NBER WORKING PAPER SERIES

GENDER REVEALS IN THE LABOR MARKET:
EVIDENCE ON GENDER SIGNALING AND STATISTICAL DISCRIMINATION IN AN ONLINE HEALTH CARE MARKET

Haoran He
David Neumark
Qian Weng

Working Paper 32929
http://www.nber.org/papers/w32929

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2024

We are grateful for helpful comments from Lorenzo Lagos, Matthew Pecenco, Yuan Tian, Yi Li, Lulu An, and conference and seminar participants at Brown University, East China Normal University, Peking University, Shandong University, Zhongnan University of Economics and Law, Zhejiang Gongshang University, the 2024 China BEEF International Workshop on Behavioral and Experimental Economics, and the 1st New Economy and New Employment Forum (2024). We thank Lulu An, Shuai Chen, Peiying Li, Jingtai Liu, Wanxin Liu, Yuhong Luo, Di Ma, Yunwei Wang, Ziyu Zhang, and especially Jingwen Xia for excellent research assistance. All remaining errors are those of the authors. Financial support from the Ministry of Education of China (Project No. 21YJA790059), the National Natural Science Foundation of China (Project Nos. 72273148, 72473010, and 72131003) and the National Program for Special Support of Top-notch Young Professionals is gratefully acknowledged. All authors made equal contributions. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Gender Reveals in the Labor Market: Evidence on Gender Signaling and Statistical Discrimination in an Online Health Care Market
Haoran He, David Neumark, and Qian Weng
NBER Working Paper No. 32929
September 2024
JEL No. I11, J16, J40, J70

ABSTRACT

A recent approach to testing for customer statistical discrimination involves studying price gaps between sellers from different gender, race, or ethnic groups and how they evolve as buyers obtain more information about seller quality. We consider a similar setting, testing for statistical discrimination against female doctors in an online health care market. But we show that this kind of analysis does not provide evidence on statistical discrimination in this setting because doctors have a choice about how strongly to signal gender. We develop a new approach to identifying statistical discrimination using doctors' choices about signaling their gender. We find evidence of statistical discrimination against female doctors in male-dominated fields, and against male doctors in female-dominated fields. In particular, female doctors mask gender more strongly initially in male-dominated fields, and male doctors do the same in female-dominated fields. But in both female- and male-dominated fields the gender gap in signaling decreases with number of customer reviews of doctors. More generally, our evidence indicates how, in some markets, sellers may be able to reduce statistical discrimination by masking their group membership.

Haoran He
Beijing Normal University
No 19, Xinjiekouwai St.
Haidian District
100875 Beiing
China
haoran.he@bnu.edu.cn

Qian Weng
Renmin University of China
No. 59 Zhongguancun Street
Haidian District
Beijing 100872
China
qian.weng@ruc.edu.cn

David Neumark
Department of Economics
University of California, Irvine
3151 Social Science Plaza
Irvine, CA 92697
and NBER
dneumark@uci.edu

1. Introduction

A recent approach to testing for customer statistical discrimination involves studying price gaps between sellers from different groups and how they evolve as buyers obtain more information about seller quality. A prominent example based on actual transaction prices is the study of statistical discrimination against racial and ethnic minority hosts on Airbnb (Laouénan and Rathelot, 2022). In a seminal earlier paper, Altonji and Pierret (2001) study how race differences in wages change with a worker's labor market experience. And more recently, Bohren et al. (2019) study gender differences in how posts (of math questions and answers) are evaluated, and how the gender differences change with the history of positive evaluations. Naturally, these kinds of analyses require that buyers (and researchers) can identify the groups to which sellers belong. However, there may be settings in which sellers are aware of potential statistical discrimination and have the ability to obscure their group membership to reduce its negative impact. For example, in the Airbnb study, hosts listing properties have a choice about whether to post a picture, and how strongly the picture signals race. In such contexts, tests of statistical discrimination based on sellers' *self-signaled* group membership may yield biased estimates and invalid tests.

In this paper, we test for gender-related customer statistical discrimination against doctors, based on the evolution of price differences as more information about doctor productivity/quality becomes available to customers. In the setting we study, sellers (doctors) can choose how strongly to signal gender, and this choice can vary over time. We first establish the biases this choice about signaling can introduce in testing for statistical discrimination. We then show how we can instead exploit choices about the signaling of gender to test for statistical discrimination, from evidence on initial differences in gender signaling and how that signaling changes over time with increasing reliability of the signal of doctor quality.

The features of the online market we study – importantly, including the ability to mask signals about gender or other group membership – are common to a wide range of online markets such as Airbnb, DoorDash, and GrubHub. In these markets, customers directly interact with and hire service providers, service providers have some choices about how to signal their characteristics, and service quality is subject to customer reviews that provide quality signals, with the information presumably becoming more reliable as reviews accumulate.[1] Thus, our findings can

[1] For example, in the Airbnb setting hosts typically include name and a photo, and hosts can choose gender- or race-neutral names, and ambiguous photos (or choose not to include one). After their stay, guests are allowed to leave a quantitative rating and a qualitative assessment regarding both the property and the host. Similar registration and customer review systems exist on DoorDash and GrubHub. Task Rabbit is like Airbnb in that service providers typically post a name and photo, and customers review them.

shed light on other gig economy jobs and our approach can be applied to investigating discrimination in these markets. Moreover, the same kind of statistical discrimination and choices about revealing group membership can also arise in traditional labor markets.[2] Some choice about signaling gender (or other group membership) is in fact a natural feature of labor markets.

We collected data for all doctors who work in highly-ranked public hospitals across China and who also provide consultation services on a nationwide online health care service platform. On this platform doctors can choose from a set of services to provide and how much to charge for each service. There is basic information that is compulsory for them to provide, such as name, hospital(s) and department(s) where they work, specialty area, professional title, and identification information for authentication. But doctors can choose whether to signal gender by reporting gender in a freestyle self-written short biography, posting a photograph, or providing neither type of information. Like other online platforms, the platform posts statistics for doctors measuring their quality, such as a comprehensive recommendation popularity measure, scores (e.g., treatment effect satisfaction, bedside manner satisfaction), and free-text reviews from patients based on experiences with the doctors. These features of the platform make it a natural setting to implement a test for statistical discrimination – in this case based on gender – along the same lines as Laouénan and Rathelot's Airbnb study. However, as noted above there are two key differences. First, we explicitly consider the likelihood that price differences may reflect gender signaling differences, making the usual test based on price differences biased. Second, we capture information about how reliably gender is signaled, which we use to construct a new approach to testing for statistical discrimination.

Our analysis of discrimination in the market for doctors is significant given pervasive evidence of gender gaps in earnings in the health sector (e.g., Gravelle et al., 2011; Jagsi et al., 2012; Jena et al., 2016). Moreover, there is suggestive evidence of discrimination that emanates from customers.[3]

---

[2] In traditional job markets one can potentially mask some information early in the process, by omitting information from resumes or applications (such as involvement in organizations that signal membership) – in a sense the opposite of what is done in correspondence studies where researchers use this kind of information to signal group membership (see, e.g., Neumark, 2018). Another example would be Asians in western labor markets using Anglicized first names, in part to reduce discrimination (e.g., Ge and Wu, 2023; Zao and Biernat, 2017).

[3] The latest Medscape report on medical specialists' earnings in the United States for 2024 shows that female specialists earn around 31 percent less than male specialists (McKenna, 2024). In a national survey of general surgery residents, 65.1% of women reported gender discrimination, and among women reporting gender discrimination 49.2% identified the source of discrimination as patients or patients' families (Hu et al., 2019).

We first lay out the conventional statistical discrimination framework (Phelps, 1972; Aigner and Cain, 1977), and show how this framework leads to a test of statistical discrimination based on price differences and how they change as quality information (reviews) accumulates (as in, e.g., Laouénan and Rathelot, 2022). However, in our context sellers (doctors) have the ability to choose how strongly to signal their group membership (gender). We then extend the conventional statistical discrimination framework to incorporate gender signaling. We show that when there is a choice about gender signaling – including doctors varying the signal as other information about them changes – the test for statistical discrimination based on price differences is no longer valid, in the sense that statistical discrimination no longer predicts the effects on which this test relies. We then show how, in this setting with gender signaling, we can test for discrimination by studying the dynamics of gender signaling.

We then implement this test using our data from the online health care market. We find evidence of statistical discrimination against female doctors in male-dominated fields (with male doctors viewed as higher quality absent other information) and statistical discrimination against male doctors in female-dominated fields (with female doctors viewed as higher quality absent other information). The evidence of this – based on our test – is that female doctors mask gender more strongly initially in male-dominated fields, and male doctors do the same in female-dominated fields. But in both female- and male-dominated fields these gender differences in signaling of gender decrease with number of customer reviews of doctors. These results are consistent with doctors who would experience statistical discrimination based on their gender choosing less informative gender signals initially, when the gender signal is most of the information customers have available. But as reviews accumulate and customers rely more on doctor-specific information, there is less incentive to mask gender.

2. Relation to existing research

Our study contributes to research that tries to test for statistical discrimination by examining whether gaps in the outcomes of interest are responsive to a quality signal and how reliable the signal is. The seminal paper by Altonji and Pierret (2001) tests for statistical discrimination on easily observable characteristics like race by asking whether, as employers learn more about productivity, race differences in outcomes decline. The paper highlights a key problem in using standard observational survey data – that researchers have to make assumptions about what employers know about workers and when they know it.

This problem has been addressed in a creative manner in research using online markets. Most notably, in our view, Laouénan and Rathelot (2022) study Airbnb postings, where it is clear what customers know initially (a photo and name of the host, from which race and ethnicity can be inferred, and property characteristics), and what they learn over time – ratings. Their analysis indicates that statistical discrimination can account fully for minority price gaps, as they are present when no reviews are posted, but disappear with large numbers of reviews. A number of other studies present similar evidence that quality signals reduce group differences in outcomes. Cui et al. (2020) show that customer reviews reduce hosts' racial discrimination against African-American guests in Airbnb. Pallais (2014) shows that employer ratings of inexperienced workers' job performance improve their future employment outcomes. Feld et al. (2022) show that providing professionals' aptitude and personality assessments reduces employers' beliefs about lower coding skills among female programmers. However, it may be more often the norm than the exception that sellers have some control over the revelation of information about their group membership. As we show, this can both complicate these tests of statistical discrimination, but also lead to a new test.

In relation to past research on gender discrimination generally, our findings are consistent with a good deal of accumulating evidence that gender discrimination is more complicated than simply discrimination against women. Rather, our evidence of discrimination against women in male-dominated fields, but discrimination against men in female-dominated fields, echoes findings from other types of studies that distinguish between evidence for more traditionally male- versus female-dominated fields (see Neumark, 2018, Riach and Rich, 2002, and Rich, 2014).

Our study also adds to a small but burgeoning literature that investigates how groups that experience discrimination potentially reduce this discrimination by altering signals of group membership. Most directly related to our work, Kang et al. (2016) conducted a lab experiment to show that racial minority workers submit more racially transparent resumes when targeting employers who state they value diversity, providing evidence that workers reveal (disadvantageous) race signals more when perceiving less discrimination.[4] Our study differs from Kang et al. in a number of ways. First, we focus on gender rather than race. Second, we obtain data from a real labor market, eliminating experimenter demand effects that can arise in laboratory experiments and increasing external validity. Third, in addition to showing how choices about gender signaling respond to anticipated discrimination, we show how these choices evolve with the reliability of the quality signal. And fourth, we tie our evidence to a direct test for statistical discrimination based on decisions about signaling group membership.

Finally, we add to existing research on customer discrimination against doctors in the

[4] Relatedly, other work has documented how members of groups anticipating discrimination choose to emphasize quality signals (e.g., Exley et al., 2024).

health care sector, which is often posited as a source of under-representation of minority doctors (Bergen, 2000; Brown et al., 2009; Lett et al., 2019) and of lower pay for minority doctors (e.g., Grisham, 2017; Ly et al., 2016). Moreover, some work in this area points to statistical discrimination. Chan (2023) finds customer racial discrimination that reduces patient willingness to pay for black and Asian doctors compared to whites, but also finds that providing signals of doctor quality reduces this discrimination by about 90%. Sarsons (2017) finds that surgeon gender influences the way signals about surgeon quality are interpreted. Based on data collected from a Chinese online health care platform, Chen (2024) finds that female doctors charge lower prices and provide fewer consultations than males, and that the penalty for females is larger for doctors who provide fewer signals about quality. To the best of our knowledge, however, no study has considered how doctors react to discrimination through signaling of group membership, nor considered the implications of choices about signaling for tests of statistical discrimination.

3. The Good Doctor Online Platform

We study data from the Good Doctor Online (hereafter GDO) platform (haodf.com). There are about 250,000 doctors who have been authenticated by the platform and provide patients with online health care services including individual online consultation;[5] 73% of these doctors work in Tier 3, Grade A hospitals, which are the hospitals with the highest level among the nine levels in the Chinese hospital classification.[6] The platform was launched in 2006 and has served over 84 million patients as of July 2023 (GDO, 2024).

