JAMA Network Open

Original Investigation | Health Policy

Effect of Physician Gender and Race on Simulated Patients' Ratings and Confidence in Their Physicians
A Randomized Trial

Rachel E. Solnick, MD, MSc; Kyle Peyton, MA, MPhil, PhD; Gordon Kraft-Todd, PhD; Basmah Safdar, MD, MSc

Abstract

IMPORTANCE Women and black physicians encounter workplace challenges because of their gender and race. It is unclear whether these individuals are assessed with lower patient satisfaction or confidence ratings compared with white male physicians.

OBJECTIVE To examine whether physician gender and race affect participant ratings in scenarios in which physician competence is challenged.

DESIGN, SETTING, AND PARTICIPANTS This randomized trial enrolled a geographically diverse sample of 3592 online respondents in the United States who were recruited from 2 crowdsourcing platforms: Amazon Mechanical Turk (n = 1741) and Lucid (n = 1851). A 2 × 2 factorial design for the gender and race of simulated physicians was conducted between March 9 and July 25, 2018. Participants were excluded before intervention if they were younger than 18 years, were pregnant, or had a history of cancer or abdominal surgical procedures.

INTERVENTIONS A clinical vignette was presented to the participant with a picture of the emergency department physician. Participants were randomly assigned to physicians with different gender and race, with 823 assigned to black women, 791 to black men, 828 to white women, and 835 to white men. A contradictory diagnosis from an online symptom checker introduced doubt about the clinical diagnosis.

MAIN OUTCOMES AND MEASURES A composite outcome (range, 0-100, with 0 representing low patient confidence and satisfaction and 100 representing the maximum on the composite scale) measured participant (1) confidence in the physician, (2) satisfaction with care, (3) likelihood to recommend the physician, (4) trust in the physician's diagnosis, and (5) likelihood to request additional tests.

RESULTS Among 3277 adult participants, complete data were available for 3215 (median age, 49 years [range, 18-89 years]; 1667 [52%] female; 2433 [76%] white). No significant differences were observed in participant satisfaction and physician confidence for the white male physician control physicians (mean composite score, 66.13 [95% CI, 64.76-67.51]) compared with white female (mean composite score, 66.50 [95% CI, 65.19-67.82]), black female (mean composite score, 67.36 [95% CI, 66.03-68.69]), and black male (mean composite score, 66.96 [95% CI, 65.55-68.36]) physicians. Machine learning with bayesian additive regression trees revealed no evidence of treatment effect heterogeneity as a function of participants' race, gender, racial prejudice, or sexism.

(continued)

Key Points

Question In a simulated clinical encounter, do participants evaluate physicians differently based on the physician's gender or race?

Findings In this randomized trial of 3592 online respondents, simulated physician gender and race did not significantly affect participant satisfaction or confidence in physician clinical judgment compared with a white male physician control.

Meaning Participants reported equal satisfaction and confidence in the simulated physicians' diagnosis and treatment plans regardless of the physician's gender or race.

+ Supplemental content

Author affiliations and article information are listed at the end of this article.

Abstract (continued)

CONCLUSIONS AND RELEVANCE No significant differences were observed for simulated patients' evaluations of female or black physicians, suggesting that bias in favor of white male physicians is negligible in survey-based measures of patient satisfaction.

TRIAL REGISTRATION ClinicalTrials.gov Identifier: NCT04190901

JAMA Network Open. 2020;3(2):e1920511. doi:10.1001/jamanetworkopen.2019.20511

Introduction

Women and minority group physicians have steadily become a larger proportion of the health care workforce during the past few decades.[1,2] However, the same groups report experiencing workplace bias from their institutions, superiors, and colleagues in the form of unfair treatment, leading to unequal compensation and career advancement.[3-7] They also report discrimination from patients; minority group physicians experience repeated microaggressions and sometimes glaring instances of racism, with patients refusing care, whereas female physicians have reported cases of gender harassment from patients.[8-12] Such treatment devalues underrepresented physician groups and negatively influences their career trajectories, professional attainment, and retention in medicine.[13-15]

Whether physicians' experience of discrimination from patients represents occasional but offensive anecdotes or signals broader systemic bias that could influence ratings of physicians remains an open question. Evidence from patient-based evaluations is mixed; studies[16-22] have variably demonstrated that patients favor male physicians, favor female physicians, or are indifferent. A meta-analysis of 45 studies,[16] mostly from the primary care setting, that pooled evaluations from more than 100 000 patients for more than 4000 physicians (one-third female) found negligible differences in patient preferences for physician gender. Although some studies[23,24] have examined the benefits of racial concordance in patient-physician relationships, there is little evidence about the causal effect of a physician's race or gender on patient-based evaluations.[23,24] It is therefore unclear whether the discrepancy in patients' preferences for race-concordant physicians is caused by differences in communication styles or choice of outpatient practice setting where patients have an opportunity to exercise their preferences.

There are also limited data on how patient preferences might influence their evaluations of physicians in emergency departments (EDs), where patients have little choice for physician gender or race.[17,22] Similarly, communication styles in real-life encounters and extant observational studies make it difficult to isolate the specific causal effect of physician gender and race on patient satisfaction.[20,25-28] Herein, we report on 2 randomized experiments that directly examined whether physicians are evaluated differently because of their gender and race using a clinical vignette in which a simulated physician's competence was challenged by an online symptom checker during an ED visit.

Methods

Study Design and Setting

This randomized trial was deemed to be exempt from review by the institutional review board at Yale University, New Haven, Connecticut, because the data were deidentified; written informed consent was obtained from participants before participation. The trial protocol is given in Supplement 1. This trial followed the Consolidated Standards of Reporting Trials (CONSORT) reporting guideline.[29]

We conducted 2 randomized experiments using online ED-based clinical vignettes that independently manipulated physician characteristics in a 2 × 2 factorial design between March 9 and

July 25, 2018. An important advantage of an ED-based design is that, unlike in primary care settings, patients in the ED do not have an opportunity to exercise their preference for a particular physician. Previous work has shown that, when given the option, patients are more likely to select a race-concordant physician and satisfaction is higher among patients who had a physician of the same race.[23] Thus, the physician assignment in the ED might expose biases that would have been filtered out by a patient's selection of their physician in other settings. Furthermore, a visit to the ED is a higher stress environment than an office setting, and studies have shown that individuals are more likely to make decisions based on racial stereotypes when experiencing a higher cognitive load.[30]

Participants

We recruited a sample of individuals in the United States (aged ≥18 years) from Amazon Mechanical Turk (MTurk) in March 2018, in which we oversampled older participants (median age, 50 years; range, 19-89 years) using MTurk features to better approximate age groups in the ED.[31] Although social experiments using MTurk have found similar treatment effects and higher data quality compared with nationally representative samples, the MTurk population tends to be younger, more liberal, and more educated than national samples.[32-35] We therefore conducted a direct replication of the first experiment on a more representative group of participants that were quota sampled to match US Census demographics using Lucid, with a participant median age of 45 years (range, 18-86 years) (study 2) in July 2018.[36]

The clinical vignette involved a diagnosis of gastroenteritis based on symptoms as evaluated by an emergency medicine physician. We excluded participants who reported pregnancy, a current or previous diagnosis of cancer, or a history of abdominal surgery. These conditions predispose patients to alternate high-risk diagnoses, which would have made the benign workup given in the vignette unrealistic to a real-world ED evaluation and potentially less credible to participants.[37] We paid all MTurk participants $1.00 in compensation. Participants in Lucid were paid directly by the vendor either in US dollars or through a points program at a similar rate. Each study took approximately 10 minutes to complete.

