

Examining How Patients Judge Their Physicians in Online Physician Reviews

by

Farrah Madanay

Public Policy
Duke University

Date: March 7, 2023

Approved:

Peter Ubel, Supervisor

M. Kate Bundorf

Aaron Kay

Richard Larrick

Dissertation submitted in partial fulfillment of
the requirements for the degree of Doctor
of Philosophy in Public Policy in the Graduate School
of Duke University

2023

ABSTRACT

Examining How Patients Judge Their Physicians in Online Physician Reviews

by

Farrah Madanay

Public Policy
Duke University

Date: March 7, 2023

Approved:

Peter Ubel, Supervisor

M. Kate Bundorf

Aaron Kay

Richard Larrick

An abstract of a dissertation submitted in partial
fulfillment of the requirements for the degree
of Doctor of Philosophy in Public Policy in the Graduate School of
Duke University

2023

Copyright by
Farrah Madanay
2023

Abstract

In three essays, this dissertation examines how patients judge their physicians in online physician reviews and whether those judgements align with traditional gender stereotypes. Specifically, I qualitatively explore patients' judgments of their physicians' interpersonal manner and technical competence, and the predominant factors within the two dimensions. I then train a machine-learning algorithm to code patients' judgments in online physician reviews at scale. Finally, I use the machine-coded sample to analyze physician gender differences in judgments received from patients and how those judgments affect physicians' review star ratings.

In Essay 1, I propose an elaborated theoretical framework to identify the predominant factors underlying patients' interpersonal manner and technical competence judgments of their physicians. This framework expands on prior grounded theory work by Lopez et al. (2012) and uses findings from a qualitative content analysis of 2,000 reviews received by distinct physicians. For this framework, I draw on a larger, new dataset of physician reviews from Healthgrades.com, one of the leading physician review websites, and use a balanced sample of reviews representing primary care physicians and surgeons, male and female physicians, and low- and high-rated reviews. I provide rich descriptions and illustrative quotations of the factors comprising interpersonal manner and technical competence, and describe factors added to and

removed from Lopez et al.'s original framework. This framework from Essay 1 demonstrates that patients value their physicians on a wide array of interpersonal manner and technical competence factors, including but not limited to bedside manner, going above and beyond, availability, knowledge, diagnostic skill, and open-mindedness about treatment.

In Essay 2, I train, test, and validate an advanced natural language processing algorithm called Robustly Optimized BERT Pre-Training Approach (i.e., RoBERTa) for classifying the presence and positive or negative valence of patients' interpersonal manner and technical competence judgments in online physician reviews. I use the 2,000 manually coded physician reviews from Essay 1 to train and test two classification models, one for interpersonal manner and one for technical competence. Both models perform with 90% accuracy, with high precision, recall, and weighted F1 scores. I validate the models using the full sample of 345,053 RoBERTa-coded reviews for 167,150 physicians by testing associations between the valence-coded judgments and review star ratings and by comparing review rating and gender analyses with extant results in the literature. The fine-tuned algorithm from Essay 2 allows us to code a large dataset of unstructured textual review data with high efficiency and accuracy, enabling subsequent large-scale text analysis.

In Essay 3, I analyze whether patients' judgments of their physicians' interpersonal manner and technical competence align with traditional gender stereotypes. Drawing on the Stereotype Content Model, I hypothesize that patients' judgments will conform with gender stereotypes, such that female physicians will be more likely to receive reviews with interpersonal manner judgments whereas male physicians will be more likely to receive reviews with technical competence judgments. Using the full sample of machine-coded reviews from Essay 2, I estimate multilevel logistic regressions to identify gender differences in interpersonal manner and technical competence judgments of physicians. Results from Essay 3 suggest that patients' judgments partly align with traditional gender stereotypes: Female physicians are more likely to receive interpersonal manner judgments, but male physicians are not more likely to receive technical competence judgments. Whether female physicians are relatively more likely to receive praise or criticism for their interpersonal manner depends on their specialty. In stereotypically warm specialties, like primary care, females are penalized for seeming cold, whereas in stereotypically technical specialties, like surgery, females are advantaged for appearing warm. Last, female physicians, in some cases, are either not rewarded as much or penalized more than their male counterparts in their star ratings when receiving positive or negative interpersonal manner and technical competence judgments.

Contents

Abstract	iv
List of Tables	xi
List of Figures	xii
Acknowledgments	xiii
1. Introduction	1
1.1 Patient-centered care	3
1.2 Healthcare consumerism	4
1.3 Measuring patient experience.....	5
1.4 Online physician reviews	6
1.5 Current research on online physician ratings and reviews.....	9
1.6 Two fundamental dimensions of social perception.....	11
1.7 Gender stereotypes along the two fundamental dimensions	13
2. Essay 1: How patients describe their physicians: Characterizing the interpersonal manner and technical competence dimensions of online physician reviews.....	16
2.1 Introduction.....	16
2.1.1 Conceptual development	19
2.2 Methods	22
2.2.1 Data collection.....	22
2.2.2 Sampling.....	22
2.2.3 Qualitative analysis.....	22

2.3 Results	23
2.3.1 Interpersonal manner.....	24
2.3.1.1 Interpersonal manner factors that aligned with those in Lopez et al.'s framework.....	26
2.3.1.2 Interpersonal manner factors added to Lopez et al.'s framework	28
2.3.1.3 Interpersonal manner factors originally in Lopez et al.'s framework that we excluded	35
2.3.2 Technical competence	38
2.3.2.1 Technical competence factors that aligned with those in Lopez et al.'s framework.....	40
2.3.2.2 Technical competence factors added to Lopez et al.'s framework	45
2.3.3 Limitations of the framework	48
2.4 Discussion.....	50
2.4.1 Conclusion.....	54
3. Essay 2: What patients like and dislike about their physicians: Developing and testing an algorithm to classify social judgments in online physician reviews	56
3.1 Introduction.....	56
3.2 Methods	58
3.2.1 Data collection.....	58
3.2.2 Coding reviews for interpersonal manner and technical competence	59
3.2.2.1 Hand-coding the training data.....	59
3.2.2.2 Training and testing the algorithm.....	60
3.3 Results	62

3.3.1 Evaluating the accuracy of the two classification models	62
3.3.2 Comparing reviews coded by RoBERTa and by hand.....	63
3.3.3 Testing the validity of the two classification models	66
3.3.3.1 Testing for associations between valence and star ratings	67
3.3.3.2 Testing whether the models reproduce prior findings.....	68
3.4 Discussion.....	70
3.4.1 Principal results	70
3.4.2 Limitations.....	74
3.4.3 Conclusion.....	75
4. Essay 3: Love them or hate them, female physicians’ personalities matter: A large-scale text analysis of online physician reviews.....	77
4.1 Introduction.....	77
4.2 Methods	82
4.2.1 Data collection.....	82
4.2.2 Classifying reviews for interpersonal manner and technical competence.....	83
4.2.3 Statistical analysis.....	84
4.3 Results	85
4.3.1 Characteristics of study sample.....	85
4.3.2 Gender differences in review judgments among physicians overall.....	89
4.3.3 Gender differences in review judgments among primary care physicians and surgeons.....	91
4.3.3.1 Primary care physicians.....	91

4.3.3.2 Surgeons	91
4.3.4 Gender differences in review star ratings among primary care physicians and surgeons who receive social judgments	93
4.3.4.1 Primary care physicians	93
4.3.4.2 Surgeons	93
4.4 Discussion.....	96
4.4.1 Finding 1: Patients were equally likely to judge male and female physicians’ technical competence	97
4.4.2 Finding 2a: Patients were more likely to judge their female physicians’ interpersonal manner.....	99
4.4.3 Finding 2b: How patients judged female physicians’ interpersonal manner depended on physicians’ specialties.....	99
4.4.4 Finding 3: Patients were less likely to reward, more likely to penalize female physicians in ratings based on interpersonal manner and technical competence ..	103
4.4.5 Limitations.....	104
4.4.6 Conclusion.....	105
5. Conclusion	107
Appendix for essay 3.....	111
References	115
Biography.....	132

List of Tables

Table 1 Interpersonal manner factors and illustrative quotations	33
Table 2 Excluded interpersonal manner factors, reasoning, and illustrative quotations .	37
Table 3 Technical competence factors and illustrative quotations.....	46
Table 4 Tuned transformer classification performance for interpersonal manner and technical competence judgments	62
Table 5 Illustrative examples of discrepancies in reviews coded by RoBERTa and by hand and reasoning underlying the discrepancies	64
Table 6 Physician characteristics.....	87
Table 7 Characteristics of physicians with and without reviews.....	88
Table 8.....	111
Table 9.....	112
Table 10.....	113

List of Figures

Figure 1 Framework of the predominant factors comprising patients' interpersonal manner and technical competence judgments in online physician reviews	50
Figure 2 Sample selection flow chart.....	59
Figure 3 Diagram showing how real physician reviews were hand-coded for the presence and valence of interpersonal manner and technical competence.....	60
Figure 4 Mean review star ratings for reviews with positive, negative, or no interpersonal manner or technical competence.....	68
Figure 5 Distribution of review star ratings.....	86
Figure 6 Percentage of reviews with positive and negative interpersonal manner judgments and positive and negative technical competence judgments.....	89
Figure 7 Odds ratio of receiving a social judgment for female physicians relative to male physicians.....	90
Figure 8.....	92
Figure 9.....	95

Acknowledgments

First, thank you, Peter, for being the best advisor I could have asked for during my PhD program. You epitomize a good mentor, always generous with your time, patient in allowing me to think through research questions and problems, and supportive of my ideas and decisions. What a joy to have learned how to do research, to teach, and to mentor from one of the most creative, curious, and compelling storytellers I know. I aspire to be as great of a mentor to others as you have been to me.

I am also indebted to my other committee members and mentors. Kate, you let me hound you with questions before you set foot at Duke, and to this day, I appreciate every opportunity to think through research with you. Aaron, thank you for letting me lean on your psychology expertise to introduce me to new theories, and to help me mull over details of experiments and measures across different projects. Rick, I was bitten by the JDM bug when I took your class my first semester, and I feel so fortunate to have benefitted from your perspective on my research ever since. Jenn and everyone in Lerner Lab, I cannot express how much I looked forward to lab each week, where I learned how to not only conduct sound experimental research but also think through theoretical and practical implications, present posters and talks, and ask thoughtful questions. Last, special thanks to Charlene Wong, who set me on the path of focusing on a PhD in health

policy. How could I not want to be in this space when I knew people like you were already here pursuing important questions with genuine passion?

Thank you to everyone who provided me with their time, feedback, and support on this dissertation. Ada Campagna, Karissa Tu, and Kelly Davis helped me manually code thousands of physician reviews and were invaluable collaborators in developing an elaborated theory. Felicia Chen helped me scrape the physician profile, rating, and review data from Healthgrades.com. John Little from Duke Libraries was critical in helping me understand how to manipulate my large dataset in R. Lastly, Andrew Trexler was kind enough to provide me with resources and guidance on how to use RoBERTa.

Finally, thank you to my friends and family. I would not have crossed the finish line in this program without my PhD cohort. Special thanks to my “accountabilibuddies,” Becca Daniels, Jane Leer, and Sarah Petry. You all were lifelines during the COVID-19 pandemic, and I am lucky to now count you as my close friends. Lee, you are a true partner. You are supportive and patient, my favorite sounding board for ideas and hypotheses, and my not-so-secret weapon for anything Excel and PowerPoint. I am so grateful to have you by my side, whether quarantining together with Callie, coaching the Unicorns, or traveling the world. Ann, John, Barry, and Dee, what a gift that with Lee came the best in-laws I could imagine. Thank you for the

support, the dinner conversations, and the extended Dallas stays. Last, Mom, Dad, Spencer, Jill, Grandma, and Grandpa, I would not be where I am or who I am today without decades of your unconditional love and support. You supported me through every twist and turn of my educational career, from my interest in art history to Holocaust studies to health policy. I may not have always clearly communicated the research I was doing, but you always asked questions and tried to understand.

1. Introduction

“She is very understanding, and caring. She took her time to explain things to me.”

-Patient of internist, 5 stars

“He performed a hernia surgery and all went very well. I was thoroughly satisfied with the procedure and recovery.”

-Patient of general surgeon, 5 stars

“Shes definitely more interested in her Minolos than she is in patients!”

-Patient of orthopedic surgeon, 1 star

“Play games with your medications. Scared of this guy.”

-Patient of family medicine physician, 1 star

When patients need to find and choose a new physician, over half turn to the Internet for help (Hanauer et al., 2014; Holliday et al., 2017; PressGaney, 2021). Patients rely on the Internet just as much or more than referrals from their healthcare provider and word of mouth (Kullgren et al., 2021; PressGaney, 2021). Patients report ratings and reviews as the most important factors to consider before choosing a physician (Hanauer et al., 2014). Since 2019, patient usage of healthcare review sites has increased by over 50% (PressGaney, 2021). Of all websites patients use to search for healthcare, Healthgrades is ranked fourth, after Google, hospital and clinic websites, and WebMD. Patients trust the accuracy of commercial online reviews more than both government ratings and traditional patient experience surveys (Hanauer et al., 2014; Holliday et al., 2017; Yaraghi et al., 2018). Although physicians’ star ratings are a key determinant in

choosing a physician, prospective patients also value the quality and helpfulness of physicians' reviews (Carbonell & Brand, 2018; PressGaney, 2021). Thus, what patients say about their physicians online should not be ignored. Online physician reviews empower patients to share their experiences (Holliday et al., 2017), influence prospective patients' decisions in selecting a physician (Burkle & Keegan, 2015b), and impact physicians' quality improvement efforts (Emmert et al., 2016).

In this dissertation, I analyze online physician reviews to better understand patients' quality judgments of their patient-physician interactions, and to discern whether patients' judgments differ based on their physicians' gender and specialty. In the following sections, I first review the literature on two prominent movements in healthcare that center patient perspectives: patient-centered medicine and healthcare consumerism. I discuss the advent of traditional patient-experience surveys and their commercial online review counterparts and trace the various streams of research on online physician reviews. I then introduce theory from social psychology on the two fundamental dimensions of social perception: warmth and competence. I examine how these dimensions relate to patient-physician interactions and undergird patients' gender stereotypes of their physicians.

Last, I present three essays exploring patients' judgments of their physicians in online reviews. In essay 1, I build on prior grounded theory with a qualitative content

analysis to develop a theoretical framework characterizing how patients describe their physicians. In essay 2, I use the qualitative coding from essay 1 to train, test, and validate an advanced machine-learning algorithm to code a large sample of reviews at scale with high accuracy, precision, and recall. Finally, in essay 3, I use the machine-coded reviews from essay 2 to conduct a large-scale text analysis to identify gender differences in patient judgments of their physicians.

1.1 Patient-centered care

The origins of patient-centered care date back to the 1950s, when U.S. psychologist Carl Rogers coined the concept to describe a trusting therapist-patient relationship (Latimer et al., 2017). Decades later, in 2001, the Institute of Medicine (IOM) enshrined patient-centeredness as a core principle of the U.S. healthcare system by naming patient-centered care as one of six aims for quality improvement in its report, *Crossing the Quality Chasm: A New Health Care System for the 21st Century* (Wolfe, 2001). The Agency for Healthcare Research and Quality (AHRQ) describes the IOM's six aims as one of the most influential frameworks for assessing quality in healthcare (AHRQ, 2022b). In addition to patient-centeredness, the aims include safety, effectiveness, timeliness, efficiency, and equity.

Patient-centered physicians provide care that is respectful of their individual patients' needs, values, and preferences (Wolfe, 2001). This concept of care shifted the

patient-physician relationship from one of paternalism, in which physicians make decisions for their patients, to one of partnership, with shared decision making (Kaba & Sooriakumaran, 2007; Park et al., 2022). The subsequent empowerment of patients in their own healthcare has improved patient outcomes and quality of life, reduced disparities in care, and increased value (Epstein et al., 2010).

1.2 Healthcare consumerism

In parallel to the patient-centeredness movement, healthcare consumerism took center stage in the 1980s and 1990s as employers used health maintenance organizations (HMOs) to control the costs of care (Thompson & Cutler, 2010). Despite slowing costs to employers, employee backlash to managed care restrictions clarified that patients, when sheltered from costs, lacked awareness of the implications of their healthcare decisions. The healthcare consumerism movement thus advocates for more informed patient decision making to curb demand for low-value services.

Healthcare consumers are characterized as patients who proactively use credible information and appropriate technology to make knowledgeable decisions about their healthcare (Carman, 2019). As a result, in 2010, Congress authorized the Patient-Centered Outcomes Research Institute (PCORI) to fund research that helps patients make more informed healthcare decisions. In the healthcare market, patients can now shop for health plans on health insurance marketplaces (Wong et al., 2019), receive

personalized out-of-pocket estimates for shoppable healthcare services (Kullgren & Fendrick, 2021), and search and compare physicians before scheduling an appointment (Murphy et al., 2019).

Healthcare consumerism and patient-centered care conceptually overlap with their emphasis on patient empowerment and decision making (Latimer et al., 2017). As a result, the expectation of both movements is for patient-centered physicians to be responsive to the needs of their patients, and patients to be informed and active co-participants in their healthcare decisions.

1.3 Measuring patient experience

Key to both patient-centeredness and healthcare consumerism movements is measuring patients' perspectives of their healthcare experiences. A year after the IOM published its "Quality Chasm" report, the Centers for Medicare and Medicaid (CMS) partnered with AHRQ to create the first standardized, national, survey of patients' perspectives of their hospital experiences. CMS implemented the survey, Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS), in 2006 and publicly reported the first results in 2008 (CMS, 2021). In 2007, AHRQ additionally released the CAHPS Clinician & Group Survey (CG-CAHPS) to measure patients' experiences with their primary and specialty care providers (AHRQ, 2015).

