Discrimination Against Doctors: A Field Experiment*

Alex Chan

Stanford University

(Click here for the most recent version)

January 25, 2023

Abstract

Discrimination against doctors is important but scantly studied. I report a field experiment which observes that customers discriminate against Black and Asian doctors when they choose healthcare providers, and that this can be substantially reduced by supplying information on physician quality. I evaluate customer preferences in the field with an online platform where cash-paying consumers can shop and book a provider for medical procedures based on a novel experimental paradigm. Actual paying customers evaluate doctor options they know to be hypothetical to be matched with a customized menu of real doctors, preserving incentives. Racial discrimination reduces patient willingness-to-pay for Black and Asian doctors by 12.7% and 8.7% of the average colonoscopy price respectively; customers are willing to travel 100–250 miles to see a white doctor instead of a Black doctor, and somewhere between 50–100 to 100–250 miles to see a white doctor instead of an Asian doctor. Providing signals of doctor quality reduces this willingness-to-pay racial gap by about 90%. Willingness-to-pay penalties on minority doctors are multiples of actual average racial quality differences and even the difference between doctors in highest and lowest quality levels. This field evidence shifts the focus beyond traditional taste-based and statistical discrimination to include behavioral mechanisms like biased beliefs and deniable prejudice. Discrimination against Black doctors are higher for non-college-graduate customers and residents in zipcodes that voted for the 2020 presidential candidate on the political right. Actual booking behavior allows cross-validation of incentive compatibility of the stated preference elicitation.†

Keywords: Discrimination; Statistical Discrimination; Conjoint Analysis

1 Introduction

Existing reports have raised concerns about customer discrimination in medicine, where patients seeking care often select from multiple physicians or other healthcare providers. Anecdotal evidence from interviews with doctors has documented reports of discrimination by patients toward Black and Asian doctors.[1] In addition, a large body of circumstantial evidence is consistent with the presence of discrimination: minority physicians are underrepresented in the U.S. medical workforce (Bergen Jr (2000), Brown et al. (2009), Lett et al. (2019), Merchant and Omary (2010)), and minority physicians earn less than white physicians.[2] Such evidence is consistent with a reduction in the marginal revenue product of minority professionals associated with customer discrimination.[3] The result

*Email: chanalex@stanford.edu. I cannot thank Loren Baker, Michelle Mello, Muriel Niederle, and Al Roth enough for their guidance and encouragement. I also thank Nick Bloom, Dave Chan, Matthew Gentzkow, David Klinowski, John List, Erzo Luttmer, Negar Matoorian, Thayer Morrill, Heather Royer, Joshua Schwartzstein, Andrei Shleifer, Colin Sullivan, Kurt Sweat, and Kenneth Train for helpful comments, and thank all the doctors and patients who participated in detailed interviews. Special thanks go to the partnering platform's CEO and Director of Business Development. This paper also benefited from comments from participants in a number of seminars (University of Chicago, Harvard Business School, Stanford University, Universidad de San Andrés, Purdue University, University of California, Santa Barbara) and conferences (NBER Race and Stratification, Becker Friedman Institute/Stigler Center, Rising Stars in Market Design Conference, International Health Economics Association Congress, Stanford Institute for Theoretical Economics Conference, Annual Conference of the American Society of Health Economists, Essen Health Conference, Economic Science Association North America/Joint initiative for Latin American Experimental Economics/World/Europe Conferences, All-California Labor Economics Conference). I want to also express my gratitude towards the prize committee of the Essen Health Conference for each recognizing this paper with their best paper award. The study was registered on the AEA RCT registry, ID AEARCTR-0006815. This project was supported by the Gerhard Casper Stanford Graduate Fellowship, grant number T32HS026128 from the Agency for Healthcare Research and Quality, and by the Dixon and Carol Doll Graduate Fellowship through a grant to the Stanford Institute for Economic Policy Research. The content is solely the responsibility of the author and does not necessarily represent the official views of the Agency for Healthcare Research and Quality.

†*JEL* Codes: J71, I11, L15, L86, M31

[1]See Filut et al. (2020) for a recent systematic review of the literature on discrimination toward physicians of color. In Filut et al. (2020), the authors noted that there is "almost no attempt to collect data on patient interactions... The lack of curiosity regarding experiences of physicians of color with discrimination from patients may reflect underlying assumptions that these physicians care solely or predominantly patients of color which have their root in U.S. history." White this statement is largely true, it is worth noting that some papers have produced some evidence of patient preference for provider race based on basic correlational evidence and focus group interviews (Gray and Stoddard (1997), Saha et al. (1999), Garcia et al. (2003)).

[2]In the U.S., Black physicians earn 13.5% less than white physicians, and Asian physicians earn 7.8% less (Grisham (2017)). Also, P Ly et al. (2016) found that in the U.S. Black male physicians have an adjusted median annual income that is 26% lower than that of their white counterparts while Black female physicians have an adjusted median annual income that is 7% lower than there white counterparts.

[3]Other factors and doctor attributes can also affect the marginal revenue product of service professionals via customer preferences. Some papers document correlations between consumer preferences and various factors like communication skills (Fung et al. (2005); Hirpa et al. (2020)), distance (Schmitt et al. (2003); Yoon et al. (2019); Mooney et al. (2000)), and other quality measures (Harris (2003); Santos et al. (2017); Kolstad and Chernew (2009); Razzouk et al. (2004); Luft et al. (1990)). Li and Hubner (2019) conduct an experiment using paid online subjects who are not actual shoppers to engage

is reduced access to a diverse set of healthcare providers. Research has suggested the positive impacts of a diverse workforce (Alsan et al. (2019)). Despite the importance of understanding customer discrimination in healthcare, there have been few analyses that provide direct evidence about discrimination or its structure.[4]

In this paper, I report a field experiment in which I study whether paying customers discriminate against Black and Asian physicians when they choose healthcare providers.[5] In addition, the experiment is designed to allow me to obtain information about whether any discrimination observed is associated with taste-based discrimination (Becker et al. (1971)), statistical discrimination (Arrow et al. (1973), Phelps (1972)), or something else altogether. Statistical discrimination arises when customers use easily observable characteristics such as race to infer the expected quality of doctors. Rational or accurate statistical discrimination assumes that customers hold rational expectations of group traits. However, customers can also hold biased priors about group traits, leading to biased belief discrimination[6] (for example, Bordalo et al. (2016), Bohren et al. (2019b), or Esponda et al. (2022)[7]). Statistical and biased belief discrimination can be eliminated when enough information about the quality of a doctor is shared with customers so that they can rely on information other than race to update their beliefs about the doctor. Taste-based discrimination theory suggests that customers choose as if there were a disutility from associating with a particular racial group. Even with perfect information, taste-based discrimination could remain. However, a choice environment with limited information about quality can provide "moral wiggle room" or plausible

in hypothetical and non-incentivized scenarios in order to evaluate the impact of online ratings on their choice of doctors and found that the subjects prefer doctors with higher ratings. In a related question of physicians referring patients to surgeons, Sarsons (2017) found that the gender of the surgeon influences the way signals about the surgeon's quality are interpreted.

[4]One reason for the dearth of studies is that it is hard to observe the full choice set faced by the customer along with all the individual doctor attributes pertinent to customer choice. Another reason is that there might not be enough variation in doctor race within the choice sets even when the data on provider options are available. Moreover, the race of a doctor can be correlated with many other factors that can explain customer choice.

[5]In the experiment, I do not differentiate between the various Asian ethnic groups and refer to East Asians (Chinese, Japanese, Korean, Okinawan, Taiwanese, Tibetan), Southeast Asians (Bruneian, Burmese, Cambodian, Filipino, Hmong, Indonesian, Laotian, Malaysian, Mien, Singaporean, Timorese, Thai, Vietnamese), and South Asians (Bangladeshi, Bhutanese, Indian, Maldivians, Nepali, Pakistani, Sri Lankan) as "Asians."

[6]Based on feedback that the author has received from various conferences, seminars, and reviewers, I will use "biased belief discrimination" to refer to this type of discrimination (sometimes referred to as inaccurate statistical discrimination, for example by Bohren et al. (2019b)) to avoid confusion with traditional statistical discrimination

[7]There is also an empirical literature that found that bias in choice can be driven by racially biased perceptions. For example, Fong and Luttmer (2011).

deniability (Dana et al. (2007), Exley (2016), Exley and Kessler (2021), Bénabou and Tirole (2006)) for someone to express antisocial behavior like taste-based discrimination. I will refer to taste-based discrimination expressed only when there is plausible deniability as deniable prejudice. A traditional critical test to determine the nature of discrimination is to see whether the introduction of an economically relevant signal closes the racial gap.[8]

To test for discrimination, I observe paying customers as they shop for a doctor to perform a colonoscopy via an established online platform. This platform has over two hundred thousand, mostly self-paying, subscribed customers (prospective patients) who can submit a request for a medical procedure and receive anonymized bids from providers across the United States. In this field setting, I show that prospective patients discriminate against Black and Asian doctors, and that discrimination fades when information about doctor quality is provided to the shopper.

My experimental design involves customers making choices from menus of hypothetical options described in terms of independently varied levels of attributes along various dimensions designed to mimic market experience (Ben-Akiva et al. (2019)).[9] Actual customers shopping for a colonoscopy were recruited to the study. Subjects reviewed hypothetical doctor profiles for their procedure with exogenous variation in price, travel distance, gender, and race. Based on their responses to the conjoint survey, subjects were offered 10 actual physician options for booking. Subjects were asked to evaluate menus of hypothetical options with the promise to match the subject to a menu of actual options based on their evaluations of the menus of hypothetical options.[10] I also observed the actual booking choices made by patients among the actual doctors offering colonoscopies, to validate the estimated discrete choice model using actual booking data from the same

[8]Previous research has used this approach to disentangle taste-based and statistical discrimination. For example, "dressing for success" in List and Gneezy (2014), signaling low search costs by declaring "I'm getting a few price quotes" (Gneezy et al. (2012)), having reviews posted on a guest's profile page on Airbnb (Cui et al. (2020)), or including quality ratings for potential mentors in a hypothetical choice setting (Gallen and Wasserman (2022)). In these experiments, the "quality signal" is introduced through an outfit signalling high socioeconomic status, a verbal statement signalling low search costs, and reviews signalling guest "quality" respectively. An alternative, creative approach is Pope and Sydnor (2011) who use evidence from online lending markets to distangle taste-based and statistical discrimination.

[9]Such designs are often referred to as choice-based conjoint analysis in the marketing literature.

[10]Hypothetical bias can arise when the experimental design fails to prompt individuals to reveal their true preferences. If incentives are not present or not strong enough to prompt subjects to reveal their true type, subjects could answer conjoint surveys strategically, provide random answers, or deviate from their true preferences for other reasons. It is worth noting that previous methodological work in economics (List et al. (2006)) has found no evidence of hypothetical bias when estimating marginal attribute values as I do in the present paper (unlike "levels" or total willingness-to-pay).

shopping instance.[11]

I randomize the study population into two groups, one that receives quality information and one that does not. I calculate estimates of willingness-to-pay for attributes based on the results of the conjoint estimates. All attributes including quality are randomized and balanced across different races for the hypothetical doctor profiles.

This experimental setting allows me to conduct two primary tests. In the first, using the patients randomized to receive no quality information, I can assess whether there is evidence for discrimination as measured by a willingness-to-pay gap between minority and white doctors. Second, I can compare the racial gap in willingness-to-pay among patients who did not receive quality information against the racial gap in willingness-to-pay for those who were randomized to receive quality information, to assess whether the provision of quality information changes the result. This is a test for mechanisms beyond traditional taste-based discrimination.

I report two main empirical findings. First, there is customer discrimination against Black and Asian doctors. Second, the majority of the observed discrimination is likely due to mechanisms beyond traditional taste-based or statistical discrimination. Customers are willing to pay a significant premium to have their colonoscopy done by a white doctor rather than a minority doctor. In particular, willingness-to-pay for Black and Asian doctors is lower than the willingness-to-pay for white doctors by 12.7% and 8.7% of the average colonoscopy price of $2122, respectively. These racial gaps in willingness-to-pay suggest that customers are willing to travel 100–250 miles just to see a white doctor for the procedure instead of a Black doctor, and somewhere between 50–100 to 100–250 miles to see a white doctor for the procedure instead of an Asian doctor. However, when quality information is provided for each doctor option, the willingness-to-pay gap for non-white providers drops significantly (to 1.2% and 1.1% of the average price for Black and Asian doctors, respectively). I validate the estimated discrete choice model using actual shopping behavior of the study subjects and find that actual doctor booking coincides with one of three options predicted to be most likely chosen in 82% of the cases.

When I evaluate sub-populations within the sample, results are largely similar to the full sample with the exception of customers segmented by education and by political lean-

[11] Provider outreach representatives at the partnering platform regularly interview providers on the platform. Provider responses indicate that all confirmed bookings via the platform converts into a clinical encounter.

ings during the 2020 presidential election. Discrimination without quality signals against the minority doctors and the magnitude of the drop in discrimination are statistically indistinguishable if we compare customer segments by age, experience with colonoscopy in the past 10 years, or actual booking decision. Willingness-to-pay gap for Black providers are higher in the treatment without provider quality signals for customers who do not have a college degree and customers residing in zipcodes that voted for the Republican presidential candidate in the 2020 election relative to those who graduated from college and those in zipcodes that voted for the Democratic candidate respectively.[12]

I then address several additional issues. First, could the results be due to information burden or distraction? I find that differences between treatment and control of the magnitudes described above seem to be present only among customers who later reported that the doctor quality signal gave them adequate information on doctor quality. This and other evidence rules out the hypothesis that observed differences are merely due to distraction given that the treatment group menus present more doctor attributes.[13]

Second, the data indicate that without quality signals, Black and Asian doctors suffer a significant willingness-to-pay penalty about two to three times larger than the willingness-to-pay penalty for 1-star (lowest quality level) relative to 5-star doctors (highest quality level). Drawing on standard models of discrimination (Aigner and Cain (1977); Bohren et al. (2019b); Fang and Moro (2011)), customers holding biased statistical models of doctor quality in relation to race can potentially account for this empirical observation.[14] However, to rely on biased beliefs alone, additional and somewhat implausible assumptions would need to be made about the magnitude of such biased beliefs or the degree of risk aversion with respect to quality. I also discuss how the exact breakdown between taste-based and statistical/biased belief discrimination might depend heavily on modeling assumptions. Other behavioral models can also explain the empirical results:

[12] See Appendix Tables 14 and 15. Willingness-to-pay gap for Black doctors are higher in the treatment without provider quality signals for customers who reside in zipcodes that voted for the Republican presidential candidate in the 2020 election relative to those who graduated from college and those in zipcodes that voted for the Democratic candidate at the 90% level, while the racial gap for Asian doctors for the same comparison groups is present but not significant at conventional levels.

[13] The lack of a difference in willingness-to-pay for various longer travel distances between the treatment and control groups offers the most cogent evidence for the absence of a distraction effect. I discuss this in Section 3.3.2 in more detail.

[14] There are multiple candidate mechanisms that could generate the biased belief discrimination implied by the data. One example, mentioned above, is stereotyping (Bordalo et al. (2016)) by a risk averse customer. Such customer can overweight representative doctor quality categories for each racial group and hold inaccurate priors of the order of magnitude implied by the data. It is beyond the scope of this experiment to distinguish which mechanism is at work in the present setting.

for instance, there might be more “moral wiggle room” to express animus or deniable prejudice towards minority doctors when there is no information about doctor quality. My preferred interpretation is that a combination of mechanisms that includes biased beliefs and deniable prejudice gave rise to the observed empirical pattern. Moreover, the key contribution of the present empirical findings is their cogent challenge to the conclusiveness of the traditional test for statistical discrimination and the approach of many field studies that considers only the simple dichotomy of traditional taste-based and statistical discrimination.

To my knowledge, this study is the first to get inside the black box of customer discrimination toward healthcare providers and to shed light on the underlying mechanisms. Customer discrimination is potentially a factor for the persistence of major issues in the medical labor markets such as discriminatory employment practices, the leaky pipeline for minority medical students, and the wage gap between white and minority providers (Freeman et al. (2016); Barr et al. (2008); Courey (2020); McGregory Jr (2013)). The finding that signals of doctor quality for customers drastically lower discrimination rules out the hypothesis that the observed customer discrimination is mainly due to taste-based discrimination. Because the reduction in discrimination is larger than what is justifiable by true quality differences, returns to providing quality information when people are miscalibrated in their beliefs are higher in the present context than providing this information if people are calibrated. More importantly, the finding also suggests provider quality information may be a potential solution to reduce labor market discrimination for minority doctors on average.

The present analysis also provides new information about the consumer behavior and preferences of the 30 million uninsured Americans (Gunja and Collins (2019)). This group consumes healthcare in widely disparate circumstances, and little has been done to understand the preferences that guide their choices. As far as I know, this is the first study that estimates demand for healthcare among uninsured Americans with respect to prices, travel distance, and provider attributes. Importantly, uninsured, cash-paying customers offer a pure form of shopping in healthcare without distortions from other payment sources.

1.1 Related Literature

This paper builds on previous work on labor market discrimination, quality signals in market design, and the design of discrete choice experiments.

Customer discrimination in labor market settings outside of healthcare has been documented in a growing literature.[15] These papers look at observed labor market outcomes such as wages to infer customer discrimination. The wage discount estimates associated with non-white workers in these studies are similar to my estimates.[16] Unlike these previous studies, this paper sheds light on customer discrimination in the healthcare labor market, an important sector that represents over 13% of the U.S. labor force or 19 million jobs accounting for $1.0 trillion in annual payroll in 2018 (Census-Bureau (2018)). Also, my randomized field experiment setting allows me to add to the literature that rely mostly on pre-existing observational data.

A body of previous experimental work on discrimination finds evidence that significant proportions of the discrimination observed in various product markets and non-market interactions[17] are statistical. The range of estimated market price differentials associated with minority sellers when there is statistical discrimination roughly brackets the estimates of willingness-to-pay differentials in the present paper.[18] Due to the heavily regulated environment and complex supply chain, the industrial organization of healthcare

[15] For example, see Kahn and Sherer (1988), Nardinelli and Simon (1990), Neumark et al. (1996), Holzer and Ihlanfeldt (1998), Leonard et al. (2010), Combes et al. (2016), Bar and Zussman (2017), Stone and Warren (1999), and Burdekin and Idson (1991).

[16] Customer preferences as a factor driving the racial wage gap is a small but growing literature. In an interesting study design, Nardinelli and Simon (1990) found that baseball cards of non-white players sell for about 10% less than the cards of white players of comparable ability. Some studies looked at the equilibrium impact of customer discrimination on prices and labor market outcomes: Kahn and Sherer (1988) find that customer demand contributed to Black baseball players getting paid 20% less, Holzer and Ihlanfeldt (1998) find that Black service employees are paid roughly 17%–26% less at establishments with only Black customers than at those with only white customers (this could be evidence of customer discrimination if white customers are more likely to discriminate or discriminate more against Black service professional), and Bar and Zussman (2017) found that the price quotes obtained by Arab service contractors are about 17% lower that those obtained by their Jewish counterparts in Israel.

[17] For example, Doleac and Stein (2013), Ayres et al. (2015), Edelman et al. (2017), Bartik and Nelson (2019), Cui et al. (2020), and Laouénan and Rathelot (2020) also leveraged online market settings, varied perceived race via skin color on an image or name of seller, and estimated race differentials in the price obtained by the seller and the probability of the seller being chosen by the buyer. For non-market interactions, see Bertrand and Duflo (2017), Gneezy et al. (2012), or List and Gneezy (2014) for well documented field experimental results.

[18] Doleac and Stein (2013) found that iPods held by a hand with darker skin generated offer prices that were 11%–12% lower on Craig's List, while Ayres et al. (2015) found that baseball cards held by a hand with darker skin sold for 20% less on eBay. Edelman et al. (2017) and Laouénan and Rathelot (2020) both looked at online marketplaces for sellers of services (like this paper) and found discrimination against minority sellers to the tune of 3% lower charges, but the latter went a step further and showed that the entire price discount vanishes when quality signals are exogenously introduced.

is not straightforward. As a result, practitioners have found economic research that rely on untested assumptions about market structure suspect. By estimating willingness-to-pay directly rather than inferring it from market prices, I am able to do so unburdened with additional assumptions about market structure. This approach allows me to get directly at the demand function and preferences of customers. Also, I go beyond this previous work and consider both biased beliefs and deniable prejudice.[19] My data reject statistical discrimination as well as the simple dichotomy of taste-versus-statistical discrimination to build a case for biased beliefs and deniable prejudice as potential key behavioral mechanisms for discrimination in the field.

A literature on quality signals and market design suggests that provision or restriction of information can affect the level of discrimination in various labor markets. For instance, there is empirical evidence that racial minorities were harmed when certain quality signals for hiring purposes were disallowed. "Ban the Box" policies that bar employers from requesting criminal records of prospective employees, for example, seem to have widened the Black-white gap in callbacks from prospective employers (Agan and Starr (2018)).[20] My results suggest that part of the differential treatment is driven by individual customers and not just corporate policies or human resource agents who might have a lower stake in the decision.

The present experimental design makes a methodological contribution to the literature on stated preference elicitation, which aims to overcome hypothetical bias not only by offering a menu of real choices that reflects hypothetical choices, but also by examining the real choices that result. In particular, it builds on the innovations in Low (2014), Kessler et al. (2019), and others.[21] The most common way researchers incentivize a conjoint choice task is by promising to deliver the product choice from one of the menus at the end of the experiment and deducting the corresponding price of the product from the payment the subject received at the beginning of the experiment (e.g. List et al. (2006)). Kessler et al. (2019) adopted a related approach and incentivized their choice

[19] An important paper by Bohren et al. (2019a) has considered and found evidence of biased belief discrimination against female authors of posts on an online QA forum.

[20] Other examples include employer credit check bans (Ballance et al. (2020); Bartik and Nelson (2019)) and pre-employment drug testing (Wozniak (2015)). What is less studied, though, is when the hiring decisions are made by individual customers, as is the case for many medical providers, lawyers, and other professional service providers.

[21] For example, earlier work by Elías et al. (2019), List et al. (2006), List and Lucking-Reiley (2002),Low (2014). There are also contemporaneous experiments that uses hypothetical preference elicitation (e.g. Sullivan (2021), Macchi (2022)).

experiment by promising to match employers to a menu of actual resumes based on their evaluations of the hypothetical resumes. While some of these pioneering works have evaluated hypothetical bias of pure hypothetical choice experiments against incentivized experiments, they have not attempted to directly assess the incentive compatibility of the incentivized choice experiments themselves against actual consumer choice within the same choice experiment. Going beyond previous approaches, my design incorporates analysis of the actual booking choice of buyers paying entirely out-of-pocket in the menu of actual doctors from a real shopping environment so that one can validate whether behavior in the incentivized hypothetical setting maps well into real shopping behavior.

The paper is organized as follows. Section 2 describes the unique context that enables the experimental design and details the empirical strategy. Sections 3 and 4 describe the data and present the key results. Section 5 outlines a simple model that distinguishes between the various sources of discrimination and helps interpret the results. Section 6 discusses the implications of this study for the literature on healthcare consumerism, discrimination, and stated preference elicitation. Section 7 concludes.

2 Experimental Design

This section presents implementation of the preference elicitation exercise combining the incentives and ecological validity of a real shopping environment that is familiar to its users with the control of a laboratory. Section 2.1 describes the context of a popular online platform for healthcare shoppers and how active customers were recruited as subjects for this experiment. In section 2.2, I outline incentives for reporting preferences in the conjoint survey and cross validate the incentive compatibility of survey design directly within the same shopping instance using booking data. Section 2.3 describes the doctor attributes and clarifies the nature of the quality signal that defines the difference between the treatment and control conditions. The hypothetical doctor profiles and the choice sets are described in Section 2.4. Section 2.5 outlines the estimation approach.

2.1 Online Shoppers for Medical Procedures and the Market Setting

I conduct preference elicitation in partnership with an online platform in the U.S. that links self-pay patients seeking non-emergency care with doctors and facilities, much the way Priceline or Hotwire connects travelers and hotels.[22] This platform charges its some 265,000 paid subscribing customers a nonrefundable $25 fee to put a procedure out for bid, this fee is in addition to the payment for the actual procedure paid directly to the providers. Doctors pay a fee of $50[23] to bid on one request for bids that come as a package deal that include the physician's fee, the facility fee, and the anesthesiologist's fee if applicable. The procedures most shopped for on this platform are colonoscopy and imaging (including MRI, CT scans etc.). Roughly six thousand providers bid regularly. Actual doctor identity and contact information are concealed from the customer as the customer evaluates bids that include price and zip code of the provider.

After clicking "Accept this Bid" for a specific doctor, the customer promptly receives the contact information for the doctor who submitted the bid, while the doctor receives the contact of the customer. The doctors have contractually committed to the upfront price equal to their accepted bid, and the customer will pay the doctor directly after booking with the doctor. The customer and doctor can arrange for the procedure to take place at a date acceptable to both, typically within six months. This platform facilitated more than 7,500 transactions resulting in over $70 million in charges in 2018. As far as I know, this market has not been explored by the economics and health policy literature that mainly relies on administrative data covering insured but not self-pay patients.

The experiment builds on the typical shopping experience of the customers of this platform where they evaluate anonymized doctor options, to naturally introduce choice sets where features characterizing each doctor option are restricted. The preference elicitation exercise is offered as a pilot tool that is similar in many ways to the usual shopping experience but different in that there are hypothetical menus of doctors to evaluate before being offered an actual menu of doctors. Furthermore, the customers gain access to

[22]Due to terms in the Data Use Agreement, the company's identity will remain private. The business model and all figures on prices, customer base, participating doctors, and transaction volumes are provided by the company and cross-checked by the author through news coverage published in Forbes, The Wall Street Journal, Washington Post, Fox News, NBC, CNN, ABC and CBS.

[23]The doctors can also opt to pay an annual $250 fee for multiple bids.

a larger set of the doctors than usually listed on the platform through the pilot as the doctors who can potentially be matched include those usually listed on the platform as well as those listed with two of the platform's competitors.

The platform sent invitation emails to its subscription base in two waves. In both waves, prospective patients were invited to use a pilot tool that gathers prices from hundreds of platform doctors as well as doctors from the platform's two main competitors to "Shop for Your Next Provider for Colonoscopy, MRI or Knee Replacement" and that "the use of this pilot tool will be a one-time-only offer for each customer....[and] we'll waive the $25 fee for the appointment you book through the pilot tool" (see Appendix Figure 12 for the censored version of this email). All subjects have used the platform to shop for doctors previously. The first wave was early January 2021 and the second at the end of February 2021. A recruitment tweet with similar content was also sent out weekly between 1 December 2020 and 31 March 2021.[24]

2.2 A Validated, Incentivized Conjoint Design

After being informed of the instructions (see Appendix Section A.3) and collecting basic information about the outside options and whether they had previously had a colonoscopy, each customer was presented with a series of menus of doctors including a "None of the above" option (selection of the "None of the above" option is modelled as selection of the outside option; see Section 2.5). Customers choose the single most preferred option in each menu instead of rating each option on desirability, even though the latter would yield more information. Daniel McFadden and others have called this choice-based conjoint elicitation (McFadden (2017), Ben-Akiva et al. (2019)), and a key benefit is its resemblance to the actual shopping experience (which is the main reason the partnering platform strongly preferred this over having customers rate each option with some cardinal or ordinal score). Each doctor offered in each menu would be described in terms of price and levels of attributes. Customers were asked to choose their most preferred option in each menu. As illustrated in Figure 1, customers were offered menus of doctors who

[24] Customers can only participate once in the pilot (experiment), and each customer participates independently without the ability to observe or influenced by what other customers are doing on the pilot. I also used location and background information from the subscription database (where possible) to validate that no co-located customers have participated together. In the analysis, I assume that the Stable Unit Treatment Value Assumption (SUTVA) is met. In other words, I assume no-interference between customers, and that I can model a customer's treatment status and not be concerned with the treatment status of every customer in the sample.

provide colonoscopies, with each doctor anomymized to resemble the platform's typical shopping experience but described in terms of price, travel distance from the customer's zip code, and a profile of the doctor including a blurred picture and gender.[25] The information provided for each doctor option is presented in a layout that would be familiar to the usual customer.

Subjects are asked to evaluate menus of hypothetical options with the understanding that the options are hypothetical and that more accurate evaluations will maximize the value of their participation. In my experiment, each customer evaluates fourteen menus with five doctor options and a "None of the above" option. Their participation incentive is that they will be matched to a specific menu of ten real booking options out of a pool of doctors aggregated in the partnering platform site as well as its two main competitors (more than doubling the usual number of doctors available to the customer). Eight of the options are doctors with the highest predicted choice probability based on particular individual preferences parameters estimated from the choice-based conjoint analysis data with mixed logit using Markov Chain Monte Carlo Hierarchical Bayes (MCMC HB) estimation (Ben-Akiva et al. (2019)). The remaining two options are randomly drawn from the remaining 331 possible doctors who offer colonoscopy services (Figure 2). By analyzing the actual booking choice in this menu of ten doctors, I can validate whether behavior in the incentivized conjoint setting maps well into real behavior.

A shortcoming of conjoint methods is that they are less reliable when the items are unfamiliar or incompletely described (McFadden (2017)). A benefit of my experimental setting is customer familiarity with the shopping experience during the experiment. Doctor information provided as customer evaluate options in the "pilot" is similar to the doctor information typically provided when customers shop on this platform. This makes the present conjoint analysis more reliable.

2.3 Doctor Attributes

Each doctor option is characterized by five attributes. The doctor options vary in race, gender, price, travel distance from the customer zip code, and quality. Race and gender are indicated by a blurry profile picture,[26] discussed in detail in Subsection 2.3.1 below.

[25]This is the baseline configuration. In the treatment group, quality "stars" are also provided. See below.

[26]Gender is also explicitly stated in the doctor profiles.

Each attribute of the doctors is selected from a realistic range based on real doctors listed on the platform and its competitor sites (see Appendix Table 4 for details on the attribute levels).

In the hypothetical doctor options, prices can take on one of ten evenly-spaced levels between \$1704.00 and \$2541.60, or the 10^{th} and 90^{th} percentiles of the actual prices for a colonoscopy from the full set of doctors on the platform and the two competitor sites. The customers are told (truthfully) that "[t]his price includes a screening or diagnostic colonoscopy with or without specimens/polyps removal by biopsy or brushing. Fees for facility and physician are include in the price. TRAVEL costs are NOT included in this price" (see Figure 1). Distance from the customer zip code can take on any of five levels: "0-10 miles", "10-50 miles", "50-100 miles", "100-250 miles," or "More than 250 miles." Finally, gender can be male or female while the profile picture has a blurred photo suggestive of either white, Black, or Asian races.

2.3.1 Indicating Doctor Race

Race is the central characteristic of interest in the present paper. I manipulate perceptions of the doctor race by using profile pictures that clearly indicate the race of the hypothetical doctor. The resolution of each profile picture is reduced so that the doctor's race shows through[27] but their attractiveness and other features are obscured (similar to Fong and Luttmer (2009)).[28]

To further ensure that detected outcome differences are due to race, I indirectly verify whether the profile pictures are similar in terms of attractiveness and other features. Besides blurring the photos, I assess the underlying unblurred photos' attractiveness. I use a popular facial symmetry detection algorithm deployed in previous publications evaluating attractiveness and economic outcomes (Dietl et al. (2020)).[29] I found no

[27]To verify that perceived race shows through, I conducted an incentivized survey on Prolific ($N = 50$) where respondents were all uninsured Americans who were asked a multiple choice question "What is the race of this doctor?" for each profile picture. On average, survey respondents match the correct race to blurred profile pictures with 93.25% accuracy.

[28]Recall that doctors are anonymized in their bids in the default shopping experience on this platform, and so the customers are reminded that "[w]e do not share detailed provider information until you have booked your appointment, therefore: We've blurred the details of each provider... Instead of the actual photo of each provider, our staff matched every one of the hundreds of actual providers to one of 24 blurred provider pictures that looks most like them" before they completed the conjoint survey and before they made the final actual booking (see Appendix Section A.4).

[29]I also rated all the profile pictures for "attractiveness" using a deep neural network based on a pre-trained ResNet-50 architecture (He et al. (2016)) that is fine-tuned with a dataset of hand-labeled

significant difference in attractiveness estimated by these measures across races (white, Black, and Asian).[30]

I also conduct an online survey of uninsured Americans via Prolific,[31] who are recruited separately from the platform's customers, where I ask the survey respondents to rate the blurred doctor profile pictures and provide subjective evaluations on perceived age and attractiveness. There is neither a significant difference in perceived age[32] (see Appendix Table 7) nor attractiveness (see Appendix Table 6) by racial group.

2.3.2 Treatment versus Control: With versus Without Quality Signal

The last doctor attribute of quality plays a key role for the test of the behavioral mechanisms giving rise to discrimination. Customers are randomized into the treatment or the control group. The groups differ in attributes they see in doctor profiles. While the control group sees doctor price, distance, gender, and profile picture as in Figure 1, the treatment group also sees stars that indicate the quality of the doctor in each profile (see Figure 3 for a side-by-side comparison). These stars range from 1 to 5 with 5 stars being the highest quality.

Customers in the treatment group see an additional informational screen explaining the stars. This screen explains that doctors with higher ratings tend to have better outcomes (fewer complications and lower patient mortality) and shows a table of ratings and other indicators of quality such as years of experience, graduation from a top 20 medical school, board certification, and customer reviews from three popular sites (Healthgrades, WebMD and Vitals). In the pool of real doctors 38% with 5 stars graduated from a top 20 medical school, and average 30 years of experience and a 4.7/5.0 patient review score from popular sites. In contrast, none of the doctors with 1 star graduated from a top 20 medical school, and they average less than 20 years of experience and a 2.2/5.0 patient review score. After these customers made their choices for all menus, I surveyed the subjects' perception of quality score by asking them "[t]o what extent do you agree or

photos. The results are similar to the validation using facial symmetry. Facial symmetry is used for validating similarity in attractiveness across races as it is a better proxy for racial-neutral attractiveness.

[30] Regression results from this validation technique is shown in Appendix Table 6

[31] Pre-registered on the AEA RCT registry, AEARCTR-0010178. This is a separate survey from the Prolific survey on perceived race.

[32] As another demonstration, I compare the main willingness-to-pay estimates from the main estimation model outlined in Section 3.3 with and without adding perceived age separately as a doctor. The estimates are statistically indistinguishable. See Appendix Table 8 for more details. The main specification do not include this perceived age variable generated separately through an unincentivized survey.

disagree with the following statement: "A provider with a higher Comprehensive Quality Score is a better provider than a provider with a lower Comprehensive Quality Score"." The subjects responded to this question using a Likert scale (1 to 5).

2.4 Choice Sets Creation and Variation

The present implementation of preference elicitation asked each customer to evaluate 14 different menus of doctor options varying multiple doctor attributes to allow for good statistical identification of the valuation of separate attributes.[33] Designs for preference elicitation that allow considerable linearly independent variation in the levels of different attributes and a considerable span of attribute levels are well documented in various textbooks (Rao et al. (2014), Raghavarao et al. (2010)). I adopted an off-the-shelf fractional factorial choice design to optimize balance, overlap, and other characteristics. See Data Appendix for the order of questions for each respondent, the order of alternatives in each choice set for each respondent, and the placement of respondents into blocks in this paper. While price is always the last attribute, the order of the other attributes presented for doctor profiles are randomized across customers.

A big advantage of the set of choice sets generated this way is its ability to help address concerns with censoring issues (e.g. if choice made on closest distance or lowest price) that might accompany real-life choice sets. Due to the fractional factorial design, choice data is generated from both menus where there is variation in each attribute as well as menus where some attribute is held constant across options within those menus (e.g. a menu where all five options are 50–100 miles away). The inclusion of the latter type of menus, as shown in the Data Appendix, can help us eliminate concerns with censoring.

I check whether balance is achieved across the exogenously varied doctor attribute levels between the treatment and control groups. The p-values for the test of equality of the levels for each one of the hypothetical doctor attributes (race, sex, price, distance) between the treatment and control groups are 0.9 or higher, consistent with a properly executed conjoint design (Rao et al. (2014)).[34]

[33]Previous choice-based conjoint surveys used by rigorous studies in economics have presented subjects with 15 or fewer menus. For instance, the example in Ben-Akiva et al. (2019) gathered data with 11 menus per subject.

[34]See Appendix Table 9 for the actual table of p-values.

2.5 Discrete Choice Model

The goal of the empirical analysis is to estimate preferences using choice data over hypothetical options. A generalized multinomial logit model with mixed (random) parameters and heteroskedasticity is estimated to allow the coefficients to vary across customers and exploit the information from the same individual making a particular sequence of choices out of the fourteen menus.

In the present discrete choice model, the customer maximizes a linear indirect utility function. Customer i's utility from choosing doctor d is

$$U_{id} = \delta_{id} + \eta_i(E_i - p_d) + e_{id} \tag{1}$$

where δ_{id} is marginal utility i gets from doctor d, η_i is marginal utility i gets from money (the price coefficient), E_i is the upper-limit of expenditure for the procedure in the allotted budget, and p_d is the price of doctor d. The term e_{id} is an unobserved random component of utility and is assumed to be distributed according to a type-1 extreme value distribution.

All customers solve a straightforward utility maximization problem: customer i chooses doctor d if and only if,

$$U_{id} \geq U_{id'}, \forall d' \in D \cup \{0\} \tag{2}$$

where D is the set of doctors and $d' = 0$ indicates the outside option.

Assuming distributional support for the error term on the real line then allows this model to rationalize any pattern of customer choices in the conjoint and obtain analytical expressions for individual choice probabilities of customers. Integrating over regions of the error space that coincide with a choice having the highest utility, I obtain choice probability of customer i choosing doctor d with the familiar multinomial logit form:

$$P_i(d) = \frac{\exp[\delta_{id} - \eta_i p_d]}{\sum_{j \in D \cup \{0\}} \exp[\delta_{ij} - \eta_i p_j]} \tag{3}$$

I further assume that marginal utility δ_{id} is a linear function of doctor attributes:

$$\delta_{id} = \mathbf{a}_d^{'} \lambda_i \tag{4}$$

where $\mathbf{a}_d$ denotes the vector of attributes of doctor d including race, gender, distance, and quality (for treatment group) and λ_i is the vector of customer specific attribute coefficients λ_{ia}.

The main empirical results and hypothesis testing are based on maximum simulated likelihood estimators (Ben-Akiva et al. (2019)).[35] More estimation details are presented in the Appendix (Section A.6).

2.5.1 Estimating Willingness-to-pay and Evaluating the Impact of Treatment

Customer utility in willingness-to-pay or money metric form is estimated to obtain more interpretable and comparable estimates for various doctor attributes (McFadden (2017)).

I estimate willingness-to-pay:

$$WTP_a = \mathbb{E}_i[\frac{\lambda_{ia}}{\eta_i}] \tag{5}$$

where WTP_a denotes the willingness-to-pay for attribute a, λ_{ia} is the coefficient from the marginal utility function estimation in (4), and η_i is the price coefficient. The main results are estimates of this metric.

To evaluate the differences of the effect on customer choice of the same attribute (e.g. Black doctor) between the treatment and control group, slightly different models for the treatment and control groups are estimated with the former having additional covariates. The effects are translated into willingness-to-pay from the logit coefficients as shown above in Equation 5. The estimates for the impact of an attribute between the treatment and control group are evaluated with a test statistic:

$$t = \frac{WTP_a^{Control} - WTP_a^{Treatment}}{\sqrt{(SE[WTP_a^{Control}])^2 + (SE[WTP_a^{Treatment}])^2}} \tag{6}$$

where $SE[WTP_a^T]$ is the standard error of WTP_a estimated from the model for treatment

[35] To allow for segmentation of customers by their preferences and allow more parameters (attributes and levels) to be estimated with smaller amounts of data collected from each customer, I also proceed with hierarchical Bayes estimation as is commonly practiced in choice-based conjoint data analysis in the marketing literature (Ben-Akiva et al. (2019), Allenby and Rossi (2006)). Recall that hierarchical Bayes estimation was also used to estimate the individual level parameters to generate match scores to identify the 10 real doctor options for customers on an on-going basis as the experiment proceeded.

condition T.[36]

3 Data and Results

3.1 Sample Population and Check for Balance

The 229 customers shopping for a colonoscopy used the pilot tool between 1 December 2020 and 31 March 2021, with 224 completing the conjoint survey.[37] The randomization approach is a simple Bernoulli trial where the unit-level probability of treatment is 0.5.[38] Of those 224 in the sample,[39] 104 were randomly assigned to the treatment condition and responded to menus with options that include quality "stars" while the other 120 customers were assigned the control condition.

Table 1 presents summary statistics, with the full sample in columns 1 and 2. On average, customers are mostly in the age range between 45 and 64 and much more likely to be uninsured than the overall U.S. population.[40] The six states most represented in the sample are Texas, Illinois, Washington, North Carolina, and Florida-Indiana in a tie. Everyone in the sample population stated that they will be paying cash out of pocket for the colonoscopy, and most have had a colonoscopy in the past 10 years.[41] On average, these customers also looked at more than four colonoscopy options other than the platform, and have an outside option with a price that is within the range of actual colonoscopy prices offered by doctors included in this experiment.