Study Procedures

Participants used their personal computers to access the study administered using the Qualtrics software platform (Qualtrics). Participants gave consent while blinded to the study objectives, and they self-reported data on demographic characteristics, health insurance, trust in physicians, self-assessed health, and frequency of ED visits in a short background survey administered before the clinical vignette. Race/ethnicity was self-reported from options based on the National Institutes of Health reporting guidelines.[38]

After completing the background survey, participants were asked to play the role of a patient reporting to the ED with symptoms consistent with gastroenteritis (eFigure 1 in Supplement 2). Accurate comprehension was assessed using recall of case details (eFigure 2 and eFigure 3 in Supplement 2).[39,40] If participants did not correctly identify case details, they were shown their symptoms a second time to enhance attentiveness and comprehension. Participants were then randomly assigned to 1 of 4 possible treatment arms that would determine the gender and race of the putative ED physician (**Figure 1**). Participants were then presented with the simulated physician's image and a written diagnosis of gastroenteritis with a conservative treatment plan, alongside a contradictory diagnosis of possible appendicitis from an Online Doc Symptom Checker (a fabrication created for the purposes of this survey experiment) with a more aggressive treatment plan (**Figure 2**). Participants then evaluated the putative physician on the primary and secondary outcome measures specified in the preanalysis plans (eMethods 4 and eMethods 7 in Supplement 2). To avoid priming participants to the goals of the study, validated measures of sexism[42,43] and racial prejudice[44,45] were asked at the end of the study after primary and secondary outcomes were measured (eMethods 5 and eFigures 12-17 in Supplement 2).

Physician Image Selection and Randomization

We created a stimulus set of physicians using images from the Chicago Face Database,[46] a research database for photographs of real human faces of varying gender/ethnicity that have been prerated by independent judges on a variety of dimensions (eg, attractiveness). Stimulus images were selected from the Chicago Face Database to minimize differences in observable traits that might influence ratings of confidence and satisfaction using the following constraints: (1) age between 27 and 39 years (younger physicians are more likely to experience discrimination),[47] (2) accurate perception of race and gender with at least 90% agreement across prerated judges, and (3) displaying neutral levels of trustworthiness and attractiveness according to prerated judges.

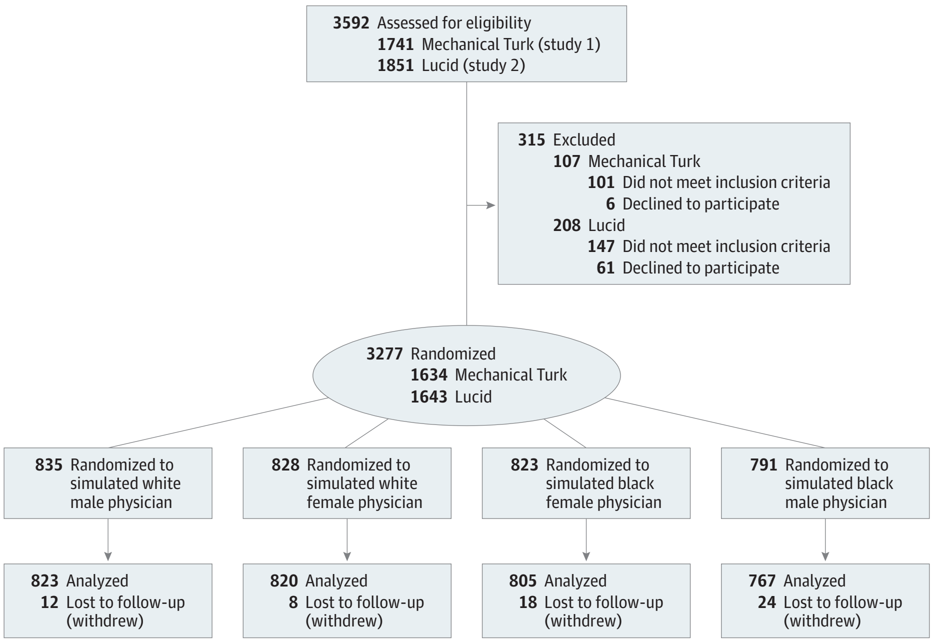
To decrease the likelihood that observed effects would be attributable to other idiosyncratic features of a particular face, we created a set of 10 images for each treatment arm (eg, 10 black men), for a total of 40 images. Simple random assignment was conducted at the participant level using the randomizer tool in Qualtrics. Each participant was first assigned to 1 of 4 possible treatment arms with equal probability: black female (n = 823), black male (n = 791), white female (n = 828), or white male (n = 835). Within each treatment arm, participants saw 1 simulated physician that was randomly selected from the set of 10 images with equal probability using simple random assignment (eMethods 1 and eMethods 2 in Supplement 2).

Primary Outcome Measures

To provide an overall summary of the effects on participant evaluations of the simulated physicians, we created a composite index (range, 0-100) by extracting the first principal component from a principal component analysis on all 5 preregistered primary outcome measures (patient confidence, patient satisfaction, likelihood to recommend, believes symptom checker, and requests more tests). Reporting a composite score for patient experience facilitates interpretability and is a method used by the Consumer Assessment of Health Plans Study survey for items, with Cronbach α coefficients greater than 0.70 indicative of good reliability.[48] All 5 primary outcome measures were highly correlated in both study 1 (α = 0.81) and study 2 (α = 0.73).

Patient confidence is the unweighted mean of participants' responses to 2 questions (from 1 [lowest] to 5 [highest] confidence): (1) "How confident are you that this physician made the correct

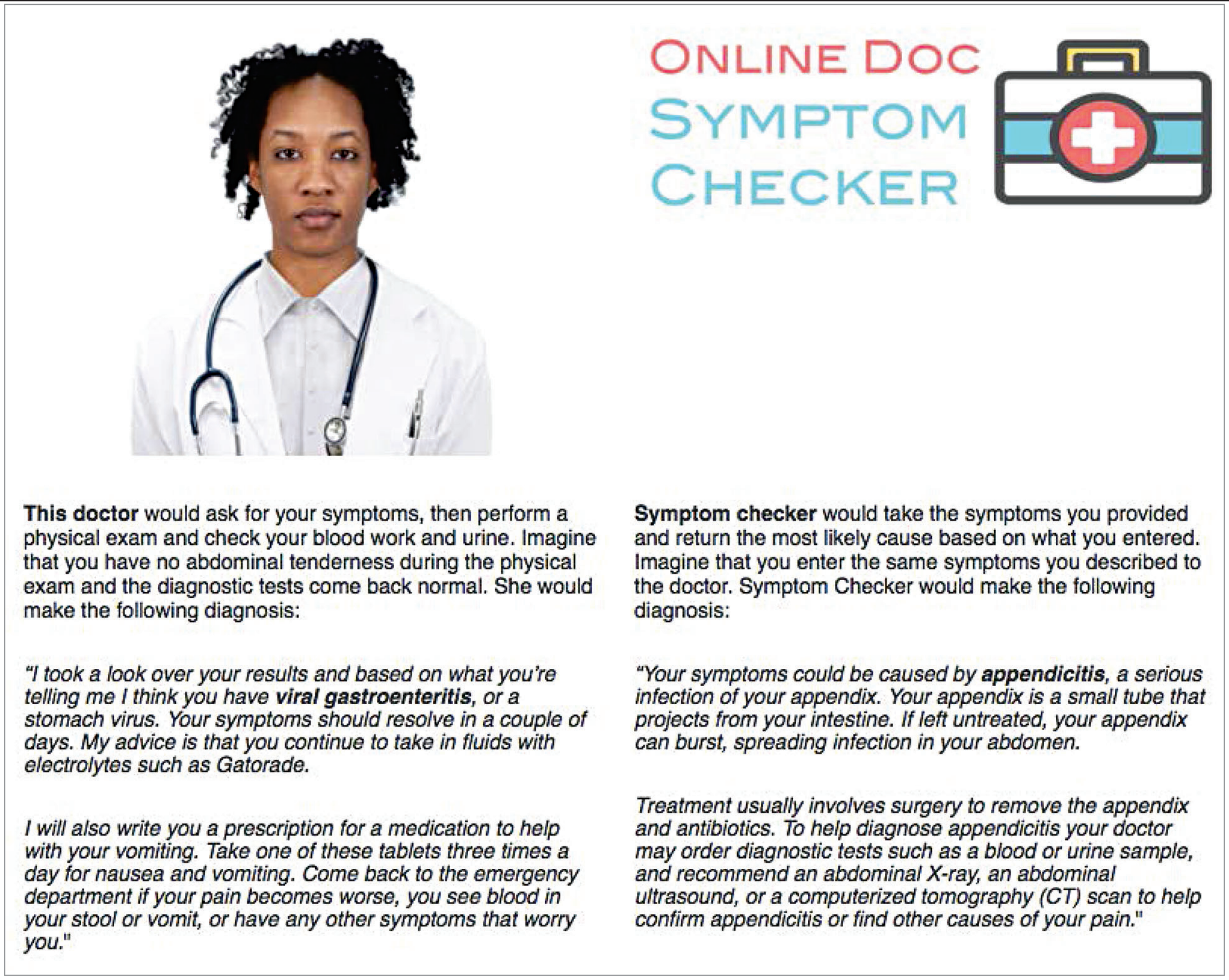
Figure 1. CONSORT Flow Diagram of Participants Through the Trial


