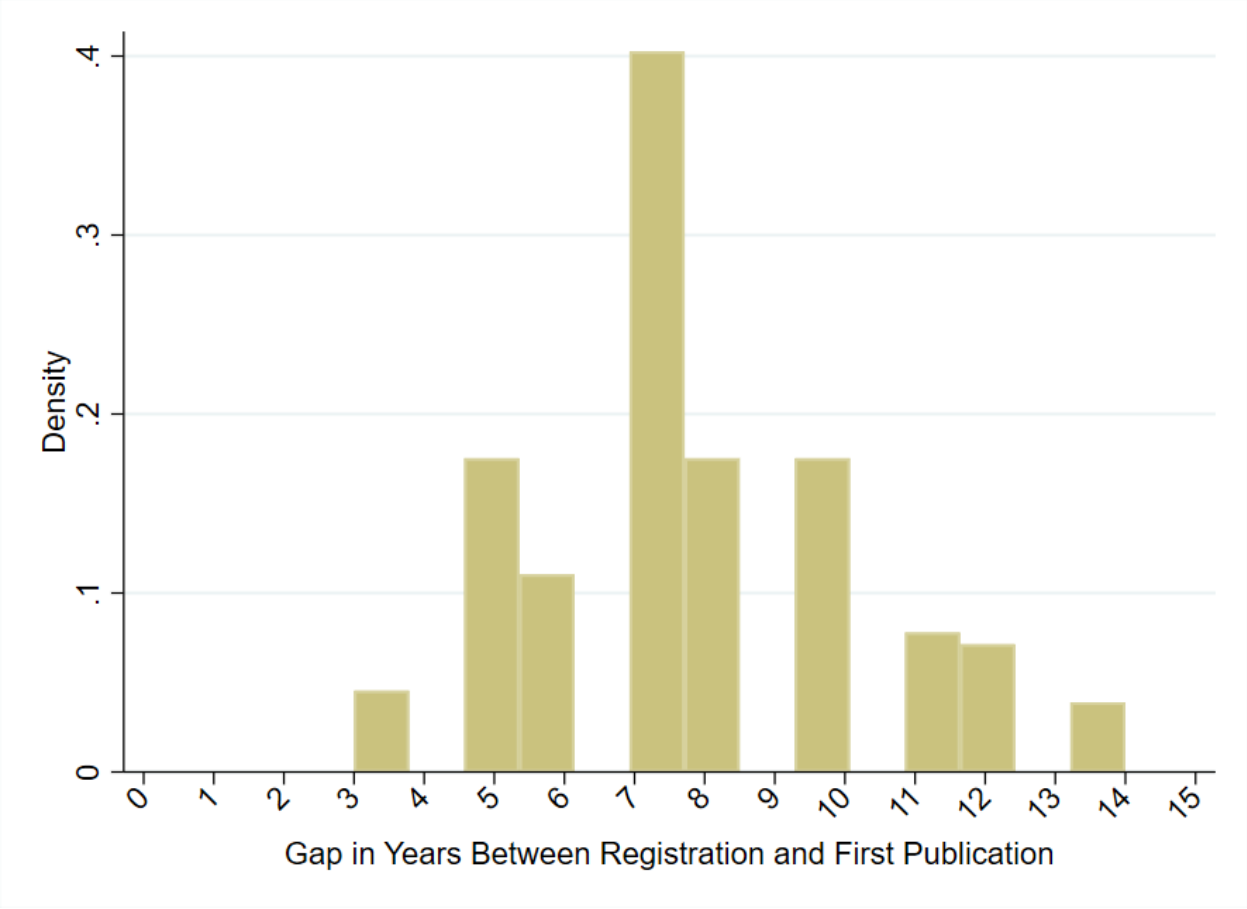
This z-score provides a frequency-adjusted measure of how strongly word w is associated with "Female"-tagged articles, allowing us to rank terms by their degree of over-representation, while down-weighting noisy differences driven by rare words. We then remove "stop-words", such as "for" or "as", which enter our lexicon but are not relevant in our context using dictionaries of stop-words published online[36]. We remove non-informative tokens but retain all medically meaningful terms, including diseases and drugs, to avoid imposing ex ante restrictions on what constitutes female-relevant research.