Health economics and health policy research have focused mostly on the insured population given the availability of administrative data, leaving about 30 million Americans

[36]The standard errors of the WTP estimates are calculated with the Delta method (assuming that WTP is normally distributed and its variance is obtained by taking a first-order-Taylor-expansion around the mean value of the variables [namely, the attribute and price coefficients] involved in the ratio expressed in Equation 5 and calculating the variance for this expression).

[37]This is about three times the number of decision-making subjects in main experiment in Kessler et al. (2019).

[38]Think of this as the case where as customers arrive at the experiment, a coin is flipped and a "heads" outcome sends the customer to the control group and a "tails" outcome sends the customer to the treatment group.

[39]This paper uses data from colonoscopy shoppers. The broader experiment included 407 customers who were shopping to get a colonoscopy, an MRI or a knee replacement. 177 shopped for an MRI and 1 shopped for a knee replacement. The MRI data is used for a different paper and also not suitable for the purposes of this paper since the MRI providers are facilities not individuals with clearly identified race and demographics. I was unable to include the shopper for knee replacement since there is only one.

[40]The U.S. uninsured population is about 10% as opposed to 90% in this sample.

[41]For the reader's reference, Fraiman et al. (2022) reported various studies that documented that just under a quarter of patients getting a colonoscopy have gotten one less than 10 years ago.

understudied. By capturing a largely uninsured population, this paper makes a contribution to the literature.[42] Furthermore, customer decision making could potentially be more fully expressed for uninsured cash-paying patients without the constraints and influences of co-pay, co-insurance, and a provider network determined by an insurance company.

The last column of Table 1 verifies that the control and treatment groups are balanced on all measured dimensions, suggesting successful randomization.

3.2 Customer Preferences in the Absence of Quality Signals

Table 2 presents the main results. Using the approach outlined in Section 2.5.1 to estimate the control group, willingness-to-pay for Black doctors is lower by 12.7% of the average colonoscopy price compared to white doctors, and lower by 8.7% of the average colonoscopy price for Asian than white doctors. In other words, in the absence of any information about doctor quality, customers were willing to pay \$270 more on average to have their colonoscopy performed by a white doctor than a Black one, and \$186 more for a white doctor than for an Asian one. These results are presented in the first column. Both of these racial differentials are statistically significant (p-values < 0.001).[43]

Without quality signals, a female doctor suffers a statistically significant \$16 willingness-to-pay penalty for gender. This is significant at conventional levels.

Travel distance also affects customer choice. The willingness-to-pay for a doctor 10-50 miles away is \$37 less than for a doctor who is 0-10 miles away, while having to travel 50-100 miles lowers willingness-to-pay by \$118. Likewise, if the doctor is 100-250 miles or more than 250 miles away from the customer, willingness-to-pay drops by \$263 and \$589 relative to a doctor who is 0-10 miles away. All distances have willingness-to-pay

[42]The present paper do not claim to have randomly sampled the uninsured population. See Appendix Table 10 for a comparison. The rare look at the behavior of the uninsured population remains an important contribution to the health economics literature.

[43]I am not aware of good data that provides the prior on whether the association between race and willingness-to-pay in absence of quality signals is true. I offer an indirect reference point for this prior by obtaining expert forecasts of this particular empirical result (discrimination against Black and Asian doctors) using the Social Science Prediction Platform (DellaVigna et al. (2019)). Based on the responses of 14 economists, 100% predicted that there will be a discrimination effect strictly larger than zero for Black race for doctors, with an unweighted average prediction of a willingness-to-pay penalty equal to 13.2% ($\approx$ 12.7%) of the average colonoscopy price; 50% predicted that there will be a discrimination effect strictly larger than zero for Asian race for doctors, with an unweighted average prediction of a willingness-to-pay premium equal to 0.6% ($> -8.7\%$) of the average colonoscopy price. This suggests that economists who made predictions via the Social Science Prediction Platform expect to observe discrimination against Black doctors but not Asian doctors in the baseline.

differences relative to 0-10 miles of distance with statistical significance of p-value less than 0.001.[44]

To put the size of the discrimination into perspective, the willingness-to-pay penalty for Black doctors is approximately equal to the willingness-to-pay penalty for having to travel 100-250 miles,[45] and the willingness-to-pay penalty for Asian doctors somewhere between 50-100 to 100-250 miles.[46]

3.3 What Difference Does a Quality Signal Make?

Figure 4 presents the willingness-to-pay for doctor quality. The willingness-to-pay for a 5 star doctor is \$72 more than a 1 star doctor, while the willingness-to-pay for a 3 star doctor is \$16 more than a 1 star doctor. The hypothesis that the willingness-to-pay for a 5 and a 4 star doctor are the same is not rejected. More stars generally increase willingness-to-pay. The relationship between stars and willingness-to-pay seems to be nonlinear (see Figure 4).[47]

Column (II) of Table 2 presents the willingness-to-pay estimates for each of the attributes for the treatment group. The coefficients for various travel distances are not statistically different in magnitudes with those recovered from the control group who did not observe the quality signal.[48]

Quality signals reduce the willingness-to-pay gap by 90% for Black doctors and 87% for Asian doctors. With signals of doctor quality, the willingness-to-pay for Black and Asian doctors are \$26 and \$24 lower than white doctors. These willingness-to-pay gaps correspond to 1.2% and 1.1% of the average colonoscopy price respectively. The test statistic from (6) applied to the willingness-to-pay for Black and Asian races yields p-

[44] Some of these coefficients might be biased upwards due to the coincidence of the experiment with the COVID-19 pandemic. This might have especially enlarged the coefficient on the "More than 250 miles" as air travel was less preferred than during normal times. It can also be argued that cheaper flights and better hotel availability during the COVID-19 pandemic could have lowered travel expenses. External validity of the results on travel distance require further evidence.

[45] This observation is confirmed by an F-test of equality between the coefficients for Black race and 100-250 miles distance.

[46] A related literature explores consumer willingness to travel to avoid fellow consumers of a different race for recreation consumption (Backstrom and Woodward (2021)).

[47] Finally, the average Likert scale rating from 5 ("Strongly Agree") to 1 ("Strongly Disagree") is 4 ("Agree") for the self-reported exit survey question "To what extent do you agree or disagree with the following statement: 'A provider with a higher Comprehensive Quality Score is a better provider than a provider with a lower Comprehensive Quality Score'," suggesting the average customer in the treatment group agrees that stars indicate better providers.

[48] The difference in the coefficients for 10–50 miles is not significant at the 5% level but significant at the 10% level.

values < 0.001,[49] suggesting that racial discrimination is significantly attenuated when a quality signal is introduced relative to when there are no direct quality signals.

This analysis suggests that customers who lack individualized information about quality might be relying on race-based assumptions about doctor quality. Racial discrimination decreased substantially for both Black and Asian doctors when the quality signal was introduced. A surprising result is that the reduction in willingness-to-pay gaps for both Black and Asian is larger than the willingness-to-pay gap between a 5 star doctor and a 1 star doctor. The test statistic from (6) rejects the hypotheses that the penalty in willingness-to-pay for being a Black or Asian doctor is the same as the penalty in willingness-to-pay for a doctor dropping from 5 stars to 1 star.

The willingness-to-pay premium associated with a high quality signal (4 or 5 stars relative to 1-3 stars) seems to be similar for white doctors and non-white doctors (overall): the difference is not statistically different at conventional levels (see Table 2 Column 8). Looking at the non-white races separately, I found that the willingness-to-pay premium associated with a high quality signal is 40.4% or \$20.6 higher for Black doctors than for white doctors. The willingness-to-pay premium associated with a high quality signal is statistically indistinguishable for Asian and white doctors. Customers perceive Black doctors who have the highest quality scores to be of higher quality than doctors of other races with similarly high scores. However, even if we factor this \$20.6 additional premium, the drop in the the Black-white gap in willingness-to-pay when the quality signal is introduced is still about 80% (or a drop larger than \$210).

Finally, Black doctors with 1 star (lowest quality score) face a lower willingness-to-pay penalty than Black doctors without a quality signal at all. This result makes it very hard to rely on the simplest, traditional versions of statistical discrimination to interpret the results.

[49] I am not aware of good data that provides the prior on whether the drop in willingness-to-pay with the introduction of quality signals is true. I offer an indirect reference point for this prior by obtaining expert forecasts of this particular empirical result using the Social Science Prediction Platform (DellaVigna et al. (2019)). Based on the responses of 14 economists, 78.6% predicted a non-zero reduction in the willingness-to-pay penalty for Black doctors when the quality signal is introduced, with the average predicted reduction in willingness-to-pay penalty equal to 49.4% of the baseline penalty under the information treatment (less than the 90% observed in the present study); 42.9% predicted a higher in willingness-to-pay for Asian doctors when the quality signal is introduced but the mean predicted change under the information treatment is 0 (different from the 87% drop observed in the present study).

3.3.1 Treatment Heterogeneity across Customer Groups

The present findings depend on characteristics of the customer. If discrimination is only driven by traditional taste-based discrimination for out-group doctors, the quality signals should not have lowered the willingness-to-pay penalty for minority doctors. Moreover, it is interesting to investigate whether discrimination differs by racial concordance between customer and doctor. Figure 6 presents evidence of such heterogeneity.[50] Both white and Black customers discriminate against Black doctors. Black doctors incur a \$78 willingness-to-pay penalty from Black patients when there is no quality signal, such a figure is much smaller than the same figure for white customers against Black doctors, which is more than four times larger. When quality signals are introduced, the point estimate for a Black patient's willingness-to-pay for a Black doctor becomes positive. While this can be evidence for a preference for racial concordance that might include benefits like improved trust (Alsan et al. (2019), Alsan and Wanamaker (2018), Alsan and Eichmeyer (2021), Idan et al. (2020)), the small sub-sample of Black customers limits precision of these estimates making the evidence of preference for racial concordance only suggestive.

Customers also show a preference for gender concordance. Figure 7 shows the willingness-to-pay for a female doctor relative to a male doctor, broken down by whether the customers themselves are female or male. Female customers are willing to pay \$85 (no signal) and \$39 (with signal) more for a female doctor while male customers are willing to pay \$82 (no signal) and \$37 (with signal) more for a male doctor. These figures are approximately 4% and 2% of the average price and are significant at conventional levels. Racial discrimination is similar across customer genders.[51]

I also segment the customers along various dimensions and consider whether discrimination differs across different customer segments. First, the levels of discrimination signal are similar for older and younger patients, with or without information about doctor quality - the willingness-to-pay racial gaps for both Black and Asian doctors are statistically indistinguishable for customers over or under 55 at conventional levels (see Appendix Table 13). Similarly, the willingness-to-pay racial gaps are statistically indistinguishable

[50] I look at heterogeneity by patient race by evaluating the white and Black sub-populations only. The segments of Asian patients, Native American patients and patients of other races do not have enough observations in my sample to yield precise enough estimates to draw reliable insights.

[51] In contrast to Fershtman and Gneezy (2001), who found that ethnic discrimination is an "entirely male phenomenon."

for customers who recall having a colonoscopy in the past 10 years and those who do not at conventional levels (see Appendix Table 16).

Differences in discrimination across segments of customers emerge when I look at college graduates with those without a college degree (see Appendix Table 14). When there are no quality signals, college graduates place significantly lower (at the 95%-level) willingness-to-pay penalties on both Black and Asian doctors than customers without a college degree. When there are quality signals, the racial gaps for Black and Asian doctors for customers who graduated from college are statistically indistinguishable from zero. For those who did not graduate from college, there remains significant willingness-to-pay penalties for Black and Asian doctors even when there are quality signals.

When I segment customers by their zipcode's political inclinations, I see differences in the level of discrimination against Black doctors when there are no quality signals (see Appendix Table 15). Customers living in zipcodes where the Republican candidate for the 2020 Presidential election won the majority of votes place a willingness-to-pay penalty of $288 on Black doctors relative to white doctors where those living in zipcodes won by the Democratic candidate who place a significantly lower (at 90% level) willingness-to-pay penalty of $203 on Black doctors (a 29.5% difference).[52] In general, customers living in Republican zipcodes places a higher willingness-to-pay penalty on Black and Asian doctors than those in Democratic zipcodes but the differences are not significant at conventional levels except for Black doctors when there are no quality signals.

3.3.2 Robustness

First, the menus of hypothetical doctors are designed so that doctor attributes are exogenously varied. This yields menus of doctors that might not match the demographic composition of real-life doctors. For example, AAMC (2020) finds that 56.2%, 5.0%, and 17.1% of active physicians in the U.S. are white, Black, and Asian respectively but some of the doctor menus presented to some customers will have over-representation of doctors of a certain race (e.g. a menu with 2 Black, 2 Asian, and 1 white doctors). To see if the main results are robust to choices among only sets of doctors with demographic make-ups that are representative of the U.S. population of doctors, I estimate the willingness-to-pay penalties (as in Section 3.3) using only data generated from customer choices within

[52] The data for which presidential candidate won each zipcode is published in Park et al. (2021).

menus of hypothetical doctors that are "representative." A representative menu of doctor option is one where there are at least three white doctors ($> 40\%$), at most two Asian doctors ($< 40\%$) and at most one Black doctor ($< 20\%$). I also require representative menus to have at least 40% male and female doctors.

The willingness-to-pay estimates using only data generated from customer choices within representative menus of hypothetical doctors, presented in Appendix Table 18, are not statistically different from willingness-to-pay estimates using all data. Importantly, using only choice data from these representative menus, I still find that there are large willingness-to-pay penalties on Black and Asian doctors when there are no quality signals and significant drops in such penalties when there are quality signals.

Second, recall that the treatment group is exposed to 5 doctor attributes while the control group is exposed to only 4 (no quality stars). While it is implausible that the difference between four and five variables can produce such huge effects, it is important to consider the possibility that the lower sensitivity to race could be due to competition for customer attention by the higher number of attributes. A related mechanism is salience (Bordalo et al. (2022)): Race might be highly salient in the no quality signal treatment but the quality signal treatment makes the quality score salient.

One way to evaluate the concerns with distraction is to look at whether coefficients for variables unlikely used as proxies for quality differ between treatment and control groups. In the exit survey, customers subjectively estimated average doctor quality of doctors based on their distances and I found that distance is not correlated with customers' subjective assessment of provider quality. This is consistent with customers not using the furthest travel distance categories as a proxy for doctor quality. Therefore, comparisons of the coefficients on travel distances between treatment and control can be used to test whether distraction was significant. The pairwise differences between the average subjective quality for each pair of distance levels have p-values > 0.9. Figure 9 shows that the coefficients for "50-100 miles", "100-250 miles" and "More than 250 miles" are all similar for the treatment when 5 doctor attributes are shown and control when 4 doctor attributes are shown, lending support to the hypothesis that the drop in the racial willingness-to-pay estimates when a quality signal was introduced was not driven by distraction by more attributes.[53]

[53]It remains possible that customers have more complex decision rules that can give rise to these results. For example, if a customer will only attend to two attributes at a time and those attributes

Another way to evaluate the issue of distraction and salience is to look at customers who are exposed to 5 attributes but did not intake information regarding doctor quality from the quality signal itself to see whether they are just like the control condition. Figure 8 shows how the main results from Figure 5 varies by customers who selected "Disagree" or "Strongly disagree" to the statement "A provider with a higher Comprehensive Quality Score is a better provider than a provider with a lower Comprehensive Quality Score" in the exit survey. Those who chose "Disagree" and "Strongly disagree" are "non-believers" of the signals and the rest "believers". For believers the results are similar to the entire population (See Figure 5). However, for non-believers the willingness-to-pay penalty did not go down to 1%. Instead the penalty for Black doctors went down to 7% of average price while the penalty for Asian doctors went down to 3%.

It is hard to conclude whether non-believers were truly not affected by the information embedded in the stars.[54] It is difficult to accurately disentangle how much of the drop in willingness-to-pay penalty for Black and Asian doctors is due to the presence of a distraction from an additional attribute in doctor profiles. Nonetheless, the two main empirical results that there is discrimination towards Black and Asian doctors and that most of the discrimination disappears upon the introduction of a quality signal, are still true comparing within the treatment group between believers and non-believers. The penalties for Black and Asian doctors are statistically significant at conventional levels, as are the differences of these penalties between the believers and non-believers. The willingness-to-pay penalty for Black doctors for believers is 8.9% of that for non-believers and the willingness-to-pay penalty for Asian doctors for believers is 26.8% of that for non-believers.

4 Validation of Estimated Preferences

This section compares predicted choice with actual choice to validate the study design. Recall that a menu of actual doctors was presented to each customer after they responded to the 14 hypothetical choice tasks. Doctor attributes that the customers can see in the

are distance and race when quality signals are not available but distance and quality when signals are available, we could see results like the ones provided here.

[54] In Appendix Table 19, one can see that non-believers place a statistically significantly higher willingness-to-pay for doctors with 5 stars relative to doctors with 1 or 2 stars despite the fact that the estimates are very imprecisely estimated.

menu of actual doctors match what they saw with the hypothetical doctors.

Each customer was matched to actual doctors at the end of the day when they completed the preference elicitation. Before each customer was matched to a set of actual doctors, I estimated mixed logit models with all the data collected until then to obtain a set of individual-level parameters for the purposing of matching customers to actual doctors for booking purposes. For the ethical reason to not disadvantage doctors' prospects of getting business due to their race, I adopted the approach of Kessler et al. (2019) and included in these models used for matching all features described in Section 2.3 except the perceived race variable.[55] Every doctor is scored based on estimated choice probability for each customer, and the 10 options of real doctors are just the 8 with the highest choice probability plus 2 randomly drawn from the remaining options without replacement.

At the end of the experiment, a mixed logit model was estimated using all the data collected from customers including all variables mentioned in Section 2.3 with hierarchical Bayes and individual-specific choice probabilities calculated for each doctor in the menus offered to every customer ("an ex post model"). A rank is assigned for each doctor within each customer menu based on their ex post predicted choice probability specific to that customer.

Figure 10 shows the percentage of actual booking choices by the 188 customers who booked with one of the 10 doctors matched to them out of 224 who completed the hypothetical preference elicitation by that option rank within their menu. Those who booked are statistically indistinguishable from those who did not on baseline observables (see Appendix Table 20).[56] Furthermore, the main results presented in Section 3.3 are similar between the customers who booked and those who did not book (see Appendix Table 17). All bars in Figure 10 show how actual bookings match up to predicted rankings of the options and predicted rankings are determined by the discrete choice model estimated using the choice data from hypothetical doctors. Within each pair of bars in Figure 10, the lighter grey bar shows the results when the specification of the discrete choice model

[55] Estimations were done with hierarchical Bayes using the Allenby-Train procedure (Allenby (1997), Train (2009), Ben-Akiva et al. (2019)) with 8,000 burn-in iterations and 8,000 iterations after burn-in.

[56] Customers who booked after being matched to 10 doctors are similar to those who did not proceed to book. Recall that the provider interviews by the platform's staff members indicate that there is no evidence that any bookings were not converted into actual clinical encounters. The customers who booked have menus of 10 actual doctors that are less likely to include any top 8 doctor options that are less than 250 miles away from them (due to each customer's idiosyncratic location relative to the doctors). See Appendix Table 20.

includes the race dummies as input variables while the darker grey bars shows results when the choice model do not include race as its arguments.

It is not surprising that most bookings are with doctors who had the highest predicted (fitted with ex post models) choice probability rank while none of the bookings are with the randomly chosen options that tend to have much lower predicted choice probability than the other 8. The individualized predicted choice probabilities predict the actual observed bookings and race does not offer additional predictive power on top of predicted choice rank based on the ex post choice model that include race dummies as arguments.[57] The top-3 accuracy score (the proportion of subjects booking with one of the predicted first, second, or third choices) and top-5 accuracy score for the fitted choice probability (using the estimated model that includes race dummies as arguments) are 82% and 93% respectively. In general, actual choice of an option is more likely the higher ranked the option is within its menu (Figure 10). Finally, the choice ranks represented by lighter grey bars in Figure 10 (predicted using doctor characteristics including race) are more predictive of actual choices than the dark grey ones (predicted using doctor characteristics excluding race), suggesting that race as an argument (as opposed to only the other doctor characteristics, e.g., only price) contributes to the estimated choice model's predictive power for actual doctor choices.

I also estimate a discrete choice model using the actual booking data alone. There are much fewer observations compared to the choice data from hypothetical choices: one menu per customer as opposed to 14 menus per customer for the latter. These estimates are reported in Appendix Table 21. The willingness-to-pay estimates are much less precise than those from the hypothetical choices given the smaller amount of information per customer, especially for the race dummies. This is due to limited variation in doctors race among actual doctors, relative to the hypothetical doctor options. Nonetheless, the estimated coefficients for all the doctor characteristics are similar in magnitudes and statistically indistinguishable between the discrete choice model estimated using hypothetical data and the model estimated using actual booking data at conventional levels.

[57] I tested whether race can predict which options are forgone in favor of a doctor option with lower ex post predicted choice probability. In particular, a linear probability model ($Y_{ic} = \psi_0 + \psi_1 Black_i + \psi_2 Asian_i + \zeta_{ic}$, where Y_{ic} is an indicator for whether option i was not chosen by customer c who selected one of the other 9 options with a lower ex post predicted choice probability (based on the discrete choice model that includes race dummies), $Black_i$ and $Asian_i$ are race dummies, and ζ_{ic} is the error term) was estimated using the 1880 doctor options presented to the 188 customers who booked with one of the 10 doctors matched to them. Neither ψ_1 nor ψ_2 is significant at conventional levels.

The discrete choice model estimated with the choice data over hypothetical options seems able to predict actual choice behavior. This is consistent with the preference elicitation mechanism being incentive-compatible.

5 Interpretation of Results

This section sets up a theoretical framework to link the empirical results to models of discrimination. To illustrate the effects of introducing a quality signal, the decision problem of customers who may be risk averse is evaluated.[58] One particular model may not explain the results. Rather, assumptions about customer beliefs and preferences leading to behavior consistent with a few stylized facts documented in the previous sections are developed.

Two key stylized facts will be addressed in this section. First, the data indicate that without quality signals, Black and Asian doctors suffer a significant willingness-to-pay penalty about two to three times larger than the willingness-to-pay penalty for 1-star relative to 5-star doctors (see Table 2). Second, the willingness-to-pay penalty for Black and Asian doctors drops by about $80\% - 90\%$ when a quality signal is introduced (see Figure 5).

5.1 A Model of Discrimination

I first present a different model from the preceding discrete choice model in order to facilitate a clear interpretation. Specifically, I assume the customer observes the group identity (i.e. race) of doctors $j \in \{B, W\}$. Doctor quality q is assumed to equal the value of the doctor's marginal product to the customer, and it is drawn from a normal distribution[59] $N(\mu_j, \sigma_j^2)$. I consider noisy unbiased signals of doctor quality, $\theta = q + \epsilon$, where ϵ is a zero-mean error that is normally distributed according to $N(0, \sigma_{\epsilon_j}^2)$.

Following Aigner and Cain (1977) and Fang and Moro (2011), I assume that customers

[58] Theoretical models for statistical discrimination that incorporate risk-averse employers have been proposed (e.g., Aigner and Cain (1977)). However, the labor economics literature has favored risk-neutral employers as a more realistic model when firms are doing the hiring (Fang and Moro (2011)). The customers making the decisions in the present paper are individual prospective patients, not firms. Therefore, I do not assume away risk aversion for their utility functions.

[59] Typical (accurate) statistical discrimination models assume that the customer knows this true underlying distribution of quality by identity group. As I will discuss in Section 5.1.1, the assumption that customers hold inaccurate statistical models will need to be explored in an attempt to reconcile the models with the key stylized facts listed at the beginning of Section 5

are risk averse with (money-metric) utility:

$$V_c(q,j) = \alpha_j - \beta \exp(-\gamma q) + \upsilon_c, \tag{7}$$

where υ_c is a utility shifter idiosyncratic to customer c (iid with mean 0) and α_j captures the preference for doctors from group j independent of quality. The difference $(\alpha_W - \alpha_B)$ is referred to as taste-based discrimination against doctors of group B relative to group W. Expected utility from booking a doctor with signal θ is given by

$$\mathbb{E}[V_c(q,j)|\theta,j] = \alpha_j - \beta \exp[-\gamma\, \mathbb{E}(q|\theta,j) + \frac{\gamma}{2} Var(q|\theta,j)]. \tag{8}$$

A customer's willingness-to-pay for a doctor is equal to the expected utility conditional on the quality signal and the group identity. Quality and signal are jointly normally distributed. Using the properties of the conditional normal distribution,[60] the willingness-to-pay conditional on the signal, θ, and group identity of the doctor, j, is

$$WTP(\theta,j) = \mathbb{E}[V_c(q,j)|\theta,j] = \alpha_j - \beta \exp[-\gamma(\frac{\sigma_j^2}{\sigma_j^2+\sigma_{\epsilon_j}^2}\theta + \frac{\sigma_{\epsilon_j}^2}{\sigma_j^2+\sigma_{\epsilon_j}^2}\mu_j) + \frac{\gamma}{2}\frac{\sigma_j^2\sigma_{\epsilon_j}^2}{\sigma_j^2+\sigma_{\epsilon_j}^2})]. \tag{11}$$

5.1.1 No signal versus quality signals

Consider the customer decision problem under the control and treatment conditions. Recall that signals of doctor quality are unbiased with a noise term normally distributed according to $N(0,\sigma_{\epsilon_j}^2)$. Under the control condition with no quality signals sent, one can think of the customer as having a signal where $\sigma_{\epsilon_j}^2 \to \infty$. It follows that willingness-to-pay can be written as

$$\lim_{\sigma_{\epsilon_j}^2 \to \infty} WTP(\theta,j) = \alpha_j - \beta \exp[-\gamma\mu_j + \frac{\gamma}{2}\sigma_j^2]. \tag{12}$$

[60] If the quality and signal are jointly normally distributed, the conditional distribution of q given θ is normal with mean equal to

$$\mathbb{E}(q|\theta,j) = \frac{\sigma_j^2}{\sigma_j^2+\sigma_{\epsilon_j}^2}\theta + \frac{\sigma_{\epsilon_j}^2}{\sigma_j^2+\sigma_{\epsilon_j}^2}\mu_j \tag{9}$$

and variance equal to

$$Var(q|\theta,j) = \frac{\sigma_j^2\sigma_{\epsilon_j}^2}{\sigma_j^2+\sigma_{\epsilon_j}^2}. \tag{10}$$

(DeGroot (2005)).

Recall that μ_j and σ_j^2 are the mean and variance of the underlying distribution of doctor quality for group j. If a customer thinks that the two groups have the same underlying quality distribution, any willingness-to-pay differences of that customer that are observed under the control condition will be attributable to taste-based discrimination $(\alpha_W - \alpha_B)$ within the context of this paper (as the second term in (12) will be differenced out).

On the other hand, customers can observe a quality signal under the treatment condition. This is modeled as a signal with a finite $\sigma_{\epsilon_j}^2$. For the sake of simplicity and a clear contrast with the control group, I assume that under the treatment condition $\sigma_{\epsilon_j}^2 = 0$ from the customer's perspective. I will discuss some implications of $\sigma_{\epsilon_j}^2$ being finite but not equal to zero below. For now, under the treatment condition where $\sigma_{\epsilon_j}^2 = 0$, willingness-to-pay can be written as:

$$\lim_{\sigma_{\epsilon_j}^2 \to 0} WTP(\theta, j) = \alpha_j - \beta \exp[-\gamma\theta], \tag{13}$$

where θ is the quality signal (number of stars). Conditional on quality signal θ, when $\sigma_{\epsilon_j}^2 = 0$, any willingness-to-pay differences of the customer that are observed under the treatment condition will be attributable to taste-based discrimination $(\alpha_W - \alpha_B)$.

Finally, using this set-up and the stylized fact that Black and Asian doctors suffer a significant willingness-to-pay penalty about two to three times larger than the willingness-to-pay penalty for 1-star relative to 5-star doctors without quality signals, the assumption of risk aversion is justified. If I assume risk neutrality, where utility is a function of expected quality (but not variance in quality), the customer will have to believe that the unconditional mean quality of white doctors is higher than the unconditional mean quality of Black and Asian doctors by 16.97 stars and 11.24 stars respectively (on a 5-star scale) to be consistent with data.[61] Such customer priors about the unconditional mean quality of doctors of different racial groups depart substantially from the actual differences in mean of doctor quality by race (measured by stars) in our pool of doctors. Under the basic set-up above, the assumption of risk averse customers seems more credible.

[61] A conference discussant had suggested that customers who are worried about extreme left-tail risk (e.g. Black doctors who have quality of effectively < -10 stars) could also give rise to the substantial racial gap when no quality signals are available. For this argument to be adequate for explaining the observed racial gap, the WTP penalty for Black doctors relative to white doctors conditional on a doctor being a 1-star doctor would have to be substantial. The actual data shows that the WTP penalty for Black doctors relative to white doctors conditional on a doctor being a 1-star doctor is only at 3.6% of the average colonoscopy price (see Column 8 of Table 2) - far smaller than required to explain the much larger drop in Black-white gap.

5.2 Biased Belief Discrimination as a Driver of Willingness-to-pay Penalty

Next, I attempt to unpack the sources of customer discrimination with the theoretical framework above and the other stylized fact mentioned in the beginning of this Section.

Traditional taste-based discrimination cannot be the whole story: If there is no statistical discrimination, biased belief discrimination, or deniable prejudice, willingness-to-pay differences between white and minority doctors in the control group (differences in willingness-to-pay in (12)) and in the treatment group (differences in (13)) would both equal $(\alpha_W - \alpha_B)$. However, the willingness-to-pay penalty for minority doctors in the control condition where no quality signal is provided is about eight to ten times larger than the willingness-to-pay penalty for minority doctors in the treatment condition where the quality signal is introduced (see Figure 5). Standard information manipulation experiments as in List (2004) have interpreted this stylized fact as evidence that most observed discrimination is statistical in nature.

What can one say about the nature of this discrimination here? Many of the previous papers trying to parse the types of discrimination assume that decision makers hold accurate beliefs about the true distribution of quality.[62] For instance, they might compare the hiring decisions to the true underlying distribution of the quality for these decisions (for example, Arnold et al. (2018)).

Access to the underlying quality distributions (Figure 11) by identity group for the doctors in this study allows me to learn the true distribution of quality for each identity group. Pairwise Kolmogorov-Smirnov tests for equality of distribution functions fail to reject the hypotheses that any pair of the racial groups have similar quality distributions in this sample: approximate p-values for the combined tests for each pair of racial groups are ≥ 0.91. The mean quality scores (μ_j) for white, Black, and Asian doctors are 2.92, 2.77, and 2.83 with variances (σ_j) 1.58, 1.86, and 1.60. The proportion of doctors in each quality level is shown in Figure 11. Given the model in (12) and (13), there is no finite parameter value γ for which the Black-white willingness-to-pay gap is three times or larger than the gap between 1 star and 5 star doctors under accurate beliefs μ_j and

[62] Practically, it is unlikely that customers have accurate beliefs about doctor quality by race. For instance, Pope and Sydnor (2011) raised doubts that respondents would have accurate priors for the probability of repayment between black and white borrowers on the specific and new platform that formed their empirical setting.

σ_j. In other words, there does not exist finite γ such that:[63]

$$\underbrace{\exp[-\gamma\mu_W + \frac{\gamma}{2}\sigma_W^2] - \exp[-\gamma\mu_B + \frac{\gamma}{2}\sigma_B^2]}_{\text{Black-White gap (accurate statistical discrimination)}} = \exp[-2.92\gamma + 1.58\frac{\gamma}{2}] - \exp[-2.77\gamma + 1.86\frac{\gamma}{2}]$$
$$= \underbrace{3(\exp[-5\gamma] - \exp[-\gamma])}_{3\times(\text{5 star-1 star gap})}. \tag{14}$$

The possibility that consumers held accurate beliefs about quality with race as a proxy could be ruled out.

Next, I examine whether biased beliefs can explain the observed disparities in willingness-to-pay, a direction proposed by Bohren et al. (2019b) and others.

Biased belief discrimination stems from incorrect beliefs held by the customers. While accurate statistical discrimination is sometimes referred to as "efficient discrimination" biased belief discrimination can give rise to inefficiencies that policymakers might want to address through information provision. In the theoretical framework, allowing for biased beliefs is modeled as allowing customers to hold subjective beliefs $\hat{\mu_j}$ and $\hat{\sigma_j}$ about the mean and variance of doctor quality for group j where $\hat{\mu_j}$ and $\hat{\sigma_j}$ are not required to equal μ_j and σ_j. In other words, customers are allowed to have a mis-specified model of the quality distribution. Consequently, the willingness-to-pay in the control condition where no quality signal is provided is now given by $\alpha_j - \beta\exp[-\gamma\hat{\mu_j} + \frac{\gamma}{2}\hat{\sigma_j}^2]$ instead of the expression in (12).

In the simple case outlined in the theoretical framework above, we still obtain willingness-to-pay penalty for doctors from group B relative to doctors from group W under the treatment condition (with quality signal) as $(\alpha_W - \alpha_B)$ from (13). With biased beliefs, the willingness-to-pay penalty for doctors from group B relative to doctors from group

[63] At any $\gamma > 0$, the level and slope of the term $\exp[-2.92\gamma+1.58\frac{\gamma}{2}]-\exp[-2.77\gamma+1.86\frac{\gamma}{2}]-3(\exp[-5\gamma]-\exp[-\gamma])$ are either both positive or the former is positive while the slope is negative and of a smaller absolute magnitude. This implies that the term above cannot converge to zero for any finite $\gamma > 0$. In other words, (14) cannot hold for any finite $\gamma > 0$.

W under the control condition is given by

$$\lim_{\sigma^2_{\epsilon_W}\to\infty} WTP(\theta,W) - \lim_{\sigma^2_{\epsilon_B}\to\infty} WTP(\theta,B) = \underbrace{(\alpha_W-\alpha_B)}_{\text{taste-based}} + \underbrace{\beta(\exp[-\gamma\hat{\mu_B}+\frac{\gamma}{2}\hat{\sigma_B}^2]-\exp[-\gamma\hat{\mu_W}+\frac{\gamma}{2}\hat{\sigma_W}^2])}_{\text{biased belief}} \tag{15}$$

where the first term represents taste-based discrimination and the second term gives the biased belief discrimination. It is possible to get a biased belief discrimination term that is positive and large relative to the taste-based discrimination term under biased beliefs. If I assume biased beliefs that are large enough (say 5 to 10 times larger than the term for taste-based discrimination), it is possible to observe willingness-to-pay penalty for Black and Asian doctors dropping by about $80-90\%$ when a quality signal is introduced.

5.2.1 Behavioral Mechanisms for Biased Beliefs

Biased beliefs described above can come from many different behavioral models. The present experiment will not conclusively identify a specific behavioral model. However, I will use one candidate mechanism to obtain "reasonable" biased beliefs to evaluate additional assumptions necessary to generate the racial gaps at magnitudes consistent with some key empirical findings.

Suppose customers are risk averse, and they deploy Kahneman and Tversky's representativeness heuristic (Tversky and Kahneman (1983)) to assess a target doctor racial group's quality. I assume that customers overweights the probability of those quality score levels that are most representative of the minority doctor racial groups relative to white doctors as in the stereotyping model formalized by Bordalo et al. (2016). A quality score level q is *representative* of group B relative to group W if it scores high on the likelihood ratio $\frac{P(q|B)}{P(q|W)}$. Stereotyping can exaggerate inter-group differences in means and variances.

To simplify my illustrative example, I will focus on Black and white doctors and proceed as if customers are evaluating doctors from these two groups only.[64] The repre-

[64]The stereotyping model in Bordalo et al. (2016) also focuses on 2-group comparisons. In the present experiment, customers have 3 groups of doctors to compare against each other. As there are no straightforward published extension of the Bordalo et al. (2016) stereotyping model to more than 2 comparison groups, I simplify my illustration so that I can directly apply their model. Future work to extend the

sentativeness of each quality score level is shown in the fourth row of Table 3. 1 star is most representative for Black doctors relative to white doctors; 5 stars is most representative for white doctors relative to Black doctors.[65]

Consider a specific stereotyping model where customers overweight or underweight quality score levels for doctor racial groups based on a representativeness rank-based discounting function. In particular, I assume that customers attach to each quality score level $q \in \{1, 2, 3, 4, 5\}$ in group j a distorted probability:

$$P^{st}(q|j) = P(q|j)\frac{\delta(rank(q))}{\sum_{q' \in \{1,2,3,4,5\}} P(q'|j)\delta(rank(q'))}, \tag{16}$$

where $P(q|j)$ is the true probability of quality score level q in group j, $rank(q) \in \{1, 2, 3, 4, 5\}$ is the representativeness ranking for quality score level q (where 1 is the most representative), and $\delta(rank(q)) \in [0, 1]$ is the discount factor applied to the odds of level q given its representativeness ranking $rank(q)$. I further assume, for simplicity, that $\delta(rank(q)) = 0.2^{rank(q)}$.[66] In this particular stereotyping model, customers discounts by a constant factor ($\delta = 0.2$) the odds of quality score level q relative to its immediate predecessors in the representativeness ranking.[67]

Stereotypical thinking distort customers' subjective probability for each q, and therefore generate inaccurate beliefs $\hat{\mu_j}$ and $\hat{\sigma_j}$ that deviate from the true quality distribution (μ_j and σ_j) for each group j. Applying this model to the empirical data on doctor quality distribution (see Figure 3), I obtain $\hat{\mu_B} = 1.65 \neq 2.77 = \mu_B$ and $\hat{\sigma_B} = 1.56 \neq 1.86 = \sigma_B$ for Black doctors and $\hat{\mu_W} = 4.35 \neq 2.92 = \mu_W$ and $\hat{\sigma_W} = 1.15 \neq 1.58 = \sigma_W$ for white doctors. With these inaccurate beliefs, I can find γ^* such that

$$\underbrace{\exp[-\gamma^* \hat{\mu_W} + \frac{\gamma^*}{2}\hat{\sigma_W}^2] - \exp[-\gamma^* \hat{\mu_B} + \frac{\gamma^*}{2}\hat{\sigma_B}^2]}_{\text{Black-White gap (biased belief discrimination)}} = \underbrace{3(\exp[-5\gamma^*] - \exp[-\gamma^*])}_{3\times(\text{5 star-1 star gap})}. \tag{17}$$

This allows the model to account for the first stylized fact presented in the beginning of

stereotyping model to multi-group comparison will be welcome.

[65] Analogous results for Asian doctors, compared to white doctors, are presented in Appendix Table 22.

[66] This is (ii) of rank-based stereotype weighing function options outlined on page 1763 in Bordalo et al. (2016).

[67] I arbitrary picked this unobservable discount factor and applied to an out-of-the-box option from Bordalo et al. (2016) to benefit the exposition. Other weighting functions used by stereotyping agents are outlined and discussed in Bordalo et al. (2016).

this section.

Plugging in the derived values for $\hat{\mu_B}$, $\hat{\sigma_B}$, $\hat{\mu_W}$, and $\hat{\sigma_W}$ to (17), I find that $\gamma^* = 8.45$.[68] A customer with expected utility described by (8) and $\gamma = 8.45$ will be indifferent between being assigned a doctor via a lottery with equal chances of drawing a 4 star doctor and a 3 star doctor and being assigned a doctor with 3.1 stars for certain.

The specific stereotyping model in this subsection is by no means the only explanation for the biased beliefs that is consistent with empirical results in the present experiment. Stereotyping models in general are just one set within a much larger space of behavioral models that can yield predictions consistent with the experimental results. In fact, as I discuss below, a takeaway from this section is that assumptions required to interpret the empirical results solely with biased beliefs might be too aggressive and ambiguous to be plausible.

5.3 Imperfect Signal Precision and Residual Willingness-to-pay Penalty for Minorities

Next, consider the implications of relaxing the assumption that the quality signal has perfect precision ($\epsilon_j^2 = 0$) under the treatment condition regarding ability to interpret the empirical estimates in this paper. To do this, consider the subjective beliefs about signal precision denoted as $\hat{\sigma_{\epsilon_j}}^2$ for identity group j, that can differ by identity group of the doctor. The willingness-to-pay penalty for doctors from group B relative to doctors from group W under the treatment condition, where a quality signal is provided, is given by

$$\begin{aligned} WTP(\theta, W) - WTP(\theta, B) = & \underbrace{(\alpha_W - \alpha_B)}_{\text{taste-based}} \\ & \underbrace{\begin{aligned} &+ \beta\{\exp[-\gamma(\frac{\hat{\sigma_B}^2}{\hat{\sigma_B}^2 + \hat{\sigma_{\epsilon_B}}^2}\theta + \frac{\hat{\sigma_{\epsilon_B}}^2}{\hat{\sigma_B}^2 + \hat{\sigma_{\epsilon_B}}^2}\hat{\mu_B}) + \frac{\gamma}{2}\frac{\hat{\sigma_B}^2 \sigma_{\hat{\epsilon_B}}^2}{\hat{\sigma_B}^2 + \hat{\sigma_{\epsilon_B}}^2})] \\ &- \exp[-\gamma(\frac{\hat{\sigma_W}^2}{\hat{\sigma_W}^2 + \hat{\sigma_{\epsilon_W}}^2}\theta + \frac{\hat{\sigma_{\epsilon_W}}^2}{\hat{\sigma_W}^2 + \hat{\sigma_{\epsilon_W}}^2}\hat{\mu_W}) + \frac{\gamma}{2}\frac{\hat{\sigma_W}^2 \hat{\sigma_{\epsilon_W}}^2}{\hat{\sigma_W}^2 + \hat{\sigma_{\epsilon_W}}^2})]\}. \end{aligned}}_{\text{biased belief}} \end{aligned} \tag{18}$$

[68]Recall that the input to utility function in (7) is quality but not money. This estimated $\gamma^* = 8.45$ is not directly comparable with γ estimates from exponential utility models of money found in the finance or macroeconomics literature.