3592 Assessed for eligibility
1741 Mechanical Turk (study 1)
1851 Lucid (study 2)
315 Excluded
107 Mechanical Turk
101 Did not meet inclusion criteria
6 Declined to participate
208 Lucid
147 Did not meet inclusion criteria
61 Declined to participate
3277 Randomized
1634 Mechanical Turk
1643 Lucid
835 Randomized to simulated white male physician
828 Randomized to simulated white female physician
823 Randomized to simulated black female physician
791 Randomized to simulated black male physician
823 Analyzed
12 Lost to follow-up (withdrew)
820 Analyzed
8 Lost to follow-up (withdrew)
805 Analyzed
18 Lost to follow-up (withdrew)
767 Analyzed
24 Lost to follow-up (withdrew)

diagnosis?" and (2) "How confident are you that this physician recommended the correct treatment plan?" *Patient satisfaction* is a single item scaled from 1 (lowest) to 10 (highest): "What number would you use to rate your care during this emergency room visit?" *Likelihood to recommend* is a single item scaled from 1 (definitely not) to 5 (definitely): "Would you recommend this physician to your friends and family?" *Believes symptom checker* is a binary response (0 [the symptom checker], 1 [the physician]) to the question: "Which diagnosis do you think is more likely to be correct?" *Requests more tests* is a single item scaled from 1 (definitely) to 5 (definitely not): "Would you ask the doctor to perform additional diagnostic tests?"

Patient confidence, believes symptom checker, and requests more tests were designed to capture the patients' confidence and willingness to challenge the physician's expertise when presented with contradictory information by an outside source of medical advice (ie, an online symptom checker). Patient satisfaction and likelihood to recommend are global ratings of satisfaction from the Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) and the Press Ganey survey, the most commonly used surveys of patient experience.[49,50] eMethods 4 in Supplement 2 gives complete details on all survey items used to construct the primary outcome measures, with screenshots of how they appeared to participants in the vignettes (eFigures 4-7 in Supplement 2) .

Figure 2. Treatment Vignette



This doctor would ask for your symptoms, then perform a physical exam and check your blood work and urine. Imagine that you have no abdominal tenderness during the physical exam and the diagnostic tests come back normal. She would make the following diagnosis:

"I took a look over your results and based on what you're telling me I think you have ***viral gastroenteritis****, or a stomach virus. Your symptoms should resolve in a couple of days. My advice is that you continue to take in fluids with electrolytes such as Gatorade.*

I will also write you a prescription for a medication to help with your vomiting. Take one of these tablets three times a day for nausea and vomiting. Come back to the emergency department if your pain becomes worse, you see blood in your stool or vomit, or have any other symptoms that worry you."

Symptom checker would take the symptoms you provided and return the most likely cause based on what you entered. Imagine that you enter the same symptoms you described to the doctor. Symptom Checker would make the following diagnosis:

"Your symptoms could be caused by ***appendicitis****, a serious infection of your appendix. Your appendix is a small tube that projects from your intestine. If left untreated, your appendix can burst, spreading infection in your abdomen.*

Treatment usually involves surgery to remove the appendix and antibiotics. To help diagnose appendicitis your doctor may order diagnostic tests such as a blood or urine sample, and recommend an abdominal X-ray, an abdominal ultrasound, or a computerized tomography (CT) scan to help confirm appendicitis or find other causes of your pain."

Photograph reproduced with permission from the Chicago Face Database.[41]

Secondary Outcome Measures

Studies 1 and 2 also included preregistered secondary outcome measures of perceived warmth and competence of the simulated physicians (eFigure 8 and eFigure 9 in Supplement 2). These measures have been used to study patient-physician relationships and capture 2 dimensions of stereotype content in social perception.[51,52] Study 1 also included a measure of the perceived fairness of a $350 charge for the ED visit (eFigure 10 in Supplement 2), and study 2 included measures for willingness to complain and sue the physician if a diagnostic error resulted in an adverse outcome (eFigure 11 in Supplement 2) because previous studies have identified increased medicolegal action against physicians who belong to minority groups.[53,54]

Statistical Analysis

We estimated treatment effects using ordinary least-squares regression on the 4-level treatment factor, with white male as the omitted reference category. To maximize the precision of estimated treatment effects, we used covariate-adjusted ordinary least squares on the stacked data set of both experiments.[55] Covariates used in adjusted regression were measured pretreatment and included participant-level demographic characteristics (age, race/ethnicity, gender, and college education), self-reported trust in physicians, mental and overall health, insurance status, unpaid medical bills, and frequency of ED visits in the previous 6 months. We also added a study fixed effect (binary indicator) to adjust for differences across studies. We reported estimated treatment effects for the composite index and the 5 underlying primary outcome measures.

To facilitate interpretation, estimates were standardized using a Glass delta, which scaled outcomes by the SD in the white male control group. Results are presented graphically with 90% CIs and 95% CIs and a margin of equivalence bound within 0.20 standard units, which corresponds to an effect size of approximately one-fifth of 1 SD. The null hypothesis of nonequivalence is rejected in favor of equivalence when a 90% CI is contained within the margin of equivalence, and the null hypothesis of no significant difference from 0 is rejected if a 95% CI excludes 0. With 700 participants per treatment arm (N = 2800), the minimum detectable effect at 80% power using a 2-sided hypothesis test (at $\alpha = .05$) is approximately 0.15 standardized units for any between-group difference. Combining the results from these 2 testing procedures assisted us in ruling out the presence of effects larger than the margin of equivalence, which was the smallest effect size of interest in this study.[56] We concluded that an estimated effect was negligible (bounded between −0.20 and 0.20 standard units) when the 90% CI was inside the margin of equivalence and the 95% CI included 0.

We also examined whether certain subgroups of participants may have been affected differently by treatment using bayesian additive regression trees (BARTs), a machine learning algorithm that estimates treatment effect heterogeneity as a function of each participant's covariate profile by including multiple potential moderators in the same model.[57,58] We preregistered this BART analysis for demographic covariates as well as measures of racial prejudice and sexism. To provide an overall summary of treatment effect heterogeneity, we plotted BART-estimated treatment effects with 95% credible intervals for each individual. Intervals that excluded 0 provided evidence in support of treatment effect heterogeneity. R version 3.5.1 (R Project for Statistical Computing) statistical software was used for statistical analyses, and the dbarts package was used for BARTs. eMethods 6 in Supplement 2 provides additional details on BART implementation.

Results

Of the 3277 randomized participants, 3215 (representing all contiguous US states) completed the survey (Figure 1). In this combined sample, participants' median age was 49 years (range, 18-89 years), 52% (1667 of 3215) were female, 76% (2433 of 3215) were white, and 10% (333 of 3215) were black. The **Table** reports background characteristics for study 1 (MTurk), study 2 (Lucid), and the pooled sample. Approximately 40% of participants in study 1 and 34% in study 2 endorsed some group-level superiority of white individuals compared with the black individuals (eMethods 5 in Supplement 2). We did not find evidence that loss to follow-up was associated with imbalance in background characteristics by treatment arm for either study (eMethods 3 and eTables 1-3 in Supplement 2).