The GDO platform provides detailed information on the doctors and on the hospitals where they work. Each doctor has a personal webpage; Figure 1 provides an example. The webpage posts basic information, including name, professional title,[7] academic title (optional), names of hospital(s)

[5] There are over 600,000 additional doctors on the platform who do not provide online health care services and hence are not part of our analysis, although they also have personal webpages and the kinds of reviews described below.

[6] According to "The Measures for the Administration of the Hospital Grade" (in Chinese) enacted by the Ministry of Health (now National Health Commission) of the People's Republic of China in 1989, Chinese public hospitals are classified into three tiers, 1 to 3, with 3 being the best; within each tier, there are three grades, A to C, with A being the best. The classification standards are based on hospital functions, tasks, facility conditions, technical construction, quality of care and scientific management, etc. (Yan, 1990).

[7] Professional titles are uniform nationally, from junior to senior at four levels, i.e., resident, attending, associate chief, and chief, as are the titles for physicians, rehabilitation engineers, technicians, nurses, examiners, and pharmacists. The promotion up one level starting from attending physicians is at least five years after working at the current level, and the years of working needed to be promoted to attending physicians depend on the level of educational degree attained. (See https://www.gov.cn/zhengce/zhengceku/2021-08/05/content_5629566.htm, in Chinese.)

and department(s) where they work, specialty areas, and outpatient schedule (optional). Critical for our research, the webpage can include a picture of the doctor, and a short biography. The doctor can decide which of these to provide, including providing neither or both.

Also, critical for our research, the platform provides an online doctor review system for all listed doctors. A patient first rates how satisfied they were with a doctor's treatment outcomes and bedside manners on a 5-level scale, from "very unsatisfied" to "very satisfied." Then the patient fills in the following compulsory items to complete a review: patient's name, disease, purpose of the consultation, what treatments were given, expenses, current disease situation, reason(s) this doctor was selected, and other open-ended comments. The number of patient reviews reflects the number of patients who have filled in every item and submitted the review. Both online and offline patients can write reviews, and reviews are anonymous. It is this count of reviews that we use to capture the reliability of the signal about doctors' service quality (just like Laouénan and Rathelot (2022) using the number of Airbnb reviews).[8] This count appears quite prominently on the webpage for each doctor (see, as an example, the "770" near the bottom of Figure 1, and the large text label to the left of it which says "[doctor name]'s patient reviews").

The website posts a few types of information about the doctors' quality of service. The most prominent indicator is labeled as "comprehensive recommendation popularity (or "popularity" for short)".[9] It ranges between 1 and 5 (low to high) and appears prominently on the webpage (see the "4.9" in the top-right corner of Figure 1). It seems most natural that patients would interpret this a summary measure of patient reviews, although information gleaned from communication with customer service (although not indicated on the platform) suggests it also assigns some weight to doctors' qualifications and other factors.[10] We treat this index as a rating of doctors, similar to the rating of properties on Airbnb, for example, and refer to this as "rating" throughout.[11]

[8] Other potential measures of reliability of doctors' quality signal include, for example, the total number of webpage visits, the number of individual online consultations, etc. However, the number of reviews is the most natural measure for our setting as the reliability of the reviews will be higher with more of them.

[9] On the platform, this changed to "recommendation popularity by patients" in 2022.

[10] The doctors' qualifications and characteristics that they provide to haodf.com include hospital level, professional title, highest educational degree, university graduated from, etc. Other factors include, for example, the number of patients served online and offline.

[11] Our 1-5 popularity measure is similar to rating measures on many online platforms. For example, a "comprehensive quality score" is used by a similar health care platform studied by Chan (2023). In addition, Chan elicited patients' perception of the quality score on this platform by asking them "[t]o what extent do you agree or disagree with the following statement: A provider with a higher Comprehensive Quality Score is a better provider than a provider with a lower Comprehensive

A doctor's personal webpage also lists other statistics about the doctor, which patients could also potentially use to evaluate doctors. These include: the treatment satisfaction rate, the bedside manner satisfaction rate, the number of thank you letters, the number of gifts, and the online service satisfaction rate.[12] The first four can be given by both online and offline patients, whereas the last can be given only by online patients. There are limitations of these other measures. For instance, the treatment effect satisfaction rate and bedside manner satisfaction rate are computed only for past two years, and are only available for 32% of the analysis sample. Similarly, the online service satisfaction rate is computed only for past 90 days and is only provided when it is higher than 60%, and thus both has limited variation and is available for only 18% of the analysis sample. In addition, at least some of these other measures appear less salient on the platform than the "popularity" rating we use.[13]

GDO provides comprehensive online health care services that are easy to access. The services include online chat consultation, phone call consultation, remote video consultation, online appointment scheduling for outpatient care, disease management after diagnosis and treatment, electronic prescriptions, popular science knowledge dissemination, and family general practice (GDO, 2024). The online consultation procedure is common to multiple channels (as is the information on doctors discussed above), such as the PC version website, the GDO app, the cellphone version website, the WeChat official account, and the WeChat applet.

To consult on the GDO platform, patients search for doctors by hospital, by disease, or by department. Having selected a doctor, patients select the consultation service they want: individual online consultation, team online consultation, appointment scheduling for outpatient, or private care service. Since we are interested in the relationship between doctors' gender, ratings

Quality Score." The response ranged 1 to 5 (low to high) and the average choice was 4, providing supporting evidence that such a measure can be used as quality signal, and that providers with higher scores are perceived as better providers than those with lower scores.

[12] The treatment satisfaction rate and bedside manner satisfaction rate are computed from the 5-level scale rating patients give doctors in the review system. Thank you letters are notes from patients to doctors expressing gratitude, while gifts can be offered to doctors to express gratitude and each gift can be up to 200 CNY. The online service satisfaction rate is computed from a binary choice rating "satisfied" or "unsatisfied" that patients give doctors after completing an online consultation (only online patients can rate).

[13] In particular, the treatment satisfaction rate and the bedside manner satisfaction rate appear near the bottom of Figure 1 below the count of patient reviews, in contrast to the popularity rating at the top of the figure. The number of thank you letters and the number of gifts appear in the statistics section at the right edge of Figure 1, and are not highlighted. The one exception is that the online service satisfaction rate appears in the top-right corner of the personal webpage, like the popularity rating. As discussed later, we show that our results are robust to incorporating information on these other potential quality signals.

(and other information about doctors – importantly, including their number of reviews), and their individual online consultation prices, we focus on individual online consultations. Once patients click online consultation on doctors' personal webpages (only clickable for doctors who provide individual online consultation), they are directed to the consultation webpage, which displays the individual consultation services available: a chat consultation (unlimited number of text questions and answers within 48 hours ("48-hour chat"), a chat for asking and answering only one question ("one-question chat"), or phone call consultation (see Figure 2). The price of each service, which is set by the doctor, is listed next to each type of consultation service, along with doctor response speed.[14] Patients then select the service.[15]

4. Data

We focus on doctors in 139 top hospitals in China listed in the 2018 Hospital Comprehensive Strength Ranking published by Fudan University's hospital ranking system, which is the most well-recognized and respected ranking for Chinese hospitals and only lists the top-20 hospitals for each of the seven regions across China.[16] These hospitals cover a wide range of locations, departments, and specialties, and usually represent the largest hospitals in China.

We collected data from publicly available webpages of the GDO website between September 15 and September 27, 2020.[17] Based on the list of doctors presented on the hospital outpatient schedule webpage in each of the selected hospitals on the GDO website, we collected all information

[14] A doctor can set multiple phone call consultation prices, such as for new and returning patients, and for different consultation period lengths, etc.

[15] Patients then provide personal information and information on their medical condition or records (which can be uploaded) and pay for the service, and after approval by GDO, wait for the doctors to agree to take the consultation. If the doctors agree, the patients and doctors can start the consultation and communication; if not, the doctors explain why and the platform wires the fee back to the patients.

[16] Fudan University's hospital ranking system is similar to the U.S. News Best Hospitals ranking, and is a public philanthropic project carried out by the Fudan University Hospital Management Research Institute. The ranking is based on ratings from thousands of doctors in the Chinese Medical Association and Physicians Association on hospitals' medical practices, quality of care, and research achievements. The seven administrative regions are Northeast China, North China, East China, South China, Central China, Northwest China, and Southwest China. The 2018 Hospital Comprehensive Strength Ranking is available at http://www.fudanmed.com/institute/news2018-2-11.aspx. Our data covers 139 hospitals because one of the 140 hospitals in this ranking system is not included on the GDO platform.

[17] One could imagine trying to crawl data in several waves to create a doctor-level panel. But the website frequently updates its anti-crawling methods, making data scraping difficult and incomplete in practice (see below for details). At least in the price regressions shown below, we are not concerned about unobserved doctor heterogeneity, because we interpret the data in terms of customers' expectations of doctor quality conditional on the information available to them, which is the same as the information available to us.

for each doctor from their personal webpages (Figure 1) and their individual online consultation pages (Figure 2, for those who provide individual online consultations). In total, 105,436 doctors were collected,[18] among whom 105,250 doctors[19] (99.82% of the original sample) have full information, and among whom 35,907 doctors provide individual online consultation.

The prices of individual online consultation services are one of our outcomes of interest, although as explained in the introduction we ultimately use gender signaling to test for statistical discrimination. The doctors on GDO who we study provide one or more of three types of online consultation services: chat consultation (48-hour chat or one-question chat), or phone call consultation. For the 48-hour chat and one-question consultations, only one price is set. Phone call consultations have up to four different prices depending on the designated length of call and also varying for new vs. returning patients. In the analysis of prices, we create observations for each pair of doctor and service types (along all of these dimensions).

Our focus is on the gender of doctors, which is not required to be revealed explicitly on GDO. Doctors' names are listed, from which customers can infer gender, but only imperfectly. Doctors need to provide their doctor licenses to the platform to register and get authenticated, so they can only use their true names rather than perhaps a more gender-neutral first name or nickname, and changing names after registration is not allowed. In addition, however, there are two ways a doctor can choose to reveal their gender more explicitly. A doctor can explicitly report gender, typically at the beginning of their short biography. A doctor can also post a picture (as in the example in Figure

[18] We crawled personal webpages and individual online consultation webpages on doctors from top to bottom of the listing on the hospital outpatient schedule webpage in each hospital. Since the website constantly rotated the order of doctors' listings in each hospital, it was possible that some doctors yet to be crawled were moved up in the listing into the set of doctors whose webpages had been crawled. In this case, these doctors would not be crawled. Conversely, downward movement of doctors would result in repeated crawling. To deal with this issue, under our resource constraint, we repeated our crawling procedure three times with some time interval in between. For those crawled multiple times, we kept the first observation. For those always omitted, we identified them based on doctor listings on the hospital outpatient schedule webpage and crawled their information. The crawling rate for each hospital was computed as the number of doctors whose information is crawled divided by the number of doctors listed on the hospital outpatient schedule webpage, averaged across all sample hospitals. The average crawling rate is 99.93%. It is less than 100% because the hospital webpage only listed doctors who provided online services on the days of crawling, and a tiny share of doctors on leave on those days were still missing.

[19] We cleaned the original sample according to the following procedure. First, for doctors who worked in multiple hospitals in the ranking list, their information was collected multiple times. In these cases, we only kept the information crawled for the first hospital in the ranking list doctors reported on personal webpages. (This dropped 117 doctors with 119 observations – 115 doctors appearing in 2 hospitals, and 2 doctors appearing in 3 hospitals.) Second, we dropped doctors whose names or hospitals worked in are missing (53 doctors). Third, we dropped doctors whose names reveal that they are teams instead of individuals (14 doctors). This left 105,250 doctors.

1). They can choose to do either or both. Doctors can freely remove or revise their short biography and pictures by logging into their accounts (at any time and with zero cost). Thus, doctors can change their gender signaling over time. Their ability to do this is central to our test of statistical discrimination based on gender signaling. We cannot observe changes directly since we scraped data at one point in time. But our data indicate how gender signaling varies with the number of reviews a doctor has at the time of scraping.

Thus, we have three means to classify doctors' gender, and try to measure what we think customers will infer based on the information available to them. We regard as most reliable the gender reported in the short biography (15,470 doctors). We also use pictures to classify gender using a facial recognition algorithm based on Baidu application programming interface (API) in Python (12,506 additional doctors who did not report gender in their biography).[20] Of the 12,506 doctors whose gender we tried to classify by pictures, the computer program could identify 11,724 doctors with full accuracy (i.e., the program was 100% certain of the gender).[21] For the remainder (782 doctors), whose faces could not be recognized by the computer program or the recognition accuracy was less than 100%, gender was then classified manually by research assistants (RAs).[22] We validate the picture classification procedure by computing the accuracy of gender identified by picture (the combined use of the computer program and manual classification) for the 10,170 doctors who also reported gender in their biography. The picture classification agreed in 99.77% of cases.

