

The Doctor Will See You Now: Physician Race and Patient Evaluations*

Tristan L. Botelho
Yale School of Management

Qingyang (Iris) Wang
Yale School of Management

Tara Sanft
Yale School of Medicine

October 2026

Management Science, forthcoming

Abstract

Patient experience surveys (e.g., HCAHPS) are increasingly used to evaluate individual physicians, shaping performance reviews, compensation, and career advancement. However, their validity as unbiased measures of physician quality is uncertain. We examine whether and when these evaluations encode racial disparities using more than 700,000 patient surveys collected in a large Northeastern U.S. health system (2018-2021), covering 4,579 physicians across seven hospitals and 33 specialties. Asian and Black physicians are about 3% and 5% less likely, respectively, than White physicians to receive “top-box” ratings. These gaps persist after accounting for endogenous patient-physician matching. Mechanism tests show that disparities are largest among early-career physicians, with gaps widening to 17% for both Asian and Black physicians, and among those with less prestigious medical training, reaching 5% and 6%, respectively. Disparities narrow when evaluation items direct patients’ attention to concrete and verifiable behaviors (e.g., protocol adherence) or to dimensions for which clearer norms exist (e.g., basic interpersonal demeanor). Overall, the patterns suggest that the organizational design of the evaluation process meaningfully shapes the extent to which patients rely on status-based expectations when evaluating physicians. We outline design and policy implications for using patient experience scores in performance management while preserving the patient voice and limiting the potential to entrench racial inequality.

Keywords: careers, evaluations, healthcare, patients, physicians, stratification

* Direct correspondence to Tristan Botelho (tristan.botelho@yale.edu). Botelho and Wang serve as co-lead authors. Ashley Bazin, Donovan Brown, and Lauren Richard provided excellent research assistance. We would also like to thank Mark Iscoe, Marissa King, as well as seminar participants at Yale University and Wharton People and Organizations for their feedback and suggestions.

Introduction

Patient experience surveys have become embedded in the career infrastructure of American medicine. Originally designed to benchmark hospital-level quality, instruments derived from the Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) survey now serve another consequential purpose: evaluating individual physicians (Manary et al. 2013, Elliott et al. 2015, Indovina et al. 2021). In a growing number of health systems, patient satisfaction scores factor into performance reviews, compensation, and public reporting (CMS 2024). In a survey of state medical society members, 78% reported that patient satisfaction ratings affected their job experience and 59% reported effects on compensation (Zgierska et al. 2014). Physicians who achieve higher “top-box” scores—the share of patients selecting the highest possible rating—receive stronger internal evaluations and often higher pay (Ogrod 1997, Greene and Hibbard 2012). Because top-box is the modal response (HCAHPS 2025), even modest shortfalls at this benchmark can push physicians into the bottom performance quartile, with tangible consequences for bonuses, patient volume, and career progression (Savel et al. 2025).

The rationale for patient evaluations is clear: Patients directly experience communication, responsiveness, and care, making their perspectives an important component of healthcare assessment. However, patients are also non-expert evaluators who face substantial information asymmetries about what constitutes good care (Ogrod 1997, Manary et al. 2013, Anhang Price et al. 2014). Underscoring this tension, patient experience measures sometimes correlate with quality, but higher scores have also been associated with shorter waiting times, care center amenities, or receipt of a requested test or prescription (Doyle et al. 2013, Larkin and Pierce 2015, Biondi et al. 2016). The pertinent question is therefore not whether patient voice should matter, but how the design of evaluation processes determines which aspects of assessments it leaves open to interpretation and whether that interpretive latitude allows physician characteristics unrelated to care quality to shape evaluative outcomes.

Status characteristics theory predicts that when quality is difficult to observe, evaluators draw on widely shared beliefs about competence associated with ascriptive characteristics such as race to form performance expectations (Berger et al. 1972, 1980, Wagner and Berger 1993). Healthcare encounters

create precisely these conditions, as even medical experts struggle to evaluate physician quality consistently (Smith et al. 1997, Landon et al. 2003, Nembhard et al. 2009, Brewer et al. 2020). Whether such status beliefs translate into measured disparities, however, is not a foregone conclusion.

On the one hand, because patients often lack the domain expertise to assess whether a diagnosis was accurate, a treatment plan appropriate, or a procedure skillfully performed, physicians' professional authority may become especially salient, potentially dampening the influence of race (Abbott 1988, Olson et al. 2022). On the other hand, that authority may not operate in a race-neutral way. Because the medical profession has historically been overrepresented by White men (Kang and Kaplan 2019), patients may internalize a racialized prototype of the physician role (Leslie et al. 2015, Rivera and Tilcsik 2019), perceiving racial minority physicians as less competent, more atypical, or foreign (Zou and Cheryan 2017, Lett et al. 2019, Chow et al. 2024). Such perceptions may lower performance expectations and produce lower patient evaluations for racial minority physicians. Complicating matters further, because patient-physician encounters are not randomly formed, observed disparities may reflect both how patients evaluate physicians and which encounters occur in the first place. We thus ask whether racial disparities emerge in patient evaluations and, if so, under what conditions they attenuate.

Prior research suggests that the organizational design of evaluation processes matters, with a large body of research examining organizational and professional settings where evaluators possess substantial domain expertise (Dobbin et al. 2015, Botelho 2024, Pedulla et al. 2026). At the same time, organizations increasingly delegate evaluative authority to non-expert audiences, such as students, customers, and patients, whose evaluations can shape consequential outcomes for professional workers (Huckman and Kelley 2013, Flaherty 2025). These evaluators often face information asymmetries, lack an independent framework for assessing underlying quality, and are less directly subject to organizational oversight. These differences raise questions about whether organizational evaluation designs can similarly constrain status-based disparities among non-expert evaluators.

A growing stream of work has begun to address this question, showing that simple features of evaluation instruments, such as the number of rating categories, can constrain or amplify disparities

arising from status-based beliefs (Rivera and Tilcsik 2019, Botelho et al. 2025). Building on this work, we examine whether broader evaluation design can reduce such disparities beyond the structure of the rating scale. Specifically, we consider formalized evaluation criteria and individuating information about quality, two commonly used features that can help expert evaluators form more accurate and consistent assessments (Fiske and Neuberg 1990, Anderson and Tomaskovic-Devey 1995, Foschi 2000, Bodenhausen et al. 2012). Both are widely present in healthcare: Hospitals routinely surface physician credentials, such as years of experience and training background, through provider profiles and appointment portals, while survey instruments structure the criteria on which care is assessed.

We argue that precisely because patients lack an independent basis for assessing quality, cues and structures provided by the evaluation environment can offer alternative anchors for assessment that compete with race-based expectations. Credentials, such as greater experience or training from more prestigious programs, need not be perfect proxies for underlying quality to shift evaluation away from categorical status beliefs and toward more individuated assessments. By a similar logic, when evaluation criteria direct attention to concrete, verifiable behaviors, they help narrow the interpretive latitude that allows status beliefs to operate. Racial disparities in evaluations should therefore attenuate when organizational arrangements better structure how quality is perceived and assessed.

To test these predictions, we draw on more than 700,000 patient experience surveys from a major integrated health system in the Northeastern United States, collected from 2018 to 2021 across 33 specialties, 7 hospitals, and 357 clinical sites. The data cover 4,579 unique physicians (MDs) practicing in emergency, inpatient, and outpatient care settings. The surveys assess physicians' protocol adherence, basic interpersonal demeanor, engagement and communication, and patients' overall satisfaction, allowing us to examine whether and how patient evaluations vary by perceived physician race.

We find sizeable and policy-relevant racial disparities. Asian and Black physicians are 3% and 5% (or 2.2 and 3.8 percentage points) less likely than White physicians to receive top-box evaluations, after accounting for patient characteristics, care contexts, and physician attributes. To address endogenous matching between patients and physicians, we conceptualize patient evaluations of physicians as a two-

stage process and construct counterfactual patient-physician interactions. We then apply two complementary approaches that adjust for sorting on observed and unobserved factors (Heckman 1976, Rubin 1990, Wooldridge 2010, Azoulay et al. 2017). The resulting estimates remain close to the baseline results, suggesting that racial disparities in patient evaluations persist beyond what can be explained by endogenous matching. Robustness checks using placebo outcomes, alternative specifications, matched-sample analysis, and sensitivity analysis further yield convergent findings.

Disparities in patient evaluations of Asian and Black physicians are concentrated among early-career physicians and those without prestigious medical school credentials. Newly practicing Asian and Black physicians are both 17% less likely than White physicians to receive top-box evaluations, corresponding to gaps of 11.4 and 11.9 percentage points, respectively. These gaps narrow by 0.7 percentage points per year of experience, although estimates for Black physicians are less precise given their smaller sample size. Among physicians with less prestigious medical training, Asian and Black physicians are 5% and 6% less likely than White physicians to receive top-box evaluations, whereas the gaps are small and statistically insignificant among top 50 U.S. medical school graduates. Notably, physician credentials are more strongly associated with racial gaps than with overall evaluation levels, suggesting that even low-information cues can compete with race-based status beliefs to shape how patients evaluate under uncertainty. Disparities are also larger when patients assess more subjective dimensions (e.g., overall assessment of physicians' skill), and narrower on dimensions that are verifiable (e.g., protocol adherence) or governed by clearer evaluative norms (e.g., basic interpersonal demeanor), allowing patients to use general knowledge rather than healthcare-specific expertise. Together, these findings suggest that both the information environment organizations create around physicians and the evaluative criteria they ask patients to apply can shape when and how racial disparities emerge.

This study documents consequential racial disparities in patient evaluations of physicians. It connects status characteristics theory to research on the design and structure of evaluation processes by showing that institutionally supplied information and evaluation criteria can constrain status-based beliefs when evaluators lack domain expertise and quality is difficult to observe (Anderson and Tomaskovic-

Devey 1995, Foschi 2000). It extends the growing literature on the structure of evaluation processes by showing that, beyond numerical rating scales, evaluation instruments themselves can shape how evaluators form and apply criteria (Rivera and Tilcsik 2019, Abraham et al. 2024, Botelho et al. 2025). Finally, it advances status research by showing that institutionally supplied cues, such as experience and educational credentials, can attenuate the relationship between race and evaluative outcomes even when they are imperfect proxies for quality (Fiske and Neuberg 1990, Bodenhausen et al. 2012).

Background and Theory

Patient Experience Surveys and Their Implications for Physicians' Careers

Since the 1980s, patient-centered care has become an increasingly influential principle in American healthcare, redefining high-quality care to include not only clinical outcomes but also whether care is “respectful of and responsive to individual patient preferences, needs, and values” (IOM 2001, p. 6). Operationalizing this principle required systematic tools to measure patient satisfaction. Patient experience surveys emerged for this purpose. In 2006, the Centers for Medicare & Medicaid Services (CMS) and the Agency for Healthcare Research and Quality (AHRQ) introduced the HCAHPS survey to provide standardized, comparable measures of patient experience on dimensions such as communication, responsiveness, and respect (Siegrist Jr 2013, AHRQ 2023, CMS 2024).

These measures were quickly incorporated by payers and administrators into healthcare accountability systems, most notably through CMS's Hospital Value-Based Purchasing program, which links a portion of Medicare reimbursement to patient experience survey results (CMS 2024). Because these results are typically summarized as top-box scores—the share of patients selecting the highest possible rating—top-box performance becomes a central organizational target, while any rating short of the highest is treated as a potential shortfall (Savel et al. 2025).

As institutional pressures to monitor and improve patient experience intensify, hospitals and health systems have increasingly used patient experience surveys not only for organizational benchmarking but also for evaluating individual physicians (Manary et al. 2013, Elliott et al. 2015, Indovina et al. 2021). These measures have thus become embedded in the career infrastructure of the

medical profession, with physicians' career outcomes growing more closely tied to their survey performance. Physicians who achieve higher top-box scores receive stronger internal assessments and, in many cases, higher compensation (Ogrod 1997, Greene and Hibbard 2012). For example, a survey of state medical society members found that patient satisfaction ratings affected job experience for 78% of physician respondents and compensation for 59% (Zgierska et al. 2014).

This high-stakes use of patient experience surveys rests on the assumption that they accurately capture physician quality, an assumption that is far from straightforward. Patients often lack the expertise to assess whether care was appropriate, making it possible for their evaluations to reflect noise unrelated to care quality (Kupfer and Bond 2012, Richman and Schulman 2022). Consistent with this concern, evidence suggests that these surveys can be shaped by patient expectations, care center amenities, or receipt of requested tests and prescriptions, independent of actual care quality (Biondi et al. 2016).

Physician Demographics and Disparities in Patient Evaluations

Of particular concern is whether evaluation outcomes reflect not only noisy behavioral signals, but also physicians' ascriptive characteristics, such as race. Status characteristics theory predicts that when quality is difficult to observe, evaluators draw on widely shared beliefs and stereotypes that link ascriptive characteristics to competence when forming performance expectations (Berger et al. 1972, 1980, Wagner and Berger 1993). Healthcare settings create precisely these conditions: Because care quality is difficult to assess consistently (Smith et al. 1997, Landon et al. 2003, Nembhard et al. 2009, Brewer et al. 2020), patients may rely on race-based status beliefs when interpreting physician performance. Whether such status beliefs translate into measured disparities, however, is not a foregone conclusion.

Patients and physicians differ in domain expertise, as patients often cannot independently assess whether a diagnosis was accurate, a treatment plan appropriate, or a procedure skillfully delivered. This information asymmetry may make physicians' professional authority especially salient, leading patients to defer to physicians' expertise rather than rely on race in evaluations (Abbott 1988, Olson et al. 2022).

For such deference to suppress racial disparities, however, professional authority would need to operate in a race-neutral way. Patients would have to extend comparable deference to physicians'

expertise regardless of whether they are interacting with an Asian, a Black, or a White physician. This assumption is difficult to sustain in medicine, a profession historically overrepresented by White men (Kang and Kaplan 2019). This historical skew may shape patients' implicit prototype of the physician role in ways that are tied to race, such that physicians who deviate from it face a penalty that their professional identity alone cannot fully offset (Leslie et al. 2015, Rivera and Tilcsik 2019).

Race may therefore continue to operate as a basis for evaluation even within a professional context. In particular, Black Americans are often perceived as less competent, whereas Asian Americans are positioned as culturally foreign (Zou and Cheryan 2017). In medicine, Black physicians remain substantially underrepresented relative to their population share, while Asian physicians' numerical representation does not eliminate perceptions of atypicality or foreignness (Lett et al. 2019, Chow et al. 2024). Both forms of departure from the prototype may contribute to status-based expectations among patients and produce less favorable evaluations independent of care quality.

H1: Racial minority physicians will be less likely to receive top-box patient satisfaction scores than their White counterparts.

Complicating matters further, patient-physician encounters are often endogenous, such that any observed disparities may reflect both how patients evaluate physicians and which encounters occur in the first place. Patients who are strongly influenced by these racialized prototypes may opt out of visiting racial minority physicians altogether, thereby attenuating observed disparities (Greene et al. 2018, de Vaan and Stuart 2022, Linde and Egede 2022). Meanwhile, racial minority physicians may be evaluated by systematically different patient pools with a lower propensity to provide top-box responses, amplifying observed disparities (Van Ryn and Burke 2000, Haviland et al. 2005). We thus examine whether patient experience surveys encode racial disparities and the extent to which those disparities persist after accounting for endogenous matching between patients and physicians.

Organizational Design of Evaluation Processes and Racial Disparities

If patient experience surveys are susceptible to the influence of race-based expectations, a key question is whether the organizational arrangements through which these evaluations are administered shape the

extent to which racial disparities emerge. Research on evaluation processes shows that status-based disparities intensify when ambiguity about how to evaluate is high, leaving evaluators to rely on their own discretion in determining what constitutes good performance and how different aspects of it should be weighed (see Abraham et al. 2024 for a review). Such discretion leaves room for racial stereotypes to shape evaluations (Berger et al. 1972, 1980, Wagner and Berger 1993).

Organizations may limit such discretion by structuring evaluation in ways that steer evaluators away from heuristic processing shaped by salient status characteristics, such as race, and toward more deliberate processing based on individuating information or concrete behavioral cues (Fiske and Neuberg 1990, Anderson and Tomaskovic-Devey 1995, Foschi 2000, Bodenhausen et al. 2012). Prior research provides empirical support for this argument: Experienced investors are less likely to rely on ascriptive characteristics when more pertinent information about quality is available (Botelho and Abraham 2017). Hiring managers and supervisors can leverage clearer evaluation criteria to produce more consistent assessments that are less susceptible to status-based bias (Castilla 2008, Dobbin et al. 2015). A substantial body of this evidence examines settings in which evaluators possess domain expertise and already have an interpretive framework for what constitutes quality (Li 2017, Botelho 2024).

At the same time, inputs from non-expert evaluators are increasingly consequential as organizations delegate evaluative authority to students, customers, and patients, whose assessments shape outcomes for professional workers (Huckman and Kelley 2013, Flaherty 2025). Whether organizational structures can constrain status-based disparities among these evaluators is less clear. Patients, for example, may recognize that more prestigious training is favorable while lacking a basis for calibrating how much such credentials should matter. From this perspective, simply structuring the evaluation process need not reduce reliance on status-based expectations.

A growing literature nevertheless shows that evaluation design can matter among non-expert audiences. Features such as the number of rating categories can constrain or amplify disparities arising from status-based beliefs (Rivera and Tilcsik 2019, Botelho et al. 2025). We extend this work beyond rating-scale structure to consider two broader features through which organizations shape evaluation: the

individuating information they provide and the criteria they ask evaluators to apply, both of which are common in healthcare. Hospital-hosted provider profiles and appointment portals routinely provide information about physicians' credentials, such as years of experience, training background, and board certifications, while patient survey instruments indicate which aspects of care matter for evaluation.

We argue that these features can shape the bases on which patients form evaluations even when they lack the expertise to independently determine whether a diagnosis was accurate or a treatment plan appropriate. Institutionally supplied cues, such as greater experience or training from more prestigious programs, may provide physician-specific information that enters into patients' assessments alongside salient social categories, such as race, thereby directing attention away from categorical status beliefs. These credentials need not be accurate proxies for care quality, nor do patients need to know how strongly they predict quality. Rather, by providing an alternative anchor for evaluation, such information may reduce the extent to which status beliefs operate.

Likewise, evaluation criteria may direct patients' attention to concrete, verifiable behaviors or to dimensions of care governed by clear norms, namely, broadly shared standards that patients can apply using general knowledge rather than healthcare-specific expertise. Such criteria may help narrow the interpretive latitude that allows status beliefs to operate. A patient with little basis for assessing what makes a skilled cardiologist may still report whether the physician explained a procedure or discussed its risks beforehand. By contrast, ambiguous evaluation items (e.g., overall assessment of a physician's skill or communication style) leave more room for patients to rely on their own interpretations of what makes a physician skilled or a good communicator, thereby increasing reliance on race-based expectations.

H2a: Racial disparities in the likelihood of receiving top-box patient satisfaction scores will be smaller when physicians possess stronger institutionally supplied cues of quality, such as greater experience or more prestigious training.

H2b: Racial disparities in the likelihood of receiving top-box patient satisfaction scores will be smaller for evaluation items that focus on concrete, verifiable behaviors or on dimensions of care governed by clearer norms.

Empirical Context

This study draws on proprietary data from a large health system in the Northeastern United States. The system delivers comprehensive healthcare services, including emergency, inpatient, and outpatient care, across seven major hospitals, 357 clinical sites, and 33 specialties. Between 2018 and 2021, it averaged approximately 130,000 annual inpatient discharges and 3.5 million outpatient encounters.

Data

The health system routinely collects patient experience data through surveys administered by Press Ganey. We obtained all patient survey responses from January 2018 to December 2021, yielding 711,553 observations. These responses capture approximately 20% of inpatient and 4% of outpatient encounters, with response rates consistent with Press Ganey's nationwide averages (Tyser et al. 2016).¹ Our unit of analysis is the patient visit, which corresponds to the level at which surveys are administered. Because most patients were surveyed only once (mean = 1.1 surveys per patient), results are substantively similar whether we analyze all patient visits or restrict the sample to unique patients (see *Robustness Checks*).

The survey instrument is based on HCAHPS and asks patients to evaluate both their overall experience and specific aspects of the visit, including hospital facilities, physicians, nurses, and other care elements. In addition, the survey includes patient demographic and socioeconomic information as well as clinical visit details, drawing on both patient self-reports and hospital administrative data.

We focus on the section dedicated to physician evaluations, which contains 5 to 11 questions depending on the care setting (see *Online Appendix A* for details).² These items primarily assess physicians' adherence to clinical protocols, basic interpersonal demeanor, and engagement and communication, as well as patients' overall assessment of the physician. To understand whether and how physician race is associated with patient responses, we linked these observations to physician-level data, including physician gender, perceived race (see *Measures* section for details), education, years in practice,

¹ The outpatient survey capture rate is lower because the health system did not survey every patient to prevent survey fatigue. Outpatient surveys are typically administered by email, which also yields lower response rates.

² Online Appendix: https://osf.io/esjvx/overview?view_only=7f76c24e405148aaa4dc125c79216974

and other National Provider Identifier (NPI) records. We excluded 137,706 observations (19% of all observations) in which patients were treated by non-MD clinicians or the attending physician was not recorded. The resulting sample includes 573,847 observations, of which 8%, 17%, and 75% are from emergency medicine, inpatient, and outpatient care settings, respectively.

Measures

Physician Evaluation Outcomes. The surveys employed a 5-point scale for each evaluation item, ranging from very poor (1), poor (2), fair (3), good (4) to very good (5). The health system averaged item-level scores to produce an overall physician evaluation score and rescaled it to a 100-point index, with 100 as the best possible outcome. Missing items were excluded when calculating the overall score. Consistent with conventions used by hospitals and administrators (HCAHPS 2025, Savel et al. 2025), we constructed a binary variable, *Top Box*, indicating whether the overall score equaled 100 (coded as 1) or not (0).

Physician Race. We measure physicians' perceived race, not self-identified race. Patient experience is most likely shaped by perceptions in the clinical visit, as physicians rarely self-disclose race and ethnicity during a visit. If patients' evaluations vary by physician race, perceived race is the construct that aligns with what patients observe, and thus the data generation process for these ratings. We measured perceived race using physicians' headshot photographs. Two research assistants independently reviewed each physician's photo from the health system's website or other public sources and coded physicians into one of four categories: Asian, Black, White, or Other (including Hispanic/Latino, Middle Eastern, other, or unknown). Interrater agreement was high (Cohen's $\kappa = 0.76$). For disagreements, the research team independently reviewed the physician's profile and determined the perceived race.