Primary Outcomes

In the combined sample (n = 3215), the unadjusted mean composite index was not statistically distinguishable for any pairwise comparison across treatment arms (white male, 66.13 [95% CI, 64.76-67.51]; black male, 66.96 [95% CI, 65.55-68.36]; black female, 67.36 [95% CI, 66.03-68.69]; white female, 66.50 [95% CI, 65.19-67.82]) (eTable 4 in Supplement 2). Estimated covariate-adjusted treatment effects (estimated against the white male control) (eTable 5 in Supplement 2) on the composite index were also not statistically distinguishable from 0 (white female, 0.03 [95% CI, −0.07 to 0.13]; black female, 0.05 [95% CI, −0.05 to 0.15]; black male, 0.06 [95 % CI, −0.04 to 0.16]). On the basis of the combined results from equivalence tests and null hypothesis tests, we found no detectable effects of physician gender and race and ruled out effects larger than within 0.20 standard units on the composite index and all underlying primary outcome measures (**Figure 3**). No significant differences were observed when study 1 and study 2 were instead analyzed separately (eTable 6 and eTable 7 in Supplement 2).

Treatment Effect Heterogeneity

Figure 4 plots BART-estimated treatment effects on the composite index for each participant as a function of their individual covariate profile for both MTurk (study 1) and Lucid (study 2) samples. The BART-estimated treatment effects were consistently indistinguishable from 0 and similar across participant samples. This analysis revealed little evidence of variation in BART-estimated treatment effects as a function of participant-level characteristics (Figure 4). The corresponding 95% credible interval did not exclude 0 in any of the cases in which a participant was estimated to have a positive (or negative) treatment effect (eTable 8 in Supplement 2). We therefore did not find compelling evidence that some subgroups of participants (eg, prejudiced white men without a college education who were aged ≥65 years) responded differently to the race and gender of simulated physicians than others.

Table. Baseline Characteristics of Study Participants[a]

Characteristic	Combined (N = 3215)	Study 1 (n = 1619)	Study 2 (n = 1596)
Age, median (range), y	49 (18-89)	50 (19-89)	45 (18-86)
Female	1667 (51.85)	873 (53.92)	794 (49.75)
College educated	1515 (47.12)	818 (50.53)	697 (43.67)
Household income below median level	2086 (66.10)	967 (59.95)	1119 (72.52)
Race/ethnicity			
White non-Hispanic	2433 (75.68)	1300 (80.30)	1133 (70.99)
Black	333 (10.36)	157 (9.70)	176 (11.03)
Hispanic	206 (6.41)	64 (3.95)	142 (8.90)
Other	243 (7.56)	98 (6.05)	145 (9.09)
Insurance			
Medicaid	404 (12.57)	34 (2.10)	370 (23.18)
Medicare	362 (11.26)	121 (7.47)	241 (15.10)
Uninsured	420 (13.06)	229 (14.14)	191 (11.97)
Unpaid medical bills	735 (22.86)	384 (23.72)	351 (21.99)
≥1 Emergency department visit in past 6 mo	609 (18.94)	240 (14.82)	369 (23.12)
Mental health, mean (SD)[b]	3.60 (1.12)	3.65 (1.11)	3.54 (1.12)
Overall health, mean (SD)[c]	3.43 (0.96)	3.44 (0.95)	3.41 (0.98)
Trust in physicians, mean (SD)[d]	3.88 (0.81)	3.89 (0.87)	3.88 (0.75)

[a] Data are presented as number (percentage) of participants unless otherwise indicated.

[b] Mean score on validated Likert score (excellent [5], very good [4], good [3], fair [2], poor [1]) for "In general, how would you rate your mental health?"

[c] Mean score on validated Likert score (excellent [5], very good [4], good [3], fair [2], poor [1]) for "In general, how would you rate your overall health?"

[d] Mean score on validated Likert score (strongly agree [7], agree [6], somewhat agree [5], neither agree nor disagree [4], somewhat disagree [3], disagree [2], strongly disagree [1]) for "How much do you agree or disagree with the following statement: All things considered, doctors in the United States can generally be trusted."

Secondary Outcomes

Perceived warmth and competence scales were created in study 1 (warmth: Cronbach α = 0.88; competence: Cronbach α = 0.88) and study 2 (warmth: Cronbach α = 0.89; competence: Cronbach α = 0.94). We did not find evidence of participant bias against black or female physicians on the secondary outcomes of perceived warmth and competence, perceived fairness of ED visit charge, or willingness to sue or complain because of misdiagnosis that resulted in a bad outcome (eMethods 7, eTable 9, and eTable 10 in Supplement 2).

Discussion

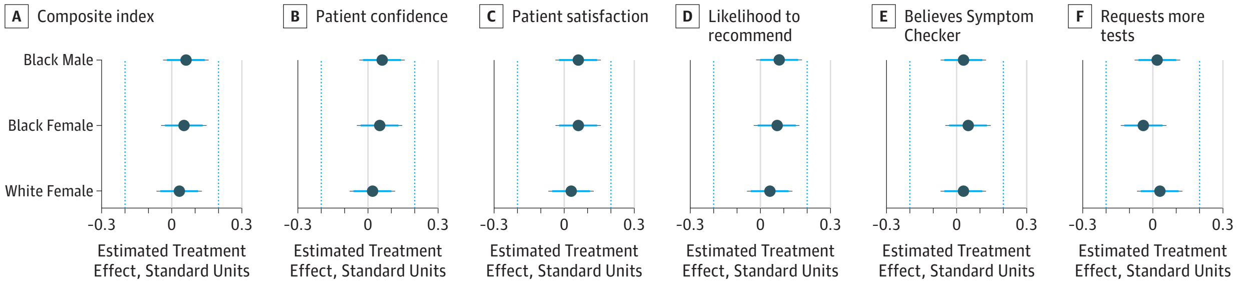
To our knowledge, these were the first large-scale, preregistered, randomized experiments to directly estimate the effect of physician gender and race on simulated patient evaluations in a diverse national sample of US participants. We found that black or female physicians were not rated lower than white male physicians in measures of simulated patient satisfaction or confidence in an ED setting. These findings suggest that survey-based measures of simulated patient confidence and satisfaction may not be systematically negatively affected by physician race and gender.

Of importance, the results reported here should not be interpreted as contradicting the lived experiences of discrimination reported by physicians from underrepresented groups. The absence of a systematic preference for white male physicians in the controlled setting does not diminish the damaging and lasting effect that even a single instance of discrimination from patients or colleagues can have on minority and female physicians.[10,12] However, we did not find compelling evidence that participants were biased against black or female physicians in the simulated interactions in our study.

The experimental designs we used address several important methodologic challenges present in previous research, including a large sample size, random assignment of physician race and gender, replication across 2 independent studies, and preregistration of primary and secondary outcomes and the statistical analyses. Unlike previous experiments that recruited smaller numbers of undergraduate or medical students as patient analogs, we recruited a large, geographically diverse pool of participants who more closely approximates typical ED patients.[31] In addition, our stimuli used multiple physician images drawn from a validated stimulus set to minimize differences in observable characteristics (eg, physician attractiveness) that may affect ratings of confidence and satisfaction independent of physician race and gender. Furthermore, we controlled for variability in the clinical content of the encounter by holding constant the simulated physicians' diagnoses and treatment plans as well as communication styles across treatment arms.