For doctors who did not signal gender via their biography or a picture, we predict gender based on names using two common methods. First, we use the gradient boosting decision tree

[20] Baidu's API for face recognition is a cloud-based technology that can identify human faces in digital images and videos, and return face frame locations and output 150 key point coordinates of faces. Based on these coordinates, face features are extracted and feature comparison is conducted using a feature model trained by a deep learning algorithm on the basis of a stored face database, which accurately identifies a variety of attributes such as gender, age, facial expression, face shape, etc. See https://intl.cloud.baidu.com/product/face.html, accessed 2023-02-06.

[21] Some doctors uploaded non-portrait pictures, such as landscape pictures, cartoon pictures, or pictures including multiple people, etc., so that in a small number of cases (20 doctors in addition to the 12,506 doctors) gender could not be classified by pictures and hence these observations were dropped from the analysis. We could have included these doctors and used the final means to predict their gender by name (detailed below), but we choose not to do so because these doctors may have provided non-portrait pictures for a reason and are different from those who do not provide pictures at all.

[22] Among those for which the algorithm was less than 100% certain, six RAs independently classified the gender of these doctors. If there was disagreement among RAs, the gender was determined based on the majority. If there was a tie, another five RAs independently classified gender, and gender was determined based on the majority of the five RAs.

(GBDT) algorithm in Python (Friedman, 2001; Brownlee, 2016).[23] This method provides only predictions of gender without a probability of accuracy. Second, we use the "ngender" package in Python to identify gender, which returns the probability indicating how accurate the classification is for each individual.[24] The predicted probability is always above 50% for the gender identified (for those below 50%, they are identified as opposite gender with probability above 50%). There are 2 names that are identified as exactly 50% male and 50% female by the algorithm. We drop these two doctors.

We compute the accuracy rate of these two methods by comparing to the gender reported in the short biography based on the sample of doctors who reported gender in their biography (15,470 doctors). Accuracy rates are not as high as for pictures – 83.99% and 83.87% for the two methods, respectively. When the two gender predictions agreed, the classification of gender based on names is unambiguous (7,201 doctors). The predictions are more likely to disagree when the predicted probability from "ngender" is lower, not surprisingly.[25] For doctors whose gender predictions from the two name-based algorithms disagree (728 doctors), we had RAs independently manually check gender of these doctors on the internet and identify gender (633 doctors).[26] We drop doctors whose gender could not be identified by both RAs, or where two RAs identified different gender and were both sure or both unsure (95 doctors). This leaves us with a final sample of 35,810 doctors with gender classified.

Our test of statistical discrimination based on gender signaling requires a classification of the accuracy of different gender signals. We base this on our findings from the classification work described above. The lower accuracy rate of identifying gender by name is central to our analysis, as it implies that using name only provides a less informative signal of gender. In contrast, because

[23] GBDT is a machine learning algorithm that combines multiple decision trees to make predictions and has high accuracy.

[24] "ngender" is a user-written package in Python to identify gender by Chinese names. See https://github.com/observerss/ngender for details.

[25] The average predicted probability is 86.4% for names for which the gender predictions agree in the two name-based algorithms (for doctors who do not signal gender by report or picture), whereas the average probability is only 60.7% for names for which the gender predictions disagree.

[26] Most doctors' gender is reported at the doctor registration website of the Ministry of Health, probably except for older doctors and doctors who work in military hospitals. For the latter cases, gender was identified from other sources such as websites of hospitals where they work. We also asked RAs to report their subjective confidence level based on the information source for gender ("sure" or "unsure"). If two RAs identified the same gender for a doctor, no matter whether they were sure or not about the source of information, we use that gender to classify doctors' gender (538 doctors); if two RAs identified different gender, with one sure and the other not sure, we use the gender from the RA who was sure (95 doctors).

the accuracy rate based on pictures is so high, we combine gender signaling by report or picture into a single category of a more (and very highly) accurate signal of gender. As a result, we classify doctors as strongly or weakly signaling gender based on whether or not they signal by report or picture, versus name only.

Finally, we construct two subsamples for female-dominated fields and male-dominated fields, defined by the shares of female doctors exceeding 70% or falling below 30%, respectively. (Note that the two subsamples exclude fields with intermediate shares of females – 56.7% of doctors. See Appendix Table A1 for details.) This classification results in 70% of female doctors working in female-dominated fields, and 92% male doctors working in male-dominated fields. The female-dominated fields are: Reproduction, Obstetrics and Gynecology, and Nutrition. The male-dominated fields are: Tuberculosis Medicine, Urology, Orthopedics, Interventional Medicine, Surgery, Occupational Medicine, Sports Medicine, Burns, and Plastic Surgery.

Tables 1a and 1b report summary statistics for the main variables used in our regression analysis, for the full sample and by gender (1a) and for each gender in the male- and female-dominated fields (1b). (See Appendix Table A2 for variable definitions.) In each of the three samples, the prices of different service types vary a good deal, with one-question chat prices lowest and phone call prices highest, as we would expect. Across the full sample and the two subsamples, 35-49% of doctors signal gender by report, and 59-71% of doctors signal by picture, resulting in 71-83% doctors signaling gender by either report or picture. Table 1b shows that the quality signal (i.e., "Rating" measured by popularity) is a little higher for women in female-dominated fields and more so for men in male-dominated fields, and that, in terms of both professional and academic titles, women rank higher in female-dominated fields, and men rank higher in male-dominated fields. These differences are potentially consistent with statistical discrimination against female doctors in male-dominated fields, and against male doctors in female-dominated fields, which figures strongly in our subsequent analysis.

In terms of gender signaling, in female-dominated fields women are more likely to signal gender strongly (0.83 vs. 0.74), whereas in male-dominated fields men are more likely to signal gender strongly (0.83 vs. 0.71). In addition, women signal gender more strongly in female-dominated fields than they do in male-dominated fields, while men signal gender more strongly in male-dominated fields than in female-dominated fields. As we show below, these patterns broadly fit our predictions of how doctors choosing how to signal gender will respond to customer statistical discrimination.

5. Testing for statistical discrimination based on prices

In this section, we discuss the strategy for testing for customer statistical discrimination based on price gaps between sellers from different groups, and present evidence from this strategy. We also argue that, in our setting, this approach likely provides biased evidence.

The model

The statistical discrimination model was developed by Phelps (1972) and Aigner and Cain (1977). This model is not tied to prejudice. Rather, statistical discrimination arises when employers have imperfect information about productivity of potential employees, and that imperfect information leads them to attribute to individuals from different groups the average productivity of their group. Moreover, the average differences can be real (accurate), or they can reflect incorrect stereotypes. In its simplest formulation, statistical discrimination can be viewed as a model of stereotyping based on assumed group averages. If wages reflect expected productivity differences, statistical discrimination implies differences in wages that work against groups with lower (perceived) average productivity. This model carries over to our setting where we simply substitute the customer (patient)-service provider (doctor) relation for the employer-worker relation by replacing wages with service prices and worker productivity with doctor quality.

Let q denote a doctor's true quality. q is assumed to be unobserved to customers. Customers observe a noisy signal of q (the rating) given by $y = q + u$, and customers know that $E(q) = \alpha$, $E(u) = 0$, $Cov(q, u) = 0$ and $Var(u) = \sigma^2$. Given this information, customers know the "quality" of the signal. Customers are assumed to be risk-neutral. They form expectations of q given y as $E(q|y)$, and pay doctors their expected true quality in the competitive market, since this will satisfy the zero (expected) profit condition, so service price $p = E(q|y)$. If q and u are jointly normally distributed, then we can use standard formulas to solve for the expectation of one normal random variable conditional on another to obtain

(1) $$p = E(q|y) = \alpha \frac{Var(u)}{Var(q)+Var(u)} + y \frac{Var(q)}{Var(q)+Var(u)}.$$

Intuitively, if $Var(u)$ approaches infinity, the value of the signal falls to zero, and $E(q|y)$ approaches α. (This can be interpreted, in our setting, as the case when there are no reviews.) On the other hand, if the signal is perfectly accurate, $Var(u) = 0$ and $E(q|y) = y$. The signal-to-noise ratio in the second term, $\frac{Var(q)}{Var(q)+Var(u)}$, denoted by γ, is the "reliability" of the signal, $0 \leq \gamma \leq 1$. Adding subscripts to indicate gender group, we obtain the price equations for female and male doctors as

(2) $$p_F = E(q_F|y_F) = \alpha_F(1 - \gamma_F) + y_F\gamma_F$$

and

(3) $$p_M = E(q_M|y_M) = \alpha_M(1 - \gamma_M) + y_M\gamma_M.$$

We might assume that customers believe (whether true or not) that average quality of female doctors is lower than that of male doctors, i.e., $\alpha_F < \alpha_M$. In this case, as long as customers have imperfect information about doctors' true quality, the "reliability" of the quality signal $\gamma < 1$, and so customers put some weight $(1 - \gamma)$ on these group averages in forming an expectation about doctor quality. For simplicity, assume γ is the same for female and male doctors, i.e., $\gamma_F = \gamma_M = \gamma$. Then for the same quality signal $(y_F = y_M)$, customers expect female doctors on average to have lower quality, and would only be willing to hire female doctors for a lower price, so we should find $p_F < p_M$, with the price difference reflecting the expected quality difference.[27] The gender price gap is then

(4) $$p_F - p_M = (\alpha_F - \alpha_M)(1 - \gamma).$$

When customers acquire more information about doctors' quality – in our setting, as doctors accumulate more reviews – the signal-to-noise ratio in the average of the ratings increases towards 1. Thus, the weight $(1 - \gamma)$ on the (perceived) group average quality diminishes, and prices become less attributable to gender per se, and more attributable to the ratings (which may or may not differ by gender).

Testing for statistical discrimination based on prices

This is precisely the idea used in Laouénan and Rathelot (2022). In particular, if we estimate a regression for price, and presume that perceived average quality of female doctors is lower absent any other information, then there should be a female pay penalty (suggesting statistical discrimination against female doctors) when there are few if any reviews, and this penalty should diminish as reviews accumulate (and perhaps eventually disappear or even reverse, depending on true average quality differences).

In order to connect the conceptual framework and the empirical specification, we can combine equations (2) and (3) to obtain

(5) $$p = [\alpha_M \cdot (1 - Female) + \alpha_F \cdot Female] \cdot [(1 - \gamma(Number)]) + Rating \cdot \gamma(Number).$$

Female indicates gender of a doctor, *Number* the doctor's number of reviews, and *Rating* the "comprehensive recommendation popularity." Note that we have now written γ as a function of $Number$. We assume $\gamma' = \frac{\partial \gamma}{\partial Number} > 0$, based on the reduction in sampling variation (noise, from the point of view of patients) as the number of reviews increases. Then, assuming for simplicity no gender difference in $Rating$, which is supported by our sample summary statistics (as shown Table 1), the gender price difference is

[27] The underlying assumption is that labor supply is perfectly inelastic, so that quantities supplied do not adjust to expected quality (and how it changes). Alternatively, it may be enough for customers to assume that male and female doctors are perfect substitutes, so that no matter how supplies change, this price difference has to prevail in equilibrium.

(6) $p_F - p_M = (\alpha_F - \alpha_M) \cdot [1 - \gamma(Number)],$

implying that any gender gap in prices will move towards zero as *Number* increases.

Based on equation (5), we can estimate the following specification using each doctor-service pair as the unit of observation:

(7) $$p_{i,j,k,l} = \beta_0 + \beta_1 Female_i + \beta_2 Female_i \times Number_i + \beta_3 Rating_i \times Number_i + \beta_4 Number_i + \beta_5 Rating_i + \Lambda X_{i,j,k,l} + \epsilon_{i,j,k,l}.$$

In equation (7), $p_{i,j,k,l}$ denotes the price for doctor i in department j at hospital k providing service l. X is a vector of control variables: service characteristics, including: indicators for consultation service types (48-hour chat, one-question chat, and 4 phone call indicators for each place of listing), returning patients, and designated call length; doctor observable characteristics, including indicators for professional title and academic title; and hospital and field characteristics, including indicators for the primary hospital and department in the ranking list, number of hospitals a doctor works in, and the position of the listed hospital in all hospitals a doctor works in. $\epsilon_{i,j,k,l}$ is an i.i.d. random variable we assume to be uncorrelated with the regressors.[28]

To test for statistical discrimination based on prices, relying on the framework above, we are interested in three estimates. The estimate of β_1 reflects the perceived gender average quality difference $(\alpha_F - \alpha_M)$ when there are zero reviews. The estimate of $\beta_1 + \beta_2 \times Number_{i,j,k,l}$ reflects the perceived average gender quality difference evaluated at a specific number of reviews which is set the same for male and female doctors, conditional on the ratings; and correspondingly, the estimate of β_2 reflects the change in the weight put on gender as reviews accumulate. And the estimate of β_3 reflects the increased weight on ratings as reviews accumulate.

The framework above suggests that if there is statistical discrimination against female doctors (i.e., $\alpha_F - \alpha_M < 0$), we should find evidence that: (i) β_1 is negative; (ii) $\beta_1 + \beta_2 \times Number_{i,j,k,l}$ approaches zero (in the case of no true quality difference; more generally, it becomes less negative or even positive) as the number of reviews increases, so β_2 is positive, assuming β_1 is negative; and (iii) β_3 is positive.