From a patient-centeredness perspective, the HCAHPS and CG-CAHPS surveys at once hold providers accountable and incentivize them to improve their quality of care. From a healthcare consumer perspective, the publicly reported CAHPS survey results allow patients to make meaningful comparisons between hospitals and physicians. For both movements, HCAHPS and CG-CAHPS empower patients to provide feedback on their patient experiences and offer trustworthy information to guide prospective patients' decision making.

As healthcare shifts toward value-based care, patient experience has become an increasingly important measure to benchmark quality. As such, CMS now ties Medicare value-based care reimbursements to the HCAHPS survey (Ranard et al., 2016). Additionally, some alternative payment models, such as the Pioneer Accountable Care Organization (ACO) and the Medicare Shared Savings Plan ACO, tie performance-based payments to CG-CAHPS results (*Frequently Asked Questions About CGCAHPS*).

1.4 Online physician reviews

The CAHPS surveys have enabled patients to compare and make decisions about their providers based on a standard set of topics. However, not all patients may be invited to take a CAHPS survey after their healthcare encounter (Health), and survey results may only be publicly available after the physician reaches a threshold of survey responses (DukeHealth). For example, Wake Forest Baptist Health only distributes 25

outpatient CG-CAHPS surveys per month per provider to a random sample of patients, and reports a return rate of 18% (Health). Although Duke Health reports a 30% return rate of its CG-CAHPS surveys, resulting in feedback from 120,000 patients per year, Duke only displays star ratings for providers who have received over 30 responses in the past 24 months (DukeHealth).

Commercial physician review websites complement traditional survey reporting with free, publicly accessible, crowdsourced ratings and reviews (Bardach et al., 2013; Schlesinger et al., 2015). Many of these commercial sites were born out of a desire to provide quality information to patients online. In 2004, before the advent of HCAHPS, Jeremy Stoppelman co-founded Yelp when he came down with a bad cold and had difficulty finding a local physician online (Richmond, 2012). Although Yelp is predominantly used by consumers to submit and read reviews of local businesses, the platform has become the most widely used, free, commercial website for hospital ratings in the U.S (Bardach et al., 2013; Ranard et al., 2016). Other commercial physician review websites have proliferated on the Internet, such as Healthgrades, Vitals, RateMDs, and ZocDoc. Established in 1998, Healthgrades became the leading comprehensive physician review website in 2014. Healthgrades populates its site using sources including claims data, crowdsourced patient surveys, and active U.S. physician profiles from the National Provider Identifier Registry (Furnas et al., 2020; Healthgrades, 2022).

As with traditional patient-experience surveys, there are key limitations to online physician reviews. Negative reviews, whether from actual or fake patients, may jeopardize a physician's reputation and patient acquisition (Frost, 2020). Yet, physicians are hindered from responding to patients because of threats to both HIPAA and trust in the physician-patient relationship (Lee, 2017). Despite physicians' concern over the consequences of negative reviews (Frost, 2020; Murphy et al., 2019), researchers have shown online reviews are overwhelmingly positive (Emmert et al., 2013; Liao et al., 2020; López et al., 2012). Physicians can also use the feedback from these reviews to improve their care, communication, workflow, and derive quality improvement measures (Emmert et al., 2016; Lee, 2017; Murphy et al., 2019).

Another limitation of online physician reviews is representativeness. Research suggests online physician reviews are predominantly written by young, educated, and affluent consumers (López et al., 2012). One study found negative online ratings were more common among physicians with more years of experience (Okike et al., 2019), which may reflect findings that younger patients, who are more likely to leave online reviews, prefer younger physicians (McKinstry & Yang, 1994). Although online physician reviews are increasing in popularity, research suggests whereas 37% of Internet users have rated goods and services online, only 3–4% have rated physicians or hospitals (Fox & Duggan, 2013). Further, one study estimated that in 2010, only 1 in 6

practicing U.S. physicians had received an online review on RateMDs (Gao et al., 2012). Although the views of a few unverified patients who write online reviews are not representative of all patient perspectives, most patients support posting narrative comments online and consider physician review websites important resources when selecting a physician (Hanauer et al., 2014; Holliday et al., 2017). Thus, it is critical to understand what information commercial reviews convey about the patient experience and how this information potentially influences prospective patient decisions.

1.5 Current research on online physician ratings and reviews

Much of the research exploring online physician ratings and reviews can be divided into four streams. In the first stream, researchers have studied whether commercial online ratings correlate with traditional patient-experience surveys. These studies generally find low-to-weak correlations between commercial ratings and traditional patient-experience surveys, such as the Press Ganey CG-CAHPS survey (Chen et al., 2018; Ryan et al., 2016; Widmer et al., 2018). However, these studies are typically conducted with small sample sizes and limited to either a single subspecialty or a single institution.

A second stream of research focuses on internal commercial rating associations between individual attributes and the global score. These studies find high, positive correlations between individual attribute ratings and the overall rating, which typically

comes from the question, “how likely would you recommend this provider?” (Bakhsh & Mesfin, 2014; Kadry et al., 2011a). For example, one study of orthopedic surgeons found five variables across four commercial websites were significantly correlated with higher overall ratings: ease of scheduling, time spent with patient, wait time, surgeon proficiency, and bedside manner (Bakhsh & Mesfin, 2014).

A third stream of research inquires whether online ratings are correlated with clinical outcomes. So far, this research has demonstrated mixed results. For example, one study found that compared to highly rated cardiac surgeons (≥ 4 stars), lower-rated cardiac surgeons were significantly associated with higher in-hospital mortality rates (Lu & Rui, 2018). However, the authors could not disentangle whether these lower-rated surgeons were truly less proficient or were physicians that elected to take on the most difficult cases, which may have resulted in poor reviews. Another study found no associations between online specialist ratings and objective quality-of-care and peer-assessment measures (Daskivich et al., 2018). However, the researchers found consistency in ratings across commercial platforms, which they suggested may mean these sites measure a latent construct unrelated to performance.

Understanding what this latent construct is leads to a fourth stream of research, which considers whether commercial physician reviews reveal other important constructs not explicitly captured by ratings. Most research within this stream focuses

on identifying what patients discuss in their narrative, unstructured comments. A qualitative analysis of reviews for female pelvic physicians found themes such as empathetic communication, accurate diagnosis, and effective management of urological conditions were associated with positive reviews (Asanad et al., 2018). Among spine surgeons, favorable reviews were related to surgical outcomes and surgeon likability, whereas negative reviews highlighted staff interactions and billing (Donnally III et al., 2018). One study comparing Yelp reviews of hospitals to HCAHPS attributes found several themes within Yelp narratives that were not present in HCAHPS, including cost of a hospital visit, insurance billing, and quality of nursing (Ranard et al., 2016). Last, a grounded theory analysis of Yelp and RateMDs reviews for primary care physicians identified four main themes: global impressions, interpersonal manner, technical competence, and system issues (López et al., 2012). This dissertation is situated within this fourth stream of research and asks two questions: how do patients judge their patient-physician interactions along the dimensions of interpersonal manner and technical competence and do those judgments align with gender stereotypes?

1.6 Two fundamental dimensions of social perception

I build on the previous grounded theory work by Lopez et al. (2012) as well theory from social psychology on how people form spontaneous impressions, judgments, and stereotypes of others. Lopez et al. identify two categories of patient

comments that relate to physicians: interpersonal manner and technical competence. The authors describe interpersonal manner as a physician's attentiveness, communication, and characteristics whereas technical competence refers to a physician's aptitude, skills, and thoroughness (López et al., 2012). These two dimensions align with the two fundamental dimensions of social perception from social psychology, warmth and competence, which underlie perceptions of others and drive group stereotypes (Cuddy et al., 2008). Prior research shows patient-centered care also requires both warmth and competence to influence patient adherence and self-care (Hall et al., 2014; Howe et al., 2019).

Extensive psychological research with people across multiple cultures shows that we universally use warmth and competence to sort and judge others (Cuddy et al., 2008; Fiske et al., 2007). Warmth traits include characteristics like friendliness, trustworthiness, and kindness; competence traits include intelligence, efficacy, and skill (Fiske et al., 2007). Although these dimensions are typically seen at odds, especially in comparative contexts (Judd et al., 2005), together they account for nearly 90% of the variance in people's judgments of others' social behaviors (Abele & Wojciszke, 2007).

Over the decades, psychologists have conceptualized these two fundamental dimensions by various nomenclatures. In the 1940s, Solomon Asch theorized that people form warm/cold and intelligent/industrious first impressions of others' personalities

(Asch, 1946). More recently, Cuddy, Fiske, and Glick's Stereotype Content Model uses warmth and competence as the central dimensions of group stereotypes (Cuddy et al., 2008), whereas Abele and Wojciszke's Dual Perspective Model describes agency and communion as the key dimensions of perceptions of self vs. others (Abele & Wojciszke, 2014). Across the different conceptualizations of the two dimensions, warmth is primary: people are more sensitive to warmth traits and make warmth judgments faster than competence judgments (Abele & Wojciszke, 2014; Cuddy et al., 2008). In this dissertation, I situate my examination of patients' judgments of their physicians within psychology's theorizing about the warmth and competence dimensions of social perception. However, I will use the broadened framework of interpersonal manner and technical competence consistent with prior research on patient reviews (López et al., 2012).

1.7 Gender stereotypes along the two fundamental dimensions

Researchers have shown warmth and competence are central dimensions of group stereotypes. Social Role Theory posits that the origins of social structural relations delineate feminine and masculine orientations (Eagly et al., 2000). These gender stereotypes follow from historical gendered divisions of labor, with women as homemakers and men as employees. Thus, we infer traits of women and men from their role behaviors, such that women are perceived as warm, communal, nurturing mothers

who take care of the home while men are competent, agentic leaders in the workplace. These gender stereotypes not only describe how women and men are (descriptive norms) but how women and men should be (prescriptive norms). Role Congruity Theory thus follows that individuals whose social roles and behaviors violate their gender stereotypes receive backlash (Eagly & Karau, 2002). Social Role Theory, however, postulates that stereotypes are dynamic and may change as roles in society shift (Diekmann & Eagly, 2000).

The Stereotype Content Model complements Social Role Theory by advancing a broader approach linking warmth and competence to group stereotypes (Cuddy et al., 2008). According to this model, most groups receive ambivalent stereotypes, with a positive evaluation of one dimension and a negative evaluation of the other. For example, women are stereotypically high in warmth but low in competence whereas men are stereotypically high in competence but low in warmth. These ambivalent stereotypes apply to other outgroups, such as model minorities and welfare recipients (Cuddy et al., 2011), which are beyond the scope of this dissertation.

Few studies of online physician reviews have examined how patient judgments align with gender stereotypes. Some studies have shown female physicians are more likely described with interpersonal language than males (Chen et al., 2021; Dunivin et al., 2020; Thawani et al., 2019), whereas two studies have additionally found male

physicians are described with more positive technical words (Haynes et al., 2021; Marrero et al., 2020). In this dissertation, I expand previous theoretical and empirical research, probing patients' gender stereotypes in reviews of physicians overall as well as within the specialties of primary care and surgery.

2. Essay 1: How patients describe their physicians: Characterizing the interpersonal manner and technical competence dimensions of online physician reviews

2.1 Introduction

Patients are increasingly turning to online physician ratings and reviews to assess physician quality (Findlay, 2016; Fox & Jones, 2009; Gao et al., 2012; Holliday et al., 2017). Between 40% and 60% of U.S. adults report ever having used a physician review website (Hanauer et al., 2014; Holliday et al., 2017; Kullgren et al., 2021). Since 2011, traffic to commercial physician review websites, including Healthgrades and Vitals, has substantially increased (Findlay, 2016). Research has shown people perceive online reviews as equally important as government clinical ratings and more important than physician qualifications when selecting a physician (Carbonell & Brand, 2018; Yaraghi et al., 2018). Most patients also trust the accuracy of commercial websites more than of health system patient-experience surveys (Holliday et al., 2017), and consider these websites as somewhat or very important resources when selecting a provider (Hanauer et al., 2014). In one study, 81% of patients reported they would choose a physician based on a positive review alone and 77% reported they would not seek care from a physician based on a negative review alone (Burkle & Keegan, 2015a).

Online physician reviews not only impact prospective patients but also physicians. Although most online reviews are positive (Emmert et al., 2013), negative reviews can harm a physician's reputation and practice (Frost, 2020). However,

physicians can use feedback provided in reviews to improve their care and communication (Lee, 2017; Murphy et al., 2019). One survey of over 2,000 physicians found that more than half of respondents used online ratings and reviews to derive quality improvement measures (Emmert et al., 2016). Most measures implemented were related to communication, scheduling, and office workflow.

Compared to traditional surveys, such as Press Ganey, commercial online reviews are easily accessible, more frequently updated, and utilized by more patients when selecting a provider (Schlesinger et al., 2015). However, commercial reviews have limitations. They can be anonymous; submitted by unverified or fraudulent patients (Frost, 2020); and represent a patient subpopulation that skews younger, more educated, and less healthy (Murphy et al., 2019; Terlutter et al., 2014). Despite their limitations, online physician reviews represent free, publicly available, voluntary patient perspectives and provide important insights into what matters to patients about their physician encounters and their medical care (Bardach et al., 2013; Peuchaud, 2020; Schlesinger et al., 2015).

Although numerical star ratings provide a quick and structured way to judge a physician's quality, the free-text reviews allow patients to detail their experiences more completely (Greaves et al., 2014; Schlesinger et al., 2015). Reviews help patients ascribe meaning to their experiences, including more richly describing their physician-patient interactions and interpretations of their physicians' skill (Schlesinger et al., 2015).

Reviews also allow patients to react emotionally to their healthcare experiences (Greaves et al., 2014). Thus, online reviews can reveal sources of perceived problems or positive experiences, and can buttress star ratings with descriptions of care-experience factors that patients value (Ranard et al., 2016; Schlesinger et al., 2015).

A few studies have sought to examine the content of patients' online reviews (Kilaru et al., 2016; Ranard et al., 2016; Wu & Tang, 2022). In one study, investigators developed a grounded theory of what patients say about their physicians on two review websites (López et al., 2012). Using a qualitative content analysis of 712 reviews received by 445 primary care physicians on Yelp.com and RateMDs.com, Lopez et al. (2012) identified that patients predominantly comment on the following four themes: global impressions, interpersonal manner, technical competence, and system issues (López et al., 2012). Other studies have built on Lopez et al.'s framework using natural language processing (NLP) approaches to conduct larger-scale analyses of the content in online physician reviews. These studies have associated Lopez et al.'s themes with positive and negative sentiments, health outcomes, overall ratings, patient choice, and gender differences (Dunivin et al., 2020; Paul et al., 2013; Wallace et al., 2014; Xu et al., 2021).

Lopez et al. provide a useful yet incomplete framework to describe the types of reviews patients write about their physicians and encounters. In this paper, we extend Lopez et al.'s framework by applying the dimensions of interpersonal manner and technical competence to a more recent, richer dataset of online reviews. We build on

Lopez et al.'s grounded theory by first analyzing a larger sample of 2,000 reviews representing 2,000 physicians. Second, we expand the physician sample to include not only primary care physicians but also surgeons, and to include physicians across the U.S., compared to Lopez et al.'s sample from four large U.S. cities (Atlanta, Chicago, New York, and San Francisco). Third, we examine a sample of reviews from a different online source, Healthgrades.com, one of the largest commercial physician review websites (Furnas et al., 2020).

2.1.1 Conceptual development

Lopez et al. (2012) identified two dimensions related to physician-patient interactions: interpersonal manner and technical competence. The investigators defined interpersonal manner as items related to physicians' characteristics, attentiveness, and communication skills, whereas technical competence included items related to physicians' perceived aptitude and thoroughness, clinical skills, and follow-up (López et al., 2012).

These dimensions of interpersonal manner and technical competence align with two fundamental dimensions of social perception prominent in social psychology literature: warmth and competence. Since the 1940s, researchers have used different labels to delineate these two dimensions. Asch (1946) coined "warm/cold" and "intelligent/industrious" to describe first impressions of people's personalities (Asch, 1946), Rosenberg, Nelson, and Vivekananthan (1968) proposed "social good/bad" and

“intellectual good/bad” in their theory of two-dimensional social judgments (Rosenberg et al., 1968), Fiske, Cuddy, and Glick (2007) referred to “warmth” and “competence” when relating these judgments to group stereotypes (Fiske et al., 2007), and Abele and Wojciszke (2007) used “communion” and “agency” to explain how people distinguish between self and others (Abele & Wojciszke, 2007). Despite the differences in nomenclature, psychologists have formed a robust consensus that these dimensions form the main components of impression formation, which together account for nearly 90% of the variance in how people spontaneously evaluate social behaviors (Abele & Wojciszke, 2007; Cuddy et al., 2008; Judd et al., 2005; Williams & Bargh, 2008; Wojciszke et al., 2009).

These two fundamental dimensions are also key components of patient-centered care. Considered by the Institute of Medicine as one of the six key elements of high-quality care, patient-centeredness requires high competence and warmth (Catalyst, 2017; De Valck et al., 2001; Epstein & Street, 2011a; Hall et al., 2014; Howe et al., 2019). Research shows patients value physicians who are patient-centered; high patient satisfaction ratings are associated with patient-centered care behaviors, such as good communication and thoroughness in care (Davis et al., 2021; Hall et al., 2014; Steiner-Hofbauer et al., 2018; Tak et al., 2015).