Our results add to a growing body of observational studies investigating patient bias against female and minority group physicians, most of which have not found evidence of systematic bias. For

Figure 3. Estimated Treatment Effects of Race and Gender of Simulated Physicians on Composite Index and Primary Outcome Measures

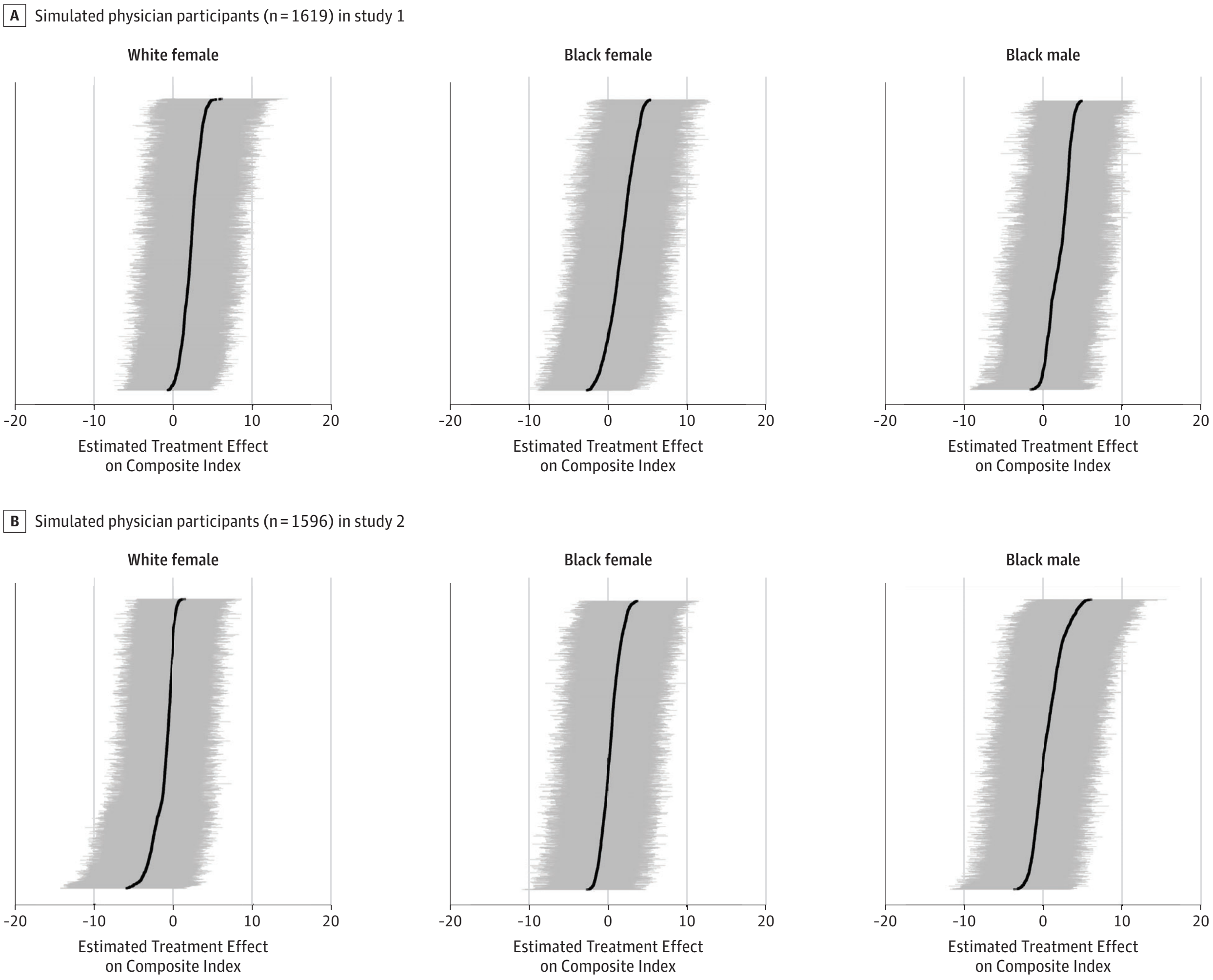


Covariate-adjusted treatment effects from ordinary least squares regression with control group (simulated white male physician). Estimates are pooled across 2 independent patient analog experiments (N = 3215) and standardized using the Glass delta, which scales outcomes by the SD in the control group. Composite index (range, 0-100) was created by extracting the first principal component from a principal component analysis on all primary outcome measures (patient confidence, patient satisfaction, likelihood to recommend, believes symptom checker, and requests more tests) (eMethods 4 and eTable 5 in Supplement 2). Dots indicate means; cyan lines, 90% CIs; gray lines, 95% CIs; and cyan dotted vertical lines, margin of equivalence within 0.20 standard units.

example, an analysis of more than 9000 Press Ganey surveys found no differences in patient satisfaction by ED physicians' gender.[59] A meta-analysis focused on observational studies in primary care settings found negligible evidence that patients favored female physicians, an effect attributed to publication bias and gender-specific differences in physician communication styles and patient-centered care.[16] Many observational studies have examined the association between patient-physician gender-concordance and patient satisfaction, but the results have been mixed across a variety of clinical settings, with some reporting gender concordance preferences and others finding evidence of gender discordance.[47,60-62]

Few studies have examined the association of physician race with overall patient evaluations, and the results have been inconsistent depending on the study design. An analysis of more than 51 000 Press Ganey and HCAHPs surveys[63] from hospital-discharged patients found no difference in patient satisfaction by physician race. In contrast, other observational surveys[23,24] have shown an association between race concordance and patients' satisfaction or feeling that a visit was participatory, such that minority group physicians were favored by their minority group patients

Figure 4. Bayesian Additive Regression Tree (BART)–Estimated Treatment Effects of Race and Gender of Simulated Physicians on Composite Index of Primary Outcomes by Study



Black dots indicate the BART-estimated treatment effect for each individual as a function of their covariate profile, ordered by effect size. Grey horizontal lines indicate 95% credible intervals; intervals that exclude 0 would provide evidence of treatment effect heterogeneity. Composite index (range, 0-100) was created by extracting the first principal component from a principal component analysis on all primary outcome measures (patient confidence, patient satisfaction, likelihood to recommend, believes symptom checker, and requests more tests) (eMethods 4 and eMethods 6 in Supplement 2).

compared with white physicians. Although minority group physicians may be preferred by minority group patients, a study found that they may experience bias by simulated white patients. In a simulation experiment, participants were shown a physician profile with a randomized name to represent a different race or gender of the physician. White participants were less likely to select a black or middle-eastern physician even though they had the same quality scores compared with the white physician counterpart.[64]

To our knowledge, the only randomized experiment on patient-physician race concordance, conducted in Oakland, California, found that black men assigned to black male physicians took more preventive health measures than black male patients assigned to white male physicians.[65] Although our experiments were not designed to detect race concordance effects (black participants were 10% of our sample), the BART analysis did not reveal compelling evidence of treatment effect heterogeneity as a function of participants' background characteristics. This finding suggests that increasing the diversity of the physician workforce is unlikely to decrease patient satisfaction and may improve quality of care for patients from underrepresented minority groups. Adequately powered experimental designs that study the effects of race and gender concordance on quality of care and health outcomes is an avenue for future research.

The importance of creating inclusive and diverse workplaces in health care cannot be overstated because more diverse teams are associated with better patient care, lower mortality, better science, and more successful organizations with higher productivity, innovation, and employee retention.[66-69] Thus, there is a renewed call to improve the status quo through institutional-level change to elevate underrepresented groups using accountability measures through organizations such as Time's Up Healthcare and Men Advocating Real Change.[70,71] Our study further supports these efforts by suggesting that patient bias against physicians may be less of a driver of workplace discrimination than these other sources.