We derive these predictions as follows. First,

(8) $\beta_1 = E(q|Female = 1) - E(q|Female = 0) = (\alpha_F - \alpha_M) \cdot [(1 - \gamma(Number)]).$

Since $\gamma < 1$, $\alpha_F - \alpha_M < 0$ (statistical discrimination against women) implies $\beta_1 < 0$.

---

[28] We can safely assume that $\epsilon_{i,j,k,l}$ is orthogonal to the controls, since patients generally have no other source of information about doctors. And if there is unobserved heterogeneity that is correlated with gender, the gender difference simply loads onto the perceived average difference by gender (i.e., this is not different from statistical discrimination).

Second,

(9) $\beta_2 = \frac{\partial(p_F - p_M)}{\partial Number} = \frac{\partial\{(\alpha_F - \alpha_M)\cdot[(1-\gamma(Number)]\}}{\partial Number} = -(\alpha_F - \alpha_M)\gamma'$,

$\gamma' > 0$ then implies that $\beta_2 > 0$.

Third, as derived from equation (5), the effect of $Rating$ on p would be increasing in $Number$:

(10) $\beta_3 = \frac{\partial p}{\partial Rating \partial Number} = \gamma' > 0$

These expected results are exactly what Laouénan and Rathelot (2022) find in their Airbnb application. However, in our setting, this is not what we find.

Evidence

Table 2 reports estimates of equation (7) for different samples. Column (1) reports the estimates for the full sample. The estimated effect of $Female$ is negative and significant, consistent with female doctors receiving lower prices when there are no reviews. The coefficient on $Female \times Number$ is positive and significant, implying that the pay penalty for female doctors declines as reviews accumulate, at least over a range up to above the 75th percentile of $Number$, before changing signs.[29] To here, the results are fully consistent with statistical discrimination. We also find a positive and significant coefficient on $Rating$, which is what we would expect. However, the coefficient on $Rating \times Number$ is negative and significant, which is not consistent with statistical discrimination, as the weight on $Rating$ should increase with the number of reviews, implying a positive coefficient on this interaction.

We present further evidence, for female- and male-dominated fields separately, in columns (2) and (3). This evidence is potentially more informative because we might expect statistical discrimination against male doctors in female-dominated fields, and against female doctors in male-dominated fields. We do not know that this is the case in this setting, of course, but we know there is evidence of hiring discrimination that follows this pattern, as discussed earlier. Moreover, we noted earlier that female doctors have higher professional and academic titles and ratings in female-dominated fields, and the same holds for male doctors in male-dominated fields. In column (3), the $Female$ coefficient is near zero and not significant in male-dominated fields. This would suggest that there is no statistical discrimination against women in male-dominated fields. In contrast, there may be statistical discrimination against women in female-dominated fields, for

[29] This is a consequence of the linear interaction, which does not allow the effect to asymptote to zero. Alternatively, the sign could change if the true average gender difference in quality (revealed with a large number of reviews) is the opposite sign of the difference in average initial beliefs about quality.

which in column (2) the $Female$ coefficient is larger negative and close to marginally significant. However, if there is no statistical discrimination against women in male-dominated fields, it is hard to explain the positive interaction on $Female \times Number$ in these fields. Overall, then, the results from these price regressions do not provide clear evidence consistent with statistical discrimination, either overall or for either female- or male-dominated fields.

The problem of testing for statistical discrimination based on prices when gender signal is a choice

The problem, however, is that this kind of evidence may not be applicable to testing for statistical discrimination in our setting, because doctors have a choice about how strongly to signal gender. As we show below, the choice about how strongly to signal gender is also likely to depend on the number of reviews, because as the average rating becomes more reliable, patients put more weight on the quality signal and less weight on the gender signal, so there is less reason to mask gender. If both the reliability of the rating *and* of the gender signal change with the number of reviews, and interact with each other endogenously, the implications derived above no longer hold; accordingly, the predictions of statistical discrimination for the price regressions (equations (8)-(10)) no longer hold.

To see this, and to interpret the empirical results in Table 2 when doctors have a choice about how strongly to signal gender, we introduce uncertainty about the gender signal. We characterize the gender signal in terms of the probability of being perceived as female, which depends on both the signal chosen and how gender-neutral the name is (detailed below in the next section). If a doctor is truly a female, we denote by θ the probability that a doctor is perceived as female, so $(1-\theta)$ is the probability that a female doctor is perceived as male. We similarly define θ and $(1-\theta)$ for male doctors.[30] Equations (2) and (3) then become

(2') $\quad p_F = E(q_F|y_F,\theta) = [\alpha_F + (1-\theta)\cdot(\alpha_M - \alpha_F)]\cdot[1-\gamma(Number)] + Rating\cdot\gamma(Number),$

and

(3') $\quad p_M = E(q_M|y_M,\theta) = [\alpha_M + (1-\theta)\cdot(\alpha_F - \alpha_M)]\cdot[1-\gamma(Number)] + Rating\cdot\gamma(Number).$

We can combine equations (2') and (3') to obtain an expanded version of equation (5):

(11) $$p = \{[\alpha_M + (1-\theta)(\alpha_F - \alpha_M)]\cdot(1-Female) + [\alpha_F + (1-\theta)(\alpha_M - \alpha_F)]\cdot Female\} \cdot[(1-\gamma(Number)]) + Rating\cdot\gamma(Number).$$

Similarly, equation (8) (again assuming for now no gender difference in $Rating$) expands to (after some algebra):

(12) $\quad \beta_1 = p_F - p_M = (\alpha_F - \alpha_M)\cdot(2\theta-1)\cdot[1-\gamma(Number)].$

[30] We could in principle allow these probabilities to depend on the true gender of the doctor, but we do not need this complication to establish the results that follow.

Since $0 \leq \theta \leq 1$, we have $-1 \leq 2\theta - 1 \leq 1$, and thus the sign of β_1 is indeterminate when the gender signal is uncertain. If $\theta > \frac{1}{2}$ then $2\theta - 1 > 0$ and $\beta_1 < 0$. Although not realistic, if $\theta < \frac{1}{2}$, then the signal confuses gender and we could find $\beta_1 > 0$. However, the estimates of β_1 in Table 2 are in fact negative.

Although we have not done so yet, we want to recognize that θ, as determined partially by the signals chosen, can be a function of the number of reviews ($Number$). We later show that $\theta' = \frac{\partial\theta}{\partial Number} > 0$ in the data (once we include the choice about gender signaling, which we have not done yet). We also predict (and test) that θ changes differently for female and male doctors depending on whether they work in female- or male-dominated fields. Thus, we can rewrite equation (12) as,

(12') $\quad \beta_1 = p_F - p_M = (\alpha_F - \alpha_M) \cdot [2\theta(Number) - 1] \cdot [1 - \gamma(Number)]$,

although we sometimes suppress the argument ($Number$) of θ below.

The derivative of the expression for the expected pay penalty for women becomes (now taking account of θ being a function of $Number$):

(13) $\quad \beta_2 = \frac{\partial(p_F - p_M)}{\partial Number} = (\alpha_F - \alpha_M)[2\theta' \cdot (1-\gamma) - (2\theta - 1) \cdot \gamma']$.

The sign of β_2 is indeterminate even if $\theta' = 0$, based on the argument following equation (12), although in this case if $\theta > \frac{1}{2}$ then $\beta_2 > 0$. However, with $\theta' > 0$, the first term in square brackets in equation (13) is positive, and thus the sign of the full expression in square brackets (and hence β_2) remains indeterminate even when $\theta > \frac{1}{2}$. Intuitively, as the number of reviews grows, the reliability of the average rating increases, which reduces the impact of the gender signal, but the reliability of the gender signal chosen also increases, which has the opposite impact, and if this impact is larger the prediction for β_2 is reversed. Thus, there is no longer a clear prediction from the expanded statistical discrimination model for the sign of the coefficient on $Female \times Number$ (β_2) in equation (7) – how the gender price difference evolves with the number of reviews. This, for example, might be why we obtain a positive estimate of β_2 for male-dominated fields even where the estimated coefficient on $Female$ is not negative.

We also noted that the estimated coefficient on β_3 in equation (7) does not align with the predictions of statistical discrimination in any columns of Table 2. To see why the prediction becomes ambiguous, we allow different reliability of the signals for female and male doctors, introducing two parameters, γ_F and γ_M, in which case equation (11) becomes (suppressing the $Number$ arguments):

(14) $$p = \{[\alpha_M + (1-\theta)(\alpha_F - \alpha_M)] \cdot (1 - Female) + [\alpha_F + (1-\theta)(\alpha_M - \alpha_F)] \cdot Female\} \cdot [(1 - \{\gamma_F \cdot \theta + (1-\theta)\gamma_M\}]) + (\gamma_F \cdot \theta + (1-\theta)\gamma_M) \cdot Rating.$$

In this case,

(15) $$\frac{\partial p}{\partial Rating} = \gamma_F \cdot \theta + \gamma_M \cdot (1-\theta),$$

and

(16) $$\frac{\partial^2 p}{\partial Rating \partial Number} = \gamma_F' \cdot \theta + \gamma_M' \cdot (1-\theta) \ + (\gamma_F - \gamma_M) \cdot \theta'.$$

The prediction from the standard statistical discrimination framework was that $\frac{\partial^2 p}{\partial Rating \partial Number} > 0$ because the average rating signal becomes more reliable as the number of reviews increases. If the gender signal does not vary with the number of reviews, then $\theta' = 0$ and equation (16) is unambiguously positive (since both $\theta > 0$ and $(1-\theta) > 0$, and the third term becomes zero). In contrast, we can see that $\frac{\partial^2 p}{\partial Rating \partial Number}$ is unambiguously more likely to be negative when $(\gamma_M - \gamma_F) \cdot \theta'$ is large. That is, this latter term is large when the gender signal becomes stronger as the number of reviews increases, and the reliability of the average rating for male doctors is viewed as higher than the reliability of the average rating for female doctors. The intuition is that when θ' is large, patients are more certain of the gender of the doctor as the number of reviews increases. When $(\gamma_M - \gamma_F)$ is also large, this makes patients discount the average rating relatively more for female doctors, and put more weight on the gender signal. Conversely, if $\gamma_M = \gamma_F$, there is no reason the increase in θ would lead to downweighting the rating, since the rating is viewed the same whether the doctor is female or male, and again the last term is zero and the expression in equation (16) (and hence β_3) is positive.

Together, this analysis explains why the predictions of the statistical discrimination framework for the regression model (equation (7)) no longer necessarily hold when doctors can choose how strongly to signal their gender (or more generally how sellers can choose how strongly to signal their group membership). We therefore next develop our test for statistical discrimination based on choices about gender signaling.

6. Using gender signaling to identify discrimination

Testing for statistical discrimination based on gender signaling

We have shown that when the signal of group membership is uncertain, and sellers can choose the strength of the signal, using prices to test for statistical discrimination is invalid. In this section, we develop an alternative test based on sellers' choices about signaling their group membership – in our case, doctors' choices about signaling their gender. The core idea is that a group that experiences statistical discrimination against it will tend to mask its group membership when customers (patients) have little information about their true quality aside from their group membership. But, consistent with the framework outlined in the previous section, as patients get more reliable information from the reviews (as the number of reviews accumulates), there is less incentive to mask gender. Thus, a group that masks gender when there are few reviews, but reveals gender more fully as the number of reviews increases, can be interpreted as experiencing – and responding to – statistical discrimination. Moreover, the incentive to mask gender can differ depending on the direction of the statistical discrimination. In particular, the incentive for women to mask gender when there are few reviews will be stronger when there is statistical discrimination against them, which we think is more likely in male-dominated fields, and the opposite could hold in female-dominated fields, where we think statistical discrimination *in favor* of female doctors is more likely.

To preview the results, our analysis based on this idea indicates that female doctors experience statistical discrimination against them in male-dominated fields, and male doctors experience statistical discrimination against them in female-dominated fields. We next flesh this test out in more detail, before turning to the evidence.

To develop this argument, we need to introduce the choice of how strongly to signal gender, which we would also expect to be affected by how gender neutral a doctor's name is. (For example, there is less incentive to mask gender – if there is such an incentive – if one's name is highly gendered). We thus expand θ – the probability of being perceived as female (or male) – to depend on both how strongly gender is signaled, and how gender neutral the name is. Since this can vary by gender, as female and male doctors may make different choices, we define $\theta_{F(F)}$ to be the probability that a female doctor is perceived as female, and $\theta_{M(M)}$ to be the probability that a male doctor is perceived as male.

We first introduce a parameter that captures the gender neutrality of a doctor's name, δ. Again, we define these as $\delta_{F(F)}$ – the probability that a female doctor is perceived as female given her name – and we similarly define $\delta_{M(M)}$ for male doctors. Consequently, for female doctors, for example, the probability that they are perceived as males given their names is $\left(1-\delta_{F(F)}\right)$. δ ranges from 0.5 to 1.

