A Study on Gender Bias in Doctor Selection

Summer 2025 - 241 Final Report
Beatrice Filart, Hannah MacDonald, Jonathan Yi

Abstract

This report details a within-subjects experiment designed to investigate the causal effect of a doctor's gender on patient selection when other factors, such as qualifications and experience, are held constant. Participants were shown profiles of doctors of varying genders and were asked to make selections for different medical specialties. The study's primary objective was to determine if a patient's choice of a provider is influenced by the doctor's gender. The secondary objective was to explore if this preference is driven by a gender concordance effect where patients prefer doctors of their own gender. The last objective was to identify if medical specialty influenced which doctor a patient chose. Our power analysis revealed that with an assumed effect size of 0.1, we will need a sample size of at least 250 participants in order to achieve 75% power. Our initial analysis indicates a significant preference for female doctors overall, while also affirming that the gender concordance effect exists.

Introduction

The patient-doctor relationship is a critical component of healthcare. While medical expertise and qualifications are important, other factors may influence a patient's decision making process. Our experiment was designed to answer the following research question: *Does the perceived gender of a doctor influence which provider a patient chooses when presented with multiple, equally qualified options?* While prior research suggests gender bias can shape professional evaluations in medical and other fields, much of this evidence relies on observational data or self-reported attitudes, which can be subject to confounding variables or social desirability bias. This experiment was designed to isolate the causal effect of perceived gender by observing choice behavior in a controlled, simulated environment.

The theoretical justification for this experiment is grounded in the literature on implicit bias. Societal stereotypes often associate professions and traits with specific genders, which can unconsciously influence decision-making. For instance, studies show persistent gender stereotyping in professional evaluations (Carnes et al., 2012; Moss-Racusin et al., 2012). More specific to healthcare, recent evidence from telehealth booking platforms shows that male patients disproportionately select male doctors, suggesting small but consistent biases in choice behavior (Feng et al., 2023). Our experiment builds on this work by creating a scenario where professional qualifications are explicitly balanced, allowing us to test if gender cues alone are sufficient to create a measurable difference in preference.

Based on this theoretical framework, we test the following hypotheses:

1. Primary hypothesis (H1): A doctor's gender alone will influence a patient's choice. We suspect this influence to result in a general preference for male doctors.
2. Secondary hypothesis (H2): This explores this potential bias further by considering the Gender Concordance effect, where we believe people are more likely to select doctors of their own gender.
3. Tertiary hypothesis (H3): This adds an additional factor where we believe that the impact of a doctor's gender on a patient's choice will vary depending on the medical specialty.
    a. For example, a patient's preference may differ in specialties that are more technical such as Cardiology as opposed to those perceived as interpersonal, like Primary Care.

Experimental Design

Study Design and Randomization

To investigate our research question, we employed a within-subjects experimental design. We chose this approach to maximize statistical power and control for baseline, participant-level preferences. In the experiment, participants were randomly assigned to a sequence of three different medical specialties drawn from a pool of 12 *(Cardiology, Dermatology, Psychiatry, Primary Care Ophthalmology, Neurology, OBGYN (Obstetrics & Gynecology), Urology, Orthopedics , Gastroenterology, Pediatrics, and Oncology)*. For each specialty, they were presented with six doctor profiles: three male-presenting and three female-presenting. To eliminate potential order effects, the on-screen position of these six profiles was also randomized for each participant in each round.

Key to our design was the balancing of doctor attributes. All doctor profiles were standardized on credentials, years of experience, and biographical content, with only name and an emoticon as gender cues being varied. We conducted a balance check on the randomized attributes, which confirmed that there was no statistically significant difference in the average years of experience between the male and female profiles presented to participants ($p > 0.05$).

Treatment and Outcome Measures

The treatment in this experiment was the perceived gender of the doctor, which was signaled implicitly through gendered names and emoticons. The primary outcome measure was a binary variable indicating if the selected doctor was male or female (*is_chosen*). This allowed us to directly measure revealed preference and test our hypotheses about gender's influence on choice. In addition to this primary outcome, we also collected post-treatment data to explore potential mechanisms behind the choices. These measures included the time a participant spent making a selection for each round and their explicit ratings of the doctors' perceived competence and trustworthiness in the final round.