Control Variables. Physicians across racial groups may vary along multiple dimensions. To better estimate racial disparities in physician evaluations, we included the following control variables. At the physician level, we controlled for gender and years in practice. We also categorized physicians' medical education into three groups based on the five-year average (2016-2020) *U.S. News & World Report* rankings: (i) top 50 U.S. medical schools, (ii) other U.S. medical schools, and (iii) non-U.S. schools. At the patient level, we controlled for gender, race, age, and insurance type (Medicaid vs. non-Medicaid).

We also included fixed effects for hospital and care setting (emergency, inpatient, or outpatient care) to absorb differences across evaluation environments.

Summary Statistics

Of the 573,847 observations in our sample, 104,685 (18% of the sample) were missing an overall physician evaluation score, suggesting that patients took the survey but left all physician evaluation items blank. We found no consistent evidence that these missing-data patterns varied by physician race (see *Robustness Checks*). Among visits with a recorded score, 73% were *Top Box* responses (i.e., score = 100; see Figure 1). This distribution aligns with national benchmarks, where scores below *Top Box* typically place physicians in the bottom quartile of performance (HCAHPS 2025). The sample includes 916 Asian (20.0% of all physicians), 160 Black (3.5%), 2,995 White (65.4%), and 508 Other (11.1%) physicians. This racial composition aligns with that of the U.S. physician workforce.³ Because over half of physicians categorized as Other had unknown race, we focused our comparisons on Asian, Black, and White physicians. During the study period, these groups received an average of 102, 128, and 140 evaluations per physician, respectively (see Table 1 for summary statistics and pairwise correlations).

[Figure 1]

[Table 1]

Empirical Strategy

Data Generation Process, Endogenous Matching, and Sample Selection Bias

Patient evaluations of physicians emerge from a two-stage process. First, patients seeking care face a choice set of physicians shaped by clinical needs, insurance coverage, geographic proximity, and other factors. Each patient ultimately receives care from a single treating physician, while alternative options remain unobserved. Second, patients are invited to complete post-visit surveys evaluating their treating physician. The resulting data thus include evaluations only from realized patient-physician interactions.⁴

³ See the U.S. Physician Workforce Data Dashboard from the Association of American Medical Colleges (AAMC).

⁴ This two-stage model simplifies a more complex reality in which patient-physician relationships vary in continuity: Some visits are one-off, whereas others occur within ongoing relationships. In our data, 17% of visits are reported to

This two-stage process means that observed evaluations reflect both who patients see and how they evaluate those physicians, making endogenous matching and the resulting sample selection bias the key identification challenge. For example, patients' race and socioeconomic status may simultaneously influence physician choice and evaluation outcomes (Van Ryn and Burke 2000, Haviland et al. 2005). Below, we outline our empirical strategies to address this challenge.

Counterfactual Patient-Physician Interactions and Simulated Choice-Set Sample

To model the selection process more explicitly, we approximated the unobserved alternatives in patients' choice sets by generating counterfactual patient-physician interactions. Specifically, for each patient visit in our sample, we enumerated all physicians active in the same medical specialty (e.g., cardiology) and care setting (e.g., inpatient care) across the health system in the visit year. Medical specialties are defined using the NPI taxonomy and were reviewed by a physician. For computational efficiency, we then randomly sampled 10 physicians from this set (excluding the treating physician) to form counterfactual patient-physician interactions for each realized visit. When fewer than 10 physicians were active in the same specialty, care setting, and visit year, all eligible physicians were included. The realized and counterfactual interactions together form a simulated choice-set sample, which we use to model realized patient-physician interactions (see *Online Appendix B* for summary statistics).

Baseline Linear Probability Models

As a baseline approach, we examined racial disparities in the likelihood of receiving a *Top Box* response with linear probability models (LPMs) of the following form:

$$Y_{ij} = \text{PhyRace}_i + \text{PhyControls}_i + \text{PatControls}_j + \text{Visit}_{ij} + \epsilon_{ij} \quad (1)$$

where i indexes physicians and j indexes patients. Y_{ij} equals one if patient j provided a *Top Box* response for physician i , and zero otherwise. PhyRace_i denotes the perceived race of physician i .

PhyControls_i and PatControls_j are vectors of physician and patient characteristics (see *Control*

be with a regular physician. Further analyses suggest that such relationships moderate racial disparities but do not eliminate them, and our results remain consistent when recurring interactions are excluded (see *Online Appendix E*).

Variables section). Visit_{ij} represents hospital and care-setting fixed effects that absorb differences across evaluation environments. As robustness checks, we also estimated specifications with medical-specialty fixed effects to further account for differences in clinical context (see *Robustness Checks*). These estimates describe racial disparities conditional on realized patient-physician interactions. Because these interactions are likely non-random, estimated differences reflect both disparities in evaluations and systematic differences in patient composition by physician race.

Inverse Probability of Treatment Weighting

To address endogenous patient-physician matching, an ideal test would compare evaluations where patients were randomly assigned to physicians of different races. In practice, such assignment is rare, and when it is most likely to occur, such as in ambulance admissions to emergency departments, interactions are often superficial and time-constrained, unlike routine interactions with general practitioners or specialists that constitute most clinical encounters. As a proxy, we used inverse probability of treatment weighting (IPTW) to construct a pseudo-population in which observed patient and physician characteristics were balanced across physician race groups, reducing selection bias from matching.

This approach assumes that patient-physician matching is determined *exclusively* by observable characteristics, such that conditioning on these determinants addresses selection bias and allows us to estimate the average treatment effect (Rubin 1990, Azoulay et al. 2017). In our context, this implies that once observable drivers of realized interactions are accounted for, patients treated by physicians of different races would have had similar expected evaluation tendencies had they been assigned to another physician within the choice set. Although this assumption is not directly testable, it is plausible in our setting: Our data contain detailed measures of patient and physician characteristics that prior research identifies as key drivers of physician choice (Greene et al. 2018, Linde and Egede 2022), and patients typically face constrained choice sets shaped by clinical need, insurance networks, and scheduling, limiting sorting on unobserved preferences (Harris 2003).

To implement IPTW, we estimated the conditional probability that a patient-physician interaction was realized using a probit model within specialty-year cells:

$$P(D_{ij} = 1|X_{ij}) = \Phi(X_{ij} + u_{sy}) \quad (2)$$

where D_{ij} indicates whether physician i treated patient j ($D_{ij} = 1$) or not ($D_{ij} = 0$), X_{ij} includes patient and physician characteristics as well as geographic proximity, and u_{sy} represents specialty-year fixed effects. To avoid extreme weights, predicted probabilities were winsorized at the 2.5% and 97.5% levels.

We constructed inverse probability of treatment weights for each patient-physician interaction as:

$$w_{ij} = \frac{\sum_{k \in T_j} \hat{P}(D_{kj} = 1|X_{kj})}{\hat{P}(D_{ij} = 1|X_{ij})} \quad (3)$$

where T_j represents the choice set for patient j . The numerator normalizes the weights within each choice set. We re-estimated Equation 1 using weighted least squares. Intuitively, this assigns lower weight to more likely interactions and higher weight to less likely ones, thereby reducing bias from observable sorting (see *Online Appendix B* for details on the selection-stage model and weight construction).

Heckman Correction

Although the IPTW estimates address selection on observables, it remains possible that patients and physicians are matched on unobserved factors that also influence evaluation outcomes. As a complementary approach, we implemented a Heckman correction model (Heckman 1976, Wooldridge 2010). Compared to the classic missing-outcome model, the key challenge in our setting is that the observed patient-physician interaction is itself endogenously formed. Accordingly, we interpret the Heckman model as a control-function approach that allows unobserved determinants of patient-physician interactions to be correlated with unobserved determinants of evaluations (Wooldridge 2015).

The Heckman framework requires an exclusion restriction: at least one variable that affects realized patient-physician interactions but does not directly affect evaluation outcomes, conditional on other covariates. We used geographic proximity as the exclusion restriction, measured as the logarithm of distance between the patient's residential ZIP code and the hospital or clinic ZIP code.

Distance plausibly affects realized patient-physician interactions because geographic proximity can factor into patients' physician choices (Keating et al. 2016, Smith et al. 2018). This is especially

relevant in our setting because all physicians belong to the same health system, limiting cross-institutional quality differences that might be correlated with distance. For over 99% of visits in our sample, patients could have been seen by feasible physicians, namely those within the same medical specialty, care setting, and year, in multiple locations, providing meaningful variation in distance (see *Online Appendix B*).

The importance of proximity may nevertheless vary across contexts: Descriptive patterns show that patients travel farther for specialist visits than for emergency or generalist care (*Online Appendix B*). We therefore focus our comparisons within similar contexts by controlling for hospital and care-setting fixed effects in the main models and medical-specialty fixed effects in robustness checks (see *Robustness Checks*). Meanwhile, distance should not directly affect evaluations. While this assumption is not directly testable, it is plausible conditional on physician, hospital, and patient characteristics; we discuss the potential threats to the exclusion restriction and robustness checks in *Online Appendix B*.

We estimated the selection equation (Equation 2) and computed the inverse Mills ratio (IMR_{ij}) for each realized patient-physician interaction. We then included the IMR_{ij} in the outcome equation:

$$Y_{ij} = \text{PhyRace}_i + \text{PhyControls}_i + \text{PatControls}_j + \text{Visit}_{ij} + IMR_{ij} + v_{ij} \quad (4)$$

Intuitively, the inverse Mills ratio captures the part of the evaluation outcome that is systematically related to the selection process. Including it adjusts for the possibility that unobserved factors simultaneously influence both realized patient-physician interactions and patient evaluations.

Although IPTW estimates and the Heckman correction cannot definitively recover causal estimates given the challenges of endogenous matching, they allow us to assess the robustness of the estimated racial disparities under different assumptions about the selection process. We discuss additional approaches to validate our findings in the *Robustness Checks* section.

Results

Descriptive Evidence on Racial Disparities and Patient-Physician Sorting

Before turning to our empirical adjustments, we report unadjusted evaluation outcomes by physician race. As shown in Table 1, 70%, 67%, and 74% of clinical visits resulted in *Top Box* responses for Asian,

Black, and White physicians, respectively. Compared to White physicians, Asian and Black physicians were 4.5 and 7.0 percentage points, or 6% and 9%, less likely to receive *Top Box* responses. These disparities correspond to 6% and 10% of the mean *Top Box* rate, and 10% and 16% of its standard deviation (mean = 73.03%; S.D. = 0.444). Consistent with H1, descriptive results suggest that racial minority physicians were less likely to receive *Top Box* responses than their White counterparts.

To better understand the identification challenge from endogenous matching, we examined whether patient-physician matching varied by race. Figure 2 plots the unadjusted probability that a patient-physician interaction was realized within the simulated choice-set sample. Racial minority patients were more likely to visit physicians of their own race. Asian patients had an 11.8% probability of visiting an Asian physician versus 8.6% for a White physician; Black patients had a 17.9% probability of visiting a Black physician versus 8.8% for a White physician. As a placebo test, we repeated this analysis among patients admitted via ambulance, where discretionary choice should be more limited, and found no comparable racial sorting patterns (Figure B2, *Online Appendix*). This evidence indicates non-random sorting along racial lines, especially where patients have greater discretion in physician choice.

[Figure 2]

Conditional on such sorting, descriptive patterns also suggest that racial disparities in patient evaluations are attenuated in same-race patient-physician pairs. Table 2 reports the evaluation outcomes by physician and patient race. For example, Asian patients gave *Top Box* responses at a rate of 67.6% when visiting Asian physicians, compared to 66.2% for White physicians and 63.3% for Black physicians; Black patients provided *Top Box* ratings at a rate of 66.4% for Black physicians, compared to 63.6% for White physicians and 61.1% for Asian physicians.

[Table 2]

This pattern does not imply that endogenous matching simply attenuates observed disparities for racial minority physicians. Rather, it may also alter physicians' exposure to patients with different socioeconomic characteristics and baseline propensities to give *Top Box* ratings. For example, 7.6% of Asian patients and 16.8% of Black patients in our sample used Medicaid or Medicaid Managed Care as

their primary insurance, compared to 3.4% of White patients. Among patients reporting residential ZIP codes, 23.4% of Asian patients and 66.6% of Black patients lived in ZIP codes in the bottom quartile by median household income, relative to 17.7% of White patients. Among those reporting education, 3.4% of Asian patients and 2.8% of Black patients did not complete high school, compared to 1.4% of White patients. Reflecting these socioeconomic differences, racial minority patients were less likely to give *Top Box* responses: Asian and Black patients had *Top Box* response rates of 66.7% and 63.2%, respectively, compared to 74.1% among White patients (Table 2). Together, these patterns complicate the interpretation of racial disparities in physician ratings and motivate the empirical adjustments that follow.

Physician Race and Top Box Responses: Baseline Linear Probability Model

To estimate racial disparities more precisely, we modeled the probability of receiving a *Top Box* response as a function of physician race using a linear probability model (Equation 1). The model was estimated on observed survey responses and controlled for physician characteristics (gender, experience, education), patient characteristics (gender, race, age, insurance), and hospital and care-setting fixed effects.

As shown in Model 1 (Table 3), Asian and Black physicians were less likely to receive *Top Box* responses than White physicians. Specifically, the estimated probability of a *Top Box* evaluation was 71.3% for Asian physicians, 69.7% for Black physicians, and 73.5% for White physicians. Compared to White physicians, Asian and Black physicians were 2.2 ($p < 0.01$) and 3.8 ($p < 0.05$) percentage points less likely to obtain *Top Box* responses, representing relative reductions of 3% and 5%. These gaps corresponded to 3% and 5% of the mean *Top Box* rate and 5% and 9% of its standard deviation (mean = 73.03%; S.D. = 0.444). The adjusted racial gaps were directionally consistent with the descriptive results and approximately half the magnitude of the unadjusted disparities.

[Table 3]

Although modest in absolute terms, these racial gaps were comparable to or larger than the effects of several other physician characteristics in the model. For example, the Asian-White gap (-2.2 percentage points) was equivalent to approximately three to four additional years of practice experience (0.6 percentage points per year) and two-thirds of the difference between graduates of non-U.S. medical

schools and those of top 50 U.S. medical schools (-3.1 percentage points). Similarly, the Black-White gap (-3.8 percentage points) corresponded to six years of experience and exceeded the magnitude of the medical school difference. Both racial gaps also exceeded the estimated gender difference, which was substantively small and only marginally significant (0.9 percentage points; $p = 0.06$). The gaps were similarly sizable relative to patient characteristics: A White physician would need to increase the share of racial minority patients by approximately 25 percentage points, or the share of Medicaid patients by roughly 50 percentage points, to match the gap faced by Asian physicians. Offsetting the gap faced by Black physicians would require shifts exceeding 40 and 80 percentage points, respectively.

Beyond statistical significance, these racial disparities carry meaningful economic consequences. Because patient satisfaction scores are included in high-stakes decisions—such as physician performance reviews and compensation—and *Top Box* is the modal outcome, even small shortfalls at this benchmark can accumulate into tangible disadvantages, including lower bonuses, reduced patient volume, and increased burnout from unfavorable social comparisons (Zgierska et al. 2014, Bond et al. 2023).

Addressing Sorting on Observables and Unobservables

To address the non-random matching between physicians and patients, we adjusted the baseline estimates using two complementary approaches. First, under the assumption that sorting operated *solely* on observed characteristics, we constructed inverse probability of treatment weights based on the simulated choice-set sample to reweight realized patient-physician interactions (Rubin 1990, Azoulay et al. 2017). Second, relaxing this assumption, we implemented a Heckman correction using geographic proximity as the exclusion restriction (Heckman 1976, Wooldridge 2010).

In both approaches, the selection stage used a probit model to predict the realization of patient-physician interactions. For each patient-physician interaction, we included physician and patient race and gender, along with their interactions, physician credentials (experience and education), patient characteristics (insurance and age), geographic proximity, and specialty-year fixed effects. The results show that, relative to all other physicians active in the same specialty and visit year, patients were more likely to visit physicians who were geographically closer and who shared their gender and race. More

experienced physicians and those who obtained their medical degrees in the U.S. were also more likely to be selected (see Table B2, *Online Appendix*). Together, these results provide evidence of non-random sorting into realized visits and corroborate the descriptive results.

Models 2 and 3 report the estimated racial gaps after incorporating the selection stage using the IPTW and Heckman correction approaches, respectively, and both closely mirror the baseline estimates. After accounting for sorting on observables, the corresponding gaps were 2.5 ($p < 0.01$) and 3.4 ($p < 0.05$) percentage points for Asian and Black physicians. After further addressing sorting on unobservables, they were 2.3 ($p < 0.05$) and 2.8 ($p = 0.09$) percentage points. While precision decreased for the Black-White gap, consistent with the smaller sample of Black physicians, the estimates remained substantively similar. Relative to the baseline model, IPTW and Heckman correction had little effect on the Asian-White gap and reduced the Black-White gap by 0.4 and 1.0 percentage points, respectively.

We replicated the IPTW and Heckman analyses using alternative specifications. As a robustness check for the Heckman correction, we added controls for hospital fixed effects and the median household income of patients' residential ZIP codes to address a potential concern with the exclusion restriction: Distance may partly proxy for institutional and socioeconomic differences that shape both selection and evaluation. These analyses yielded substantively similar results (*Online Appendix B*). Together, they show that although patient-physician matching is endogenous, sorting on both observables and unobservables does not explain away the racial disparities in *Top Box* responses. The persistence of disparities across models suggests that racial minority physicians face systematic disadvantages in patient evaluations.

Physician Experience, Education, and Racial Disparities

Having established that patients are less likely to give Asian and Black physicians *Top Box* responses, we next explore whether these racial disparities vary with the strength of institutionally supplied cues of physician quality. Healthcare organizations often provide information about physicians' years in practice, training background, and board certification through provider profiles and appointment portals. While these are imperfect proxies for underlying quality, prior research suggests patients consult them when choosing physicians (Kozikowski et al. 2022). Such cues may also be inferred from age or credentials

displayed in clinical settings even when patients do not actively seek them out. We predict that stronger cues, such as greater experience or more prestigious training, will reduce patients' reliance on race-based expectations and attenuate racial disparities in ratings.

We first considered physician experience as a cue that may reduce the influence of race on patient evaluations. We measured physician experience using *Years in Practice*, proxied by years since physicians' first NPI registration. In our sample, racial minority physicians had fewer years of practice than White physicians. On average, Asian, Black, and White physicians began practicing in 2010, 2009, and 2008, respectively (see Table 1). We introduced *Years in Practice* as a moderator in Equation 1 and re-estimated the relationship between physician race and *Top Box* responses. Figure 3 presents the results.

[Figure 3]

Consistent with our expectations, racial gaps narrowed as physicians gained experience. Among physicians with zero years of experience, Asian and Black physicians were 11.4 ($p < 0.001$) and 11.9 ($p < 0.05$) percentage points (or approximately 17% in both cases) less likely to receive *Top Box* responses relative to White physicians. These gaps narrowed by 0.7 percentage points for Asian and Black physicians for each additional year of experience ($p < 0.001$ and $p = 0.14$, respectively), although the estimated gradient for Black physicians became less precise given the smaller sample size. For physicians with 15 years in practice, the racial gaps were negligible (Table C1, *Online Appendix*).

By a similar logic, we also examined physicians' educational prestige, measured by their MD-granting institution. Specifically, we calculated each school's average *U.S. News & World Report* ranking over 2016-2020 and created a binary indicator, *Educational Prestige*, equal to one if the physician graduated from a top 50 U.S. medical school, a cutoff often used to distinguish more prestigious institutions (Gao et al. 2012). Physicians who attended schools ranked outside the top 50, as well as international institutions that are likely less familiar to the average patient in a U.S. context, were coded as zero. In our sample, Asian and Black physicians were less likely than White physicians to have graduated from a top 50 U.S. medical school: 36% of Asian physicians, 40% of Black physicians, and 50% of White physicians graduated from a top 50 U.S. medical school (Table 1).

We introduced *Educational Prestige* as a moderator in the baseline model (Equation 1) and presented the results in Figure 4. Among physicians who did not graduate from a top 50 U.S. medical school, Asian and Black physicians were 3.4 ($p < 0.001$) and 4.7 ($p < 0.05$) percentage points less likely, respectively, to receive *Top Box* responses than White physicians, equivalent to gaps of 5% and 6%. By contrast, among physicians who graduated from a top 50 U.S. medical school, these gaps were more modest in magnitude, at 1.9 and 2.2 percentage points, and statistically indistinguishable from zero. More formally, physician race was significantly associated with *Top Box* responses among physicians without top 50 medical school training, but its interactions with *Educational Prestige* were not, suggesting that the racial gaps emerged mainly among physicians without prestigious training, although the difference in racial gaps across the two groups did not itself reach statistical significance (Table C1, *Online Appendix*).

[Figure 4]

Three features make these results noteworthy. First, they contextualize the average effects in Table 3. Although the overall average gaps were modest—Asian and Black physicians were 2.2-3.8 percentage points less likely than White peers to receive a *Top Box* rating—racial disparities were substantially larger among early-career physicians and among those without prestigious educational credentials, providing evidence consistent with the moderating role of institutionally supplied cues. For example, at career entry, Asian and Black physicians faced penalties of 11.4 and 11.9 percentage points, respectively—or three to five times the average gaps—which only became negligible around year 15. Similarly, racial disparities were concentrated among physicians without prestigious medical-school training and were substantially attenuated among those with such credentials.

Second, these patterns point to structural disadvantages that both reflect and reinforce existing racial inequality in the medical profession. Credentials such as educational prestige may themselves be unequally distributed, with racial minority physicians facing more constrained access to them (Boatright et al. 2017, Nguyen et al. 2022). The fact that such credentials moderate racial gaps even within the same health system, where all physicians have already cleared a common selection bar, suggests that the cues attenuating disparities are themselves embedded in broader structures of inequality. Further, larger gaps

among early-career physicians mean that racial minority physicians endure years of systematically lower evaluations before those disparities narrow, and these early shortfalls may compound into persistent differences in pay, recognition, and opportunities even if evaluation gaps diminish later on.

It is also notable that although physician experience and education meaningfully moderate racial gaps, they are only weakly associated with overall evaluation outcomes. Among White physicians, each additional year of experience increased the probability of receiving a *Top Box* response by 0.4 percentage points ($p < 0.001$), and graduating from a top 50 U.S. medical school was associated with a 1.0 percentage point increase ($p = 0.08$). This suggests that experience and education matter less as direct signals of quality than as alternative anchors that shape patients' evaluations.