We then characterize gender signaling as a doctor's choice about the share of the unknown probability about one's gender to eliminate by signaling gender more strongly. We parameterize the gender signal for female doctors as the choice of $\pi_{F(F)}$, which ranges between 0 and 1, defined so that the probability a female doctor is perceived as female is

(17) $\theta_{F(F)} = 1 - \left(1 - \pi_{F(F)}\right) \cdot \left(1 - \delta_{F(F)}\right).$

Thus, the higher is $\pi_{F(F)}$, the higher is the probability that a female doctor is perceived as female; e.g., when $\pi_{F(F)} = 1$, $\theta_{F(F)} = 1$. We define the parameter $\pi_{M(M)}$ symmetrically for male doctors. $\pi_{F(F)}$ and $\pi_{M(M)}$ are the choices about gender signaling that female and male doctors make – the basis of our proposed test for statistical discrimination when this signal is a choice.

We can then augment the price equations (2') and (3') to include these perceived probabilities of the gender of doctors:

(2'') $$p_F = E(q_F|y_F,\theta) = \left[\alpha_F + \left(\left(1-\pi_{F(F)}\right)\cdot\left(1-\delta_{F(F)}\right)\right)\cdot(\alpha_M-\alpha_F)\right][1-\gamma_F] + Rating\cdot\gamma_F,$$

and

(3'') $$p_M = E(q_M|y_M,\theta) = \left[\alpha_M + \left(\left(1-\pi_{M(M)}\right)\cdot\left(1-\delta_{M(M)}\right)\right)\cdot(\alpha_F-\alpha_M)\right][1-\gamma_M] + Rating\cdot\gamma_M.$$

For example, if a female doctor signals her gender with certainty, then $\pi_{F(F)} = 1$ and $[\alpha_F + \left(\left(1-\pi_{F(F)}\right)\cdot\left(1-\delta_{F(F)}\right)\right)\cdot(\alpha_M-\alpha_F)]$ collapses to α_F. In contrast, if the only information is the name, then this expression reduces to $\{\alpha_F + \left(1-\delta_{F(F)}\right)\cdot(\alpha_M-\alpha_F)\}$, with the weight on α_F higher the more strongly her name signals female gender.

It is clear from equations (2'') and (3'') that a doctor whose gender is perceived as less productive has an incentive to mask gender. For example, for female doctors

(18) $$\frac{\partial p_F}{\partial \pi_{F(F)}} = \left(1-\delta_{F(F)}\right)\cdot(\alpha_F-\alpha_M)(1-\gamma_F)$$

and for men

(19) $$\frac{\partial p_M}{\partial \pi_{M(M)}} = \left(1-\delta_{M(M)}\right)\cdot(\alpha_M-\alpha_F)(1-\gamma_M).$$

Suppose there is statistical discrimination against female doctors, or $\alpha_F < \alpha_M$. In this case, $\frac{\partial p_F}{\partial \pi_{F(F)}} < 0$, so female doctors earn a higher price by masking their gender – i.e., choosing a smaller value of $\pi_{F(F)}$. That is because by masking gender, female doctors are penalized less by the average lower quality for female than male doctors that patients assume. In contrast, if there is statistical discrimination against male doctors, $\alpha_F > \alpha_M$, then $\frac{\partial p_F}{\partial \pi_{F(F)}} > 0$, implying that female doctors earn a higher price by signaling their gender more strongly. Similar implications for signaling gender by male doctors follow.

We can derive other predictions. First, how does the choice to signal gender more strongly vary with the number of reviews? From equations (18) and (19),

(20) $$\frac{\partial^2 p_F}{\partial \pi_{F(F)}\partial Number} = -\left(1-\delta_{F(F)}\right)\cdot(\alpha_F-\alpha_M)\gamma_F{}'$$

and

(21) $\frac{\partial^2 p_M}{\partial \pi_{M(M)} \partial Number} = -\left(1-\delta_{M(M)}\right) \cdot (\alpha_M - \alpha_F)\gamma_M{}'$.

If there is statistical discrimination against female doctors ($\alpha_F < \alpha_M$), then $\frac{\partial^2 p_F}{\partial \pi_{F(F)} \partial Number} > 0$ and $\frac{\partial^2 p_M}{\partial \pi_{M(M)} \partial Number} < 0$, implying that as the number of reviews increases, there is less incentive for women to mask their gender. The reverse is true if there is statistical discrimination against male doctors, and there is less incentive for them to mask their gender as reviews accumulate.

The equations provide the bases for our tests of statistical discrimination based on gender signaling, and our inference of the direction of statistical discrimination. In particular, the group that experiences statistical discrimination against it based on gender should signal gender more weakly (mask gender more strongly) initially, but has less incentive to do so as ratings accumulate, and hence average ratings become more reliable.[31]

We test for statistical discrimination by estimating the following model using each doctor as the unit of observation, where $Signal_{i,j,k}$ is a dummy variable for a strong signal (picture, report, or both).

(22) $$Signal_{i,j,k} = \alpha + \beta_1 Female_i + \beta_2 Female_i \times Number_i + \beta_3 Number_i + \beta_4 Rating_i + \beta_5 Rating_i \times Number_i + \Lambda X_{i,j,k} + \epsilon_{i,j,k}.$$

All independent variables and subscripts are defined the same as in equation (7); the only difference is that we have a single observation per doctor, and hence drop the controls for service type.

We are interested in the estimates of β_1 and β_2. As shown above, if there is statistical discrimination against women, we should find that when there are fewer reviews female doctors are less likely to signal gender strongly, implying $\beta_1 < 0$. But as reviews accumulate, there is less incentive for women to mask gender as less weight is put on assumed average quality difference ($\alpha_F - \alpha_M$). Thus, when there is statistical discrimination against women, we should find $\beta_2 > 0$.

We do not know whether, in general, there is statistical discrimination against female doctors. Thus, we are agnostic about the predictions when we use all the data. However, we believe it is more likely that there is statistical discrimination against female doctors in male-dominated

[31] In addition, reviews may themselves provide information about the gender of doctors, making masking gender less useful as reviews accumulate, although we did not scrape the content of reviews nor do we know whether and how thoroughly patients read them. One could also imagine some positive return to conveying gender accurately, which becomes relatively more decisive as reviews accumulate.

fields, and statistical discrimination against male doctors in female-dominated fields. We thus test these predictions for the choice about signaling gender for female- and male-dominated fields separately. The prediction is that $\beta_1 > 0$ and $\beta_2 < 0$ in female-dominated fields, but $\beta_1 < 0$ and $\beta_2 > 0$ in male-dominated fields. In either case, one gender has less incentive to signal their gender strongly when patients put more weight on the higher assumed quality for the other gender, and this incentive to mask gender declines as reviews accumulate.

Finally, note from equations (18)-(21) that all of the derivatives for the choice of gender signaling are larger (in absolute value) when $\delta_{F(F)}$ or $\delta_{M(M)}$ are closer to 0.5 (their minimum). This gives rise to a testable implication from varying the sample based on how gender-neutral names are. In the other direction, these derivatives approach 0 as $\delta_{F(F)}$ or $\delta_{M(M)}$ approach 1 – which means that the name more strongly signals gender (with 1 implying the name perfectly signals gender). This implies what might be viewed as a falsification test. In particular, when the name is highly gendered there should be little or no difference in how gender influences the choice to signal gender strongly, nor should the number of reviews influence this choice.

Evidence

We begin with some simple visual evidence. Figures 3a and 3b depict how the share of doctors strongly signaling gender by report and/or picture evolves with the number of reviews in female- and male-dominated fields, respectively, for the samples for which we estimate equation (22). (As explained above, the predictions for the pooled data are less clear, so we do not show the pooled figure although we report regression results below.) We divide the data into bins that each contain about 5% of observations of the samples.[32] Figure 3a shows that in female-dominated fields, the share of female doctors strongly signaling gender starts from around 60% when there are no reviews. This is nearly one-third higher than the share for male doctors in these fields. The share increases with number of reviews for both women and men, although it increases faster for men and the gap largely closes by the time the number of reviews exceeds 30 or so. Both of these results are consistent with what we predict for the choice of signaling gender by female and male doctors in the presence of statistical discrimination against male doctors in female-dominated fields. (The first is consistent with $\beta_1 > 0$, and the second with $\beta_2 < 0$.) The opposite holds in Figure 3b for male-dominated fields. Here – consistent with statistical discrimination against women in these fields – women are initially (with few or no reviews) less likely to strongly signal

[32] We bin the observations this way because there are far more observations with low numbers of reviews (as the figures show), so bins defined based on equal ranges of the number of reviews would have a very uneven distribution of the number of doctors across bins.

gender than men, while they increase the likelihood of signaling gender more quickly than men as reviews accumulate.

Turning to regression evidence, Table 3 reports the estimates of equation (21) for different samples defined along two dimensions. First, columns (1)-(3) report regression results for all doctors, and then for doctors who work in female-dominated or male-dominated fields, as defined earlier based on the share of female doctors. For all fields (column (1)), the estimated coefficients on $Female$ (β_1) and $Female \times Number$ (β_2) are both positive but neither is statistically significant. These results do not align with the predictions of statistical discrimination for gender signaling. However, if statistical discrimination differs in female- and male-dominated fields, then the implications for the estimates in column (1) differ for different parts of the sample and hence are not clear.

In contrast, in female-dominated fields (column (2)), the estimated effect of $Female$ is positive and significant, and the coefficient on $Female \times Number$ is negative and significant, which is consistent with how male doctors would respond to statistical discrimination in favor of female doctors in these fields, with male doctors masking their gender more strongly initially ($\beta_1 > 0$) but then the difference between male and female doctors in gender signaling diminishing as the number of reviews grows ($\beta_2 < 0$). This evidence is thus consistent with statistical discrimination against male doctors in female-dominated fields. In contrast, in male-dominated fields (column (3)), the opposite is observed, i.e., the estimated effect of $Female$ is negative and significant, and the coefficient on $Female \times Number$ is positive and significant. This evidence is consistent with statistical discrimination against female doctors in male-dominated fields.

In columns (1')-(3') we restrict the samples in all three cases to doctors whose names are more gender-neutral. We define gender-neutral names as those with less than 80% probability of accuracy as to either gender, based on the "ngender" package in Python. Recall that the prediction is that the absolute values of the estimates of β_1 and β_2 should be larger (in absolute value) for doctors with gender-neutral names in the presence of statistical discrimination, since there is more gain from initially masking gender (by the group against which there is statistical discrimination). This prediction is borne out in the data. In both columns (2') and (3') we continue to find evidence consistent with statistical discrimination against male doctors in female-dominated fields ($\beta_1 > 0$ and ($\beta_2 < 0$), and vice versa for female doctors in male-dominated fields ($\beta_1 < 0$) and ($\beta_2 > 0$), but the estimated magnitudes are always larger in absolute value in columns (2') and (3') compared to columns (2) and (3), respectively.

Additional analyses

We next report on some additional analyses that provide more evidence on the validity of our assumptions, explore the robustness of our findings, and test additional implications of the theory.

Confirming evidence

We have interpreted our results on gender signaling as consistent with statistical discrimination against male doctors in female-dominated fields, and against female doctors in male-dominated fields. In Table 4, we present additional evidence consistent with this interpretation. Specifically, we focus attention on fields that are not clearly either female- or male-dominated – those with the share of female doctors between 40% and 60%. The absolute magnitudes of the estimated coefficients on $Female$ and $Female \times Number$ in Table 4 are much smaller than the corresponding estimates in columns (2)-(3) and (2')-(3') in Table 3, and three of the four estimates are not statistically significant. Interpreted through the lens of our test, this would imply that there is not clear evidence of statistical discrimination against either gender in the fields that have roughly equal shares of female and male doctors. Assuming either that women tend to enter fields where there is not statistical discrimination against them, and similarly for men, or that patients' statistical discrimination is based on the "norm" for which gender dominates a field, this evidence bolsters the interpretation of our evidence as reflecting statistical discrimination against men in female-dominated fields, and against women in male-dominated fields.

Stricter definitions of female- and male-dominated fields and gender-neutral names

In Table 5 we alter the definitions of female- and male-dominated fields as well as the criterion for gender-neutral names. We impose a higher cutoff for female- and male-dominated fields (73% female for the former and 20% female for the latter).[33] And we tighten the definition of gender-neutral names to those with less than 70% probability of accuracy as to either gender. We interpret both of these changes as getting us closer, in a sense, to the "ideal experiment," i.e., fields that are closer to 100% female or male, where the statistical discrimination in favor of or against women should be more stark, and names that are closer to truly gender neutral, where the incentive to mask gender initially may be stronger for those against whom patients statistically discriminate. We might generally expect stronger results (larger magnitudes in absolute value)

[33] Our initial inclination was to simply move these from 70% and 30% to 80% and 20%. However, there is only one department (67 doctors, see Table A1) with the share of female doctors over 80%, so we used a cutoff for female-dominated fields of 73% to capture the two departments with the largest share of female doctors. This adds the very large Obstetrics and Gynecology field to the set of female-dominated fields.

for the estimates of β_1 and β_2, although we cannot be sure because the sample changes (and the sample size reduction tends to make the results statistically weaker).