Limitations

This study has limitations. First, the experimental designs used written clinical vignettes in a hypothetical ED interaction in which neither the physician nor the patient were real. Simulated encounters cannot capture important characteristics of real-world interactions that might shape patient-physician interactions, such as nonverbal communication and communication style. However, the use of lay participants to play the role of a patient analog is supported by a meta-analysis of communication studies that showed a large overlap between patient analogs and patient perceptions of a clinical encounter.[72] Furthermore, case vignettes using written descriptions show that physicians make similar assessments from vignettes as they do in real clinical encounters.[73] The benefit of simulation designs is that they control for complexities introduced by real-world interactions, such as practice styles, which have been found to be independently associated with patient ratings.[25,27]

Furthermore, the use of an ED setting may limit the generalizability of the findings reported here to other clinical contexts. Unlike other contexts, the ED is a unique environment where patient-physician relationships are brief and episodic, and physicians cannot be chosen by patients in advance. Investigating the role that the length of the patient-physician relationship and patient choice play in determining patient satisfaction is an important area for future research.

Moreover, we chose a low-acuity clinical vignette. It is possible that a high-stake encounter could have elicited a different response. However, in our second experiment (study 2), we extended the vignette and the participant was told they underwent emergency surgery because of a misdiagnosis by the physician and had to stay in the intensive care unit. We did not find evidence of race or gender biases in the extent to which patients desired retribution in terms of willingness to sue or complain for the physician's error. Still, conducting similar studies of situations in which patients experience greater stress or cognitive load is another important avenue for future research.

In addition, it is theoretically possible that some participants may have discerned the purpose of the study and censored their prejudice against female and black physicians to appear more socially desirable, thereby attenuating estimated treatment effects. However, all participants were blinded to the study objectives, and to our knowledge, there is no empirical support for such threats to inference in randomized survey experiments conducted in the anonymous online environment.[74,75] In addition, our respondents willingly disclosed racial and/or gender prejudice on related measures; for example, approximately 40% of participants in study 1 and 34% of participants in study 2 endorsed some group-level superiority of white individuals vs black individuals (eMethods 5 in the Supplement 2). We did not find evidence that these characteristics were predictive of heterogeneous effects in the BART analysis.

Conclusions

Using large, survey-based experiments of a simulated ED encounter, we found no detectable effects of physicians' race or gender on simulated patients' confidence and satisfaction. These results suggest that institutional biases and workplace dynamics, including potential discrimination from leadership, peers, and staff, may play a greater role in the bias experienced by women and minority group physicians in the ED than clinical encounters with patients.

ARTICLE INFORMATION

Accepted for Publication: November 20, 2019.

Published: February 21, 2020. doi:10.1001/jamanetworkopen.2019.20511

Corresponding Author: Basmah Safdar, MD, MSc, Department of Emergency Medicine, Yale School of Medicine, 464 Congress Ave, Ste 260, New Haven, CT 06519 (basmah.safdar@yale.edu).

Author Affiliations: National Clinical Scholars Program, Institute for Healthcare Policy and Innovation, University of Michigan, Ann Arbor (Solnick); Department of Emergency Medicine, University of Michigan Medical School, Ann Arbor (Solnick); Yale Law School, New Haven, Connecticut (Peyton); Department of Psychology, Boston College, Chestnut Hill, Massachusetts (Kraft-Todd); Department of Emergency Medicine, Yale School of Medicine, New Haven, Connecticut (Safdar).

Author Contributions: Drs Solnick and Safdar had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Concept and design: All authors.

Acquisition, analysis, or interpretation of data: All authors.

Drafting of the manuscript: All authors.

Critical revision of the manuscript for important intellectual content: All authors.

Statistical analysis: Peyton, Kraft-Todd.

Obtained funding: Solnick, Safdar.

Administrative, technical, or material support: Solnick, Safdar.

Supervision: Solnick, Safdar.

Conflict of Interest Disclosures: None reported.

Funding/Support: The study was funded by the 2017 Joan F. Giambalvo Fund for the Advancement of Women research grants program from the American Medical Association Foundation.

Role of the Funder/Sponsor: The funding organization had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Data Sharing Statement: See Supplement 3.

Meeting Presentation: This study was presented at the Society of Academic Emergency Medicine Annual Meeting, May 15, 2019; Las Vegas, Nevada; and at the American Medical Association Annual Meeting, Women Physicians Section Business Meeting; June 9, 2019; Chicago, Illinois.

REFERENCES

1. Association of American Medical Colleges. Table 1: medical students, selected years, 1965-2015. https://www.aamc.org/system/files/reports/1/2015table1.pdf. Published 2016. Accessed February 11, 2019.

2. Association of American Medical Colleges. Table A-14.1: race/ethnicity responses (alone and in combination) of applicants to US medical schools, 2015-2016 through 2019-2020. https://www.aamc.org/system/files/2019-11/2019_FACTS_Table_A-14.1.pdf. Accessed January 2, 2020.

3. Jagsi R, Griffith KA, Jones R, Perumalswami CR, Ubel P, Stewart A. Sexual harassment and discrimination experiences of academic medical faculty. *JAMA*. 2016;315(19):2120-2121. doi:10.1001/jama.2016.2188

4. Jagsi R, Griffith KA, Stewart A, Sambuco D, DeCastro R, Ubel PA. Gender differences in the salaries of physician researchers. *JAMA*. 2012;307(22):2410-2417. doi:10.1001/jama.2012.6183

5. Madsen TE, Linden JA, Rounds K, et al. Current status of gender and racial/ethnic disparities among academic emergency medicine physicians. *Acad Emerg Med*. 2017;24(10):1182-1192. doi:10.1111/acem.13269

6. Fang D, Moy E, Colburn L, Hurley J. Racial and ethnic disparities in faculty promotion in academic medicine. *JAMA*. 2000;284(9):1085-1092. doi:10.1001/jama.284.9.1085

7. Pololi LH, Civian JT, Brennan RT, Dottolo AL, Krupat E. Experiencing the culture of academic medicine: gender matters, a national study. *J Gen Intern Med*. 2013;28(2):201-207. doi:10.1007/s11606-012-2207-1

8. Jenner S, Djermester P, Prügl J, Kurmeyer C, Oertelt-Prigione S. Prevalence of sexual harassment in academic medicine. *JAMA Intern Med*. 2019;179(1):108-111. doi:10.1001/jamainternmed.2018.4859

9. National Academies of Sciences, Engineering, and Medicine. *Sexual Harassment of Women: Climate, Culture, and Consequences in Academic Sciences, Engineering, and Medicine*. Washington, DC: National Academies Press; 2018.

10. Osseo-Asare A, Balasuriya L, Huot SJ, et al. Minority resident physicians' views on the role of race/ethnicity in their training experiences in the workplace. *JAMA Netw Open*. 2018;1(5):e182723. doi:10.1001/jamanetworkopen.2018.2723

11. Nunez-Smith M, Curry LA, Bigby J, Berg D, Krumholz HM, Bradley EH. Impact of race on the professional lives of physicians of African descent. *Ann Intern Med*. 2007;146(1):45-51. doi:10.7326/0003-4819-146-1-200701020-00008

12. Wheeler M, de Bourmont S, Paul-Emile K, et al. Physician and trainee experiences with patient bias [published online October 28, 2019]. *JAMA Intern Med*. 2019. doi:10.1001/jamainternmed.2019.4122

13. Choo EK, van Dis J, Kass D. Time's up for medicine? only time will tell. *N Engl J Med*. 2018;379(17):1592-1593. doi:10.1056/NEJMp1809351

14. Nunez-Smith M, Pilgrim N, Wynia M, et al. Health care workplace discrimination and physician turnover. *J Natl Med Assoc*. 2009;101(12):1274-1282. doi:10.1016/S0027-9684(15)31139-1

15. McMurray JE, Linzer M, Konrad TR, Douglas J, Shugerman R, Nelson K; The SGIM Career Satisfaction Study Group. The work lives of women physicians: results from the physician work life study. *J Gen Intern Med*. 2000;15(6):372-380. doi:10.1046/j.1525-1497.2000.9908009.x

16. Hall JA, Blanch-Hartigan D, Roter DL. Patients' satisfaction with male versus female physicians: a meta-analysis. *Med Care*. 2011;49(7):611-617. doi:10.1097/MLR.0b013e318213c03f

17. Schindelheim GL, Jerrard DA, Witting M. Patient preference for emergency physician age and gender. *Am J Emerg Med*. 2004;22(6):503. doi:10.1016/j.ajem.2004.07.011

18. Cousin G, Schmid Mast M, Jaunin-Stalder N. When physician-expressed uncertainty leads to patient dissatisfaction: a gender study. *Med Educ*. 2013;47(9):923-931. doi:10.1111/medu.12237

19. Rogo-Gupta LJ, Haunschild C, Altamirano J, Maldonado YA, Fassiotto M. Physician gender is associated with Press Ganey patient satisfaction scores in outpatient gynecology. *Womens Health Issues*. 2018;28(3):281-285. doi:10.1016/j.whi.2018.01.001

20. Bertakis KD, Franks P, Azari R. Effects of physician gender on patient satisfaction. *J Am Med Womens Assoc (1972)*. 2003;58(2):69-75.