Additional analyses further support the robustness of these results. First, because physicians who receive lower evaluation scores may be more likely to exit early, the moderating effect of experience could partly reflect selective survival (Zgierska et al. 2014, Bond et al. 2023). We therefore examined whether racial disparities narrowed as the same physicians gained experience. The within-physician estimates were directionally consistent but not statistically significant, likely reflecting the limited four-year span of our data (Table C2, *Online Appendix*). Second, results were robust to alternative definitions of prestige. Similar patterns emerged when we used top-20 and top-40 cutoffs and when we measured medical school rank using total NIH funding from 2016-2020 (Table C2, *Online Appendix*). Finally, when the interactions of physician race with experience and educational prestige were included simultaneously, the substantive patterns remained similar (Table C1, *Online Appendix*).⁵

Together, these results suggest that racial disparities in patient evaluations were smaller when institutionally supplied cues made physician competence more salient to patients (H2a). One caveat is that our data do not capture whether patients actively sought out or noticed these cues, so the estimated moderation may differ from settings where exposure is directly observed. Nevertheless, the systematic

⁵ Further analysis suggests that physician experience more strongly attenuates racial disparities among physicians from non-Top-50 medical schools than among those with prestigious educational credentials, although this difference (i.e., the three-way interaction) is not statistically significant.

variation in racial gaps across experience and educational prestige is difficult to reconcile with underlying differences in the quality of care provided by racial minority physicians, which do not readily explain why these gaps narrow as such cues become stronger. The findings are more consistent with the explanation that even low-information cues such as experience and educational credentials may provide alternative anchors for assessing physician quality, reducing reliance on race-based expectations.

Evaluation Items and Racial Disparities

We further explore whether racial disparities vary with the extent to which evaluation items direct patients' attention to concrete behaviors or to dimensions of care governed by clearer norms. If some items require relatively little subjective inference about physician quality while others leave greater room for personal interpretation, racial disparities should be smaller on the former.

To examine this possibility, we estimated Equation 1 separately for each evaluation item, replacing the overall evaluation outcome with item-specific *Top Box* indicators. Figure 5 plots the Asian-White and Black-White gaps for each evaluation item, measured as the percentage-point difference in the probability of receiving a *Top Box* response relative to White physicians. Positive (negative) values indicate higher (lower) probabilities (see also Table D1, *Online Appendix* for estimates and adjustments for multiple hypothesis testing).

[Figure 5]

To characterize variation across evaluation items more systematically, we grouped the evaluation items into four categories based on two organizing principles: (i) the concreteness and verifiability of the behavior the item asks about, and (ii) the degree to which clearer evaluative norms exist for the dimension being assessed. Concreteness captures whether an item directs patients to a specific, observable behavior rather than a broader impression, while clarity of norms captures whether patients can rely on widely shared standards rather than healthcare-specific experience to infer what constitutes adequate performance (Weech-Maldonado et al. 2008, Weinick et al. 2011). These dimensions are related but need not coincide: For example, whether a physician “explained things in a way you understood” refers to an observable behavior but leaves discretion over what constitutes an adequate explanation, whereas concern

for a patient's privacy is less tied to a discrete action but governed by clearer everyday norms. Together, these principles yield four categories: protocol adherence, basic interpersonal demeanor, engagement and communication, and overall assessment (see Table D2 in *Online Appendix* for item-level classification).

As shown in Figure 5, Asian and Black physicians received evaluations broadly similar to those of White physicians on items capturing *protocol adherence* that refer to relatively concrete actions, such as whether physicians explained procedures before and after treatment ("MD explained why procedure mattered" and "MD explained results after"), and whether they provided opportunities for patients to ask questions ("MD provided chance to ask questions"). Racial disparities were also modest on items assessing physicians' *basic interpersonal demeanor*. Although these items are less tied to discrete actions, they are governed by relatively clear everyday norms that require little healthcare-specific knowledge to apply. For example, gaps on items such as "MD's concern for your privacy" and "MD treated you with respect" were small and statistically indistinguishable from zero.

By contrast, racial disparities were substantial on items capturing physicians' *engagement and communication* with patients, such as whether they spent enough time, explained things in an understandable way, and kept patients informed ("Time MD spent with you," "MD explained in a way you understood," and "MD kept you informed"). Although these items also concern interpersonal behavior, what constitutes sufficient time, an understandable explanation, or effective communication is less clearly defined, leaving greater room for interpretation in the healthcare context.

Disparities were particularly pronounced on highly subjective items that invited patients to form an *overall assessment* of the physician: their confidence in the physician's skill and their likelihood of recommending the physician. Asian physicians were 3.3 ($p < 0.001$) and 2.1 ($p < 0.01$) percentage points less likely than White physicians to receive *Top Box* responses on confidence-in-skill and likelihood-to-recommend ratings, respectively, while Black physicians were 5.2 ($p < 0.01$) and 3.3 ($p = 0.054$) percentage points less likely to receive *Top Box* responses on these items. These differences corresponded to relative reductions of 4% and 2% for Asian physicians, and 6% and 4% for Black physicians. While some intra-category variation exists, the overall pattern is consistent with our expectations (H2b): Racial

disparities are smaller when the evaluative criteria require less discretionary judgment, as on protocol adherence and basic interpersonal demeanor items.

We treat these category labels as analytic groupings rather than mutually exclusive bins, as some items could plausibly fit more than one category. Accordingly, we present item-level disparities (Figure 5) before providing a post hoc analysis of category-level patterns. The results are robust to scoring items continuously on the two organizing principles rather than assigning them to discrete categories (*Online Appendix D*). We also examine this pattern more systematically using an internal meta-analysis across categories and by care setting, addressing the possibility that these differences reflect not only the nature of the items themselves but also the care settings in which they were used. These additional analyses yield directionally consistent results (*Online Appendix D*).

Importantly, the item exhibiting the largest racial gap—confidence in the physician’s skill—was widely used across care settings, making it less likely that the result was driven by a particular setting. It is also notable that this item is among the least behaviorally anchored, asking patients to assess physician competence without reference to any specific behavior. As such, it provides a particularly strong case of how interpretive latitude may allow race-based expectations to shape evaluations (H2b).

Together, our analyses show that Asian and Black physicians were less likely than White physicians to receive *Top Box* evaluations (H1). These disparities cannot be fully explained by non-random sorting and instead are consistent with race-based expectations shaping evaluations. The gaps were especially pronounced among early-career physicians, those without prestigious educational credentials (H2a), and on evaluation items requiring greater discretion from patients (H2b).

Robustness Checks

Racial Disparities by Patient Race

We examined whether the racial disparities in physician evaluations varied by patient race. As shown in Figure 6, the largest disparities emerged among White patients: Asian and Black physicians were 2.4 and 4.7 percentage points, or 3% and 6%, less likely than White physicians to receive *Top Box* responses (both with $p < 0.01$). By contrast, evaluations by Asian and Black patients did not differ significantly by

physician race; if anything, these patients evaluated ingroup physicians slightly more favorably. The patterns were similar under the IPTW and Heckman adjustments (see Table E1 in *Online Appendix*).

[Figure 6]

These results have two broader implications. First, the overall disparities documented in our main analyses appear to be driven primarily by White patients, who carry disproportionate weight because they constitute the largest share of the patient population. Second, this pattern is difficult to reconcile with a pure quality-difference explanation: If racial minority physicians systematically provided lower-quality care, lower evaluations should appear more consistently across patient groups. Instead, the concentration of disparities among White patients is more consistent with racial differences in evaluation.

Placebo Tests Using Evaluations of Other Care Aspects

To assess whether the estimated racial disparities in physician evaluations were artifacts of unobserved patient characteristics or sorting patterns, we also conducted placebo tests using the main linear probability model (Equation 1) on patients' ratings of other aspects of care: (i) access to care, (ii) the facility, (iii) nurses, and (iv) the care team's attention to personal concerns. We found that racial disparities appeared mainly in domains where spillovers from the patient-physician interaction were plausibly present, such as ratings of nurses and the care team's attention to personal concerns, and not in domains more loosely related to the treating physician, such as access to care and the facility (*Online Appendix E*). These results provide further evidence that patients evaluate physicians differently by race.

Assessing Survey Nonresponse Bias

Although our main analyses adjust for non-random matching between physicians and patients, they do not directly address potential survey nonresponse bias. For example, patients may be more likely to complete the survey after negative experiences with racial minority physicians than after similar experiences with White physicians, which could mechanically inflate observed racial gaps.

We therefore conducted three additional checks (*Online Appendix E*). First, as a proxy, we examined whether patients who answered at least one survey question were differentially likely to complete the physician evaluation section by physician race. Conditional on survey participation, we

found no consistent evidence that completion patterns varied systematically between White and racial minority physicians. Second, we examined the distribution of evaluation scores. If survey nonresponse bias primarily arose because dissatisfied patients were more likely to complete surveys after visits with racial minority physicians, the Asian-White and Black-White gaps should be concentrated at the lower end of the evaluation scale. Instead, the cumulative distribution functions show that the gaps persist across the score range rather than appearing only in the lower tail. Third, we conducted a subsample analysis after excluding low-scoring evaluations (i.e., below 25 or 60) and found similar racial disparities.

These analyses, however, do not directly rule out the possibility that dissatisfied patients are more likely to voice their opinions through the survey after visiting racial minority physicians, since doing so would require observing the satisfaction levels of patients who did not respond. Our results nevertheless suggest that such nonresponse bias is unlikely to be the sole driver of the observed racial disparities. Importantly, even if such bias exists, the observed racial gaps remain substantively meaningful because reported satisfaction, not unobserved satisfaction among nonrespondents, is used to assess physicians.

Sensitivity Analysis, Alternative Specifications, and Matched-Sample Analysis

We further assessed the robustness of our main results through a sensitivity analysis. Following Oster (2019), we quantified how strong unobserved confounders would need to be to eliminate the estimated racial disparities. Assuming that confounders were as influential as the observed controls and increased the model's explanatory power by 30%, the estimated disparities remained large and statistically significant. To fully explain away the disparities, confounders would need to be 3.2 times as important as the observed controls for the Asian-White gap and 5.8 times as important for the Black-White gap.

We also replicated the main analyses using alternative specifications and samples, including (i) replacing care-setting fixed effects with more granular medical-specialty fixed effects, (ii) using continuous evaluation scores, (iii) estimating probit models, (iv) restricting the sample to each patient's first survey response, and (v) using coarsened exact matching (CEM) to balance White and racial minority physicians (Iacus et al. 2012). Results were substantively similar across these analyses, suggesting that omitted variable bias and modeling choices are unlikely to account for the observed racial

disparities. We further examined whether differences in patients' health status and prior relationships with treating physicians helped explain the observed disparities using subsamples with information on these variables (11% and 27% of all observations). Although these factors were associated with overall evaluation outcomes, controlling for them did not eliminate the racial gaps (see *Online Appendix E*).

Discussion

Patient experience surveys have become a critical tool for evaluating perceived care quality in the era of patient-centered care (IOM 2001, Zgierska et al. 2014, CMS 2024). Although originally designed to benchmark hospital-level performance, they are now widely used to assess individual physicians (Manary et al. 2013, Elliott et al. 2015, Indovina et al. 2021). Using proprietary and detailed administrative patient experience data from a large Northeastern U.S. health system, we document racial disparities in these evaluations: Compared to White physicians, Asian and Black physicians are 3% and 5% less likely to receive top-box responses. Because top-box is the modal response, even modest shortfalls can push physicians into the bottom performance quartile, with tangible consequences for performance reviews, compensation plans, and career advancement opportunities (Zgierska et al. 2014). These estimates remain substantively similar after accounting for endogenous patient-physician sorting on both observed and unobserved factors, suggesting that the disparities persist beyond non-random matching alone.

We also find that racial disparities in evaluations attenuate when organizational arrangements better structure how quality is perceived and assessed. Gaps are larger among early-career physicians and those with less prestigious training, and when patients are asked to assess more subjective dimensions of care. They are narrower when institutionally supplied cues provide stronger signals of physician quality and when evaluation criteria direct attention to more concrete and verifiable behaviors or to dimensions of care for which clearer norms exist. Together, these findings suggest that institutionally supplied cues and evaluative structures can provide alternative anchors for assessing physician quality, reducing the extent to which evaluations are shaped by physician race.

Limitations

An important limitation of this study is that we do not observe objective care quality, so we cannot definitively separate race-based expectations from true care quality differences. In line with the broader difficulty of measuring care quality (Smith et al. 1997, Landon et al. 2003, Nembhard et al. 2009, Brewer et al. 2020), we rely on diagnostic tests: Disparities are largest when patients evaluate more subjective items, when institutionally supplied cues provide weaker signals of physician quality, and when racial minority physicians are evaluated by White patients—patterns more consistent with status-based evaluation than with systematic quality differences. Moreover, we cannot find any theoretical basis or field evidence that care quality differs by physician race, on average.

Second, despite the richness of our data, several other limitations remain. Although our data allow for comparisons across racial categories, they do not support a more fine-grained intersectional analysis of race and gender with sufficient precision. The fixed effects we use sharply reduce the effective sample size once intersectionality is considered, particularly for Black female and Black male physicians. Meanwhile, because our sample comes from a single health system, generalizability to other health systems and periods remains uncertain. In addition, because our theoretical question concerns how patients perceive physicians, we measure physician race as perceived rather than self-identified. However, perceived race may not always correspond to physicians' own racial identification. Finally, although we observe an extensive set of patient, physician, and visit characteristics, additional factors, such as patient health and prior relationships with physicians, may also shape evaluations. While our IPTW and Heckman approaches aim to account for endogenous selection and yield similar estimates, both rely on constructed rather than observed patient choice sets. In particular, we cannot observe how often realized visits involved active choices across locations, a limitation especially relevant in medical specialties where distance is less predictive of matching.

Theoretical Contributions

This study provides systematic field evidence on racial disparities in patient evaluations of physicians. While prior research has raised the possibility that patient satisfaction surveys may capture not only care

quality but also race-based expectations, direct evidence remains limited. Scholars in economics, management, psychology, and sociology have approached this question indirectly, by analyzing pay gaps or inferring patient attitudes through interviews, vignette experiments, and theoretical models (Sinclair and Kunda 1999, Wingfield and Chavez 2020). Healthcare researchers, with greater access to patient evaluation data, have primarily analyzed gender disparities; studies of racial disparities are limited and typically confined to specific clinics or specialties (Garcia et al. 2019, Takeshita et al. 2020, Kim et al. 2024). Although valuable, these studies are not designed to yield systematic estimates, in part because context, evaluation criteria, and patient-physician matching vary across studies (Chown 2020). By examining patient evaluations across a large integrated health system while accounting for endogenous patient-physician matching, we document consequential racial disparities in physician evaluations.

This study advances status-based explanations of inequality by integrating status characteristics theory with the structure of evaluations literature. Status characteristics theory argues that diffuse status characteristics, such as race, shape performance expectations under uncertainty (Berger et al. 1972, 1980, Wagner and Berger 1993). We show that in a professional domain with high information asymmetry, namely patients evaluating physicians, such expectations depend in part on how evaluation is organized. Institutionally supplied cues can provide alternative bases for assessment that compete with category-based expectations (Fiske and Neuberg 1990, Bodenhausen et al. 2012), and more formalized and concrete evaluation criteria constrain the discretion that allows racial stereotypes to shape outcomes (Anderson and Tomaskovic-Devey 1995, Dobbin et al. 2015). Together, these arrangements can structure how evaluators assess quality and reduce status-based disparities when evaluators lack domain expertise.

In doing so, we also contribute to the growing literature on the structure of evaluation processes. Prior work has identified design features that can reduce disparities by reducing evaluative ambiguity, such as avoiding scales that evoke an idealized image of the “perfect” candidate or reducing the number of categories evaluators must navigate (Rivera and Tilcsik 2019, Botelho et al. 2025). We extend this work by showing that, beyond rating scales themselves, the concreteness and verifiability of evaluation items also shape the extent to which evaluators rely on race-based expectations. Racial disparities are

larger when patients are asked to assess more subjective dimensions of care, but smaller when evaluation items direct attention to more verifiable behaviors or to dimensions of care for which clearer norms exist.

Finally, the study advances status research by showing how institutionally supplied cues shape evaluation when non-expert evaluators must assess professional services. A substantial body of research shows that individuating information can reduce reliance on category-based expectations among professional evaluators (Fiske and Neuberg 1990, Bodenhausen et al. 2012, Botelho and Abraham 2017), while organizations increasingly delegate evaluative authority to customers, students, patients, and other non-expert audiences. We show that in such settings, institutionally supplied cues such as physician experience and educational prestige can provide alternative anchors for assessment: They meaningfully moderate racial disparities despite being only weakly associated with overall evaluation levels, suggesting that even imperfect proxies for quality can compete with race-based expectations and support more individuated evaluations. These findings suggest that when organizations cannot rely on expert evaluators (Li 2017, Botelho 2024), the design of evaluation processes becomes especially consequential.

Future Research

Our study opens several paths for future research. One priority is replication across health systems to assess the generalizability of these patterns across organizations and patient populations. Future work could also combine patient evaluations with more direct measures of care quality to better distinguish evaluative disparities from underlying performance differences. In addition, longitudinal designs linking patient evaluations to physician career outcomes would help clarify how early disadvantages in high-stakes ratings accumulate into longer-term inequality in pay, recognition, and career advancement.

Beyond documenting racial disparities, future work can use field experiments to test how evaluation design shapes bias, including comparisons of subjective versus behaviorally anchored items, alternative scale structures, different item orders and rater prompts, and credential-salience interventions such as brief physician bios in survey invitations. More broadly, the status-sensitive evaluation framework we advance, centered on ambiguity and individuating signals, should be tested across contexts to assess its portability as a lever for building fairer and more effective evaluation systems.

Practical Implications

For hospitals and administrators, our findings point to several practical levers for reducing racial differences in patient experience surveys. One promising pathway lies in survey design. Items that direct attention to concrete and verifiable behaviors or to dimensions of care for which clearer norms exist can reduce reliance on stereotypes. Where inherently subjective dimensions must be measured (e.g., communication), ambiguity can be lowered by providing clear definitions and brief examples, rather than leaving patients to apply idiosyncratic standards. Beyond question design, organizations can periodically audit items, flag those that show systematic group-based gaps unrelated to clinical quality, and down-weight or remove those items when they function differently across physicians.

A second lever involves making performance-relevant cues more salient. Because identity cues can shape evaluations, interventions that foreground performance-relevant credentials can redirect attention away from demographics. Analogous to the way “doctor” identity badges reduced gender differences (Olson et al. 2022), including brief physician profiles with survey invitations—summarizing certifications and experience, or simply highlighting physicians’ professional authority—may redirect patient attention from demographic characteristics toward performance-relevant indicators. Of note, this intervention works by limiting the influence of a race-based heuristic through the introduction of a credential-based one. It may therefore narrow racial disparities without necessarily improving the accuracy of evaluations as measures of care quality. Meanwhile, because physicians from different racial groups may differ systematically in the types of credentials they possess, hospitals and survey administrators should carefully consider which credentials to highlight and how to present them without introducing new sources of inequality into the evaluation process.

Overall, our findings caution against using patient experience surveys as standalone measures of physician performance and strategic decision-making. Without careful calibration, tying compensation, promotion, contract renewal, or opportunity structure more broadly to patient satisfaction scores risks amplifying existing structural barriers and distorting incentives.

[Link to Online Appendix](#)

References

- Abbott A (1988) *The system of professions: An essay on the division of expert labor* (University of Chicago press).
- Abraham M, Botelho TL, Lamont-Dobbin G (2024) The (re)production of inequality in evaluations: A unifying framework outlining the drivers of gender and racial differences in evaluative outcomes. *Research in Organizational Behavior* 44:100207.
- Agency for Healthcare Research and Quality (AHRQ) (2023) CAHPS surveys and tools to advance patient-centered care.
- Anderson CD, Tomaskovic-Devey D (1995) Patriarchal pressures: An exploration of organizational processes that exacerbate and erode gender earnings inequality. *Work and Occupations* 22(3):328–356.
- Anhang Price R, Elliott MN, Zaslavsky AM, Hays RD, Lehrman WG, Rybowski L, Edgman-Levitan S, Cleary PD (2014) Examining the role of patient experience surveys in measuring health care quality. *Medical Care Research and Review* 71(5):522–554.
- Azoulay P, Liu CC, Stuart TE (2017) Social influence given (partially) deliberate matching: Career imprints in the creation of academic entrepreneurs. *American Journal of Sociology* 122(4):1223–1271.
- Berger J, Cohen BP, Zelditch Jr M (1972) Status characteristics and social interaction. *American Sociological Review*:241–255.
- Berger J, Rosenholtz SJ, Zelditch M (1980) Status organizing processes. *Annual Review of Sociology* 6:479–508.
- Biondi EA, Hall M, Leonard MS, Pirraglia PA, Alverson BK (2016) Association between resource utilization and patient satisfaction at a tertiary care medical center. *Journal of Hospital Medicine* 11(11):785–791.
- Boatright D, Ross D, O'Connor P, Moore E, Nunez-Smith M (2017) Racial disparities in medical student membership in the Alpha Omega Alpha Honor Society. *JAMA Internal Medicine* 177(5):659–665.
- Bodenhausen GV, Kang SK, Peery D (2012) Social categorization and the perception of social groups. *The Sage handbook of social cognition*:311–329.
- Bond AM, Casalino LP, Tai-Seale M, Unruh MA, Zhang M, Qian Y, Kronick R (2023) Physician turnover in the United States. *Annals of Internal Medicine* 176(7):896–903.
- Botelho TL (2024) From audience to evaluator: When visibility into prior evaluations leads to convergence or divergence in subsequent evaluations among professionals. *Organization Science* 35(5):1682–1703.
- Botelho TL, Abraham M (2017) Pursuing quality: How search costs and uncertainty magnify gender-based double standards in a multistage evaluation process. *Administrative Science Quarterly* 62(4):698–730.
- Botelho TL, Jun S, Humes D, DeCelles KA (2025) Scale dichotomization reduces customer racial discrimination and income inequality. *Nature* 639:395--403.
- Brewer A, Osborne M, Mueller AS, O'Connor DM, Dayal A, Arora VM (2020) Who gets the benefit of the doubt? Performance evaluations, medical errors, and the production of gender inequality in emergency medical education. *American Sociological Review* 85(2):247–270.
- Castilla EJ (2008) Gender, race, and meritocracy in organizational careers. *American Journal of Sociology* 113(6):1479–1526.
- Centers for Medicare & Medicaid Services (CMS) (2024) Hospital Value-Based Purchasing.
- Chow CJ, Wadsworth R, Ryujin D, Vo M, Thomas JK (2024) “Where are you really from?": a qualitative study of Asian American medical provider experiences. *Advances in Health Sciences Education*:1–20.
- Chown J (2020) Financial incentives and professionals' work tasks: The moderating effects of jurisdictional dominance and prominence. *Organization Science* 31(4):887–908.