Measuring Tool

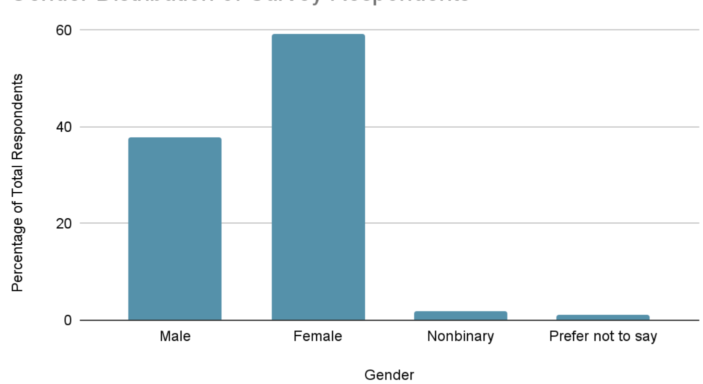
We implemented our experiment using a custom-built website designed to serve as our primary measuring tool. This platform emulated a more realistic doctor selection process, where participants were presented with a series of profiles and asked to make a choice for a given medical specialty rather than participating in a survey setting. By designing our own website, we achieved a high degree of experimental control. This allowed us to precisely manipulate the treatment, the perceived gender of the doctor, while balancing all other attributes, such as qualifications, years of experience, and biographical details. This control was critical for isolating the causal effect of gender on patient preference. Furthermore, the interactive and familiar format of the website was intended to increase the validity of the task, encouraging participants to make choices in a way that more closely mirrors a real-world online selection environment.

Study Population

Our study population consisted of adults recruited through convenience sampling via online and personal networks. After data cleaning and analysis, a total of 164 participants completed all three rounds of the survey. The demographic composition of our sample is summarized in the figures within this section and tables listed below.

Figure 1

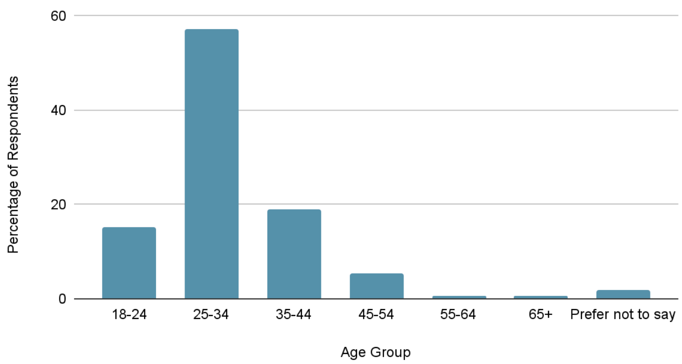
Gender Distribution of Survey Respondents



We can see from Figure 1, the sample was predominantly female, with women comprising 63.4% (n=104) of participants, while men comprised 34.1% (n=56). A small number of participants identified as non-binary (1.2%) or preferred not to state their gender (1.2%). Figure 2 showed the age distribution was heavily skewed towards younger adults; a significant majority of participants (72.0%, n=118) fell within the 25-34 age group. The next largest group was the 35-44 age bracket, accounting for 11.0% of the sample.

Figure 2

Age Distribution of Survey Respondents



It is important to note the limitations regarding the external validity of our findings due to this sample composition. The noticeable gender and age imbalances mean our sample is not representative of the general patient population. Consequently, our findings are most generalizable to women in the 25-34 age group, and caution should be exercised when extending these conclusions to other demographic groups.

Gender Distribution Table (N=164)

Gender	Percentage
Female	59.1%
Male	37.8%
Non-binary	1.8%
Prefer not to say	1.2%

Age Distribution Table (N=164)

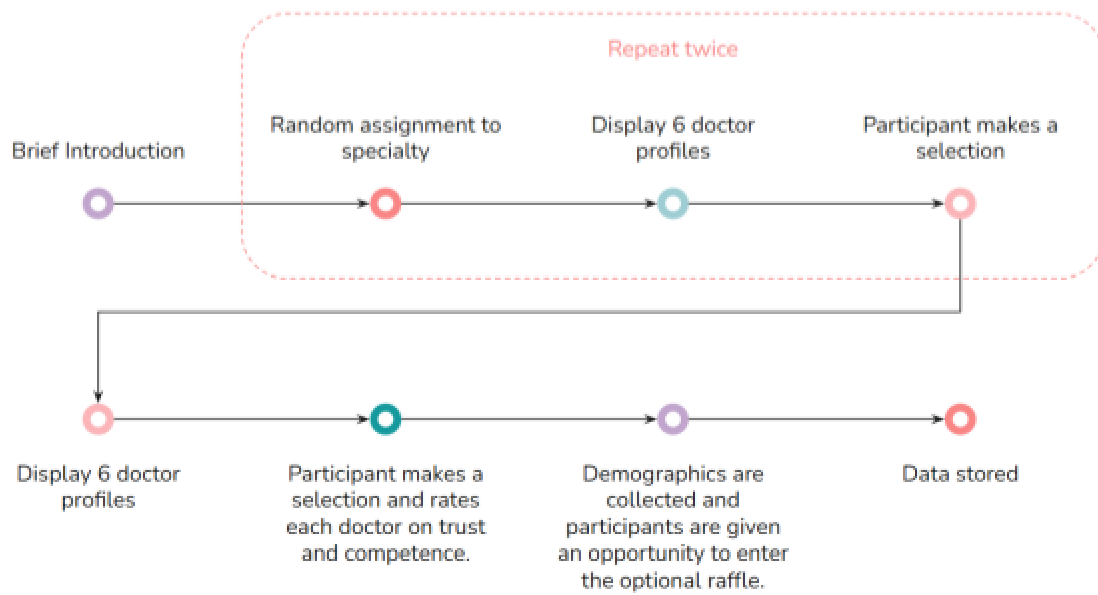
Age Group	Percentage
18-24	15.2%
25-34	57.3%