To separate words that are associated with health conditions and anatomy on the one hand and potential references to the gender of study participants on the other hand, we separate our initial list into terms that likely refer to the gender of participants (for example, "Female") and/or appear in group M (Named Groups, Persons [M01]) of MeSH terms and that indicate a gender, such as "Women" and "Pregnant People". This subset is categorized as words that likely refer to the gender of study participants. The remaining words in our original lexicon are then categorized as words pertaining to health conditions or anatomy.

[36] See:Sarica and Luo (2021), https://pythonspot.com/nltk-stop-words/ and https://github.com/igorbrigadir/stopwords

A.2 Figures & Tables

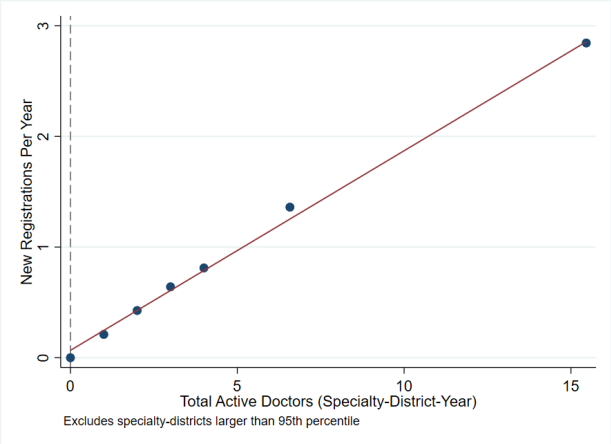
Figure A1: Gap In Years Between Doctor Registration and First Publication



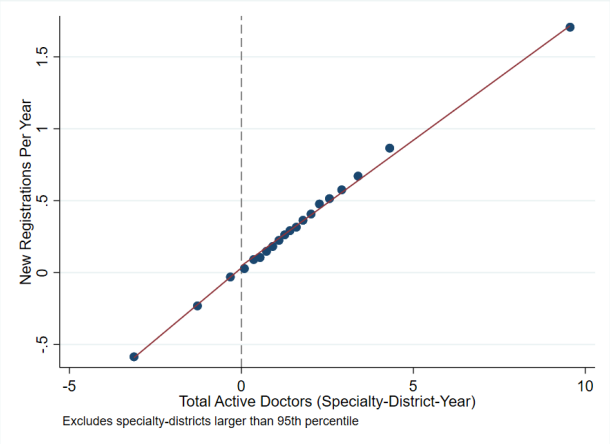
Note: The graph plots an the gap in years between a doctor's year of registration and first publication. The data are based on matches between medical doctor registration data and PubMed articles.

Figure A2: Constant Turnover Rates in Specialty-District Units

(a) Raw Turnover



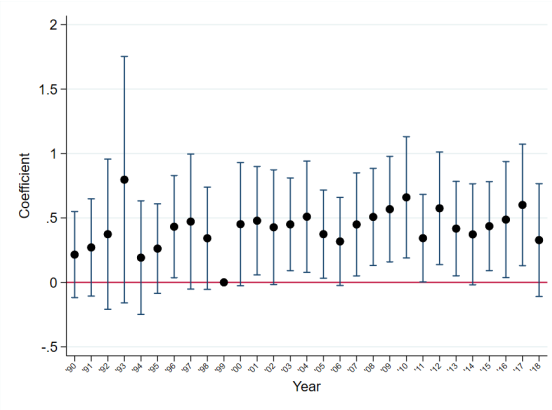
(b) With Specialty & District Fixed Effects