Lopez et al.’s grounded theory approach to characterizing online physician reviews confirms the relevance of psychology’s two fundamental dimensions to

evaluations of healthcare providers. However, Lopez et al.'s interpersonal manner and technical competence dimensions expand beyond psychology's warmth and competence dimensions by incorporating behaviors and skills distinct to physicians. For example, physicians' empathy and knowledge fall within the fundamental dimensions of warmth and competence, whereas time spent with physician during appointments and referrals are not warmth and competence concepts but are important aspects of physicians' interpersonal manner and technical competence. Thus, in this paper, we use Lopez et al.'s nomenclature of interpersonal manner and technical competence.

Lopez et al. provide a compelling initial framework of the interpersonal manner and technical competence factors patients evoke in comments about their physicians. Starting with Lopez et al.'s theory, we conduct our own qualitative content analysis and present an elaborated theoretical framework to distinguish the range of factors patients use to evaluate both primary care physicians and surgeons in online reviews. Our goal is to expand our understanding of how patients judge their physicians according to the dimensions of interpersonal manner and technical competence. We do this first by providing thicker descriptions of the factors comprising the two dimensions accompanied by fuller sets of illustrative examples. Second, we add factors not initially captured in Lopez et al.'s framework but that we observed patients describe in their reviews about their physician-patient interactions. Last, we exclude factors from Lopez et al.'s original framework that do not clearly represent either dimension.

2.2 Methods

2.2.1 Data collection

We scraped the physician profiles, including both rating and review data, from Healthgrades.com, one of the largest commercial online physician review platforms. We collected primary care physician (PCP) profiles listed under family medicine, internal medicine, and pediatrics, and surgeon profiles listed under general surgery; orthopedic surgery; and cosmetic, plastic, and reconstructive surgery. We scraped up to 20 reviews per physician, each associated with a 1-to-5-star rating.

2.2.2 Sampling

We conducted purposive random sampling to attain an equal number of reviews for PCPs and surgeons, for female and male physicians, and for reviews associated with low-star (≤ 3 stars) and high-star (≥ 4 stars) ratings.

2.2.3 Qualitative analysis

We manually coded each review for the presence and valence of interpersonal manner and technical competence. To develop our coding scheme, we started with Lopez et al.'s grounded theory. Lopez et al. developed their framework by coding reviews of PCPs practicing within four U.S. big cities (López et al., 2012). With our aim to characterize the reviews of both PCPs and surgeons practicing across the U.S., we also conducted our own qualitative content analysis to capture themes that were pertinent to both PCP- and surgeon-patient relationships.

Four investigators iteratively added to and refined the coding scheme through four rounds of double coding. After each round, the team met to discuss coding differences, resolve discrepancies, and reach group consensus on hard-to-code comments. We used a fifth round of double coding with 300 reviews to assess inter-rater reliability. After achieving high inter-rater reliability (Cohen's κ range 0.74–0.85), the remaining reviews were independently coded, with 10% being double coded to ensure continued high inter-rater reliability (Cohen's κ range 0.80–0.92). We used Stata for inter-rater reliability analyses.

After coding the full sample of reviews, we evaluated the reviews as a group to identify subthemes within interpersonal manner and technical competence. We used inductive reasoning to discern recurrent, unifying concepts, and developed a taxonomy through consensus (Bradley et al., 2007). Predominant themes are presented in this paper, including those that align with Lopez et al.'s original themes and those that we removed from or added to Lopez et al.'s coding scheme.

2.3 Results

We coded 2,000 reviews, each received by a unique physician, from Healthgrades.com. Based on our sampling design, our reviews were evenly distributed across the following categories: highly rated female PCPs, highly rated male PCPs, highly rated female surgeons, highly rated male surgeons, lowly rated female PCPs,

lowly rated male PCPs, lowly rated female surgeons, lowly rated male surgeons (n=250 or 12.5% each).

Approximately 68.4% (n=1,368) of the patient reviews expressed any judgment of the physician's interpersonal manner. Of those reviews, 56.5% (n=773) praised the physician's interpersonal manner. In contrast, 56.5% (n=1,130) of patient reviews expressed any judgment of the physician's technical competence and 56.0% (n=633) of those praised the physician's technical competence.

2.3.1 Interpersonal manner

We observed three ways in which patients described the interpersonal manner of their physicians: attitude and character, behavior, and communication.

Patients' comments about their physicians' attitude and character focused on how their physicians connected with them and recognized them as individuals with concerns and needs. Patients cared about their physicians' personalities, that their physician showed genuine concern for them as a patient, and that their physician exhibited decorum. Patients found their physicians personable, empathetic, caring, and respectful or, conversely, arrogant, rude, indifferent, and dismissive. Three of Lopez et al.'s factors fell into this category: empathetic, friendly, and nonjudgmental/condescending. In addition, we observed patients describe their physicians' bedside manner, professionalism, and caring when commenting on their physicians' attitude and character.

When patients described their physicians' behavior, they focused on their physicians' displays of personal engagement and actions taken to strengthen the patient-physician relationship. Spending time with patients, reassuring patients, exceeding expectations, and being available when needed improved physicians' relationships with their patients, whereas rushing through appointments, ignoring fears, deprioritizing the patient, and doing the bare minimum to care for the patient degraded the patient-physician relationship. Two of Lopez et al.'s factors fell into this category: time spent with physician during appointment and puts patient at ease. We added three factors that patients emphasized when commenting on their physicians' behavior: going above and beyond, prioritizes patient, and availability and access.

When patients commented on their physicians' communication, they described what, if any, information their physicians gave, how their physicians talked to them, and whether their physicians listened. Physicians' communication included informing the patient, explaining medical information, listening, answering questions, and conversation style. Two of Lopez et al.'s factors fell into the communication category: physician explains and physician listens. We additionally observed patients who commented on their physicians' communication described whether their physicians kept them informed and whether they appreciated their physicians' conversation style.

Although these overarching themes are an addition to Lopez et al.'s original taxonomy, most of the factors of interpersonal manner identified by Lopez et al. fall within these themes.

2.3.1.1 Interpersonal manner factors that aligned with those in Lopez et al.'s framework

Attitude and character

Within attitude and character, we identified three factors that matched those in Lopez et al.: empathy, friendly, and nonjudgmental/condescending.

Patients cared that their physicians showed empathy and took their concerns seriously. When patients described their physicians as empathetic or sympathetic, they perceived their physician as compassionate and understanding of their health issues. For example, "she showed compassion and [was] very relatable." In contrast, physicians who were not empathetic were described as insensitive or lacking the ability to share patients' feelings of pain or concern, such as, "She wasn't very sympathetic to the pain my mom endured from the needle."

Patients perceived their physicians' friendly or unfriendly demeanor as reflective of their physicians' attitude toward them. Physicians reviewed as friendly received praise for being charismatic, kind, and approachable, such as, "he was both patient and personable." Oppositely, patients criticized physicians who were unfriendly, cold, or arrogant, such as, "he emanates and holds a Godlike attitude towards himself."

Patients described their physicians as nonjudgmental when their physicians acknowledged and respected their health concerns and values. Patients appreciated their physicians who listened to and validated their concerns and preferences, such as, “never talks down to you...really listens to what you have to say.” In contrast, patients who described their physicians as condescending felt their physicians had patronized them and either invalidated or dismissed their experiences and concerns. For example, “If you want an MD to gaslight you about your concerns, he is your man.”

Behavior

Two of Lopez et al.’s original factors fell within our theme of behavior: time spent with physician during appointment and puts patient at ease.

Patients judged their physicians depending on how much time their physician spent with them in their patient-physician encounters. Patients appreciated when their physicians did not hurry through the appointment and took their time to listen, answer questions, and explain, such as, “always spends time discussing issues and concerns.” In contrast, patients disliked when their physicians rushed through the appointment or seemed to purposely truncate conversations, such as, “he takes on so many clients that our convos always felt rushed.”

Patients cared about their physicians’ ability to put them at ease when they expressed anxiety about their health and upcoming procedures and treatments. Patients valued physicians who reassured and soothed them, like, “alleviates fears and concerns, before and after surgery.” Oppositely, patients recognized when their physicians either

ignored their fears and doubts or actively distressed them, exacerbating their fears, such as, “invalidated my fears and concerns.”

Communication

Two of our factors within communication matched those found in Lopez et al.’s framework: physician explains and physician listens.

Patients praised physicians who provided clear explanations and gave detailed answers to their questions, such as, “he does a great job of explaining medical related topics in laymen terms,” and, “she gave detailed answers to my questions about resuming a number of activities.” In contrast, patients criticized physicians who insufficiently explained and answered questions, or who explained confusingly, such as, “she ignores questions, even when they have been asked multiple times.”

When patients described their health concerns and issues, they noted whether their physicians listened. Patients appreciated when their physicians listened, such as “she listens to the patient concerns,” and disliked when their physicians did not listen, such as, “she did not listen to me or acknowledge my statements.”

2.3.1.2 Interpersonal manner factors added to Lopez et al.’s framework

Attitude and character

In addition to the factors Lopez et al. identified, our analysis picked up three factors related to attitude and character. These factors included bedside manner, professionalism, and caring.

Patients commented on their physicians' bedside manner as a broad characterization of their physicians' approach to them as patients. A term unique to healthcare, bedside manner describes how healthcare providers personally interact with patients at their patients' bedside. Patients perceived good bedside manner, such as, "she has the most wonderful bedside manner," or bad bedside manner, such as, "this doctor has very poor bedside manner and interactions with his patients."

Patients described their physicians' professionalism when their physicians conducted themselves appropriately in patient-physician interactions. Patients either praised their physicians for being respectful and well-mannered, such as, "she is professional, but easy to talk to," or criticized their physicians for being rude and acting inappropriately, such as, "he wouldn't take accountability, and actually yelled at me multiple times when I questioned his decisions." Our inclusion of professionalism as an interpersonal manner factor departs from Lopez et al.'s framework. Lopez et al. characterized comments about professionalism as expressing a vague, overall sentiment about the physician or the healthcare visit; however, we observed comments about professionalism focused on the physician's polished manner.

When patients described their physicians as caring, they commented on how their physicians showed they were invested in their wellbeing and in helping them maintain or improve their health, such as, "truly cares about his patients and wants to

see them get better.” In contrast, patients criticized physicians who showed indifference toward their wellbeing, such as, “he couldn’t care less how she was doing.”

Behavior

Within behavior, we added three factors that were absent from Lopez et al.’s framework: going above and beyond, prioritizes patient, and availability and access.

Patients admired their physicians who went above and beyond to improve their patient care experience. Patients praised their physicians who incorporated personal touches into the care experience, such as remembering personal details about their lives, and supporting their health journeys as if they were close friends. For example, “she walked in pre op to check on me and was wearing a TEAM LORI shirt that I had given her while becoming so close and trusting her on my journey.” In contrast, patients judged physicians who only did the bare minimum of what was necessary to provide patient care. For example, “transferred My Mother in law to a therapy hospital in Cochran she was there 20 days, no visit, no call, no nothing from him.”

Patients appreciated when their physicians treated them as the main priority, such as “he has made me feel as if I was his top priority, even though I knew he had many others that needed his expertise and care even more than I did.” Patients, however, disliked their physicians who seemed to prioritize other factors above them. Patients criticized their physicians when they perceived their physicians cared more about themselves, the money they made from the visit, or pursuits outside of the office,

such as, “two out of three appointments this woman cancelled...beautiful weather had more allure than taking care of her patients.”

Patients cared about whether their physicians were available when needed and showed up on time for appointments. When patients commented on their physicians’ availability, they described access to their physician both in the office and outside of the office for questions, such as, “ALWAYS available, either in person or via message,” or, “does not have the time to take care of us. We have only been able to see her on two occasions in 3 years, and that was because we made them MONTHS in advance.” Patients also valued physicians who were punctual to their appointments, such as, “we never have to wait long to see him,” and criticized physicians with long wait times, such as, “he makes us wait hours after our appointment time because he thinks his time is more important than ours.”

Our inclusion of availability and access departs from Lopez et al.’s framework. Lopez et al. characterized comments about appointment access and appointment wait time as system issues; however, we observed that patients interpreted availability and punctuality as reflections of their physicians’ priorities and investment in the patient-physician relationship.

Communication

We included two additional factors to Lopez et al.’s framework: physician informs and conversation style.

Patients valued physicians who kept them informed and in the loop about their health issues, their treatment, and the latest medical research, such as, “always provides up to date pediatric information.” In contrast, patients criticized their physicians who withheld information, keeping them in the dark about their medical care, such as, “I had to take initiative and ask questions otherwise she would not have much to say.”

Patients cared about not only what information they received from their physicians but also the way their physicians conveyed that information. Patients judged their physicians’ conversation styles based on their own preferences. Some patients appreciated physicians who were straight shooters, such as, “She does not fuss but just tells the consequences in a frank way, and lets me decide,” whereas other patients found those forthright physicians to be too blunt, such as, “unfortunately, her manner was brusque and noncommunicative.” Patients also appreciated communication styles that matched their own personal backgrounds, such as, “he spoke to my husband like one country man speaking to another country man and it made him so very comfortable.” Table 1 presents our interpersonal manner factors and illustrative quotations.

Table 1 Interpersonal manner factors and illustrative quotations

	<i>Interpersonal manner factors</i>	<i>Illustrative quotations</i>
Attitude and character	Empathetic	<ul style="list-style-type: none"> • Very sympathetic to your needs • You're instead met with a complete lack of empathy
	Friendly	<ul style="list-style-type: none"> • Very nice guy • She acts superior to people in her care and cuts them off
	Nonjudgmental/ Condescending	<ul style="list-style-type: none"> • She was attentive and took my concerns seriously • Very dismissive with his patients' opinions and wishes
	Bedside manner ^a	<ul style="list-style-type: none"> • Her bedside manner is exceptional • His bed side manner does not even exist
	Professionalism ^a	<ul style="list-style-type: none"> • I was treated with respect and professional attitude by her at all times • Very unprofessional, rude with a gangster attitude
Behavior	Caring ^a	<ul style="list-style-type: none"> • Truly cares about his patients and wants to see them get better • No regards for patient well being
	Time spent with physician during appointment Puts patient at ease	<ul style="list-style-type: none"> • Never makes you feel rushed • Doctor spent a total of 5 minutes with me • He is comforting and prays with you prior to surgery • Didn't do anything to relax my mom or to explain what to expect during the procedure
	Going above and beyond ^a	<ul style="list-style-type: none"> • She tends to remember what's going on in my life and asks about how things are going • He has to be reminded why you are there, then he looks at his computer and still treats you like his is doing

		you a favor
Communication	Prioritizes patient ^a	<ul style="list-style-type: none"> • She makes you feel as if you were her one and only patient” • Did not listen to me when I was going over my problems like he use to, and is clearly only about getting his paycheck now
	Availability and access ^a	<ul style="list-style-type: none"> • Our daughter loves him and he's always available when needed • She is always on time, which is not something you find often in the medical field • She was never available to me when needed so I spent most of my time seeing docs at Urgent Care
	Physician explains	<ul style="list-style-type: none"> • She went over my biopsy results as well as the surgery that needed to be done and exactly what the surgery entailed, all in detail • Did not answer any of my questions in terms I could understand
	Physician listens	<ul style="list-style-type: none"> • Listens to patients carefully • Provider lacks listening skills
	Physician informs ^a	<ul style="list-style-type: none"> • He kept my family informed during the procedure & after the procedure • He never mentioned I might not get normal use of my arm and that I could be in pain after. I'd asked about side effects twice
	Conversation style ^a	<ul style="list-style-type: none"> • Straight forward no games • She sometimes came across abrupt and brash

Notes:
^aFactor added to Lopez et al.'s framework

2.3.1.3 Interpersonal manner factors originally in Lopez et al.'s framework that we excluded

We excluded four factors from Lopez et al.'s original framework: longevity of relationship with physician, helpful, trustworthy, and competent. Lopez et al. characterized all four factors as interpersonal manner; we removed three because they did not unambiguously reflect the interpersonal manner domain and one because it was mischaracterized as interpersonal manner instead of technical competence. Longevity of relationship with physician, helpful, and trustworthy could fall in either the interpersonal manner or technical competence domains, depending on their context in the review.

Patients mentioned their longevity of relationship with a physician in both complimentary and critical reviews. In both cases, longevity was typically nested within a broader comment about why the patient appreciated or disliked their physician, which could relate to the physician's interpersonal manner, technical competence, both, or neither. Some patients related their relationship longevity to their physician's interpersonal manner, such as, "has been our family doctor for over 20 years. He's never in a hurry, always answering our questions fully." Other patients related their relationship longevity to their physician's technical competence, such as, "has been treating my mother for several years. Due to his neglect and imcompatence [sic] my mother has lost her leg." Patients also highlighted their physicians' interpersonal manner and technical competence together as reasons for their longevity of relationship,

such as, "I've been her patient for over 6 years now. She has a wonderful personality and has always given me outstanding results."

Patients described their physicians as helpful or trustworthy both in terms of interpersonal manner and technical competence. For example, patients appreciated when their physicians were helpful in communicating and reassuring them, such as, "really helped put my fears to rest about my surgery." Patients also perceived their physicians as helpful in offering solutions, and in identifying and treating their health conditions, such as, "was helpful with outlining options for my shoulder injury."

Patients' comments, however, could also be unclear as to whether they found their physicians' interpersonal manner or technical competence helpful or unhelpful.