Columns (1)-(4) in Table 5 correspond to columns (2), (3), (2') and (3') in Table 3, but with stricter definitions of female- and male-dominated fields. The results still hold. One difference is that the estimate of β_2 (the coefficient on $Female \times Number$) is not statistically significant in column (3). The estimated magnitudes are sometimes larger (e.g., the estimated coefficients of β_2 in columns (1) and (2) of Table 5 vs. columns (2) and (3) of Table 3).

In columns (5) and (6) we revert to the definition of female- and male-dominated fields from Table 3 but impose the stricter criterion for gender-neutral names. Again, the signs of the estimated coefficients are all consistent with statistical discrimination against male doctors in female-dominated fields, and vice versa. Two estimated absolute magnitudes in column (6) are larger than the corresponding estimates in column (3') of Table 3. Overall then, the results are robust to tightening the definitions of female- and male-dominated fields and gender-neutral names.

Alternative measures of quality signal

In the main analysis we use comprehensive recommendation popularity as the measure for doctors' quality of service. We next explore the robustness of our findings to using alternative measures. As discussed earlier, there are five other statistics patients could potentially use to evaluate doctors although we think these are less preferable and less salient: the treatment satisfaction rate, the bedside manner satisfaction rate, the number of thank you letters, the number of gifts, and the online service satisfaction rate. Since, however, only the first four ratings are provided by both online and offline patients, and only these ratings are not conditional on exceeding a certain value, we use only the first four measures. We construct a standardized average of these other ratings. We first standardize each of these four variables based on the available observations, and then compute the average over the non-missing standardized rating variables for each doctor. The results using this alternative rating are reported in Table 6. The results are nearly identical to those in corresponding columns in Table 3, which shows the robustness of our results.

Alternative measures of reliability of quality signal

In the main analysis we use the number of patient reviews as the measure for reliability of doctors' quality signal. There is an alternative measure, which is the number of online consultations. While this statistic also appears on the top-right corner of the personal webpage (see Figure 1) and can be viewed (incorrectly) as the number of reviews by patients, it is actually not directly related to patient reviews. Nonetheless, Table 7 reports the same specifications as Table 3 using this alternative measure. The results are robust. Note that the estimated coefficients on

$Female \times Number$ are smaller because the number on online consultations is much higher than the number of patient reviews (see Appendix Table A3a and Table 1a for summary statistics).

Falsification test

Finally, we perform the falsification test described earlier. We restrict the sample to those with highly-gendered names (above 90% probability of accuracy). We expect no gender difference in gender signaling overall or in the female- or male-dominated fields, since the doctors' names already by and large reveal their gender. The estimates in Table 8, which should be contrasted with the results for those with more gender-neutral names in, e.g., columns (1')-(3') in Table 3, bear this out. The estimated coefficients on $Female$ and $Female \times Number$ are all close to zero and insignificant, consistent with no relationship between gender signaling, statistical discrimination, and the reliability of the quality signal when name already signals gender very strongly.

7. Conclusion and discussion

We study data on an online market for doctors in China, testing for statistical discrimination by patients based on doctor gender. We have prices for doctors' online services, quality ratings, and the number of reviews on which ratings are based, and different types of information on their gender that doctors provide. We develop a new approach to testing for statistical discrimination based on doctors' choices about how strongly to signal their gender. This approach is applicable in many markets and perhaps especially online markets, because economic agents sometimes have discretion over how much information to reveal about their membership in groups that may either suffer from or benefit from statistical discrimination. For example, in an online market with photographs of sellers, there may be a choice about how clearly the photograph reveals gender, race, etc., or even whether to include a photograph.[34] In our setting, doctors have the choice of providing additional information that more strongly signals gender – in addition to their name – via either a picture, a biography, or both.

We first lay out the conventional statistical discrimination framework and how it underlies a test of statistical discrimination based on gender information, price differences by gender, and how these price differences evolve as reviews accumulate (as in the application in a different context in Laouénan and Rathelot, 2022). We then show when doctors can choose how strongly to signal their gender – an issue we think is likely endemic to these kinds of tests – the statistical discrimination model does not have clear predictions for price differences by gender and how they

[34] As an example, in research using LinkedIn data, Berry et al. (2024) document that photographs on LinkedIn profiles sometimes have varying lighting, or sometimes show more than one person in a photo.

evolve.

We then extend the conventional statistical discrimination framework to incorporate gender signaling, and show how to test for statistical discrimination by studying the dynamics of gender signaling for female and male doctors. We find evidence of statistical discrimination against female doctors in male-dominated fields (with male doctors viewed as higher quality absent other information), and statistical discrimination against male doctors in female-dominated fields (where female doctors are viewed as higher quality absent other information). The evidence of this is that female doctors mask gender more strongly initially in male-dominated fields, and male doctors do the same in female-dominated fields. But in both female- and male-dominated fields these gender differences in signaling of gender decrease with number of customer reviews of doctors. In other words, female doctors in male-dominated fields, and male doctors in female-dominated fields, choose less informative gender signals initially, when the gender signal is most of the information customers have. But as reviews accumulate and customers rely more on a doctor-specific quality signal, there is less incentive to mask gender.

Our evidence of statistical discrimination against women in male-dominated fields and against men in female-dominated fields is consistent with some other evidence on discrimination. In particular, although not about statistical discrimination per se, field experiment evidence from correspondence studies points to hiring discrimination against women in male-dominated jobs, and against men in female-dominated jobs (and, also consistent with our evidence, an absence of discrimination in more integrated jobs).[35]

Finally, we want to emphasize that the evidence of statistical discrimination against women in male-dominated fields and against men in female-dominated fields does not imply a sort of "neutrality" that does not on net harm women. There are far more medical positions in fields with a high share of male doctors than a high share of female doctors (Appendix Table A1). Thus, statistical discrimination against female doctors in male-dominated fields may pose a significant barrier to women entering many fields of medicine.

[35] See Neumark (2018) and in particular Riach and Rich (2002) and Rich (2014).

References

Aigner, D., and Cain, G. (1977). Statistical theories of discrimination in labor markets. Industrial and Labor Relations Review, 30(2), pp. 175-187.

Altonji, J., and Pierret, C. (2001). Employer learning and statistical discrimination. The Quarterly Journal of Economics, 116(1), pp. 313-350.

Bergen, Jr, S.S., 2000. Underrepresented minorities in medicine. JAMA, 284(9), pp.1138-1139.

Berry, A., Maloney, M., and Neumark, D. (2024). The missing link? Using LinkedIn data to measure race, ethnic, and gender differences in employment outcomes at individual companies. NBER Working Paper No. 32294.

Bohren, J., Imas, A., and Rosenberg, M. (2019). The dynamics of discrimination: Theory and evidence. American Economic Review, 109(10), pp. 3395-3436.

Brown, T., Liu, J.X., and Scheffler, R.M. (2009). Does the under or overrepresentation of minority physicians across geographical areas affect the location decisions of minority physicians? Health Services Research, 44 (4), pp. 1290–1308.

Brownlee, J. (2016). A gentle introduction to the gradient boosting algorithm for machine learning. [online] Available at: https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/ [Accessed 30 Apr. 2024]

Chan, A. (2023). Discrimination against doctors: A field experiment. Unpublished paper

Chen, Y. (2024). Does the gig economy discriminate against women? Evidence from physicians in China. Journal of Development Economics, 169, 103275.

Cui, R., Li, J., and Zhang, D. (2020). Reducing discrimination with reviews in the sharing economy: Evidence from field experiments on Airbnb. Management Science, 66(3), pp. 1071-1094.

Exley, C.L. et al. (2024). Information-optional policies and the gender concealment gap. NBER Working Paper No. 32350.

Feld, J., Ip, E., Leibbrandt, A., and Vecci, J. (2022). Identifying and overcoming gender barriers in tech: A field experiment on inaccurate statistical discrimination. CESifo Working Paper No. 9970.

Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, pp. 1189-1232.

GDO (good doctor online), (2024). Introduction about good doctor online (in Chinese). Available at: https://www.haodf.com/info/aboutus.php. [Accessed 6 Feb. 2023].

Ge, Q., and Wu, S. (2023). How do you say your name? Difficult-to-pronounce names and labor market outcomes. Unpublished paper.

Gravelle, H., Hole, A., and Santos, R. (2011). Measuring and testing for gender discrimination in physician pay: English family doctors. Journal of Health Economics, 30(4), pp. 660-674.

Grisham, Sarah, (2017). Medscape Physician Compensation Report 2017. Available at: https://www.medscape.com/slideshow/compensation-2017-overview-6008547. [Accessed 5 Apr. 2023]

Hu, Y. et al. (2019). Discrimination, abuse, harassment, and burnout in surgical residency training. New England Journal of Medicine, 381(18), pp. 1741-1752.

Jagsi, R. et al. (2012). Gender differences in the salaries of physician researchers. JAMA, 307(22), pp. 2410-2417.

Jena, A., Olenski, A., and Blumenthal, D. (2016). Sex differences in physician salary in US public medical schools. JAMA Internal Medicine, 176(9), pp. 1294-1304.

Kang, S., DeCelles, K., Tilcsik, A., and Jun, S. (2016). Whitened résumés: Race and self-presentation in the labor market. Administrative Science Quarterly, 61(3), pp. 469-502.

Laouénan, M., and Rathelot, R. (2022). Can information reduce ethnic discrimination? Evidence from Airbnb. American Economic Journal: Applied Economics, 14(1), 107-32.

Lett, E., et al. (2019). Trends in racial/ethnic representation among US medical students. JAMA network open, 2 (9), pp. 1910490–1910490.

Ly, D.P., Seabury, S.A., and Jena, A.B. (2016). Differences in incomes of physicians in the United States by race and sex, BMJ: British Medical Journal, p.353.

McKenna, Jon. 2024. Medscape Physician Compensation Report 2024: Bigger checks, yet doctors still see an underpaid profession. Available at: https://www.medscape.com/slideshow/2024-compensation-overview-6017073#2 (requires login).

Neumark, D. (2018). Experimental research on labor market discrimination. Journal of Economic Literature, 56(3), pp. 799-866.

Pallais, A. (2014). Inefficient hiring in entry-level labor markets. American Economic Review, 104(11), pp. 3565-3599.

Phelps, E. (1972). The statistical theory of racism and sexism. American Economic Review, 62(4), pp. 659-661.

Riach, P.A., and Rich, J. (2002). Field experiments of discrimination in the market place. Economic Journal, 112(483), pp. 480-518.

Rich, J. (2014). What do field experiments of discrimination in markets tell us? A meta analysis of studies conducted since 2000. IZA Discussion Paper No. 8584.

Sarsons, H. (2017). Interpreting signals in the labor market: evidence from medical referrals. Unpublished paper.

Yan, W. (1990). Introduction to nursing management standards and evaluation methods at various levels of hospitals (Trial implementation). Chinese Journal of Nursing (in Chinese), 25(05), pp. 209-210.

Zao, X., and Biernat, M. (2017). “Welcome to the U.S.” but “change your name”? Adopting Anglo names and discrimination. Journal of Experimental Social Psychology, 70, pp. 59-68.

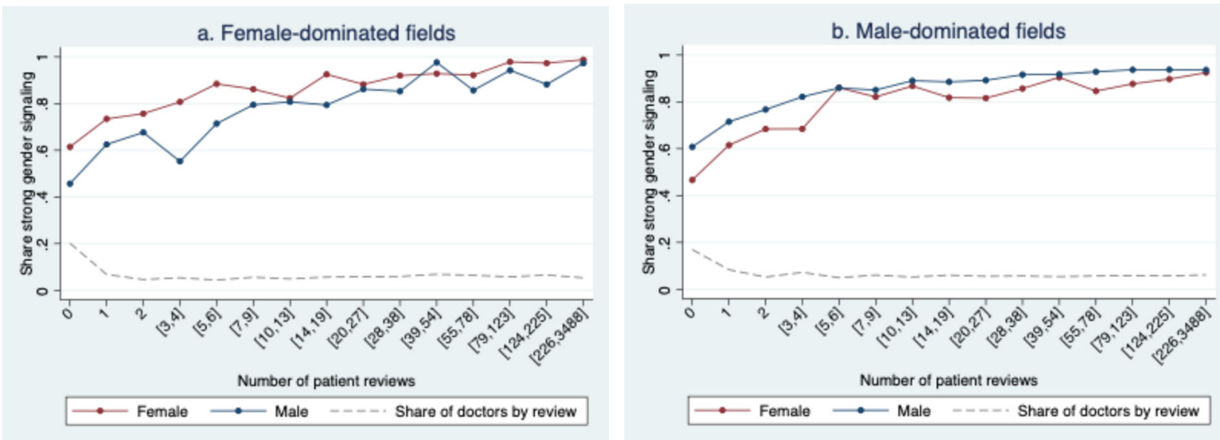
Figure 1. Example: screenshot of a doctor's personal webpage

Note: We inserted English translations of key items in text boxes.

Figure 2. An example of doctor's individual online consultation page

Note: We inserted English translations of key items in text boxes.