35-44	18.9%
45-54	5.5%
55-64	0.6%
65+	0.6%

Experiment Flow

The experiment was structured to guide participants through a seamless, multi-stage process. Upon entering the study, participants were presented with a brief introduction outlining the task. The experiment then proceeded in three distinct rounds. In each round, a medical specialty was randomly assigned to the participant, and they were shown six doctor profiles from which they were asked to select one. After completing all three selection rounds, participants entered a final phase where they were asked to rate all six doctors from the last round on perceived competence and trustworthiness. Following this rating task, demographic information was collected, and participants were given an optional opportunity to enter a raffle as a thank you for their time. The illustration below provides a visual summary of this progression (Figure 3).

Figure 3



Power Calculation

To determine the necessary sample size for our experiment, we conducted a power analysis. We assumed a modest but realistic effect size of 0.1, which translates to participants choosing a male doctor 55% of the time instead of the 50% expected by chance under the null hypothesis. This assumption was based on prior research showing persistent, subtle gender biases in professional evaluations.

We simulated power for three different study designs across a range of sample sizes (50-400 participants).

- Basic Design: A simple design where each participant is assigned to a single specialty.
- Repeated-Specialty Design: A within-subjects design where each participant makes selections in three different specialties.
- Added Features Design: This builds on the repeated-specialty design by including covariates such as perceived competence and trust to improve the precision of the model.

Our simulations showed that the Basic Design was significantly underpowered, achieving only about 30-35% power with 200 participants. In contrast, both the Repeated-Specialty Design and the Added Features Design achieved approximately 70% power with 200 participants. To reach our target of at least 75% power, the simulations indicated that a minimum of 250 participants would be required for the more powerful designs.

Analysis

Data

The dataset used for this analysis consists of 2,952 observations from 164 participants who completed all three rounds of the experiment. Each observation corresponds to one of the six doctor profiles presented in a given round.

- Outcome Variable: The primary outcome we measured is *is_chosen*, a binary variable indicating whether a doctor profile was selected (TRUE) or not (FALSE).
- Treatment Variable: The treatment, as described in the Experimental Details, is the *doctor_gender*. In our models, this is an operationalized factor with two levels: "male" (the reference category) and "female".

- Covariate Features: To improve the precision of our estimates, we included important pre-treatment covariates: *participant_gender*, *participant_age*, and *participant_education*.

Models

Our modeling strategy is progressive, following some of the scenarios outlined in our power analysis. We begin with a simple model to estimate the treatment-control contrast and then build upon it to correctly model the data structure and improve precision. As our outcome is binary, we use logistic regression models.

<u>Model 1 & 2: Main Effect of Doctor Gender</u>

1. simple_model <- glm(is_chosen ~ doctor_gender)
2. mixed_model_covariates <- glmer(is_chosen ~ doctor_gender + participant_gender + participant_age + participant_education + (1 | participant_id))

Our initial analysis focused on the primary hypothesis: whether a doctor's gender influences patient choice. For our model 1, we first ran a simple logistic regression model looking only at the effect of *doctor_gender*. The result showed a statistically significant preference for female doctors ($p < 0.001$).

To properly account for our repeated-specialty (within-subjects) design, our model 2 is a mixed-effects logistic regression. This model includes a random intercept for each participant to control for their baseline tendencies, ensuring we are comparing how the same person responds to different doctor genders. After adding participant demographic covariates to increase precision, the main finding remained robust. As shown in our stargazer table (Table 1), the effect of a doctor's gender was highly statistically significant ($p < 0.001$). After converting the log-odds coefficient (0.413), we found that female doctors had 51% higher odds of being selected compared to male doctors, confirming our primary hypothesis.