In Section 5.2, the observed willingness-to-pay penalty for minority doctors in the treatment condition ("residual") is taste-based discrimination. While the observed discriminatory behavior was decomposed into taste-based discrimination and biased belief discrimination with perfect signal precision, such decomposition might not be viable without additional assumptions about unobservable customer beliefs.

If one assumes customers subjectively believe that minority doctors have lower mean quality ($\hat{\mu_B} < \hat{\mu_W}$), higher variability in quality ($\hat{\sigma_B} > \hat{\sigma_W}$) and weakly lower signal precision given the same quality scoring system ($\hat{\sigma_{\epsilon_B}}^2 \geq \hat{\sigma_{\epsilon_W}}^2$), the biased belief discrimination term in (18) will be positive. In this case, it is possible that the willingness-to-pay penalty for minority doctors in both the treatment and control conditions are entirely due to biased belief discrimination (the biased belief discrimination terms of (18) and (15)) while taste-based discrimination is equal to zero ($\alpha_W - \alpha_B = 0$). On the other hand, if customers subjectively believe minority doctors have higher mean quality ($\hat{\mu_B} > \hat{\mu_W}$), lower variability in quality ($\hat{\sigma_B} < \hat{\sigma_W}$) and weakly higher signal precision given the same quality scoring system ($\hat{\sigma_{\epsilon_B}}^2 \leq \hat{\sigma_{\epsilon_W}}^2$), the biased belief discrimination term in (18) will be negative with the biased belief discrimination favoring the minority group when the imprecise signal is introduced. A negative and large enough biased belief discrimination term can explain the stylized fact where the observed overall willingness-to-pay penalty for minority doctors is much smaller in the treatment condition compared to the control condition, even if the willingness-to-pay penalty in the control group is entirely accounted for by traditional taste-based discrimination. Although the former case is more reasonable given the previous literature on customer discrimination, it is clear that any other case with different mixtures of taste-based and biased belief discrimination could be consistent with observed data in this experiment.[69] The data tell us: traditional taste-based discrimination alone cannot be the whole story with biased belief discrimination showing up prominently in either the control or treatment or both conditions taking on either sign. It is therefore worth noting that interpretation of the data will ultimately be tied to assumptions about signal precision and subjective beliefs.

[69]Depending on the relative sizes of the biased belief discrimination terms in Equations 15 and 18.

5.4 Deniable Prejudice

In the no quality signal treatment, a substantial degree of risk aversion with respect to quality ($\gamma^* = 8.45$ or customers being indifferent between a 50/50 lottery between a 3-star and 4-star doctor and a 3.1-star doctor for certain) is required to account for the observed discrimination using behavioral models like stereotypes. In the same time, as mentioned in Section 5.2, the true quality distributions are statistically indistinguishable. These forces us to make very strong and potentially implausible assumptions[70] if we rely solely on biased belief-generating mechanisms like stereotypes.

Another source of the racial gap is one where a noisier quality signal (or the lack of quality information) allows customers "moral wiggle room" (Exley and Kessler (2021)) to express or magnify their taste-based discrimination or prejudice against minority doctors.[71] Likewise, customers might use the moral wiggle room to express their in-group preferences for doctors based on gender. In the present case, we can express deniable prejudice by replacing α_j in Section 5.1 by a function $\alpha_j(\sigma^2_{\epsilon_j})$ where $\frac{d(\alpha_j(\sigma^2_{\epsilon_j}))}{\sigma^2_{\epsilon_j}} \leq 0$ if j is a minority group and $\frac{d(\alpha_W(\sigma^2_{\epsilon W}))}{\sigma^2_{\epsilon_W}} \geq 0$ (j if white). I replace the equation for willingness-to-pay in Equation 11 by

$$WTP(\theta, j) = \alpha_j(\sigma^2_{\epsilon_j}) - \beta \exp[-\gamma(\frac{\sigma^2_j}{\sigma^2_j + \sigma^2_{\epsilon_j}}\theta + \frac{\sigma^2_{\epsilon_j}}{\sigma^2_j + \sigma^2_{\epsilon_j}}\mu_j) + \frac{\gamma}{2}\frac{\sigma^2_j \sigma^2_{\epsilon_j}}{\sigma^2_j + \sigma^2_{\epsilon_j}})], \tag{19}$$

deniable prejudice is $(\alpha_W(\sigma^2_{\epsilon_W}) - \alpha_B(\sigma^2_{\epsilon_B}))$.[72]

In Figure 6, I show that Black patients place a willingness-to-pay penalty on Black doctors when there is no quality signal but yield a point estimate is suggestive of a willingness-to-pay premium when there is a quality. First, biased beliefs towards in-

[70]For instance, that the customers have a very precise, shared knowledge of the underlying quality distributions and the non-monotonic representativeness of quality scores between Black and white doctors (that 1 star is most representative for Black doctors relative to white doctors, followed by 4 stars; 5 stars is most representative for white doctors relative to Black doctors, followed by 3 stars).

[71]To check if "moral wiggle room" might come in other flavors, I estimate the preferences using only menus where there is at least one white doctor who dominates minorities in the same menu on at least one attribute. Customers might use the difference in that one attribute as an excuse. However, the willingness-to-pay penalties associated with less desirable levels of doctor attributes for each attribute (e.g. farther travel distances, lower quality) that are estimated using only data from menus where at least one white doctor dominates minorities in that attribute are not statistically distinguishable from the willingness-to-pay penalties estimated using the full data. I will focus on deniable prejudice where customers might have exploited the "moral wiggle room" from the lack of quality information.

[72]For illustrative purposes, assuming that $\sigma^2_{\epsilon_W} = \sigma^2_{\epsilon_B} = \sigma^2_\epsilon$, if $\lim_{\sigma^2_\epsilon \to 0}(\alpha_W(\sigma^2_\epsilon) - \alpha_B(\sigma^2_\epsilon)) = \$25.803$ and $\lim_{\sigma^2_\epsilon \to \infty}(\alpha_W(\sigma^2_\epsilon) - \alpha_B(\sigma^2_\epsilon)) = \$270.133$, deniable prejudice can fully account for the empirical results on the Black-white gap.

group members have been well-documented in some settings (e.g. Fershtman and Gneezy (2001)). Second, the willingness-to-pay penalty from Black customers is statistically not distinguishable from that of a 4-star-versus-1-star quality difference at conventional levels. We could assume that there is significant taste-based discrimination by Black customers towards Black doctors, and focus on deniable prejudice as the mechanism. However, for Black customers, biased beliefs that Black doctors have three to four fewer stars than white doctors can also give rise to the observed empirical result.

Similarly, the empirical results on the preference for gender concordance in Figure 7 can be readily accounted for by a mechanism like deniable prejudice. It can also be reconciled with only biased belief discrimination if male and female customers happen to have exactly opposing biased beliefs of a similar absolute magnitude.

It is beyond the scope of this paper to select a specific behavioral model to rationalize the empirical results. It is best to acknowledge that a mixture of mechanisms are at work. This mixture likely involve not just the traditional mechanisms of taste-based and statistical discrimination, even when we consider biased beliefs (e.g. deniable prejudice is also plausible). Importantly, this and the discussions in the previous sections establish that the traditional test for statistical discrimination (that relies solely on the provision of a quality signal) has limitations on its ability to definitively decompose the nature of discrimination. The present paper makes a clear methodological challenge to the conclusions drawn by previous work that consider only the dichotomy of traditional taste-based discrimination and statistical discrimination.

6 Discussion

6.1 Implications for Healthcare Labor Markets

Since Arrow (1963), economists have been aware consumers face difficulties determining the quality of services provided by doctors because of asymmetric information. In this paper, customers of healthcare are found to use race as a proxy for quality in light of asymmetric information. Most doctors in the U.S. are employed by hospitals and groups, not directly negotiating prices with customers. However, lower willingness-to-pay associated with minority doctors could make them less attractive to employers. Furthermore, we need to ensure minority doctor entry is not deterred in specialties where trust is important for better population health outcomes (Alsan et al. (2019)). Such specialties are also likely those in which customer discrimination can play the largest role as in primary care or oncology. Specialty choice is responsive to wages (Nicholson and Propper (2011)). Should the willingness-to-pay penalties translate into lower wages for minority doctors, diversity of the workforce in these specialties will be affected.

A key result in this paper is that the provision of a quality signal increases the relative willingness-to-pay for Black and Asian doctors on average. While the magnitude of the willingness-to-pay penalties is similar to the raw wage gaps for these minority groups relative to the average white doctors, it is hard to directly map the impact of information provision to actual effects on the racial wage gap or labor force participation by minorities in medicine due to the industrial organization of medicine and unknown general equilibrium effects. Nevertheless, this paper strongly suggests that the provision of provider quality information can effectively reduce much of the discrimination in healthcare labor markets by customers.[73]

While quality signals decrease discrimination to Black and Asian doctors as a group by $80\%-90\%$, doctors within each racial group with poor quality ratings might find it harder to get customers at the same price as their peers with better quality ratings. And while the average quality scores of Black and Asian doctors do not differ from white doctors for the labor pool in the marketplace of the present experiment, Black and Asian doctor ratings could be biased downwards due to racial preferences from majority patients in the broader healthcare market (Cooper et al. (2003)). In this case, quality ratings might help perpetuate racial biases held by customers. What one can confidently conclude from this paper is that provision of unbiased quality signals can reduce racial disparities by lowering expressed deniable prejudice and biased belief discrimination. Caution must be applied, however, when drawing welfare conclusions when adopting information provision as a policy.

The present results have implications for the general problems of customer discrimination and for discrimination in the healthcare sector. As more online marketplaces emerge for healthcare, clear market design implications are provided in this paper for platforms

[73]Skeptical readers might even claim that as VIC informs respondents that the menu options (including the race of the doctors) are hypothetical, one should not describe the observed results as evidence for some form of traditional statistical or biased belief discrimination at play. Similar to Kessler et al. (2019), I will not be able to explore potential forms of behavioral models to satisfy such skeptical readers. However, even the skeptical readers will agree that I found evidence that provision of quality information can reduce discrimination in healthcare labor markets.

and publicly run marketplaces that aspire to provide a discrimination-free environment. Online platforms create new markets by eliminating search friction and facilitating transactions. Even online, service professionals face discrimination. While the study setting is a specialized market compared to the majority of insured healthcare customers, the unique nature of the setting enabled study design to causally investigate the important topic of customer discrimination. Previous work had suggested that by "reducing the salience of race, platforms could reduce discrimination" (Fisman and Luca (2016)). In light of the evidence presented in this paper, market designs that favor better information flow regarding quality of providers can also reduce discrimination. And since both white and minority patients might have a preference for doctor racial concordance,[74] the more welfare-improving design might be closer to the treatment in my experiment than suppressing provider race information in online platforms.

6.2 A Methodological Point on the Validated Incentivized Conjoint Approach

This paper introduced the novel methodology to overcome hypothetical bias while eliciting preferences. My design simulates market conditions via an experimental design and offers superior recovery of preferences compared to a direct stated preference elicitation (Louviere and Hensher (1983)). The value of analyses of the present kind is most prominent in many markets where there is very little variation in key product attributes (e.g., prices in the short run) with the set of existing products representing only a very sparse set of points in the product characteristic space (Allenby et al. (2019)). The biggest advantage is that it enables the researcher to avoid "endogeneity" problems as all variation in both product attributes and price is exogenous and usable to estimate preference.

Choice-based conjoint analysis is one of the most popular tools in marketing research with an estimated 14,000 studies conducted yearly by various industry and academic research bodies (Allenby et al. (2019)). Despite this, the uptake in economics is limited. Economics experiments generally aspire to have incentive compatibility properties so that the subjects reveal their true preferences due to stakes aligned with their decisions. Hypothetical bias limits the uptake of conjoint analysis by economists. There are more

[74] See the suggestive evidence in Section 3.3.1. Also, evidence from Alsan and Wanamaker (2018) and Alsan et al. (2019) would suggest that racial concordance between patient and doctor can enhance trust between them, especially if the patient is Black.

conjoint studies conducted every year by researchers and businesses than randomized clinical trials conducted by medical researchers. Further development of the present method will provide a useful empirical tool for economists to complement methods using observational data.

Practically, many tech companies can avail themselves of my methodology to improve the performance of their offerings when accurate predictions of customer preferences is critical to their success. While usage data can tell platforms a lot about customer preferences, periodically eliciting preferences from customers to adjust presented contents can help companies such as social media platforms obtain information orthogonal to their usual data sources much like the company that partnered with me on this experiment.

Naturally, the methodology has limitations. First, it can only be used to estimate demand and cannot be used to compute market equilibrium outcomes such as market prices (wages for doctors in this paper) or equilibrium product positioning in characteristic space. Without supply assumptions and cost information, one is limited about statements on equilibrium outcomes and welfare. Second, preference elicitation surveys can present an unnatural environment with respect to the typical shopping experience. Furthermore, there is the burden on the researcher to adequately specify all product attributes and describe them to the respondents in a clear and meaningful way. In the present unique setting, this second limitation is less concerning as the design and the way options are presented to customers are very similar to the typical shopping experience on the platform.

As this design informs respondents that the menu options are hypothetical and might be used in research, one might worry about experimenter demand effects. This concern is likely not significant as customers are shopping in the privacy of their own home to identify the best match (doctor) to whom they will be paying cash out of their own pocket. As discussed above, the individual choice models estimated with the responses to the hypothetical menus predicted the actual booking choice nicely.

7 Conclusion

Racial gaps in hiring, promotion, and pay persist in many industries even for high skilled professionals. This paper consider possible mechanisms that contributes to these gaps.

Preferences elicited from actual healthcare shoppers trying to book a doctor without a strong quality signal on doctor quality show that customers are willing to pay a significant premium to have their colonoscopy done by a white doctor rather than a Black or Asian doctor. Willingness-to-pay is lowered by 12.7% of the average colonoscopy price for Black doctors and by 8.7% for Asian doctors. But when quality information in form of "stars" is provided for each provider, the willingness-to-pay gap for non-white providers dropped to 1.2% for Black doctors and and 1.1% for Asian doctors.

Results cannot rule out the presences of behavioral mechanisms like deniable prejudice where prejudice is expressed only when there is an excuse, and biased belief discrimination where customers of healthcare use mis-specified statistical models of provider quality. While the data suggest the majority of discrimination could be not due to traditional taste-based discrimination, it is worth noting that we cannot interpret this as a lack of animus against minorities. Customers can exhibit deniable prejudice. On the other hand, customers may develop biased beliefs due to such animus. I can say that a large part of the discrimination vanished upon the provision of credible information on the relevant quality distributions.

Decomposing the nature of customer discrimination towards doctors and other service professionals has pivotal implications for policy. If the discrimination is due to customers forming biased statistical models about minority provider service quality or to customers using the lack of information as an excuse to express prejudice, then interventions that can revise and correct those biased beliefs while enriching the decision architecture with quality information to eliminate excuses for discrimination would have impact. Future studies of discrimination need to go beyond traditional taste-based and statistical discrimination and evaluate a broader set of candidate behavioral mechanisms to stay relevant for policy.

Lastly, the present novel preference elicitation methodology can broadly study individual preferences in a way economists find reliable, incentive compatible and cross-validated. Just as randomized controlled trials have come to be widely adopted by economists, similar preference elicitation surveys can become as much a part of the empirical economist arsenal as they are for marketing scholars. The novel design proposed can be a step in that direction.

References

AAMC, "Diversity in Medicine: Facts and Figures 2019," 2020.

Agan, Amanda and Sonja Starr, "Ban the box, criminal records, and racial discrimination: A field experiment," *The Quarterly Journal of Economics*, 2018, *133* (1), 191–235.

Aigner, Dennis J and Glen G Cain, "Statistical theories of discrimination in labor markets," *Ilr Review*, 1977, *30* (2), 175–187.

Allenby, G, "An introduction to hierarchical Bayesian modeling," in "Tutorial notes, Advanced Research Techniques Forum, American Marketing Association" 1997.

Allenby, Greg M and Peter E Rossi, "Hierarchical bayes models," *The handbook of marketing research: Uses, misuses, and future advances*, 2006, pp. 418–440.

_ , Nino Hardt, and Peter E Rossi, "Economic foundations of conjoint analysis," in "Handbook of the Economics of Marketing," Vol. 1, Elsevier, 2019, pp. 151–192.

Alsan, Marcella and Marianne Wanamaker, "Tuskegee and the health of black men," *The quarterly journal of economics*, 2018, *133* (1), 407–455.

_ and Sarah Eichmeyer, "Persuasion in Medicine: Messaging to Increase Vaccine Demand," Technical Report, National Bureau of Economic Research 2021.

_ , Owen Garrick, and Grant Graziani, "Does diversity matter for health? Experimental evidence from Oakland," *American Economic Review*, 2019, *109* (12), 4071–4111.

Arnold, David, Will Dobbie, and Crystal S Yang, "Racial bias in bail decisions," *The Quarterly Journal of Economics*, 2018, *133* (4), 1885–1932.

Arrow, Kenneth et al., "The theory of discrimination, discrimination in labor markets," *Achenfelter, A. Ress (eds.), Princeton–New Jersey*, 1973.

Arrow, Kenneth J., "Uncertainty and the Welfare Economics of Medical Care," *American Economic Review*, 1963, *53* (5), 941–973.

Ayres, Ian, Mahzarin Banaji, and Christine Jolls, "Race effects on eBay," *The RAND Journal of Economics*, 2015, *46* (4), 891–917.

Backstrom, Jesse and Richard T Woodward, "Racial Distance Matters Too: Home Bias and Racial Influence in Recreation Decision Making," *Available at SSRN 3818265*, 2021.

Ballance, Joshua, Robert Clifford, and Daniel Shoag, "“No more credit score”: Employer credit check bans and signal substitution," *Labour Economics*, 2020, *63*, 101769.

Bar, Revital and Asaf Zussman, "Customer discrimination: evidence from Israel," *Journal of Labor Economics*, 2017, *35* (4), 1031–1059.

Barr, Donald A, Maria Elena Gonzalez, and Stanley F Wanat, "The leaky pipeline: Factors associated with early decline in interest in premedical studies among underrepresented minority undergraduate students," *Academic Medicine*, 2008, *83* (5), 503–511.

Bartik, Alexander W and Scott T Nelson, "For Online Publication: Appendix for Deleting a Signal: Evidence from Pre-Employment Credit Checks," 2019.

Becker, Gary S et al., "The Economics of Discrimination," *University of Chicago Press Economics Books*, 1971.

Ben-Akiva, Moshe Emanuel, Daniel McFadden, Kenneth Train et al., *Foundations of stated preference elicitation: Consumer behavior and choice-based conjoint analysis*, Now, 2019.

Bénabou, Roland and Jean Tirole, "Incentives and prosocial behavior," *American economic review*, 2006, *96* (5), 1652–1678.

Bertrand, Marianne and Esther Duflo, "Field experiments on discrimination," *Handbook of economic field experiments*, 2017, *1*, 309–393.

Bohren, J Aislinn, Alex Imas, and Michael Rosenberg, "The dynamics of discrimination: Theory and evidence," *American economic review*, 2019, *109* (10), 3395–3436.

— , **Kareem Haggag, Alex Imas, and Devin G Pope**, "Inaccurate statistical discrimination," Technical Report, National Bureau of Economic Research 2019.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer, "Stereotypes," *The Quarterly Journal of Economics*, 2016, *131* (4), 1753–1794.

— , **Nicola Gennaioli, and Andrei Shleifer**, "Salience," *Annual Review of Economics*, 2022, *14*, 521–544.

Brown, Timothy, Jenny X Liu, and Richard M Scheffler, "Does the Under- or Overrepresentation of Minority Physicians across Geographical Areas Affect the Location Decisions of Minority Physicians?," *Health services research*, 2009, *44* (4), 1290–1308.

Burdekin, Richard CK and Todd L Idson, "Customer preferences, attendance and the racial structure of professional basketball teams," *Applied Economics*, 1991, *23* (1), 179–186.

Census-Bureau, U.S., "2018 Census," U.S. Department of Commerce May 2018.

Combes, Pierre-Philippe, Bruno Decreuse, Morgane Laouenan, and Alain Trannoy, "Customer discrimination and employment outcomes: theory and evidence from the french labor market," *Journal of Labor Economics*, 2016, *34* (1), 107–160.

Cooper, Lisa A, Debra L Roter, Rachel L Johnson, Daniel E Ford, Donald M Steinwachs, and Neil R Powe, "Patient-centered communication, ratings of care, and concordance of patient and physician race," *Annals of internal medicine*, 2003, *139* (11), 907–915.

Courey, Gabriel, "Gender and Racial Wage Differentials in Nonprofit Hospitals," *LABOUR*, 2020, *34* (4), 373–398.

Cui, Ruomeng, Jun Li, and Dennis J Zhang, "Reducing discrimination with reviews in the sharing economy: Evidence from field experiments on Airbnb," *Management Science*, 2020, *66* (3), 1071–1094.

Dana, Jason, Roberto A Weber, and Jason Xi Kuang, "Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness," *Economic Theory*, 2007, *33* (1), 67–80.

DeGroot, Morris H, *Optimal statistical decisions*, Vol. 82, John Wiley & Sons, 2005.

DellaVigna, Stefano, Devin Pope, and Eva Vivalt, "Predict science to improve science," *Science*, 2019, *366* (6464), 428–429.

Dietl, Helmut, Anil Özdemir, and Andrew Rendall, "The role of facial attractiveness in tennis TV-viewership," *Sport Management Review*, 2020, *23* (3), 521–535.

Doleac, Jennifer L and Luke CD Stein, "The visible hand: Race and online market outcomes," *The Economic Journal*, 2013, *123* (572), F469–F492.

Edelman, Benjamin, Michael Luca, and Dan Svirsky, "Racial discrimination in the sharing economy: Evidence from a field experiment," *American Economic Journal: Applied Economics*, 2017, *9* (2), 1–22.

Elías, Julio J, Nicola Lacetera, and Mario Macis, "Paying for kidneys? a randomized survey and choice experiment," *American Economic Review*, 2019, *109* (8), 2855–88.

Esponda, Ignacio, Ryan Oprea, and Sevgi Yuksel, "Discrimination Without Reason: Biases in Statistical Discrimination," 2022.

Exley, Christine L, "Excusing selfishness in charitable giving: The role of risk," *The Review of Economic Studies*, 2016, *83* (2), 587–628.

_ **and Judd B Kessler**, "Information Avoidance and Image Concerns," Technical Report, National Bureau of Economic Research 2021.

Fang, Hanming and Andrea Moro, "Theories of statistical discrimination and affirmative action: A survey," *Handbook of social economics*, 2011, *1*, 133–200.

Fershtman, Chaim and Uri Gneezy, "Discrimination in a segmented society: An experimental approach," *The Quarterly Journal of Economics*, 2001, *116* (1), 351–377.

Filut, Amarette, Madelyn Alvarez, and Molly Carnes, "Discrimination toward physicians of color: a systematic review," *Journal of the National Medical Association*, 2020, *112* (2), 117–140.

Fisman, Raymond and Michael Luca, "Fixing discrimination in online marketplaces," *Harvard business review*, 2016, *94* (12), 88–95.

Fong, Christina M and Erzo FP Luttmer, "What determines giving to Hurricane Katrina victims? Experimental evidence on racial group loyalty," *American Economic Journal: Applied Economics*, 2009, *1* (2), 64–87.

_ **and** _ , "Do fairness and race matter in generosity? Evidence from a nationally representative charity experiment," *Journal of Public Economics*, 2011, *95* (5-6), 372–394.

Fraiman, Joseph, Shannon Brownlee, Michael A Stoto, Kenneth W Lin, and Alison N Huffstetler, "An Estimate of the US Rate of Overuse of Screening Colonoscopy: a Systematic Review," *Journal of general internal medicine*, 2022, pp. 1–9.

Freeman, Brandi Kaye, Alden Landry, Robert Trevino, David Grande, and Judy A Shea, "Understanding the leaky pipeline: perceived barriers to pursuing a career in medicine or dentistry among underrepresented-in-medicine undergraduate students," *Academic Medicine*, 2016, *91* (7), 987–993.

Fung, Constance H, Marc N Elliott, Ron D Hays, Katherine L Kahn, David E Kanouse, Elizabeth A McGlynn, Mark D Spranca, and Paul G Shekelle, "Patients' preferences for technical versus interpersonal quality when selecting a primary care physician," *Health services research*, 2005, *40* (4), 957–977.

Gallen, Yana and Melanie Wasserman, "Does Information Affect Homophily?," 2022.

Garcia, Jorge A, Debora A Paterniti, Patrick S Romano, and Richard L Kravitz, "Patient preferences for physician characteristics in university-based primary care clinics," *Ethnicity and Disease*, 2003, *13* (2), 259–267.

Gennaioli, Nicola and Andrei Shleifer, "What comes to mind," *The Quarterly journal of economics*, 2010, *125* (4), 1399–1433.

Gneezy, Uri, John List, and Michael K Price, "Toward an understanding of why people discriminate: Evidence from a series of natural field experiments," Technical Report, National Bureau of Economic Research 2012.

Gray, Bradley and Jeffrey J Stoddard, "Patient-physician pairing: does racial and ethnic congruity influence selection of a regular physician?," *Journal of community health*, 1997, *22* (4), 247–259.

Grisham, Sarah, "Medscape physician compensation report 2017," *Medscape website. https://www. medscape. com/slideshow/compensation-2017-overview-6008547. Published April*, 2017, *5.*

Gunja, Munira Z and Sara R Collins, "Who are the remaining uninsured, and why do they lack coverage," *Commonwealth Fund. https://www. commonwealthfund. org/publications/issue-briefs/2019/aug/who-are-remaining-uninsured-and-why-do-they-lack-coverage*, 2019.

Harris, Katherine M, "How do patients choose physicians? Evidence from a national survey of enrollees in employment-related health plans," *Health services research*, 2003, *38* (2), 711–732.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in "Proceedings of the IEEE conference on computer vision and pattern recognition" 2016, pp. 770–778.

Hirpa, Meron, Tinsay Woreta, Hilena Addis, and Sosena Kebede, "What matters to patients? A timely question for value-based care," *Plos one*, 2020, *15* (7), e0227845.

Holzer, Harry J and Keith R Ihlanfeldt, "Customer discrimination and employment outcomes for minority workers," *The Quarterly Journal of Economics*, 1998, *113* (3), 835–867.

Idan, Edna, Anlu Xing, Javarcia Ivory, and Marcella Alsan, "Sociodemographic correlates of medical mistrust among African American men living in the East Bay," *Journal of health care for the poor and underserved*, 2020, *31* (1), 115–127.

Jr, Richard Carey McGregory, "An analysis of black-white wage differences in nursing: wage gap or wage premium?," *The Review of Black Political Economy*, 2013, *40* (1), 31–37.

Jr, Stanley S Bergen, "Underrepresented minorities in medicine," *JAMA*, 2000, *284* (9), 1138–1139.

Kahn, Lawrence M and Peter D Sherer, "Racial differences in professional basketball players' compensation," *Journal of Labor Economics*, 1988, *6* (1), 40–61.

Kessler, Judd B, Corinne Low, and Colin D Sullivan, "Incentivized resume rating: Eliciting employer preferences without deception," *American Economic Review*, 2019, *109* (11), 3713–44.

Kolstad, Jonathan T and Michael E Chernew, "Quality and consumer decision making in the market for health insurance and health care services," *Medical care research and review*, 2009, *66* (1_suppl), 28S–52S.

Laouénan, Morgane and Roland Rathelot, "Can information reduce ethnic discrimination? evidence from airbnb," *American Economic Journal: Applied Economics*, 2020.

Leonard, Jonathan S, David I Levine, and Laura Giuliano, "Customer discrimination," *The Review of Economics and Statistics*, 2010, *92* (3), 670–678.

Lett, Lanair Amaad, H Moses Murdock, Whitney U Orji, Jaya Aysola, and Ronnie Sebro, "Trends in racial/ethnic representation among US medical students," *JAMA network open*, 2019, *2* (9), e1910490–e1910490.

Li, Siyue and Austin Hubner, "The impact of web-based ratings on patient choice of a primary care physician versus a specialist: randomized controlled experiment," *Journal of medical Internet research*, 2019, *21* (6), e11188.

List, John A, "The nature and extent of discrimination in the marketplace: Evidence from the field," *The Quarterly Journal of Economics*, 2004, *119* (1), 49–89.

_ **and David Lucking-Reiley**, "The effects of seed money and refunds on charitable giving: Experimental evidence from a university capital campaign," *Journal of political Economy*, 2002, *110* (1), 215–233.

— , **Paramita Sinha, and Michael H Taylor**, "Using choice experiments to value non-market goods and services: evidence from field experiments," *Advances in economic analysis & policy*, 2006, *5* (2).

List, John and Uri Gneezy, *The why axis: Hidden motives and the undiscovered economics of everyday life*, Random House, 2014.

Louviere, Jordan J and David A Hensher, "Using discrete choice models with experimental design data to forecast consumer demand for a unique cultural event," *Journal of Consumer research*, 1983, *10* (3), 348–361.

Low, Corinne, "Pricing the biological clock: Reproductive capital on the us marriage market," *Manuscript, Wharton School of Business*, 2014, *17.*

Luft, Harold S, Deborah W Garnick, David H Mark, Deborah J Peltzman, Ciaran S Phibbs, Erik Lichtenberg, and Stephen J McPhee, "Does quality influence choice of hospital?," *Jama*, 1990, *263* (21), 2899–2906.

Ly, Dan P, Seth A Seabury, and Anupam B Jena, "Differences in incomes of physicians in the United States by race and sex," *BMJ: British Medical Journal*, 2016, *353.*

Macchi, Elisa, "Worth your weight: experimental evidence on the benefits of obesity in low-income countries," Technical Report, Working Paper 2022.

McFadden, Daniel, "Stated preference methods and their applicability to environmental use and non-use valuations," in "Contingent Valuation of Environmental Goods," Edward Elgar Publishing, 2017.

Merchant, Juanita L and M Bishr Omary, "Underrepresentation of underrepresented minorities in academic medicine: the need to enhance the pipeline and the pipe," 2010.

Mooney, Cathleen, Jack Zwanziger, Ciaran S Phibbs, and Susan Schmitt, "Is travel distance a barrier to veterans' use of VA hospitals for medical surgical care?," *Social science & medicine*, 2000, *50* (12), 1743–1755.

Nardinelli, Clark and Curtis Simon, "Customer racial discrimination in the market for memorabilia: The case of baseball," *The Quarterly Journal of Economics*, 1990, *105* (3), 575–595.

Neumark, David, Roy J Bank, and Kyle D Van Nort, "Sex discrimination in restaurant hiring: An audit study," *The Quarterly journal of economics*, 1996, *111* (3), 915–941.

Nicholson, Sean and Carol Propper, "Medical workforce," in "Handbook of health economics," Vol. 2, Elsevier, 2011, pp. 873–925.

Park, Alice, Charlie Smart, Rumsey Taylor, and Watkins Miles, "An extremely detailed map of the 2020 election," *The New York Times*, 2021.

Phelps, Edmund S, "The statistical theory of racism and sexism," *The american economic review*, 1972, *62* (4), 659–661.

Pope, Devin G and Justin R Sydnor, "What's in a Picture? Evidence of Discrimination from Prosper. com," *Journal of Human resources*, 2011, *46* (1), 53–92.

Raghavarao, Damaraju, James B Wiley, and Pallavi Chitturi, *Choice-based conjoint analysis: models and designs*, CRC Press, 2010.

Rao, Vithala R et al., *Applied conjoint analysis*, Springer, 2014.

Razzouk, Nabil, Victoria Seitz, and Janet Marie Webb, "What's important in choosing a primary care physician: an analysis of consumer response," *International Journal of Health Care Quality Assurance*, 2004.

Saha, Somnath, Miriam Komaromy, Thomas D Koepsell, and Andrew B Bindman, "Patient-physician racial concordance and the perceived quality and use of health care," *Archives of internal medicine*, 1999, *159* (9), 997–1004.

Santos, Rita, Hugh Gravelle, and Carol Propper, "Does quality affect patients' choice of doctor? Evidence from England," *The Economic Journal*, 2017, *127* (600), 445–494.

Sarsons, Heather, "Interpreting signals in the labor market: evidence from medical referrals," *Job Market Paper*, 2017.

Schmitt, Susan K, Ciaran S Phibbs, and John D Piette, "The influence of distance on utilization of outpatient mental health aftercare following inpatient substance abuse treatment," *Addictive Behaviors*, 2003, *28* (6), 1183–1192.

Stone, Eric W and Ronald S Warren, "Customer discrimination in professional basketball: evidence from the trading-card market," *Applied Economics*, 1999, *31* (6), 679–685.

Sullivan, Colin D, "Eliciting preferences over life and death: experimental evidence from organ transplantation," 2021.

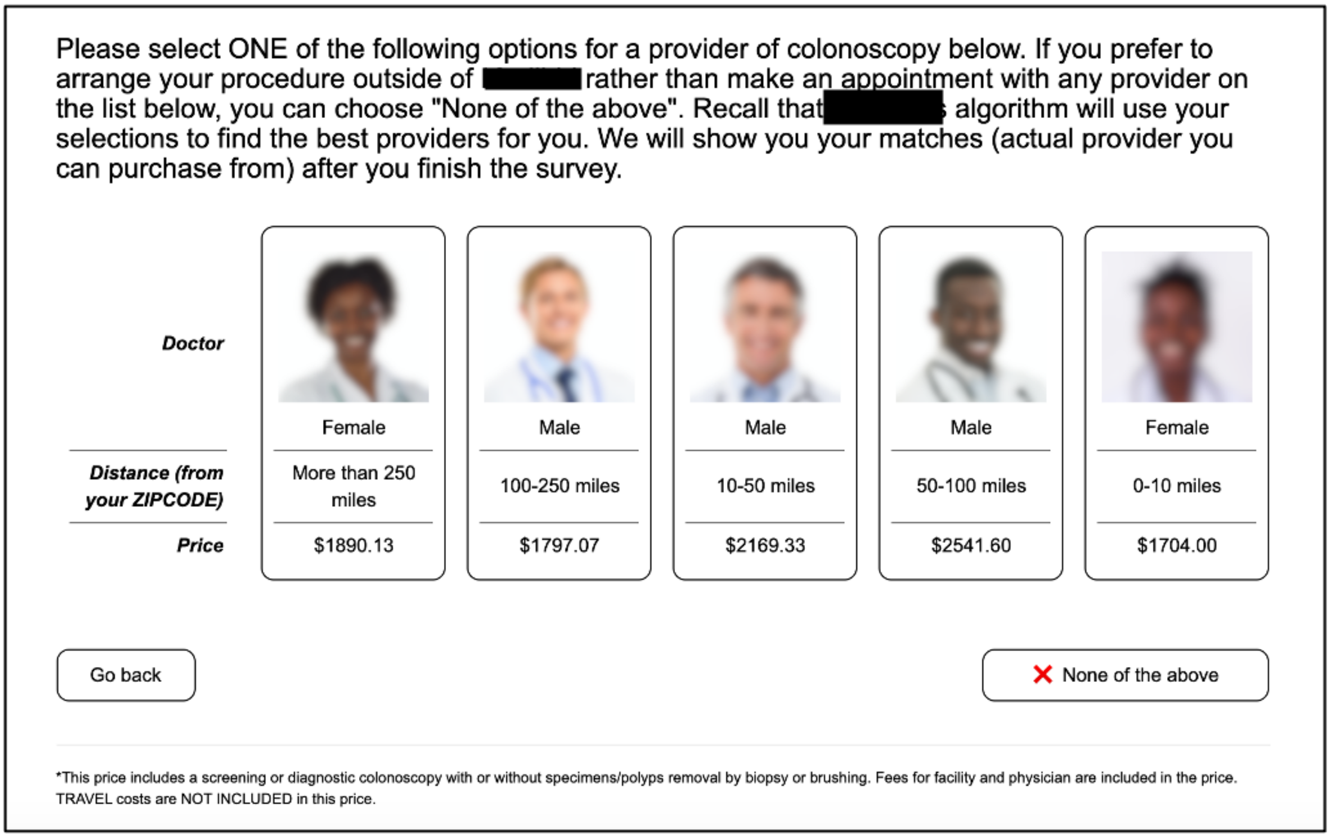
Train, Kenneth E, *Discrete choice methods with simulation*, Cambridge university press, 2009.

Tversky, Amos and Daniel Kahneman, "Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment.," *Psychological review*, 1983, *90* (4), 293.

Wozniak, Abigail, "Discrimination and the effects of drug testing on black employment," *Review of Economics and Statistics*, 2015, *97* (3), 548–566.

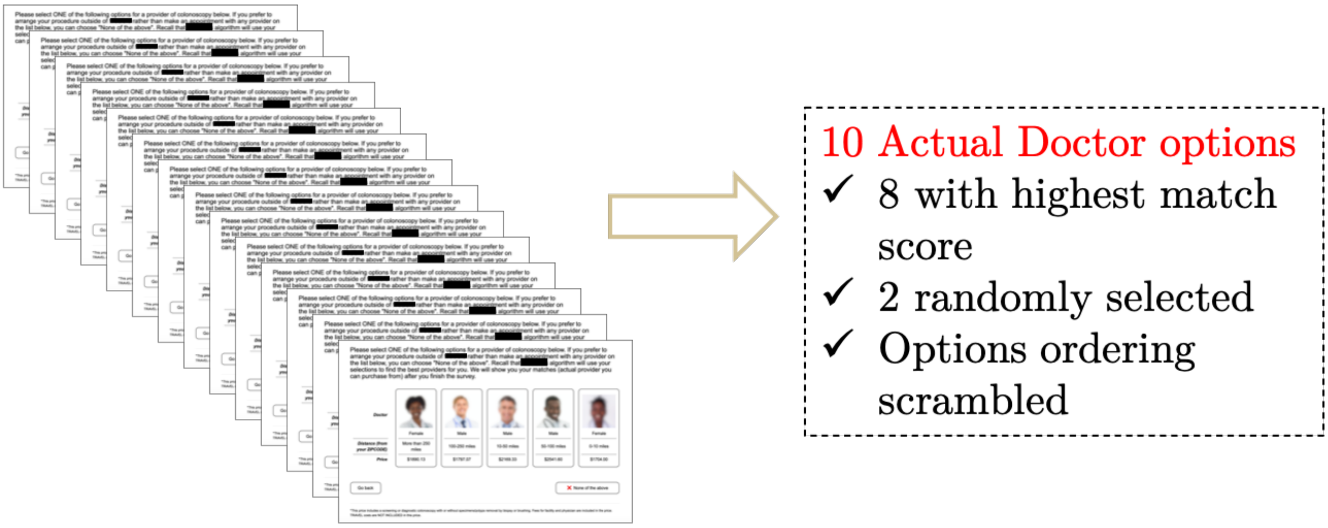
Yoon, Jean, Megan E Vanneman, Sharon K Dally, Amal N Trivedi, and Ciaran S Phibbs, "Veterans' reliance on VA care by type of service and distance to VA for nonelderly VA-Medicaid dual enrollees," *Medical care*, 2019, *57* (3), 225.