Figure 3. Evolution of share of doctors signaling gender with number of reviews by gender



Notes: we cut number of reviews into bins that have around 5% of doctors based on the analysis sample – if one single number of review has more than 5% doctors, it is a bin by itself; if one single number of reviews has fewer 5% of doctors, it is combined with the next number, done successively until the sum of shares in that bin is around 5%. Figure 3a and 3b are produced for the regression sample for female- and male-dominated fields, respectively. The solid lines show the share of doctors strongly signaling gender within each bin. The dashed lines show the share of doctors within each bin.

Table 1a. Summary statistics for regression variables the full sample and by gender

	All doctors		Female doctors		Male doctors	
	Obs.	Mean (Std. Err.)	Obs.	Mean (Std. Err.)	Obs.	Mean (Std. Err.)
Price	90,457	51.36 (0.25)	29,675	50.80 (0.42)	60,782	51.64 (0.31)
48-hour chat price	31,938	51.27 (0.42)	10,689	52.05 (0.74)	21,249	50.87 (0.51)
One-question chat price	31,467	27.64 (0.22)	10,522	27.38 (0.37)	20,945	27.77 (0.28)
Phone call price	27,052	79.08 (0.57)	8,464	78.32 (0.98)	18,588	79.42 (0.69)
Female	35,777	0.33 (0.002)	n.a.	n.a.	n.a.	n.a.
Signal by report	35,777	0.43 (0.003)	11,899	0.45 (0.005)	23,878	0.42 (0.003)
Signal by picture	35,777	0.63 (0.003)	11,899	0.59 (0.005)	23,878	0.66 (0.003)
Signal by report and/or picture	35,777	0.78 (0.002)	11,899	0.76 (0.004)	23,878	0.79 (0.003)
Rating(popularity)	35,777	3.38 (0.002)	11,899	3.34 (0.004)	23,878	3.40 (0.003)
Number of patient reviews	35,777	48.73 (0.70)	11,899	36.12 (0.93)	23,878	55.01 (0.93)
Chief physician	35,777	0.24 (0.002)	11,899	0.23 (0.004)	23,878	0.25 (0.003)
Associate chief physician	35,777	0.31 (0.002)	11,899	0.29 (0.004)	23,878	0.32 (0.003)
Attending physician	35,777	0.33 (0.002)	11,899	0.34 (0.004)	23,878	0.32 (0.003)
Resident physician	35,777	0.11 (0.002)	11,899	0.13 (0.003)	23,878	0.10 (0.002)
No professional title reported	35,777	0.002 (0.0002)	11,899	0.002 (0.0004)	23,878	0.002 (0.0003)
Professor	35,777	0.14 (0.002)	11,899	0.12 (0.003)	23,878	0.15 (0.002)
Associate Professor	35,777	0.17 (0.002)	11,899	0.15 (0.003)	23,878	0.18 (0.002)
Assistant Professor	35,777	0.09 (0.002)	11,899	0.08 (0.002)	23,878	0.10 (0.002)
Teaching assistant	35,777	0.01 (0.0005)	11,899	0.01 (0.0008)	23,878	0.01 (0.0006)
No academic title reported	35,777	0.59 (0.003)	11,899	0.65 (0.004)	23,878	0.56 (0.003)
Returning patient	90,457	0.001 (0.0001)	29,675	0.002 (0.0002)	60,782	0.001 (0.0001)
Designated call length	90,457	3.27 (0.02)	29,675	3.11 (0.03)	60,782	3.35 (0.02)

Notes: Summary statistics for the full sample and by gender for the regression sample. Summary statistics for some variables that vary little by gender and are not consequential are not shown, such as number of hospitals a doctor works in and the position of the listed hospital.

Table 1b. Summary statistics for regression variables for gender fields and by gender

	Female-dominated fields (share female > 70%)				Male-dominated fields (share female < 30%)			
	Female doctors		Male doctors		Female doctors		Male doctors	
	Obs.	Mean (Std. Err.)	Obs.	Mean (Std. Err.)	Obs.	Mean (Std. Err.)	Obs.	Mean (Std. Err.)
Price	4,252	58.21	1,557	59.91	2,560	38.43	31,537	47.88
		(1.39)		(2.22)		(1.05)		(0.39)
48-hour chat price	1,495	59.93	533	60.25	947	37.72	10,996	46.25
		(2.66)		(3.81)		(1.64)		(0.62)
One-question chat price	1,464	31.94	521	30.88	931	21.07	10,874	25.98
		(1.23)		(1.87)		(0.87)		(0.37)
Phone call price	1,293	85.98	503	89.61	682	63.1	9,667	74.36
		(2.86)		(4.88)		(2.73)		(0.91)
Signal by report	1,660	0.49	602	0.36	1,010	0.35	12,223	0.42
		(0.01)		(0.02)		(0.02)		(0.004)
Signal by picture	1,660	0.68	602	0.65	1,010	0.61	12,223	0.71
		(0.01)		(0.02)		(0.02)		(0.004)
Signal by report and/or picture	1,660	0.83	602	0.74	1,010	0.71	12,223	0.83
		(0.01)		(0.02)		(0.01)		(0.003)
Rating(popularity)	1,660	3.39	602	3.37	1,010	3.30	12,223	3.42
		(0.01)		(0.02)		(0.01)		(0.004)
Number of patient reviews	1,660	53.5	602	56.30	1,010	43.95	12,223	58.86
		(2.91)		(5.34)		(4.58)		(1.33)
Chief physician	1,660	0.26	602	0.18	1,010	0.12	12,223	0.25
		(0.01)		(0.02)		(0.01)		(0.004)
Associate chief physician	1,660	0.28	602	0.26	1,010	0.25	12,223	0.33
		(0.01)		(0.02)		(0.01)		(0.004)
Attending physician	1,660	0.33	602	0.40	1,010	0.45	12,223	0.32
		(0.01)		(0.02)		(0.02)		(0.004)
Resident physician	1,660	0.12	602	0.16	1,010	0.18	12,223	0.10
		(0.01)		(0.02)		(0.01)		(0.003)
No professional title reported	1,660	0.002	602	0.002	1,010	0.001	12,223	0.001
		(0.001)		(0.002)		(0.001)		(0.0003)
Professor	1,660	0.14	602	0.11	1,010	0.07	12,223	0.15
		(0.01)		(0.01)		(0.01)		(0.003)
Associate Professor	1,660	0.15	602	0.15	1,010	0.11	12,223	0.20
		(0.01)		(0.01)		(0.01)		(0.004)
Assistant Professor	1,660	0.07	602	0.11	1,010	0.10	12,223	0.11
		(0.01)		(0.01)		(0.01)		(0.003)
Teaching assistant	1,660	0.01	602	0.01	1,010	0.01	12,223	0.01
		(0.002)		(0.004)		(0.003)		(0.001)
No academic title reported	1,660	0.63	602	0.62	1,010	0.72	12,223	0.54
		(0.01)		(0.02)		(0.01)		(0.005)
Returning patient	4,252	0.002	1,557	0.003	2,560	0.0004	31,537	0.001
		(0.0007)		(0.001)		(0.0004)		(0.0001)
Designated call length	4,252	3.33	1,557	3.59	2,560	2.86	31,537	3.33
		(0.08)		(0.14)		(0.10)		(0.03)

Notes: Summary statistics for female- and male-dominated fields, by gender for regression sample. See notes to Table 1a.

Table 2. Price regressions

	All fields	Female-dominated fields (female > 70%)	Male-dominated fields (female < 30%)
Dependent var: price (inverse hyperbolic sine)	(1)	(2)	(3)
Female	**-0.0160****	**-0.0461**	**-0.0056**
	(0.0079)	**(0.0309)**	**(0.0212)**
Female × (Number/100)	**0.0430*****	**0.0586*****	**0.0313*****
	(0.0065)	**(0.0198)**	**(0.0116)**
Rating × (Number/100)	-0.1339***	-0.1457***	-0.1050***
	(0.0064)	(0.0198)	(0.0089)
Number/100	0.6374***	0.6900***	0.5259***
	(0.0299)	(0.0930)	(0.0427)
Rating	0.6599***	0.5695***	0.5780***
	(0.0141)	(0.0536)	(0.0225)
Service characteristics	Yes	Yes	Yes
Doctor characteristics	Yes	Yes	Yes
Hospital and department characteristics	Yes	Yes	Yes
Observations	90,457	5,809	34,097
R-squared	0.5937	0.6763	0.5522
Estimated effects of Female at Number of review quartiles			
Female + Female × p25 (Number/100)	-0.0155	-0.0455	-0.0053
Female + Female × p50 (Number/100)	-0.0129	-0.0391	-0.0022
Female + Female × p75 (Number/100)	0.0004	-0.0156	0.0097
Female + Female × p100(Number/100)	1.4852	0.8671	0.8942

Notes: The table reports estimates of equation (7) in the top panel, and estimated gender difference at quartiles of number of reviews in the bottom panel. The dependent variable is the price of an online consultation service. Since zero price appears in our data and can also be a meaningful data point, we keep observations for doctors with zero prices in the analysis sample and compute the inverse hyperbolic sine, which approximates the natural logarithm. The unit of observation is a doctor-service pair. Estimated coefficients of control variables are not reported. These include: service characteristics; doctor characteristics; and hospital and department characteristics. Variable definitions are listed in Appendix Table A2. Standard errors clustered at the doctor level are reported in parentheses. ** and *** indicate statistical significance at the 5%, and 1% levels, respectively.

Table 3. Gender signaling regressions

	All fields	Female-dominated fields (female > 70%)	Male-dominated fields (female < 30%)	All fields	Female-dominated fields (female > 70%)	Male-dominated fields (female < 30%)
		All names		Gender-neutral names (probability male or female < 80%)		
Dependent variable: strong gender signal (report and/or picture)	(1)	(2)	(3)	(1')	(2')	(3')
Female	**0.0014**	**0.0791*** **	**-0.0700*** **	**-0.0190***	**0.1465*** **	**-0.1597*** **
	(0.0053)	**(0.0215)**	**(0.0140)**	**(0.0097)**	**(0.0487)**	**(0.0231)**
Female × (Number/100)	**0.0029**	**-0.0159***	**0.0127*** **	**0.0068**	**-0.0404** **	**0.0296*** **
	(0.0030)	**(0.0081)**	**(0.0037)**	**(0.0048)**	**(0.0179)**	**(0.0099)**
Rating × (Number/100)	-0.0372***	-0.0459***	-0.0306***	-0.0455***	-0.0554***	-0.0335***
	(0.0024)	(0.0081)	(0.0032)	(0.0055)	(0.0138)	(0.0084)
Number/100	0.1616***	0.2144***	0.1342***	0.1920***	0.2786***	0.1446***
	(0.0110)	(0.0387)	(0.0152)	(0.0228)	(0.0648)	(0.0359)
Rating	0.1448***	0.1672***	0.1208***	0.1654***	0.1918***	0.1327***
	(0.0065)	(0.0258)	(0.0097)	(0.0129)	(0.0508)	(0.0230)
Doctor characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Hospital and department characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Observations	35,777	2,262	13,233	10,397	837	3,277
R-squared	0.2274	0.2390	0.2443	0.2273	0.3278	0.2777

Notes: The table reports linear probability estimates for whether gender signaling by report and/or picture or not. The unit of observation is a doctor. Estimated coefficients of control variables are not reported. These include: doctor characteristics; and hospital and department characteristics. Variable definitions are listed in Appendix Table A2. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

Table 4. Gender signaling regressions, changing sample to gender-neutral fields with 40%-60% of female doctors

	Gender-neutral fields (female = 40%-60%)	Gender-neutral fields (female = 40%-60%)
	All names	Gender-neutral names (probability male or female < 80%)
Dependent variable: strong gender signal (report and/or picture)	(1)	(2)
Female	**0.0159****	**0.0034**
	(0.0069)	**(0.0128)**
Female × (Number/100)	**-0.0043**	**0.0103**
	(0.0052)	**(0.0077)**
Rating × (Number/100)	-0.0445***	-0.0665***
	(0.0044)	(0.0077)
Number/100	0.1944***	0.2755***
	(0.0191)	(0.0335)
Rating	0.1672***	0.1822***
	(0.0106)	(0.0193)
Doctor characteristics	Yes	Yes
Hospital and department characteristics	Yes	Yes
Observations	15,880	4,953
R-squared	0.2319	0.2411

Notes: See notes to Table 3. The only difference is that the sample is changed to gender-neutral fields defined by the share of female doctors between 40% and 60%.