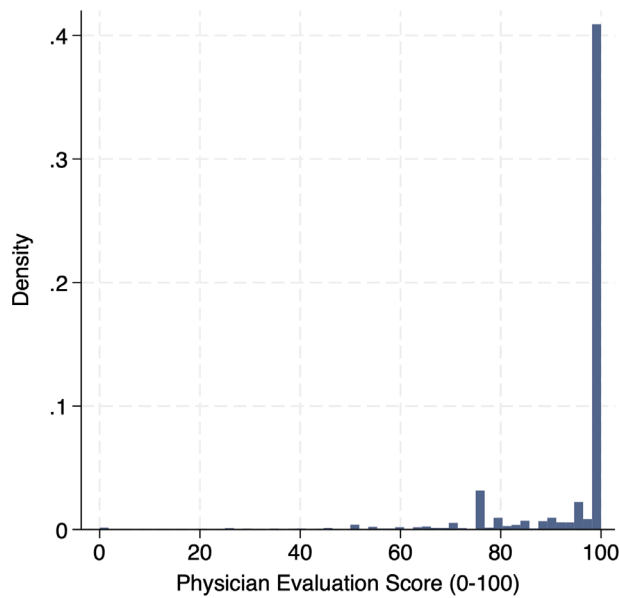
- Dobbin F, Schrage D, Kalev A (2015) Rage against the iron cage: The varied effects of bureaucratic personnel reforms on diversity. *American Sociological Review* 80(5):1014–1044.
- Doyle C, Lennox L, Bell D (2013) A systematic review of evidence on the links between patient experience and clinical safety and effectiveness. *BMJ Open* 3(1):e001570.
- Elliott MN, Cohea CW, Lehrman WG, Goldstein EH, Cleary PD, Giordano LA, Beckett MK, Zaslavsky AM (2015) Accelerating improvement and narrowing gaps: trends in patients' experiences with hospital care reflected in HCAHPS public reporting. *Health Services Research* 50(6):1850–1867.
- Fiske S, Neuberg S (1990) A continuum model of impression formation, from category-based to individuating processes: Influences on information and motivation on attention and interpretation. *Advances in Experimental Social Psychology* 23:1–74.
- Flaherty C (2025) Survey Says: Students Are Customers. *Inside Higher Ed* <https://tinyurl.com/2jt7wmx3>.
- Foschi M (2000) Double standards for competence: Theory and research. *Annual Review of Sociology* 26(1):21–42.
- Gao GG, McCullough JS, Agarwal R, Jha AK (2012) A changing landscape of physician quality reporting: analysis of patients' online ratings of their physicians over a 5-year period. *Journal of Medical Internet Research* 14(1):e38.
- Garcia LC, Chung S, Liao L, Altamirano J, Fassiotto M, Maldonado B, Heidenreich P, Palaniappan L (2019) Comparison of outpatient satisfaction survey scores for Asian physicians and non-Hispanic white physicians. *JAMA Network Open* 2(2):e190027–e190027.
- Greene J, Hibbard JH (2012) Why does patient activation matter? An examination of the relationships between patient activation and health-related outcomes. *Journal of General Internal Medicine* 27(5):520–526.
- Greene J, Hibbard JH, Sacks RM (2018) Does the race/ethnicity or gender of a physician's name impact patient selection of the physician? *Journal of the National Medical Association* 110(3):206–211.
- Harris KM (2003) How do patients choose physicians? Evidence from a national survey of enrollees in employment-related health plans. *Health Services Research* 38(2):711–732.
- Haviland MG, Morales LS, Dial TH, Pincus HA (2005) Race/ethnicity, socioeconomic status, and satisfaction with health care. *American Journal of Medical Quality* 20(4):195–203.
- HCAHPS (2025) Summary of HCAHPS Survey Results. <https://tinyurl.com/2sa4e24a>.
- Heckman JJ (1976) The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement, Volume 5, Number 4* 475–492.
- Huckman RS, Kelley MA (2013) Public reporting, consumerism, and patient empowerment. *New England Journal of Medicine* 369(20):1875–1877.
- Iacus SM, King G, Porro G (2012) Causal inference without balance checking: Coarsened exact matching. *Political Analysis* 20(1):1–24.
- Indovina KA, Keniston A, Manchala V, Burden M (2021) Predictors of a top-box patient experience: a retrospective observational study of HCAHPS data at a safety net institution. *Journal of Patient Experience* 8:23743735211034342.
- Institute of Medicine (US) Committee on Quality of Health Care in America (2001) *Crossing the quality chasm: a new health system for the 21st century* (National Academies Press).
- Kang SK, Kaplan S (2019) Working toward gender diversity and inclusion in medicine: myths and solutions. *The Lancet* 393(10171):579–586.
- Keating NL, Kouri EM, He Y, Freedman RA, Volya R, Zaslavsky AM (2016) Location isn't everything: proximity, hospital characteristics, choice of hospital, and disparities for breast cancer surgery patients. *Health Services Research* 51(4):1561–1583.
- Kim Y hun, Bihun R, Kim L, Skjaerlund J, Cole K, Ouk M, Wilshusen L, Warsame R (2024) Patient Perceptions of East and Southeast Asian Physicians Amidst Anti-Asian Rhetoric During the COVID-19 Pandemic. *Mayo Clinic Proceedings (Elsevier)*, 1945–1955.

- Kozikowski A, Morton-Rias D, Mauldin S, Jeffery C, Kavanaugh K, Barnhill G (2022) Choosing a provider: what factors matter most to consumers and patients? *Journal of Patient Experience* 9:23743735221074175.
- Kupfer JM, Bond EU (2012) Patient satisfaction and patient-centered care: necessary but not equal. *JAMA* 308(2):139–140.
- Landon BE, Normand SLT, Blumenthal D, Daley J (2003) Physician clinical performance assessment: prospects and barriers. *JAMA* 290(9):1183–1189.
- Larkin I, Pierce L (2015) Compensation and employee misconduct: The inseparability of productive and counterproductive behavior in firms. *Organizational wrongdoing: Key perspectives and new directions*:1–27.
- Leslie SJ, Cimpian A, Meyer M, Freeland E (2015) Expectations of brilliance underlie gender distributions across academic disciplines. *Science* 347(6219):262–265.
- Lett E, Murdock HM, Orji WU, Aysola J, Sebro R (2019) Trends in racial/ethnic representation among US medical students. *JAMA Network Open* 2(9):e1910490–e1910490.
- Li D (2017) Expertise versus Bias in Evaluation: Evidence from the NIH. *American Economic Journal: Applied Economics* 9(2):60–92.
- Linde S, Egede LE (2022) Association of county race and ethnicity characteristics with number of insurance carriers and insurance network breadth. *JAMA Network Open* 5(4):e227404.
- Manary MP, Boulding W, Staelin R, Glickman SW (2013) The patient experience and health outcomes. *New England Journal of Medicine* 368(3):201–203.
- Nembhard IM, Alexander JA, Hoff TJ, Ramanujam R (2009) Why does the quality of health care continue to lag? Insights from management research. *The Academy of Management Perspectives*:24–42.
- Nguyen M, Chaudhry SI, Desai MM, Hajduk AM, McDade WA, Fancher TL, Boatright D (2022) Rates of medical student placement into graduate medical education by sex, race and ethnicity, and socioeconomic status, 2018–2021. *JAMA Network Open* 5(8):e2229243.
- Ogrod ES (1997) Compensation and quality: A physician’s view. *Health Affairs* 16(3):82–86.
- Olson EM, Dines VA, Ryan SM, Halvorsen AJ, Long TR, Price DL, Thompson RH, Tollefson MM, Van Gompel JJ, Oxentenko AS (2022) Physician identification badges: a multispecialty quality improvement study to address professional misidentification and bias. *Mayo Clinic Proceedings* (Elsevier), 658–667.
- Oster E (2019) Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics* 37(2):187–204.
- Pedulla DS, Dias F, Wullert KE, Muñoz J (2026) Racial discrimination in context: The role of organizational policies and practices in hiring discrimination. *American Journal of Sociology* 131(4):918–971.
- Richman BD, Schulman KA (2022) Are patient satisfaction instruments harming both patients and physicians? *JAMA* 328(22):2209–2210.
- Rivera LA, Tilsik A (2019) Scaling down inequality: Rating scales, gender bias, and the architecture of evaluation. *American Sociological Review* 84(2):248–274.
- Rubin DB (1990) Formal mode of statistical inference for causal effects. *Journal of Statistical Planning and Inference* 25(3):279–292.
- Savel RH, Benson P, Collins C, Gongireddy S, Oquendo C, Gyekye K, Villar-Trinidad E, Fennimore J, Akunyili I (2025) Independent risk factors for less than “top box” doctor communication patient experience survey scores in an urban teaching hospital. *American Journal of Medical Quality* 40(4):140–145.
- Siegrist Jr RB (2013) Patient satisfaction: history, myths, and misperceptions. *AMA Journal of Ethics* 15(11):982–987.
- Sinclair L, Kunda Z (1999) Reactions to a black professional: motivated inhibition and activation of conflicting stereotypes. *Journal of Personality and Social Psychology* 77(5):885.

- Smith H, Currie C, Chaiwuttisak P, Kyprianou A (2018) Patient choice modelling: how do patients choose their hospitals? *Health Care Management Science* 21(2):259–268.
- Smith MA, Atherly AJ, Kane RL, Pacala JT (1997) Peer review of the quality of care: reliability and sources of variability for outcome and process assessments. *JAMA* 278(19):1573–1578.
- Takeshita J, Wang S, Loren AW, Mitra N, Shults J, Shin DB, Sawinski DL (2020) Association of racial/ethnic and gender concordance between patients and physicians with patient experience ratings. *JAMA Network Open* 3(11):e2024583–e2024583.
- Tyser A, Abtahi A, McFadden M, Presson A (2016) Evidence of non-response bias in the Press-Ganey patient satisfaction survey. *BMC Health Services Research* 16(1):350.
- de Vaan M, Stuart T (2022) Gender in the Markets for Expertise. *American Sociological Review* 87(3):443–477.
- Van Ryn M, Burke J (2000) The effect of patient race and socio-economic status on physicians' perceptions of patients. *Social Science & Medicine* 50(6):813–828.
- Wagner DG, Berger J (1993) *Status characteristics theory: The growth of a program*. (Stanford University Press).
- Weech-Maldonado R, Elliott MN, Oluwole A, Schiller KC, Hays RD (2008) Survey response style and differential use of CAHPS rating scales by Hispanics. *Medical Care* 46(9):963.
- Weinick RM, Elliott MN, Volandes AE, Lopez L, Burkhart Q, Schlesinger M (2011) Using standardized encounters to understand reported racial/ethnic disparities in patient experiences with care. *Health Services Research* 46(2):491–509.
- Wingfield AH, Chavez K (2020) Getting in, getting hired, getting sideways looks: Organizational hierarchy and perceptions of racial discrimination. *American Sociological Review* 85(1):31–57.
- Wooldridge JM (2010) *Econometric analysis of cross section and panel data* (MIT press).
- Wooldridge JM (2015) Control function methods in applied econometrics. *Journal of Human Resources* 50(2):420–445.
- Zgierska A, Rabago D, Miller MM (2014) Impact of patient satisfaction ratings on physicians and clinical care. *Patient Preference and Adherence*:437–446.
- Zou LX, Cheryan S (2017) Two axes of subordination: A new model of racial position. *Journal of Personality and Social Psychology* 112(5):696.

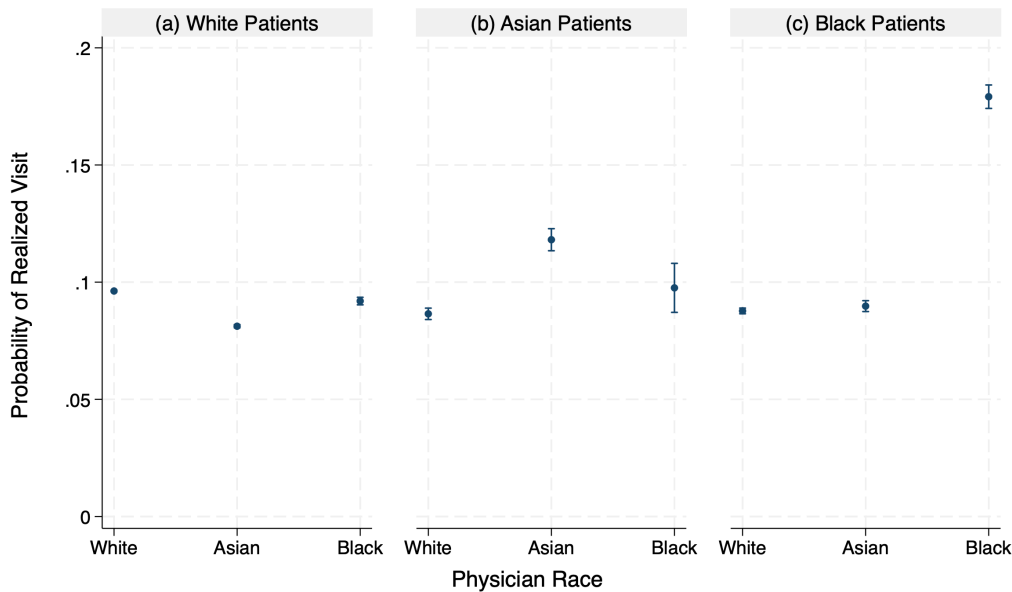
Figures and Tables

Figure 1. Distribution of Physician Evaluation Score



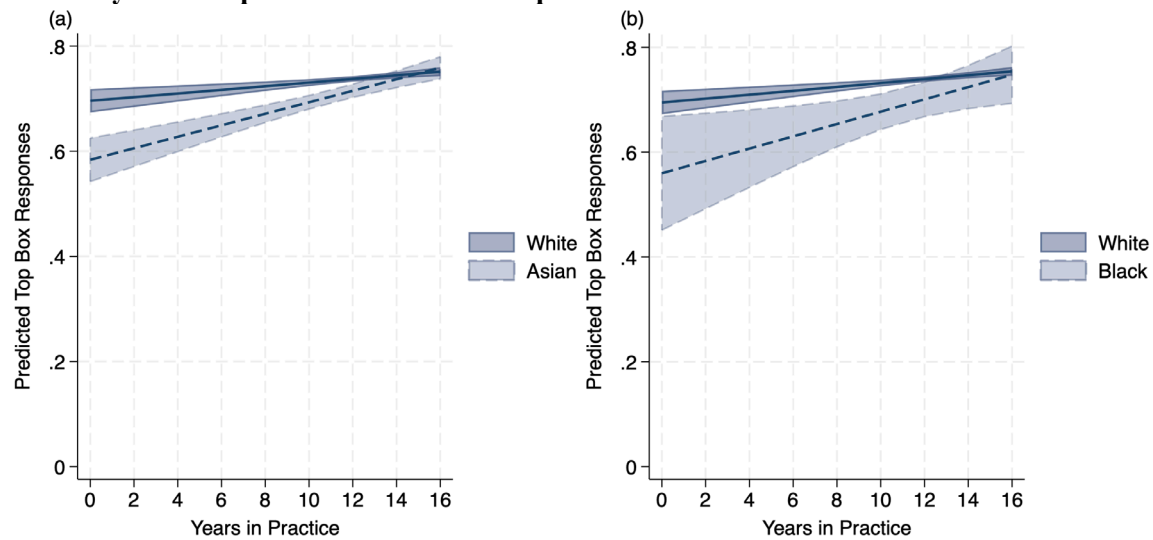
Note. This plot shows the distribution of the overall physician evaluation score (100-point index; 100 indicates the best possible outcome, i.e., *Top Box*).

Figure 2. Probability of Visit by Patient and Physician Race



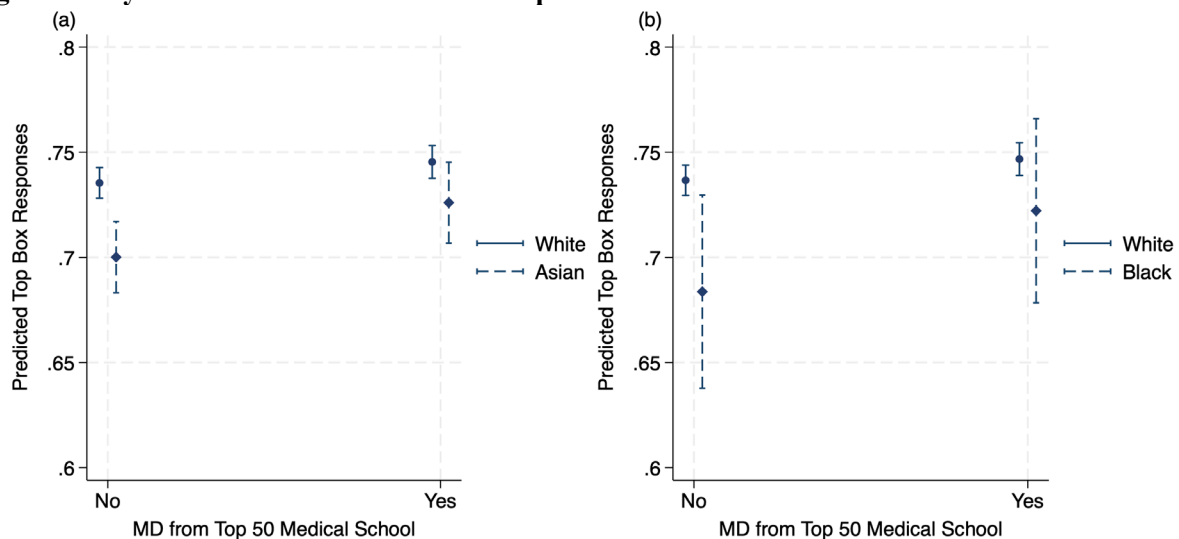
Note. The figure plots the probability that a patient-physician pair in the simulated choice-set sample is a realized match, shown separately by physician race for (a) White, (b) Asian, and (c) Black patients. The simulated choice set combines each realized visit with up to 10 counterfactual physicians drawn from the same specialty, care setting, and year. Estimates are descriptive and do not include additional controls beyond specialty-year fixed effects. Error bars represent 95% confidence intervals.

Figure 3. Physician Experience and Racial Disparities



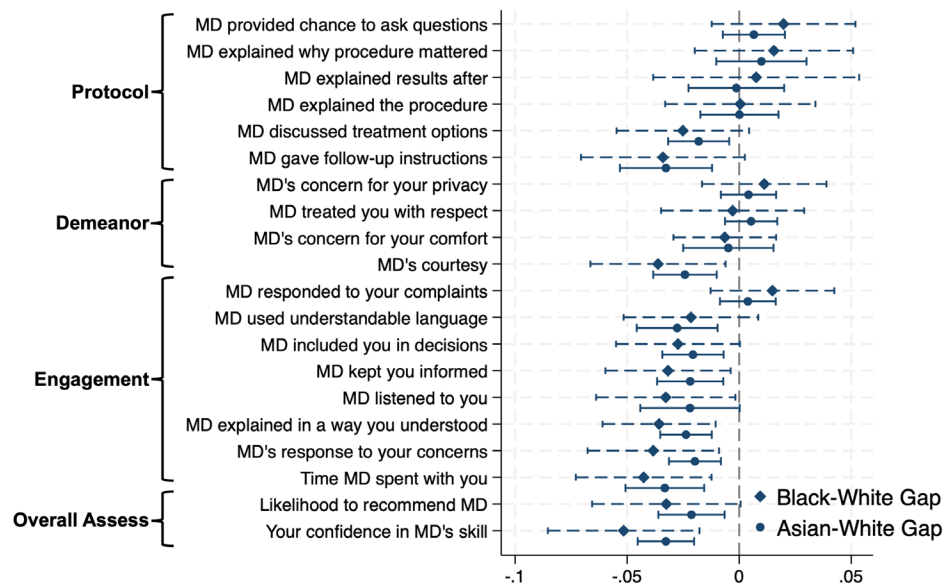
Note. The figure plots predicted probabilities of receiving a *Top Box* response by physician race across years in practice. Panel (a) compares White and Asian physicians, and panel (b) compares White and Black physicians. Estimates come from linear probability models that interact physician race with years in practice and include controls for physician characteristics, patient characteristics, and hospital and care-setting fixed effects. Standard errors are clustered at the physician level. Shaded areas indicate 95% confidence intervals. See also Table C1 in *Online Appendix*.

Figure 4. Physician Education and Racial Disparities



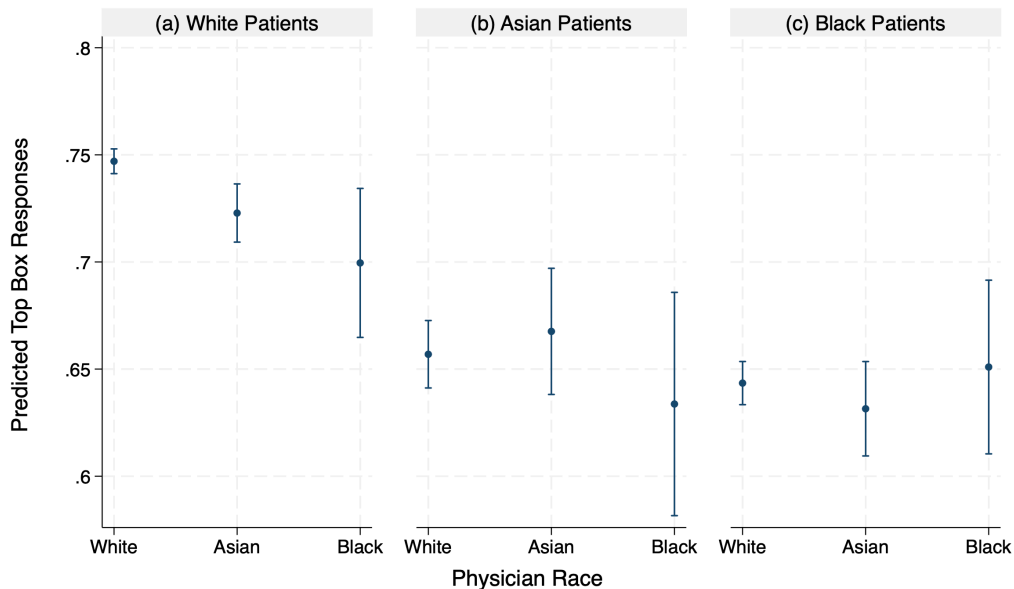
Note. The figure plots predicted probabilities of receiving a *Top Box* response by physician race across medical school type. Panel (a) compares White and Asian physicians, and panel (b) compares White and Black physicians. Estimates come from linear probability models that interact physician race with medical school type (Top-50 vs. non-Top-50) and include controls for physician characteristics, patient characteristics, and hospital and care-setting fixed effects. Standard errors are clustered at the physician level. Error bars represent 95% confidence intervals. See also Table C1 in *Online Appendix*.

Figure 5. Evaluation Items and Racial Disparities



Note. The figure shows the probability that Asian and Black physicians receive a *Top Box* response relative to White physicians by evaluation item. A positive (negative) value indicates a higher (lower) probability. Effects are estimated with linear probability models controlling for physician characteristics, patient characteristics, and hospital and care-setting fixed effects. Standard errors are clustered at the physician level. Error bars represent 95% confidence intervals. See also Table D1 in *Online Appendix*.