Examples included, "Most unhelpful medical professional who I have ever encountered," and, "she's helped me more than I can explain."

As with helpful, patients described their physicians' trustworthiness as both reflective of their physicians' interpersonal manner and technical competence. Patients trusted or felt confident in their physicians when their physicians honestly communicated and genuinely listened. For example, "I feel like he is very trustworthy person/doctor and truly listens to what you have to say." Patients also expressed trust in their physicians, or lack thereof, based on the physicians' ability to provide quality care and medical advice, such as, "did not see me as an inpatient after readmission for surgical complications. do not trust your body to this person." At other times, however,

patients described their physician as trustworthy or untrustworthy without elaborating on either the physician’s interpersonal manner or technical competence, such as, “amazing doctor you can truly trust.”

Last, when patients mentioned their physicians’ competence, they described their physicians’ technical competence rather than interpersonal manner. Patients perceived their physicians as competent when physicians demonstrated knowledge of their patients’ health issues or successfully diagnosed or treated their patients. Examples included, “very competent in diagnosis,” or, in contrast, “was incompetent. My operation was a disaster.” Table 2 presents Lopez et al.’s interpersonal factors, which we excluded from our framework, along with our reasoning and illustrative quotations.

Table 2 Excluded interpersonal manner factors, reasoning, and illustrative quotations

<i>Interpersonal manner factors</i>	<i>Reasoning and illustrative quotations</i>
Longevity of relationship with physician	<p>Could relate to interpersonal manner or technical competence</p> <ul style="list-style-type: none"> • IM: I have been a patient of Dr. Hickey for close to 20 years. He is a Phenomenal Doctor who deeply cares about his patients. • TC: I have been Dr. Zavala's patient for 3 years and have always had good experiences with her. She is thorough and knowledgeable. • Ambiguous: I have been going to Dr Nelson for over 3 years. He is by far the best doctor i have ever had.

Helpful	<p>Could relate to interpersonal manner or technical competence</p> <ul style="list-style-type: none"> • IM: He was very helpful with answering my questions. • TC: Dr. Douglas is very knowledgeable, and has helped me control not only my diabetes, but my weight. • Ambiguous: Most unhelpful medical professional who I have ever encountered.
Trustworthy	<p>Could relate to interpersonal manner or technical competence</p> <ul style="list-style-type: none"> • IM: I feel she's trustworthy in telling me what I need to hear, not necessarily what I want to hear. • TC: Trustworthy. I felt extremely comfortable & at ease knowing this doctor would be removing the body & tail of my pancreas along with my spleen. • Ambiguous: This guy is an untrustworthy quack
Competent	<p>Relates to technical competence</p> <ul style="list-style-type: none"> • TC: Really competent and knowledgeable doctor that was able to understand my symptoms and recommend a plan of action that works. • TC: One of the least competent doctors I have ever visited. I have been misdiagnosed by him on multiple occasions

Notes:
IM = interpersonal manner
TC = technical competence

2.3.2 Technical competence

Patients described the technical competence of their physicians using the following three categories: Expertise, treatment approach, and outcomes.

When patients commented on their physicians' expertise, they focused on whether their physicians had acquired the expected medical training, practice, and experience in the field to execute quality care. Physicians who demonstrated expertise were perceived as knowledgeable, capable of identifying and defining patients' health issues, and both efficient and thorough in their work. Physicians deficient in expertise seemed inexperienced, uninformed, and careless. Three of Lopez et al.'s factors fell into this category: knowledgeable, efficient, and detailed. We additionally observed patients describe their physicians' diagnostic skill when commenting on their physicians' expertise.

Patients who discussed their physician's treatment approach reflected on what their physicians did to maintain or improve their health. Physicians' treatment approach included offering a treatment plan, being open-minded about treatment, providing timely and appropriate referrals, and following up to ensure the patient remained in good health post-treatment. Four of Lopez et al.'s factors fell into the treatment approach category: clinical skills, complementary alternative medicine, referrals, and follow-up. Clinical skills may refer to several activities, including taking a patient history, conducting a physical examination, diagnosing, or performing a procedure. The illustrative comments Lopez et al. had characterized as clinical skills related to physicians' abilities to formulate and offer treatment plans. We therefore relabeled clinical skills as treatment offered to better reflect this treatment approach factor. We

also relabeled complementary alternative medicine as open-minded about treatment. We observed that patients commented on their physicians' open-mindedness to not only alternative medicine options but also treatments suggested by their patients and other physicians as well as outside-the-box medical treatments.

Patients who commented on their outcomes described how they perceived the impact of their physician's health care. Patients perceived good outcomes when their physicians' decisions or treatments improved their health, and perceived poor outcomes when those decisions or treatments either did not improve or harmed their health. Two of Lopez et al.'s factors fell into this category: perceived poor decision making and perceived successfulness of treatment. We relabeled perceived poor decision making to incorporate both perceived good and bad decision making.

Although these overarching themes are an addition to Lopez et al.'s original taxonomy, all the factors Lopez identified as technical competence fall within one of these three themes.

2.3.2.1 Technical competence factors that aligned with those in Lopez et al.'s framework

Expertise

Within expertise, we identified three factors that matched those in Lopez et al.: knowledgeable, efficient, and detailed.

Patients cared that their physicians were knowledgeable in their health issues and in the treatments and procedures their physicians performed, such as, "the

treatment was expertly and perfectly done,” or, “poor knowledge of eczema and alternative treatments.” Patients also commented on their physicians’ experience as a signal of their physicians’ knowledge gained through working in the profession for a period. For example, “very skilled, experienced surgeon,” or oppositely, “young and clueless.”

Patients described their physicians as efficient when they perceived their physicians’ prompt ability to address their issues, such as through diagnosis or procedure. Patients praised their physicians for fixing the problem quickly, such as, “she repaired my severely broken ankle, she was very efficient and made the experience as best as possible.” In contrast, patients critiqued their physicians for appearing slow and disorganized, such as, “This Dr and his staff appear to be completely unorganized, which is not only irresponsible, but dangerous.”

Patients judged their physicians as detailed when their physicians demonstrated thoroughness in their assessments and procedures, such as meticulous examinations and diligent reviews of their patient history. Patients also noticed when their physicians lacked sufficient attention to avoid errors, such as overlooking key patient medical information, or carelessness in ordering prescriptions and tests. For example, “quite impressed by his quality of care and the attention to details,” or, in contrast, “filled out the perscription [sic] wrong and where it said date he put my mothers birthdate.”

Treatment approach

All factors we identified under treatment approach matched those in Lopez et al.'s framework. The four factors include: treatment offered, open-minded about treatment, follow up, and referrals. Lopez et al. had named treatment offered as "clinical skills," and had named open-minded about treatment as "complementary alternative medicine." We changed how these two factors were characterized to add more specificity and to provide an expanded definition, respectively.

When patients commented on the treatment offered by their physicians, they described their physicians as formulating and executing targeted treatment plans that were in their patients' best interest or as failing to provide effective treatment plans. Patients complimented their physicians for diligently tinkering to provide the best treatment, such as "we started with drug treatment to find the best coarse [sic], which he modified over the years to where I now take minimal medication." Patients also expressed dismay when their physicians either offered no plan or formulated a course of treatment that was ineffective, such as, "offers no solutions or plan of action."

Patients judged their physicians' open-mindedness about treatment based on whether their physicians integrated information from a variety of sources, including patient preferences, second opinions, the latest research, and alternative medicine, to inform their recommendations. Patients praised their physicians for being appropriately flexible and innovative in offering a course of action, such as "He is also refreshingly open to some alternative supplements which I wanted to use as an adjunct to traditional

treatment,” or “she gave options for treatment rather than just one recommendation.” In contrast, patients criticized physicians with narrow or biased opinions about the patient’s treatment plan, such as “She was curt and dismissive about less invasive options, which gave me enough pause to take a break on pursuing a remedy.”

Patients commented about whether their physicians provided prompt and appropriate referrals as needed, either for a second opinion or to another provider more equipped to address their issue. Patients wanted referrals when they asked for them, and they wanted referrals to providers who could effectively identify and treat their health issues. For example, “when I began having foot problems he referred me to a brace maker.” In contrast, patients disliked when their physicians failed to make timely referrals or made referrals to providers the patients had already consulted, who were not in network, or who could not treat their condition, such as, “she doesn’t give referrals to specialists as quickly as she should because she wants to figure out the problem.”

Patients expected their physicians to not only treat their health issues but also follow-up with them to ensure proper healing. Patients also cared that their physicians maintained their patient-physician relationship post-visit or post-procedure. For example, “her and her team made sure to follow up/visit often during my hospital stay and were very responsive during my healing process!” In contrast, patients critiqued

their physicians for not dedicating time to further observe them after treatment, such as, “she never checked in with any of her patients.”

Outcomes

Our two factors identified under outcomes matched those in Lopez et al.’s original framework: perceived decision making and perceived successfulness of treatment. In Lopez et al.’s framework, perceived decision making was called “perceived poor decision making.” We changed the name of this factor to encompass both good and bad decision making.

Patients cared about whether their physicians’ decision-making and recommendations benefited or would benefit their health. Patients appreciated when their physicians offered targeted recommendations, such as, “he always reviewed the results of the current test results: the good and the not so good, and advises accordingly,” and disliked when their physicians’ recommendations led to harm, such as “he told me to double my exercises. As a result, I overworked the knee.” When reading these comments, we adopted patients’ viewpoints in how they perceived their physicians’ decision making. If a physician did not prescribe a medication or chose to recommend a less-intensive treatment than the patient desired, the physician’s decision may have been appropriate. However, if the patient perceived poor decision making, we coded the review comment accordingly, such as, “Nothing showed on MRI so he said there was nothing there. Was unwilling to go digging around in there.”

When patients commented on their perceived successfulness of treatment, they described whether their physicians helped them achieve their desired health results. Patients either expressed satisfaction with their treatment outcomes, such as, “one day after surgery and I am already walking better than I have for the last year,” or perceived their physician failed to adequately remedy their health issue, such as, “very unhappy with work that was done and what was promised. Issue never fixed.”

2.3.2.2 Technical competence factors added to Lopez et al.’s framework

Expertise

We added one factor to Lopez et al.’s framework: diagnostic skill.

Patients appreciated their physicians’ diagnostic skill when their physicians diagnosed their condition with timeliness or accuracy, such as “I had been misdiagnosed with a long standing cough for 2 years, and she was able to within a few minutes ascertain that the cause of the cough was the medication I was taking.” On the other hand, patients criticized their physicians who either missed opportunities to detect a health issue or misdiagnosed their condition. For example, “I was hospitalized for a heart condition only to find out later that my resting heart rate and blood pressure had been astronomical for years and he had never mentioned anything.”

Although we expanded the characterizations of the factors within treatment approach and outcomes, we did not add new factors to Lopez et al.’s original framework. We also did not exclude any factors from Lopez et al.’s framework. Table 3 presents our technical competence factors and illustrative quotations.

Table 3 Technical competence factors and illustrative quotations

	<i>Technical competence factors</i>	<i>Illustrative quotations</i>
Expertise	Knowledgeable	<ul style="list-style-type: none"> • She is very knowledgeable in her field • He had to Google information about my condition during my appt
	Efficient	<ul style="list-style-type: none"> • He was efficient in assessing my situation, evaluating me, explaining my options, and formulating a treatment plan • He is the slowest doctor I have ever seen
	Detailed	<ul style="list-style-type: none"> • Extremely thorough in her examinations/ordering tests when necessary • He never looked at my chart or history and suggested I follow up with a doctor I had already seen
	Diagnostic skill ^a	<ul style="list-style-type: none"> • Made a quick and accurate diagnosis, and made an arrangement for surgery quickly losing no time • This is the most incompetent to say the least doctor i've ever seen. she couldn't diagnose redness my little one had, throwing at me random guesses.
Treatment approach	Treatment offered ^b	<ul style="list-style-type: none"> • She will investigate and provide the best solutions until the issue is resolved • He prescribed the wrong

Outcomes	Open-minded about treatment ^c	<p>medications and the wrong treatment plans</p> <ul style="list-style-type: none"> • When we felt hopeless with health issues; she researched and thought outside the box • He is not willing to pursue other options if they are not his suggestions
	Follow up	<ul style="list-style-type: none"> • He even called to follow up after my son's appointment to ensure that he was not having any issues • No follow up appt after my procedure, I was told to contact only if I had problems
	Referrals	<ul style="list-style-type: none"> • Her prompt referral to a vascular surgeon probably saved my life. • She has also referred me to physicians that do not accept Medicare
	Perceived decision making ^d	<ul style="list-style-type: none"> • My case was a hard one he did all he could then transferred me to a university hospital which was the best decision • Gave me medical advice that, had I taken it, would have guaranteed my cancer progressing making surgery unnecessary
	Perceived successfulness of treatment	<ul style="list-style-type: none"> • Within 3 weeks, my cough disappeared • She should have never said she could perform this surgery because it was a miserable outcome

Notes:

^aFactor added to Lopez et al.'s framework

^bTreatment offered is characterized in Lopez et al. as "Clinical Skills." Factor relabeled to specifically describe the distinctive basic care component.

^cOpen-minded about treatment is characterized in Lopez et al. as "Complementary Alternative Medicine." Our factor includes comments about the physician's flexibility and innovative thinking or, oppositely, the physician's unwillingness to consider alternatives.

^dPerceived decision making is characterized in Lopez et al. as "Perceived Poor Decision Making." Factor relabeled to include comments reflecting both good and bad decision making.

2.3.3 Limitations of the framework

For several reviews, it was difficult for us to come to a consensus about categorizing a review as interpersonal manner or technical competence because of two broad issues: First, the two domains are tightly linked and not mutually exclusive. Second, review language was often ambiguous, leading to different interpretations of a comment as reflecting interpersonal manner, technical competence, both, or neither.

Many reviews were clearly about both interpersonal manner and technical competence (e.g., "personable and very knowledgeable"). However, because these domains are not mutually exclusive, sometimes we had trouble discerning whether the review was about interpersonal manner or technical competence. For example, the review comment, "She gave me no discharge instructions [on] how to take care of incisions," could be about the physician's communication (i.e., interpersonal manner), or about the physician's thoroughness (i.e., technical competence). Similarly, patients at times perceived follow-up care not only as a part of their physicians' treatment

approach, but also as a way for physicians to show their character and genuine investment in their well-being. For example, “absolutely cares about his patients and has good follow up with patients” links the physician’s follow up with his caring character. In contrast, “never available in the office for follow ups. I have issues after a cosmetic procedure and can't get in to discuss,” emphasizes poor follow-up care as a reflection of the physician’s availability to the patient.

Ambiguous review language also affected whether we could confidently classify a review comment as interpersonal manner or technical competence. In cases where language could be attributed to either category, we decided not to code the review as either. For example, “he helped me,” without additional context, could refer to the physician helping the patient understand the treatment (i.e., interpersonal manner) or to the physician successfully treating the patient’s condition (i.e., technical competence).

Last, we chose not to code review comments about system factors or staff unless the patient linked those comments to an evaluation of the physician’s interpersonal manner or technical competence. For example, “long wait” was not coded, whereas “Wait time is always long, he doesn't care” was coded as interpersonal manner. Additionally, “staff very rude” was not coded, whereas “If he is responsible for office and the way it operates, he gets a o. In my opinion office is operated so poorly it might put a patient at risk!!” was coded as technical competence. Figure 1 presents our

elaborated framework of interpersonal manner and technical competence judgments and their predominant factors.

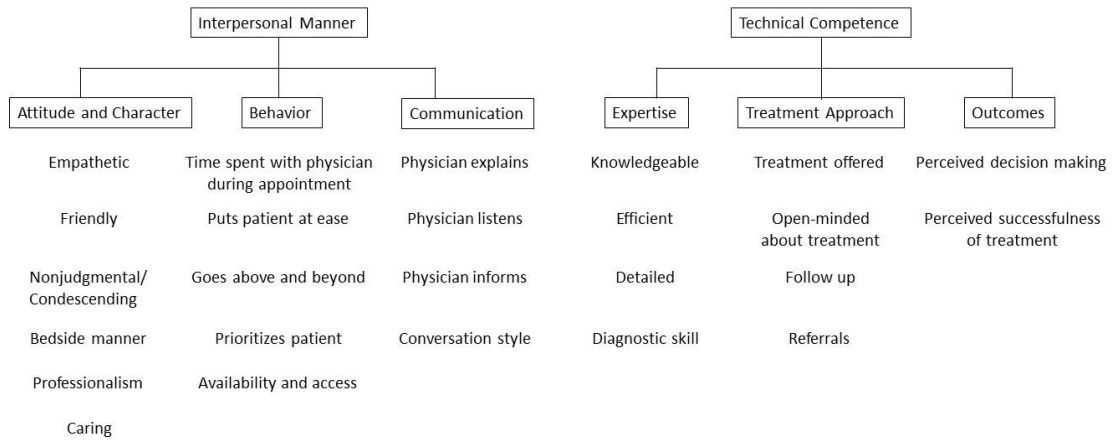


Figure 1 Framework of the predominant factors comprising patients' interpersonal manner and technical competence judgments in online physician reviews

2.4 Discussion

Online physician reviews allow patients to evaluate the quality of their physicians by detailing their patient-physician interactions in their own words (Emmert et al., 2013). These reviews offer meaning and emotion to patients' numerical ratings, helping physicians better understand why a patient liked or disliked them (Greaves et al., 2014; Schlesinger et al., 2015). These reviews not only provide accessible, candid information to prospective patients but also actionable feedback to help physicians improve their quality (Emmert et al., 2013; Hanauer et al., 2014).