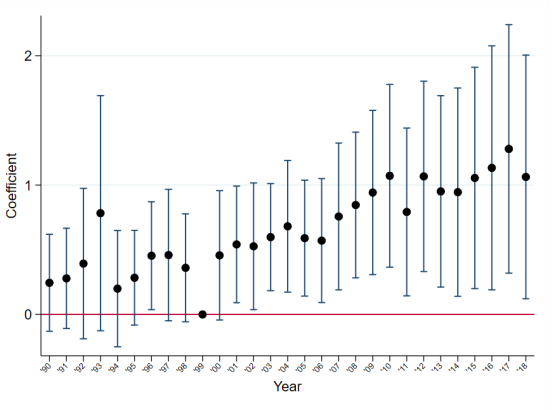
Note: The graphs plot binscatters of new medical registrations against the size of the medical specialty–district unit. Data are from medical registrations, 1952-2018.

Figure A3: Event Study - Female Tags in Articles, Single Years

(a) Reduced Form

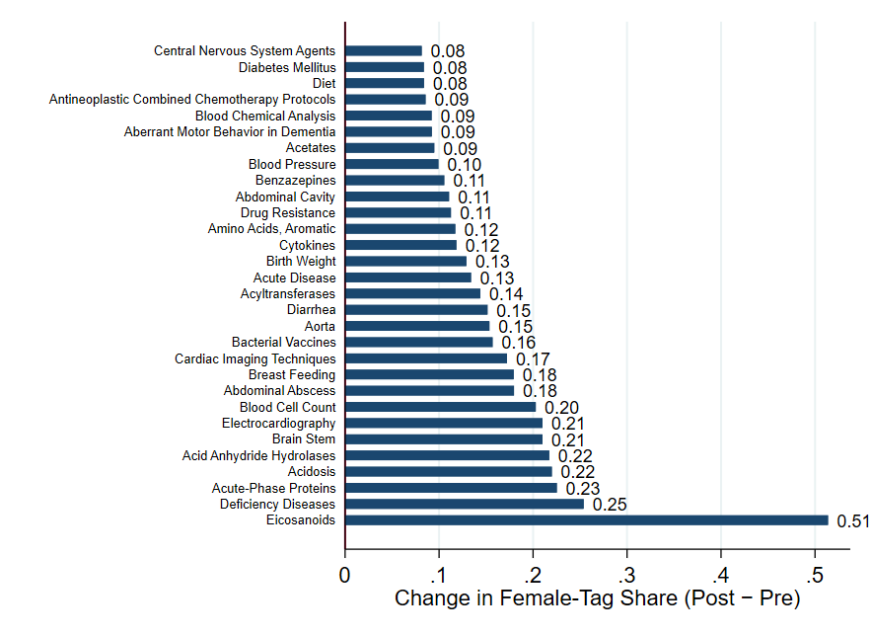


(b) With Specialty Trends



Note: The graph plots event studies for the likelihood of a "Female" tag in article based on equation 3 and PubMed data. The event study ranges from 1988 to 2018, and includes fixed effects for specialty, year and district. Standard errors are clustered by specialty and district. Round markers represent point estimates and vertical lines represent confidence intervals. The base period is 1998-1999.