Figure 6. Patient Race and Racial Disparities



Note. The figure shows racial disparities in the likelihood of receiving a *Top Box* response by patient race. Effects are estimated with linear probability models controlling for physician characteristics, patient characteristics, and hospital and care-setting fixed effects. Standard errors are clustered at the physician level. Error bars represent 95% confidence intervals. See also Table E1 in *Online Appendix*.

Table 1. Summary Statistics and Pairwise Correlations

Panel A: Summary Statistics					
Physician Race	Asian	Black	White	Other	Total
Physician Count	916	160	2,995	508	4,579
Female (%)	45%	61%	37%	39%	40%
Physician Edu: U.S. Excl. Top 50 (%)	22%	36%	35%	22%	31%
Physician Edu: Non-U.S. (%)	42%	24%	15%	45%	24%
NPI Registration Year (Mean)	2010	2009	2008	2009	2008
Total Evaluation Count	93,393	20,493	419,607	40,354	573,847
Evaluation Per Physician	102	128	140	79	125
Top Box (%)	69.85%	67.42%	74.38%	69.61%	73.03%
Mean Patient Age	62.72	58.79	61.52	60.49	61.55
Female Patient (%)	58.87%	65.95%	56.98%	58.92%	57.74%
White Patient (%)	67.27%	61.58%	67.73%	65.21%	67.26%
Asian Patient (%)	2.03%	1.61%	1.22%	1.37%	1.37%
Black Patient (%)	6.31%	12.33%	4.92%	6.91%	5.55%
Medicaid Patient (%)	5.33%	7.51%	4.59%	5.96%	4.91%

Panel B: Pairwise Correlations														
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
(1) Top Box	1.00													
(2) Physician Race: Asian	-0.03	1.00												
(3) Physician Race: Black	-0.03	-0.09	1.00											
(4) Physician Race: White	0.05	-0.73	-0.32	1.00										
(5) Years in Practice	0.08	-0.13	-0.06	0.18	1.00									
(6) Physician Edu: U.S. Excl. Top 50	0.01	-0.07	0.06	0.04	-0.02	1.00								
(7) Physician Edu: Non-U.S.	-0.04	0.27	-0.03	-0.27	-0.10	-0.31	1.00							
(8) Female Physician	0.02	0.06	0.10	-0.11	-0.11	-0.03	0.01	1.00						
(9) Patient Age	0.02	0.04	-0.02	-0.02	0.07	-0.04	0.02	-0.14	1.00					
(10) Female Patient	-0.01	0.01	0.03	-0.02	-0.01	-0.01	0.01	0.19	-0.10	1.00				
(11) Patient Race: Asian	-0.02	0.02	0.00	-0.02	0.00	0.00	0.00	0.02	-0.06	0.02	1.00			
(12) Patient Race: Black	-0.06	0.01	0.06	-0.05	-0.03	0.00	0.02	0.02	-0.09	0.06	-0.03	1.00		
(13) Patient Race: White	0.04	-0.01	-0.03	0.03	0.12	0.00	-0.01	-0.02	0.20	-0.04	-0.22	-0.45	1.00	
(14) Medicaid Patient	-0.05	0.01	0.02	-0.03	-0.08	0.01	0.01	0.02	-0.26	0.03	0.01	0.13	-0.17	1.00

Table 2. Patient Characteristics by Race

	Patient Race		
	Asian	Black	White
Top Box Responses (%)	66.7% [65.6%, 67.7%]	63.2% [62.7%, 63.7%]	74.1% [73.9%, 74.2%]
Top Box Responses by Physician Race (%)			
Asian Physician	67.6% [65.5%, 69.8%]	61.1% [59.9%, 62.3%]	70.8% [70.5%, 71.2%]
Black Physician	63.3% [58.0%, 68.6%]	66.4% [64.5%, 68.3%]	67.9% [67.1%, 68.8%]
White Physician	66.2% [65.0%, 67.4%]	63.6% [63.0%, 64.2%]	75.4% [75.2%, 75.6%]
Medicaid Usage (%)	7.6% [7.1%, 8.1%]	16.8% [16.6%, 17.0%]	3.4% [3.4%, 3.5%]
Bottom Quartile Household Income (%)	23.4% [22.4%, 24.4%]	66.6% [66.1%, 67.1%]	17.7% [17.5%, 17.8%]
Below High School (%)	3.4% [2.8%, 4.1%]	2.8% [2.5%, 3.2%]	1.4% [1.3%, 1.5%]

Note. The table reports patient characteristics by patient race. Bracketed values report 95% confidence intervals. *Medicaid Usage* indicates whether Medicaid or Medicaid Managed Care was recorded as the patient's primary insurance at the time of visit. *Bottom Quartile Household Income* indicates whether the patient's residential ZIP code falls within the lowest quartile of median household income in the sample. *Below High School* indicates whether the patient reported not completing high school. Education and residential ZIP code information are not available for all patients; 450,309 (78%) visits include non-missing ZIP code information and 153,286 (27%) include non-missing education information.

Table 3. Physician Race and Top Box Responses

Model	LPM (1)	Top Box IPTW (2)	Heckman (3)
Physician Race			
Asian	-0.022** (0.007)	-0.025** (0.008)	-0.023* (0.009)
Black	-0.038* (0.016)	-0.034* (0.017)	-0.028+ (0.017)
Other/Unknown	-0.004 (0.008)	-0.006 (0.010)	-0.007 (0.010)
Female Physician	0.009+ (0.005)	0.008 (0.006)	0.007 (0.006)
Physician Years in Practice	0.006*** (0.001)	0.007*** (0.001)	0.007*** (0.001)
Physician Education			
Non-Top-50 U.S. Medical School	-0.003 (0.005)	-0.005 (0.006)	-0.003 (0.006)
Non-U.S. Medical School	-0.031*** (0.007)	-0.037*** (0.009)	-0.037*** (0.009)
Patient Race			
Asian	-0.076*** (0.006)	-0.080*** (0.009)	-0.077*** (0.007)
Black	-0.088*** (0.004)	-0.083*** (0.005)	-0.085*** (0.005)
Other/Unknown	-0.022*** (0.003)	-0.024*** (0.004)	-0.028*** (0.004)
Female Patient	-0.007** (0.002)	-0.005 (0.003)	-0.001 (0.003)
Medicaid Patient	-0.046*** (0.005)	-0.046*** (0.006)	-0.048*** (0.005)
Patient Age	-0.000** (0.000)	-0.000*** (0.000)	-0.000*** (0.000)
Constant	0.690*** (0.012)	0.680*** (0.015)	0.373*** (0.039)
Hospital FE	Y	Y	Y
Care-Setting FE	Y	Y	Y
N =	410,479	280,547	3,127,853
R-squared	0.045	0.061	
Log pseudolikelihood			-983,156
ρ			0.069*** (0.007)

Note. Model 1 estimates the likelihood of *Top Box* responses by physician race using a linear probability model (LPM) based on observed evaluations. Model 2 reweights observations using inverse probability of treatment weights (IPTW) to adjust for sorting on observables. Model 3 applies Heckman correction that allows unobserved determinants of sorting and evaluations to be correlated. All models control for other physician characteristics (gender, credentials), patient characteristics (gender, race, age, insurance), hospital fixed effects, and care-setting fixed effects. The IPTW and Heckman adjustments additionally require patients' and physicians' ZIP codes to calculate geographic proximity, which is available for 69% of the full sample. White physicians, White patients, and physician education at top 50 U.S. medical schools are the reference categories. Standard errors are clustered at the physician level. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, + $p < 0.10$

Online Appendix for “The Doctor Will See You Now: Physician Race and Patient Evaluations”

A. Additional Information on Survey Measures

The health system administers patient surveys across six departments: emergency, medical practice, oncology, other outpatient, ambulatory surgery, and other inpatient. Table A1 presents the detailed survey items by department. Because the six-department classification primarily reflects administrative structure, we consolidate these into three care-setting groups for analysis: emergency department, outpatient (medical practice, oncology, and other outpatient), and inpatient (ambulatory surgery and other inpatient). This grouping has been reviewed by a physician.

[Table A1]

B. Addressing Endogenous Matching between Patients and Physicians

B.1 Construction of Simulated Choice Sets

To address endogenous patient-physician matching, we first approximate the set of physician choices available to patients in a given clinical context. Patients presumably consider some set of physicians when seeking care (i.e., the *actual choice set*), which we cannot observe directly. We approximate this actual choice set by identifying all physicians active in the same specialty, care setting, and visit year, which we refer to as the *full feasible set*. For computational efficiency in estimating the selection models, we retain each realized interaction and randomly sample up to 10 counterfactual physicians from its full feasible set, producing the *simulated choice-set* sample used in our analyses.

Table B1 reports summary statistics and pairwise correlations for the resulting 5,144,604 patient-physician interactions in the simulated choice-set sample, 9.12% of which are realized interactions.

[Table B1]

The simulated choice-set sample captures meaningful racial heterogeneity among physicians, as patients could plausibly have encountered physicians from multiple racial groups within the same specialty, care setting, and year. As shown in Figure B1, 92% of visit-level choice sets include both White

and racial minority physicians. Disaggregating by racial group, 87% include both White and Asian physicians, and 30% include both White and Black physicians.

[Figure B1]

B.2 Supplementary Details on IPTW Estimation

A key challenge of our baseline linear probability model (LPM) is that it is estimated only on realized patient-physician interactions. Specifically, consider the model

$$Y_{ij} = \alpha + \beta \text{PhyRace}_i + X'_{ij}\theta + \epsilon_{ij}$$

where Y_{ij} indicates whether patient j gives physician i a *Top Box* response, PhyRace_i denotes physician race, and X_{ij} includes other physician, patient, and visit characteristics.

We observe Y_{ij} only when the patient-physician interaction is realized ($D_{ij} = 1$). Because

$$E[Y_{ij} | \text{PhyRace}_i, X_{ij}, D_{ij} = 1] = \alpha + \beta \text{PhyRace}_i + X'_{ij}\theta + E[\epsilon_{ij} | \text{PhyRace}_i, X_{ij}, D_{ij} = 1]$$

the LPM estimator is consistent only if

$$E[\epsilon_{ij} | \text{PhyRace}_i, X_{ij}, D_{ij} = 1] = 0$$

The mean-independence condition may fail even if the control vector X_{ij} captures all relevant characteristics that influence patient evaluation, such that no omitted variable bias exists in the full population (that is, $E[\epsilon_{ij} | \text{PhyRace}_i, X_{ij}] = 0$). Because realized interactions are not randomly drawn from the full set of feasible patient-physician interactions, conditioning on realization can systematically alter the distribution of patient and physician characteristics across physician racial groups, creating selection on observables.

IPTW addresses this source of bias by reweighting realized patient-physician interactions according to their inverse probability of realization. We first estimate the probability that a patient-physician interaction is realized as a function of observed patient characteristics, physician characteristics, and geographic proximity within specialty-year cells using a probit model. Patients are more likely to visit physicians who are geographically proximate and share their gender and race; more experienced physicians and those with U.S. medical degrees are also more likely to be selected (Table B2).

[Table B2]

Figure B2 provides a benchmark for these sorting patterns using the ambulance-admitted subsample, in which patients arrive by ambulance and have relatively limited discretion over physician selection. The figure plots the probability of a realized interaction across patient and physician racial groups. Compared with the full sample, racial sorting is weaker in this reduced-discretion setting, consistent with patient choice contributing to the sorting observed in realized interactions.

[Figure B2]

We then assign greater weight to realized interactions that have a lower predicted probability of being realized, using normalized inverse-probability weights. Intuitively, this procedure reconstructs a pseudo-population in which the distribution of observed patient-physician interactions more closely resembles the full set of feasible interactions. For illustration, Figure B3 plots the inverse-probability weights by physician and patient race. As expected, patient-physician interactions that are more likely to occur, namely racially concordant pairs, receive lower weights.

[Figure B3]

The IPTW adjustment relies on two key assumptions. First, it requires common support: Each patient-physician interaction in the feasible choice set must have a positive probability of being realized. To reduce the influence of extreme weights and improve overlap, we winsorize the predicted probabilities at the 2.5% and 97.5% levels. Second, IPTW assumes selection on observables: Conditional on observed patient, physician, and visit characteristics and geographic proximity, no unobserved factors jointly affect whether an interaction is realized and how the physician would be evaluated. Although this assumption cannot be tested directly, our rich set of covariates helps mitigate this concern.

B.3 Supplementary Details on Heckman Correction

To address the possibility that the selection-on-observables assumption may be violated, we also implement a Heckman correction model, which allows the unobserved determinants of sorting and evaluation to be correlated. Formally, the model specifies the latent propensity for a patient-physician interaction to be realized as

$$D_{ij}^* = Z'_{ij}\gamma + v_{ij}, \quad D_{ij} = 1[D_{ij}^* > 0]$$

where Z_{ij} includes at least one exclusion restriction—a variable that shifts the probability that a patient-physician interaction is realized but, conditional on the interaction occurring, does not directly affect the patient evaluation outcome.

When the unobserved determinants of sorting and evaluation are correlated, such that $\text{Cov}(v_{ij}, \epsilon_{ij}) \neq 0$, estimation of the LPM based on realized interactions will be biased. The Heckman model addresses this challenge by imposing the assumption that (v_{ij}, ϵ_{ij}) follow a jointly normal distribution with correlation ρ , and by augmenting the outcome equation with the inverse Mills ratio

$$\text{IMR}_{ij} = \phi(Z'_{ij}\gamma) / \Phi(Z'_{ij}\gamma)$$

evaluated at the first-stage probit estimates, so that

$$E[Y_{ij} | \text{PhyRace}_i, X_{ij}, D_{ij} = 1] = \alpha + \beta \text{PhyRace}_i + X'_{ij}\theta + \rho\sigma_\epsilon \text{IMR}_{ij}$$

The final term captures the extent to which selection into the realized sample is driven by unobserved factors that are correlated with the evaluation outcome. In this sense, it serves as a control function for non-random patient-physician matching that is not fully explained by observed covariates.

We use geographic proximity, defined as the logarithm of distance between physicians and patients, as the exclusion restriction. For distance to serve as a valid exclusion restriction, it should vary meaningfully among physicians available to a patient, predict which patient-physician interaction is realized, and not directly affect evaluations through other pathways.

Regarding the first condition, patients could plausibly have visited physicians across multiple locations in our setting. In the full feasible set, more than 99% of visits have alternative physicians in the same specialty, care setting, and year across more than one location, and 83% have alternatives across six or more locations (Table B3). This geographic variation is preserved after downsampling to the simulated choice set: More than 99% of visit-level choice sets span more than one location, 77% span four or more locations, and 29% span six or more locations.

[Table B3]

Second, distance predicts whether a patient is matched to a physician. As shown in Table B2, the logarithm of patient-physician distance is negatively associated with the realized-interaction indicator, suggesting that realized interactions are more likely to involve geographically proximate physicians. Third, conditional on the realized interaction and the controls in the outcome equation, distance should not independently affect how patients evaluate their physicians. Although this assumption is not directly testable, it is plausible in our setting given the rich set of patient, physician, and visit characteristics included in the model.

Because the role of distance in physician choice may differ across specialties, we examine this heterogeneity in two ways. Table B4 reports the average patient-physician distance by specialty in the simulated choice set, separately for realized and counterfactual interactions. Patients travel farther for specialist visits than for emergency care or generalist care (e.g., Emergency Medicine, General Internal Medicine, Family Medicine), with more discretionary specialties (e.g., Dermatology, Plastic and Reconstructive Surgery) involving even longer travel distances.

[Table B4]

We then examine whether this heterogeneity translates into differences in the selection process. Table B5 reports an alternative selection-stage specification that interacts specialty discretion with distance, physician experience, and educational prestige. Family Medicine, Emergency Medicine, General Internal Medicine, and related generalist fields are classified as non-discretionary. Distance is less predictive of whether an interaction is realized but remains a substantively meaningful predictor, supporting the relevance condition for its use as the exclusion restriction. To further account for heterogeneity across evaluative environments, the main analyses include hospital and care-setting fixed effects, while robustness checks replace care-setting fixed effects with the more granular medical-specialty fixed effects.

[Table B5]

A separate concern is that geographic proximity may proxy for patient characteristics that also influence evaluations. For example, patients willing to travel farther may differ in socioeconomic

background or expectations. To assess this concern more directly, we control for the median household income of patients' residential ZIP codes in both the selection and outcome stages, which helps account for institutional and socioeconomic differences that may jointly shape willingness to travel and evaluation behavior. The results remain similar with these controls (Table B6).

[Table B6]

B.4 Alternative Choice-Set Constructions and Alternative Selection-Stage Models

We consider two alternative choice-set constructions. First, we pair each realized visit with up to 20 rather than 10 counterfactual physicians. Second, we allow counterfactual physicians to be drawn from the same specialty and care setting within one year before or after the visit year. Re-estimating the IPTW and Heckman analyses under both alternative constructions yields similar results (Table B7).

[Table B7]

We also consider two alternative selection-stage models. First, in addition to interactions between physician race and patient race, we allow physician race to interact with physician credentials (education and experience), patient characteristics (patient age and insurance class), and geographic proximity. Second, instead of modeling patient-physician demographic similarity using fully interacted race and gender terms, we use binary indicators for racial concordance and gender concordance. Re-estimating the IPTW and Heckman analyses under both alternative models yields consistent results (Table B8).

[Table B8]

C. Supplementary Results for Physician Experience, Education, and Racial Disparities

Table C1 presents the detailed estimates for analyses of how physician experience and educational prestige moderate racial disparities in patient evaluations. Models 1 and 2 examine the interactions of physician race with experience and educational prestige separately, while Model 3 includes both sets of interactions simultaneously, with the estimated patterns remaining directionally consistent.

[Table C1]

Table C2 presents robustness checks for these analyses. Model 1 examines within-physician changes in *Top Box* responses as physicians gain experience using physician fixed effects. Models 2-5 use

alternative measures of *Educational Prestige* based on two medical-school rankings. Our primary measure uses *U.S. News & World Report* rankings from 2016-2020 (Models 2 and 3), while the alternative measure uses total National Institutes of Health (NIH) funding over 2016-2020 (Models 4 and 5). For each ranking, we consider alternative cutoffs at the top 20 and top 40 schools. The estimated patterns are robust across these alternative specifications, although the precision of the estimates varies.

[Table C2]

D. Supplementary Results for Evaluation Items and Racial Disparities

Table D1 presents item-level estimates of racial disparities across the 20 evaluation items, along with original and Holm-adjusted p -values. Because testing multiple evaluation items increases the likelihood of false positives, we use the Holm step-down procedure (Holm 1979) to control the familywise error rate (FWER). The procedure orders all p -values from smallest to largest ($p_1 \leq p_2 \leq \dots \leq p_{20}$) and sequentially adjusts them according to the number of hypotheses remaining to be tested, applying the strongest correction to the smallest p -values and progressively less stringent corrections thereafter. This approach controls the probability of making one or more Type I errors across the full set of item-level tests. The results remain consistent after adjustment. Notably, both Asian and Black physicians remain less likely to receive *Top Box* ratings on the evaluation item asking about patients' confidence in physician skill (Holm-adjusted $p < 0.001$ for Asian physicians and Holm-adjusted $p = 0.056$ for Black physicians).

[Table D1]

Table D2 presents the categorization of the evaluation items along two dimensions: (i) the concreteness and verifiability of the behavior being assessed, and (ii) the extent to which clear evaluative norms exist for that dimension of care. These principles yield four broad analytic categories: protocol adherence, basic interpersonal demeanor, engagement and communication, and overall assessment. Protocol-adherence items are both concrete and governed by clear norms, as they refer to identifiable procedural behaviors, such as explaining why a procedure mattered. Basic interpersonal-demeanor items are less behaviorally specific but are governed by relatively clear and widely shared norms, such as treating patients with respect or showing concern for privacy. By contrast, engagement-and-

communication items, such as whether the physician spent enough time, explained things in a way the patient understood, or kept the patient informed, require greater patient discretion in judging what constitutes effective communication in a medical setting. Overall-assessment items, such as confidence in the physician's skill and willingness to recommend the physician, are the least behaviorally anchored and ask patients to interpret physician quality without reference to any specific behavior.

[Table D2]

Figure D1 reports an internal meta-analysis to better summarize the racial disparities across categories. Specifically, we treat each item-level estimate as a unit of observation and compute category-level averages, weighting by the inverse of the variance so that more precisely estimated effects receive greater weight. The category-level estimates mirror the pattern in the item-level analysis: Racial disparities are smaller for protocol-adherence and basic-interpersonal-demeanor items and larger for overall-assessment items for both Asian-White and Black-White comparisons.

[Figure D1]

We conduct two additional analyses to assess the robustness of this interpretation. First, because some evaluation items are used in different care settings, Table D3 presents care-setting-specific meta-analyses for emergency, inpatient, and outpatient settings. Results are broadly consistent across care settings except in emergency care, where more time-constrained encounters may limit opportunities to observe differentiated physician behaviors. Second, because some items could plausibly fit multiple categories, Table D4 uses continuous concreteness and clarity scores rather than the four-category classification. Results are also consistent with the main analysis, with larger racial disparities for items that are less concrete and governed by less clear evaluative norms. Together, these analyses suggest that the pattern is not driven solely by care-setting differences or the categorization approach.

[Table D3]

[Table D4]

E. Robustness Checks

E.1 Racial Disparities by Patient Race

Table E1 examines whether racial disparities in patient evaluations vary by patient race using interactions between patient and physician race. Parallel to our main analyses, we estimate three versions of this specification: a baseline LPM, an IPTW-adjusted model, and a Heckman correction model. The results suggest that racial disparities are largest when Asian and Black physicians are evaluated by White patients.

[Table E1]

E.2 Placebo Tests

Table E2 reports placebo analyses using patient ratings of other dimensions of the care experience: facility quality, access, staff attentiveness to personal concerns, and nursing care. Facility and access ratings are primarily determined by hospital- or system-level factors and should therefore be less directly related to individual physician characteristics. By contrast, staff attentiveness and nursing care may still reflect physician race indirectly because patients may view physicians as part of the broader care team or because physician characteristics may shape interactions with other staff.

We construct *Top Box* indicators for each outcome and replicate the main analysis. Consistent with these expectations, we find no evidence of racial disparities in facility or access ratings, whereas disparities emerge for staff attentiveness and nursing care, suggesting that racial disparities are concentrated in domains where physician race is more likely to be salient.

[Table E2]

E.3 Survey Nonresponse Bias

Tables E3, E4, and Figure E1 assess potential survey nonresponse bias. Table E3 tests whether, among patients who answered at least one survey item, completion of the physician evaluation section varies by physician race. Without controls, there are no significant differences in missing evaluation rates by physician race (Model 1). After controlling for patient and physician characteristics and adding hospital and care-setting fixed effects, Asian and Black physicians are 1.2 ($p < 0.01$) and 0.6 ($p = 0.29$) percentage points less likely to have missing evaluations than White physicians (Model 2). Overall, we find no consistent evidence that missing-data patterns vary by physician race.