By examining the content of online physician reviews, researchers can better understand patients' perspectives about their care and gain insights on how to enhance the patient-physician relationship. In this study, we propose an elaborated theoretical framework characterizing patients' interpersonal manner and technical competence judgments of their physicians. We developed this framework by expanding on the grounded theory of Lopez et al. and by conducting our own qualitative content analysis of 2,000 online physician reviews.

We expanded on Lopez et al.'s framework in the following ways: First, we derived our qualitative coding insights from a new, larger dataset of reviews received by physicians practicing across the U.S., with balanced representation of PCPs and surgeons, male and female physicians, and high- and low-star reviews. Second, we added factors to Lopez et al.'s framework to capture patients' judgments of both their PCPs and surgeons more comprehensively and we removed factors that were too ambiguous to categorize. Last, we provided thicker descriptions and more illustrative examples of the factors both included and excluded from the framework.

As medicine shifts to a more patient-centered approach, physicians are tasked with recognizing the patient as an individual with preferences, and with providing treatment expertise that aligns with those preferences (Howe et al., 2019). We focus on interpersonal manner and technical competence because these dimensions not only reflect core concepts of patient-centeredness, but also are within physicians' direct

control. In our study, we found that patients valued physicians for characteristics like their empathy, availability, and clear explanations. Patients also desired physicians who were knowledgeable, open-minded about treatment options, and provided referrals when appropriate. In contrast, patients were dissatisfied with physicians who exhibited poor interpersonal and technical skills, like those who rushed them through appointments and were condescending, who missed medical chart details and misdiagnosed them.

This study has several limitations. First, physicians' interpersonal manner and technical competence are not the only dimensions patients value when judging the quality of their healthcare experiences. Lopez et al. and other researchers have shown patients care about and comment on system issues, such as staff, location, parking, cost, and office environment (Ko et al., 2019; Lagu et al., 2010; López et al., 2012). Future research should continue to elaborate and refine these frameworks to better understand predominant system-level issues.

Second, we did not capture demographic information on the patients who submitted these reviews. Unlike in traditional patient-experience surveys, patients can submit online reviews anonymously or under pseudonyms. Although most commercial review websites do not verify that reviewers are real patients, some websites, like ZocDoc, are starting to change this by only posting reviews from patients who book through their system or are contacted after their appointment by third-party survey

providers (ZocDoc). However, despite physicians' fears that online review websites provide opportunities for anonymous patients (and non-patients) to criticize their care, most ratings and reviews of physicians are positive (Kadry et al., 2011b; Lagu et al., 2010; López et al., 2012; Saliba & Black, 2009).

Despite the drawbacks of online physician reviews, there are distinct benefits to commercial physician review websites. Unlike with traditional surveys, patients can evaluate their physicians online based on one-time or repeated interactions, providing a more holistic view of their patient-physician interactions. The selection bias in who volunteers an online review also differs from that of who completes a traditional patient-experience survey. Whereas older, less educated patients tend to respond to traditional surveys, younger, more educated patients tend to review their physicians online (Lasek et al., 1997; López et al., 2012; Terlutter et al., 2014). Thus, online reviews may capture patient perspectives not otherwise represented in traditional assessments. Last, patients who respond to traditional surveys may demonstrate social desirability bias and avoid offering criticism for fear of retribution (Anhang Price et al., 2022; Badejo et al., 2022). Research shows ratings published by healthcare institutions are more favorable toward physicians and less dispersed than corresponding ratings on commercial websites (Kordzadeh, 2019). Allowing proactive, anonymous patient reviews online may therefore minimize potential response biases.

Both traditional patient-experience surveys and commercial review websites are limited in not only what questions they pose but also their validity (Daskivich et al., 2018; McGrath et al., 2018; Pines et al., 2018; Sitzia, 1999). Thus, online physician reviews may reveal additional factors patients care about that are not captured in either traditional survey or commercial review measures (Ranard et al., 2016). Future research should examine how factors from physician reviews can be aggregated to provide more digestible survey measures and should test how the inclusion of different survey measures impacts patient-experience scores both online and in traditional surveys.

2.4.1 Conclusion

Online physician reviews are a publicly accessible resource to tap into patient perspectives of their medical encounters in their own words. By exploring online physician reviews, researchers can better understand how patients judge their patient experiences and the characteristics they like and dislike in their physicians. In this study, we developed an elaborated theoretical framework characterizing patients' interpersonal manner and technical competence judgments of their physicians. This framework drew on prior grounded theory research and a qualitative content analysis of reviews received by a broad sample of physicians on one of the largest commercial physician review websites (López et al., 2012). This framework has implications for future researchers seeking to analyze physician reviews, for physicians seeking to draw

insights from unstructured feedback, and for policymakers seeking to improve their survey measures to better reflect physician characteristics patients value.

3. Essay 2: What patients like and dislike about their physicians: Developing and testing an algorithm to classify social judgments in online physician reviews

3.1 Introduction

Patients increasingly turn to commercial physician rating and review websites to discuss their patient experiences (Guo et al., 2021) and provide feedback to hospitals and providers (Kilaru et al., 2014; López et al., 2012). Patient-authored reviews on these websites may capture factors of the patient experience not otherwise found in traditional patient experience surveys (e.g., Press Ganey) or academic research (e.g., interviews, questionnaires), such as insurance processing and appointment scheduling (Ranard et al., 2016; Xu et al., 2021). These websites have therefore gained increased attention among researchers seeking to better understand what patients care about and how commercial review data compare to other healthcare quality measures. For example, researchers analyzing commercial hospital reviews identified topics discussed by patients that were not covered in the current Hospital Consumer Assessment of Healthcare Providers and Systems (i.e., HCAHPS) survey, like nurse quality, staff compassion, and the technical aspects of care (Ranard et al., 2016). Other researchers found negative commercial reviews of surgeons focused on surgeon-independent factors, such as wait times and office staff, suggesting patients may consider factors beyond the patient-physician interaction when assessing quality (Trehan et al., 2016).

Online physician review websites potentially impact both patient choice and physician care quality. Some prospective patients rely on online physician ratings and reviews to help them choose physicians (Burkle & Keegan, 2015b; Murphy et al., 2019; Xu et al., 2021). Research shows people prefer words to numbers and can easier comprehend review narratives over quantitative ratings (Emmert et al., 2016). Additionally, physicians use patient feedback conveyed in online reviews to implement and improve quality measures, particularly related to patient communication (Emmert et al., 2016).

The unstructured narrative text, however, presents a challenge for researchers seeking to make inferences from reviews. Methods previously used to identify judgments within reviews include hand-coding (López et al., 2012) and dictionary-based approaches (Chen et al., 2021; Dunivin et al., 2020). Hand-coding approaches, however, are time and resource intensive, which limits sample size (Nelson et al., 2021). Likewise, dictionary-based methods, like Linguistic Inquiry and Word Count, use a context-independent bag-of-words approach, which may overlook misspellings, colloquialisms, and keywords and phrases not captured in prebuilt dictionaries (Ballard et al., 2022; Li et al., 2020).

In this paper, we present measures of precision, recall, and accuracy for a new natural language processing (NLP) algorithm, designed to identify the presence and valence of two social judgments in online physician reviews: interpersonal manner and

technical competence. We use a Robustly Optimized BERT Pre-Training Approach (i.e., RoBERTa), which we trained to classify our two social judgments in reviews and which has been successfully applied in other classification contexts (e.g., Twitter) (Ballard et al., 2022; Oliveira et al., 2022). We validate this algorithm by correlating results with review star ratings and by comparing results with those found in prior literature.

3.2 Methods

3.2.1 Data collection

We scraped physician profile, rating, and review data published on Healthgrades.com in April 2020. We collected primary care physician profiles associated with family medicine, internal medicine, and pediatrics, and surgeon profiles associated with general surgery; orthopedic surgery; and cosmetic, plastic, and reconstructive surgery. Healthgrades.com has a physician profile for every U.S. physician with an active profile listed on the National Provider Identifier Registry (Healthgrades Frequently Asked Questions (FAQs)). In addition to physician profile characteristics, we scraped rating information and up to 20 of the most recent reviews per physician. On Healthgrades.com, patients can elect to submit a star rating alone (i.e., 1–5 stars, no fractions) or a star rating accompanied by a written review. The study was approved by our institutional review board and all data collected were publicly available and aggregated for research purposes. Our final sample included 345,053 reviews submitted for 167,150 physicians. Figure 2 shows a flow chart of our sample selection.

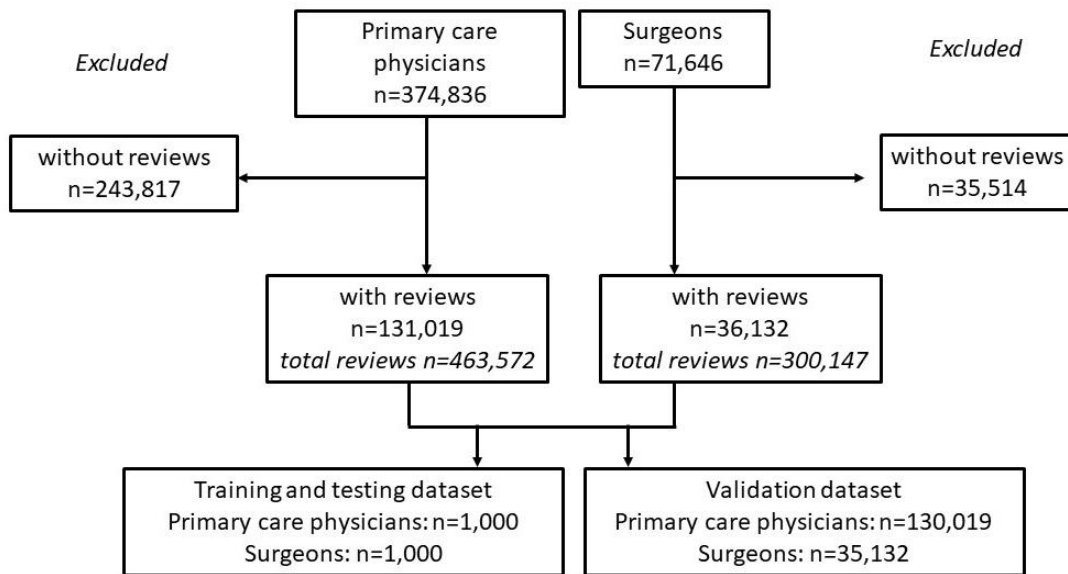


Figure 2 Sample selection flow chart

3.2.2 Coding reviews for interpersonal manner and technical competence

3.2.2.1 Hand-coding the training data

We trained a coding algorithm using a gold standard dataset of rigorously hand-coded physician reviews (Nelson et al., 2021). We purposely sampled 2,000 random reviews for equal representation of primary care physicians and surgeons, female and male physicians, and low-star (≤ 3 stars) and high-star (≥ 4 stars) review ratings.

We coded each review for the presence or absence of interpersonal manner and technical competence. Reviews could be coded for the presence of only one domain, both domains, and neither domain. Once we indicated the presence of a judgment

domain, we then coded the valence of the judgment as positive or negative. If we did not code the presence of a judgment domain, we would not have a valence indicated for that judgment. Figure 3 provides a diagram with illustrative examples showing how we hand-coded the presence and valence of the two social judgment domains.

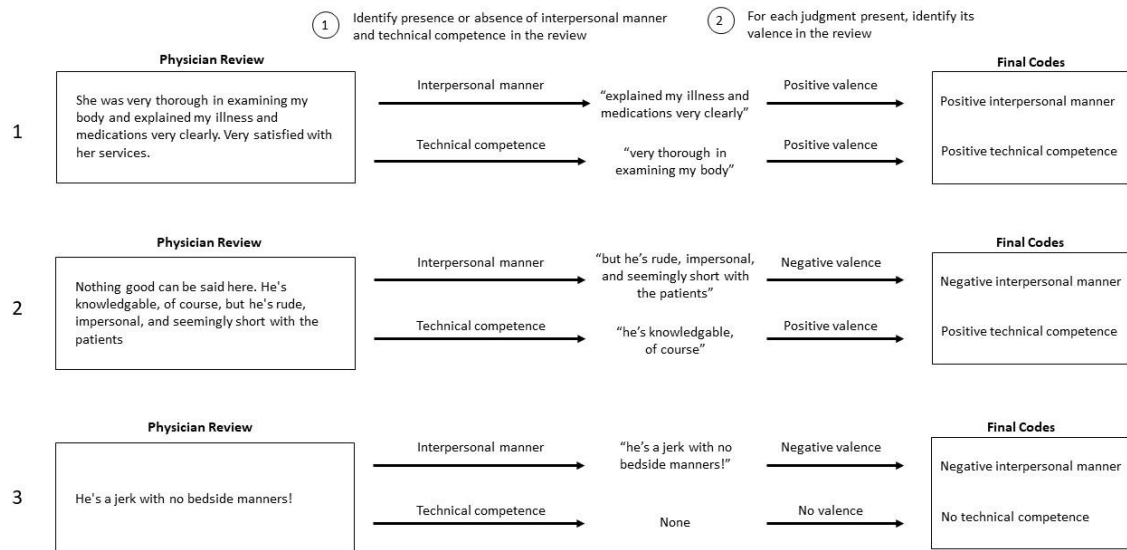


Figure 3 Diagram showing how real physician reviews were hand-coded for the presence and valence of interpersonal manner and technical competence

3.2.2.2 Training and testing the algorithm

To code interpersonal manner and technical competence judgments in the sample of 345,053 reviews, we first used our hand-coded data to train a transformer classification model called RoBERTa. Transformers are neural-network systems that use vectors to capture the meanings of words in context and are the main architecture

underlying advanced NLP models (Ballard et al., 2022; Vaswani et al., 2017). These state-of-the-art NLP models improve upon prior NLP classification approaches, such as those based on dictionaries or fixed embeddings (Ballard et al., 2022). Specifically, RoBERTa builds bidirectional context-aware embeddings such that the vector representing the word changes depending on its context in the text (Ballard et al., 2022; Liao et al., 2021). RoBERTa is pretrained on a large corpus of text and can be fine-tuned with one additional level of pretraining data for specific classification tasks. In our study, each review used to train RoBERTa has its own sequence embedding and helps fine-tune the model to code the reviews for the presence and valence of interpersonal manner and technical competence.

We fine-tuned two models, one for classifying interpersonal manner and one for technical competence. We tuned each model using 1,600 (80%) reviews randomly sampled from our hand-coded training data. We completed six iterations through our training data for both models. We used a test dataset, or a set of 200 (10%) hand-coded reviews held out of the training data, to evaluate each model's fit on the training dataset while further fine-tuning the model. Finally, we used a new set of 200 (10%) hand-coded reviews to provide an unbiased evaluation of the classification performance of each fully trained model for interpersonal manner and technical competence judgments.

After training and testing the two models using the 2,000 hand-coded reviews, we applied the fully trained models to all the reviews in our dataset. This resulted in a

dataset with 345,053 reviews coded for the presence and valence of interpersonal manner and technical competence judgments. We used Python and Google Colab to train RoBERTa on our judgment classification tasks and code our full sample of reviews.

3.3 Results

3.3.1 Evaluating the accuracy of the two classification models

Our two classification models were highly accurate. The out-of-sample predictive accuracy for the interpersonal manner model was 90% with weighted F1 score of 0.89 (range 0.82–0.95), precision of 0.89 (range 0.85–0.94), and recall of 0.90 (range 0.80–0.96). The out-of-sample predictive accuracy for the technical competence model was also 90%, with weighted F1 score of 0.90 (range 0.90–0.92), precision of 0.91 (range 0.88–0.95), and recall of 0.90 (range 0.85–0.95). Table 4 details the classification performance metrics for the interpersonal manner and technical competence models.

Table 4 Tuned transformer classification performance for interpersonal manner and technical competence judgments

<i>Interpersonal Manner Model</i>			
Valence	Precision	Recall	F1 score
No interpersonal manner	0.85	0.80	0.82
Negative interpersonal manner	0.88	0.89	0.88
Positive interpersonal manner	0.94	0.96	0.95
Accuracy			0.90
Macro Avg	0.89	0.88	0.88
Weighted Avg	0.89	0.90	0.89

<i>Technical Competence Model</i>			
Valence	Precision	Recall	F1 score
No technical competence	0.88	0.91	0.90
Negative technical competence	0.95	0.85	0.90
Positive technical competence	0.89	0.95	0.92
Accuracy			0.90
Macro Avg	0.91	0.90	0.90
Weighted Avg	0.91	0.90	0.90

Notes:

Precision = number of true positives divided by the sum of true positives and false positives. Recall = number of true positives divided by the sum of true positives and false negatives. F1 score = harmonic mean of precision and recall, given by $2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

(Ballard et al., 2022). Classification performance is based on a comparison to an evaluation dataset of 200 reviews hand-coded by our team of researchers.

3.3.2 Comparing reviews coded by RoBERTa and by hand

As part of our final sample of 345,053 reviews, the 2,000 hand-coded reviews from our training dataset were re-coded by RoBERTa. The inter-rater reliability between our hand-coding and RoBERTa was Cohen's $\kappa = 0.96$ for both interpersonal manner and technical competence. Comparing the RoBERTa codes with the original hand codes for these reviews, we found only 107 (5.4%) reviews had coding discrepancies. Of those, 49 (2.5%) reviews had the same interpersonal manner code but different technical competence code, 57 (2.9%) had the same technical competence code but different interpersonal manner code, and 1 (.05%) had both different interpersonal manner and

technical competence codes. Table 5 shows illustrative examples of coding discrepancies in our RoBERTa-coded reviews and our hand-coded reviews from the training dataset.