Figure 1: Single Example of Conjoint Menu of Doctors



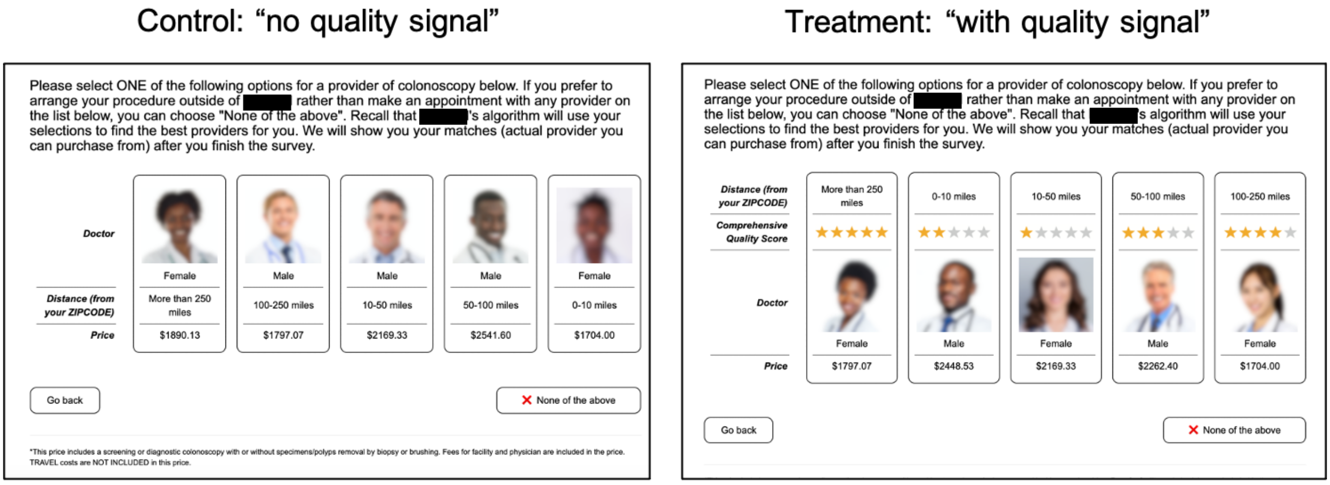
Notes: This figure shows an example of a menu of hypothetical doctor options, with 5 doctor options and a "None of the above" option, that form the basis of the choice experiments used in this paper. Each customer/respondent will encounter 14 menus like this one but with exogenously varied doctor features like price, profile picture, travel distance and sex of the doctor. The customer will have to click one of the options for each menu to proceed in the experiment. The name of the online platform that partnered with me on this project is Blacked out in this figure to maintain privacy for the platform per the Data Use Agreement.

Figure 2: Matching Conjoint Responses to Actual Doctor Options



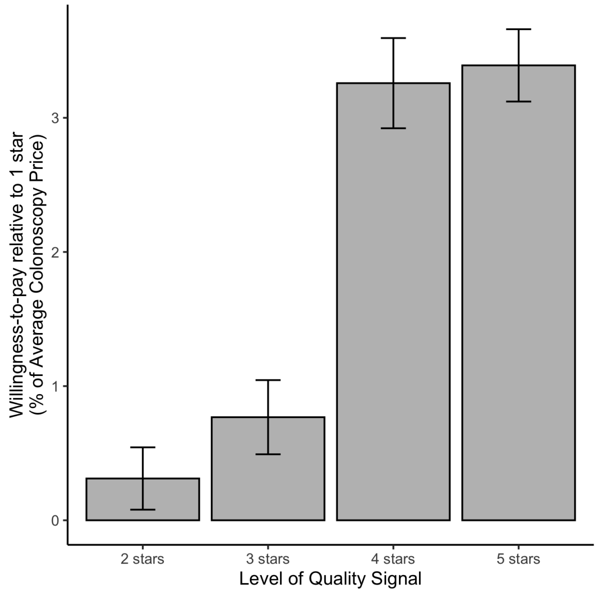
Notes: This figure illustrate the mechanism with which the conjoint survey with hypothetical doctor profiles is incentivized. Based on the customer's choice across the 14 menus of hypothetical doctor options, they will be matched to 10 actual doctors that they can purchase from. For each customer, 8 of the actual doctors will be the ones with the highest predicted choice probability based on the estimated preference parameters of that customer using their responses to the 14 menus while 2 actual doctors will be randomly drawn from the remaining doctors. On the menus in this figure, the name of the online platform that partnered with me on this project is Blacked out in this figure to maintain privacy for the platform per the Data Use Agreement.

Figure 3: Single Menu of Doctor Options for Under Control and Treatment Conditions



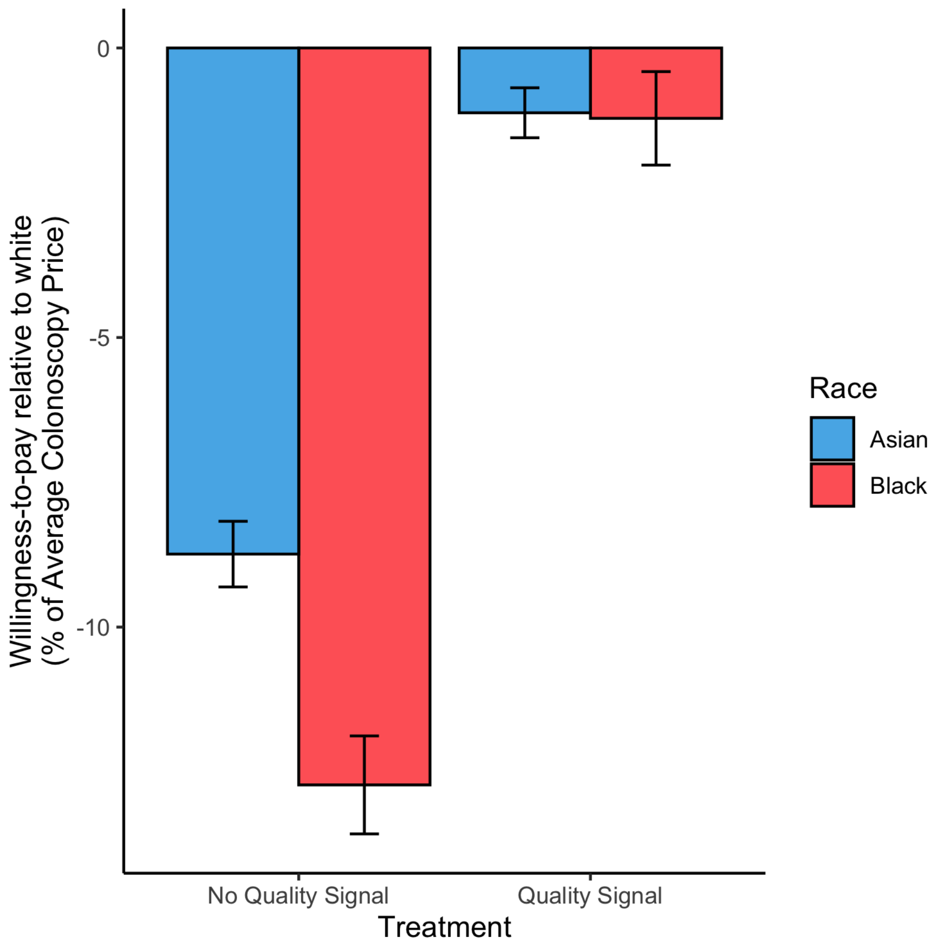
Notes: This figure shows the difference between the menus of hypothetical doctors that will be presented by each customer in the control group (left panel) and the treatment group (right panel). The treatment group choice menus feature an additional doctor attribute: the quality “stars” as shown on the right panel. Except for the price, the ordering from top to bottom of each doctor attribute (profile picture and sex, travel distance, and stars) is randomized across individuals. The name of the online platform that partnered with me on this project is Blacked out in this figure to maintain privacy for the platform per the Data Use Agreement.

Figure 4: Willingness-to-pay "Premium" by Number of Stars Relative to Doctors with 1 Star as Quality Signal



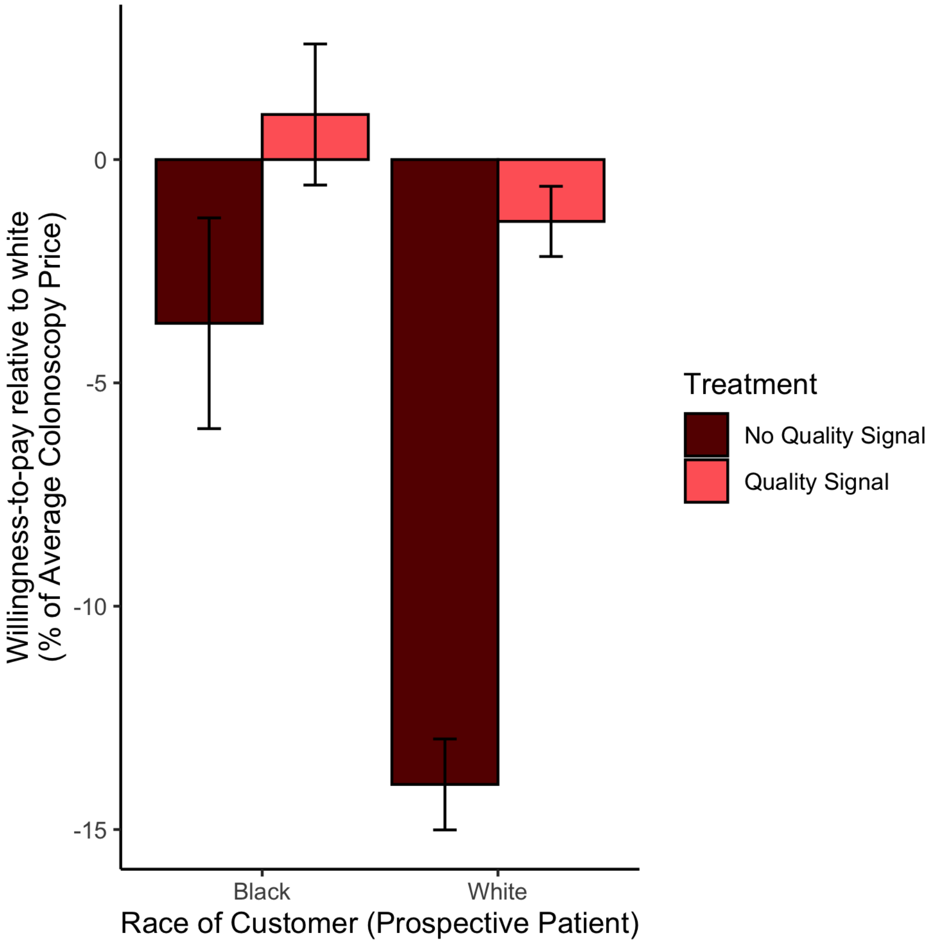
Notes: This figure presents the estimated parameter results for different star-levels (relative to 1 star) from the primary choice model from Section 2.5 expressed in terms of willingness-to-pay calculated as outlined in Section 2.5.1, using the treatment group data. It reports the results for the 104 customers in the "Quality Signal" treatment in the main sample ($N = 8,736$). The willingness-to-pay estimates are based on maximum simulated likelihood estimators from the multinomial logit model with mixed parameters, described in Section 2.5. Standard errors of the WTP estimates are calculated using the Delta method. The height of the bars represents willingness-to-pay coefficients as a percentage of the average price of a colonoscopy in my sample ($2122.80). The error bars show the 95% confidence intervals.

Figure 5: Willingness-to-pay "Penalty" by Race Relative to White Doctors



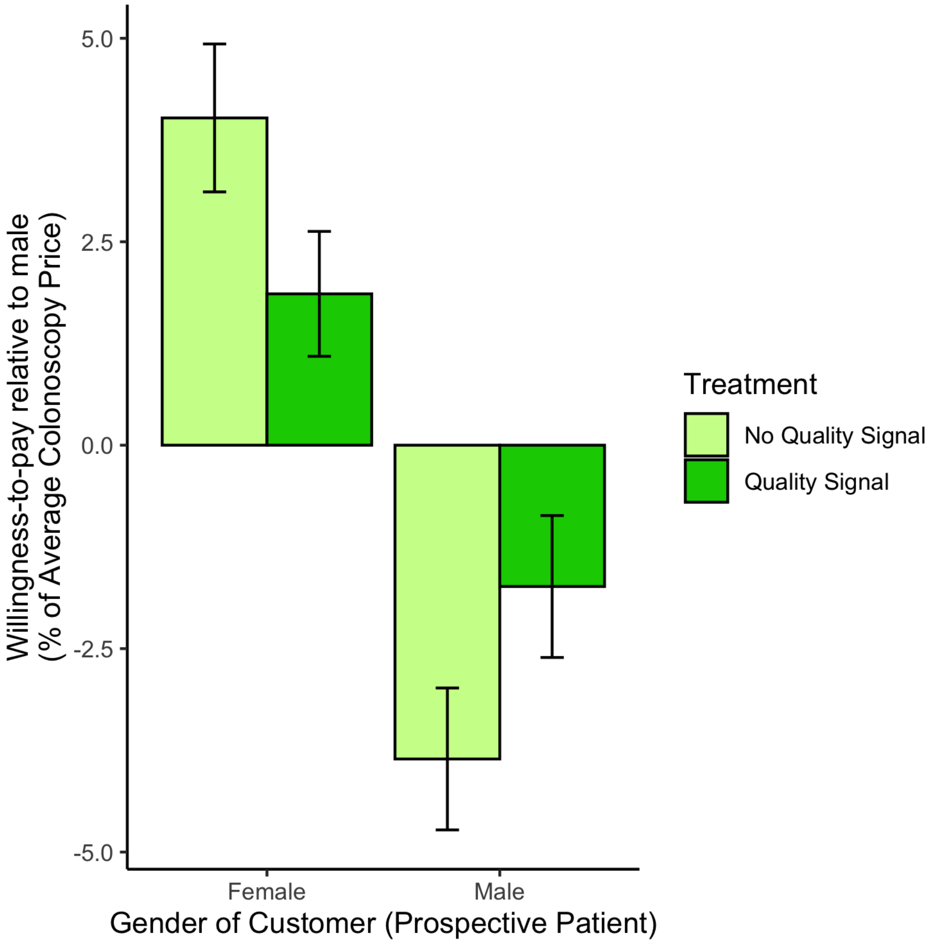
Notes: This figure presents the estimated parameter results for Black and Asian race (relative to white race) from the primary choice model from Section 2.5 expressed in terms of willingness-to-pay calculated as outlined in Section 2.5.1, using the control group and treatment groups. It reports the results for the 224 customers in the main sample ($N = 18,816$: 10,080 for the "No Quality Signal" bars and 8,736 for the "Quality Signal" bars), by treatment group. The willingness-to-pay estimates are based on maximum simulated likelihood estimators from the multinomial logit model with mixed parameters, described in Section 2.5. Standard errors of the WTP estimates are calculated using the Delta method. The height of the bars represents willingness-to-pay coefficients as a percentage of the average price of a colonoscopy in my sample (\$2122.80). The error bars show the 95% confidence intervals.

Figure 6: Comparison of Black and White Customers' Responses



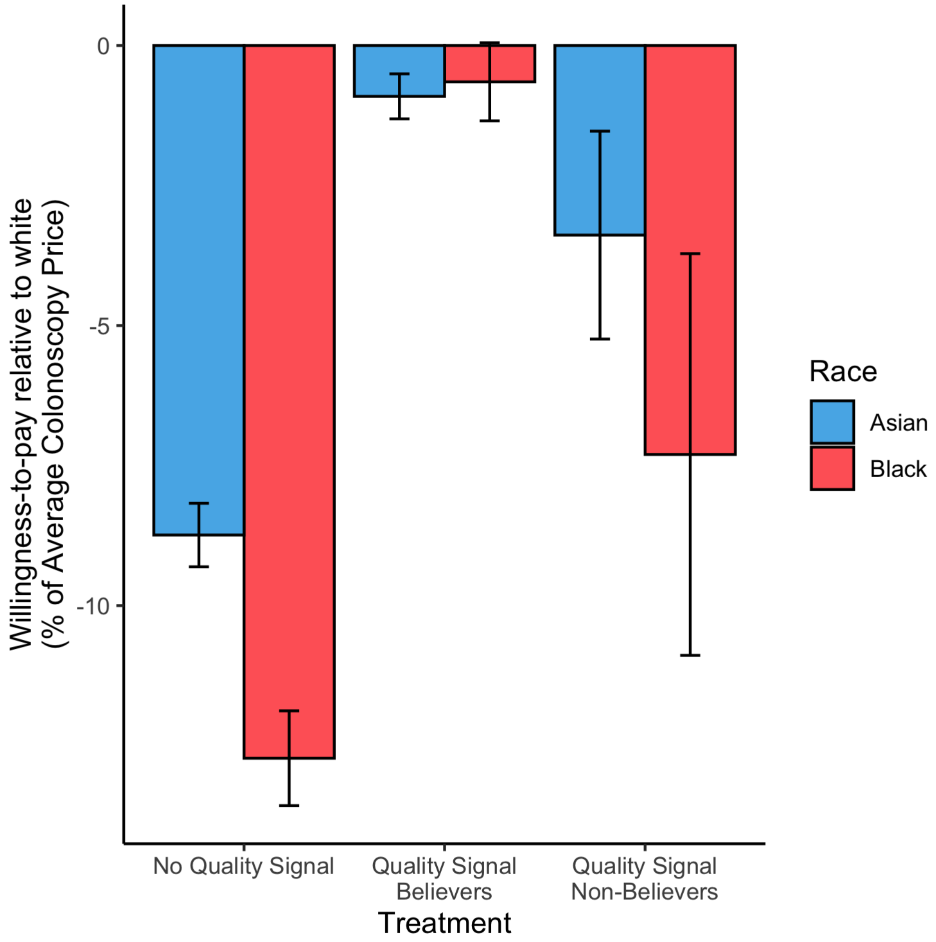
Notes: This figure presents the estimated parameter results for Black race (relative to white race) from the primary choice model from Section 2.5 expressed in terms of willingness-to-pay calculated as outlined in Section 2.5.1, using the control group and treatment groups segmented by the customer/respondent's race (Black or white). It reports the results for the 208 customers (185 white and 23 Black customers) in the main sample ($N = 17,472$: 1,176 for Black in "No Quality Signal" treatment, 756 for Black in "Quality Signal" treatment, 8,064 for white in "No Quality Signal" treatment, and 7,476 for white in "Quality Signal" treatment), by treatment group. The willingness-to-pay estimates are based on maximum simulated likelihood estimators from the multinomial logit model with mixed parameters, described in Section 2.5. Standard errors of the WTP estimates are calculated using the Delta method. The height of the bars represents willingness-to-pay coefficients as a percentage of the average price of a colonoscopy in my sample ($2122.80). The error bars show the 95% confidence intervals.

Figure 7: Comparison of Female Doctor Willingness-to-pay "Penalty" Relative to Male by Customer Gender



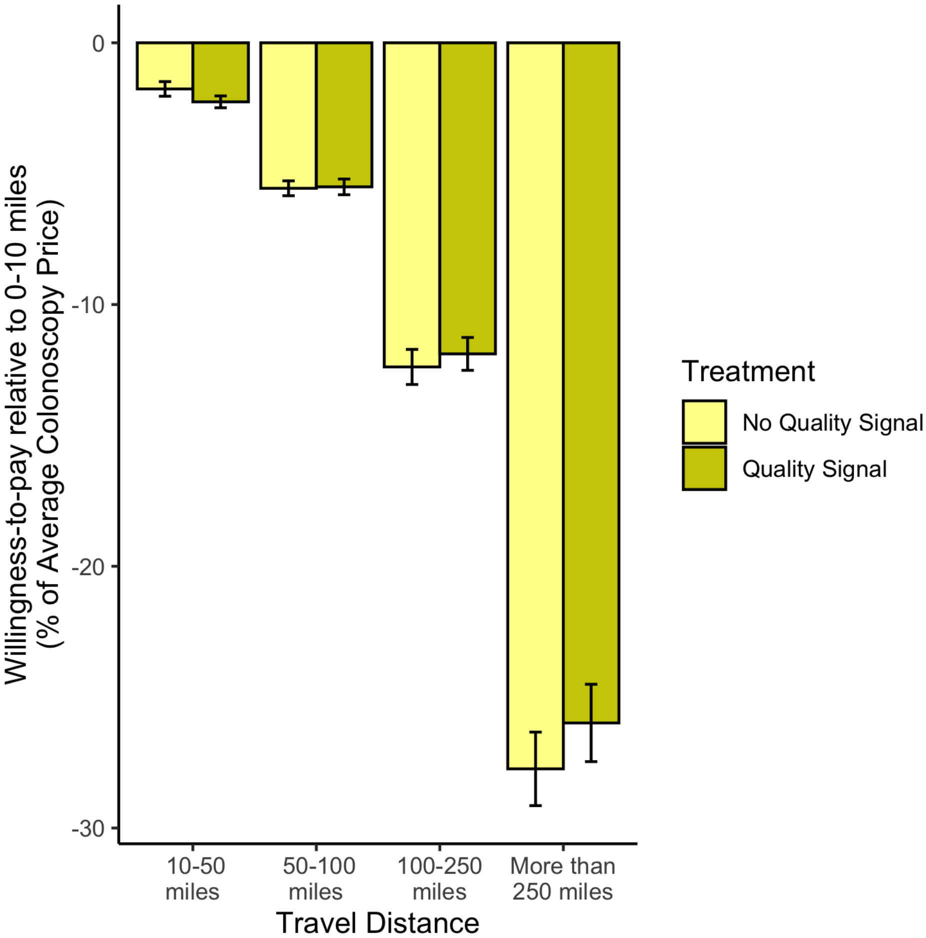
Notes: This figure presents the estimated parameter results for female gender (relative to male) from the primary choice model from Section 2.5 expressed in terms of willingness-to-pay calculated as outlined in Section 2.5.1, using the control group and treatment groups segmented by the customer/respondent's gender. It reports the results for the 224 customers (103 female and 121 male) in the main sample ($N = 18,816$: 4,368 for female in "No Quality Signal" treatment, 4,284 for female in "Quality Signal" treatment, 5,712 for male in "No Quality Signal" treatment, and 4,452 for male in "Quality Signal" treatment), by treatment group. The willingness-to-pay estimates are based on maximum simulated likelihood estimators from the multinomial logit model with mixed parameters, described in Section 2.5. Standard errors of the WTP estimates are calculated using the Delta method. The height of the bars represents willingness-to-pay coefficients as a percentage of the average price of a colonoscopy in my sample ($2122.80). The error bars show the 95% confidence intervals.

Figure 8: Comparison of "Believers" of the Quality Signals versus "Non-Believers"



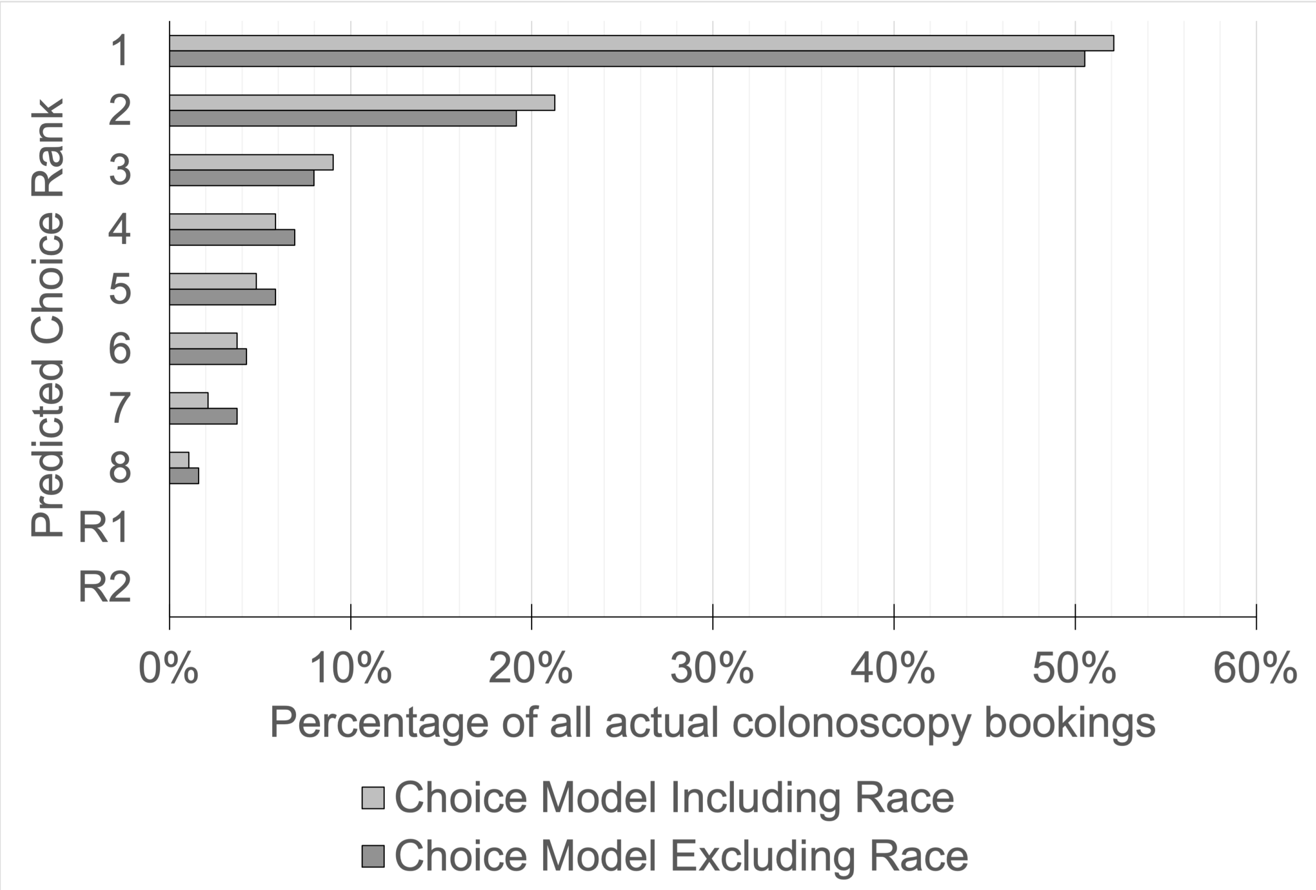
Notes: This figure presents the estimated parameter results for Black race (relative to white race) from the primary choice model from Section 2.5 expressed in terms of willingness-to-pay calculated as outlined in Section 2.5.1, using the control group, the "Believers" sample in the treatment group, and the "Non-Believers" sample in the treatment group. It reports the results for the 224 customers in the main sample ($N = 18,816$: 10,080 for the "No Quality Signal" bars, 7,812 for the "Quality Signal Believers" bars, and 924 for the "Quality Signal Non-Believers" bars), by treatment group. The willingness-to-pay estimates are based on maximum simulated likelihood estimators from the multinomial logit model with mixed parameters, described in Section 2.5. Standard errors of the WTP estimates are calculated using the Delta method. "Non-Believers" are defined as customers who responded that they "Strongly Disagree" or "Disagree" that "A provider with a higher Comprehensive Quality Score is a better provider than a provider with a lower Comprehensive Quality Score". The height of the bars represents willingness-to-pay coefficients as a percentage of the average price of a colonoscopy in my sample ($2122.80). The error bars show the 95% confidence intervals.

Figure 9: Comparison of Willingness-to-pay for Different Travel Distances With and Without Quality Signals



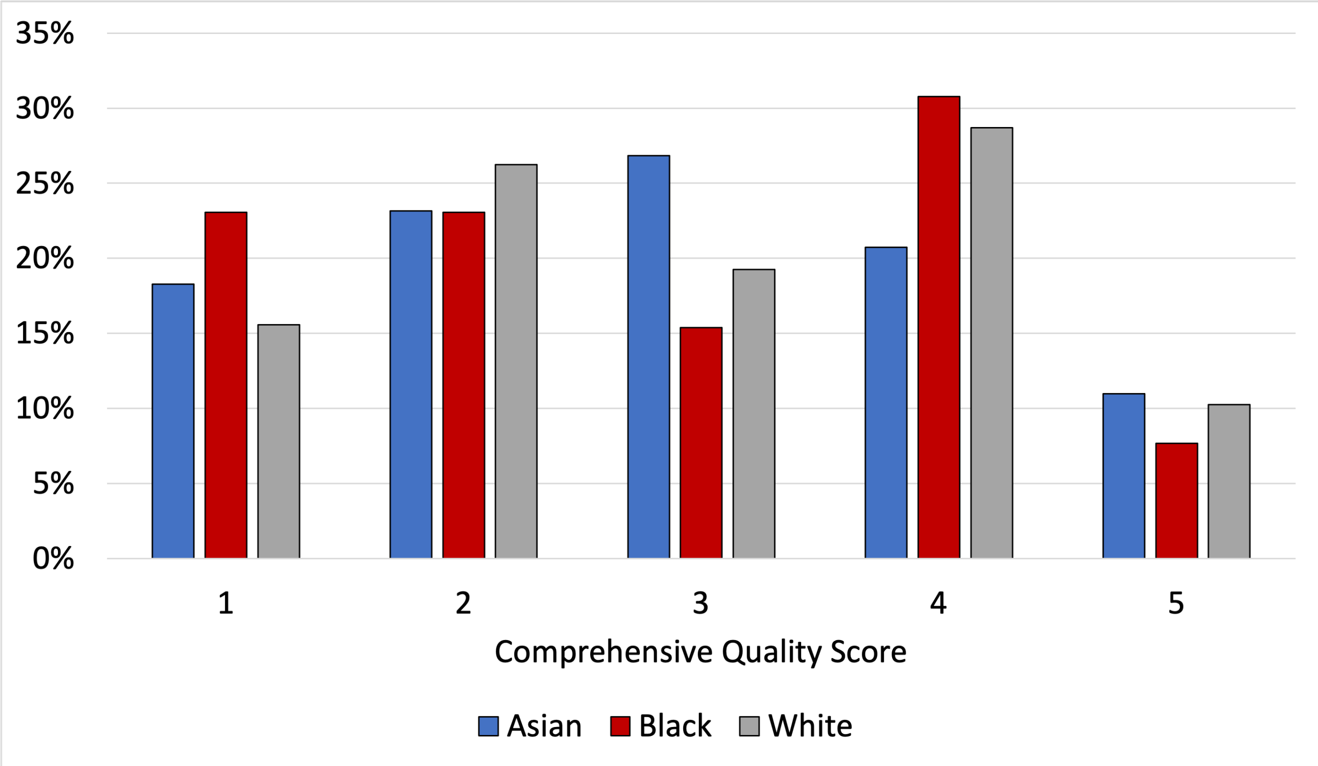
Notes: This figure presents the estimated parameter results for different travel distances (relative to "0-10 miles") from the primary choice model from Section 2.5 expressed in terms of willingness-to-pay calculated as outlined in Section 2.5.1, using the control and treatment group data. It reports the results for the 224 customers in the main sample ($N = 18,816$: 10,080 in the "No Quality Signal" arm and 8,736 in the "Quality Signal" arm), by treatment group. The willingness-to-pay estimates are based on maximum simulated likelihood estimators from the multinomial logit model with mixed parameters, described in Section 2.5. Standard errors of the WTP estimates are calculated using the Delta method. The height of the bars represents willingness-to-pay coefficients as a percentage of the average price of a colonoscopy in my sample ($2122.80). The error bars show the 95% confidence intervals.

Figure 10: Actual Booking Choice by Predicted Choice Probability Using Individual Coefficients from Hypothetical Choice Data



Notes: This figure presents the percentage of actual doctor booking by the rank (within the actual menu of 10 booking options) of doctor based on predicted choice probability individual coefficients estimated from choices made in the 14 hypothetical menus in the validated incentivized conjoint set-up. The lighter grey bar is based on the choice ranking predicted using the individual discrete choice model that includes all the doctor characteristics as predictive variables (price, gender, distance, and race for control group customers; price, gender, distance, quality, and race for treatment group customers); The lighter grey bar is based on the choice ranking predicted using the individual discrete choice model that includes all the doctor characteristics except race dummies as predictive variables (price, gender, and distance for control group customers; price, gender, distance, and quality for treatment group customers). Both models are estimated using the choice data from the 14 hypothetical menus, and the coefficients for each doctor characteristic is different between the models as the model specification is different. It reports the results for the 188 customers in the main sample who booked a doctor for colonoscopy. Individual coefficients are estimated with a mixed logit model using hierarchical Bayes estimation. The rank number 1 through 8 are the highest ranking doctors for each customer, R1 and R2 are two random doctors chosen from the remaining doctors in the pool and offered alongside the 8 to form the actual menu of 10 doctors.

Figure 11: Quality Distribution for Doctors by Race



Notes: This figure presents the empirical quality distribution of the actual doctors included in this study. There are totally 339 doctors offering colonoscopies, including 82 Asian doctors, 13 Black doctors, and 244 white doctors. The height of each bar represents the proportion of doctors at each level of comprehensive quality score, $P(q|Asian)$, $P(q|Black)$ or $P(q|White)$.

Table 1: Summary Statistics and Balance Checks

	Total		No Signal		Quality Signal		
	mean	s.d.	mean	s.d.	mean	s.d.	Difference p-value
Female	0.46	0.50	0.43	0.50	0.49	0.50	0.395
Black	0.10	0.30	0.12	0.32	0.09	0.28	0.461
Asian	0.06	0.23	0.06	0.24	0.06	0.23	0.984
Hispanic	0.10	0.30	0.08	0.28	0.12	0.32	0.424
White	0.83	0.38	0.80	0.40	0.86	0.35	0.274
Age under 45	0.07	0.25	0.05	0.22	0.09	0.28	0.277
Age between 45 and 54	0.45	0.50	0.47	0.50	0.42	0.50	0.515
Age between 55 and 64	0.46	0.50	0.45	0.50	0.48	0.50	0.647
High School Graduate	0.80	0.40	0.80	0.40	0.81	0.40	0.886
College Graduate	0.29	0.46	0.29	0.46	0.30	0.46	0.917
Married	0.50	0.50	0.50	0.50	0.51	0.50	0.886
Currently employed	0.54	0.50	0.53	0.50	0.57	0.50	0.528
Currenly Self-employed	0.20	0.40	0.17	0.37	0.24	0.43	0.171
Uninsured	0.90	0.30	0.89	0.31	0.91	0.28	0.587
# of outside options evaluated	4.42	3.17	4.19	2.85	4.68	3.48	0.249
Price of outside option	2535	660	2563	560	2502	758	0.489
Colonoscopy in past 10 years	0.79	0.41	0.80	0.40	0.77	0.42	0.578
Booked one of matched options	0.84	0.37	0.82	0.39	0.87	0.34	0.324
Patients	224		120		104		
N	18,816		10,080		8,736		

Notes: This table reports the background characteristics of the 224 customers in the main sample, pooled and by treatment group. "Female" indicates the share of female sex; "Black," "Asian," "Hispanic," and "White" indicate the shares of customers belonging to each of these categories. Age data was recorded in intervals, "Age under 45," "Age between 45 and 54," and "Age between 55 and 64" indicate shares of customers in these age buckets. "High School Graduate" indicate the share of customers who graduated from high school, "College Graduate" indicate indicate the share of customers who reported that they have a bachelor's or associate degree. "Married" indicate the share of customers who are currently married (not single, divored, separated, or widowed). "Currently employed" indicate the share who are employed part-time or full-time (not including self-employment while "Currently Self-employed" indicate the share who are self-employed. "Uninsured" indicates the share of customers who selected "NOT COVERED by any health insurance plan" in the survey. "# of outside options evaluated" indicates the number of colonoscopy providers that the customer has explored outside of the pilot's tool, while "Price of outside option" is the self-reported price that the customer expects to pay out-of-pocket (in USD). "Colonoscopy in past 10 years" indicates the share of customers who reported that they have had a colonoscopy in the past 10 years. "Colonoscopy in past 10 years" indicates the share of customers who reported that they have had a colonoscopy in the past 10 years. "Booked one of matched options" indicates the share who booked one of the 10 real doctor options matched to them through the experiment. Table shows averages ("mean") and standard deviations ("s.d."). The Difference p-value column reports the p-value for the test of equality between the treatment and control groups. Stars indicate whether this difference is significant. I do not find any statistically significant differences for all these differences.

Table 2: Choice Model Parameter Estimates in Willingness-to-Pay

Attribute	Control	Treatment						
		All star levels	4/5 stars ...	+ 2/3 stars	... + race interactions			
					Black and Asian		Non-white	
a	WTP_a^C	WTP_a^T	$WTP_a^{T,HQ}$	$WTP_a^{T,HMQ}$	$WTP_a^{T,HQ\cdot race,V1}$	$WTP_a^{T,HMQ\cdot race,V1}$	$WTP_a^{T,HQ\cdot race,V2}$	$WTP_a^{T,HMQ\cdot race,V2}$
Black	-270.133	-25.803	-25.700	-29.353	-43.519	-75.885	-27.516	-64.458
	(9.164)	(8.739)	(8.657)	(8.739)	(10.635)	(12.729)	(10.141)	(11.931)
Asian	-185.529	-23.757	-24.260	-24.265	-21.243	-42.012	-26.020	-36.562
	(6.139)	(4.672)	(4.610)	(4.622)	(6.107)	(7.023)	(5.770)	(6.344)
Female	-15.598	0.529	0.538	0.486	0.503	0.459	0.473	0.464
	(7.227)	(6.413)	(6.355)	(6.377)	(6.341)	(6.456)	(6.362)	(6.399)
10-50 miles	-37.409	-47.888	-47.432	-46.570	-47.432	-42.616	-48.344	-40.859
	(3.028)	(2.454)	(2.399)	(2.431)	(2.423)	(2.523)	(2.415)	(2.516)
50-100 miles	-117.994	-116.837	-103.776	-115.838	-102.165	-114.840	-103.060	-114.440
	(2.879)	(3.255)	(3.183)	(3.235)	(3.204)	(3.347)	(3.204)	(3.347)
100-250 miles	-262.882	-252.300	-249.423	-248.826	-252.079	-251.866	-251.636	-249.694
	(6.587)	(6.798)	(6.632)	(6.736)	(6.715)	(6.963)	(6.694)	(6.942)
More than 250 miles	-588.866	-551.620	-540.375	-542.832	-544.324	-542.093	-545.323	-542.282
	(13.133)	(16.004)	(13.649)	(15.252)	(14.366)	(15.216)	(14.366)	(15.539)
5 stars		71.981						
		(2.920)						
4 stars		69.167						
		(3.641)						
3 stars		16.312						
		(2.994)						
2 stars		6.612						
		(2.516)						
High Quality (4 or 5 stars)			55.924	70.898	51.004	61.518	51.168	60.607
			(2.033)	(2.998)	(3.636)	(5.067)	(3.636)	(5.472)
High Quality × Black					20.616	38.775		
					(5.355)	(7.403)		
High Quality × Asian					-7.679	-9.913		
					(5.388)	(7.831)		
High Quality × Non-white							6.543	12.895
							(4.661)	(6.482)
Medium Quality (2 or 3 stars)				21.740		23.009		23.018
				(2.942)		(5.100)		(4.947)
Medium Quality × Black						7.360		
						(7.864)		
Medium Quality × Asian						-11.104		
						(7.732)		
Medium Quality × Non-white								4.380
								(6.581)
N	10,080	8,736	8,736	8,736	8,736	8,736	8,736	8,736

Notes: This table reports the results for the 224 customers in the main sample, by treatment group. This table presents the estimated parameter results for the primary choice model from Section 2.5 expressed in terms of willingness-to-pay calculated as outlined in Section 2.5.1. All willingness-to-pay coefficients are in dollar units with standard errors for parameters given in parentheses. Marginal utility is assumed to be a linear function of doctor attributes and the parameters represents the willingness-to-pay relative to the reference attribute level. Standard errors of the WTP estimates are calculated using the Delta method. For race, the reference level is white race for the doctor. For gender, the reference level is male. For distance, the reference level is "0-10 miles" distance between the doctor's zip code and the customer's zip code. For quality stars, 1 star is the reference level for column 2. For quality stars, less than 4 stars is the reference level for columns 3-5 and less than 2 stars is the reference level for columns 6-8. The first column labelled "Control" displays the results for the control group. The second column labelled "Treatment - All star levels" displays the main results for the treatment group (shown also in Figure 4). The third column displays results for the treatment group with "High Quality (4 or 5 stars)" as the only variable for quality. The fourth column is similar to the second except that 2 additional variables: interaction terms between Black and high quality and between Asian and high quality are added. The fifth column is similar to the third except that there is only one interaction term between high quality and non-white race. The sixth column displays results for the treatment group with "Medium Quality (2 or 3 stars)" and "High Quality (4 or 5 stars)" as the variables for quality. The seventh column is similar to the third except that 2 additional variables: interaction terms between Black and medium quality and between Asian and medium quality are added. The eighth column is similar to the fourth except that there is one additional variable: interaction term between medium quality and non-white race.

Table 3: The Quality Distributions for Black and White Doctors

Quality Score	1 star	2 stars	3 stars	4 stars	5 stars
Black Doctors	23.08%	23.08%	15.38%	30.77%	7.69%
White Doctors	15.57%	26.23%	19.26%	28.69%	10.25%
Representativeness for Black vs. White $\frac{P(q\|Black)}{P(q\|White)}$	1.48	0.88	0.80	1.07	0.75

Notes: This table presents the empirical quality distribution of the actual doctors included in this study. There are totally 339 doctors offering colonoscopies, including 82 Asian doctors, 13 Black doctors, and 244 white doctors. The first two rows of this table presents the proportion of doctors at each level of comprehensive quality score, $P(q|Black)$. The last row presents representativeness of each quality level for Black race given comparison group white race, which is defined as the likelihood ratio $\frac{P(q|Black)}{P(q|White)}$ (following Bordalo et al. (2016) and Gennaioli and Shleifer (2010)).

A Appendix

A.1 Experimental Details

In this Section, I describe the design of my experiment in detail, including recruitment materials (A.2), and the VIC survey construction (A.3).