[Table E3]

Table E4 assesses whether the results are sensitive to the inclusion of low-scoring evaluations. If nonresponse bias arises primarily because dissatisfied patients are more likely to complete surveys after visits with racial minority physicians, the observed disparities should be concentrated among the lowest ratings. We therefore replicate the main analyses after excluding evaluations with scores below 25 and, separately, below 60. The estimated racial gaps remain similar across these specifications.

[Table E4]

Figure E1 provides a complementary test by plotting the full distributions of evaluation scores by physician race. The racial disparities persist across the score range rather than being concentrated in the lower tail. While we cannot fully rule out survey nonresponse bias, these analyses suggest that it is unlikely to be the sole driver of the observed racial disparities.

[Figure E1]

E.4 Sensitivity Analysis

Table E5 reports a sensitivity analysis for potential omitted-variable bias. The analysis assesses how strongly unobserved confounders would need to relate to physician race and evaluation outcomes, relative to the observed controls, to eliminate the estimated disparities. Following Oster (2019), we assume that unobserved confounders are as influential as the full set of observed controls ($\delta = 1$) and that including them would increase the model's explanatory power by 30% ($R^2_{\text{Max}} = 1.3 \times R^2$). Under these assumptions, the estimated racial gaps remain statistically significant. To fully explain away the disparities, unobserved confounders would need to be approximately 3.2 times as important as the observed controls for the Asian-White gap and 5.8 times as important for the Black-White gap.

[Table E5]

E.5 Alternative Specifications

Tables E6 and E7 assess the robustness of the findings to alternative specifications and samples. Table E6 replaces care-setting fixed effects with more granular medical-specialty fixed effects to account for heterogeneity in evaluative environments across specialties. The estimated racial disparities remain

similar, although precision declines as the more granular fixed effects reduce the number of racial minority physicians contributing to within-specialty comparisons.

[Table E6]

Table E7 considers three additional specifications. Model 1 restricts the sample to each patient's first survey response, addressing the possibility that patients who appear multiple times disproportionately influence the results. Model 2 estimates the *Top Box* outcome using a probit rather than a linear probability model, and Model 3 uses the continuous evaluation score rather than the *Top Box* indicator. The results remain similar across specifications.

[Table E7]

Table E8 examines whether patient health and prior patient-physician relationships help explain the observed disparities. Because these measures are available only for subsets of the sample, we first re-estimate the baseline specification within each relevant subsample before adding the additional control and its interaction with physician race.

[Table E8]

Models 1-3 examine patient health, measured using self-reported health from 1 ("poor") to 5 ("excellent") and available for 11% of responses. One concern is that racial minority physicians may disproportionately treat patients in worse health, who may in turn report lower satisfaction. The racial gaps remain evident in this restricted sample (Model 1) and are only modestly reduced after controlling for patient health (Model 2). Patients reporting better health provide more favorable evaluations. Interactions between patient health and physician race show no consistent evidence that racial disparities are concentrated among patients in better or worse health (Model 3).

Models 4-6 examine prior patient-physician relationships using an indicator for whether the patient regularly visits the treating physician, available for 27% of responses. Prior relationships may shape evaluations through greater familiarity, trust, or continuity of care and may also be unevenly distributed across physicians. The racial gaps are again evident in this restricted sample (Model 4) and remain similar after controlling for regular interaction (Model 5). Patients who regularly see the physician

provide more favorable evaluations overall. This suggests that regular interaction may attenuate racial disparities, although the estimates are imprecise and do not reach conventional levels of statistical significance (Model 6).

E.6 Matched-Sample Analysis

Table E9 applies coarsened exact matching (CEM) to improve balance in observed characteristics between White and racial minority physicians (Ho et al. 2007, Iacus et al. 2012). Given the consistent disparities faced by both Asian and Black physicians across all specifications, we combine them into a single minority category for this analysis. We construct matched strata based on physician characteristics (gender, education, years in practice) and patient characteristics (gender, race, age, insurance type).¹ This procedure matches 83,681 racial minority physician evaluations to 295,316 White physician evaluations across 1,919 strata. Covariate balance before and after matching is presented in Figure E2.

[Figure E2]

We estimate racial disparities based on the full sample (Model 1) and matched sample (Model 2). Both estimates yield consistent results: Racial minority physicians are 2.5 to 3.1 percentage points less likely to receive *Top Box* responses than their White peers ($p < 0.001$).

[Table E9]

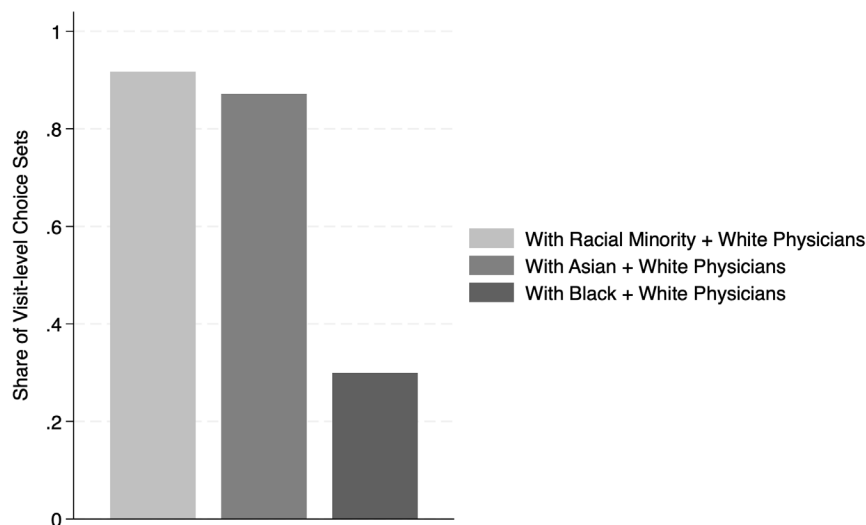
References

- Ho DE, Imai K, King G, Stuart EA (2007) Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* 15(3):199–236.
- Holm S (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*:65–70.
- Iacus SM, King G, Porro G (2012) Causal inference without balance checking: Coarsened exact matching. *Political Analysis* 20(1):1–24.
- Oster E (2019) Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics* 37(2):187–204.

¹ Physicians' education is measured using the U.S. News ranking of the medical school where they obtained their MD degree. The variable is coarsened with cutoff points at ranks 10, 20, 50, and 100. Years in practice is coarsened with cutoff points at 4, 8, and 12 years. Patient age is coarsened into 10-year bins between ages 30 and 70.

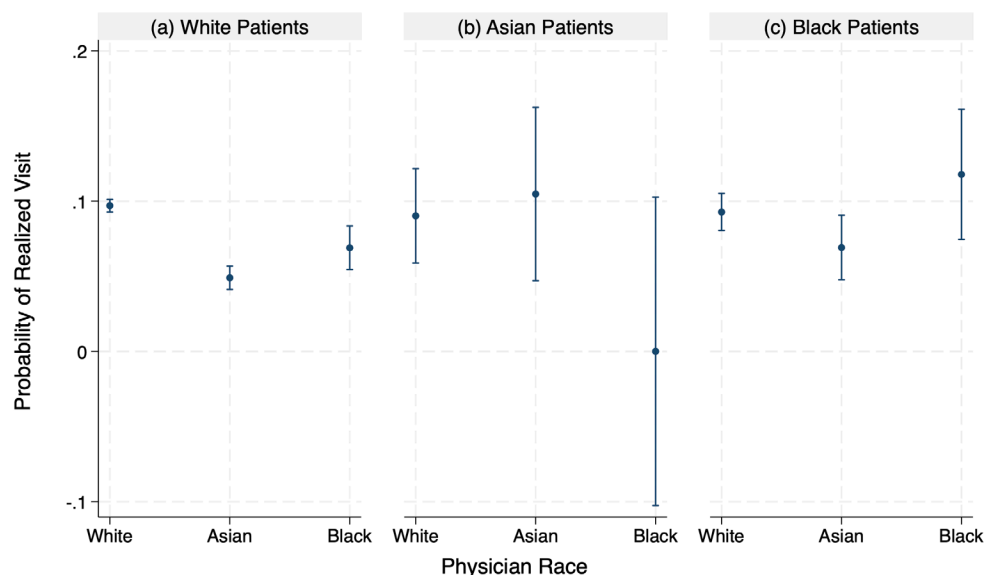
Figures and Tables

Figure B1. Racial Composition of the Simulated Choice Set



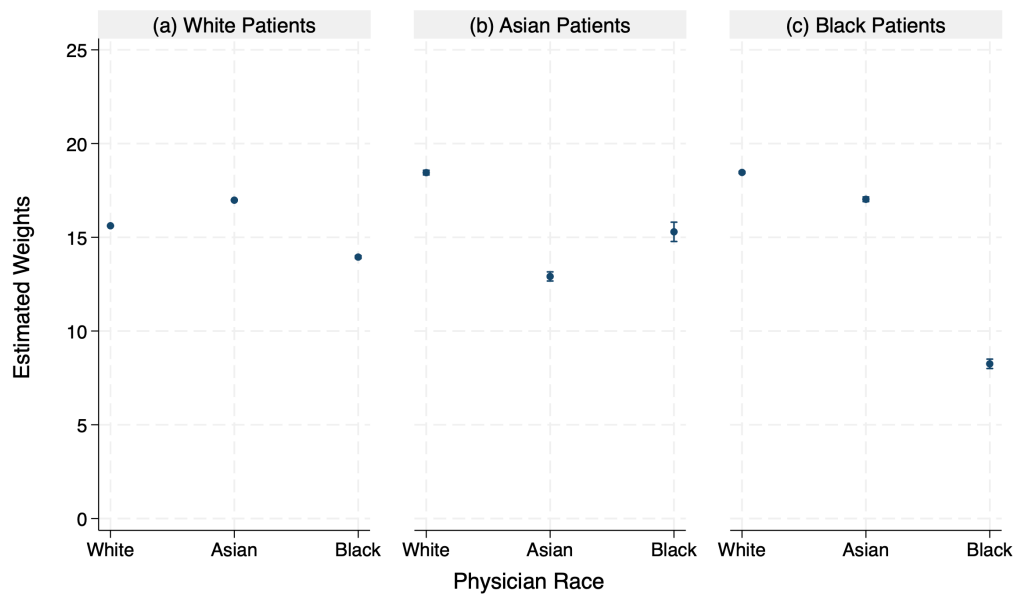
Note. The figure summarizes the racial composition of the simulated choice set. For each observed visit, the simulated choice set includes the treating physician and sampled counterfactual physicians who were active in the same specialty, care setting, and year. Bars report the share of visit-level choice sets that include both White and racial minority physicians (92%), both White and Asian physicians (87%), and both White and Black physicians (30%).

Figure B2. Probability of Visit by Patient and Physician Race: Ambulance Admitted



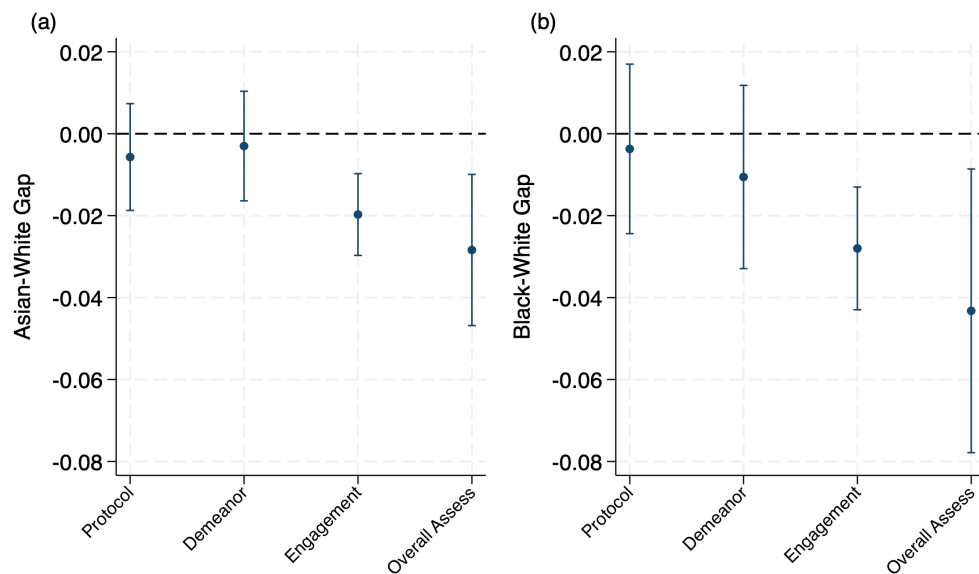
Note. The figure plots, among the subsample of visits in which patients were admitted through ambulance, the probability that a patient-physician pair in the simulated choice-set sample is realized, shown by physician race for (a) White, (b) Asian, and (c) Black patients. The simulated choice set combines each realized visit with up to 10 counterfactual physicians drawn from the same specialty, care setting, and year. Estimates are based only on patients admitted through ambulance. These estimates are descriptive and include no additional controls beyond specialty-year fixed effects. Error bars indicate 95% confidence intervals.

Figure B3. Inverse Probability of Treatment Weights by Patient and Physician Race



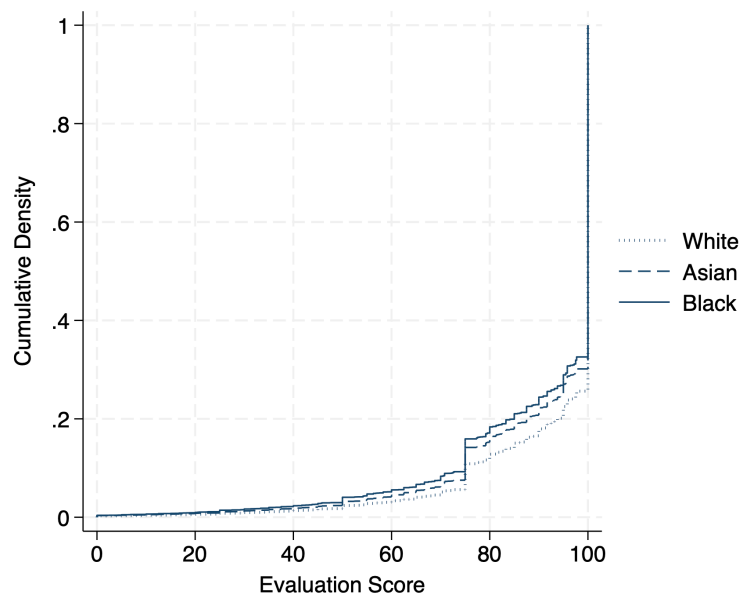
Note. The figure plots inverse probability of treatment weights by physician race for (a) White, (b) Asian, and (c) Black patients. Estimates are descriptive and do not include additional controls beyond specialty-year fixed effects. Error bars represent 95% confidence intervals.

Figure D1. Evaluation Items and Racial Disparities (Internal Meta-Analysis)



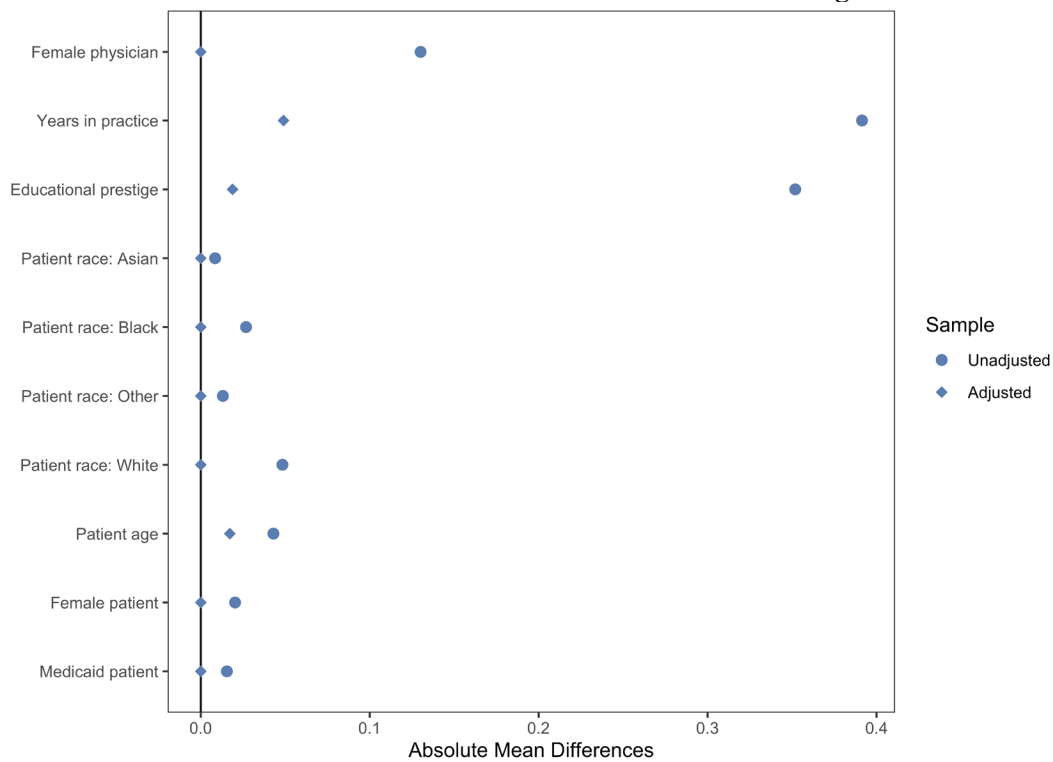
Note. The figure presents an internal meta-analysis of racial disparities by evaluation category. Positive (negative) estimates indicate higher (lower) probabilities of receiving a *Top Box* response relative to White physicians. Individual item-level effect sizes are weighted by the inverse of their variance. Error bars represent 95% confidence intervals.

Figure E1. Empirical Cumulative Distribution of Evaluation Scores by Physician Race



Note. This figure plots the empirical cumulative distribution of physicians' overall evaluation scores (0-100 scale, with higher values indicating more favorable outcomes). The curves show the cumulative proportion of visits at or below each score for White, Asian, and Black physicians.

Figure E2. Covariate Balance Before and After Coarsened Exact Matching



Note. Each point represents the absolute standardized mean difference between treated (Asian and Black physicians) and control (White physicians) groups on the listed covariates. Circles indicate mean differences before matching; diamonds show balance after coarsened exact matching (CEM). Values near 0 reflect improved covariate balance between treated and control observations.

Table A1. Physician Evaluation Items

Item Name	Item Description	Item Included in					
		Emergency	Medical Practice	Outpatient Oncology	Other Outpatient	Inpatient Ambulatory Surgery	Other Inpatient
MD's courtesy	Courtesy of the medical providers	Y	Y		Y	Y	Y
MD listened to you	How well the medical providers took the time to listen to you	Y					
MD included you in decisions	How well the medical providers included you in decisions about your treatment	Y	Y	Y			
MD kept you informed	Medical providers' concern to keep you informed about your treatment	Y	Y	Y			Y
MD's concern for your comfort	Medical providers' concern for your comfort while treating you	Y					
MD's response to your concerns	Concern the medical provider showed for your questions or worries		Y	Y	Y	Y	Y
MD explained in a way you understood	Explanations the medical provider gave you about your problem or condition		Y		Y		
MD discussed treatment options	Medical provider's discussion of any proposed treatment (options, risks, benefits, etc.)		Y	Y			
MD gave follow-up instructions	Instructions the medical provider gave you about follow-up cares (if any)		Y				
MD used understandable language	Degree to which medical provider talked with you using words you could understand		Y				
Time MD spent with you	Amount of time the medical provider spent with you		Y				Y
MD provided chance to ask questions	Medical provider provided opportunity to ask questions				Y		
MD responded to your complaints	Response to concerns/ complaints made during your visit				Y		
MD treated you with respect	Medical provider treated you with respect and dignity				Y		
MD explained the procedure	Medical provider's explanation about what the procedure would be like					Y	
MD explained results after	Information the medical provider shared about how the procedure went					Y	
MD explained why procedure mattered	Medical provider's explanation about why this procedure was important to your health					Y	
MD's concern for your privacy	Our concern for your privacy				Y		
Your confidence in MD's skill	Your confidence in the medical provider		Y	Y	Y		Y
Likelihood to recommend MD	Likelihood of your recommending this medical provider to others		Y				

Table B1. Summary Statistics and Pairwise Correlations for the Simulated Choice Set

Panel A: Summary Statistics															
Physician Race	Asian	Black	White	Other	Total										
Total Interaction Count	933,233	170,295	3,594,128	446,948	5,144,604										
Interaction Per Physician	1,019	1,064	1,200	880	1,124										
Realized (%)	8.34%	10.20%	9.49%	7.37%	9.12%										
Log(Distance)	2.91	2.83	2.98	2.96	2.96										
Mean Patient Age	62.00	60.12	60.76	60.43	60.93										
Female Patient (%)	57.25%	60.83%	58.00%	59.05%	58.05%										
White Patient (%)	76.28%	75.10%	75.19%	75.64%	75.42%										
Asian Patient (%)	1.59%	1.77%	1.54%	1.70%	1.57%										
Black Patient (%)	6.48%	7.66%	6.10%	7.08%	6.31%										
Medicaid Patient (%)	5.95%	6.65%	5.82%	6.14%	5.90%										

Panel B: Pairwise Correlations															
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
(1) Actual	1.00														
(2) Physician Race: Asian	-0.01	1.00													
(3) Physician Race: Black	0.01	-0.09	1.00												
(4) Physician Race: White	0.02	-0.70	-0.30	1.00											
(5) Years in Practice	0.05	-0.14	-0.06	0.20	1.00										
(6) Physician Edu: U.S. Excl. Top 50	0.01	-0.07	0.01	0.09	-0.01	1.00									
(7) Physician Edu: Non-U.S.	-0.01	0.20	0.00	-0.27	-0.09	-0.33	1.00								
(8) Female Physician	-0.01	0.06	0.10	-0.10	-0.11	0.00	0.01	1.00							
(9) Log(Distance)	-0.27	-0.02	-0.02	0.03	0.03	0.01	-0.01	-0.01	1.00						
(10) Patient Age	0.00	0.03	0.00	-0.02	0.00	-0.04	0.03	-0.10	-0.01	1.00					
(11) Female Patient	0.00	-0.01	0.01	0.00	0.01	0.01	0.01	0.09	-0.02	-0.11	1.00				
(12) Patient Race: Asian	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.01	-0.01	-0.07	0.02	1.00			
(13) Patient Race: Black	0.00	0.01	0.01	-0.02	0.00	0.01	0.01	0.02	-0.09	-0.09	0.06	-0.03	1.00		
(14) Patient Race: White	0.02	0.01	0.00	-0.01	0.06	0.00	0.01	-0.01	0.04	0.20	-0.03	-0.23	-0.46	1.00	
(15) Medicaid Patient	0.00	0.00	0.01	0.00	-0.02	0.01	-0.01	0.00	-0.05	-0.25	0.02	0.01	0.12	-0.17	1.00

Table B2. Endogenous Sorting between Patients and Physicians

	Realized Visit (1)
Log(Distance)	-0.471*** (0.001)
Asian Physician	-0.040*** (0.003)
Black Physician	0.054*** (0.006)
Other/Unknown Race Physician	-0.146*** (0.005)
Asian Patient	-0.137*** (0.010)
Black Patient	-0.244*** (0.005)
Other/Unknown Race Patient	-0.071*** (0.005)
Physician $\times$ Patient Race	
Asian $\times$ Asian	0.283*** (0.020)
Asian $\times$ Black	0.087*** (0.011)
Asian $\times$ Other	0.071*** (0.011)
Black $\times$ Asian	0.061 (0.040)
Black $\times$ Black	0.372*** (0.018)
Black $\times$ Other	0.117*** (0.019)
Other $\times$ Asian	-0.030 (0.032)
Other $\times$ Black	0.115*** (0.015)
Other $\times$ Other	0.160*** (0.014)
Female Physician	-0.212*** (0.004)
Female Patient	-0.118*** (0.003)
Female Physician $\times$ Female Patient	0.299*** (0.005)
Physician Years in Practice	0.031*** (0.000)
Physician from Non-Top-50 U.S. Medical School	0.052*** (0.002)
Physician from Non-U.S. Medical School	-0.024*** (0.003)
Medicaid Patient	-0.073*** (0.005)
Patient Age	-0.001*** (0.000)
Constant	-0.345*** (0.012)
Specialty-Year FE	Y
N =	3,127,825
Log pseudolikelihood	-819,394

Note. Model 1 is a probit model with robust standard errors estimated on the simulated choice-set sample. The dependent variable indicates whether a physician-patient interaction is realized. The model includes patient characteristics, physician characteristics, interactions between patient and physician race and gender, geographic proximity, and specialty-by-year fixed effects. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, + $p < 0.10$.