Figure A4: Top Conditions with Changes in Female-Inclusiveness



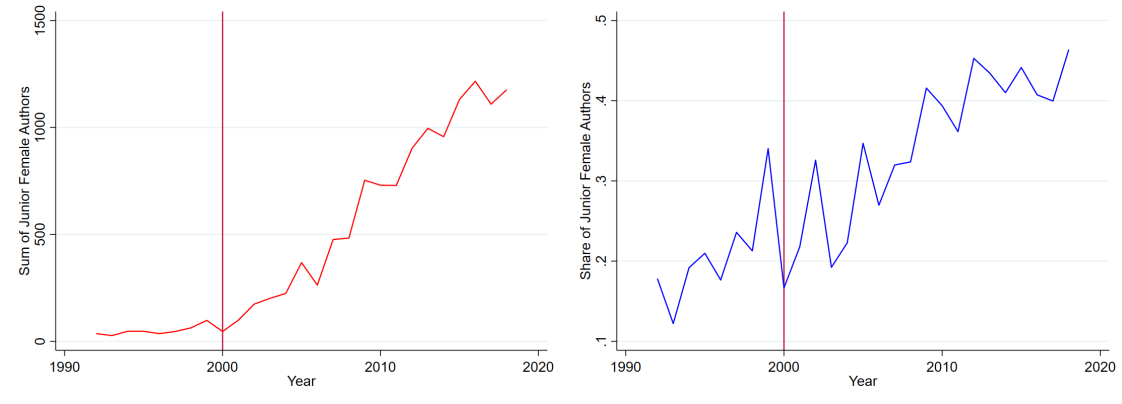
Note: The figure plots the top 30 MeSH terms with the highest change in female-inclusiveness, indicated by the presence of the "Female"-Tag. Data are based on our sample of PubMed data from 1988-2018.

Figure A5: Growth in Female Authors in PubMed Articles



Note: The graphs plot the increase in female authors across a discrete version of the treatment variable, based on the mean level of treatment. Data are from PubMed for 1992-2018.

Figure A6: Growth in Female Junior Authors in PubMed Articles



(a) Number of Female Junior Authors **(b)** Female Share of Junior Authors

Note: The graphs plot the increase in the number of female junior authors and the female share of junior authors. Data are from PubMed for 1992-2018.

Table A1: Top 10 MeSH Terms

Major MeSH Term	Frequency
Receptors	2,573
Diabetes Mellitus	1,721
Carcinoma (sub-type)	1,632
Models	1,461
Antibodies	1,449
Polymorphism	1,429
Mutation	1,403
Tuberculosis	1,339
Medical	1,282
Attitudes	1,283
Observations	185,212

Note: The table presents the 10 most frequent unique Major MeSH terms (medical subject headings) among PubMed articles with at least one Pakistani author. An article may contain multiple Major MeSH terms.

Table A2: Female Tagged Research - Difference-in-Differences

	Probability Female in			
	MeSH Term DD	MeSH Term^{+} DD	MeSH Term^{human} DD	MeSH Term$^{human,+}$ DD
	(1)	(2)	(3)	(4)
Δ Female Share $\times$ Post	0.116**	0.018	0.133**	0.028
	(0.058)	(0.045)	(0.064)	(0.043)
Observations	7,128	7,128	7,128	7,128
Mean DV	0.68	0.19	0.67	0.19

Notes: This table presents results from DD regressions of the change in female share by specialty and district on articles' female relevance. In columns 1 and 3, the outcome is an indicator for a "Female" MeSH tag. In columns 2 and 4, the outcome is an indicator for a "Female" tag and *not* a "Male" tag in the MeSH term, indicated by $^{+}$. The "human" superscript in columns 3 and 4 indicates that the MeSH term includes a "Human" tag. All regressions include specialty, district and year fixed effects. Standard errors are clustered by specialty and district. Stars indicate significant differences: *p $<$ 0.10, **p $<$ 0.05, ***p $<$ 0.01.

Table A3: Robustness - Articles

	Female Tag (1)	Female Tag^{human} (2)
A: Non-Medical Doctor Researchers		
Δ Female Share $\times$ Post	-0.294**	-0.140
	(0.142)	(0.159)
Observations	9,647	9,647
R^2	0.28	0.28
B: Placebo Treatment Year=1990		
Δ Female Share$\times$ Post	6.185	8.342
	(8.820)	(11.583)
Observations	1,190	1,190
R^2	0.58	0.56
C: Remove Small Clusters		
Δ Female Share$\times$ Post	0.492**	0.477**
	(0.216)	(0.219)
Observations	6,995	6,995
R^2	0.68	0.67
D: Trends in District and Specialty		
Δ Female Share$\times$ Post	0.702*	0.760*
	(0.426)	(0.451)
Observations	7,128	7,128
R^2	0.68	0.67

Notes: This table presents robustness checks for our main article level results. Standard errors are clustered by specialty and district. Stars indicate significant differences: *p $<$ 0.10, **p $<$ 0.05, ***p $<$ 0.01.