A.2 Recruitment Materials

The partnering platform sent emails to its subscription base to offer an opportunity to use a pilot tool to shop for one of three medical procedures: colonoscopy, MRI or knee replacement. The pilot tool allows each customer a one-time opportunity to get provider options with prices without having to pay the usual fee, rather than a replacement for the usual channel. The recruitment email for customers, shown in Figure 12, was sent to the subscribed customer base to recruit customers who want to "Shop for Your Next Provider for Colonoscopy, MRI or Knee Replacement" and that "the use of this pilot tool will be a one-time-only offer for each customer....[and] we'll waive the $25 fee for the appointment you book through the pilot tool." This email was sent out twice in the beginning of 2021 spaced 7 weeks apart. A Tweet was also posted weekly from the week of November 30 2020 to the week of March 29 2021 (Figure 13).

[redacted] Pilot: Shop for Your Next Provider for Colonoscopy, MRI or Knee Replacement

[redacted]@[redacted]
Thu 1/14/2021 4:08 PM
To: [redacted]

Dear Customer,

[redacted] is piloting [redacted]–Choice to help you find the best doctor for your needs. We will survey your preferences in this tool and identify providers that you can purchase your procedure from. With this pilot tool, we search the web for hundreds of providers across [redacted], [redacted], and more, giving you more options than any individual website to choose from. Are you shopping for a colonoscopy, MRI or knee replacement? If so, you can use this pilot tool. Currently, the use of this pilot tool will be a one-time-only offer for each customer.

After you complete a few questions that will take about 15 minutes, [redacted] will then find the best providers across all major websites for you based on your responses, using [redacted]'s newly developed matching algorithm. We will email you with 10 real providers within 48 hours after you finish. The more carefully you complete these questions, the better we will be able to match you with real provider options.

If you participate in this pilot, we'll waive the $25 fee for the appointment you book through the pilot tool. When you are ready to book a provider for the procedure you are looking for, please take the tool by clicking the link here:

https:/[redacted].com/[redacted]

The pilot tool will be available until February 28, 2021.

Happy Shopping!

[redacted]

Figure 12: Recruitment email sent to partner platform's customer base

Figure 13: Recruitment "tweet" sent from partner platform's Twitter account

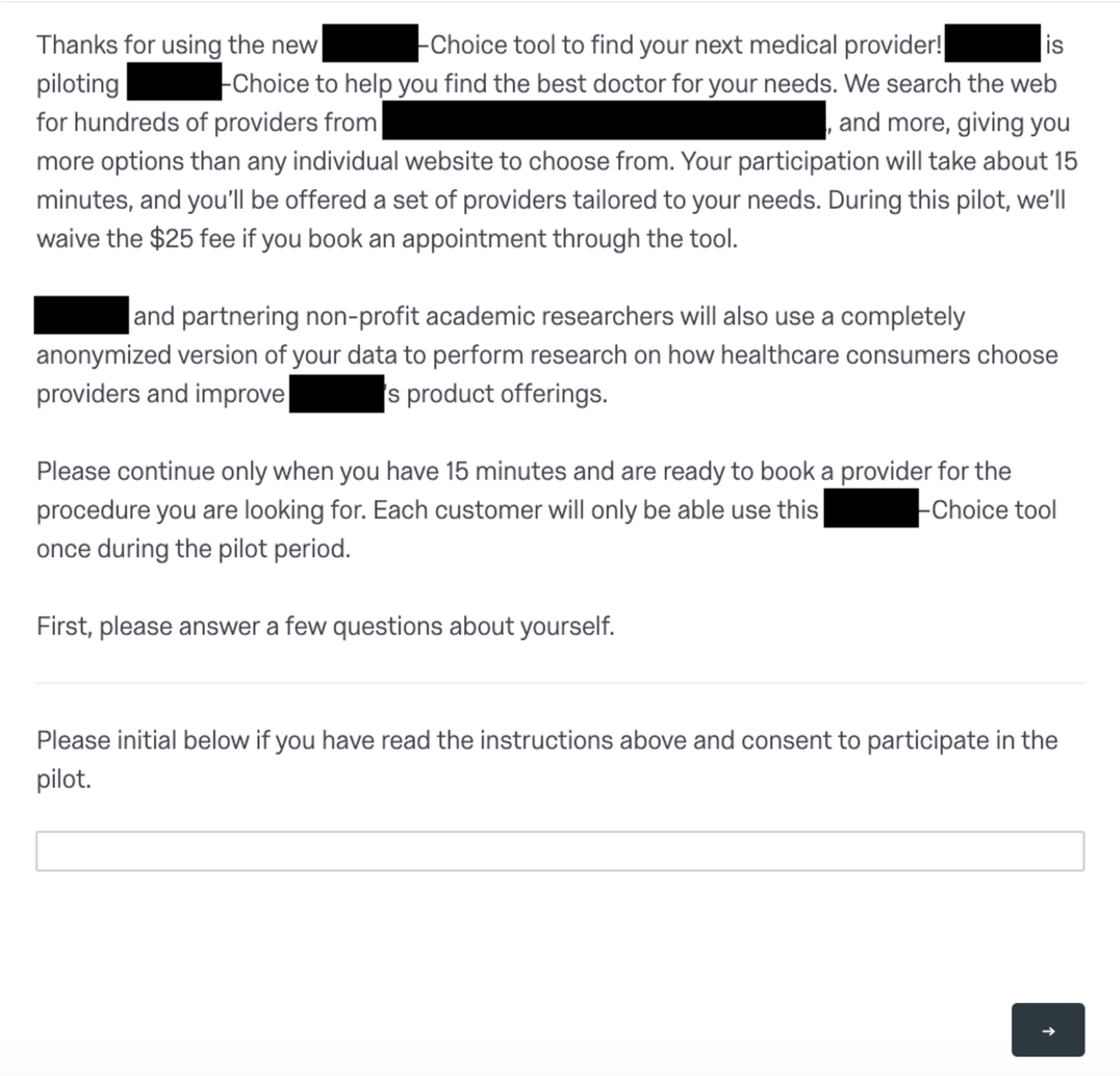
A.3 Experiment Design

Here, I outline the experiment design including the entry and exit surveys as well as the conjoint instruments. The entry survey provides information on the "pilot" shopping tool, and gather information from the customers on things like the procedure that is being shopped for, insurance status, and outside options. After the entry survey, the customers who will be redirected to the conjoint survey and randomized (via Bernoulli draws as subjects arrive) into either the treatment or control group. After the completion of the conjoint survey, the customer will complete an exit survey where demographic information is gathered. Questions about the customers' own demographic information like race and gender are deliberately delayed until after the customer have evaluated all the doctor profiles in the conjoint survey.

This appendix Section should serve to provide additional details about the actual experience of subjects of this experiment as they shop via the "pilot" tool. I will share the screens broken down into the following four subsections.

A.3.1 Entry Survey

I constructed the survey tool using Qualtrics software for customers who clicked the URL to participate in the "pilot" to shop for their provider to access from a web browser. Upon opening the survey link, respondents must initial on the instructions page (see Figure 14) to continue.



Thanks for using the new [redacted]-Choice tool to find your next medical provider! [redacted] is piloting [redacted]-Choice to help you find the best doctor for your needs. We search the web for hundreds of providers from [redacted], and more, giving you more options than any individual website to choose from. Your participation will take about 15 minutes, and you'll be offered a set of providers tailored to your needs. During this pilot, we'll waive the $25 fee if you book an appointment through the tool.

[redacted] and partnering non-profit academic researchers will also use a completely anonymized version of your data to perform research on how healthcare consumers choose providers and improve [redacted]'s product offerings.

Please continue only when you have 15 minutes and are ready to book a provider for the procedure you are looking for. Each customer will only be able use this [redacted]-Choice tool once during the pilot period.

First, please answer a few questions about yourself.

Please initial below if you have read the instructions above and consent to participate in the pilot.

→

Figure 14: Instruction page

Then, the subjects are asked to answer a series of questions listed below before they are redirected to the conjoint survey itself. These question gather information about the customer on (in the order of appearance): source of information about the pilot (Figure 20), location (Figure 16), age (Figure 17), education (Figure 18), marital status (Figure 19), employment status (Figure 20), insurance status (Figure 21), method of payment (Figure 22), and what procedure is being shopped for (Figure 23).

If the customer selected the option to indicate that they are shopping for a procedure other than colonoscopy, MRI, or knee replacement, they will be re-directed to a screen to indicate that they are not eligible for the "pilot" and then redirected to the regular shopping landing page on the online platform's website.

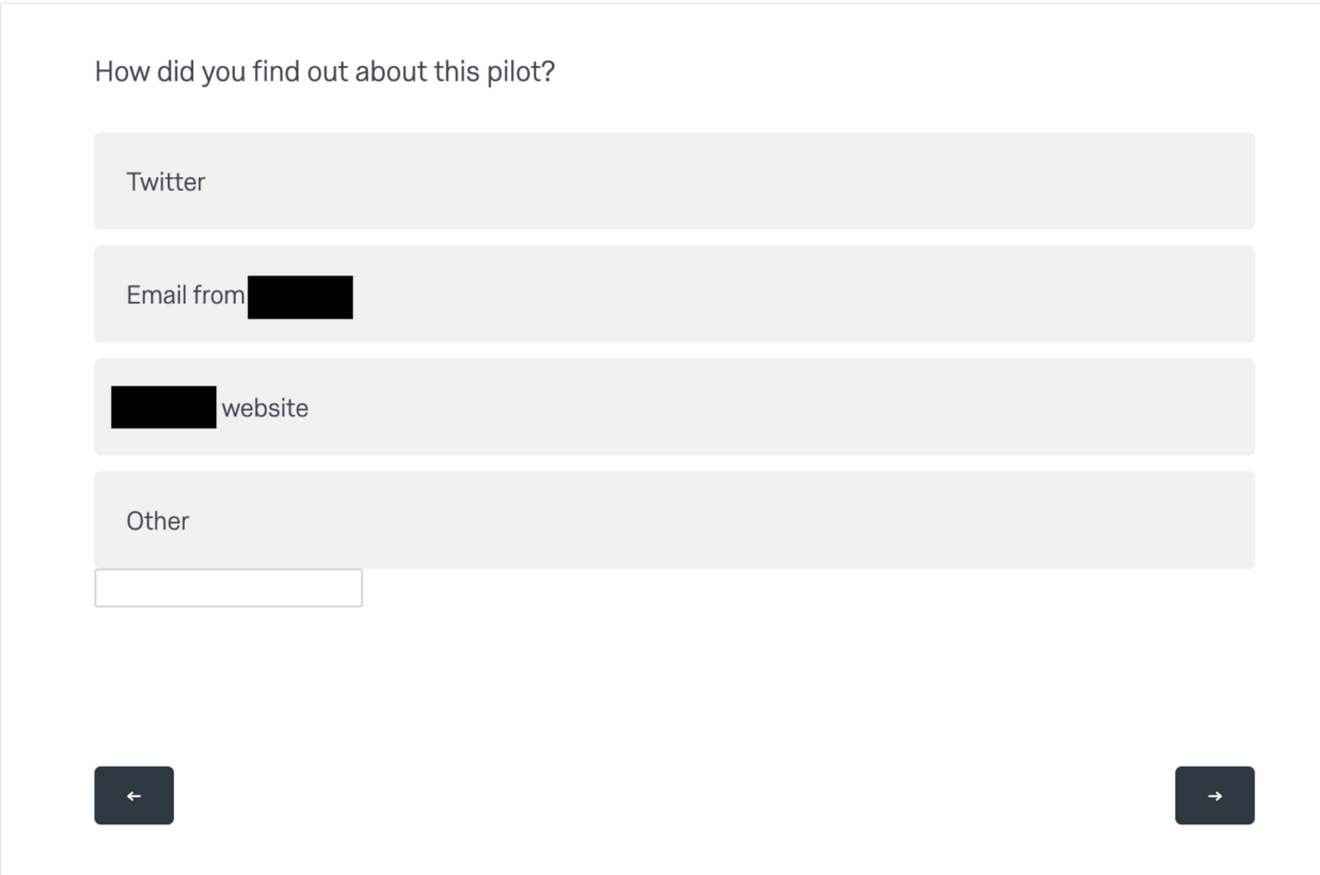


Figure 15: Entry survey question about how the subject heard about the "pilot"

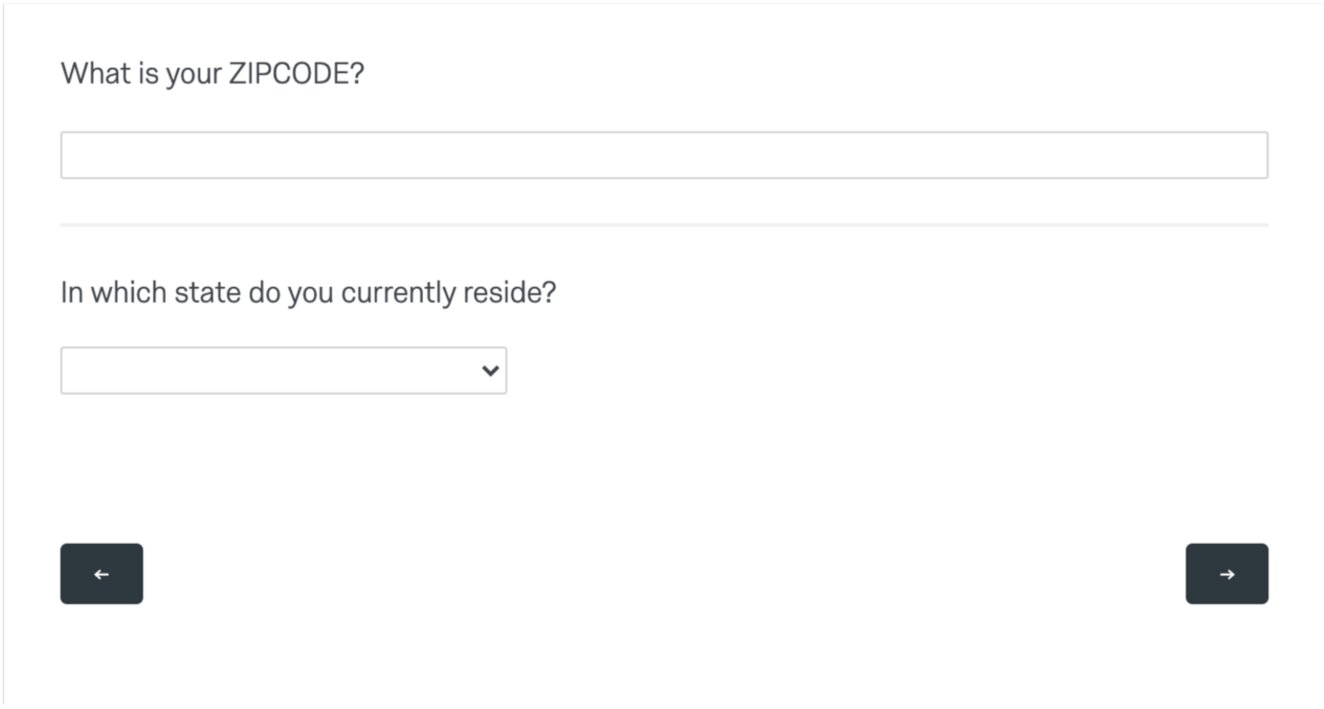


Figure 16: Entry survey question about how the subject's address

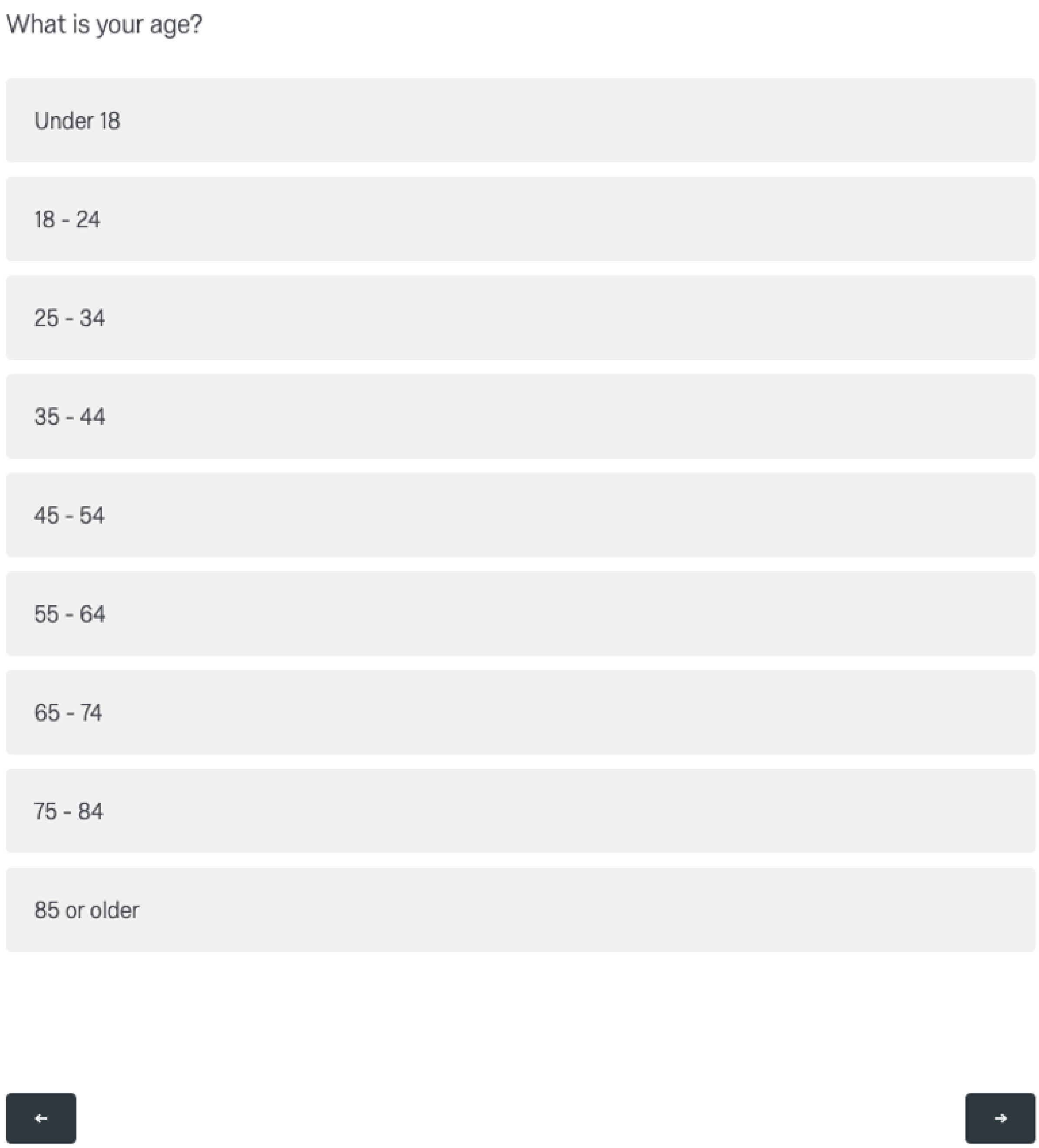
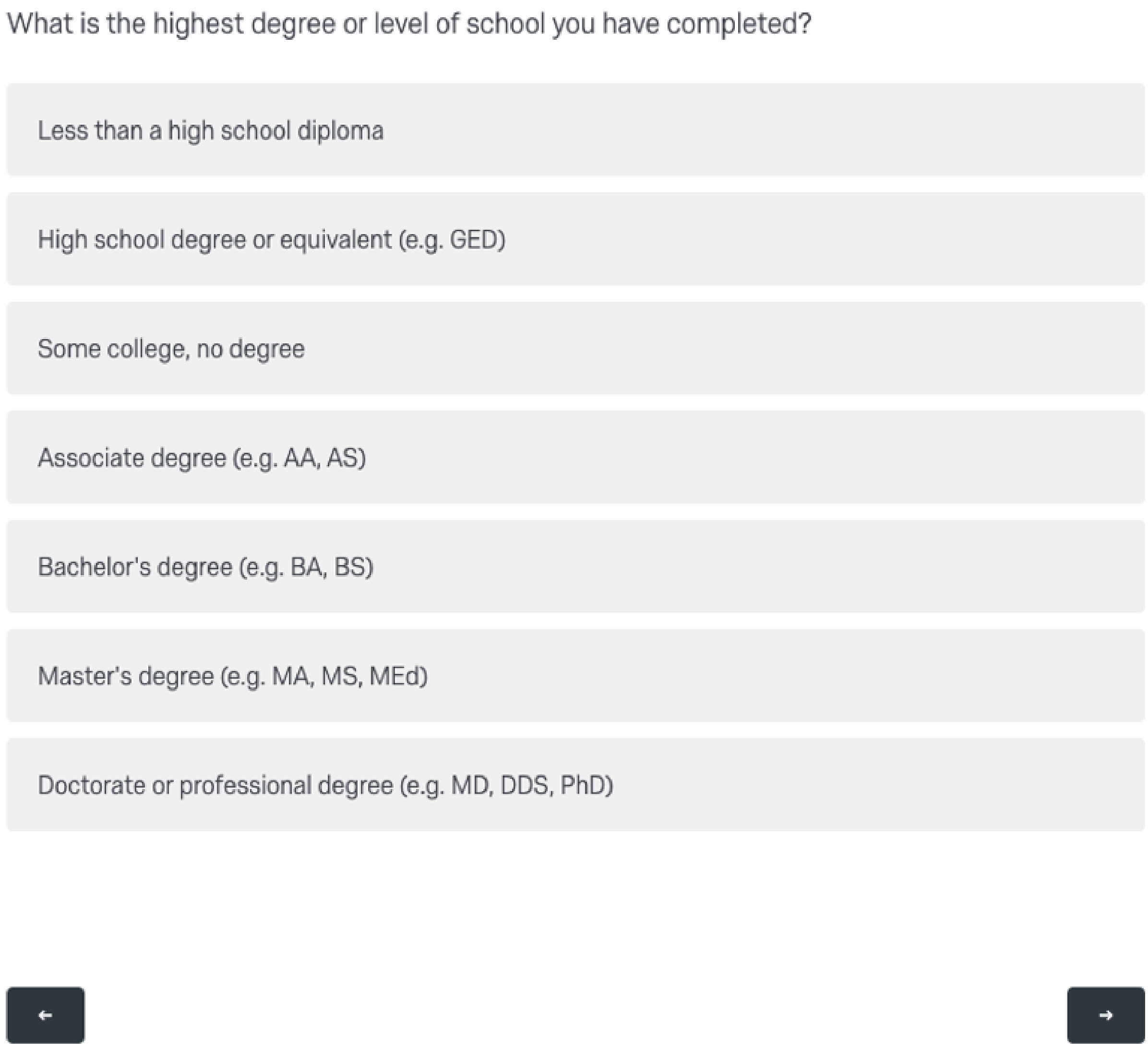


Figure 17: Entry survey question about how the subject's age


What is the highest degree or level of school you have completed?

- Less than a high school diploma
- High school degree or equivalent (e.g. GED)
- Some college, no degree
- Associate degree (e.g. AA, AS)
- Bachelor's degree (e.g. BA, BS)
- Master's degree (e.g. MA, MS, MEd)
- Doctorate or professional degree (e.g. MD, DDS, PhD)

Figure 18: Entry survey question about how the subject's highest level of education

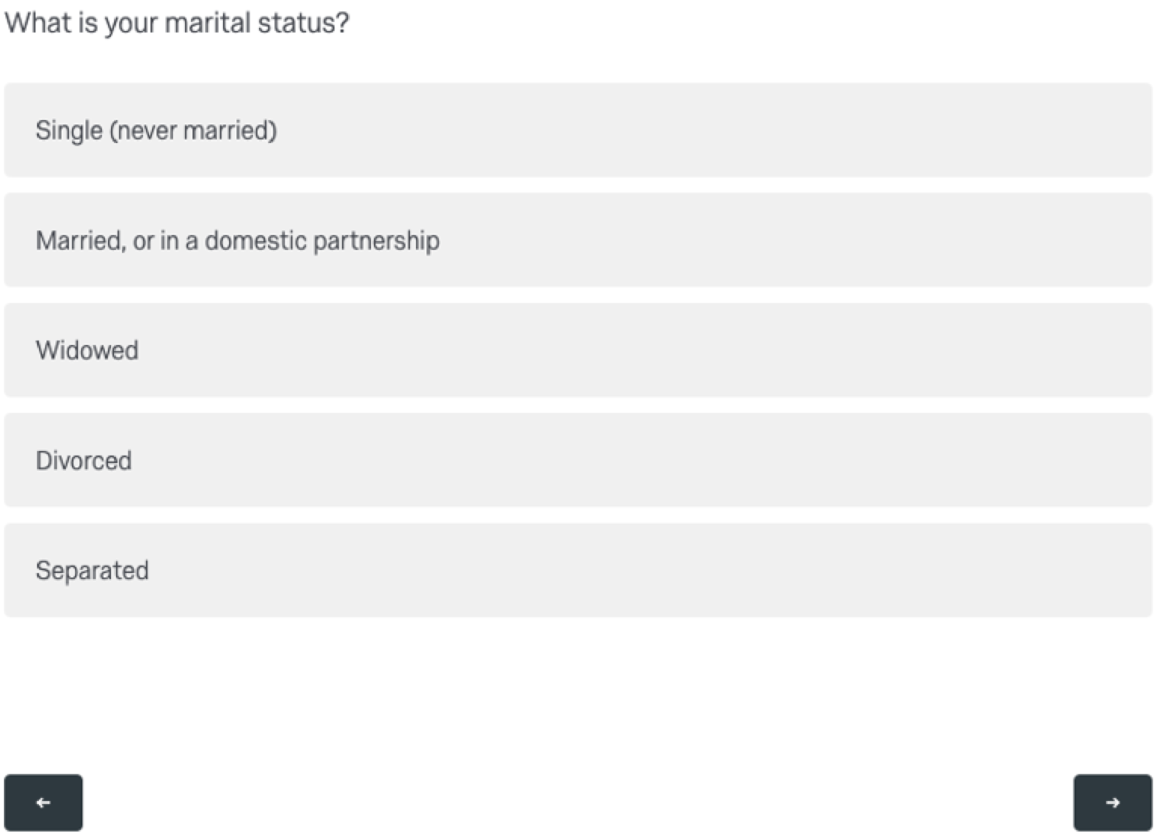


Figure 19: Entry survey question about how the subject's marital status

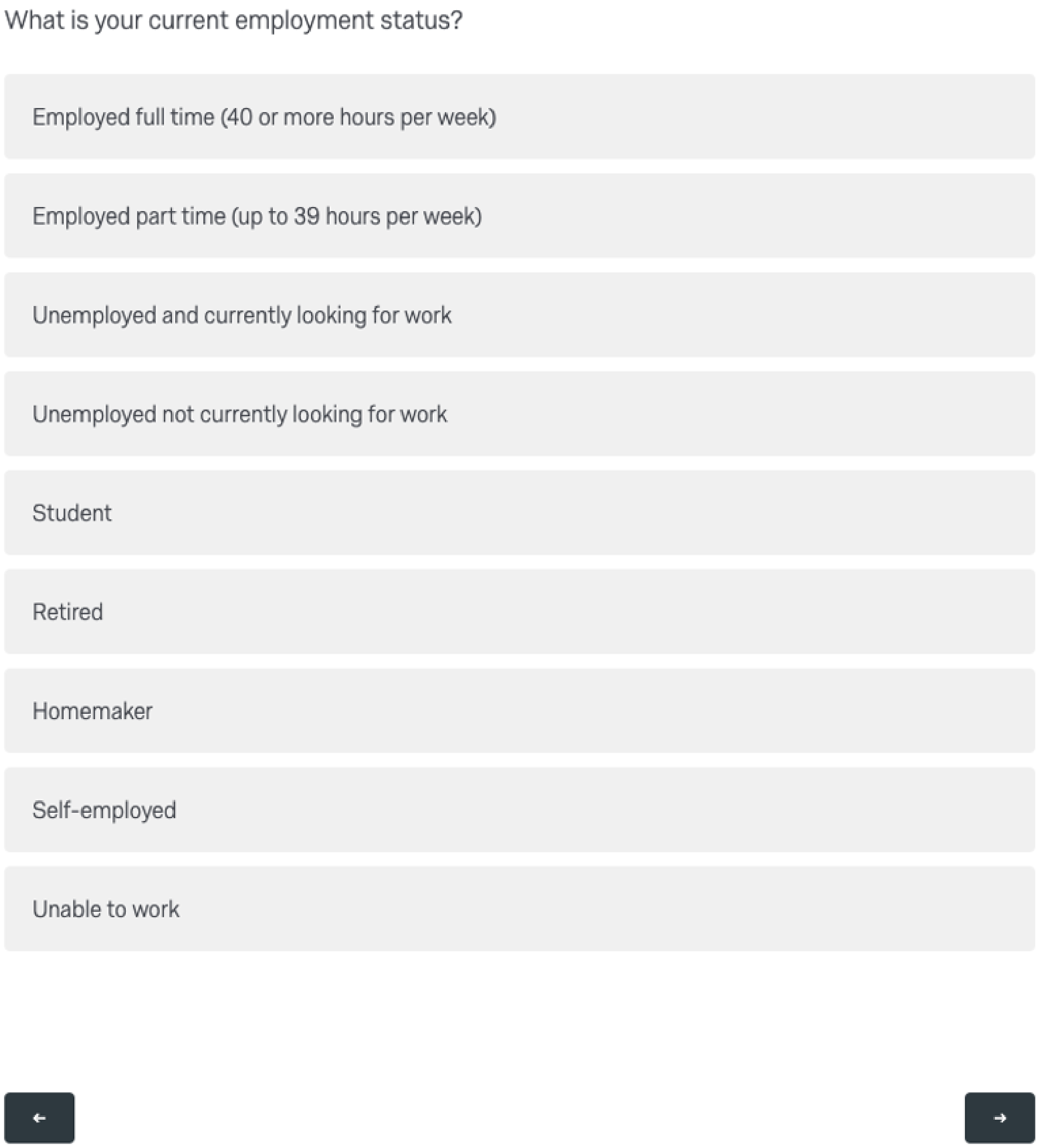


Figure 20: Entry survey question about how the subject's employment status

Are you CURRENTLY covered by any of the following types of health insurance or health coverage plans?

- NOT COVERED by any health insurance plan
- Insurance through a current or former employer or union (of yourself or another family member)
- Insurance purchased directly from an insurance company (by yourself or another family member)
- Medicare, for people 65 and older, or people with certain disabilities
- Medicaid, Medical Assistance, or any kind of government-assistance plan for those with low incomes or a disability
- TRICARE or other military health care
- Veterans Affairs (including if you have ever used or enrolled for Veterans Affairs health care)
- Indian Health Service
- Any other type of health insurance or health coverage plan

Figure 21: Entry survey question about how the subject's insurance status

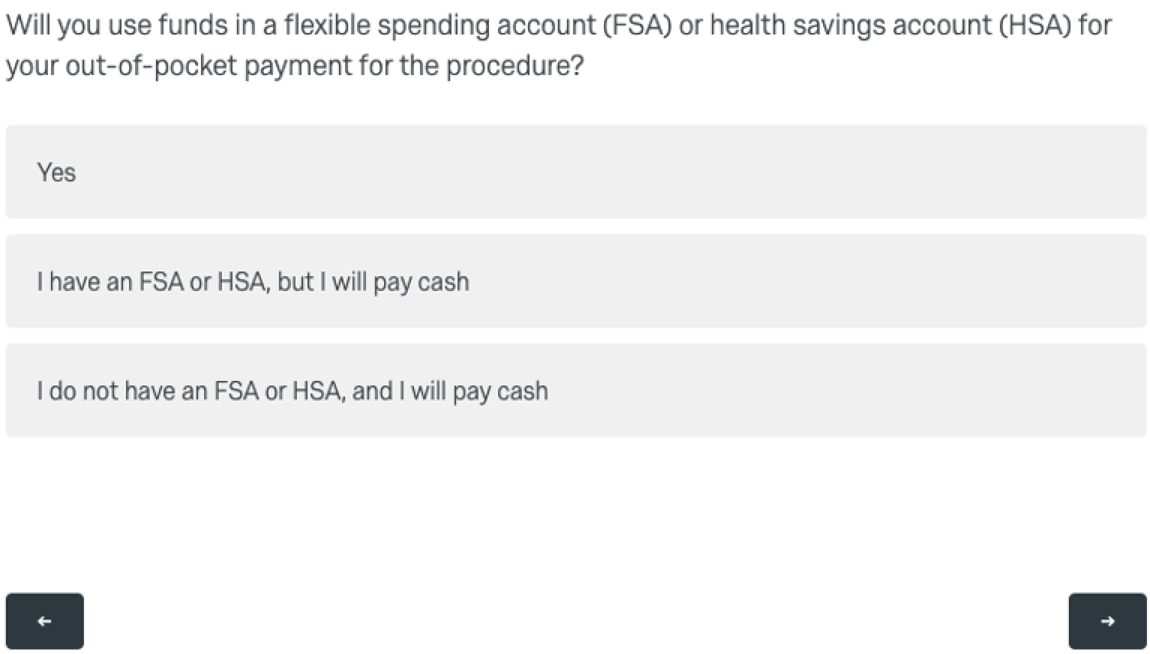


Figure 22: Entry survey question about the subject's method of payment

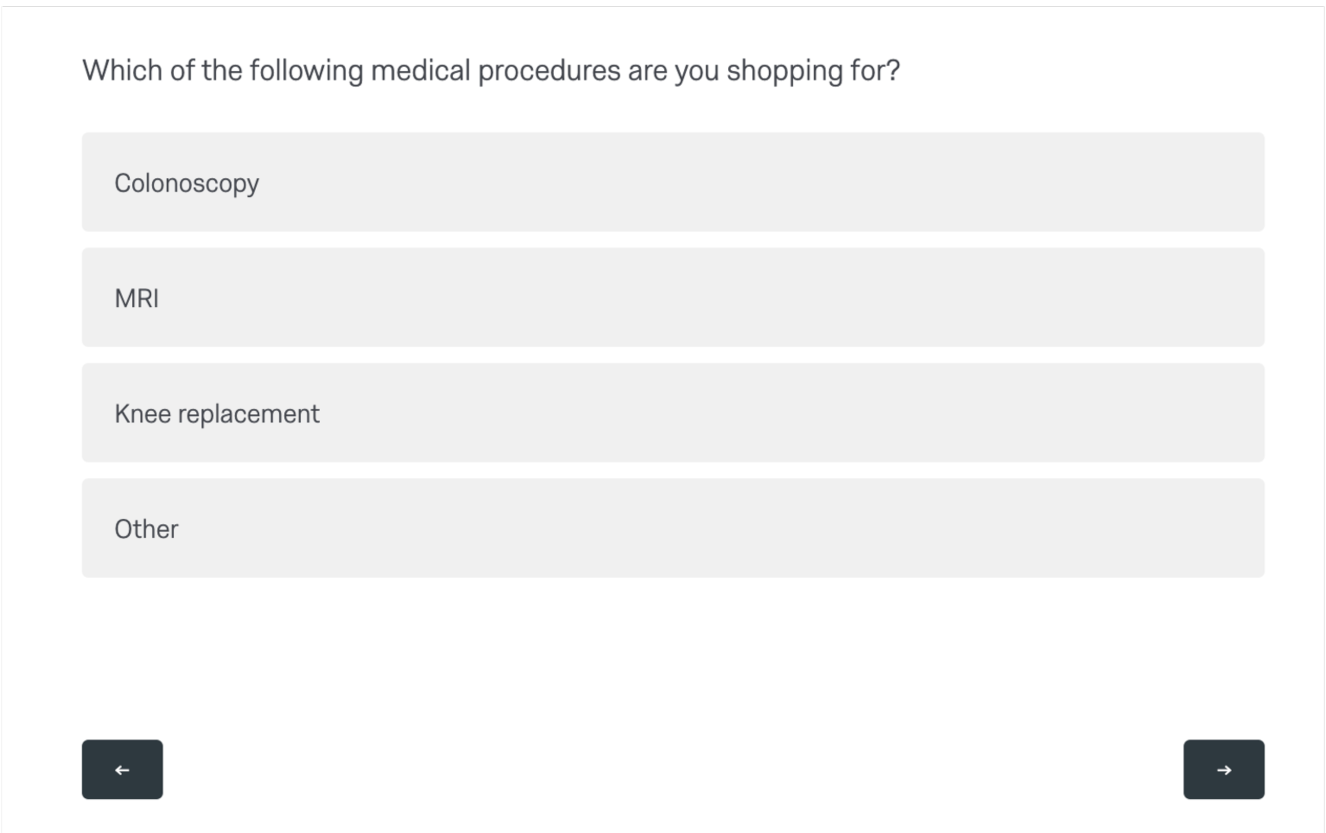


Figure 23: Entry survey question about which procedure the subject is shopping for

The subjects in the experiment in this present paper are the ones who had selected "colonoscopy" as the medical procedure that they are shopping for. They will then answer a few questions before getting redirected to the conjoint survey itself.

These last few questions inquire the subjects about their outside options: the time/effort used to identify a provider for the colonoscopy prior to using the "pilot" tool (Figure 24), number of options explored outside of the "pilot" (Figure 25), and the expected price and distance of the customer's outside options (Figure 26). After these questions, the subjects are shown a screen that describes the nature of the conjoint survey and explained how the responses to the conjoint survey maps to actual provider options that will be shared with them shortly after the completion of the survey (Figure 27).

There are customers who participated in the "pilot" to shop for either an MRI or a knee replacement. As mentioned in the main text of the paper, I do not use that data in the present study. The MRI customer data is used for a separate paper, while there is only one customer who used the pilot for knee replacement.

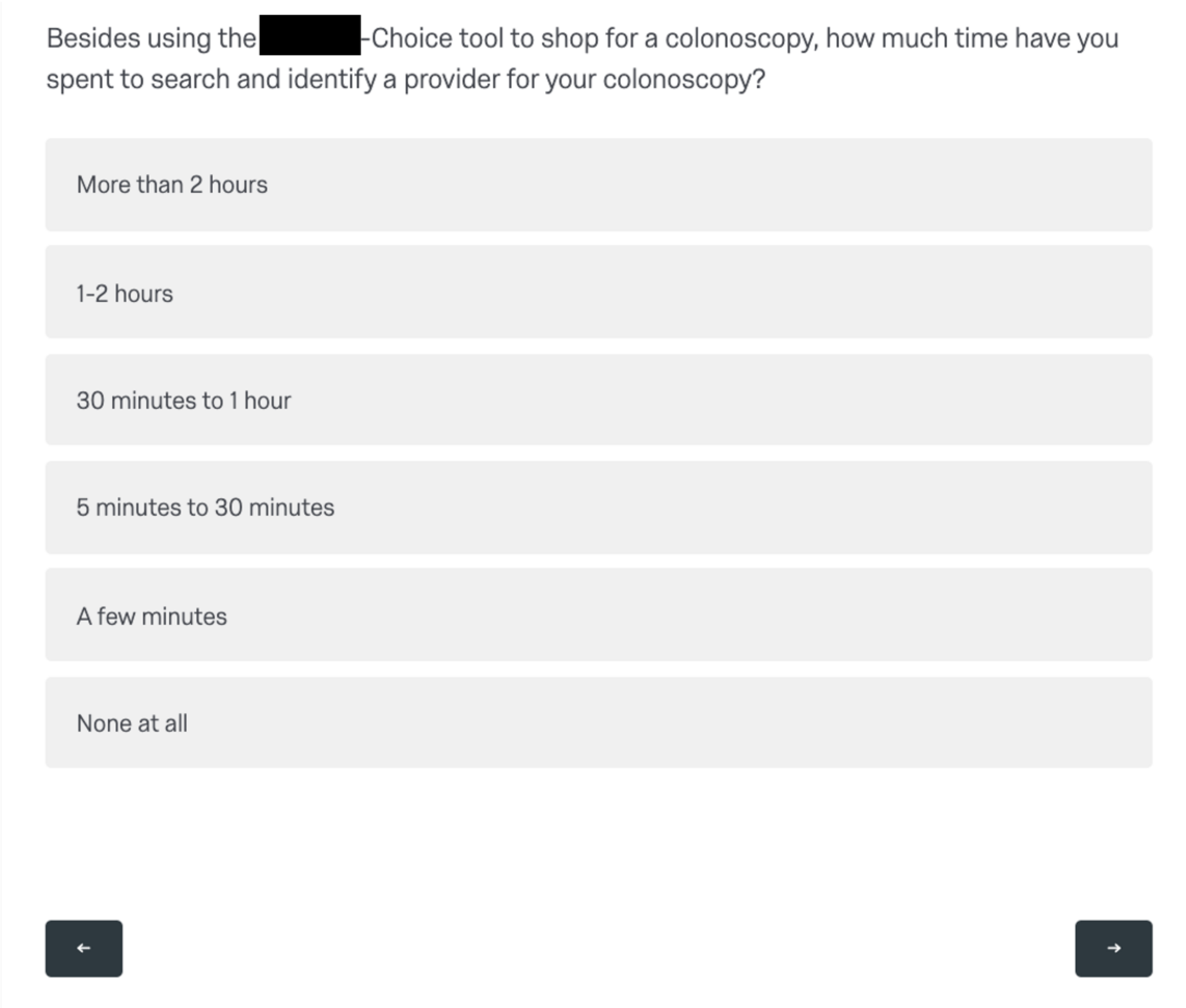


Figure 24: Entry survey question about how the subject's search effort before using the "pilot" tool

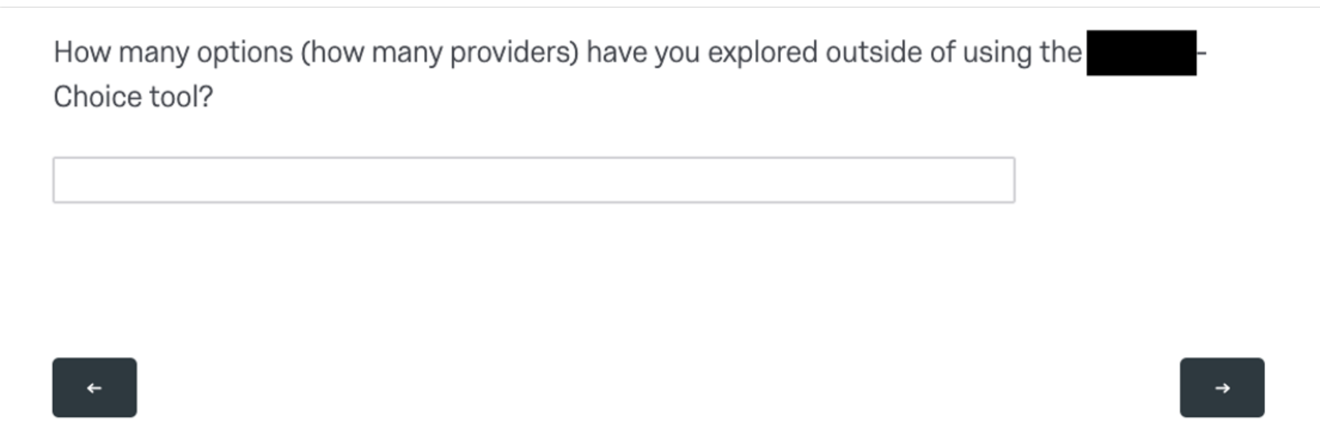


Figure 25: Entry survey question about the number of doctor options explored outside of the "pilot"

Based on your best estimate, how much would you and your insurance have to pay a provider for a colonoscopy in your zipcode? Please write down the SUM of what you have to pay (out-of-pocket, co-pay, co-insurance) and what the insurance plan will pay to the provider (in US$):

If you were to NOT use the [redacted]-Choice tool and shop by yourself to purchase your colonoscopy from a provider outside of the [redacted]-Choice tool, how much do you expect to pay out-of-pocket? (in US$)

If you were to NOT use the [redacted]-Choice tool and shop by yourself to purchase your colonoscopy from a provider outside of the [redacted]-Choice tool, how far would you travel to get the procedure?