Table 5 Illustrative examples of discrepancies in reviews coded by RoBERTa and by hand and reasoning underlying the discrepancies

<i>Review</i>	<i>RoBERTa coding</i>		<i>Hand-coding with reasoning</i>	
	Interpersonal manner	Technical competence	Interpersonal manner	Technical competence
#metoo. He knows what he did.	0	0	- Hints at sexual assault by the physician	0
When someone is scared, I Think the Dr. should try to comfort them instead of telling them that when I finish this procedure , you will no longer be my patient . I'll refer you to someone else. So I wish him the best.	+	0	- Feels physician did not provide comfort	0
Unique combination of an extremely clean office, caring staff, artistic skill. The staff got me in the door due to the pricing, the place was clean and the ladies were nice. The doctor was able to place the injections in all the right places to cover some folds I noticed when I would smile. I'm happy with how things are looking since I started	+	+	0 Discusses the interpersonal manner of the staff, not the physician	+

coming here.

Walking on the second day.
My daughter also! I would
send anyone I know his way..
We're from out of town and
after using a local doctor here
and having to do it all over
again the experience was
amazing and recovery was
shorter

Dr. Goldberg is technically
proficient, but I felt she
wasn't really in my corner.
She would say one thing to
my face, then put something
else in her written notes to
make me look like bad or like
I wasn't in compliance with
taking thyroid medication, for
example. I know a lot about
natural health and also have a
Naturopath. She probably
didn't like that fact, even
though I was transparent
about the care I received from
my Naturopath and followed
many of Goldberg's
recommendations.

I was dreading my
procedure, as all I had heard
were negative things from
everyone who had this
procedure in the past.
However, they were wrong!
The doctor and the entire staff
were gracious, comforting
and kind, and actually made
this a good experience. So
happy I chose Dr. Meghan

0

0

0

+

Perceives
treatment as a
success

-

-

-

+

Describes
physician as
technically
proficient

+

+

+

0

Does not
discuss
physician's
expertise,
treatment, or
outcomes

<p>and her crew to work on me!</p> <p>Very similar to all the one star ratings, if 0 stars were an option I'd choose that. The follow up on patients are non-existent, which makes it very obvious that the surgeon just wants \$. The staff is always rude. I wish they would treat their patients and their family members how they would like their own to be treated.</p> <p>My mom has had two infections where they removed her lymph nodes after trying to call them about this several times, we took her to a different doctor to have the site drained</p>	<p>0</p>	<p>0</p>	<p>-</p> <p>Feels physician prioritized money over their care</p>	<p>-</p> <p>Critiques physician's lack of follow-up care</p>
--	----------	----------	---	--

3.3.3 Testing the validity of the two classification models

We test the validity of our classification models using the full sample of RoBERTa-coded reviews. We validate our models in two ways: First, we relate our valence coding with the review star ratings (i.e., 1–5-star ratings submitted with each review), with the expectation that positive-valence judgments would be associated with higher star ratings and negative-valence judgments would be associated with lower star ratings. Second, we compare our findings with prior literature on social judgments of physicians in online physician reviews.

3.3.3.1 Testing for associations between valence and star ratings

Using multilevel linear regressions, we analyzed associations between interpersonal manner and technical competence judgment valences and review star ratings. We found evidence of construct validity: Positive valences for both interpersonal manner and technical competence were significantly positively associated with review star ratings whereas negative valences for both social judgments were significantly negatively associated with review star ratings. Compared to reviews with no or negative judgment, reviews with positive interpersonal manner were associated with 1.82 more stars (95% CI [1.81, 1.83], $p < .001$) and reviews with positive technical competence were associated with 1.50 more stars (95% CI [1.49, 1.51], $p < .001$). In contrast, compared to reviews with no or positive judgment, reviews with negative interpersonal manner were associated with 3.30 fewer stars (95% CI [-3.31, -3.29], $p < .001$) and reviews with negative technical competence were associated with 3.00 fewer stars (95% CI [-3.01, -2.98], $p < .001$).

Figure 4 displays mean review star ratings for each judgment domain.

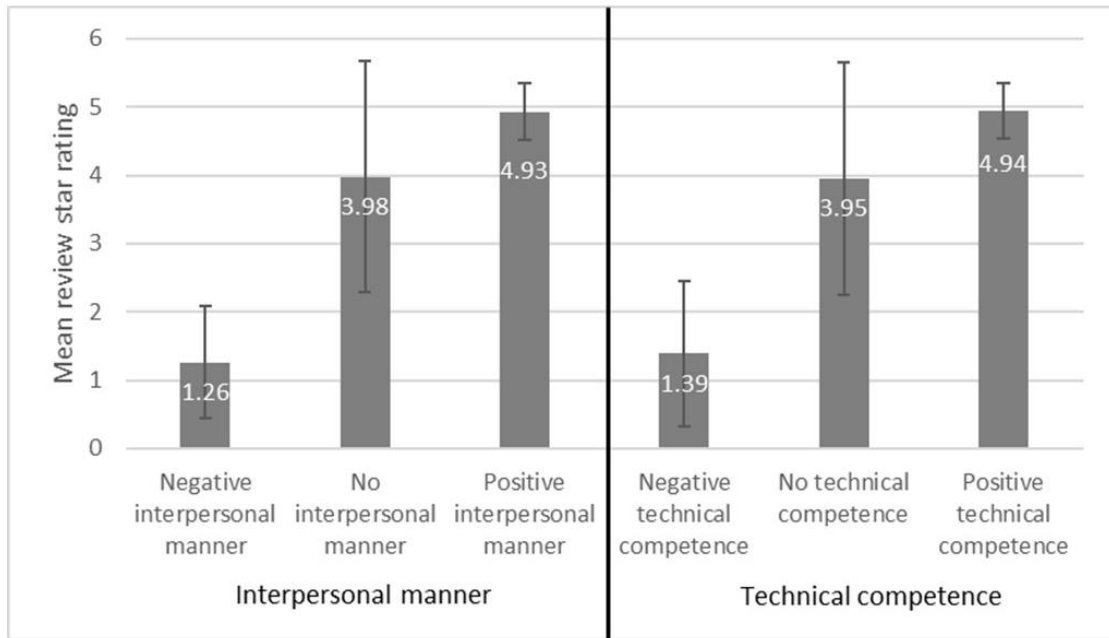


Figure 4 Mean review star ratings for reviews with positive, negative, or no interpersonal manner or technical competence

3.3.3.2 Testing whether the models reproduce prior findings

One study determined 69% of interpersonal manner reviews and 80% of technical competence reviews were positive (López et al., 2012). We identified a similar pattern of majority positive reviews; 81% of reviews mentioning interpersonal manner and 82% of reviews mentioning technical competence were positive. Another study reported physicians who received reviews with interpersonal manner language were at least 2.39 times more likely to receive a 5-star review rating (Chen et al., 2021). We similarly found physicians who received interpersonal manner reviews had 1.69 times

the odds of receiving a 5-star review rating (95% CI [1.65, 1.73], $p < .001$). When controlling for physician gender, specialty, age, and practicing state, as well as review word count, physicians with interpersonal manner reviews continued to have higher odds of receiving a 5-star review rating (OR: 2.22, 95% CI [2.17, 2.28], $p < .001$).

Prior research also showed female physicians, compared to male physicians, had higher odds of receiving reviews mentioning interpersonal manner (Chen et al., 2021; Dunivin et al., 2020). Our findings support these results: We determined female physicians had 1.56 times the odds of receiving a review mentioning interpersonal manner than male physicians (95% CI [1.53, 1.59], $p < .001$). When including physician age, specialty, practicing state, and word count controls, female physicians still had significantly higher odds of receiving an interpersonal manner review (OR: 1.19, 95% CI [1.17, 1.22], $p < .001$).

One group of investigators demonstrated female physicians were more likely than male physicians to receive both reviews praising and reviews criticizing their interpersonal manner (Dunivin et al., 2020). Consistent with these results, we found female physicians had 1.40 times the odds of receiving a negative review about their interpersonal manner than male physicians (95% CI [1.36, 1.44], $p < .001$). This significant difference remained when including controls (OR: 1.25, 95% CI [1.21, 1.29], $p < .001$). Female physicians also had 1.18 times the odds of receiving a positive interpersonal manner review than male physicians (95% CI [1.15, 1.20], $p < .001$); however, this gender

difference was not significant when adding controls (OR: 1.02, 95% CI [1.00, 1.04], $p=.052$).

Likewise, another study concluded highly rated male physicians were 1.48 times more likely to receive reviews describing technical competence whereas highly rated female physicians were 2.11 times more likely to receive reviews describing interpersonal manner (Haynes et al., 2021). We found similar results: Highly rated female physicians had 1.76 higher odds of receiving an interpersonal manner review (95% CI [1.72, 1.81], $p<.001$), which was still significant after including controls (OR: 1.25, 95% CI: [1.22, 1.29], $p<.001$). Highly rated male physicians had 1.33 higher odds of receiving a technical competence review (95% CI [1.30, 1.36], $p<.001$); however, with controls, this gender difference flipped, such that females were more likely to receive a technical competence review (OR: 0.95, 95% CI [0.93, 0.98], $p<.001$).

3.4 Discussion

3.4.1 Principal results

Our two classification models identified the presence and valence of interpersonal manner and technical competence judgments with high precision, recall, and accuracy. Our models identified these two social judgments from a broad training dataset inclusive of reviews for female and male physicians, for primary care physicians and surgeons, and for low- and high-star-rated reviews.

Our interpersonal manner and technical competence models underperformed on classifying reviews with no judgment and negative valence relative to reviews with positive valence. However, the overall predictive accuracy of both models (90%) was higher than the rate of hand-coding agreement among the four investigators (Cohen's κ = 0.84 and 0.77 for interpersonal manner and technical competence, respectively).

Our models produced classification metrics comparable to those found in other studies that use fine-tuned RoBERTa algorithms for coding tasks. For example, researchers who used RoBERTa to detect polarizing vs. nonpolarizing rhetoric in tweets written by Congressmembers reported a model with 90% predictive accuracy and weighted F1 score of 0.9 (Ballard et al., 2022). They compared their results to the Valence Aware Dictionary and Sentiment Reasoner, a dictionary-based sentiment analysis model, which demonstrated a 68% accuracy and F1 score of 0.75. Another study that used RoBERTa to classify five classes of mental illness in Reddit posts reported a model with an F1 score of 0.86 (Murarka et al., 2021). Last, researchers who forecasted star ratings from physician reviews written on RateMDs.com demonstrated an 84.6% accuracy and a mean F1 score of 0.83 with their RoBERTa model, which outperformed other NLP models (Jhaveri et al., 2022). In comparison, our interpersonal manner and technical competence models were each 90% accurate with weighted F1 scores of 0.89 and 0.90, respectively.

Although we did not compare our own RoBERTa models to other NLP algorithms, our accuracy scores perform equal to or better than prior methods used to code social judgments in online physician reviews. For example, in one study, investigators who hand-coded reviews for four broad thematic categories, including interpersonal manner and technical competence, reported an interrater reliability range of κ 0.8–1.0 (López et al., 2012). Another study, which used dictionary-based text analysis to code for positive and negative soft skills reported a mean accuracy of 0.76 (range 0.42–0.92) (Dunivin et al., 2020). Last, the rates of hand-coding agreement for interpersonal manner and technical competence among our own four investigators were Cohen’s κ of 0.84 and 0.77, respectively.

Our NLP classification models overcome several limitations of prior research using hand-coding and dictionary-based tools to identify the prevalence and valence of judgments in online physician reviews. Hand-coding, although considered the gold standard, is time intensive, which limits scalability. Prior studies using multiple coders could only analyze data from sample sizes of fewer than 1,000 physician reviews (López et al., 2012; Marrero et al., 2020). Dictionary-based approaches enable analysis of larger samples but are restricted by the keywords contained in their dictionaries. Dictionary-based or bag-of-words models may overlook misspellings and jargon (e.g., he butchered my surgery), leading to false negatives. They may also misidentify judgments about non-physician staff (e.g., front desk worker, nurse) or pick up words that have different

meanings in different contexts (e.g., he is thorough in his examinations vs. she gives thorough explanations), leading to false positives. Dictionary-based models also have difficulty distinguishing between words used positively or with negations (e.g., she was smart vs. she was not smart), which complicates valence estimates. Prior research on physician reviews using dictionary-based models could not determine valence, and only coded reviews that contained at least one pre-selected dictionary keyword (Chen et al., 2021; Dunivin et al., 2020; Gupta & Jordan, 2022; Saifee et al., 2022).

We validated our classification models by examining associations between our coded judgment valence and review star ratings and by comparing our coded judgments to findings from prior studies. We found positive interpersonal manner and technical competence judgments were associated with higher review star ratings whereas negative judgments were associated with lower review star ratings. Additionally, we found similar patterns of results with other studies that have examined the presence and valence of social judgments in online physician reviews. Future research should examine how both interpersonal manner and technical competence judgments vary depending on both physician gender and specialty.

We demonstrate that fine-tuning RoBERTa classification models to code interpersonal manner and technical competence judgments in online physician reviews offers a scalable, reliable, and accurate method for analyzing unstructured textual review data. To our knowledge, we are the first to use an advanced NLP algorithm to

code a large dataset of online physician reviews for social judgments. This algorithm successfully coded online physician reviews, suggesting that RoBERTa may also be used to code similar unstructured text, including reviews from other commercial physician review websites (e.g., RateMDs, ZocDoc) and from traditional Press Ganey patient-experience surveys. Future research is needed to ascertain the effectiveness of RoBERTa models with more far-afield text, such as crowdfunding campaigns, social media posts, and recommendation letters.

3.4.2 Limitations

There are several limitations to our study. First, the algorithm was trained on 2,000 hand-coded reviews. Although hand-coding is considered the gold standard and team members rigorously followed a coding framework, biases in how individual coders identified interpersonal manner and technical competence judgments may have influenced the NLP classification models. The imperfect interrater reliability present within the hand-coded dataset is evidence of differences between coders, which may have complicated the fine-tuning of the models. The 2,000 hand-coded reviews also represented only 0.6% of all reviews in the final dataset; thus, our models could have been overfitted to this relatively small training dataset.

Second, since the RoBERTa algorithm was not trained on a prebuilt dictionary but on a reference set of hand-coded reviews, it is difficult to determine specific words and phrases the models used when classifying interpersonal manner and technical

competence reviews. Third, although our models were not limited by prebuilt dictionaries, we only trained the models on reviews written in English. Reviews written in other languages, such as Spanish, were not translated and thus received codes of no interpersonal manner and no technical competence, despite potentially describing either social judgment. Reviews written in languages other than English, however, represented a small proportion of the total reviews in our sample.

Fourth, we excluded other judgments besides physicians' interpersonal manner and technical competence from our coding. Other judgments included global remarks that were categorized as neither interpersonal manner nor technical competence (e.g., "Would definitely recommend to others!" or "The worst") and system-level comments about the office, staff, or other aspects of the healthcare experience (e.g., "dingy building" or "his assistant was the most wonderful person I have ever met"). Last, although RoBERTa offers a more advanced NLP algorithm than dictionary-based methods, the algorithm may still not recognize cultural jargon. The first illustrative example of Table 5, in which RoBERTa did not recognize the connotation of the #metoo reference, demonstrates this limitation.

3.4.3 Conclusion

We coded a large dataset of online physician reviews for the presence and valence of social judgments using RoBERTa, a pretrained NLP classification algorithm. We trained and tested our models using a gold standard dataset of hand-coded reviews

and demonstrated that our models accurately and reliably coded interpersonal manner and technical competence. We also validated the algorithm by comparing our machine-coded dataset with review star ratings and results from prior literature. The RoBERTa algorithm overcomes text analysis limitations present in previous work by identifying social judgments in a broad range of physician reviews accurately and at scale.

4. Essay 3: Love them or hate them, female physicians' personalities matter: A large-scale text analysis of online physician reviews

4.1 Introduction

Patients judge their physicians on a variety of criteria. Patients value physicians who are empathetic, friendly, thorough, and knowledgeable, and dislike when physicians are inattentive and poor decision makers (Bendapudi et al., 2006; López et al., 2012; Menendez et al., 2015; Simsekler et al., 2021). Researchers have shown these criteria tend to fall into two overarching dimensions: interpersonal manner and technical competence (Howe et al., 2019; López et al., 2012). A physician's interpersonal manner includes their communication, care, and personal engagement, whereas their technical competence encompasses their efficiency, skill, and knowledge (Howe et al., 2019).

Research shows these two judgment dimensions undergird patient-centered care (Hall et al., 2014; Howe et al., 2019). Listed by the Institute of Medicine as one of the six key elements of high-quality care, patient-centered care is responsive to patient preferences and needs, and defines outcomes in terms of patients' values.(Catalyst, 2017; Epstein & Street, 2011b). To improve patient care quality and measure physicians' patient-centeredness, health systems and payers have implemented patient-experience surveys, which ask patients to provide feedback on their patient-physician interactions. Patients rate their physicians on interpersonal manner and technical competence items, such as how well their provider listened, explained, knew their medical history, and

followed-up with results (AHRQ, 2018, 2022a). These survey results, in part, also become tied to provider reimbursements, incentives, and promotions (CMS, 2023; Liao et al., 2020; Ranard et al., 2016).