Table B3. Distribution of Feasible Physician Locations by Medical Specialty

Medical Specialty	Number of Locations with Feasible Physicians					
	1	2	3	4	5	6+
Dermatology	0%	0%	0%	0%	0%	99%
Emergency Medicine	0%	0%	0%	0%	1%	99%
Endocrinology, Diabetes, & Metabolism	0%	1%	1%	0%	0%	99%
Family Medicine	0%	0%	1%	2%	1%	96%
Pulmonary Disease, Critical Care and Sleep Medicine	0%	0%	0%	2%	3%	94%
General Internal Medicine	0%	0%	0%	0%	6%	94%
Pediatrics	0%	0%	1%	5%	0%	93%
Neurology	2%	2%	2%	2%	0%	92%
Cardiology (Internal Medicine)	0%	0%	2%	2%	4%	92%
Rheumatology (Internal Medicine)	1%	0%	0%	0%	9%	89%
Orthopedic Surgery	0%	0%	0%	1%	12%	86%
Surgery	0%	0%	0%	0%	14%	86%
Urology	0%	0%	0%	5%	15%	80%
Obstetrics & Gynecology	0%	0%	0%	0%	29%	71%
Hospitalist	0%	0%	4%	24%	3%	69%
Geriatric Medicine (Internal Medicine)	1%	12%	12%	9%	1%	64%
Gastroenterology (Internal Medicine)	0%	0%	8%	16%	13%	64%
Ophthalmology	0%	0%	1%	35%	14%	50%
Physical Medicine & Rehabilitation	9%	1%	14%	5%	26%	46%
Neurological Surgery	0%	0%	0%	37%	29%	34%
Plastic and Reconstructive Surgery (Surgery)	0%	1%	13%	31%	27%	27%
Radiology	0%	2%	2%	19%	51%	26%
Anesthesiology	7%	0%	13%	27%	31%	21%
Allergy and Immunology	1%	12%	5%	43%	35%	4%
Nephrology (Internal Medicine)	0%	0%	6%	65%	27%	1%
Infectious Disease (Internal Medicine)	12%	1%	0%	62%	24%	1%
Clinical Genetics	31%	43%	25%	1%	0%	0%
Oncology (Internal Medicine)	0%	87%	9%	3%	0%	0%
Oral & Maxillofacial	44%	56%	0%	0%	0%	0%
Other Internal Medicine	0%	52%	48%	0%	0%	0%
Pathology	0%	100%	0%	0%	0%	0%
Preventive Medicine	33%	6%	58%	3%	0%	0%
Total	0%	1%	1%	5%	9%	83%

Note. The table reports the distribution of the number of distinct physician locations in the full feasible set for each realized visit, separately by medical specialty. For each visit, the full feasible set includes all physicians active in the same specialty, care setting, and year, without the downsampling used to construct the simulated choice set.

Locations are identified using unique patient-physician distances, such that physicians at the same distance from the patient are treated as practicing at the same location. Percentages are calculated within specialty and may not sum to 100% because of rounding.

Table B4. Patient-Physician Distance in Realized and Counterfactual Interactions by Specialty

Medical Specialty	Mean Log(Distance)	
	Realized Interaction	Counterfactual Interaction
Other Internal Medicine	1.7	4.2
Family Medicine	1.7	3.4
General Internal Medicine	1.7	3.0
Emergency Medicine	1.8	2.9
Geriatric Medicine (Internal Medicine)	1.8	2.9
Hospitalist	1.9	2.7
Cardiology (Internal Medicine)	1.9	3.0
Anesthesiology	1.9	3.1
Gastroenterology (Internal Medicine)	2.0	2.9
Rheumatology (Internal Medicine)	2.0	3.0
Endocrinology, Diabetes, & Metabolism (Internal Medicine)	2.0	3.3
Oncology (Internal Medicine)	2.0	3.3
Obstetrics & Gynecology	2.1	3.1
Pediatrics	2.1	3.2
Pulmonary Disease, Critical Care and Sleep Medicine (Internal Medicine)	2.1	3.0
Radiology	2.1	3.1
Infectious Disease (Internal Medicine)	2.1	2.6
Physical Medicine & Rehabilitation	2.1	3.0
Ophthalmology	2.2	3.0
Urology	2.2	3.1
Surgery	2.3	3.0
Orthopedic Surgery	2.3	3.3
Nephrology (Internal Medicine)	2.4	3.0
Dermatology	2.4	2.9
Allergy and Immunology (Internal Medicine)	2.4	3.1
Neurology	2.4	3.1
Preventive Medicine	2.5	3.2
Neurological Surgery	2.6	3.4
Plastic and Reconstructive Surgery (Surgery)	2.6	3.1
Oral & Maxillofacial Surgery	2.7	3.6
Clinical Genetics	2.7	3.2
Pathology	4.3	4.3

Note. The table reports mean log physician-patient distance by medical specialty separately for realized and counterfactual interactions. For each observed visit, counterfactual interactions are constructed by pairing the patient with sampled physicians who were active in the same specialty, care setting, and year as the treating physician. Lower values indicate closer geographic proximity between patients and physicians.

Table B5. Patient-Physician Sorting by Specialty Discretion

	Realized Visit (1)
Log(Distance)	-0.541*** (0.002)
Physician Years in Practice	0.031*** (0.001)
Physician Education	
Non-Top-50 U.S. Medical School	0.192*** (0.005)
Non-U.S. Medical School	0.095*** (0.005)
Discretionary Specialty	-0.132*** (0.011)
Discretionary Specialty $\times$ Log(Distance)	0.114*** (0.003)
Discretionary Specialty $\times$ Physician Years in Practice	-0.000 (0.001)
Discretionary Specialty $\times$ Physician Education	
Non-Top-50 U.S. Medical School	-0.199*** (0.005)
Non-U.S. Medical School	-0.167*** (0.006)
Control Variables	
Physician Race	Y
Patient Race	Y
Physician $\times$ Patient Race	Y
Female Physician	Y
Female Patient	Y
Female Physician $\times$ Female Patient	Y
Medicaid Patient	Y
Patient Age	Y
Year FE	Y
Constant	-0.141*** (0.017)
N =	3,158,911
Log pseudolikelihood	-826,793

Note. Model 1 is a probit model with robust standard errors estimated on the simulated choice-set sample. The dependent variable indicates whether a physician-patient tie is realized. The specification parallels Table B2 but additionally interacts specialty discretion with geographic proximity, physician experience, and physician education. Discretionary Specialty is a binary indicator. Family Medicine, Emergency Medicine, General Internal Medicine, and related generalist fields are classified as non-discretionary; all other specialties are classified as discretionary.

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, + $p < 0.10$.

Table B6. Alternative Specification for Heckman Correction

Model	Heckman (1)
Physician Race	
Asian	-0.023* (0.009)
Black	-0.028 (0.017)
Other/Unknown	-0.004 (0.010)
Female Physician	0.006 (0.006)
Physician Years in Practice	0.006*** (0.001)
Physician Education	
Non-Top-50 U.S. Medical School	-0.004 (0.006)
Non-U.S. Medical School	-0.038*** (0.008)
Patient Race	
Asian	-0.076*** (0.007)
Black	-0.075*** (0.005)
Other/Unknown	-0.023*** (0.004)
Female Patient	-0.000 (0.003)
Medicaid Patient	-0.043*** (0.005)
Patient Age	-0.000*** (0.000)
Patient Residential Zip Code	0.000***
Median Household Income	(0.000)
Constant	0.367*** (0.037)
Hospital FE	Y
Care-Setting FE	Y
N =	3,125,949
Log pseudolikelihood	-923,964
ρ	0.063*** (0.007)

Note. Model 1 estimates the likelihood of *Top Box* responses by physician race using a Heckman correction parallel to the main results in Table 3, Model 3, but additionally controls for hospital fixed effects and the median household income of the patient's residential ZIP code in both the selection and outcome stages. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, + $p < 0.10$.

Table B7. Alternative Simulated Choice Set, IPTW, and Heckman

Sample Model	Enlarged Choice Set		Coarsened Year Matching	
	IPTW (1)	Heckman (2)	IPTW (3)	Heckman (4)
Physician Race				
Asian	-0.026** (0.008)	-0.023* (0.009)	-0.025** (0.008)	-0.024* (0.009)
Black	-0.033* (0.017)	-0.028+ (0.017)	-0.033* (0.016)	-0.028 (0.017)
Other/Unknown	-0.007 (0.010)	-0.007 (0.010)	-0.006 (0.010)	-0.008 (0.010)
Female Physician	0.007 (0.006)	0.007 (0.006)	0.007 (0.006)	0.006 (0.006)
Physician Years in Practice	0.007*** (0.001)	0.007*** (0.001)	0.007*** (0.001)	0.007*** (0.001)
Physician Education				
Non-Top-50 U.S. Medical School	-0.005 (0.006)	-0.003 (0.006)	-0.006 (0.006)	-0.003 (0.006)
Non-U.S. Medical School	-0.039*** (0.009)	-0.037*** (0.009)	-0.038*** (0.009)	-0.038*** (0.009)
Patient Race				
Asian	-0.073*** (0.008)	-0.077*** (0.007)	-0.075*** (0.009)	-0.077*** (0.007)
Black	-0.083*** (0.005)	-0.085*** (0.005)	-0.081*** (0.005)	-0.084*** (0.005)
Other/Unknown	-0.025*** (0.004)	-0.028*** (0.004)	-0.020*** (0.005)	-0.028*** (0.004)
Female Patient	-0.004 (0.003)	-0.001 (0.003)	-0.004 (0.003)	-0.001 (0.003)
Medicaid Patient	-0.047*** (0.006)	-0.048*** (0.005)	-0.046*** (0.006)	-0.048*** (0.005)
Patient Age	-0.000*** (0.000)	-0.000*** (0.000)	-0.001*** (0.000)	-0.000*** (0.000)
Constant	0.677*** (0.015)	0.362*** (0.040)	0.680*** (0.015)	0.372*** (0.039)
Hospital FE	Y	Y	Y	Y
Care-Setting FE	Y	Y	Y	Y
N =	280,547	5,936,443	280,571	2,900,878
R-squared	0.062		0.064	
Log pseudolikelihood		-1,160,280		-948,902
ρ		0.072*** (0.008)		0.068*** (0.007)

Note. Models 1 and 2 use the enlarged choice-set sample. Models 3 and 4 use the coarsened year-matching choice-set sample. Models 1 and 3 use IPTW to adjust for sorting on observables. Models 2 and 4 apply a Heckman correction. Across all models, the selection stage models the probability that a physician-patient interaction is realized using a probit model, controlling for patient characteristics, physician characteristics, patient-physician race and gender interactions, geographic proximity, and specialty-by-year fixed effects (Table B2). The outcome stage estimates the likelihood of *Top Box* responses by physician race, controlling for physician characteristics, patient characteristics, hospital fixed effects, and care-setting fixed effects. Standard errors are clustered at the physician level. White physicians, White patients, and physician education at top 50 U.S. medical schools are the reference categories. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, + $p < 0.10$.

Table B8. Alternative Selection-Stage Model, IPTW, and Heckman

Selection-Stage Model	Fully Interacted		Race and Gender Concordance	
	IPTW (1)	Heckman (2)	IPTW (3)	Heckman (4)
Physician Race				
Asian	-0.025** (0.008)	-0.024* (0.009)	-0.024** (0.008)	-0.026** (0.009)
Black	-0.040* (0.018)	-0.028+ (0.017)	-0.034* (0.017)	-0.034* (0.017)
Other/Unknown	-0.008 (0.010)	-0.006 (0.010)	-0.007 (0.010)	-0.007 (0.010)
Female Physician	0.009 (0.006)	0.007 (0.006)	0.009 (0.006)	0.008 (0.006)
Physician Years in Practice	0.007*** (0.001)	0.007*** (0.001)	0.007*** (0.001)	0.007*** (0.001)
Physician Education				
Non-Top-50 U.S. Medical School	-0.005 (0.006)	-0.003 (0.006)	-0.004 (0.006)	-0.003 (0.006)
Non-U.S. Medical School	-0.037*** (0.009)	-0.037*** (0.009)	-0.037*** (0.009)	-0.036*** (0.009)
Patient Race				
Asian	-0.080*** (0.009)	-0.077*** (0.007)	-0.078*** (0.009)	-0.077*** (0.007)
Black	-0.083*** (0.005)	-0.085*** (0.005)	-0.082*** (0.005)	-0.082*** (0.005)
Other/Unknown	-0.023*** (0.004)	-0.028*** (0.004)	-0.024*** (0.004)	-0.030*** (0.004)
Female Patient	-0.005 (0.003)	-0.001 (0.003)	-0.005 (0.003)	-0.002 (0.003)
Medicaid Patient	-0.046*** (0.006)	-0.048*** (0.005)	-0.046*** (0.005)	-0.048*** (0.005)
Patient Age	-0.000*** (0.000)	-0.000*** (0.000)	-0.000*** (0.000)	-0.000*** (0.000)
Constant	0.680*** (0.015)	0.376*** (0.039)	0.678*** (0.015)	0.369*** (0.039)
Hospital FE	Y	Y	Y	Y
Care-Setting FE	Y	Y	Y	Y
N =	280,547	3,127,853	280,571	3,434,464
R-squared	0.061		0.061	
Log pseudolikelihood		-981,512		-1,000,168
ρ		0.066*** (0.007)		0.074*** (0.007)

Note. Models 1 and 3 parallel the IPTW analysis in the main results (Table 3, Model 2) but use two alternative specifications for the selection stage. Model 1 adds interactions between physician race and physician experience, education, patient age, and insurance class. Model 3 replaces the physician-patient race and gender interactions with binary indicators for whether the physician-patient pair shares the same race and the same gender, respectively. Models 2 and 4 present Heckman corrections based on these alternative selection-stage specifications. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, + $p < 0.10$.

Table C1. Physician Experience, Education, and Racial Disparities

		Top Box	
	(1)	(2)	(3)
Physician Race			
Asian	-0.114*** (0.023)	-0.034*** (0.009)	-0.117*** (0.024)
Black	-0.119* (0.056)	-0.047* (0.023)	-0.117* (0.054)
Other/Unknown	-0.104*** (0.027)	-0.007 (0.011)	-0.103*** (0.028)
Years in Practice	0.004*** (0.001)	0.006*** (0.001)	0.004*** (0.001)
Educational Prestige	0.012** (0.005)	0.010+ (0.006)	0.011+ (0.005)
Years in Practice × Physician Race			
Asian	0.007*** (0.002)		0.007*** (0.002)
Black	0.007 (0.005)		0.006 (0.005)
Other/Unknown	0.008*** (0.002)		0.008*** (0.002)
Educational Prestige × Physician Race			
Asian		0.015 (0.014)	0.012 (0.014)
Black		0.025 (0.032)	0.017 (0.034)
Other/Unknown		-0.006 (0.016)	-0.009 (0.016)
Female Physician	0.009+ (0.005)	0.009+ (0.005)	0.009+ (0.005)
Patient Race			
Asian	-0.076*** (0.006)	-0.076*** (0.006)	-0.076*** (0.006)
Black	-0.089*** (0.004)	-0.089*** (0.004)	-0.089*** (0.004)
Other/Unknown	-0.023*** (0.003)	-0.023*** (0.003)	-0.023*** (0.003)
Female Patient	-0.008** (0.002)	-0.007** (0.002)	-0.008** (0.002)
Medicaid Patient	-0.046*** (0.005)	-0.046*** (0.005)	-0.046*** (0.005)
Patient Age	-0.000** (0.000)	-0.000** (0.000)	-0.000** (0.000)
Constant	0.711*** (0.012)	0.680*** (0.011)	0.711*** (0.012)
N =	410,479	410,479	410,479
R-squared	0.045	0.045	0.045

Note. All models are linear probability models with robust standard errors clustered at the physician level. The models control for physician characteristics (gender and credentials), patient characteristics (gender, race, age, and insurance), and hospital and care-setting fixed effects. White physicians, White patients, and physicians outside the educational-prestige category are the reference categories. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, + $p < 0.10$.

Table C2. Robustness Checks for Physician Experience, Education, and Racial Disparities

Educational Prestige Definition	Top 50 US News Rank (1)	Top 20 US News Rank (2)	Top Box		
			Top 40 US News Rank (3)	Top 20 US NIH Rank (4)	Top 40 US NIH Rank (5)
Physician Race					
Asian		-0.030*** (0.008)	-0.030** (0.009)	-0.034*** (0.008)	-0.030*** (0.009)
Black		-0.045* (0.021)	-0.047* (0.022)	-0.048* (0.021)	-0.044+ (0.023)
Other/Unknown		-0.004 (0.010)	-0.004 (0.011)	-0.008 (0.009)	-0.006 (0.010)
Years in Practice	0.009*** (0.001)	0.006*** (0.001)	0.006*** (0.001)	0.006*** (0.001)	0.006*** (0.001)
Educational Prestige		0.018** (0.006)	0.018** (0.006)	0.009 (0.006)	0.018** (0.006)
Years in Practice × Physician Race					
Asian	0.004 (0.003)				
Black	0.002 (0.006)				
Other/Unknown	0.001 (0.003)				
Educational Prestige × Physician Race					
Asian		0.004 (0.016)	0.006 (0.014)	0.015 (0.017)	0.006 (0.015)
Black		0.021 (0.033)	0.028 (0.032)	0.030 (0.034)	0.014 (0.033)
Other/Unknown		-0.026 (0.018)	-0.017 (0.016)	-0.015 (0.020)	-0.012 (0.016)
Female Physician		0.009+ (0.005)	0.009+ (0.005)	0.009+ (0.005)	0.009+ (0.005)
Patient Race					
Asian	-0.071*** (0.006)	-0.076*** (0.006)	-0.076*** (0.006)	-0.076*** (0.006)	-0.076*** (0.006)
Black	-0.072*** (0.004)	-0.089*** (0.004)	-0.089*** (0.004)	-0.089*** (0.004)	-0.089*** (0.004)
Other/Unknown	-0.015*** (0.003)	-0.022*** (0.003)	-0.022*** (0.003)	-0.023*** (0.003)	-0.022*** (0.003)
Female Patient	-0.012*** (0.002)	-0.007** (0.002)	-0.007** (0.002)	-0.008** (0.002)	-0.008** (0.002)
Medicaid Patient	-0.028*** (0.004)	-0.046*** (0.005)	-0.046*** (0.005)	-0.046*** (0.005)	-0.046*** (0.005)
Patient Age	0.001*** (0.000)	-0.000** (0.000)	-0.000** (0.000)	-0.000** (0.000)	-0.000** (0.000)
Constant	0.587*** (0.013)	0.680*** (0.011)	0.678*** (0.011)	0.683*** (0.011)	0.680*** (0.011)
Physician FE	Y	N	N	N	N
N =	416,497	410,479	410,479	410,479	410,479
R-squared	0.085	0.045	0.045	0.045	0.045

Note. All models are linear probability models with robust standard errors clustered at the physician level. Model 1 includes physician fixed effects and controls for patient characteristics (gender, race, age, and insurance). Models 2-5 do not include physician fixed effects but control for physician characteristics (gender and credentials), patient characteristics, and hospital and care-setting fixed effects. White physicians, White patients, and physicians outside the educational-prestige category are the reference categories. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, + $p < 0.10$.

Table D1. Racial Disparities by Evaluation Items

Evaluation Item	Physician Race: Black				Physician Race: Asian			
	Coef.	S.E.	<i>p</i>	Holm-adjusted <i>p</i>	Coef.	S.E.	<i>p</i>	Holm-adjusted <i>p</i>
MD provided chance to ask questions	0.020	0.016	0.226	1.000	0.007	0.007	0.355	1.000
MD explained why procedure mattered	0.015	0.018	0.390	1.000	0.010	0.010	0.334	1.000
MD explained results after	0.008	0.023	0.747	1.000	-0.001	0.011	0.905	1.000
MD explained the procedure	0.000	0.017	0.978	1.000	0.000	0.009	0.991	1.000
MD discussed treatment options	-0.025	0.015	0.095	0.954	-0.018	0.007	0.009	0.093
MD gave follow-up instructions	-0.034	0.019	0.068	0.748	-0.033	0.010	0.002	0.028
MD's concern for your privacy	0.011	0.014	0.430	1.000	0.004	0.006	0.509	1.000
MD treated you with respect	-0.003	0.016	0.857	1.000	0.005	0.006	0.372	1.000
MD's concern for your comfort	-0.006	0.012	0.581	1.000	-0.005	0.010	0.636	1.000
MD's courtesy	-0.036	0.015	0.019	0.297	-0.024	0.007	0.001	0.014
MD responded to your complaints	0.015	0.014	0.292	1.000	0.004	0.006	0.545	1.000
MD used understandable language	-0.022	0.015	0.160	1.000	-0.028	0.009	0.003	0.037
MD included you in decisions	-0.027	0.014	0.053	0.683	-0.021	0.007	0.003	0.041
MD kept you informed	-0.032	0.014	0.026	0.388	-0.022	0.008	0.004	0.044
MD listened to you	-0.033	0.016	0.039	0.541	-0.022	0.011	0.052	0.471
MD explained in a way you understood	-0.036	0.013	0.005	0.104	-0.024	0.006	0.000	0.001
MD's response to your concerns	-0.038	0.015	0.010	0.177	-0.020	0.006	0.001	0.014
Time MD spent with you	-0.043	0.015	0.006	0.104	-0.033	0.009	0.000	0.004
Likelihood to recommend MD	-0.033	0.017	0.054	0.683	-0.021	0.008	0.005	0.052
Your confidence in MD's skill	-0.052	0.017	0.003	0.056	-0.033	0.006	0.000	0.000

Note. The table shows the probability that Asian and Black physicians receive a *Top Box* response on each evaluation item relative to White physicians. A positive (negative) value indicates a higher (lower) probability. Effects are estimated with linear probability models controlling for physician characteristics, patient characteristics, and hospital and care-setting fixed effects. Standard errors are robust and clustered at the physician level.