- 0-10 miles
- 10-50 miles
- 50-100 miles
- 100-250 miles
- More than 250 miles

Figure 26: Entry survey question about the doctor options explored outside of the "pilot" (e.g. price and distance of the outside option)

Next, we'll find the best providers for your needs!

The following 12-14 questions will let us know what you're looking for in a provider. After you complete the 12-14 questions, we'll email you a list of 10 providers and instructions on how to make your booking. The more carefully you complete these 12-14 questions, the better we will be able to match you with real provider options.

In each question, you will choose between five hypothetical provider options. Just pick the option you like best, or click "None of the above" if none of the options is acceptable.

[redacted] will then find the best providers for you based on your responses, using [redacted]'s newly developed matching algorithm. We will email you with 10 real providers within 48 hours after you finish.

In the field below, please enter the email address where you would like to receive the 10 real providers options and instructions to complete the booking with one of them.

Happy shopping!

Figure 27: Entry survey screen to record email and reiterate instructions of the "pilot"

A.3.2 Conjoint Survey

The conjoint survey where customers evaluated 14 hypothetical doctors generates the data for the main analysis in the paper.

There are two slightly different version of the conjoint survey, and customers will only see one version. Which version of the conjoint survey the customer would see depends on whether they are randomized in the treatment or the control group.

The conjoint survey constitute screens where customer are informed about the doctor attributes they will see on the doctor profiles, followed by the 14 menus, and then the exit survey (described in the next subsection).

I present the conjoint survey screen-by-screen as a subject would experience it. If the screens for the treatment and control groups differ, I will show the control group version first followed by the treatment group version before moving to the next screen.

First, both treatment arms will see an introduction screen (Figure 28), followed by a screen introducing the attributes of the doctor options in the menus the customer would see. The one for the control group (Figure 29) is different from the one for the treatment group (Figure 30), as the latter has one more attribute than the former.

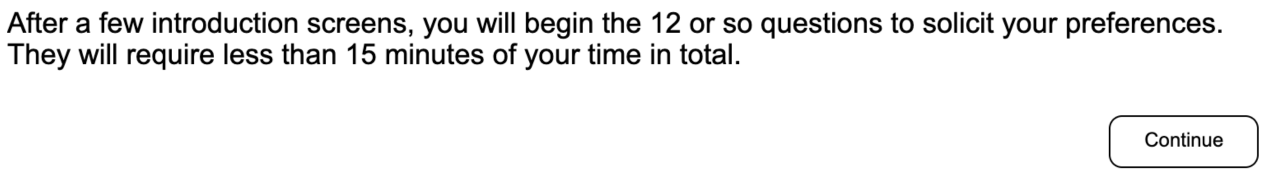


Figure 28: Launch screen for conjoint survey

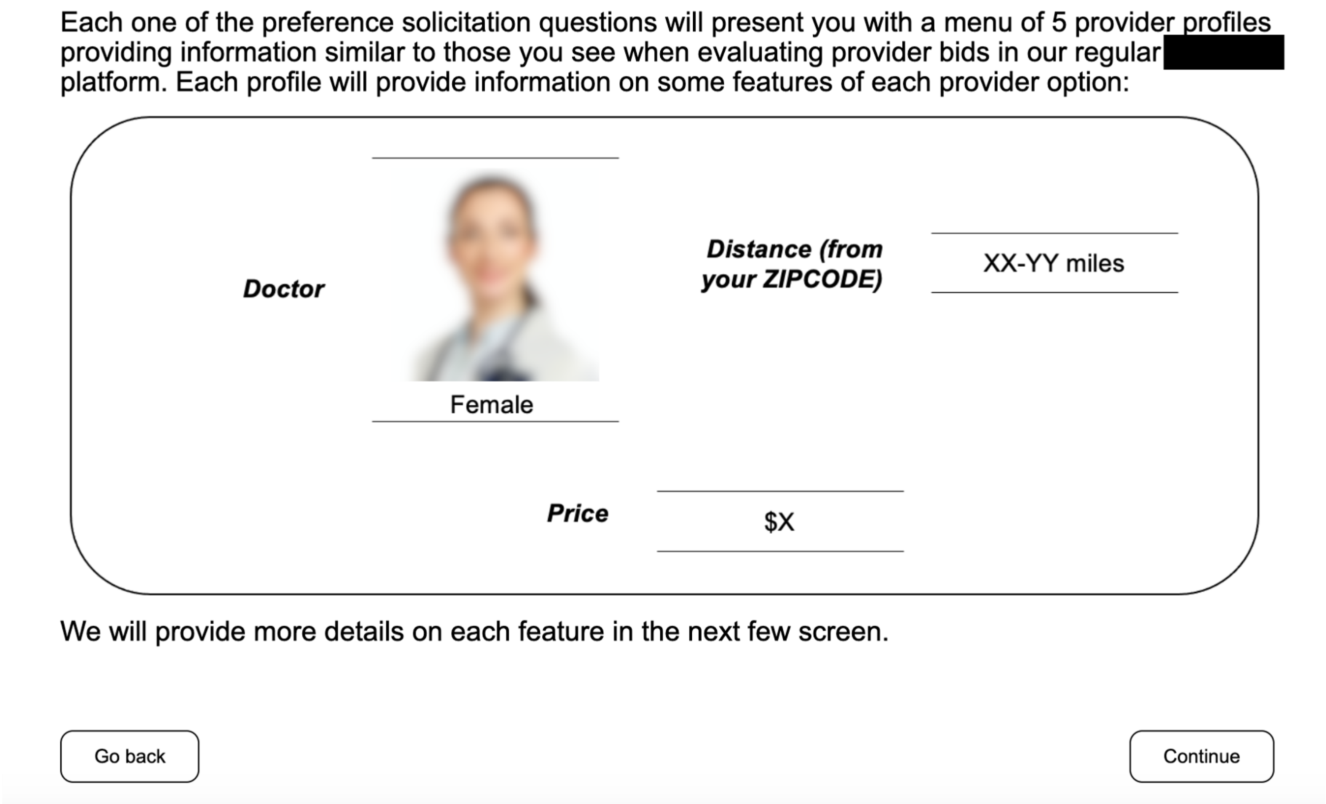


Figure 29: Screen introducing attributes for control group

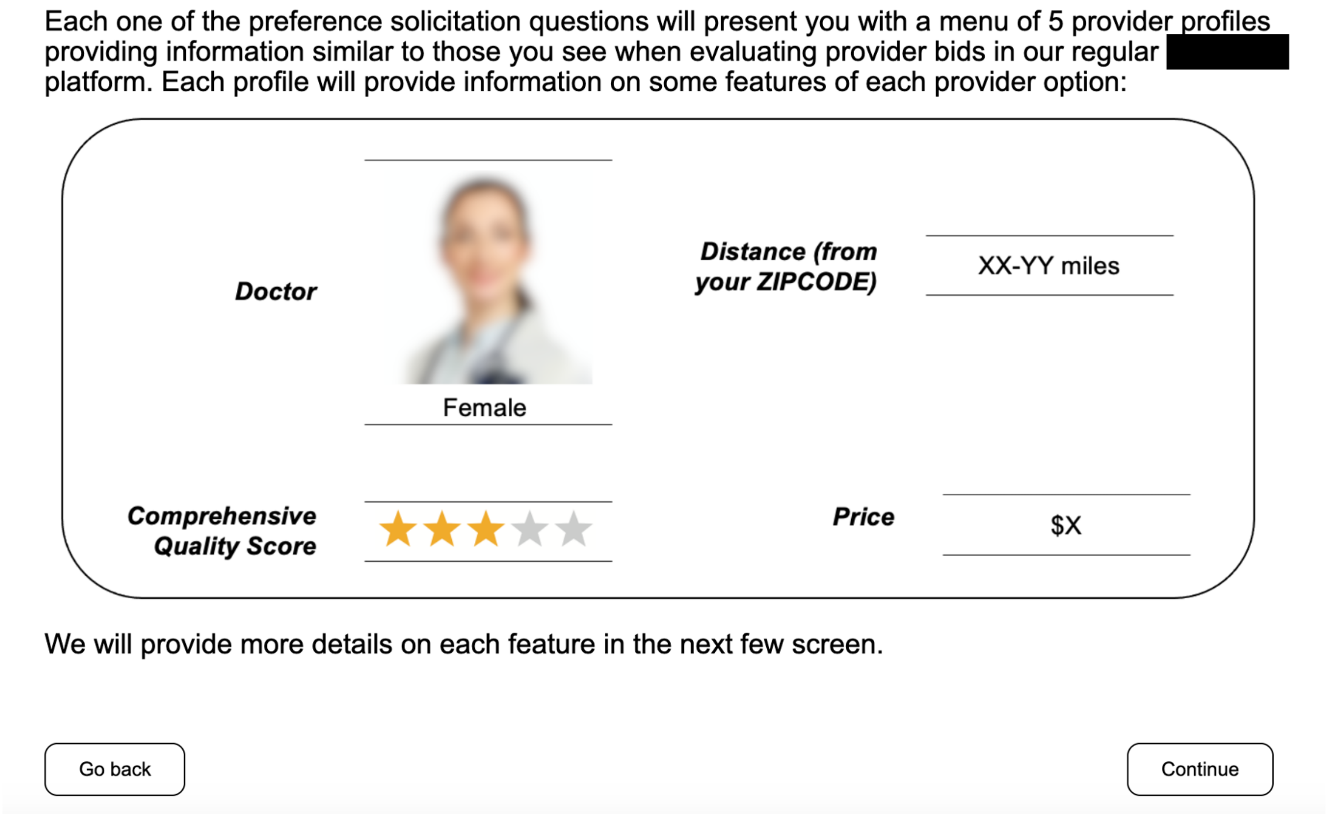


Figure 30: Screen introducing attributes for treatment group

Information about each attribute will be presented with its own screen. The order with which these screens are presented is randomized across subjects (e.g. one subject might see the information about price first, distance second and the doctor profile picture and gender last but another subject might see distance first and other attributes in a different order). Before the subject can move on from one screen to the next, they have to pass an attention test by picking the correct color (yellow, green, and blue). The ordering of these colors (ordering of the buttons) are randomized across subjects and screens.

Both treatment and control group subjects will see a screen clarifying the doctor attributes of "price," "doctor" (which include a doctor profile picture and gender), and "distance from your ZIPCODE." The treatment group will also see an additional screen clarifying the doctor attribute of "Comprehensive Quality Score," the quality signal.

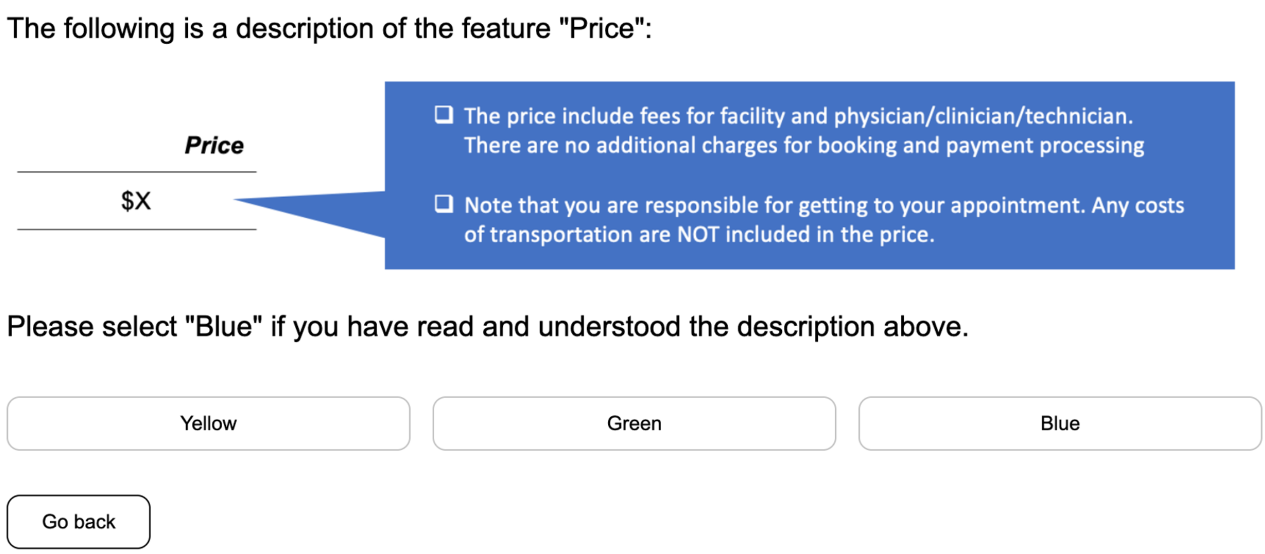


Figure 31: Information screen for doctor attribute of "Price"

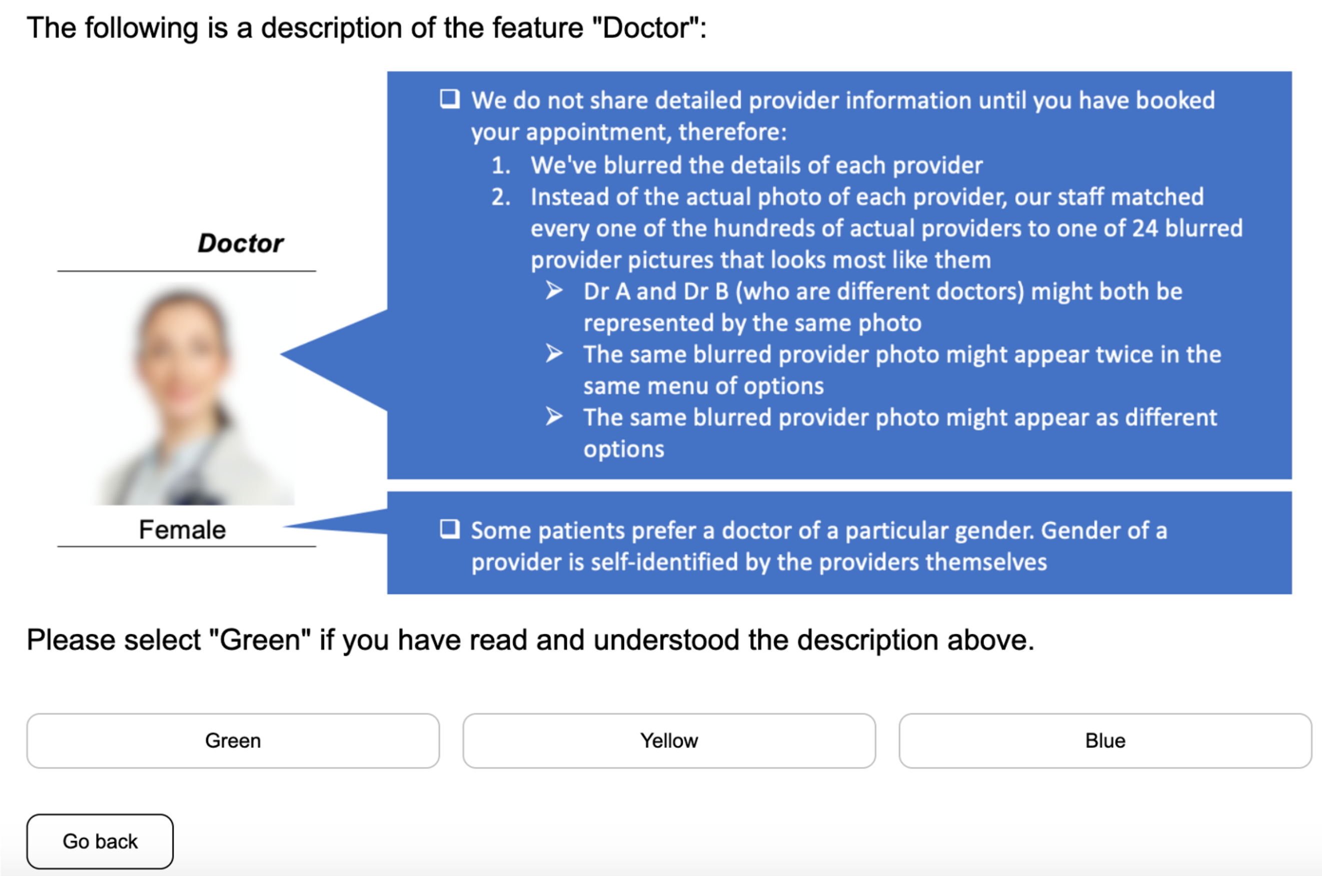


Figure 32: Information screen for doctor attribute of "Doctor": including the profile picture and gender

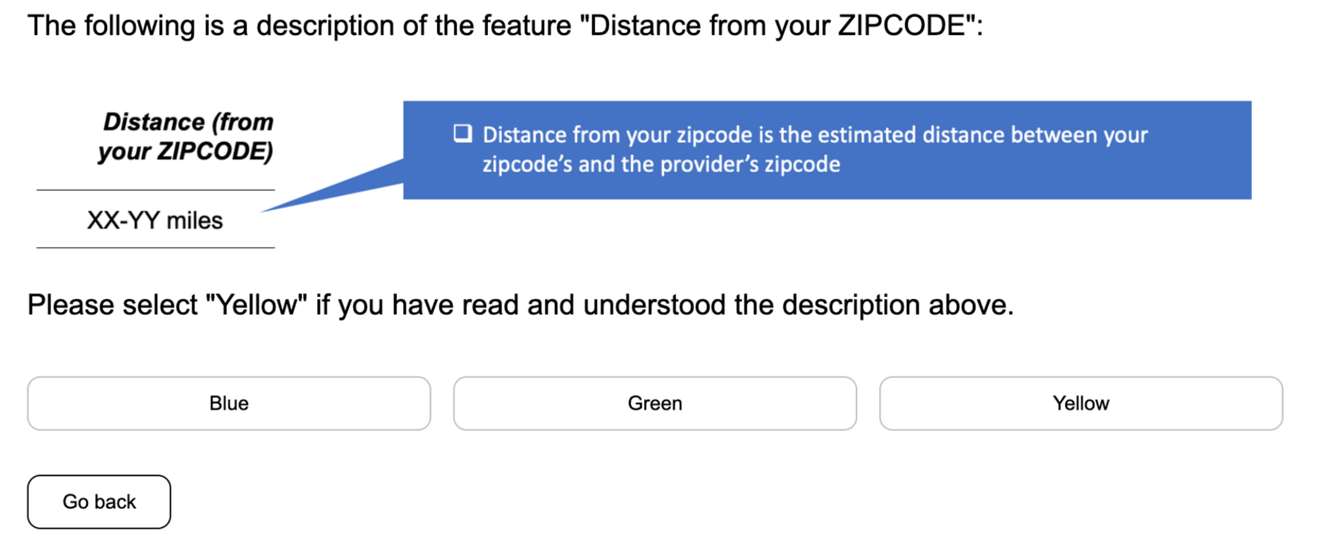
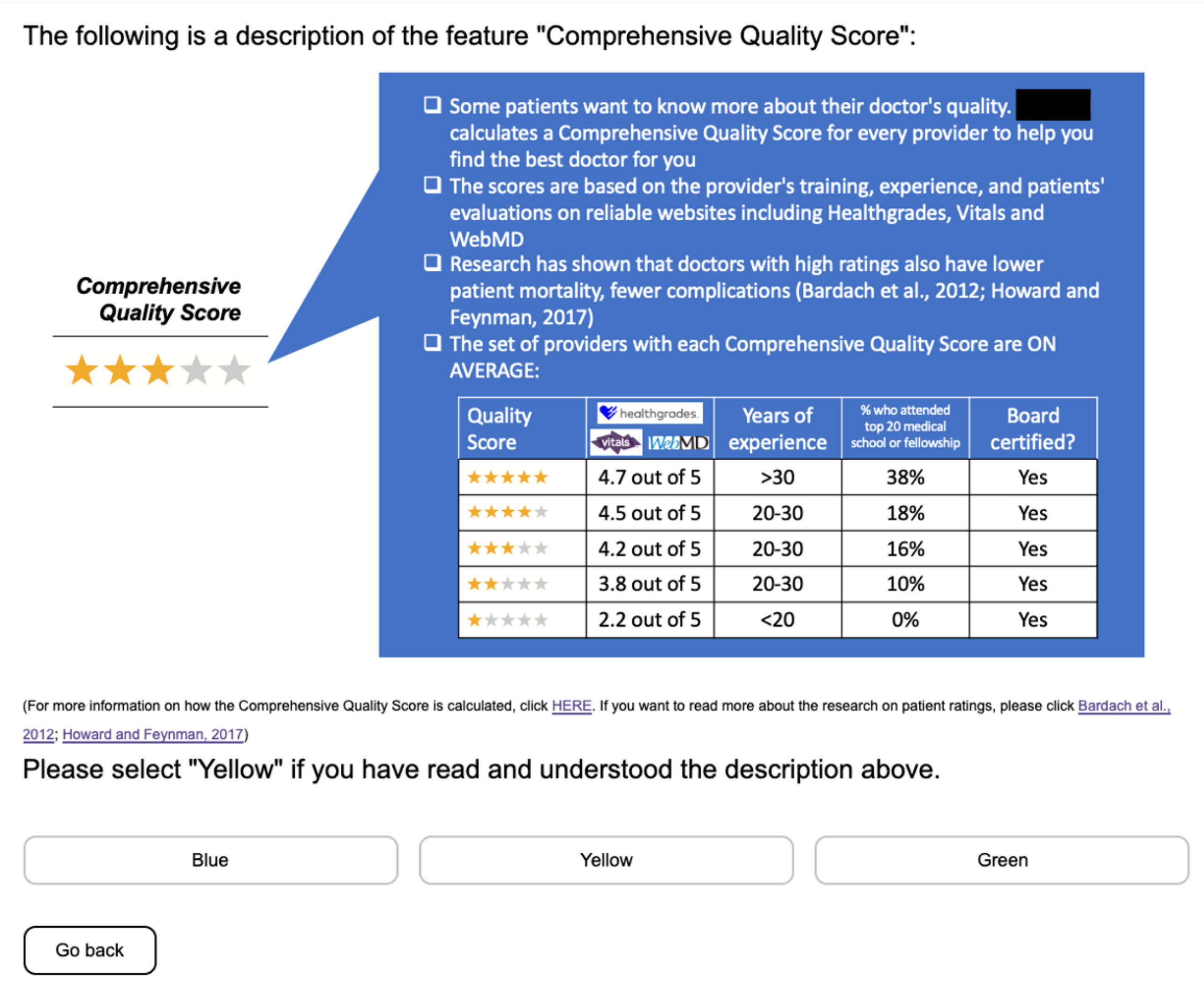


Figure 33: Information screen for doctor attribute of Travel Distance


The following is a description of the feature "Comprehensive Quality Score":

Comprehensive Quality Score

★★★★★

- Some patients want to know more about their doctor's quality. [redacted] calculates a Comprehensive Quality Score for every provider to help you find the best doctor for you
- The scores are based on the provider's training, experience, and patients' evaluations on reliable websites including Healthgrades, Vitals and WebMD
- Research has shown that doctors with high ratings also have lower patient mortality, fewer complications (Bardach et al., 2012; Howard and Feynman, 2017)
- The set of providers with each Comprehensive Quality Score are ON AVERAGE:

Quality Score	healthgrades vitals WebMD	Years of experience	% who attended top 20 medical school or fellowship	Board certified?
★★★★★	4.7 out of 5	>30	38%	Yes
★★★★☆	4.5 out of 5	20-30	18%	Yes
★★★☆☆	4.2 out of 5	20-30	16%	Yes
★★☆☆☆	3.8 out of 5	20-30	10%	Yes
★☆☆☆☆	2.2 out of 5	<20	0%	Yes

(For more information on how the Comprehensive Quality Score is calculated, click HERE. If you want to read more about the research on patient ratings, please click Bardach et al., 2012; Howard and Feynman, 2017)

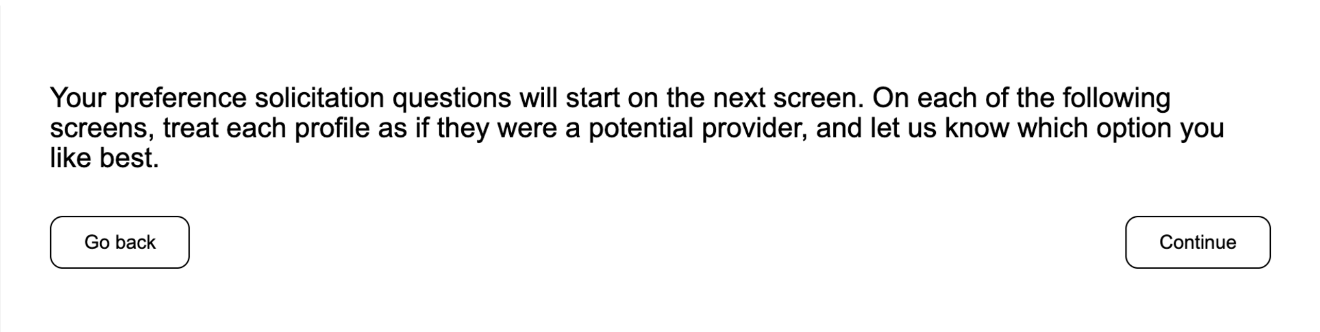
Please select "Yellow" if you have read and understood the description above.

Blue | Yellow | Green

Go back

Figure 34: Information screen for doctor attribute of "Comprehensive Quality Score"

After the screens to help customers comprehend the doctor attributes, a screen alerts the customers that the "preference solicitation questions" will start, this is the last screen before the customers evaluate the 14 menus of hypothetical doctor profiles (see Figure 35).


Your preference solicitation questions will start on the next screen. On each of the following screens, treat each profile as if they were a potential provider, and let us know which option you like best.

Go back | Continue

Figure 35: Screen alerting subjects that conjoint survey is about to start

In the key part of the conjoint survey (and VIC), the customers evaluate 14 menus. Each menu will have 5 doctor options and a button for "None of the above." The ordering of the attributes from top to bottom of a doctor profile is randomized across subjects but price is always presented as the bottom attribute. The ordering of attributes from top to bottom is the same within each subject's set of 14 menus. An example of the control group's menus is presented in Figure 36. An example of the treatment group's menus is presented in Figure 37.

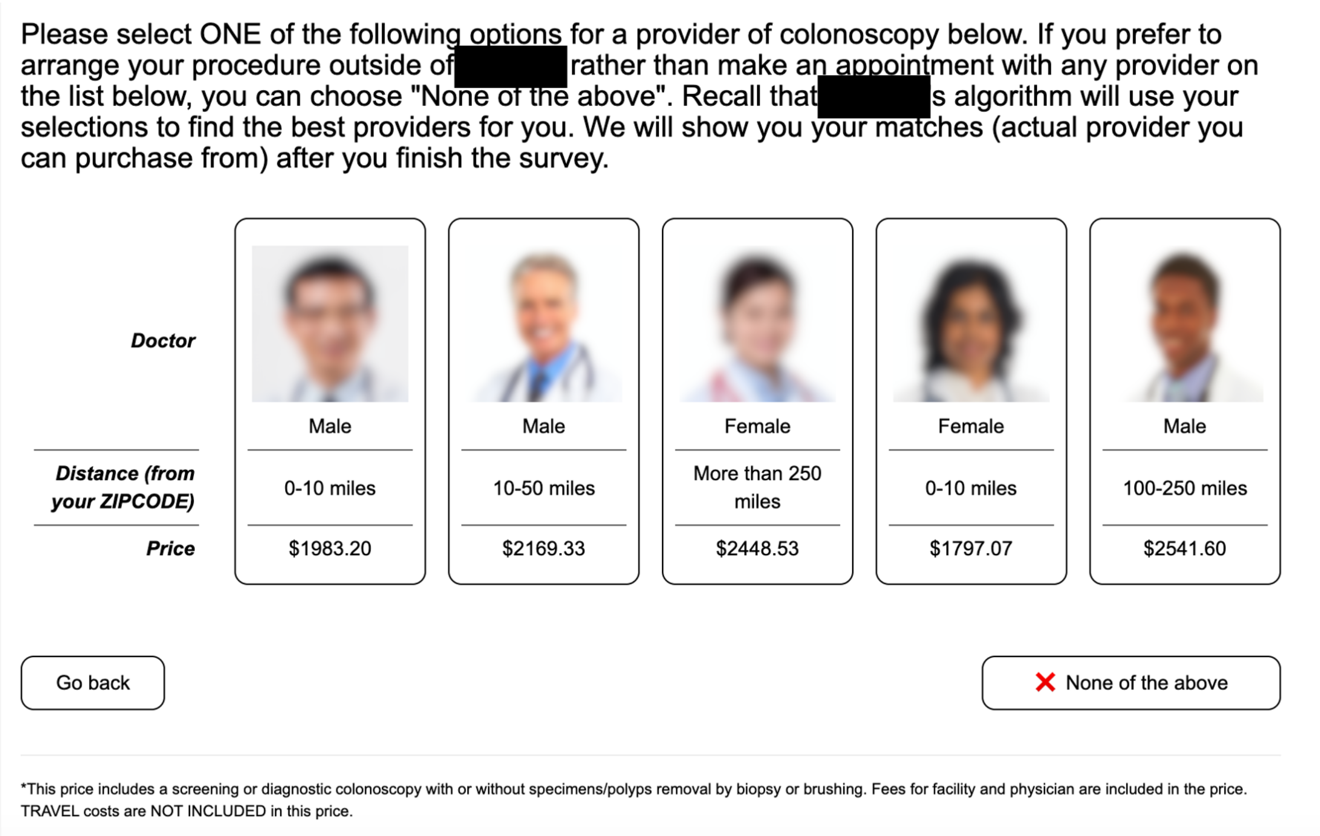


Figure 36: Information screen for doctor attribute of "Comprehensive Quality Score"

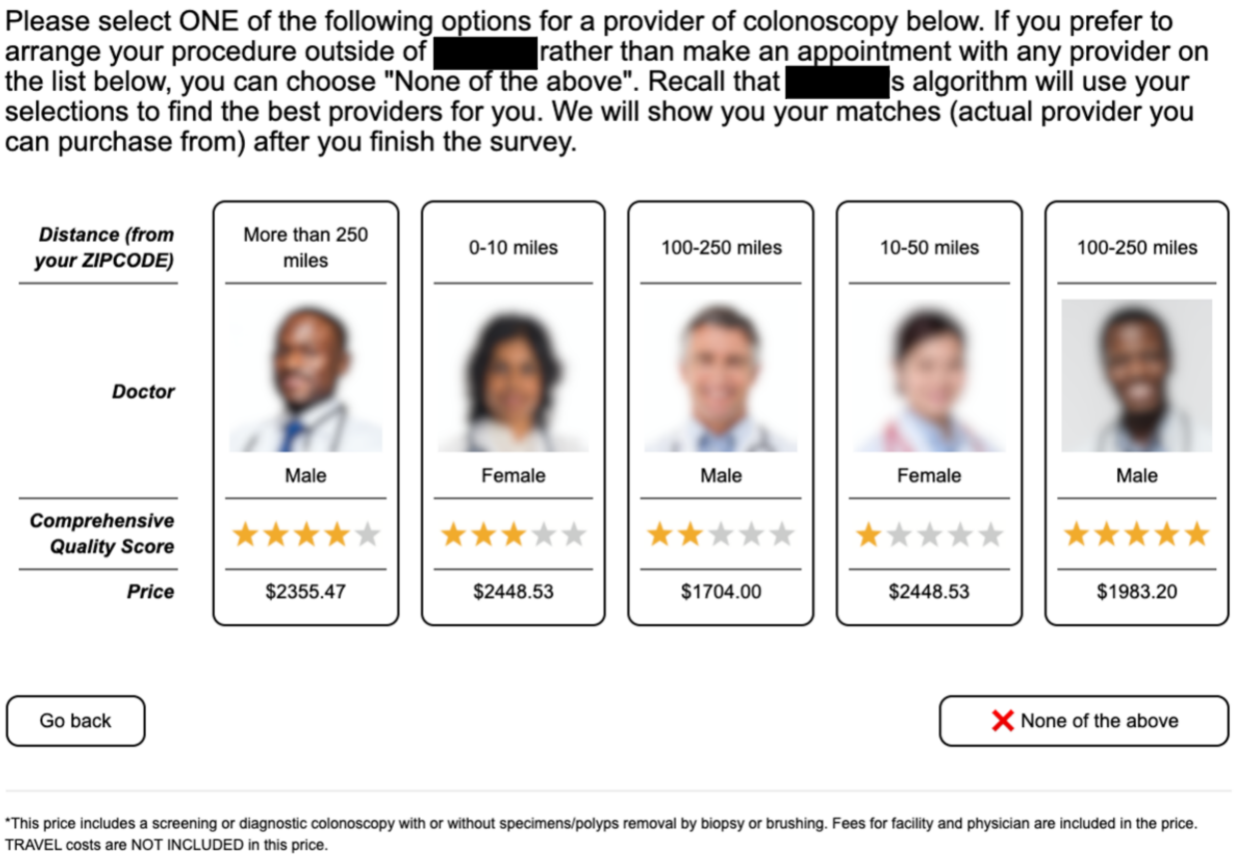


Figure 37: Information screen for doctor attribute of “Comprehensive Quality Score”

A.3.3 Exit Survey

After finishing the evaluation of the 14 menus, each customer is asked to describe their approach in a free text box (Figure 38). The provision of comments is optional.

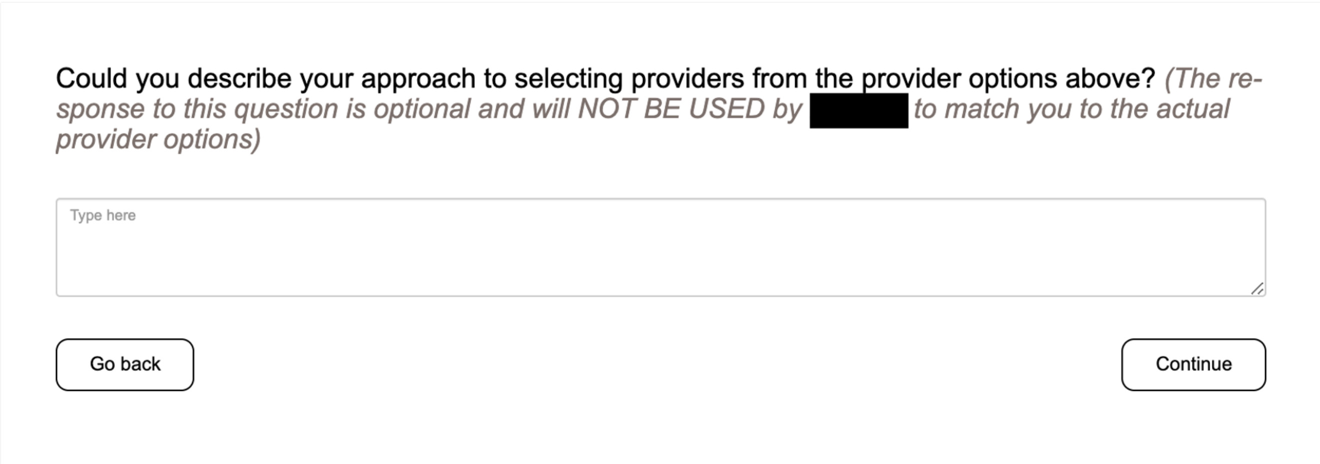


Figure 38: Screen asking subjects for free-text description of their approach

After the comment box, subjects from both treatment and control groups are asked to subjectively evaluate the quality of providers from various distances from their zipcode based on their own experience (Figure 39).

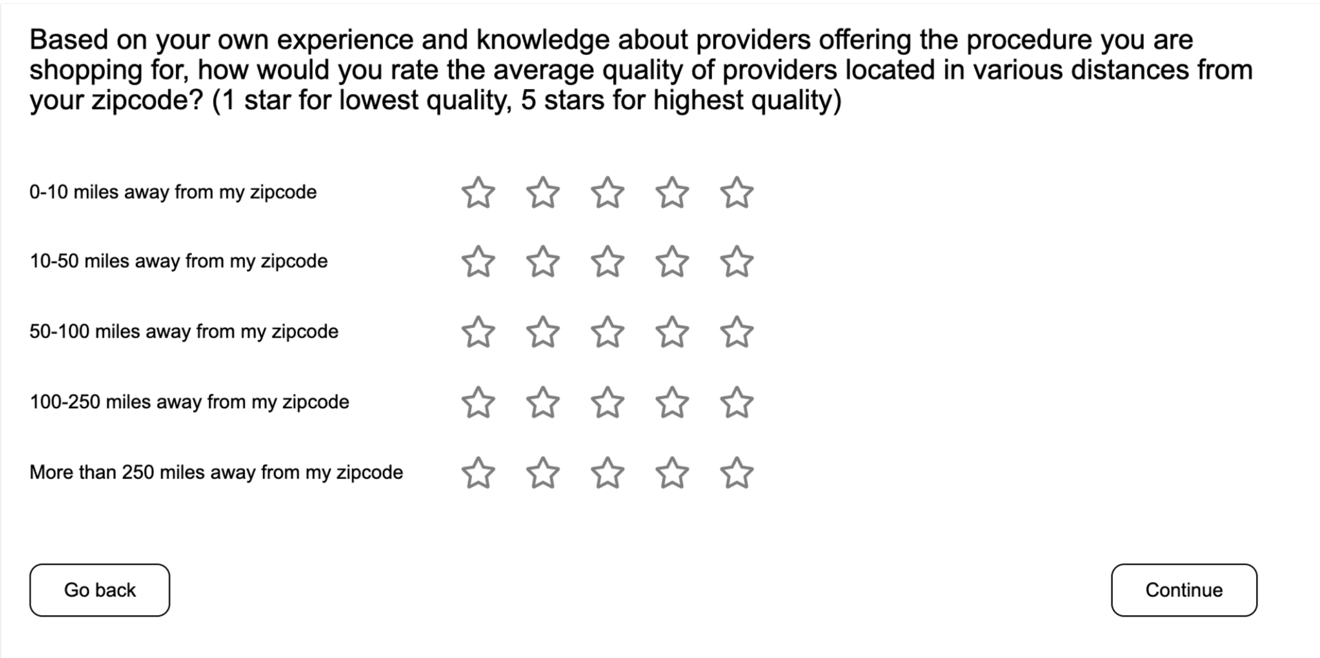


Figure 39: Screen asking customer's expected correlation between distance and quality

For subjects in the treatment group, they are asked 3 Likert scale questions for their opinion on the Comprehensive Quality Score (Figures 40, 41, and 42).