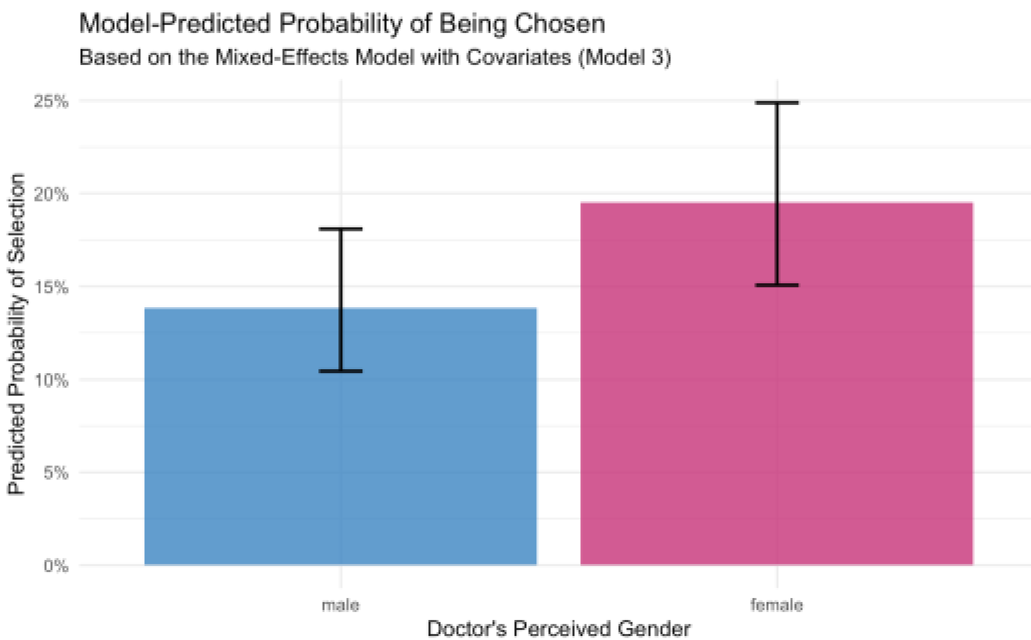
Table 1

Key Results: Mixed-Effects Logistic Regression Models

	Dependent variable:	
	logistic	generalized linear mixed-effects
	ATE Model	With Covariates
	(1)	(2)
Doctor Gender (Female)	0.413 (0.100) p = 0.00004	0.413 (0.100) p = 0.00004
Participant Gender (Male)		0.000 (0.106) p = 1.000
Observations	2,952	2,952
Log Likelihood	-1,321.416	-1,321.416
Note:		*p<0.1; **p<0.05; ***p<0.01

To visualize this main effect, we plotted the model-predicted probabilities in Figure 4. The plot clearly shows that the predicted probability of selection is significantly higher for female doctors than for male doctors, providing an illustration of the main finding from our second model.

Figure 4



<u>Model 3: Heterogeneous Effects by Participant Gender (Gender Concordance)</u>

3. model_gender_interaction <- glmer(is_chosen ~ doctor_gender * participant_gender + participant_age + participant_education + (1 | participant_id))

While the main effect was clear, we hypothesized that this was driven by a gender concordance effect. To test this, we introduced an interaction term between *doctor_gender* and *participant_gender* in our mixed-effects model. The results, shown in Table 2, confirmed a large and highly significant interaction effect ($p < 0.001$). The model shows that for our baseline group of female participants, there was a very strong preference for female doctors (log-odds of 0.975). In contrast, the significant negative interaction term for male participants (-1.286) indicates a complete reversal of

Table 2

Fixed effects:

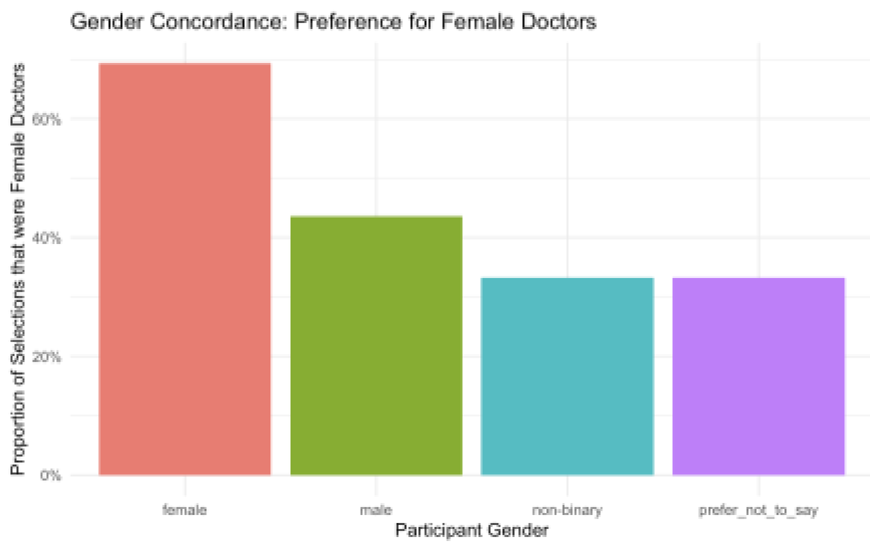
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.176e+00	1.793e-01	-12.132	< 2e-16	***
doctor_genderfemale	9.753e-01	1.377e-01	7.084	1.40e-12	***
participant_gendermale	7.138e-01	1.576e-01	4.529	5.94e-06	***
participant_gendernon-binary	9.230e-01	4.812e-01	1.918	0.0551	.
participant_genderprefer_not_to_say	9.230e-01	9.348e-01	0.987	0.3234	
participant_age25-34	-1.784e-13	1.553e-01	0.000	1.0000	
participant_age35-44	-1.560e-13	1.830e-01	0.000	1.0000	
participant_age45-54	-1.624e-13	2.558e-01	0.000	1.0000	
participant_age55-64	-1.987e-13	6.606e-01	0.000	1.0000	
participant_age65+	-1.652e-13	7.458e-01	0.000	1.0000	
participant_ageprefer_not_to_say	-1.638e-13	6.650e-01	0.000	1.0000	
participant_educationhigh_school	-1.289e-13	2.412e-01	0.000	1.0000	
participant_educationmasters	-3.647e-14	1.147e-01	0.000	1.0000	
participant_educationother	-3.377e-14	3.578e-01	0.000	1.0000	
participant_educationphd_doctorate	-4.847e-14	2.387e-01	0.000	1.0000	
participant_educationprefer_not_to_say	-1.219e-13	5.452e-01	0.000	1.0000	
doctor_genderfemale:participant_gendermale	-1.286e+00	2.124e-01	-6.056	1.40e-09	***
doctor_genderfemale:participant_gendernon-binary	-1.802e+00	7.800e-01	-2.310	0.0209	*
doctor_genderfemale:participant_genderprefer_not_to_say	-1.802e+00	9.502e-01	-1.896	0.0579	.

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

this preference, resulting in male participants being significantly more likely to choose male doctors.

To make this interaction intuitive, we visualized the model's predicted probabilities in Figure 5. The plot clearly shows the two groups' opposing preferences: the line for female participants shows a high probability of selecting a female doctor, while the line for male participants shows a high probability of selecting a male doctor. This visualization provides strong support for our second hypothesis and explains that the overall preference for female doctors in our sample was driven by the larger number of female participants exhibiting this concordance effect.

Figure 5



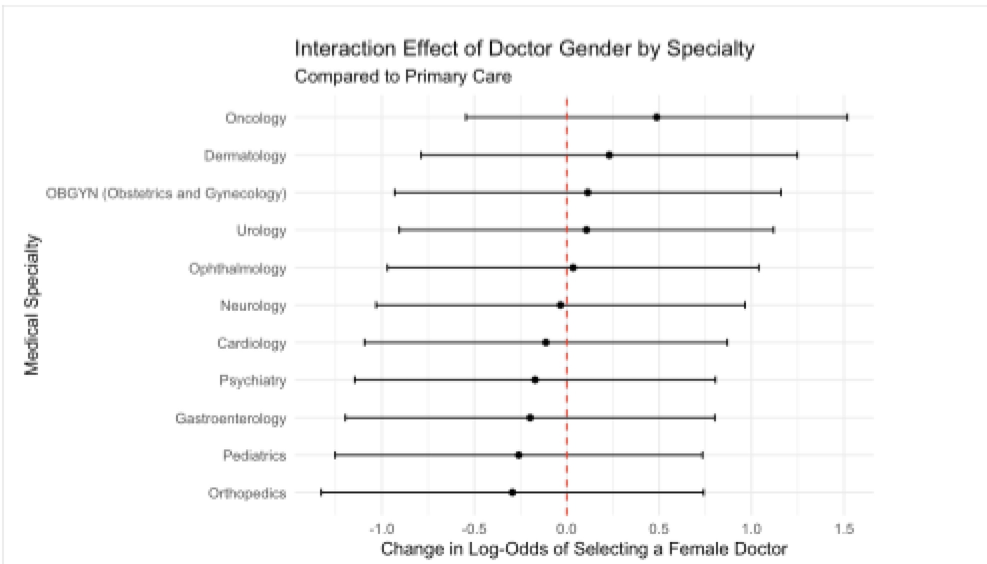
Model 4: Heterogeneous Effects by Medical Specialty