Table D2. Categorization of Evaluation Items

Evaluation Item	Category	Concreteness Score	Clarity Score
MD provided chance to ask questions	Protocol Adherence	3	3
MD explained why procedure mattered		3	3
MD explained results after		3	3
MD explained the procedure		3	3
MD discussed treatment options		3	3
MD gave follow-up instructions		3	3
MD's concern for your privacy	Basic Interpersonal Demeanor	2	3
MD treated you with respect		2	3
MD's concern for your comfort		2	3
MD's courtesy		2	3
MD responded to your complaints	Engagement and Communication	2	2
MD used understandable language		2	1
MD included you in decisions		2	1
MD kept you informed		2	1
MD listened to you		2	2
MD explained in a way you understood		2	1
MD's response to your concerns		2	2
Time MD spent with you		2	1
Likelihood to recommend MD	Overall Assessment	1	1
Your confidence in MD's skill		1	1

Note. The table presents our classification of physician evaluation items. Concreteness captures the extent to which an item refers to a specific, observable physician behavior; clarity captures the extent to which widely shared evaluative norms exist for the dimension being assessed. Both scores range from 1 to 3, with higher values indicating greater concreteness or clearer norms.

Table D3. Racial Disparities across Evaluation Items and Care Setting (Internal Meta-Analysis)

Care Setting	Emergency		Inpatient		Outpatient	
Outcome	Asian-White	Black-White	Asian-White	Black-White	Asian-White	Black-White
Variable	Gap	Gap	Gap	Gap	Gap	Gap
	(1)	(2)	(3)	(4)	(5)	(6)
Basic Interpersonal Demeanor			-0.032* (0.010)	-0.070** (0.009)	0.011 (0.010)	0.010 (0.015)
Engagement and Communication	0.002 (0.002)	0.013 (0.007)	-0.039** (0.007)	-0.067*** (0.005)	-0.004 (0.009)	-0.006 (0.013)
Overall Assessment			-0.051** (0.010)	-0.089*** (0.008)	-0.010 (0.011)	-0.017 (0.018)
Constant	-0.005+ (0.002)	-0.017+ (0.006)	0.003 (0.004)	0.008+ (0.004)	-0.011 (0.008)	-0.012 (0.011)
Number of Evaluation Items	5	5	8	8	15	15
R-squared	0.202	0.507	0.923	0.985	0.304	0.199

Note. The table reports results from internal meta-analyses of racial disparities across evaluation categories by care setting (Models 1 and 2 for emergency care, Models 3 and 4 for inpatient care, Models 5 and 6 for outpatient care). Item-level Asian-White and Black-White gaps in the probability of receiving a *Top Box* response are regressed on evaluation-category indicators, weighted by the inverse of their variance. Basic Interpersonal Demeanor is the omitted category for emergency care, and Protocol Adherence is the omitted category for inpatient and outpatient care. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, + $p < 0.10$.

Table D4. Racial Disparities by Concreteness and Clarity of Evaluation Items

Outcome	Asian-White	Black-White	Asian-White	Black-White
Variable	Gap	Gap	Gap	Gap
	(1)	(2)	(3)	(4)
Concreteness Score	0.010+ (0.005)	0.018* (0.008)		
Clarity Score			0.011** (0.003)	0.014** (0.004)
Constant	-0.035** (0.012)	-0.057** (0.018)	-0.035*** (0.006)	-0.048*** (0.010)
Number of Evaluation Items	20	20	20	20
R-squared	0.174	0.209	0.444	0.367

Note. The table reports item-level regressions relating racial disparities in patient evaluations to the concreteness and clarity of the evaluation items. Each item-level gap is estimated separately, and the regressions are weighted by the inverse of its variance. Concreteness Score captures the extent to which an item refers to a specific, observable physician behavior, and Clarity Score captures the extent to which widely shared evaluative norms exist for the dimension being assessed. Both scores range from 1 to 3, with higher values indicating greater concreteness or clearer evaluative norms. Standard errors are reported in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, + $p < 0.10$.

Table E1. Physician Race, Patient Race, and Top Box Responses

Model	LPM (1)	Top Box IPTW (2)	Heckman (3)
Asian Physician	-0.024** (0.007)	-0.029*** (0.009)	-0.027** (0.010)
Black Physician	-0.047** (0.018)	-0.042* (0.018)	-0.039* (0.018)
Other/Unknown Race Physician	-0.006 (0.009)	-0.007 (0.010)	-0.010 (0.010)
Asian Patient	-0.090*** (0.008)	-0.094*** (0.011)	-0.095*** (0.009)
Black Patient	-0.095*** (0.005)	-0.089*** (0.006)	-0.092*** (0.006)
Other/Unknown Race Patient	-0.025*** (0.004)	-0.029*** (0.005)	-0.034*** (0.005)
Physician × Patient Race: Asian × Asian	0.035* (0.016)	0.054** (0.021)	0.053** (0.018)
Asian × Black	0.012 (0.012)	0.019 (0.014)	0.010 (0.013)
Asian × Other	0.010 (0.009)	0.013 (0.013)	0.019+ (0.011)
Black × Asian	0.024 (0.026)	0.000 (0.042)	0.017 (0.026)
Black × Black	0.055** (0.019)	0.061** (0.023)	0.063** (0.021)
Black × Other	0.022 (0.020)	0.027 (0.021)	0.027 (0.020)
Other × Asian	0.071** (0.023)	0.045 (0.039)	0.074* (0.029)
Other × Black	0.010 (0.013)	-0.010 (0.015)	0.004 (0.014)
Other × Other	0.007 (0.012)	0.011 (0.016)	0.015 (0.014)
Female Physician	0.006 (0.006)	0.005 (0.007)	-0.004 (0.007)
Female Patient	-0.009** (0.003)	-0.007+ (0.003)	-0.006+ (0.003)
Female Physician × Female Patient	0.005 (0.005)	0.006 (0.006)	0.016** (0.006)
Physician Years in Practice	0.006*** (0.001)	0.007*** (0.001)	0.007*** (0.001)
Physician Education: Non-Top-50 U.S.	-0.002 (0.005)	-0.005 (0.006)	-0.003 (0.006)
Non-U.S. Medical School	-0.031*** (0.007)	-0.037*** (0.009)	-0.037*** (0.009)
Medicaid Patient	-0.046*** (0.005)	-0.046*** (0.006)	-0.047*** (0.005)
Patient Age	-0.000** (0.000)	-0.000*** (0.000)	-0.000*** (0.000)
Constant	0.692*** (0.012)	0.683*** (0.015)	0.376*** (0.039)
Hospital and Care-Setting FE	Y	Y	Y
N =	410,479	280,547	3,127,853
R-squared (or Log pseudolikelihood)	0.045	0.061	-983,121
ρ			0.073*** (0.008)

Note. Models 1, 2, and 3 replicate Models 1, 2, and 3 of Table 3 but additionally include interactions between physician and patient gender and race. White physicians, White patients, and physicians educated at top 50 U.S. medical schools are the reference categories. Standard errors are clustered at the physician level. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, + $p < 0.10$

Table E2. Physician Race and Non-Physician Evaluation Outcomes

	Top Box Ratings on			
	Facility	Access	Staff Attentiveness	Nurse
	(1)	(2)	(3)	(4)
Physician Race				
Asian	-0.015 (0.011)	-0.008 (0.006)	-0.013* (0.006)	-0.012* (0.006)
Black	-0.043 (0.032)	-0.015 (0.015)	-0.020 (0.013)	-0.027* (0.013)
Other/Unknown	0.010 (0.016)	-0.005 (0.008)	-0.006 (0.007)	-0.012 (0.008)
Female Physician	-0.016+ (0.009)	-0.003 (0.005)	0.004 (0.004)	0.003 (0.004)
Physician Years in Practice	0.013*** (0.001)	0.007*** (0.001)	0.007*** (0.001)	0.006*** (0.001)
Physician Education				
Non-Top-50 U.S. Medical School	-0.007 (0.009)	-0.000 (0.005)	-0.003 (0.005)	-0.004 (0.005)
Non-U.S. Medical School	-0.053*** (0.013)	-0.013+ (0.007)	-0.020** (0.006)	-0.020** (0.007)
Patient Race				
Asian	-0.049*** (0.010)	-0.071*** (0.007)	-0.095*** (0.007)	-0.111*** (0.007)
Black	-0.021*** (0.006)	-0.042*** (0.004)	-0.091*** (0.004)	-0.080*** (0.004)
Other/Unknown	0.029*** (0.006)	-0.016*** (0.003)	-0.023*** (0.003)	-0.029*** (0.003)
Female Patient	-0.006 (0.004)	-0.006* (0.003)	-0.008*** (0.002)	-0.023*** (0.002)
Medicaid Patient	0.025*** (0.006)	-0.006 (0.004)	-0.024*** (0.004)	-0.032*** (0.004)
Patient Age	0.000 (0.000)	0.001*** (0.000)	0.000+ (0.000)	0.000 (0.000)
Constant	0.423*** (0.021)	0.479*** (0.012)	0.595*** (0.011)	0.674*** (0.011)
N =	198,698	400,176	365,956	328,643
R-squared	0.064	0.045	0.075	0.043

Note. All models are linear probability models with robust standard errors clustered at the physician level. The models control for physician characteristics, patient characteristics, and hospital and care-setting fixed effects. White physicians, White patients, and physician education at top 50 U.S. medical schools are the reference categories. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, + $p < 0.10$

Table E3. Probability of a Missing Physician Evaluation by Physician Race

	Missing Physician Evaluation	
	(1)	(2)
Physician Race		
Asian	-0.021 (0.021)	-0.012** (0.004)
Black	-0.035 (0.041)	-0.006 (0.006)
Other/Unknown	-0.003 (0.030)	-0.011* (0.005)
Female Physician		-0.011*** (0.003)
Physician Years in Practice		-0.006*** (0.001)
Physician Education		
Non-Top-50 U.S. Medical School		-0.005 (0.003)
Non-U.S. Medical School		-0.004 (0.004)
Patient Race		
Asian		-0.001 (0.003)
Black		-0.002 (0.002)
Other/Unknown		-0.001 (0.002)
Female Patient		0.010*** (0.001)
Medicaid Patient		0.012*** (0.002)
Patient Age		0.001*** (0.000)
Constant	0.187*** (0.010)	0.100*** (0.012)
Hospital FE		Y
Care-Setting FE		Y
N =	573,847	447,144
R-squared	0.001	0.135

Note. Both models are linear probability models with robust standard errors clustered at the physician level. Model 1 reports descriptive results by physician race. Model 2 additionally controls for physician characteristics, patient characteristics, and hospital and care-setting fixed effects. White physicians, White patients, and physician education at top 50 U.S. medical schools are the reference categories. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, + $p < 0.10$

Table E4. Physician Race and Top Box Responses (Excluding Extremely Negative Evaluations)

Sample	Top Box	
	Evaluation Score ≥ 25 (1)	Evaluation Score ≥ 60 (2)
Physician Race		
Asian	-0.020** (0.007)	-0.017** (0.006)
Black	-0.036* (0.016)	-0.030* (0.014)
Other/Unknown	-0.003 (0.008)	-0.001 (0.007)
Female Physician	0.008+ (0.005)	0.007+ (0.004)
Physician Years in Practice	0.006*** (0.001)	0.006*** (0.001)
Physician Education		
Non-Top-50 U.S. Medical School	-0.002 (0.005)	-0.002 (0.005)
Non-U.S. Medical School	-0.031*** (0.007)	-0.027*** (0.006)
Patient Race		
Asian	-0.078*** (0.006)	-0.071*** (0.006)
Black	-0.088*** (0.004)	-0.085*** (0.004)
Other/Unknown	-0.022*** (0.003)	-0.022*** (0.003)
Female Patient	-0.006* (0.002)	-0.000 (0.002)
Medicaid Patient	-0.040*** (0.004)	-0.029*** (0.004)
Patient Age	-0.000*** (0.000)	-0.001*** (0.000)
Constant	0.701*** (0.012)	0.734*** (0.010)
Hospital FE	Y	Y
Care-Setting FE	Y	Y
N =	407,503	396,078
R-squared	0.043	0.035

Note. Both models estimate the likelihood of *Top Box* responses by physician race using a linear probability model, controlling for other physician characteristics (gender, credentials), patient characteristics (gender, race, age, insurance), hospital fixed effects, and care-setting fixed effects. Model 1 is based on observed evaluations with scores of 25 or higher. Model 2 is based on observed evaluations with scores of 60 or higher. White physicians, White patients, and physician education at top 50 U.S. medical schools are the reference categories. Standard errors are clustered at the physician level. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, + $p < 0.10$

Table E5. Oster Bounds for Estimated Racial Disparities

	$\delta = 1$			δ s.t. $\beta = 0$
	β_{adj}	Confidence Interval	Reject Null?	
Physician Race: Asian	-0.016	[-.0199, -.0115]	Yes	3.194
Physician Race: Black	-0.032	[-.0403, -.0235]	Yes	5.813

Note. The table reports Oster bounds with $R^2_{Max} = 1.3 \times R^2$. β_{adj} is the bias-adjusted coefficient at $\delta = 1$; confidence intervals are estimated using bootstrap standard errors. δ s.t. $\beta = 0$ reports δ required to shrink the adjusted coefficient to zero.

Table E6. Physician Race and Top Box Responses: Medical-Specialty Fixed Effects

Model	LPM (1)	Top Box IPTW (2)	Heckman (3)
Physician Race			
Asian	-0.017* (0.007)	-0.015+ (0.008)	-0.014 (0.009)
Black	-0.029+ (0.016)	-0.026+ (0.016)	-0.023 (0.016)
Other/Unknown	0.001 (0.008)	0.002 (0.009)	0.004 (0.010)
Female Physician	0.015** (0.005)	0.015* (0.006)	0.012* (0.006)
Physician Years in Practice	0.005*** (0.001)	0.006*** (0.001)	0.006*** (0.001)
Physician Education			
Non-Top-50 U.S. Medical School	0.003 (0.006)	-0.000 (0.006)	0.002 (0.006)
Non-U.S. Medical School	-0.020** (0.007)	-0.020* (0.008)	-0.023** (0.009)
Patient Race			
Asian	-0.074*** (0.006)	-0.079*** (0.009)	-0.076*** (0.007)
Black	-0.086*** (0.004)	-0.079*** (0.005)	-0.081*** (0.005)
Other/Unknown	-0.023*** (0.003)	-0.026*** (0.004)	-0.030*** (0.004)
Female Patient	-0.007** (0.002)	-0.007** (0.003)	-0.005+ (0.002)
Medicaid Patient	-0.041*** (0.004)	-0.041*** (0.005)	-0.042*** (0.005)
Patient Age	0.000 (0.000)	-0.000 (0.000)	0.000 (0.000)
Constant	0.670*** (0.012)	0.664*** (0.015)	0.359*** (0.046)
Hospital FE	Y	Y	Y
Medical-Specialty FE	Y	Y	Y
N =	405,766	280,547	3,127,853
R-squared	0.047	0.068	
Log pseudolikelihood			-982,446
ρ			0.060*** (0.006)

Note. Model 1 estimates the likelihood of *Top Box* responses by physician race using a linear probability model (LPM) based on observed evaluations. Model 2 reweights observations using inverse probability of treatment weights (IPTW) to adjust for sorting on observables. Model 3 applies Heckman correction that allows unobserved determinants of sorting and evaluations to be correlated. All models control for other physician characteristics (gender, credentials), patient characteristics (gender, race, age, insurance), hospital fixed effects, and medical-specialty fixed effects. The IPTW and Heckman adjustments additionally require patients' and physicians' ZIP codes to calculate geographic proximity, which is available for 69% of the full sample. White physicians, White patients, and physician education at top 50 U.S. medical schools are the reference categories. Standard errors are clustered at the physician level. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, + $p < 0.10$

Table E7. Physician Race and Top Box Responses: Alternative Specifications

Sample Outcome Model	Unique Patients Top Box LPM (1)	All Visits Top Box Probit (2)	All Visits Evaluation Score OLS (3)
Physician Race			
Asian	-0.022** (0.007)	-0.067** (0.021)	-0.859*** (0.227)
Black	-0.038* (0.018)	-0.114* (0.047)	-1.425* (0.587)
Other/Unknown	-0.004 (0.009)	-0.014 (0.025)	-0.297 (0.291)
Female Physician	0.008 (0.005)	0.029+ (0.015)	0.344* (0.161)
Physician Years in Practice	0.006*** (0.001)	0.018*** (0.002)	0.168*** (0.028)
Physician Education			
Non-Top-50 U.S. Medical School	-0.003 (0.006)	-0.009 (0.017)	-0.118 (0.176)
Non-U.S. Medical School	-0.031*** (0.007)	-0.096*** (0.021)	-0.898*** (0.225)
Patient Race			
Asian	-0.077*** (0.007)	-0.231*** (0.018)	-1.933*** (0.203)
Black	-0.088*** (0.004)	-0.259*** (0.012)	-2.285*** (0.145)
Other/Unknown	-0.020*** (0.003)	-0.073*** (0.010)	-0.439*** (0.107)
Female Patient	-0.005* (0.002)	-0.021** (0.007)	-0.696*** (0.076)
Medicaid Patient	-0.042*** (0.005)	-0.126*** (0.013)	-2.560*** (0.198)
Patient Age	-0.000*** (0.000)	-0.001* (0.000)	0.017*** (0.003)
Constant	0.689*** (0.012)	-0.191* (0.091)	91.823*** (0.425)
Hospital FE	Y	Y	Y
Care-Setting FE	Y	Y	Y
N =	361,788	410,479	410,479
R-squared	0.048		0.056
Log pseudolikelihood		-230,393	

Note. All models estimate the evaluation outcome by physician race while controlling for physician characteristics, patient characteristics, hospital fixed effects, and care-setting fixed effects. Models 2 and 3 are based on all observed responses, whereas Model 1 is restricted to the first survey response for each unique patient. Models 1 and 2 use *Top Box* as the outcome, whereas Model 3 uses the raw evaluation score. Models 1 and 3 are estimated using linear models, whereas Model 2 is estimated using a probit model. White physicians, White patients, and physician education at top 50 U.S. medical schools are the reference categories. Standard errors are clustered at the physician level. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, + $p < 0.10$.

Table E8. Physician Race and Top Box Responses: Patient Health and Prior Relationship

Sample	Top Box					
	Self-Rated Health Available (1)	Self-Rated Health Available (2)	Self-Rated Health Available (3)	Regular Visit Indicator Available (4)	Regular Visit Indicator Available (5)	Regular Visit Indicator Available (6)
Physician Race						
Asian	-0.037*** (0.011)	-0.028** (0.010)	0.006 (0.020)	-0.043* (0.017)	-0.040* (0.016)	-0.053** (0.019)
Black	-0.052** (0.019)	-0.040* (0.018)	-0.032 (0.035)	-0.059 (0.038)	-0.057 (0.038)	-0.066 (0.040)
Other/Unknown	-0.023 (0.015)	-0.014 (0.013)	0.019 (0.023)	-0.017 (0.023)	-0.016 (0.024)	0.001 (0.029)
Self-Rated Health		0.091*** (0.002)	0.095*** (0.003)			
Regular Interaction					0.082*** (0.006)	0.078*** (0.007)
Self-Rated Health × Physician Race						
Asian			-0.010+ (0.005)			
Black			-0.002 (0.010)			
Other/Unknown			-0.010 (0.006)			
Regular Interaction × Physician Race						
Asian						0.031+ (0.017)
Black						0.020 (0.052)
Other/Unknown						-0.038 (0.029)
Control Variables						
Physician Gender	Y	Y	Y	Y	Y	Y
Physician Years in Practice	Y	Y	Y	Y	Y	Y
Physician Education	Y	Y	Y	Y	Y	Y
Patient Gender	Y	Y	Y	Y	Y	Y
Patient Race	Y	Y	Y	Y	Y	Y
Patient Age	Y	Y	Y	Y	Y	Y
Patient Insurance Type	Y	Y	Y	Y	Y	Y
Constant	0.585*** (0.021)	0.179*** (0.021)	0.166*** (0.022)	0.628*** (0.033)	0.618*** (0.033)	0.621*** (0.033)
Hospital FE	Y	Y	Y	Y	Y	Y
Care-Setting FE	Y	Y	Y	Y	Y	Y
N =	58,950	58,950	58,950	23,136	23,136	23,136
R-squared	0.028	0.060	0.061	0.012	0.021	0.021

Note. All models estimate the likelihood of a *Top Box* response by physician race while controlling for physician characteristics, patient characteristics, hospital fixed effects, and care-setting fixed effects. Models 1-3 are estimated on the subsample of responses in which patients report a self-rated health score. Model 2 additionally controls for self-rated health, and Model 3 further includes its interaction with physician race. Models 4-6 are estimated on the subsample of responses in which patients indicate whether they regularly visit the treating physician. Model 5 additionally controls for the regular-visit indicator, and Model 6 further includes its interaction with physician race. White physicians, White patients, and physician education at top 50 U.S. medical schools are the reference categories. Standard errors are clustered at the physician level. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, + $p < 0.10$.

Table E9. Racial Disparities on Matched Sample

Sample	Top Box	
	Full (1)	Matched (2)
Physician Race		
Asian or Black	-0.025*** (0.007)	-0.031*** (0.007)
Control Variables		
Physician Gender	Y	
Physician Years in Practice	Y	
Physician Education	Y	
Patient Gender	Y	
Patient Race	Y	
Patient Age	Y	
Patient Insurance Type	Y	
Matched Strata		Y
Hospital FE	Y	Y
Care-Setting FE	Y	Y
N =	381,832	378,997
R-squared	0.044	0.053

Note. Both models are linear probability models with robust standard errors clustered at the physician level. Model 1 uses the full sample and controls for physician characteristics, patient characteristics, and hospital and care-setting fixed effects. Model 2 uses the matched sample and includes matched-strata, hospital, and care-setting fixed effects. White physicians are the reference category. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, + $p < 0.10$