Table A4: Female Tagged Research - Extensive and Intensive Margins

	Probability Female in			
	MeSH Term	MeSH Term^h	MeSH Term	MeSH Term^h
	DDIV		Reduced Form	
	(1)	(2)	(3)	(4)
Δ Female Share $\times$ Post	0.557*	0.596*		
	(0.320)	(0.334)		
FemaleShare $> 0\times$ Post	0.114	0.111	0.111	0.105
	(0.101)	(0.106)	(0.072)	(0.074)
FemaleShare $\times$ Post			0.113**	0.115**
			(0.052)	(0.053)
Observations	7,206	7,206	7,196	7,196
Sanderson-Windmeijer F-stat	9.30	9.30		
Kleibergen-Paap F-stat	2.17	2.17		
A-R p-value	0.08	0.09		
Mean DV	0.68	0.67	0.68	0.67

Notes: This table presents results from DDIV regressions of the change in female share by specialty and district on articles' female relevance. We use 1-(FemaleShare)$_{sd}$ in reduced form regressions instead of FemaleShare$_{sd}$ to keep the direction of the results consistent between reduced form and 2SLS results and event studies. In columns 1 and 3, the outcome is an indicator for a "Female" MeSH tag. In columns 2 and 4, the outcome is an indicator for a "Female" tag and *not* a "Male" tag in the MeSH term, indicated by $^+$. The "h" superscript in columns 3 and 4 indicates that the MeSH term includes a "Human" tag. All regressions include specialty, district and year fixed effects. Standard errors are clustered by specialty and district. Stars indicate significant differences: *p $<$ 0.10, **p $<$ 0.05, ***p $<$ 0.01.

Table A5: Incumbent and Post-Reform Female Author Collaborations

	"Female" Tag					
	With Junior Coauthor			No Junior Coauthor		
	Female Author	Male Author	Triple Difference	Female Author	Male Author	Triple Difference
	DDIV	DDIV	DDIV	DDIV	DDIV	DDIV
	(1)	(2)	(3)	(4)	(5)	(6)
Δ Female Share × Post	0.549	-9.885	-3.847	0.367	0.868	0.705
	(0.372)	(58.287)	(16.677)	(0.296)	(8.350)	(5.201)
Δ Female Share × Post × Female			4.898			-0.669
			(16.658)			(5.386)
Observations	1,146	2,673	3,819	1,146	2,673	3,819
Sanderson-Windmeijer F-stat	19.13	0.03	0.19	19.13	0.03	0.19
Kleibergen-Paap F-stat	9.46	0.01	0.01	9.46	0.01	0.01
SY 10% maximal IV size	7.03	7.03	7.03		7.03	7.03
A-R p-value	0.13	0.31	0.09	0.33	0.54	0.38
Mean DV	0.15	0.13	0.14		0.07	0.48
Author FE	✓	✓	✓	✓	✓	✓

Notes: This table presents results from DDIV regressions of the probability that an article's MeSH term includes a "Female" tag and was coauthored with at least 1 junior author. In columns 1 and 4, the sample is restricted to articles with a matched incumbent female authors, and in columns 2 and 5, to matched incumbent male authors. to articles by pre-reform female authors, and in column 3, articles by pre-reform male authors. Columns 3 and 6 represent IV triple differences models. SY maximal 10% IV size refers to the Stock-Yogo maximal IV size for weak instruments (Stock and Yogo, 2002). All regressions include author, year, specialty and district fixed effects. Standard errors are clustered by specialty and district. Stars indicate significant differences: *p $<$ 0.10, **p $<$ 0.05, ***p $<$ 0.01.