4. model_specialty_interaction <- glmer(is_chosen ~ doctor_gender * specialty + participant_gender + participant_age + participant_education + (1 | participant_id))

To test our third hypothesis, we explored whether the effect of a doctor's gender varied by medical specialty. Figure 6 displays the raw proportion of female doctors selected within each specialty. Visually, there appears to be some variation; for instance, the selection rate for female doctors in Oncology (over 70%) appears much higher than in Cardiology (around 50%).

However, to determine if these observed differences are statistically significant or likely due to random chance, we ran a mixed-effects model that included an interaction term between *doctor_gender* and *specialty*. Contrary to the visual suggestion of the raw data, the model's results showed that none of the interaction terms were statistically significant. This means that after accounting for the overall preference and the variance between participants, we do not have sufficient evidence to conclude that the gender effect is truly different in any specific specialty compared to our baseline of Primary Care. Therefore, while the data hints at potential differences, our formal statistical test does not provide support for our third hypothesis. The effect of doctor gender appears to be relatively consistent across the different medical fields in our study.

Figure 6



<u>Post-Treatment Analysis: Competence and Trust</u>

Finally, to explore the mechanism behind these choices, we analyzed the post-treatment outcomes of perceived competence and trust. We ran two linear mixed-effects models and found no statistically significant difference in how participants rated male versus female doctors on either measure. The average ratings were nearly identical for both genders. This disconnect between biased choices and unbiased explicit ratings suggests the selection preference is likely driven by an implicit bias rather than a conscious belief that one gender is more competent or trustworthy.

Conclusion

Our experiment supports the hypothesis that a doctor's perceived gender influences patient choice, even when qualifications and experience are held constant. Contrary to our initial expectation of a general preference for male doctors, participants overall showed a significant preference for female doctors. This effect was strongly shaped by gender concordance: female participants tended to select female doctors, while male participants preferred male doctors. Specialty-level variation was limited, with the strongest preference for female doctors observed in Oncology.

Interestingly, participants' explicit ratings of competence and trust did not differ by doctor gender, suggesting that these preferences may be driven by implicit bias rather than conscious evaluation.

While the within-subjects design strengthened internal validity, limitations such as a relatively small sample size, reliance on convenience sampling, demographic imbalance toward younger female participants, and the artificial nature of the simulated online setting restrict the external generalizability of our findings. Future research should aim to recruit a larger and more diverse sample, incorporate real-world behavioral data, and explore whether similar patterns emerge in live clinical settings.

Overall, these findings contribute to the growing body of evidence on gender bias in healthcare, underscoring the importance of awareness and mitigation strategies to ensure equitable patient-provider matching.

Reference

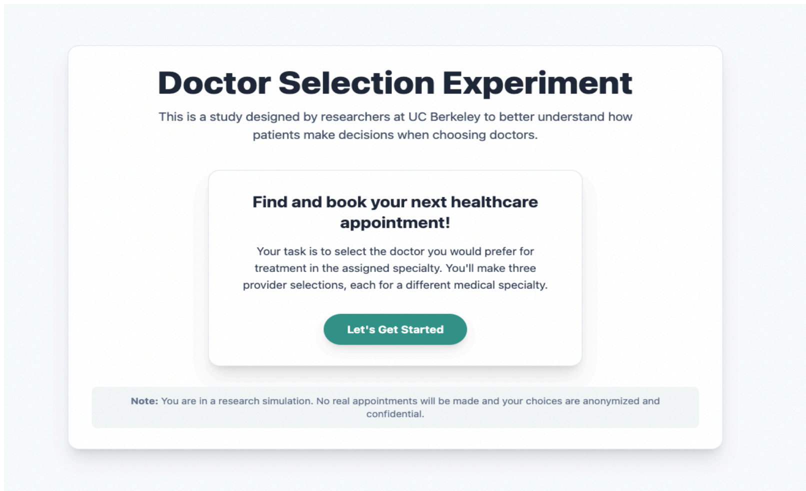
Carnes, M., Devine, P. G., Isaac, C., Baier Manwell, L., Ford, C. E., Byars-Winston, A., Fine, E., & Sheridan, J. (2012). Promoting institutional change through bias literacy. *Journal of Diversity in Higher Education*, *5*(2), 63–77. https://doi.org/10.1037/a0028128