21. Gerbert B, Berg-Smith S, Mancuso M, et al. Video study of physician selection: preferences in the face of diversity. *J Fam Pract*. 2003;52(7):552-559.

22. Nolen HA, Moore JX, Rodgers JB, Wang HE, Walter LA. Patient preference for physician gender in the emergency department. *Yale J Biol Med*. 2016;89(2):131-142.

23. Laveist TA, Nuru-Jeter A. Is doctor-patient race concordance associated with greater satisfaction with care? *J Health Soc Behav*. 2002;43(3):296-306. doi:10.2307/3090205

24. Cooper-Patrick L, Gallo JJ, Gonzales JJ, et al. Race, gender, and partnership in the patient-physician relationship. *JAMA*. 1999;282(6):583-589. doi:10.1001/jama.282.6.583

25. Roter DL, Hall JA, Aoki Y. Physician gender effects in medical communication: a meta-analytic review. *JAMA*. 2002;288(6):756-764. doi:10.1001/jama.288.6.756

26. Schmid Mast M, Hall JA, Roter DL. Disentangling physician sex and physician communication style: their effects on patient satisfaction in a virtual medical visit. *Patient Educ Couns*. 2007;68(1):16-22. doi:10.1016/j.pec.2007.03.020

27. Hall JA, Roter DL, Blanch-Hartigan D, Mast MS, Pitegoff CA. How patient-centered do female physicians need to be? analogue patients' satisfaction with male and female physicians' identical behaviors. *Health Commun*. 2015;30(9):894-900. doi:10.1080/10410236.2014.900892

28. Hall JA, Gulbrandsen P, Dahl FA. Physician gender, physician patient-centered behavior, and patient satisfaction: a study in three practice settings within a hospital. *Patient Educ Couns*. 2014;95(3):313-318. doi:10.1016/j.pec.2014.03.015

29. Moher D, Hopewell S, Schulz KF, et al; Consolidated Standards of Reporting Trials Group. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *J Clin Epidemiol*. 2010;63(8):e1-e37. doi:10.1016/j.jclinepi.2010.03.004

30. Burgess DJ. Are providers more likely to contribute to healthcare disparities under high levels of cognitive load? how features of the healthcare setting may lead to biases in medical decision making. *Med Decis Making*. 2010;30(2):246-257. doi:10.1177/0272989X09341751

31. Rui P. Kang K. National hospital ambulatory medical care survey: 2015 emergency department summary tables. https://www.cdc.gov/nchs/data/nhamcs/web_tables/2015_ed_web_tables.pdf. Published 2015. Accessed February 11, 2018.

32. Coppock A. Generalizing from survey experiments conducted on Mechanical Turk: a replication approach. *Polit Sci Res and Methods.* 2019;7(3):613-628. doi:10.1017/psrm.2018.10

33. Behrend TS, Sharek DJ, Meade AW, Wiebe EN. The viability of crowdsourcing for survey research. *Behav Res Methods*. 2011;43(3):800-813. doi:10.3758/s13428-011-0081-0

34. Berinsky AJ, Huber GA, Lenz GS. Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Polit Anal*. 2012;20(3):351-368. doi:10.1093/pan/mpr057

35. Horton JJ, Rand DG, Zeckhauser RJ. The online laboratory: conducting experiments in a real labor market. *Exp Econ*. 2011;14(3):399-425. doi:10.1007/s10683-011-9273-9

36. Coppock A, McClellan OA. Validating the demographic, political, psychological, and experimental results obtained from a new source of online survey respondents. *Res Polit*. 2019;6(1):1-14. doi:10.1177/2053168018822174

37. Hang BS, Bork S, Ditkoff J, Long B, Koyfman A. Nausea and vomiting. In: Tintinalli JE, Stapczynski JS, Ma OJ, et al, eds. *Tintinalli's Emergency Medicine: A Comprehensive Study Guide*. 8th ed. New York, NY: McGraw-Hill Companies; 2011.

38. US Department of Health and Human Services, National Institutes of Health, Office of Extramural Research. NIH policy on reporting race and ethnicity data: subjects in clinical research. 2001. https://grants.nih.gov/grants/guide/notice-files/not-od-01-053.html. Accessed January 3, 2020.

39. Oppenheimer DM, Meyvis T, Davidenko N. Instructional manipulation checks: detecting satisficing to increase statistical power. *J Exp Soc Psychol*. 2009;45(4):867-872. doi:10.1016/j.jesp.2009.03.009

40. Meade AW, Craig SB. Identifying careless responses in survey data. *Psychol Methods*. 2012;17(3):437-455. doi:10.1037/a0028085

41. Ma DS, Correll J, Wittenbrink B. The Chicago face database: A free stimulus set of faces and norming data. *Behav Res Methods*. 2015;47(4):1122-1135. doi:10.3758/s13428-014-0532-5

42. Glick P, Fiske ST. Hostile and benevolent sexism: measuring ambivalent sexist attitudes toward women. *Psychol Women Q*. 1997;21(1):119-135. doi:10.1111/j.1471-6402.1997.tb00104.x

43. Kunst JR, Fischer R, Sidanius J, Thomsen L. Preferences for group dominance track and mediate the effects of macro-level social inequality and violence across societies. *Proc Natl Acad Sci U S A*. 2017;114(21):5407-5412. doi:10.1073/pnas.1616572114

44. Huddy L, Feldman S. On assessing the political effects of racial prejudice. *Annu Rev Polit Sci*. 2009;12(1):423-447. doi:10.1146/annurev.polisci.11.062906.070752

45. Peyton K, Huber GA. Do survey measures of racial prejudice predict racial discrimination? experimental evidence on anti-black discrimination. *SocArXiv*. Published online April 18, 2018.

46. Ma DS, Correll J, Wittenbrink B. The Chicago Face Database: a free stimulus set of faces and norming data. *Behav Res Methods*. 2015;47(4):1122-1135. doi:10.3758/s13428-014-0532-5

47. Hall JA, Irish JT, Roter DL, Ehrlich CM, Miller LH. Satisfaction, gender, and communication in medical visits. *Med Care*. 1994;32(12):1216-1231. doi:10.1097/00005650-199412000-00005

48. Hargraves JL, Hays RD, Cleary PD. Psychometric properties of the Consumer Assessment of Health Plans Study (CAHPS) 2.0 adult core survey. *Health Serv Res*. 2003;38(6, pt 1):1509-1527. doi:10.1111/j.1475-6773.2003.00190.x

49. Centers for Medicare & Medicaid Services. HCAHPS: patients' perspectives of care survey. Page last modified, October 15, 2019. https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-instruments/hospitalqualityinits/hospitalHCAHPS.html. Accessed February 15, 2019.