Last, these types of patient judgments align with a broader framework from psychology on the two fundamental dimensions of social perception. According to the Stereotype Content Model, interpersonal impressions and group stereotypes form spontaneously along the dimensions of warmth and competence (Cuddy et al., 2008). Like interpersonal manner, warmth includes characteristics like friendliness, trustworthiness, and kindness; like technical competence, competence traits include intelligence, efficacy, and skill (Fiske et al., 2007). These dimensions have consistently emerged throughout social psychology literature, albeit under different labels, such as communal and agentic (Abele & Wojciszke, 2007), or warm/cold and intelligent/industrious (Asch, 1946). These two fundamental dimensions also account for nearly 90% of the variance in perceptions of social behaviors (Abele & Wojciszke, 2007), although warmth is dominant in impression formation. Warmth judgments not only carry more weight in evaluative judgments but also mainly decide the positivity or negativity of impressions (Abele & Wojciszke, 2007; Fiske et al., 2007; Wojciszke et al., 1998).

These fundamental dimensions are also central to group stereotypes. The Stereotype Content Model suggests we form ambivalent stereotypes of people, such that

men are expected to be high in competence and low in warmth, whereas women are expected to be high in warmth and low in competence (Cuddy et al., 2008). Additionally, warmth and competence traits are related to medical specialties. Surgeons are perceived as quick decision-makers, and skilled at fixing patients, while also arrogant, pushy, and assertive (Wainwright et al., 2022). The stereotypical surgeon is a man, and the field is regarded as an “old boys club” (Hill et al., 2014). These stereotypes reflect surgery’s statistical male dominance. In a 2021 AAMC report of active US physicians, orthopedic surgery represented the surgical specialty with the highest percentage of men at 94%; general surgery had the lowest percentage of men at 77% (AAMC, 2021). In contrast, men are less dominant, and in some cases, in the minority, among primary care specialties. For example, men comprise about 60% of family medicine and internal medicine physicians, and 35% of pediatricians. Stereotypes are less prevalent among primary care physicians (PCPs), but core to the culture of primary care are factors such as responsiveness, care, and communication (Grant et al., 2014).

Patients may value their physician’s interpersonal manner or technical competence more based on their physician’s specialty. For example, surgical patients may care about their surgeon’s procedural skill and less about their personality (Ashton-James et al., 2019). In contrast, primary care patients may prioritize physicians who are good communicators and are not condescending (Detz et al., 2013). However, issues may arise when patients value one dimension over the other when judging physicians of

different genders. Prior research suggests gender stereotypes in workplace evaluations disproportionately, negatively affect women. For example, gender stereotypes have shown to hinder women's hiring and promotion in academia (Kreitzer & Sweet-Cushman, 2021; Madera et al., 2009; Mitchell & Martin, 2018), support for women political candidates (Dolan, 2010), and women's workplace advancement (Correll et al., 2020; Lyness & Heilman, 2006). Gender stereotypes may similarly impact gender equity in medicine.

Online physician reviews provide an opportunity to study how patients judge their physicians and whether those judgments align with gender stereotypes. Unlike with star ratings, the unstructured text of these reviews allow patients to detail their experiences more completely (Greaves et al., 2014; Schlesinger et al., 2015), enabling patients to ascribe meaning and emotion to their patient-physician interactions (Greaves et al., 2014; Schlesinger et al., 2015).

Few studies have focused on potential gender stereotypes in patient judgments of physicians. Some studies have found that female physicians overall were more likely described with interpersonal language (Chen et al., 2021; Dunivin et al., 2020; Haynes et al., 2021; Thawani et al., 2019), whereas one study also found highly rated male physicians were more likely described with technical words (Haynes et al., 2021). Similarly, in a study of surgeons only, researchers showed female surgeons were more

likely praised for their social interaction whereas males were more likely praised for their technical skill (Marrero et al., 2020).

In this paper, we sought to address important limitations of previous research on gender stereotypes in physician reviews. First, we improved on the methodology of these prior studies using a trained machine-coding algorithm, a larger sample of reviews, and statistical procedures accounting for nested data. Prior researchers coded reviews using either hand coding, which limited sample size, or using pre-specified word dictionaries, which limited accuracy and valence identification (Chen et al., 2021; Dunivin et al., 2020; López et al., 2012). These studies also either restricted physician samples to select geographic locations (e.g., a single hospital or large cities), or limited the number of reviews analyzed per physician (Chen et al., 2021; Marrero et al., 2020).

Second, we assessed gender differences in reviews for physicians overall, as well as for PCPs versus surgeons. Prior studies investigated gender differences either among all physicians regardless of practicing specialty or within a single specialty subsample. Third, we examined gender differences not only in whether a physician received any interpersonal manner or technical competence judgment, but also based on the valence of that judgment. Most prior research focused exclusively on how patients described their physicians using interpersonal manner language; few studies also analyzed technical competence language (Haynes et al., 2021; Marrero et al., 2020). Additionally, we analyzed gender differences in how physicians' review star ratings responded to

patients' interpersonal manner and technical competence judgments. Whereas prior research has demonstrated gender differences in star ratings (Dunivin et al., 2020; Saifee et al., 2022; Thawani et al., 2019), few studies have interrogated how patients' judgments in online reviews are associated with gender differences in review ratings (Chen et al., 2021). Last, we framed our examination of gender differences in online physician reviews within social psychology theories about gender stereotypes, specifically informed by the Stereotype Content Model (Cuddy et al., 2008).

In line with the Stereotype Content Model, we hypothesized that interpersonal manner and technical competence judgments in online physician reviews would conform with traditional gender stereotypes. Thus, the purpose of this study was to interrogate how patients judged the interpersonal manner and technical competence of their physicians, whether those judgments aligned with gender stereotypes, and how those judgments affected physicians' review star ratings.

4.2 Methods

4.2.1 Data collection

We scraped physician profile, rating, and review data from Healthgrades.com in April 2020. Healthgrades is one of the largest commercial online physician rating and review websites and reports having a physician profile for every U.S. physician with an active profile on the National Provider Identifier Registry (Healthgrades, 2022). To examine gender and specialty differences in physician reviews at scale, we collected PCP

profiles associated with family medicine, internal medicine, and pediatrics, and surgeon profiles associated with general surgery; orthopedic surgery; and cosmetic, plastic, and reconstructive surgery. Physician profile characteristics collected included physician name, gender, age, practicing specialty, and primary office city and state. Rating variables were overall score and number of star ratings. Review variables included up to 20 of the physician's most recent reviews and accompanying star ratings, review submission dates, reviewer names and locations, and counts of people who flagged the reviews as helpful or unhelpful. The study was approved by our institutional review board and all data collected from Healthgrades were publicly available and aggregated for research purposes. We collected a total of 446,475 Healthgrades physician profiles and excluded physicians who did not receive at least one review.

4.2.2 Classifying reviews for interpersonal manner and technical competence

We used a sample of hand-coded reviews to train a classification algorithm called Robustly Optimized BERT Pre-Training Approach (i.e., RoBERTa) to code all reviews for the presence and valence of interpersonal manner and technical competence at scale. We previously tested and validated our advanced machine-learning algorithm, which was fine-tuned to develop two classification models for interpersonal manner and technical competence. The two models demonstrated 90% accuracy, with high precision, recall, and weighted F1 scores. For a detailed description of how we trained, tested, and

validated our classification models, please see Essay 2. We used Python and Google Colab to train and test our RoBERTa models and code our reviews.

4.2.3 Statistical analysis

To estimate gender and specialty differences in social judgments of physicians in online physician reviews, we used multilevel logistic models. In our initial set of regressions, we had six binary dependent variables of interest: (1) any judgment of interpersonal manner, (2) positive interpersonal manner, (3) negative interpersonal manner, (4) any judgment of technical competence, (5) positive technical competence, and (6) negative technical competence. The any judgment variables took the value of 1 for the presence of the judgment and 0 for absence. The positive judgment variables took the value of 1 for positive valence and 0 for negative valence or no judgment; we coded the negative judgment variables inversely. In our next set of regressions, we had two binary dependent variables of interest: (1) high-star review ratings, and (2) low-star review ratings. The high-star rating variable took the value of 1 for a 4- or 5-star review, and 0 for a 1–3-star review; the low-star rating variable was coded inversely. The main independent variable of interest was physician gender (female, male). Healthgrades did not list nonbinary or other genders on its website. Regressions controlled for physician age category, physician practicing state, and review word count. In models examining all physicians, regressions additionally controlled for physician specialty. Subsequent models stratified by physician specialty (PCP, surgery).

In the first set of regressions, we analyzed gender differences in social judgments received by physicians overall. In the second set of regressions, we stratified our sample based on primary care and surgery specialties. Our multilevel logistic models estimated the odds of judgment presence, while accounting for dependence of reviews nested within physicians (Sommet & Morselli, 2017). These models allowed us to estimate the odds ratios and predicted probabilities of a female physician receiving a social judgment in a review relative to their male counterpart. In the third set of regressions, we analyzed gender differences in receipt of high-star (4–5 star) and low-star (1–3 star) review ratings among physicians who received social judgments, stratified by specialty. These multilevel logistic models estimated the odds ratios and predicted probabilities of a female physician with a positive or a negative social judgment receiving a review star rating relative to their male counterpart. We used Stata to run our regressions.

4.3 Results

4.3.1 Characteristics of study sample

After excluding physicians who did not receive at least one review, our final sample included 345,053 reviews received by 167,150 U.S. physicians on Healthgrades.com. Reviews were submitted from 2016 to 2020. Most reviews (77%) had five-star ratings, followed by one-star ratings (19%). Figure 5 shows the distribution of review star ratings.

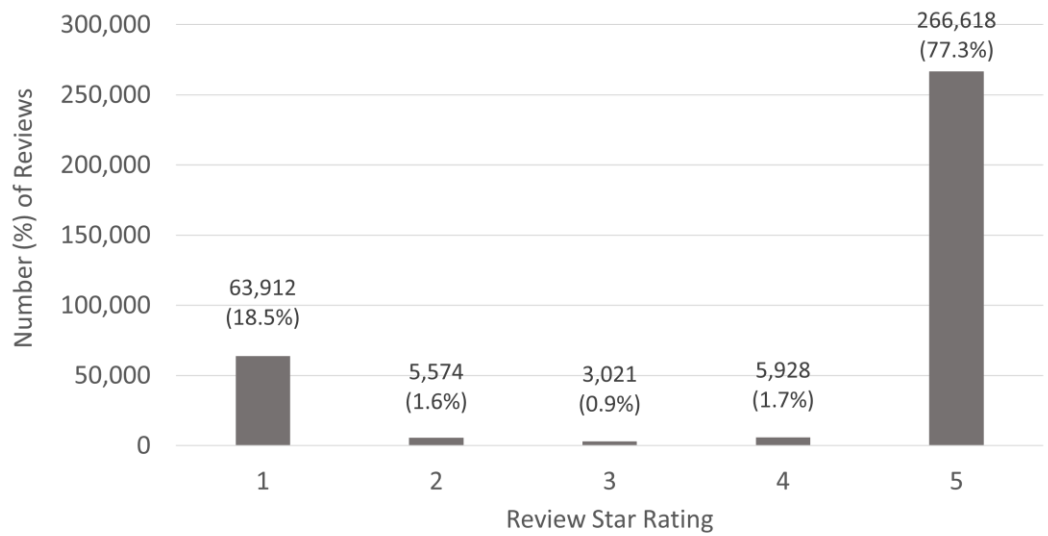


Figure 5 Distribution of review star ratings

Although there were more PCPs than surgeons in our sample (78% vs. 22%), the two specialties received nearly equal shares of reviews (49% received by PCPs vs. 51% received by surgeons). The mean length of a review with an interpersonal manner judgment was 52.80 words, and the mean length of a review with a technical competence judgment was 58.01 words. Reviews with positive interpersonal manner and technical competence judgments were approximately 20 words shorter in length than their negative counterparts. Table 6 presents physician characteristics as well as descriptive review statistics.

Table 6 Physician characteristics

	<i>Female Primary Care Physicians</i>	<i>Male Primary Care Physicians</i>	<i>Female Surgeons</i>	<i>Male Surgeons</i>	<i>Total</i>
Physicians n (%)	55,814 (33.4)	75,204 (45.0)	4,246 (2.5)	31,886 (19.1)	167,150 (100)
Reviews n (%)	75,941 (22.0)	93,326 (27.1)	13,769 (4.0)	162,017 (47.0)	345,053 (100)
Mean Reviews per Physician M [SD]	2.93 [4.06]	2.42 [3.58]	9.82 [7.33]	12.15 [7.21]	7.40 [7.45]
Review Rating M [SD]	3.94 [1.72]	4.01 [1.68]	4.38 [1.41]	4.37 [1.42]	4.18 [1.58]
Physician Age M [SD]	44.96 [19.16]	53.46 [18.44]	42.63 [20.33]	52.04 [19.23]	50.07 [19.30]

Notes:

We scraped a maximum of 20 reviews per physician, even though some physicians in our dataset could have had more than 20 reviews on Healthgrades.com.

Physicians who received at least one review were different from physicians who received star ratings but no written reviews. Compared to physicians who did not receive reviews, physicians with at least one review had, on average, slightly lower overall ratings (4.05 vs. 3.95) and three times as many ratings (3.44 vs. 12.85). Among physicians who received at least one review, male surgeons accounted for nearly one-third of all ratings. In contrast, male surgeons received about 14% of all ratings among physicians without reviews. Table 7 shows descriptive statistics of physicians with at

least one review compared to physicians excluded from our sample for receiving no reviews.

Table 7 Characteristics of physicians with and without reviews

	<i>Physicians with Reviews</i> N = 167,150	<i>Physicians without Reviews</i> N = 85,000
Physician Gender, n (%)	167,150 (100)	85,000 (100)
Female	60,060 (35.9)	32,220 (37.9)
Male	107,090 (64.1)	52,780 (62.1)
Physician Specialty, n (%)	167,150 (100)	85,000 (100)
Primary Care	131,018 (78.4)	73,607 (86.6)
Surgery	26,132 (21.6)	11,393 (13.4)
Physician Age, M [SD]	50.07 [19.30]	54.59 [19.54]
Number of Ratings, n (%)	2,147,770 (100)	292,477 (100)
Female PCP	563,946 (26.3)	100,591 (34.4)
Male PCP	861,409 (40.1)	147,725 (50.5)
Female Surgeon	60,212 (2.8)	4,656 (1.6)
Male Surgeon	662,203 (30.8)	39,505 (13.5)
Mean Number of Ratings per Physician, M [SD], range	12.85 [18.94], 1–1186	3.44 [3.82], 1–345
Female PCP	10.10 [11.87], 1–613	3.26 [3.32], 1–69
Male PCP	11.45 [13.76], 1–911	3.45 [3.53], 1–220
Female Surgeon	14.18 [29.79], 1–1186	3.37 [3.32], 1–24
Male Surgeon	20.77 [31.45], 1–1068	3.95 [5.89], 1–345
Mean Rating, M [SD]	3.95 [0.97]	4.05 [1.26]
Female PCP	3.85 [1.04]	4.05 [1.27]
Male PCP	3.92 [0.98]	4.04 [1.27]
Female Surgeon	4.25 [0.85]	4.29 [1.12]
Male Surgeon	4.15 [0.81]	4.06 [1.24]

Approximately three-fourths of all reviews in our sample had any mention of interpersonal manner. About 60% of reviews had positive interpersonal manner valence and 14% had negative interpersonal manner valence. Additionally, over 60% of reviews in our sample had any mention of technical competence. Over half the reviews had positive technical competence and 11% had negative technical competence. Figure 6 shows the percentage breakdown of interpersonal manner and technical competence reviews.

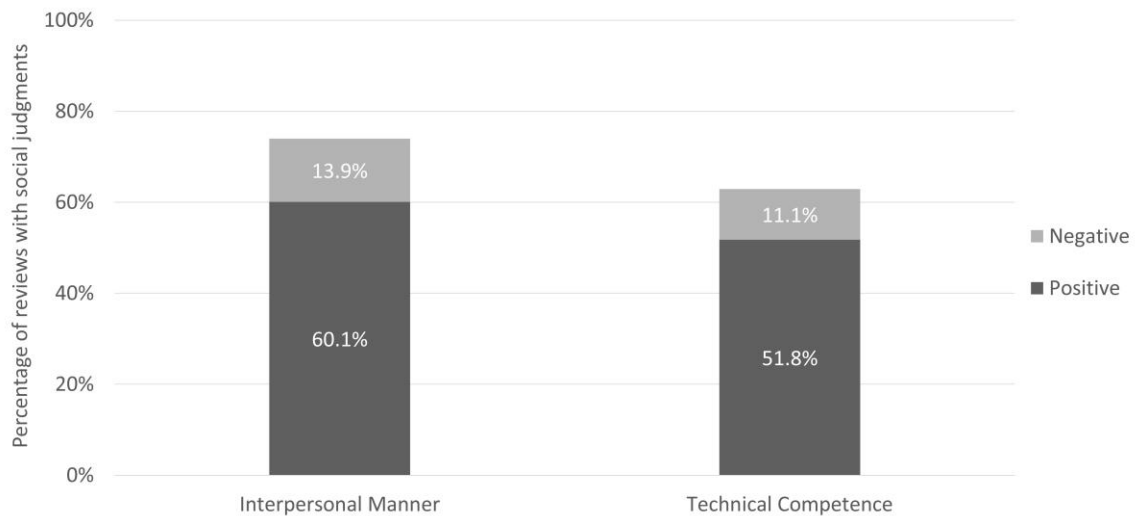


Figure 6 Percentage of reviews with positive and negative interpersonal manner judgments and positive and negative technical competence judgments

4.3.2 Gender differences in review judgments among physicians overall

Compared to male physicians, female physicians had 19% increased odds of receiving a review with any interpersonal manner judgment (95% CI [1.17, 1.22], $p < .001$)

and 25% increased odds of receiving a negative interpersonal manner judgment (95% CI [1.21, 1.29], $p < .001$). Female and male physicians had equal odds of receiving a review with any technical competence judgment (OR: 1.00, 95% CI [0.98, 1.03], $p = .052$), but female physicians had slightly higher odds of receiving a negative technical competence judgment (OR: 1.11, 95% CI [1.07, 1.15], $p < .001$; Figure 7).