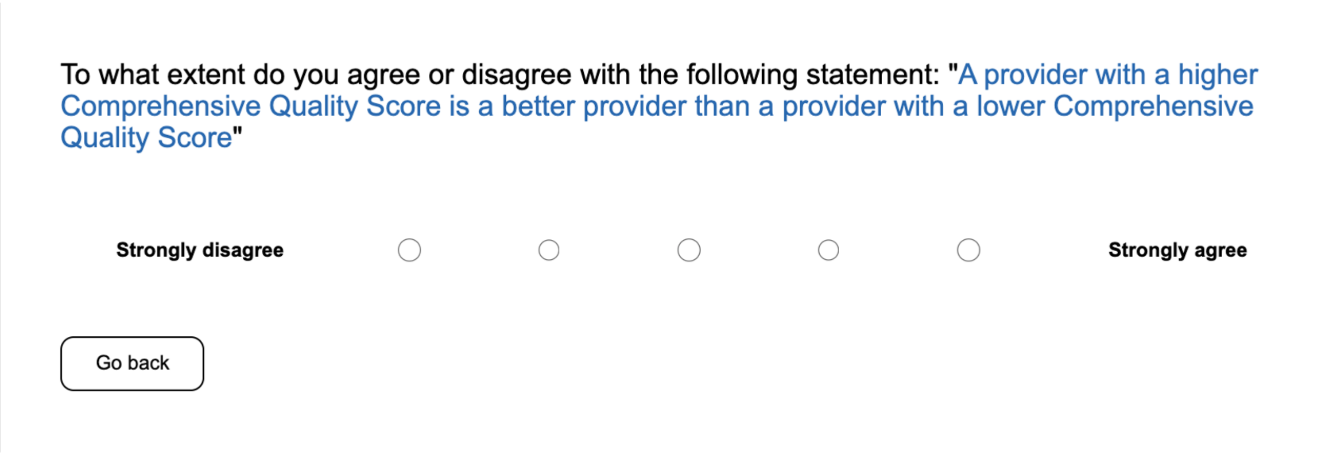


Figure 40: First Likert scale question on quality score for treatment group subjects

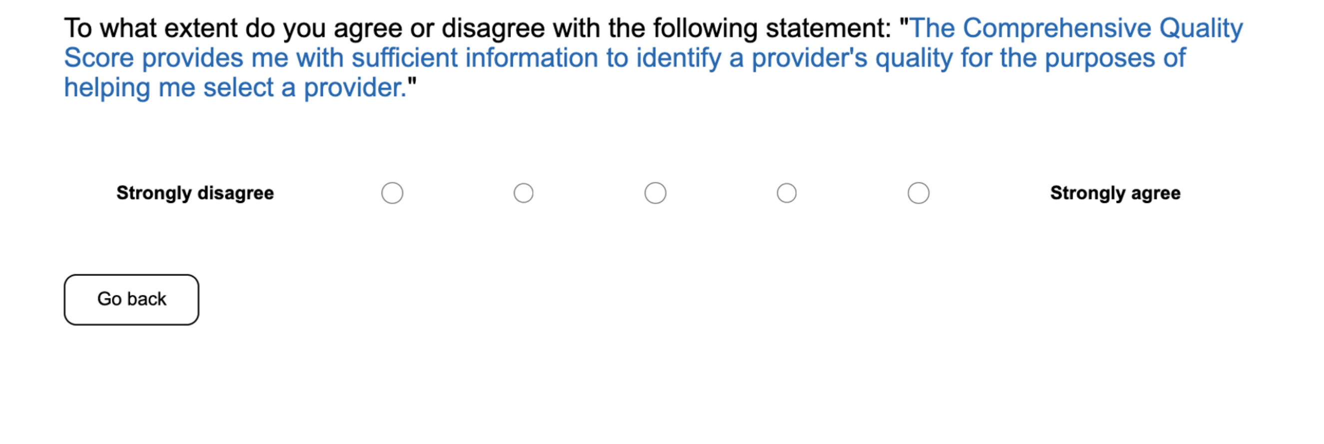


Figure 41: Second Likert scale question on quality score for treatment group subjects

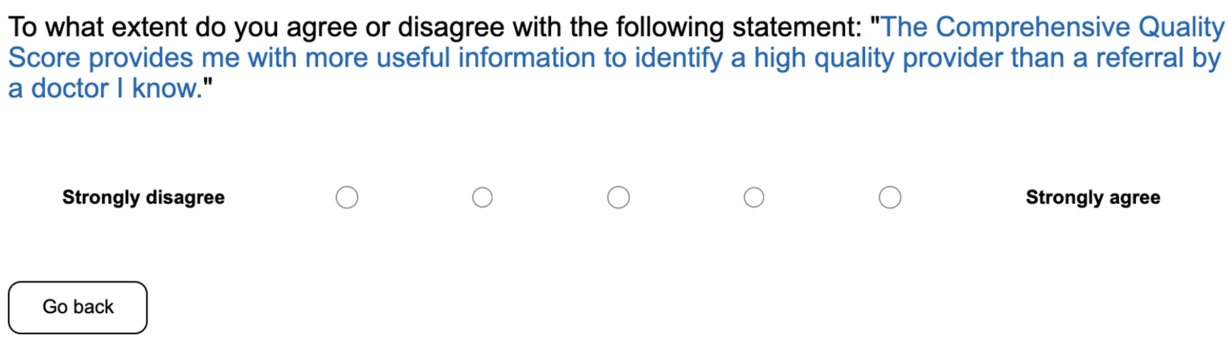


Figure 42: Third Likert scale question on quality score for treatment group subjects

Before the subjects finish the conjoint survey, I gathered information on their gender and race as well. This happen across 3 consecutive screens (Figures 43, 44 and 45). This information is gathered for both treatment and control group.

Figure 43: Exit survey question about gender

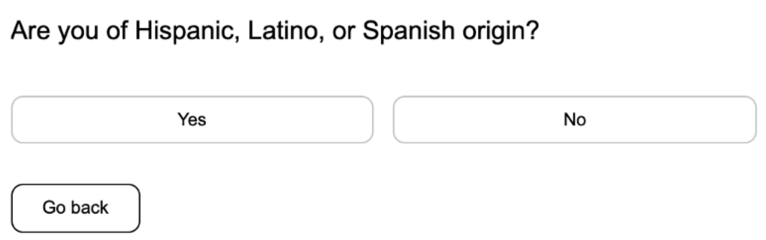


Figure 44: Exit survey question about ethnicity

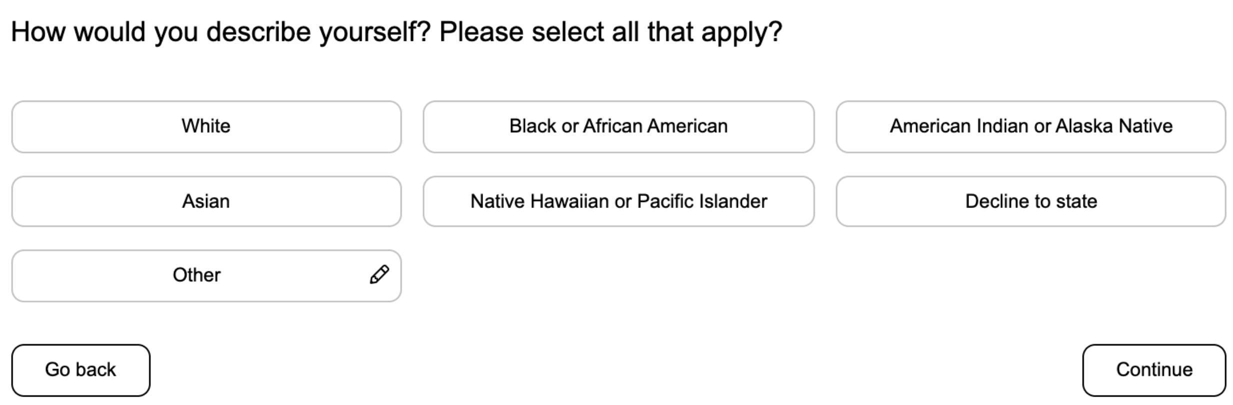
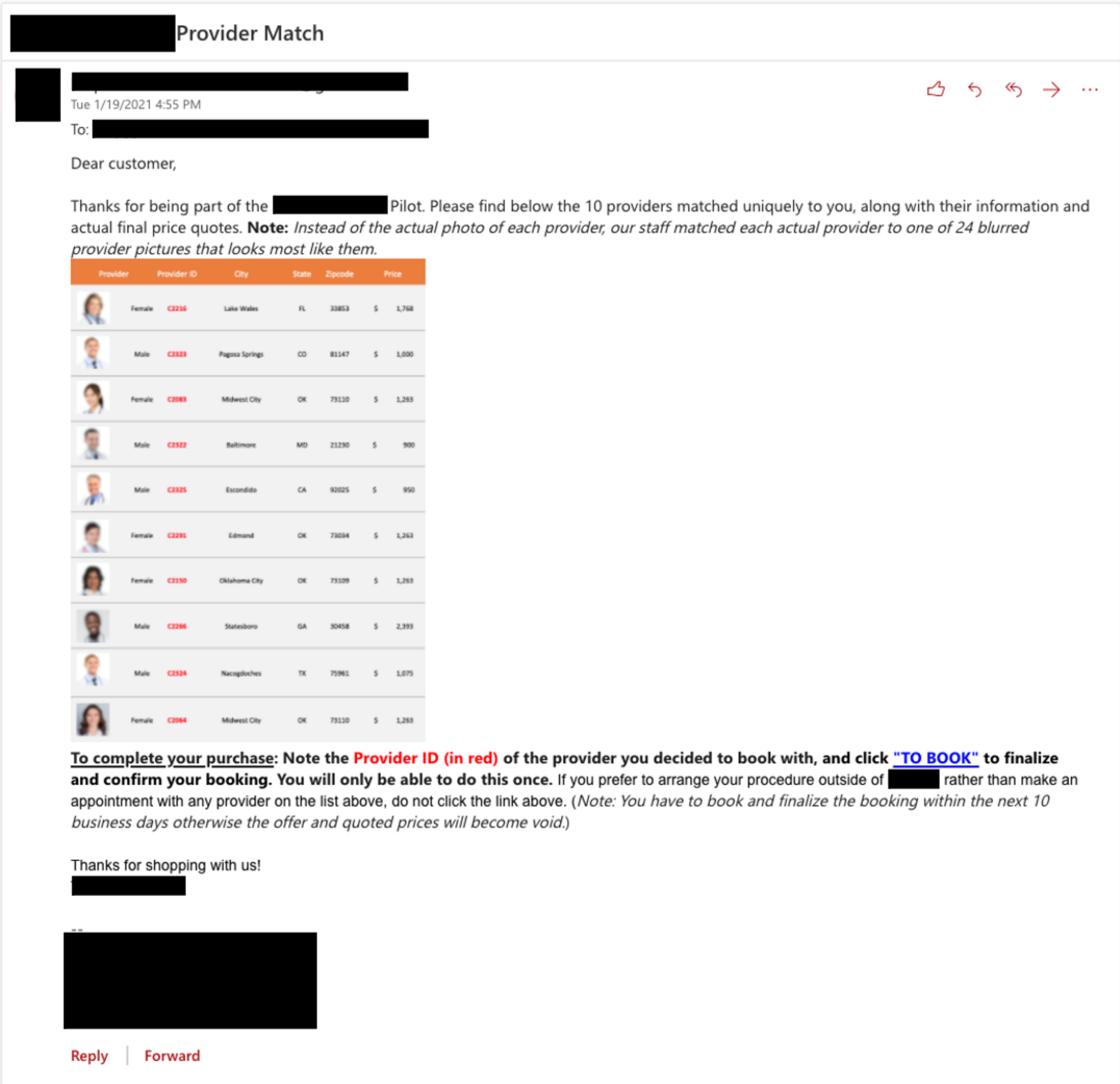


Figure 45: Exit survey question about race

A.3.4 Actual Booking Options

Within 48 hours after each subject finished the entry survey, conjoint survey, and exit survey, they will received a customized email (see Figure 46) listing the 10 actual doctors matched to them. The subjects were told that the offers were only valid for the next 10 business days and have to click on a link to get redirected to an online booking agreement form to complete the booking.



[redacted] Provider Match

[redacted]
Tue 1/19/2021 4:55 PM
To: [redacted]

Dear customer,

Thanks for being part of the [redacted] Pilot. Please find below the 10 providers matched uniquely to you, along with their information and actual final price quotes. **Note:** *Instead of the actual photo of each provider, our staff matched each actual provider to one of 24 blurred provider pictures that looks most like them.*

Provider	Provider ID	City	State	Zipcode	Price
Female	C2216	Lake Wales	FL	33853	$ 1,768
Male	C2323	Pagosa Springs	CO	81147	$ 1,000
Female	C2083	Midwest City	OK	73110	$ 1,263
Male	C2322	Baltimore	MD	21230	$ 900
Male	C2325	Escondido	CA	92025	$ 950
Female	C2291	Edmond	OK	73034	$ 1,263
Female	C2150	Oklahoma City	OK	73109	$ 1,263
Male	C2266	Statesboro	GA	30458	$ 2,393
Male	C2324	Nacogdoches	TX	75961	$ 1,075
Female	C2064	Midwest City	OK	73110	$ 1,263

To complete your purchase: Note the Provider ID (in red) of the provider you decided to book with, and click "TO BOOK" to finalize and confirm your booking. You will only be able to do this once. If you prefer to arrange your procedure outside of [redacted] rather than make an appointment with any provider on the list above, do not click the link above. (*Note: You have to book and finalize the booking within the next 10 business days otherwise the offer and quoted prices will become void.*)

Thanks for shopping with us!
[redacted]

--
[redacted]

Reply | Forward

Figure 46: Email to customer with actual doctor options for booking

A.4 Hypothetical Doctor Profile Pictures

A hypothetical doctor profile picture appears as an element on each profile in the menus (Figure 47). I manipulate perceptions of the doctor race by using profile pictures that clearly indicate the race of the hypothetical doctor. The resolution of each profile picture is reduced so that the doctor's race shows through but their attractiveness and other features are obscured. The profile pictures are chosen such that they have a similar degree of facial symmetry across races (see Table 6).

To see if certain pictures or features drive all the observed discrimination for each of the minority races, I also run regressions that yields results similar to Table 2 where I include each minority doctor photo separately as a dummy variable (e.g. Black Male doctor 1). All but one estimate for photos of the same race and gender are statistically indistinguishable from each other holding the treatment arm constant. An exception is Asian Male doctor 3, or the third Asian male doctor from the left in Figure 47, whose photo yielded significantly larger penalties than others in the same race and gender. In particular, using the fourth picture from the left on the first row in Figure 47 (a white male doctor) as reference, the range of willingness-to-pay premia/penalties for each group of profile pictures are as follows:

- White Male doctors (3 profiles): \$11.8 to \$46.5 without quality signals and \$24.0 to \$80.0 with quality signals.
- White Female doctors (4 profiles): $-$\$29.7 to \$37.0 without quality signals and \$19.0 to \$37.9 with quality signals.
- Black Male doctors (4 profiles): $-$\$212.9 to $-$\$286.4 without quality signals and $-$\$1.9 to \$36.9 with quality signals.
- Black Female doctors (4 profiles): $-$\$241.9 to $-$\$283.7 without quality signals and $-$\$14.7 to \$49.8 with quality signals.
- Asian Male doctors (4 profiles): $-$\$147.4 to $-$\$161.0 from the three profiles excluding Asian Male doctor 3 without quality signals and \$4.9 to \$47.4 from the three profiles excluding Asian Male doctor 3 with quality signals. When there is no quality signal, the WTP penalty for Asian Male doctor 3 is $-$\$201.5 (which is a significantly higher WTP penalty than Asian Male doctors 2 and 4 at the 5%

level and Asian Male doctor 1 at the 10% level). When there is a quality signal, the WTP penalty for Asian Male doctor 3 is −\$11.7 (which is a significantly higher WTP penalty than Asian Male doctors 1, 2 and 4 at the 10% level).

- Asian Female doctors (4 profiles): −\$135.2 to −\$180.8 without quality signals and −\$6.3 to \$25.6 with quality signals.

This suggests that my main results are largely robust to the doctor profile picture selection.

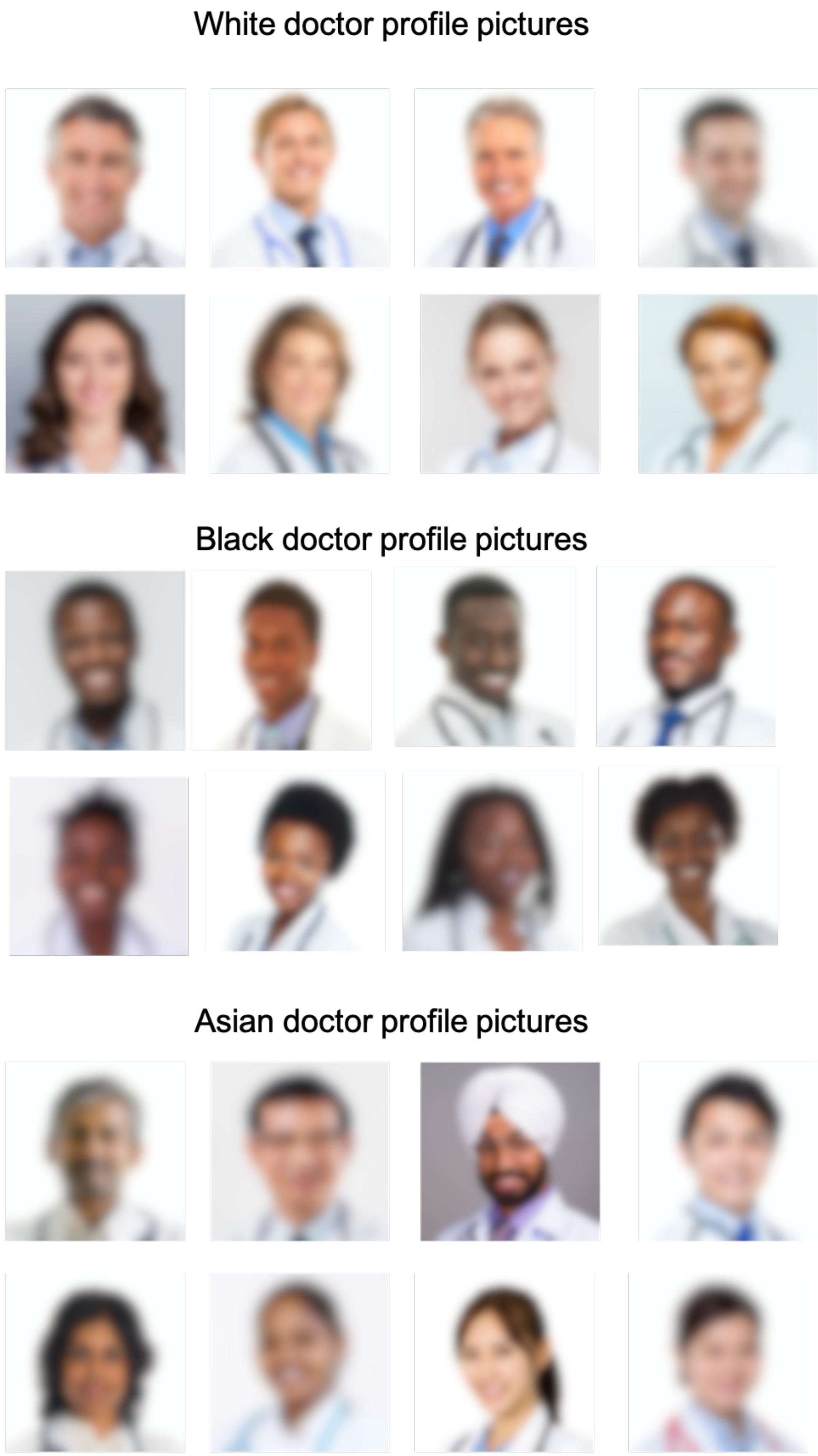


Figure 47: Blurred doctor profile pictures indicative of race

A.5 Customer Descriptions on Choice Approach

In this Section, I present some word clouds for the free text description submitted by the subjects from the control and treatment groups. This is an optional field so not every customer submitted text. “Price” is prominent in both control and treatment groups. The word cloud for the control group is shown in Figure 48 and the word cloud for the treatment group is shown in Figure 49.

Control

Figure 48: Word cloud based on customer description on their choice approach in the control group

Treatment

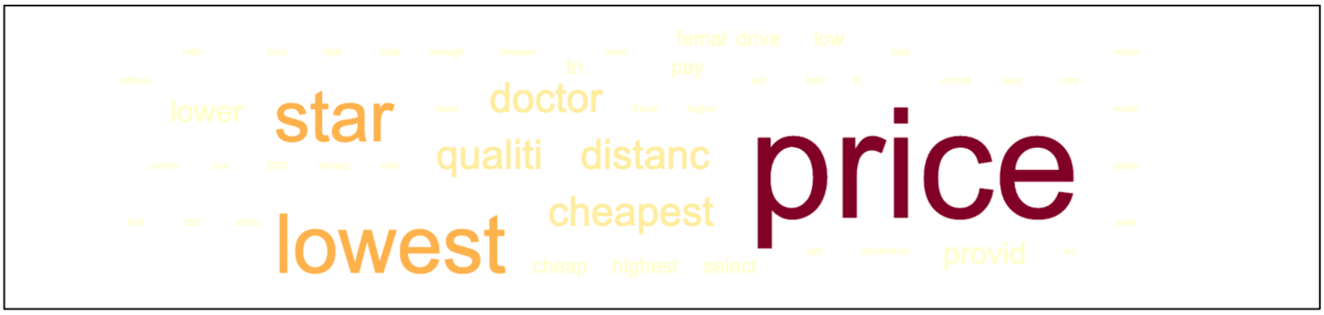


Figure 49: Word cloud based on customer description on their choice approach in the treatment group

A.6 Maximum Simulated Likelihood Estimators

In Section 2.5, I obtain choice probability of customer i choosing doctor d with the familiar multinomial logit form:

$$P_i(d) = \frac{\exp[\delta_{id} - \eta_i p_d]}{\sum_{j \in D \cup \{0\}} \exp[\delta_{ij} - \eta_i p_j]} \tag{20}$$

and I further assume that marginal utility δ_{id} is a linear function of doctor attributes:

$$\delta_{id} = \mathbf{a}_d^{'} \lambda_i \tag{21}$$

where $\mathbf{a}_d$ denotes the vector of attributes of doctor d including race, gender, distance, and quality (for treatment group). λ_i is the vector of attribute coefficients λ_{ia}. To simplify notation, I will call the vector of attribute coefficient plus the price coefficient λ and refer to the observable attributes and price of customer i's doctor option d in menu m as x_{idm}. I will also assume that this discrete choice model is parameterized with deep parameters θ.

I estimate a mixed logit model to allow the coefficients in the model to vary across customers. Considering that the customers make several choices (14 menus, denoted as $m \in \{1, 2, ..., 14\}$), the probability of a particular sequence of choices by customer i is given by:

$$S_i = \int \prod_{m=1}^{14} \prod_{d=0}^{D} [\frac{\exp[x_{idm}^{'} \lambda]}{\sum_{j \in D \cup \{0\}} \exp[x_{jdm}^{'} \lambda]}]^{y_{idm}} f(\lambda|\theta) d\lambda, \tag{22}$$

where $f(\lambda|\theta)$, which depends on deep parameters θ, is the density function of λ. And $y_{idm} = 1$ if customer i chose alternative d in menu m and $y_{idm} = 0$ otherwise.

The main empirical results and hypothesis testing are based on maximum simulated likelihood estimators (Ben-Akiva et al. (2019)). In the present paper, I assumed that marginal utility is a linear function of doctor attributes and the λ parameters are estimated by maximizing the simulated log-likelihood function using the data from customers 1 through N:

$$SLL = \sum_{i=1}^{N} \ln\{\frac{1}{R} \sum_{r=1}^{R} \prod_{m=1}^{14} \prod_{d=0}^{D} [\frac{\exp[x_{idm}^{'} \lambda_i^{[r]}]}{\sum_{j \in D \cup \{0\}} \exp[x_{jdm}^{'} \lambda_i^{[r]}]}]^{y_{idm}}\}, \tag{23}$$

where $\lambda_i^{[r]}$ is the r^{th} draw for customer i from the distribution of λ.

B Supplementary Tables

Table 4: Attribute Levels Used for Doctor Profiles in the Conjoint Surveys

Doctor Attribute	Levels	*Remarks*
Price	$1704.00	10^{th}*-percentile of actual prices in real doctor sample*
	$1797.07	
	$1890.13	
	$1983.20	
	$2076.27	
	$2169.33	
	$2262.40	
	$2355.47	
	$2448.53	
	$2541.60	90^{th}*-percentile of actual prices in real doctor sample*
Doctor Demographic	Black Male	*4 different profile pictures*
	Black Female	*4 different profile pictures*
	White Male	*4 different profile pictures*
	White Female	*4 different profile pictures*
	Asian Male	*4 different profile pictures*
	Asian Female	*4 different profile pictures*
Travel Distance	0-10 miles	*Distance b/w customer and provider zipcodes*
	10-50 miles	*Distance b/w customer and provider zipcodes*
	50-100 miles	*Distance b/w customer and provider zipcodes*
	100-250 miles	*Distance b/w customer and provider zipcodes*
	250 miles or more	*Distance b/w customer and provider zipcodes*
Comprehensive Quality Score	1 star	*Lowest quality level*
	2 stars	
	3 stars	
	4 stars	
	5 stars	*Highest quality level*

Notes: This table lists all the attributes and levels of the attributes that are used to design the conjoint survey doctor profiles. There are 6000 different attribute level combinations that are possible for the doctor profiles. Subjects cannot be exposed to every one of these profiles. In the experiment, I deploy allow considerable linearly independent variation in the levels of different attributes and a considerable span of attribute levels with an off-the-shelf fractional factorial choice design to optimize balance, overlap, and other characteristics.

Table 5: Comparing Facial Symmetry Across Race for Doctor Profile Pictures

	(1) Facial Symmetry	(2) Facial Symmetry
Black	0.375	0.375
	(7.291)	(7.303)
Asian	1.750	1.750
	(7.291)	(7.303)
Female		5.750
		(5.963)
N	24	24
R^2	0.003	0.047
White Average	59.250	59.250
Black Average	59.625	59.625
Asian Average	61.000	61.000

Standard errors in parentheses
$^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$

Notes: This table shows the balance on facial symmetry (as a proxy of race-neutral attractiveness) across races in the doctor profiles used in the conjoint surveys. The facial symmetry score is calculated based on the unblurred pictures. I estimate correlation between race of doctor profile picture and facial symmetry with the model $Symmetry_i = \beta_0 + \beta_1 1\{Black\} + \beta_2 1\{Asian\} + \epsilon_i$, in column (1). Standard errors for coefficients are given in parentheses. I control for gender of the doctor profile pictures in the estimation model for column (2).

Table 6: Comparing Subjective Attractiveness Ratings Across Race for Doctor Profile Pictures

	(1) Attractiveness	(2) Attractiveness
Black	0.113	0.113
	(0.423)	(0.409)
Asian	-0.273	-0.273
	(0.423)	(0.409)
Female		0.520
		(0.334)
N	24	24
R^2	0.040	0.144
White Average	5.870	5.870
Black Average	5.983	5.983
Asian Average	5.598	5.598

Standard errors in parentheses
$^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$

Notes: This table shows the balance on perceived age across races in the doctor profiles used in the conjoint surveys. The perceived age is based on a survey I conducted with uninsured Americans (N=50) via the platform Prolific where the respondents reviewed the blurred doctor pictures and was asked “In your opinion, how attractive is this doctor’s appearance?”. The survey respondents respondents responded via a Likert scale: 1 being “Not attractive at all” and 10 being “Extremely attractive”. I estimate correlation between race of doctor profile picture and perceived age with the model $Attractiveness_i = \beta_0 + \beta_1 1\{Black\} + \beta_2 1\{Asian\} + \epsilon_i$, in column (1), and the model $Attractiveness_i = \beta_0 + \beta_1 1\{Black\} + \beta_2 1\{Asian\} + \beta_3 1\{Female\} + \epsilon_i$, in column (2). Standard errors for coefficients are given in parentheses.

Table 7: Comparing Perceived Age Across Race for Doctor Profile Pictures

	(1) Perceived Age	(2) Perceived Age
Black	-5.913	-5.913
	(3.919)	(3.661)
Asian	-1.778	-1.778
	(3.919)	(3.661)
Female		-6.023
		(2.989)
N	24	24
R^2	0.103	0.254
White Average	42.680	42.680
Black Average	36.768	36.768
Asian Average	40.903	40.903

Standard errors in parentheses
$^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$

Notes: This table shows the balance on perceived age across races in the doctor profiles used in the conjoint surveys. The perceived age is based on a survey I conducted with uninsured Americans (N=50) via the platform Prolific where the respondents reviewed the blurred doctor pictures and was asked "How old do you think this doctor is?". I estimate correlation between race of doctor profile picture and perceived age with the model $Age_i = \beta_0 + \beta_1 1\{Black\} + \beta_2 1\{Asian\} + \epsilon_i$, in column (1), and the model $Age_i = \beta_0 + \beta_1 1\{Black\} + \beta_2 1\{Asian\} + \beta_3 1\{Female\} + \epsilon_i$, in column (2). Standard errors for coefficients are given in parentheses.

Table 8: Choice Model Parameter Estimates in Willingness-to-Pay With and Without Age Controls

Attribute	Control	Treatment	Control	Treatment
Black	-270.133	-25.803	-266.570	-29.687
	(9.164)	(8.739)	(9.109)	(9.809)
Asian	-185.529	-23.757	-183.142	-26.136
	(6.139)	(4.672)	(6.005)	(4.961)
Female	-15.598	0.529	-8.547	0.288
	(7.227)	(6.413)	(7.247)	(7.285)
10-50 miles	-37.409	-47.888	-38.694	-48.090
	(3.028)	(2.454)	(2.961)	(2.575)
50-100 miles	-117.994	-116.837	-113.595	-122.650
	(2.879)	(3.255)	(2.818)	(3.462)
100-250 miles	-262.882	-252.300	-258.252	-263.913
	(6.587)	(6.798)	(6.444)	(7.325)
More than 250 miles	-588.866	-551.620	-576.437	-564.991
	(13.133)	(16.004)	(13.190)	(17.353)
5 stars		71.981		70.597
		(2.920)		(3.096)
4 stars		69.167		67.423
		(3.641)		(3.855)
3 stars		16.312		15.902
		(2.994)		(3.157)
2 stars		6.612		6.306
		(2.516)		(2.649)
Perceived age			1.151	-0.466
			(0.697)	(0.613)
N	10,080	8,736	10,080	8,736

Notes: This table presents the estimated parameter results for the primary choice model from Section 2.5 expressed in terms of willingness-to-pay calculated as outlined in Section 2.5.1. The estimates are based on the data generated by the 224 customers as they evaluate the 14 menus of hypothetical doctor options (5 doctor options per menu) each. The perceived age is based on a survey I conducted with uninsured Americans (N=50) via the platform Prolific where the respondents reviewed the blurred doctor pictures and was asked "How old do you think this doctor is?" (see also Appendix Table 7). All willingness-to-pay coefficients are in dollar units with standard errors for parameters given in parentheses. Marginal utility is assumed to be a linear function of doctor attributes and the parameters represents the willingness-to-pay relative to the reference attribute level. For race, the reference level is white race for the doctor. For gender, the reference level is male. For distance, the reference level is "0-10 miles" distance between the doctor's zip code and the customer's zip code. For quality stars, 1 star is the reference level. The t-statistic is calculated as described in Section 2.5.1.

Table 9: Balance Check for Conjoint Survey Profiles

	Total		Control		Treatment		
	mean	s.d.	mean	s.d.	mean	s.d.	Difference p-value
Black	0.34	0.47	0.34	0.47	0.34	0.47	0.981
White	0.33	0.47	0.33	0.47	0.33	0.47	0.964
Asian	0.33	0.47	0.33	0.47	0.33	0.47	0.945
Female	0.50	0.50	0.50	0.50	0.50	0.50	0.976
Black Male	0.17	0.37	0.17	0.37	0.17	0.37	0.988
Black Female	0.17	0.37	0.17	0.37	0.17	0.37	0.963
White Male	0.17	0.37	0.17	0.37	0.17	0.38	0.740
White Female	0.16	0.37	0.17	0.37	0.16	0.37	0.695
Asian Male	0.16	0.37	0.16	0.37	0.16	0.37	0.696
Asian Female	0.17	0.37	0.17	0.37	0.17	0.38	0.635
Price	2122.79	266.95	2122.42	267.47	2123.22	266.37	0.852
Distance 0-10 Miles	0.20	0.40	0.20	0.40	0.20	0.40	0.954
Distance 10-50 Miles	0.20	0.40	0.20	0.40	0.20	0.40	0.961
Distance 50-100 Miles	0.20	0.40	0.20	0.40	0.20	0.40	0.902
Distance 100-250 Miles	0.20	0.40	0.20	0.40	0.20	0.40	0.901
Distance Over 250 Miles	0.20	0.40	0.20	0.40	0.20	0.40	0.915
Patients	224		120		104		
N	18,816		10,080		8,736		

Notes: This table reports the attribute levels of the 18,816 hypothetical profiles across 224 customers in the conjoint survey, pooled and by treatment group. “Black,” “Asian,” and “White” indicate the shares of doctor profiles belonging to each of these categories; “Female” indicates the share of female sex for the doctor profiles; likewise, “Black Male,” “Black Female,” “White Male,” “White Female,” “Asian Male,” and “Asian Female” indicate the shares of doctor profiles belonging to each of these categories. “Price” is the average price of the doctor profiles. The “Distance” rows indicate the shares of doctor profiles belonging to each of these distance categories. The Difference p-value column reports the p-value for the test of equality between the treatment and control groups. Stars indicate whether this difference is significant. As one would expect from a well executed conjoint design, I do not find any statistically significant differences for all these differences.

Table 10: Summary Statistics of Sample Versus U.S. Uninsured Under 65 and General Populations

	Experimental Sample	Uninsured Under-65	General Population
Female	0.46	0.45	0.51
Black	0.10	0.13	0.14
Asian	0.06	0.04	0.06
Hispanic	0.10	0.38	0.19
White	0.83	0.41	0.76
Age under 45	0.07	0.72	0.60^{+}
Age between 45 and 54	0.45	0.16	0.12^{+}
Age between 55 and 64	0.46	0.12	0.13^{+}
High School Graduate	0.80	0.76	0.89^{++}
College Graduate	0.29	0.13	0.33^{++}
Employed Full-time	0.60	0.73	0.77^{+++}
Employed Part-time	0.14	0.12	0.15^{+++}
Non-worker	0.26	0.15	0.08^{+++}
Uninsured	0.90	1.00	0.09
N	224	28.9M	331.5M

Notes: This table reports the background characteristics of the 224 customers in the main experiment sample, compared to two larger populations in the U.S: the under-65 uninsured population and the general population. The data on the uninsured population under-65 population come from the Kaiser Family Foundation 2019 data (https://www.kff.org/report-section/key-facts-about-the-uninsured-population-appendix/). The data on the general population come from the US Census (V2021) and the 2020 Current Population Survey from the Bureau of Labor Statistics. "Female" indicates the share of female sex; "Black," "Asian," "Hispanic," and "White" indicate the shares of people belonging to each of these categories. Age data was recorded in intervals, "Age under 45," "Age between 45 and 54," and "Age between 55 and 64" indicate shares of people in these age buckets. "High School Graduate" indicate the share of people who graduated from high school, "College Graduate" indicate indicate the share of people who reported that they have a bachelor's or associate degree. "Married" indicate the share of people who are currently married (not single, divored, separated, or widowed). "Currently employed" indicate the share who are employed part-time or full-time (not including self-employment while "Currently Self-employed" indicate the share who are self-employed. "Uninsured" indicates the share of people who are not covered by any health insurance plan. $^{+}$ from 2020. $^{++}$ as a percentage of the population over 25. $^{+++}$ as a percentage of the civilian labor force.

Table 11: Choice Model Parameter Estimates in Willingness-to-Pay Broken Down by Race Segments (Black and White)

Attribute	Black Patients		White Patients	
	Control	Treatment	Control	Treatment
a	$WTP_a^{C,Black}$	$WTP_a^{T,Black}$	$WTP_a^{C,white}$	$WTP_a^{T,white}$
Black	-77.808	21.425	-296.996	-29.386
	(25.554)	(16.819)	(11.028)	(8.752)
Asian	-101.558	-13.154	-198.336	-24.864
	(17.263)	(12.581)	(7.123)	(4.791)
Female	-22.694	-27.515	-16.428	3.783
	(22.839)	(15.671)	(8.610)	(6.543)
10-50 miles	-58.704	-34.627	-35.265	-49.459
	(10.540)	(6.558)	(3.837)	(2.390)
50-100 miles	-127.319	-111.704	-117.167	-117.029
	(13.293)	(16.653)	(3.462)	(3.584)
100-250 miles	-252.136	-269.950	-257.211	-253.020
	(24.383)	(32.056)	(7.089)	(7.537)
More than 250 miles	-595.366	-567.257	-577.198	-551.991
	(53.173)	(70.206)	(15.416)	(16.617)
5 stars		74.228		72.035
		(8.927)		(2.917)
4 stars		61.797		69.867
		(9.957)		(3.815)
3 stars		9.300		16.389
		(9.504)		(3.218)
2 stars		0.793		7.058
		(6.076)		(3.320)
N	1,176	756	8,064	7,476

Notes: This table reports the results for the 23 Black and 185 white customers in sample, by treatment group. This table presents the estimated parameter results for the primary choice model from Section 2.5 expressed in terms of willingness-to-pay calculated as outlined in Section 2.5.1. All willingness-to-pay coefficients are in dollar units with standard errors for parameters given in parentheses. Marginal utility is assumed to be a linear function of doctor attributes and the parameters represents the willingness-to-pay relative to the reference attribute level. For race, the reference level is white race for the doctor. For gender, the reference level is male. For distance, the reference level is "0-10 miles" distance between the doctor's zip code and the customer's zip code. For quality stars, 1 star is the reference level.

Table 12: Choice Model Parameter Estimates in Willingness-to-Pay Broken Down by Gender Segments (Female and Male)

Attribute	Female Patients		Male Patients	
	Control	Treatment	Control	Treatment
a	$WTP_a^{C,female}$	$WTP_a^{T,female}$	$WTP_a^{C,male}$	$WTP_a^{T,male}$
Black	-270.147	-32.374	-270.125	-19.494
	(13.210)	(12.160)	(14.410)	(10.589)
Asian	-193.795	-21.263	-180.106	-26.151
	(8.918)	(7.028)	(7.115)	(4.989)
Female	85.370	39.480	-81.852	-36.866
	(9.845)	(8.903)	(9.444)	(7.899)
10-50 miles	-55.739	-54.229	-25.381	-41.799
	(4.653)	(3.336)	(3.886)	(3.313)
50-100 miles	-122.466	-113.761	-115.060	-119.790
	(6.004)	(4.142)	(3.651)	(4.931)
100-250 miles	-292.051	-255.389	-243.743	-249.335
	(13.927)	(9.884)	(7.775)	(9.528)
More than 250 miles	-670.749	-568.394	-535.135	-535.517
	(29.230)	(23.010)	(14.738)	(19.739)
5 stars		69.831		74.045
		(4.610)		(3.378)
4 stars		70.319		68.060
		(5.809)		(4.145)
3 stars		16.728		15.913
		(4.074)		(3.977)
2 stars		14.708		-1.161
		(3.819)		(3.745)
N	4,368	4,284	5,712	4,452

Notes: This table reports the results for the 103 female and 121 male customers in full sample, by treatment group. This table presents the estimated parameter results for the primary choice model from Section 2.5 expressed in terms of willingness-to-pay calculated as outlined in Section 2.5.1. All willingness-to-pay coefficients are in dollar units with standard errors for parameters given in parentheses. Marginal utility is assumed to be a linear function of doctor attributes and the parameters represents the willingness-to-pay relative to the reference attribute level. For race, the reference level is white race for the doctor. For gender, the reference level is male. For distance, the reference level is “0-10 miles” distance between the doctor’s zip code and the customer’s zip code. For quality stars, 1 star is the reference level.