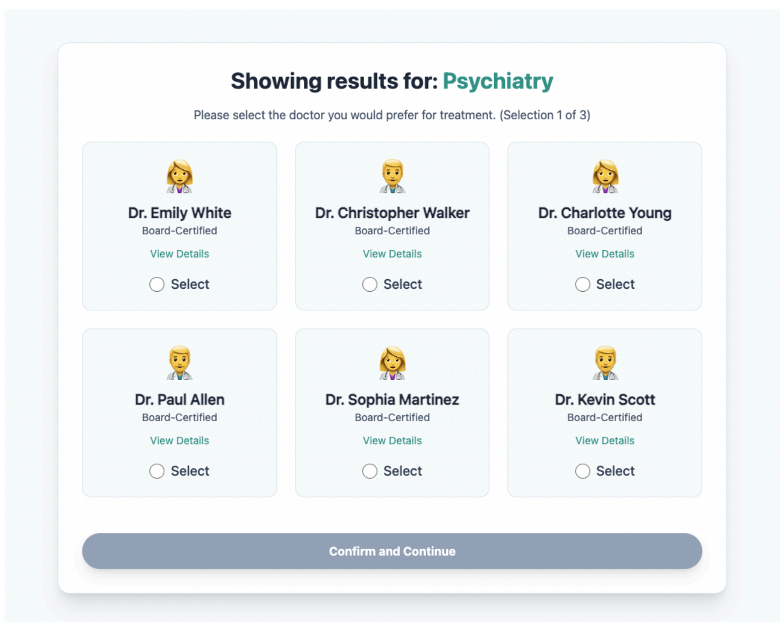
Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, *109*(41), 16474–16479. https://doi.org/10.1073/pnas.1211286109

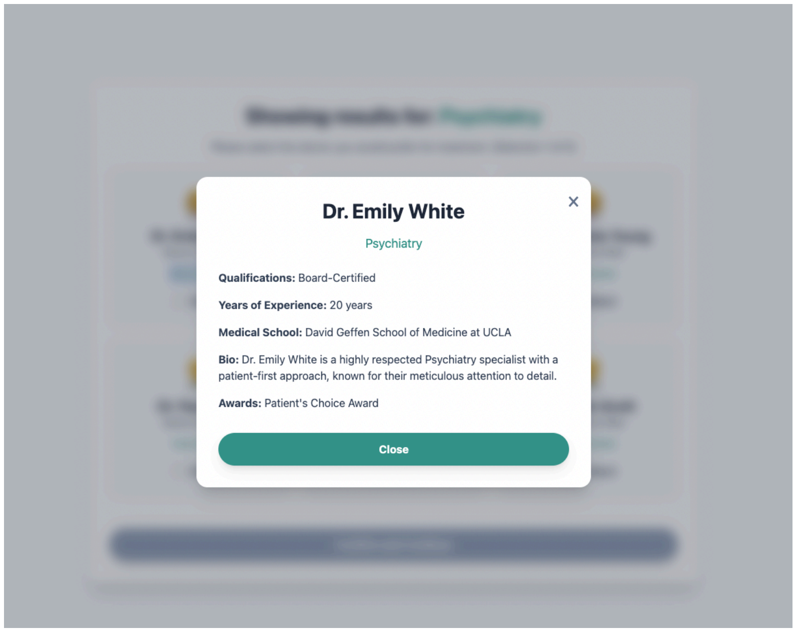
Feng, J., et al. (2023). *Doctor selection behavior on digital health platforms*. [Unpublished manuscript].