50. Presson AP, Zhang C, Abtahi AM, Kean J, Hung M, Tyser AR. Psychometric properties of the Press Ganey outpatient medical practice survey. *Health Qual Life Outcomes*. 2017;15(1):32. doi:10.1186/s12955-017-0610-3

51. Kraft-Todd GT, Reinero DA, Kelley JM, Heberlein AS, Baer L, Riess H. Empathic nonverbal behavior increases ratings of both warmth and competence in a medical context. *PLoS One*. 2017;12(5):e0177758. doi:10.1371/journal.pone.0177758

52. Fiske ST, Cuddy AJ, Glick P, Xu J. A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *J Pers Soc Psychol*. 2002;82(6):878-902. doi:10.1037/0022-3514.82.6.878

53. Cardarelli R, Licciardone JC, Ramirez G. Predicting risk for disciplinary action by a state medical board. *Tex Med*. 2004;100(1):84-90.

54. Rogers P. *Demographics of Disciplinary Action by the Medical Board of California (2003-2013)*. Sacramento: California Research Bureau; January 2017.

55. Lin W. Agnostic notes on regression adjustments to experimental data: reexamining Freedman's critique. *Ann Appl Stat*. 2013;7(1):295-318. doi:10.1214/12-AOAS583

56. Lakens D, Scheel AM, Isager PM. Equivalence testing for psychological research: a tutorial. *Adv Methods Pract Psychol Sci*. 2018;1(2):259-269. doi:10.1177/2515245918770963

57. Green DP, Kern HL. Modeling heterogeneous treatment effects in survey experiments with bayesian additive regression trees. *Public Opin Q*. 2012;76(3):491-511. doi:10.1093/poq/nfs036

58. Chipman HA, George EI, McCulloch RE. BART: bayesian additive regression trees. *Ann Appl Stat*. 2010;4(1):266-298. doi:10.1214/09-AOAS285

59. Milano A, Dalawari P, McGregor AJ, et al. Emergency department evaluation of patient satisfaction: does physician gender impact Press Ganey scores? a multicenter study. *Am J Emerg Med*. 2018;36(9):1708-1709. doi:10.1016/j.ajem.2018.01.067

60. Street RL Jr, O'Malley KJ, Cooper LA, Haidet P. Understanding concordance in patient-physician relationships: personal and ethnic dimensions of shared identity. *Ann Fam Med*. 2008;6(3):198-205. doi:10.1370/afm.821

61. Derose KP, Hays RD, McCaffrey DF, Baker DW. Does physician gender affect satisfaction of men and women visiting the emergency department? *J Gen Intern Med*. 2001;16(4):218-226. doi:10.1046/j.1525-1497.2001.016004218.x

62. Schmittdiel J, Grumbach K, Selby JV, Quesenberry CP Jr. Effect of physician and patient gender concordance on patient satisfaction and preventive care practices. *J Gen Intern Med*. 2000;15(11):761-769. doi:10.1046/j.1525-1497.2000.91156.x

63. Chen JG, Zou B, Shuster J. Relationship between patient satisfaction and physician characteristics. *J Patient Exp*. 2017;4(4):177-184. doi:10.1177/2374373517714453

64. Greene J, Hibbard JH, Sacks RM. Does the race/ethnicity or gender of a physician's name impact patient selection of the physician? *J Natl Med Assoc*. 2018;110(3):206-211. doi:10.1016/j.jnma.2017.05.010

65. Alsan M, Garrick O, Graziani GC. Does diversity matter for health? experimental evidence from Oakland. Working paper 24787. *Natl Bureau Econ Res*. https://www.nber.org/papers/w24787?utm_campaign=ntw&utm_medium=email&utm_source=ntw. Published 2018. Accessed September, 21, 2018.

66. Tsugawa Y, Jena AB, Figueroa JF, Orav EJ, Blumenthal DM, Jha AK. Comparison of hospital mortality and readmission rates for Medicare patients treated by male vs female physicians. *JAMA Intern Med*. 2017;177(2):206-213. doi:10.1001/jamainternmed.2016.7875

67. Shannon G, Jansen M, Williams K, et al. Gender equality in science, medicine, and global health: where are we at and why does it matter? *Lancet*. 2019;393(10171):560-569. doi:10.1016/S0140-6736(18)33135-0

68. Greenwood BN, Carnahan S, Huang L. Patient-physician gender concordance and increased mortality among female heart attack patients. *Proc Natl Acad Sci U S A*. 2018;115(34):8569-8574. doi:10.1073/pnas.1800097115

69. Morgan Stanley. An investor's guide to gender diversity. https://www.morganstanley.com/ideas/gender-diversity-investor-guide. Published January 17, 2017. Accessed April 25, 2019.

70. The Lancet. Feminism is for everybody. *Lancet*. 2019;393(10171):493. doi:10.1016/S0140-6736(19)30239-9

71. Catalyst. Men advocating real change. https://www.catalyst.org/marc/. Published 2019. Accessed April 25, 2019.

72. van Vliet LM, van der Wall E, Albada A, Spreeuwenberg PMM, Verheul W, Bensing JM. The validity of using analogue patients in practitioner-patient communication research: systematic review and meta-analysis. *J Gen Intern Med*. 2012;27(11):1528-1543. doi:10.1007/s11606-012-2111-8

73. Kirwan JR, Chaput de Saintonge DM, Joyce CR, Currey HL. Clinical judgment in rheumatoid arthritis. I. rheumatologists' opinions and the development of 'paper patients'. *Ann Rheum Dis*. 1983;42(6):644-647. doi:10.1136/ard.42.6.644

74. Kreuter F, Presser S, Tourangeau R. Social desirability bias in CATI, IVR, and web surveys: the effects of mode and question sensitivity. *Public Opin Q*. 2008;72(5):847-865. doi:10.1093/poq/nfn063

75. de Quidt J, Haushofer J, Roth C. Measuring and bounding experimenter demand. *Am Econ Rev*. 2018;108(11): 3266-3302. doi:10.1257/aer.20171330

SUPPLEMENT 1.

Trial Protocol

SUPPLEMENT 2.

eMethods 1. Randomization and Estimation Procedures
eMethods 2. Additional Design Details
eMethods 3. Covariate Balance
eMethods 4. Primary Outcomes
eMethods 5. Survey Measures of Racial Prejudice and Sexism
eMethods 6. BART Estimated Treatment Effects
eMethods 7. Secondary Outcomes
eFigure 1. Scenario Instructions 1 of 2
eFigure 2. Attention Check Drag and Drop (With Correct Responses Displayed)
eFigure 3. Scenario Instructions 2 of 2
eFigure 4. Patient Confidence Measure in Study 1
eFigure 5. Patient Satisfaction Measure in Study 1
eFigure 6. Patient Confidence Measure in Study 2
eFigure 7. Patient Satisfaction Measure in Study 2
eFigure 8. Warmth and Competence in Study 1
eFigure 9. Warmth and Competence in Study 2
eFigure 10. Fairness of Visit in Study 1
eFigure 11. Willingness to Publish Doctor Error in Study 2
eFigure 12. Example of Explicit Prejudice Survey Item Used in Qualtrics
eFigure 13. Distribution of Scores on Prejudice Items in Study 1
eFigure 14. Distribution of Scores on Prejudice Items in Study 2
eFigure 15. Distribution of Scores on Sexism Items in Study 1
eFigure 16. Distribution of Scores on Hostile Sexism Items in Study 2
eFigure 17. Distribution of Scores on Benevolent Sexism Items in Study 2
eTable 1. Background Characteristics by Treatment Group in Study 1
eTable 2. Background Characteristics by Treatment Group in Study 2
eTable 3. Randomization Inference (RI) for Covariate Balance
eTable 4. Average Patient Evaluation Scores on Composite Index of Primary Outcomes
eTable 5. Estimated Treatment Effects on Primary Outcomes in Combined Sample
eTable 6. Estimated Treatment Effects on Primary Outcomes in Study 1
eTable 7. Estimated Treatment Effects on Primary Outcomes in Study 2
eTable 8. Summary Statistics for BART-Estimated Treatment Effects on Composite Index (0-100)
eTable 9. Estimated Treatment Effects on Secondary Outcomes in Study 1
eTable 10. Estimated Treatment Effects on Secondary Outcomes in Study 2
eReferences.

SUPPLEMENT 3.

Data Sharing Statement