Table 13: Choice Model Parameter Estimates in Willingness-to-Pay Broken Down by Age Segments (Under 55 and 55 or Above)

Attribute	Under 55		55 or Over	
	Control	Treatment	Control	Treatment
a	$WTP_a^{C,U55}$	$WTP_a^{T,U55}$	$WTP_a^{C,O55}$	$WTP_a^{T,O55}$
Black	-256.620	-4.920	-276.848	-21.554
	(21.006)	(13.940)	(24.274)	(16.779)
Asian	-175.272	-15.398	-179.414	-21.421
	(14.172)	(9.157)	(16.300)	(12.949)
Female	-12.921	-6.166	-28.402	2.443
	(21.006)	(12.778)	(20.018)	(16.153)
10-50 miles	-36.835	-43.689	-33.277	-45.760
	(6.758)	(7.196)	(7.612)	(5.599)
50-100 miles	-109.237	-112.835	-114.248	-119.818
	(7.623)	(7.682)	(6.804)	(8.136)
100-250 miles	-242.990	-247.143	-247.994	-253.910
	(17.437)	(20.064)	(15.558)	(15.185)
More than 250 miles	-554.275	-513.908	-572.246	-574.347
	(36.672)	(64.182)	(26.028)	(22.346)
5 stars		72.011		64.764
		(3.238)		(8.842)
4 stars		80.635		69.385
		(3.670)		(4.878)
3 stars		15.718		19.189
		(6.116)		(6.220)
2 stars		5.466		4.793
		(5.398)		(3.901)
N	5,208	4,452	4,872	4,284

Notes: This table reports the results for the 115 customers under the age of 55 and 109 customers age 55 or over in full sample, by treatment group. This table presents the estimated parameter results for the primary choice model from Section 2.5 expressed in terms of willingness-to-pay calculated as outlined in Section 2.5.1. All willingness-to-pay coefficients are in dollar units with standard errors for parameters given in parentheses. Marginal utility is assumed to be a linear function of doctor attributes and the parameters represents the willingness-to-pay relative to the reference attribute level. For race, the reference level is white race for the doctor. For gender, the reference level is male. For distance, the reference level is "0-10 miles" distance between the doctor's zip code and the customer's zip code. For quality stars, 1 star is the reference level.

Table 14: Choice Model Parameter Estimates in Willingness-to-Pay Broken Down by Education-level Segments (College-Graduates and Non-College-Graduates)

Attribute	College Grads		Non-College Grad	
	Control	Treatment	Control	Treatment
a	$WTP_a^{C,college}$	$WTP_a^{T,college}$	$WTP_a^{C,NCollege}$	$WTP_a^{T,NCollege}$
Black	-186.605	21.827	-299.294	-36.143
	(36.376)	(13.956)	(17.835)	(13.970)
Asian	-138.233	-5.007	-195.165	-27.853
	(15.347)	(12.067)	(12.475)	(10.182)
Female	-13.787	2.365	-31.462	-8.049
	(24.990)	(13.919)	(17.682)	(13.887)
10-50 miles	-43.245	-41.508	-32.180	-47.186
	(9.222)	(4.475)	(5.176)	(5.452)
50-100 miles	-116.062	-103.777	-111.349	-126.981
	(8.975)	(4.994)	(5.724)	(8.567)
100-250 miles	-255.792	-219.944	-244.391	-269.112
	(20.389)	(15.642)	(11.944)	(17.756)
More than 250 miles	-572.246	-489.028	-564.243	-579.575
	(46.869)	(41.917)	(25.123)	(30.678)
5 stars		71.961		71.946
		(6.656)		(3.272)
4 stars		79.565		70.427
		(4.216)		(6.710)
3 stars		14.746		19.126
		(5.295)		(5.633)
2 stars		11.843		2.186
		(8.063)		(4.934)
N	2,940	2,604	7,140	6,132

Notes: This table reports the results for the 66 customers who has a college degree and 158 customers without a college degree in full sample, by treatment group. This table presents the estimated parameter results for the primary choice model from Section 2.5 expressed in terms of willingness-to-pay calculated as outlined in Section 2.5.1. All willingness-to-pay coefficients are in dollar units with standard errors for parameters given in parentheses. Marginal utility is assumed to be a linear function of doctor attributes and the parameters represents the willingness-to-pay relative to the reference attribute level. For race, the reference level is white race for the doctor. For gender, the reference level is male. For distance, the reference level is "0-10 miles" distance between the doctor's zip code and the customer's zip code. For quality stars, 1 star is the reference level.

Table 15: Choice Model Parameter Estimates in Willingness-to-Pay Broken Down by Zipcode Political Inclination (Those Who Voted for Republican Presidential Candidate and Those Who Voted for Democratic Candidate in 2022)

Attribute	Republican 2020		Democrat 2020	
	Control	Treatment	Control	Treatment
a	$WTP_a^{C,Rep}$	$WTP_a^{T,Rep}$	$WTP_a^{C,Dem}$	$WTP_a^{T,Dem}$
Black	-288.204	-19.733	-202.686	10.345
	(15.906)	(11.700)	(34.605)	(18.594)
Asian	-185.823	-20.550	-149.386	-15.048
	(10.337)	(6.567)	(21.362)	(18.594)
Female	-32.217	-7.603	27.503	12.611
	(13.673)	(11.506)	(27.950)	(17.893)
10-50 miles	-32.101	-47.196	-42.361	-42.208
	(5.823)	(4.337)	(8.802)	(6.968)
50-100 miles	-110.763	-119.165	-121.186	-105.731
	(6.080)	(5.495)	(7.112)	(13.484)
100-250 miles	-244.804	-246.009	-250.556	-264.345
	(12.725)	(14.622)	(25.724)	(22.400)
More than 250 miles	-554.620	-552.600	-592.275	-577.891
	(22.373)	(26.584)	(44.072)	(58.257)
5 stars		68.086		78.315
		(4.033)		(3.308)
4 stars		71.983		76.917
		(5.339)		(6.986)
3 stars		11.146		32.611
		(4.887)		(4.888)
2 stars		3.590		11.643
		(4.214)		(7.792)
N	7,560	6,384	2,520	2,352

Notes: This table reports the results for the 166 customers who lived in zipcodes that voted for the Republican candidate and 58 customers who lived in zipcodes that voted for the Democratic candidate, in the 2020 US Presidential Election, in full sample, by treatment group. This table presents the estimated parameter results for the primary choice model from Section 2.5 expressed in terms of willingness-to-pay calculated as outlined in Section 2.5.1. All willingness-to-pay coefficients are in dollar units with standard errors for parameters given in parentheses. Marginal utility is assumed to be a linear function of doctor attributes and the parameters represents the willingness-to-pay relative to the reference attribute level. For race, the reference level is white race for the doctor. For gender, the reference level is male. For distance, the reference level is "0-10 miles" distance between the doctor's zip code and the customer's zip code. For quality stars, 1 star is the reference level.

Table 16: Choice Model Parameter Estimates in Willingness-to-Pay Broken Down by Previous Colonoscopy Utilization (Those Who Had Colonoscopy Past 10 years and Those Who Did Not)

Attribute	Colonoscopy Last 10Y		No Colonoscopy Last 10Y	
	Control	Treatment	Control	Treatment
a	$WTP_a^{C,colo10Y}$	$WTP_a^{T,colo10Y}$	$WTP_a^{C,NoColo}$	$WTP_a^{T,NoColo}$
Black	-272.945	-6.340	-252.097	-15.331
	(19.956)	(11.759)	(52.329)	(17.937)
Asian	-181.601	-17.870	-166.346	-24.993
	(12.603)	(8.684)	(22.617)	(16.469)
Female	-17.142	1.377	-25.174	-9.255
	(17.127)	(11.391)	(42.096)	(17.925)
10-50 miles	-34.038	-47.650	-33.277	-38.144
	(7.626)	(5.064)	(9.900)	(5.967)
50-100 miles	-111.349	-112.038	-122.361	-121.036
	(6.135)	(7.250)	(11.673)	(17.214)
100-250 miles	-246.660	-242.793	-251.965	-266.769
	(14.050)	(15.301)	(17.805)	(39.559)
More than 250 miles	-564.243	-534.215	-575.370	-583.632
	(21.685)	(35.607)	(29.359)	(62.536)
5 stars		69.806		74.052
		(4.293)		(7.735)
4 stars		74.677		73.910
		(4.748)		(8.891)
3 stars		16.364		20.023
		(4.017)		(6.309)
2 stars		6.317		2.805
		(3.683)		(9.932)
N	8,064	6,720	2,016	2,016

Notes: This table reports the results for the 176 customers who have had a colonoscopy in the past 10 years and 48 who either did not have a colonoscopy or do not remember if they did have a colonoscopy in the past 10 years (39 did not, 9 did not remember or do not know) in full sample, by treatment group. This table presents the estimated parameter results for the primary choice model from Section 2.5 expressed in terms of willingness-to-pay calculated as outlined in Section 2.5.1. All willingness-to-pay coefficients are in dollar units with standard errors for parameters given in parentheses. Marginal utility is assumed to be a linear function of doctor attributes and the parameters represents the willingness-to-pay relative to the reference attribute level. For race, the reference level is white race for the doctor. For gender, the reference level is male. For distance, the reference level is "0-10 miles" distance between the doctor's zip code and the customer's zip code. For quality stars, 1 star is the reference level.

Table 17: Choice Model Parameter Estimates in Willingness-to-Pay Broken Down by Actual Booking Status (Booked and Not Booked)

Attribute	Booked		Did Not Book	
	Control	Treatment	Control	Treatment
a	$WTP_a^{C,female}$	$WTP_a^{T,female}$	$WTP_a^{C,male}$	$WTP_a^{T,male}$
Black	-276.576	-13.351	-229.867	12.009
	(18.458)	(10.166)	(47.340)	(15.809)
Asian	-182.337	-20.563	-161.397	-12.242
	(10.499)	(7.239)	(16.771)	(13.645)
Female	-22.530	-3.564	-0.419	5.466
	(14.965)	(9.162)	(38.330)	(14.799)
10-50 miles	-33.468	-44.459	-37.969	-48.824
	(6.287)	(5.709)	(10.050)	(10.949)
50-100 miles	-113.303	-117.307	-107.203	-112.835
	(5.581)	(5.692)	(16.100)	(15.564)
100-250 miles	-245.938	-250.602	-249.317	-241.584
	(11.155)	(18.600)	(33.159)	(36.830)
More than 250 miles	-561.328	-560.022	-582.317	-508.534
	(19.789)	(25.692)	(37.037)	(58.881)
5 stars		71.946		82.094
		(2.444)		(9.461)
4 stars		72.681		85.493
		(4.948)		(11.266)
3 stars		17.809		18.404
		(4.144)		(7.788)
2 stars		3.758		12.221
		(3.305)		(8.759)
N	8,232	7,560	1,848	1,176

Notes: This table reports the results for the 188 customers who eventually booked with one of the actual doctors offered at the end of the experiment and 36 who did not book in full sample, by treatment group. This table presents the estimated parameter results for the primary choice model from Section 2.5 expressed in terms of willingness-to-pay calculated as outlined in Section 2.5.1. All willingness-to-pay coefficients are in dollar units with standard errors for parameters given in parentheses. Marginal utility is assumed to be a linear function of doctor attributes and the parameters represents the willingness-to-pay relative to the reference attribute level. For race, the reference level is white race for the doctor. For gender, the reference level is male. For distance, the reference level is "0-10 miles" distance between the doctor's zip code and the customer's zip code. For quality stars, 1 star is the reference level.

Table 18: Choice Model Parameter Estimates in Willingness-to-Pay Using ONLY "Representative" Doctor Menus

Attribute	Control	Treatment	t-Statistic
a	$WTP_a^{Control,Rep}$	$WTP_a^{Treatment,Rep}$	
Black	-266.285	-8.079	-2.737
	(76.996)	(54.516)	
Asian	-179.371	-18.773	-2.508
	(49.653)	(40.444)	
Female	-2.598	-1.138	-0.018
	(60.062)	(54.291)	
10-50 miles	-33.277	-44.559	0.556
	(15.856)	(12.638)	
50-100 miles	-133.654	-116.473	-0.463
	(26.872)	(25.542)	
100-250 miles	-246.669	-248.994	0.036
	(53.812)	(35.921)	
More than 250 miles	-575.266	-558.086	-0.129
	(87.523)	(100.607)	
5 stars		71.954	
		(14.656)	
4 stars		74.677	
		(25.105)	
3 stars		18.106	
		(16.613)	
2 stars		4.885	
		(21.335)	
N	906	1,122	

Notes: This table presents the estimated parameter results for the primary choice model from Section 2.5 expressed in terms of willingness-to-pay calculated as outlined in Section 2.5.1. Instead of the 18,816 observations that come from 14 menus of doctor options for each of the 224 customers, the results in this data is estimated using only the subset of the data where the doctor menus are "representatives." A menu of hypothetical doctors is defined as representative if there is a majority of white doctors (at least 3 white doctors), at most 2 Asian doctors, at most 1 Black doctor, and at least 40% of doctors of each gender (ie. 3 white 1 Asian 1 Black, 3 White 2 Asian, 4 white 1 Black, 4 white 1 Asian menus with at least 2 male and 2 female doctors). There are 338 (151 for No Quality and 187 for Quality Signal) representative menus across 172 customers. The full set of menus across all customers is available in the online Data Appendix. All willingness-to-pay coefficients are in dollar units with standard errors for parameters given in parentheses. Marginal utility is assumed to be a linear function of doctor attributes and the parameters represents the willingness-to-pay relative to the reference attribute level. For race, the reference level is white race for the doctor. For gender, the reference level is male. For distance, the reference level is "0-10 miles" distance between the doctor's zip code and the customer's zip code. For quality stars, 1 star is the reference level. The t-statistic is calculated as described in Section 2.5.1.

Table 19: Choice Model Parameter Estimates in Willingness-to-Pay Broken Down by Believers versus Non-Believers

Attribute	Control	Treatment (sub-samples)	
		Believers	Non-Believers
a	WTP_a^C	$WTP_a^{T,Believ}$	$WTP_a^{T,NBeliev}$
Black	-270.133	-13.763	-155.009
	(9.164)	(8.188)	(37.968)
Asian	-185.529	-19.275	-71.850
	(6.139)	(4.401)	(19.695)
Female	-15.598	1.283	-7.563
	(7.227)	(6.262)	(29.227)
10-50 miles	-37.409	-46.538	-62.371
	(3.028)	(2.566)	(15.098)
50-100 miles	-117.994	-113.204	-155.825
	(2.879)	(3.014)	(20.520)
100-250 miles	-262.883	-250.106	-275.847
	(6.587)	(6.590)	(41.856)
More than 250 miles	-588.866	-542.791	-646.376
	(13.133)	(15.224)	(92.269)
5 stars		73.655	54.013
		(2.847)	(13.976)
4 stars		71.470	44.447
		(3.913)	(9.282)
3 stars		16.389	15.488
		(3.198)	(14.223)
2 stars		8.277	-11.259
		(2.997)	(11.522)
N	10,080	7,812	924

Notes: This table reports the results for the 224 customers in the main sample, by treatment group and by whether a treatment group subject "believes" the quality signal. Customers are categorized based their their response to the Likert scale question: "A provider with a higher Comprehensive Quality Score is a better provider than aprovider with a lower Comprehensive Quality Score" in the exit survey. Those who chose "Disagree" and "Strongly disagree" are non-believers of the signals and the rest "believers." This table presents the estimated parameter results for the primary choice model from Section 2.5 expressed in terms of willingness-to-pay calculated as outlined in Section 2.5.1. All willingness-to-pay coefficients are in dollar units with standard errors for parameters given in parentheses. Marginal utility is assumed to be a linear function of doctor attributes and the parameters represents the willingness-to-pay relative to the reference attribute level. For race, the reference level is white race for the doctor. For gender, the reference level is male. For distance, the reference level is "0-10 miles" distance between the doctor's zip code and the customer's zip code. For quality stars, 1 star is the reference level.

Table 20: Summary Statistics and Balance Checks for Customers Who Booked and Those Who Did Not

	Total		Didn't Book		Booked		
	mean	sd	mean	sd	mean	sd	Difference p-value
Female	0.46	0.50	0.53	0.51	0.45	0.50	0.37
Black	0.10	0.30	0.17	0.38	0.09	0.29	0.17
Asian	0.06	0.23	0.00	0.00	0.07	0.25	0.10
Hispanic	0.10	0.30	0.08	0.28	0.10	0.30	0.74
White	0.83	0.38	0.78	0.42	0.84	0.37	0.41
Age under 45	0.07	0.25	0.03	0.17	0.07	0.26	0.31
Age between 45 and 54	0.45	0.50	0.42	0.50	0.45	0.50	0.70
Age between 55 and 64	0.46	0.50	0.53	0.51	0.45	0.50	0.41
High School Graduate	0.80	0.40	0.83	0.38	0.80	0.40	0.63
College Graduate	0.29	0.46	0.36	0.49	0.28	0.45	0.34
Married	0.50	0.50	0.58	0.50	0.49	0.50	0.30
Currently employed	0.54	0.50	0.44	0.50	0.56	0.50	0.19
Currently Self-employed	0.20	0.40	0.22	0.42	0.20	0.40	0.73
Uninsured	0.90	0.30	0.86	0.35	0.91	0.29	0.37
No. of outside options evaluated	4.42	3.18	3.83	2.35	4.53	3.31	0.23
Price of outside option	2535	660	2430	636	2555	617	0.27
Colonoscopy in past 10 years	0.79	0.41	0.72	0.45	0.80	0.40	0.31
Quality Signal treatment	0.46	0.50	0.42	0.50	0.47	0.50	0.53
Actual Doctor Options:							
Pred. 1st Choice Price	1007	180	969	143	1014	185	0.18
Pred. Top 3 Price	1028	150	1003	102	1033	157	0.27
Pred. Top 8 Price	1134	93	1123	32	1136	101	0.44
Random Options Price	2212	253	2241	282	2207	247	0.46
Pred. 1st Choice Quality	3.89	1.31	4.33	1.05	3.82	1.34	0.16
Pred. Top 3 Quality	3.61	0.50	3.58	0.64	3.61	0.48	0.80
Pred. Top 8 Quality	3.67	0.25	3.69	0.25	3.67	0.25	0.77
Random Options Quality	3.05	0.87	3.43	0.53	2.99	0.90	0.07*
Pred. 1st Choice Female	0.07	0.25	0.08	0.28	0.06	0.25	0.67
Pred. Top 3 Female Share	0.06	0.18	0.06	0.17	0.06	0.18	0.98
Pred. Top 8 Female Share	0.17	0.21	0.30	0.28	0.15	0.18	0.00***
Random Options Female Share	0.28	0.30	0.24	0.33	0.28	0.30	0.38
Pred. 1st Choice Non-White	0.10	0.30	0.00	0.00	0.12	0.33	0.03**
Pred. Top 3 Non-White Share	0.12	0.20	0.10	0.19	0.12	0.20	0.54
Pred. Top 8 Non-White Share	0.33	0.11	0.37	0.12	0.33	0.11	0.02**
Random Options Non-White Share	0.22	0.28	0.22	0.25	0.22	0.29	0.98
Pred. 1st Choice Under 250 miles	0.42	0.49	0.25	0.44	0.45	0.50	0.03**
Any Pred. Top 3 Under 250 miles	0.48	0.50	0.25	0.44	0.53	0.50	0.00***
Any Pred. Top 8 Under 250 miles	0.52	0.50	0.25	0.44	0.57	0.50	0.00***
Any Random Options Under 250 miles	0.17	0.38	0.28	0.45	0.15	0.36	0.07*
Patients	224		36		188		

Notes: This table reports the background characteristics of the 224 customers in the main sample, pooled and by treatment group. "Female" indicates the share of female sex; "Black," "Asian," "Hispanic," and "White" indicate the shares of customers belonging to each of these categories. Age data was recorded in intervals, "Age under 45," "Age between 45 and 54," and "Age between 55 and 64" indicate shares of customers in these age buckets. "High School Graduate" indicate the share of customers who graduated from high school, "College Graduate" indicate the share of customers who reported that they have a bachelor's or associate degree. "Married" indicate the share of customers who are currently married (not single, divorced, separated, or widowed). "Currently employed" indicate the share who are employed part-time or full-time (not including self-employment while "Currently Self-employed" indicate the share who are self-employed. "Uninsured" indicates the share of customers who selected "NOT COVERED by any health insurance plan" in the survey. "# of outside options evaluated" indicates the number of colonoscopy providers that the customer has explored outside of the pilot's tool, while "Price of outside option" is the self-reported price that the customer expects to pay out-of-pocket (in USD). "Colonoscopy in past 10 years" indicates the share of customers who reported that they have had a colonoscopy in the past 10 years. "Colonoscopy in past 10 years" indicates the share of customers who reported that they have had a colonoscopy in the past 10 years. "Quality signal treatment" indicates the share who were assigned to the Quality Signal treatment the experiment. The final set of rows, in groups of four indicate the averages for (1) the predicted first choice in the menu of 10 doctors presented to the customer; (2) the predicted top 3 choices; (3) the predicted top 8 choices (all the choices assigned based on estimated individual preferences; and (4) the two random options inserted to the menu of 10, respectively. The first group of four row presents the average price, second group presents the average quality score (for the 104 customers in the Quality Signal treatment: 89 booked and 15 did not book), the third group presents the share of female doctors among the options, the fourth presents the share of non-white minorities doctors (Black or Asian) among the doctor options, and the fifth group presents the share of customers who have at least one option that is located less than 250 miles away among the actual doctors. Table shows averages ("mean") and standard deviations ("s.d."). The Difference p-value column reports the p-value for the test of equality between the treatment and control groups. Stars indicate whether this difference is significant.

Table 21: Comparing Choice Model Parameter Estimates in Willingness-to-Pay Between Choices Over Hypothetical Options and Choices Over Actual Doctors

Attribute	Hypothetical Doctor Options		Actual Bookings	
	(Control)	(Treatment)	(Control)	(Treatment)
Black	-270.133	-25.803	-314.221	-19.354
	(9.164)	(8.739)	(102.291)	(70.198)
Asian	-185.529	-23.757	-200.970	-29.784
	(6.139)	(4.672)	(60.396)	(52.225)
Female	-15.598	0.529	-21.784	-5.077
	(7.227)	(6.413)	(73.988)	(70.225)
10-50 miles	-37.409	-47.888	-38.220	-45.439
	(3.028)	(2.454)	(19.794)	(15.773)
50-100 miles	-117.994	-116.837	-107.701	-120.704
	(2.879)	(3.255)	(30.063)	(33.272)
100-250 miles	-262.882	-252.300	-242.476	-246.503
	(6.587)	(6.798)	(58.212)	(57.829)
More than 250 miles	-588.866	-551.620	-617.585	-613.207
	(13.133)	(16.004)	(98.722)	(111.37)
5 stars		71.981		73.399
		(2.920)		(18.126)
4 stars		69.167		74.410
		(3.641)		(31.445)
3 stars		16.312		16.545
		(2.994)		(26.985)
2 stars		6.612		4.320
		(2.516)		(23.765)
N	10,080	8,736	1,078	990

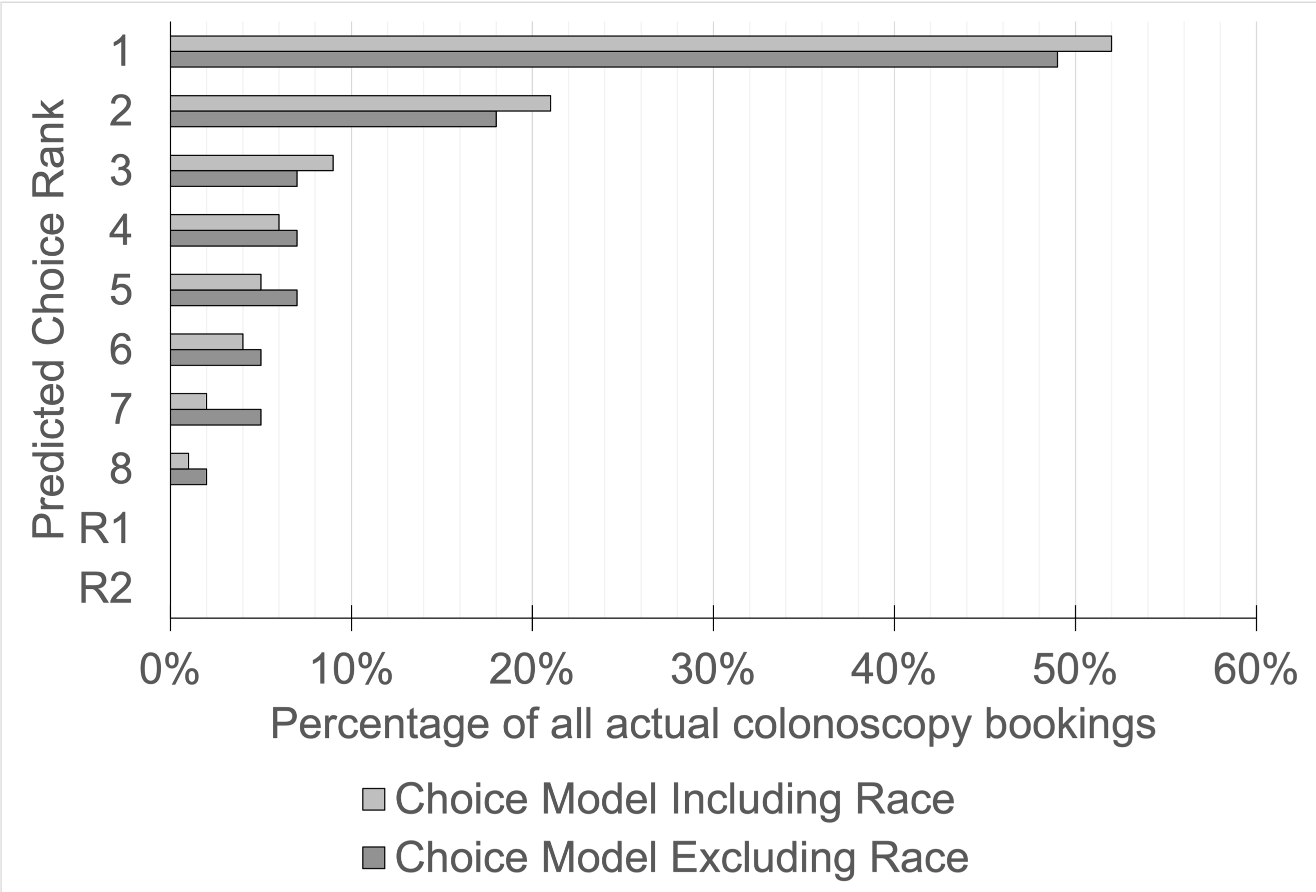
Notes: This table presents the estimated parameter results for the primary choice model from Section 2.5 expressed in terms of willingness-to-pay calculated as outlined in Section 2.5.1. The first two columns present the estimates using the data generated by the 224 customers as they evaluate the 14 menus of hypothetical doctor options (5 doctor options per menu) each, while the last two columns present the estimates using the data from the actual booking decision of customers who evaluated one menu (10 doctor options per menu) each. All willingness-to-pay coefficients are in dollar units with standard errors for parameters given in parentheses. Marginal utility is assumed to be a linear function of doctor attributes and the parameters represents the willingness-to-pay relative to the reference attribute level. For race, the reference level is white race for the doctor. For gender, the reference level is male. For distance, the reference level is "0-10 miles" distance between the doctor's zip code and the customer's zip code. For quality stars, 1 star is the reference level. The t-statistic is calculated as described in Section 2.5.1.

Table 22: The Quality Distributions for Asian and White Doctors

Quality Score	1 star	2 stars	3 stars	4 stars	5 stars
Asian Doctors	18.29%	23.17%	26.83%	20.73%	10.98%
White Doctors	15.57%	26.23%	19.26%	28.69%	10.25%
Representativeness for Asian vs. White $\frac{P(q\|Asian)}{P(q\|White)}$	1.17	0.88	1.39	0.72	1.07

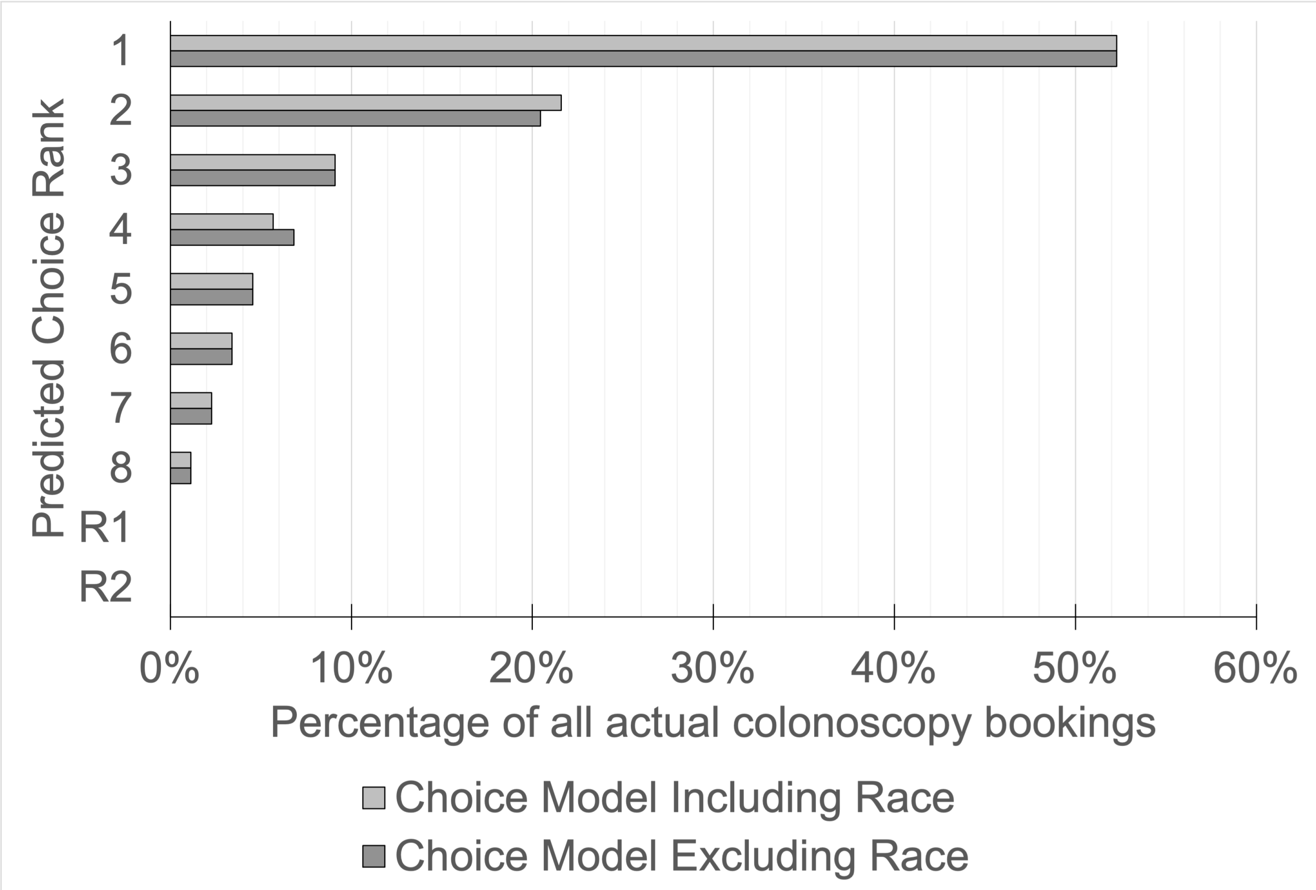
Notes: This table presents the empirical quality distribution of the actual doctors included in this study. There are totally 339 doctors offering colonoscopies, including 82 Asian doctors, 13 Black doctors, and 244 white doctors. The first two rows of this table presents the proportion of doctors at each level of comprehensive quality score, $P(q|Asian)$. The last row presents representativeness of each quality level for Asian race given comparison group white race, which is defined as the likelihood ratio $\frac{P(q|Asian)}{P(q|White)}$ (following Bordalo et al. (2016) and Gennaioli and Shleifer (2010)).

Figure 50: Actual Booking Choice by Predicted Choice Probability Using Individual Coefficients from Hypothetical Choice Data (No Quality Signal Treatment Customers only)



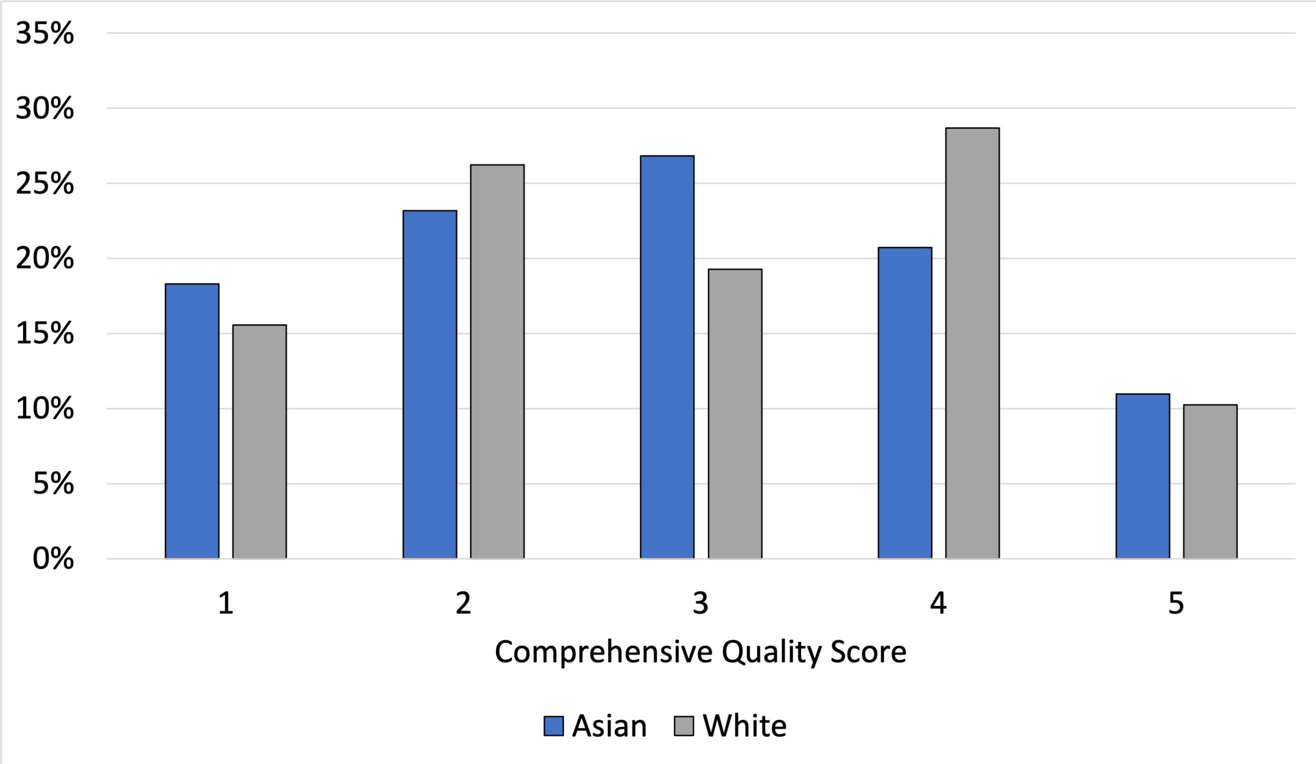
Notes: This figure presents the percentage of actual doctor booking by the rank (within the actual menu of 10 booking options) of doctor based on predicted choice probability individual coefficients estimated from choices made in the 14 hypothetical menus in the validated incentivized conjoint set-up. The lighter grey bar is based on the choice ranking predicted using the individual discrete choice model that includes all the doctor characteristics as predictive variables (price, gender, distance, and race for control group customers; price, gender, distance, quality, and race for treatment group customers); The lighter grey bar is based on the choice ranking predicted using the individual discrete choice model that includes all the doctor characteristics except race dummies as predictive variables (price, gender, and distance for control group customers; price, gender, distance, and quality for treatment group customers). Both models are estimated using the choice data from the 14 hypothetical menus, and the coefficients for each doctor characteristic is different between the models as the model specification is different. It reports the results for the 100 customers in the "No Quality Signal" treatment arm in the main sample who booked a doctor for colonoscopy. Individual coefficients are estimated with a mixed logit model using hierarchical Bayes estimation. The rank number 1 through 8 are the highest ranking doctors for each customer, R1 and R2 are two random doctors chosen from the remaining doctors in the pool and offered alongside the 8 to form the actual menu of 10 doctors.

Figure 51: Actual Booking Choice by Predicted Choice Probability Using Individual Coefficients from Hypothetical Choice Data (Quality Signal Treatment Customers only)



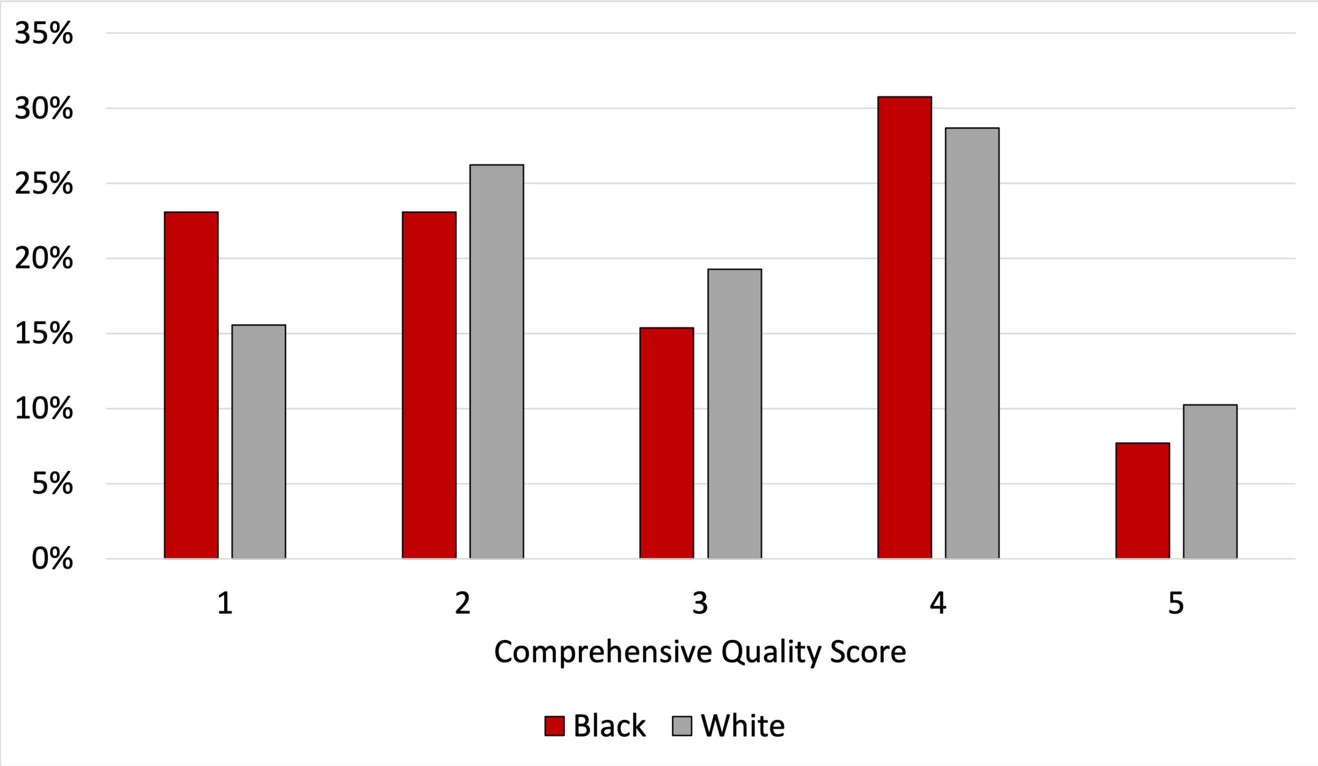
Notes: This figure presents the percentage of actual doctor booking by the rank (within the actual menu of 10 booking options) of doctor based on predicted choice probability individual coefficients estimated from choices made in the 14 hypothetical menus in the validated incentivized conjoint set-up. The lighter grey bar is based on the choice ranking predicted using the individual discrete choice model that includes all the doctor characteristics as predictive variables (price, gender, distance, and race for control group customers; price, gender, distance, quality, and race for treatment group customers); The lighter grey bar is based on the choice ranking predicted using the individual discrete choice model that includes all the doctor characteristics except race dummies as predictive variables (price, gender, and distance for control group customers; price, gender, distance, and quality for treatment group customers). Both models are estimated using the choice data from the 14 hypothetical menus, and the coefficients for each doctor characteristic is different between the models as the model specification is different. It reports the results for the 88 customers in the "Quality Signal" treatment arm in the main sample who booked a doctor for colonoscopy. Individual coefficients are estimated with a mixed logit model using hierarchical Bayes estimation. The rank number 1 through 8 are the highest ranking doctors for each customer, R1 and R2 are two random doctors chosen from the remaining doctors in the pool and offered alongside the 8 to form the actual menu of 10 doctors.

Figure 52: Quality Distribution for White and Asian Doctors



Notes: This figure presents the empirical quality distribution of the actual Asian and white doctors included in this study. There are totally 339 doctors offering colonoscopies, including 82 Asian doctors, 13 Black doctors, and 244 white doctors. The height of each bar represents the proportion of doctors at each level of comprehensive quality score, $P(q|Asian)$ or $P(q|White)$.

Figure 53: Quality Distribution for White and Black Doctors



Notes: This figure presents the empirical quality distribution of the actual Black and white doctors included in this study. There are totally 339 doctors offering colonoscopies, including 82 Asian doctors, 13 Black doctors, and 244 white doctors. The height of each bar represents the proportion of doctors at each level of comprehensive quality score, $P(q|Black)$ or $P(q|White)$.