Table 5. Gender signaling regressions, varying definitions of female- and male-dominated fields and definition of gender-neutral names

	Female-dominated fields (female > 73%)	Male-dominated fields (female < 20%)	Female-dominated fields (female > 73%)	Male-dominated fields (female < 20%)	Female-dominated fields (female > 70%)	Male-dominated fields (female < 30%)
	All names		Gender-neutral names (probability male or female < 80%)		Gender-neutral names (probability male or female < 70%)	
Dependent variable: strong gender signal (report and/or picture)	(1)	(2)	(3)	(4)	(5)	(6)
Female	**0.0790*****	**-0.0705*****	**0.1019***	**-0.1640*****	**0.1243***	**-0.2233*****
	(0.0258)	**(0.0142)**	**(0.0611)**	**(0.0235)**	**(0.0662)**	**(0.0298)**
Female × (Number/100)	**-0.0208****	**0.0132*****	**-0.0288**	**0.0310*****	**-0.0325**	**0.0443*****
	(0.0087)	**(0.0037)**	**(0.0202)**	**(0.0101)**	**(0.0262)**	**(0.0141)**
Rating × (Number/100)	-0.0524***	-0.0305***	-0.0584***	-0.0336***	-0.0369**	-0.0476***
	(0.0092)	(0.0032)	(0.0144)	(0.0085)	(0.0173)	(0.0140)
Number/100	0.2500***	0.1334***	0.2939***	0.1446***	0.2033**	0.2065***
	(0.0438)	(0.0152)	(0.0673)	(0.0363)	(0.0892)	(0.0594)
Rating	0.1580***	0.1223***	0.1192**	0.1350***	0.0901	0.1435***
	(0.0298)	(0.0098)	(0.0584)	(0.0232)	(0.0653)	(0.0312)
Doctor characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Hospital and department characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,668	13,059	629	3,225	510	1,907
R-squared	0.2655	0.2412	0.3676	0.2752	0.3854	0.3105

Notes: See notes to Table 3. The only difference is the use of stricter definitions of female-and male-dominated fields and gender-neutral names. We chose a cutoff of 73% female to define female-dominated fields, rather than 80%, because there is only one department (67 doctors, see Table A1) with share of female doctors over 80%. Choosing a cutoff of 73% thus adds the very large Obstetrics and Gynecology field to the set of female-dominated fields.

Table 6. Gender signaling regressions, changing rating measure to average of standardized rating measures

	All fields	Female-dominated fields (female > 70%)	Male-dominated fields (female < 30%)	All fields	Female-dominated fields (female > 70%)	Male-dominated fields (female < 30%)
	All names			Gender-neutral names (probability male or female < 80%)		
Dependent variable: strong gender signal (report and/or picture)	(1)	(2)	(3)	(1')	(2')	(3')
Female	**-0.0001**	**0.0781*** **	**-0.0695*** **	**-0.0202** **	**0.1444*** **	**-0.1606*** **
	(0.0053)	**(0.0216)**	**(0.0140)**	**(0.0097)**	**(0.0491)**	**(0.0233)**
Female × (Number/100)	**0.0025**	**-0.0159** **	**0.0078** **	**0.0060**	**-0.0379* **	**0.0270** **
	(0.0026)	**(0.0081)**	**(0.0039)**	**(0.0043)**	**(0.0202)**	**(0.0122)**
Rating × (Number/100)	-0.0009*	-0.0014	-0.0012**	-0.0011	-0.0005	-0.0008
	(0.0005)	(0.0012)	(0.0005)	(0.0010)	(0.0016)	(0.0014)
Number/100	-0.0062**	0.0120	-0.0029	-0.0094*	0.0360*	-0.0053
	(0.0026)	(0.0094)	(0.0035)	(0.0051)	(0.0209)	(0.0077)
Rating	0.1373***	0.1472***	0.1156***	0.1517***	0.1543***	0.1262***
	(0.0068)	(0.0258)	(0.0106)	(0.0128)	(0.0495)	(0.0233)
Doctor characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Hospital and department characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Observations	35,777	2,262	13,233	10,397	837	3,277
R-squared	0.2244	0.2345	0.2416	0.2237	0.3216	0.2750

Notes: See notes to Table 3. The only difference is that the rating measure is changed to the average of four standardized other possible rating measures. We first standardize four other statistics patients could potentially use to evaluate doctors: the treatment satisfaction rate, the bedside manner satisfaction rate, the number of thank you letters, and the number of gifts. Each variable is standardized to have mean 0 and variance 1 based on the available observations in the analysis sample. We then compute the average over the non-missing standardized rating variables for each doctor.

Table 7. Gender signaling regressions, changing number of reviews measure to number of individual online consultations

	All fields	Female-dominated fields (female > 70%)	Male-dominated fields (female < 30%)	All fields	Female-dominated fields (female > 70%)	Male-dominated fields (female < 30%)
	All names			Gender-neutral names (probability male or female < 80%)		
Dependent variable: strong gender signal (report and/or picture)	(1)	(2)	(3)	(1')	(2')	(3')
Female	**0.0029**	**0.0822*****	**-0.0715*****	**-0.0165***	**0.1451*****	**-0.1610*****
	(0.0053)	**(0.0218)**	**(0.0142)**	**(0.0096)**	**(0.0501)**	**(0.0234)**
Female × (Number/100)	**-0.0001**	**-0.0010****	**0.0012*****	**-0.00002**	**-0.0014**	**0.0028*****
	(0.0002)	**(0.0005)**	**(0.0004)**	**(0.0003)**	**(0.0014)**	**(0.0011)**
Popularity × (Number/100)	-0.0022***	-0.0022***	-0.0021***	-0.0025***	-0.0037***	-0.0018***
	(0.0002)	(0.0004)	(0.0002)	(0.0003)	(0.0010)	(0.0005)
Number/100	0.0091***	0.0101***	0.0085***	0.0104***	0.0164***	0.0072***
	(0.0008)	(0.0019)	(0.0009)	(0.0014)	(0.0042)	(0.0021)
Rating	0.1496***	0.1725***	0.1285***	0.1665***	0.2116***	0.1376***
	(0.0059)	(0.0249)	(0.0086)	(0.0114)	(0.0487)	(0.0194)
Doctor characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Hospital and department characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Observations	35,777	2,262	13,233	10,397	837	3,277
R-squared	0.2274	0.2396	0.2443	0.2269	0.3295	0.2773

Notes: See notes to Table 3. The only difference is that the number of reviews measure is changed to the number of individual online consultations.

Table 8. Gender signaling regressions, sample restricted to strongly-gendered names (probability female or male > 90%)

	All fields	Female-dominated fields (female > 70%)	Male-dominated fields (female < 30%)
	Strongly gendered names (probability female or male > 90%)		
Dependent variable: strong gender signal (report and/or picture)	(1)	(2)	(3)
Female	**-0.0143***	**0.0055**	**-0.0130**
	(0.0083)	**(0.0318)**	**(0.0239)**
Female × (Number/100)	**0.0043**	**0.0002**	**0.0080**
	(0.0054)	**(0.0145)**	**(0.0053)**
Rating × (Number/100)	-0.0333***	-0.0603***	-0.0304***
	(0.0033)	(0.0168)	(0.0043)
Number/100	0.1443***	0.2554***	0.1324***
	(0.0149)	(0.0756)	(0.0204)
Rating	0.1371***	0.1909***	0.1189***
	(0.0092)	(0.0416)	(0.0134)
Doctor characteristics	Yes	Yes	Yes
Hospital and department characteristics	Yes	Yes	Yes
Observations	16,849	910	6,777
R-squared	0.2323	0.3060	0.2503

Notes: The table reports sensitivity analysis for the test of statistical discrimination based on gender signaling regressions. See notes to Table 3. The only difference is that the sample is changed to doctors with highly-gendered names (above 90% probability of accuracy).

Appendix: Additional tables

Table A1. Share of female doctors by department

	Number of doctors	Proportion female
Tuberculosis Medicine	1	0.000
Urology	99	0.010
Orthopedics	2,491	0.027
Interventional Medicine	387	0.083
Surgery	10,003	0.086
Occupational Medicine	9	0.111
Sports Medicine	81	0.173
Burns and Plastic Surgery	174	0.207
Anesthesiology	494	0.312
Otorhinolaryngology(ENT)	1,246	0.339
Dentistry	1,153	0.376
Other Departments	869	0.391
Traditional Chinese Medicine	644	0.398
Pathology	94	0.404
Medical Imaging	668	0.412
Oncology	1,958	0.413
Aesthetic Dermatology	37	0.432
Rehabilitation Medicine	348	0.434
Pediatrics	2,490	0.465
Internal Medicine	7,381	0.483
Infectious Diseases	410	0.498
Integrative Medicine	169	0.503
Psychiatry and Psychology	315	0.524
Dermatology and Venereology	946	0.577
Ophthalmology	1,077	0.579
Reproduction	594	0.702
Obstetrics and Gynecology	1,605	0.740
Nutrition	67	0.881

Notes: Departments are ranked in ascending order of proportion female.

Table A2. Variable definition

Variable	Definition
Price	Price of all online consultations
48-hour chat price	Price of a chat consultation with unlimited number of text questions and answers within 48 hours
One-question chat price	Price of a chat consultation for asking and answering only one set of question
Phone call price	Price of phone call consultation depending on the designated length of call and new vs. returning patients
Female	=1 if a doctor is a female, = 0 if male. It is classified by report first, then picture, then name.
Signal by report	=1 if a doctor signals gender by report in short bio, = 0 if not.
Signal by picture	=1 if a doctor signals gender by picture, = 0 if not.
Signal by report and/or picture	=1 if a doctor signals gender by report in short bio or by picture, = 0 if not.
Rating (popularity)	The comprehensive recommendation popularity ranges between 1 to 5 (low to high). It is a summary measure of patient reviews and also assigns weight to doctors' qualifications and characteristics including hospital level, professional title, highest educational degree, university graduated from, and other factors.
Number of reviews	The number of patient reviews reflects the number of patients who have filled in every item of the online doctor review system and submitted the review for online or offline service.
Professional title	Professional titles are from junior to senior at four levels, i.e., resident, attending, associate chief, and chief, as are the titles for physicians, rehabilitation engineers, technicians, nurses, examiners, and pharmacists. For doctors who do not report professional tiles, we refer them to as "no professional title reported."
Academic title	Academic titles are from junior to senior at four levels, i.e., teaching assistant, assistant professor, associate professor, professor, as are the titles for researchers. For doctors who do not report academic titles, we refer them to as "no academic title reported."
Number of hospitals	How many hospitals a doctor works in.
Position of listed hospital	If a doctor works in multiple hospitals, in which position the doctor lists the hospital in the ranking list. If the doctor also works in multiple hospitals in the ranking list, this indicates the position of the first listed hospital.
Retuning patient	= 1 if a patient is not new to a doctor (sees the doctor for the second time or more times), = 0 if not. It can only be equal to 1 if a price is for phone call consultation and it is not for new patients, and is always equal to zero for prices of chat consultations or for phone call consultation price for new patients.
Designated call length	Designated period length of call (in minutes) in phone call consultation

Table A3a. Summary statistics for alternative rating and review measures for the full sample and by gender

Sample	Full		Female doctors		Male doctors	
Variable	Obs.	Mean (Std. Err.)	Obs.	Mean (Std. Err.)	Obs.	Mean (Std. Err.)
Treatment satisfaction rate	11,341	99.14	3,381	98.98	7,960	99.20
		(0.03)		(0.06)		(0.03)
Bedside manner satisfaction rate	11,341	99.32	3,381	99.19	7,960	99.37
		(0.03)		(0.06)		(0.03)
Number of thank you letters	35,777	23.51	11,899	16.42	23,878	27.04
		(0.36)		(0.45)		(0.49)
Number of gifts	35,777	46.83	11,899	36.08	23,878	52.19
		(0.97)		(1.49)		(1.25)
Online service satisfaction rate	6,443	98.57	2,189	98.55	4,254	98.58
		(0.05)		(0.08)		(0.06)
Number of online consultations	35,777	667.17	11,899	544.45	23,878	728.33
		(10.54)		(14.51)		(14.02)

Table A3b. Summary statistics for alternative rating and review measures for gender fields and by gender

	Female-dominated fields (female > 70%), female doctors		Female-dominated fields (female > 70%), male doctors		Male-dominated fields (female < 30%), female doctors		Male-dominated fields (female < 30%), male doctors	
Variable	Obs.	Mean (Std. Err.)	Obs.	Mean (Std. Err.)	Obs.	Mean (Std. Err.)	Obs.	Mean (Std. Err.)
Treatment satisfaction rate	592	98.84	200	99.06	331	98.97	4,550	99.30
		(0.17)		(0.24)		(0.19)		(0.04)
Bedside manner satisfaction rate	592	98.83	200	99.26	331	99.30	4,550	99.41
		(0.18)		(0.22)		(0.14)		(0.04)
Number of thank you letters	1,660	23.82	602	26.03	1,010	21.18	12,223	30.02
		(1.44)		(2.89)		(2.22)		(0.71)
Number of gifts	1,660	73.52	602	72.24	1,010	31.55	12,223	48.86
		(7.38)		(8.11)		(3.58)		(1.62)
Online service satisfaction rate	482	98.53	155	98.01	145	98.18	2,076	98.54
		(0.19)		(0.34)		(0.37)		(0.09)
Number of online consultations	1,660	1,003.53	602	1,237.38	1,010	513.89	12,223	644.81
		(52.49)		(112.82)		(48.44)		(16.78)