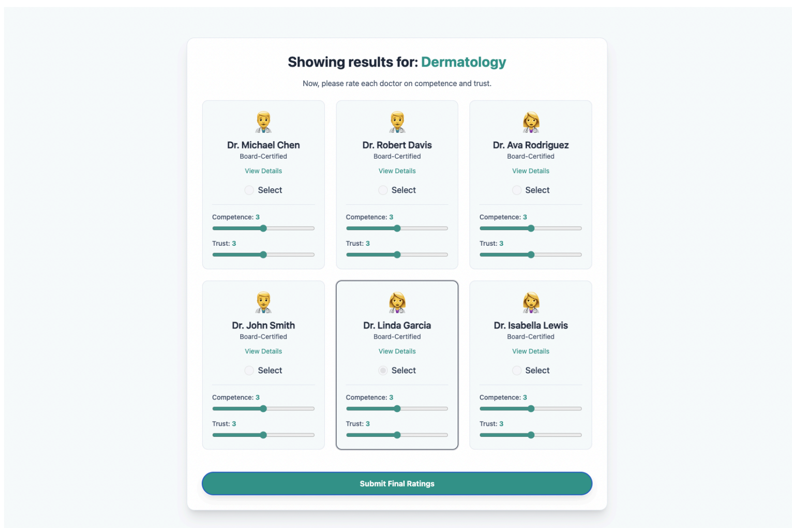
Appendix

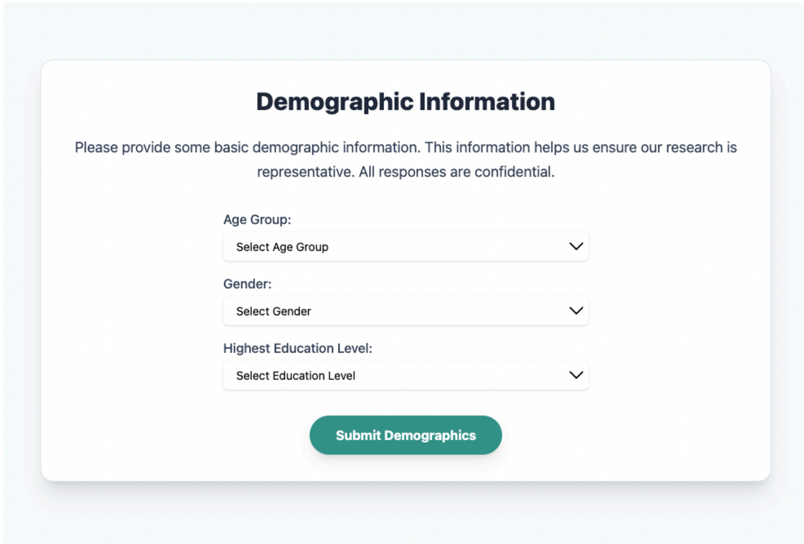
Website

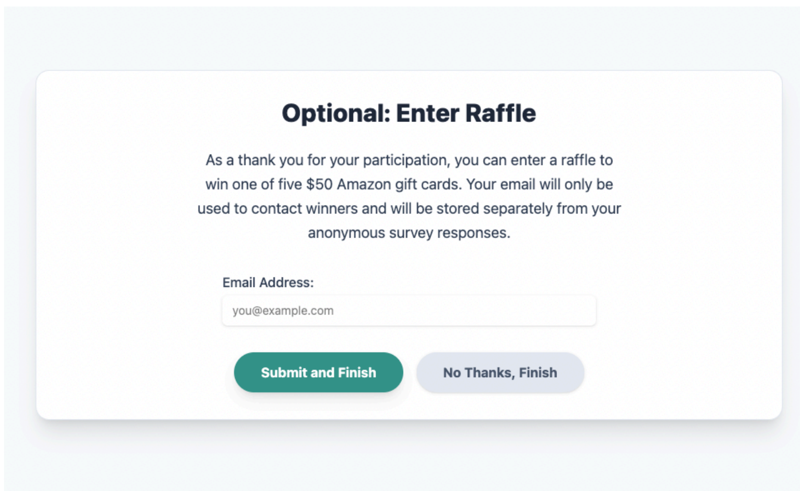





Dr. Emily White
Psychiatry
Qualifications: Board-Certified
Years of Experience: 20 years
Medical School: David Geffen School of Medicine at UCLA
Bio: Dr. Emily White is a highly respected Psychiatry specialist with a patient-first approach, known for their meticulous attention to detail.
Awards: Patient's Choice Award
Close


Showing results for: Dermatology
Now, please rate each doctor on competence and trust.
Dr. Michael Chen
Board-Certified
View Details
Select
Competence: 3
Trust: 3
Dr. Robert Davis
Board-Certified
View Details
Select
Competence: 3
Trust: 3
Dr. Ava Rodriguez
Board-Certified
View Details
Select
Competence: 3
Trust: 3
Dr. John Smith
Board-Certified
View Details
Select
Competence: 3
Trust: 3
Dr. Linda Garcia
Board-Certified
View Details
Select
Competence: 3
Trust: 3
Dr. Isabella Lewis
Board-Certified
View Details
Select
Competence: 3
Trust: 3
Submit Final Ratings


Demographic Information
Please provide some basic demographic information. This information helps us ensure our research is representative. All responses are confidential.
Age Group:
Select Age Group
Gender:
Select Gender
Highest Education Level:
Select Education Level
Submit Demographics


Optional: Enter Raffle
As a thank you for your participation, you can enter a raffle to win one of five $50 Amazon gift cards. Your email will only be used to contact winners and will be stored separately from your anonymous survey responses.
Email Address:
you@example.com
Submit and Finish
No Thanks, Finish