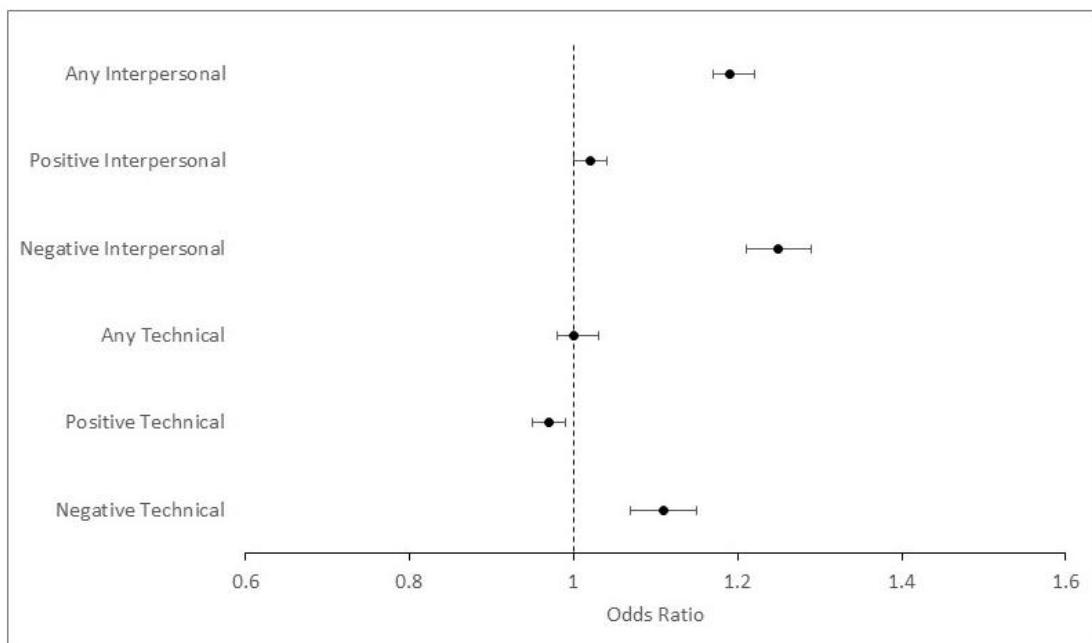


Figure 7 Odds ratio of receiving a social judgment for female physicians relative to male physicians

4.3.3 Gender differences in review judgments among primary care physicians and surgeons

4.3.3.1 Primary care physicians

Compared to male PCPs, female PCPs had 14% higher odds of receiving a review with any interpersonal manner judgment (95% CI [1.11, 1.17], $p < .001$) and 24% higher odds of receiving a negative interpersonal manner judgment (95% CI [1.19, 1.28], $p < .001$). Female and male PCPs had equal odds of receiving a review with any technical competence judgment (OR: 1.00, 95% CI [0.98, 1.03], $p = .733$). However, female PCPs had slightly higher odds of receiving a negative technical competence judgment (OR: 1.10, 95% CI [1.06, 1.14], $p < .001$; Figure 8a).

4.3.3.2 Surgeons

Compared to male surgeons, female surgeons had 35% higher odds of receiving a review with any interpersonal manner judgment (95% CI [1.29, 1.42], $p < .001$) and 30% higher odds of receiving a positive interpersonal manner review (95% CI [1.24, 1.36], $p < .001$). Female and male surgeons had equal odds of receiving a review with any technical competence judgment (OR: 1.02, 95% CI [0.97, 1.107], $p = .409$; Figure 8b).

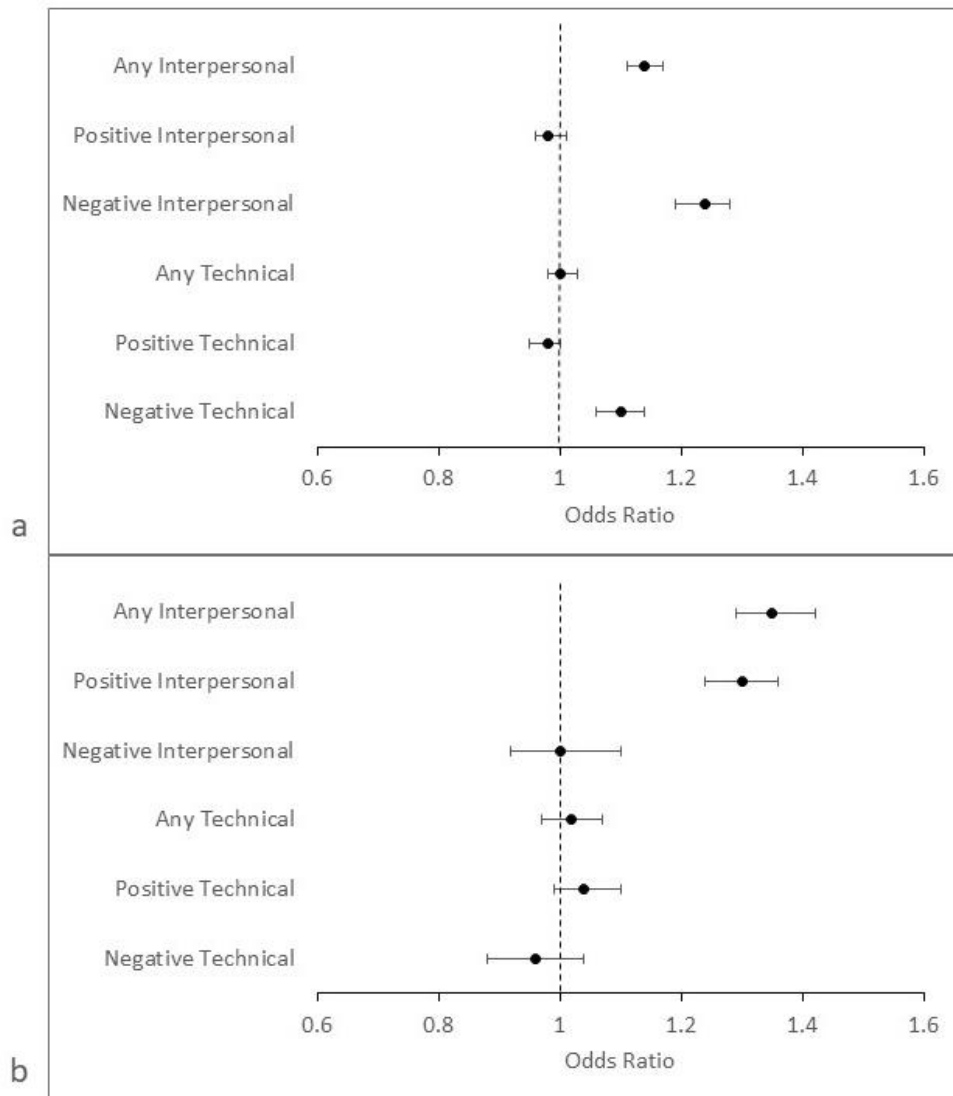


Figure 8 (a) Odds ratio of receiving a social judgment for female PCPs relative to male PCPs; (b) Odds ratio of receiving a social judgment for female surgeons relative to male surgeons

4.3.4 Gender differences in review star ratings among primary care physicians and surgeons who receive social judgments

4.3.4.1 Primary care physicians

Among PCPs who received positive interpersonal manner judgments, female and male PCPs had equal odds of receiving a high-star (4- or 5-star) rating (OR: 1.01, 95% CI [0.90, 1.13], $p=.835$). Among PCPs who received positive technical competence judgments, female PCPs had 17% decreased odds of receiving a high-star rating (95% CI [0.71, 0.96], $p=.014$; Figure 9a).

Among PCPs who received negative interpersonal manner judgments, female PCPs had 60% increased odds of receiving a low-star (1-, 2-, or 3-star) rating (95% CI [1.37, 1.88], $p<.001$). Among PCPs who received negative technical competence judgments, female PCPs had 67% increased odds of receiving a low-star rating (95% CI [1.38, 2.02], $p<.001$; Figure 9b).

4.3.4.2 Surgeons

Among surgeons who received positive interpersonal manner judgments, female and male surgeons had equal odds of receiving a high-star rating (OR: 1.09, 95% CI [0.83, 1.44], $p=.532$). Likewise, among surgeons who received positive technical competence judgments, female and male surgeons had equal odds of receiving a high-star rating (OR: 1.04, 95% CI [0.75, 1.43], $p=.819$; Figure 9a).

Among surgeons who received negative interpersonal manner judgments, female and male surgeons had equal odds of receiving a low-star rating (OR: 1.11, 95%

CI [0.78, 1.59], $p=.563$). However, among surgeons who received negative technical competence judgments, female surgeons had 52% increased odds of receiving a low-star rating (95% CI [1.14, 2.04], $p=.005$; Figure 9b). See Appendix for detailed regression results underlying Figures 7 through 9.

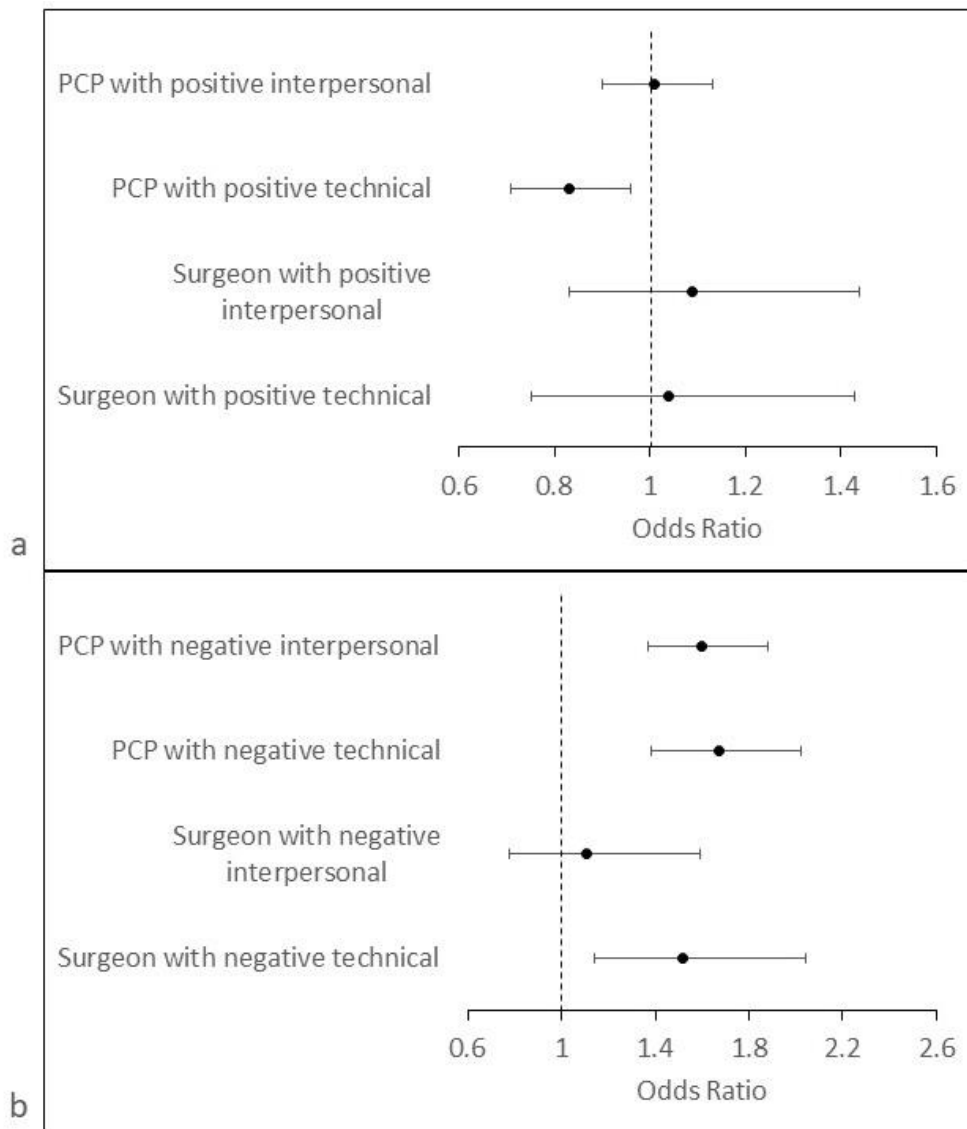


Figure 9(a) Odds ratio of receiving a high-star rating for female PCPs and female surgeons who receive positive social judgments, relative to their male counterparts; (b) Odds ratio of receiving a low-star rating for female PCPs and female surgeons who receive negative social judgments, relative to their male counterparts

4.4 Discussion

Online physician reviews allow patients to provide immediate, publicly available evaluations of their physicians, which prospective patients use and trust when choosing among physicians (Gao et al., 2012; Holliday et al., 2017; Hong et al., 2019). By analyzing these narrative reviews, we aimed to discern whether patients' judgments of their physicians' interpersonal manner and technical competence aligned with gender stereotypes.

Our results provide insight into who is more likely to receive which social judgments, and how these social judgments affect review star ratings. We demonstrated three main findings. First, male physicians were not more likely than females to receive patient judgments of their technical competence. Second, female physicians were more likely than males to receive patient judgments of their interpersonal manner. However, whether female physicians were relatively more likely to receive patient praise or criticism for their interpersonal manner depended on their specialty. Third, female PCPs and surgeons were penalized more than their male counterparts in their star ratings when they received criticism for their interpersonal manner or technical competence.

To our knowledge, our study evaluates the largest number of online physician reviews to date and is the first to examine gender differences in interpersonal manner and technical competence judgments received by PCPs and surgeons. Unlike numerical ratings, narrative reviews allow patients to describe their physician-patient interactions

more completely, and to express both positive and negative emotions (Greaves et al., 2014; Schlesinger et al., 2015). Thus, these reviews provide a rich resource to understand how patients judge their physicians and whether those judgments align with gender stereotypes. In this study, we used an advanced machine-learning algorithm to code the presence and valence of these social judgments in reviews at scale with high accuracy, precision, and recall. Our algorithm allowed us to overcome limitations present in previous studies that used hand-coding and dictionary-based methods (Chen et al., 2021; Dunivin et al., 2020; Haynes et al., 2021; Marrero et al., 2020).

4.4.1 Finding 1: Patients were equally likely to judge male and female physicians' technical competence

Inconsistent with our hypothesis, we found male physicians were not more likely to receive reviews with technical competence judgments than females. This finding was consistent across all physicians, among PCPs only, and among surgeons only.

Few prior studies have examined gender differences in patients' technical competence judgments of their physicians in reviews. One study among all physicians found patients more often ascribed agentic qualities to high-rated male physicians compared to females (Haynes et al., 2021). Another study found male surgeons were more likely than females to receive positive reviews of their technical skill (Marrero et al., 2020). In contrast, we found no substantial gender differences in physicians' receipt of any technical competence judgments or positive technical competence judgments. We

did, however, find that female physicians overall and female PCPs were more likely to receive criticism of their technical competence.

Although we found male and female physicians were equally likely to receive praise for technical competence, we did not find gender equality in criticism. More research is needed to discern why female PCPs, but not female surgeons, were more likely than their male counterparts to receive criticism for their technical competence. One explanation could be that female PCPs have worse technical skills than male PCPs. Another interpretation could be that patients more easily noticed poor technical performance in female PCPs compared to male PCPs because poor performance confirmed patients' stereotype bias that women are not as technically competent (Fiske et al., 2018). Alternatively, female PCPs' poor interpersonal manner may have influenced negative impressions of their technical competence. Impression formation research suggests warmth judgments mainly decide the positivity or negativity of first impressions (Abele & Wojciszke, 2007; Fiske et al., 2007; Wojciszke et al., 1998); thus female PCPs, who were more likely than males to receive negative interpersonal manner judgments, may have also been more likely to receive negative technical competence judgments. We did not observe a similar gender difference in negative interpersonal manner judgments among surgeons.

4.4.2 Finding 2a: Patients were more likely to judge their female physicians' interpersonal manner

Consistent with our hypothesis, we found female physicians overall, as well as female PCPs and female surgeons, were more likely to receive interpersonal manner judgments than their male counterparts. This finding corroborates prior studies on physician reviews, which similarly found female physicians were more likely to receive interpersonal manner judgements (Chen et al., 2021; Dunivin et al., 2020; Thawani et al., 2019). Our finding is also consistent with evaluations of women compared to men in other fields, such as in faculty recommendation letters (Madera et al., 2009).

4.4.3 Finding 2b: How patients judged female physicians' interpersonal manner depended on physicians' specialties

Our machine-learning algorithm allowed us to explore judgment valence at scale, advancing beyond prior studies, which relied on review star ratings as proxies for positive and negative sentiment (Dunivin et al., 2020; Haynes et al., 2021; Marrero et al., 2020). We found that female physicians overall and female PCPs were more likely criticized for interpersonal manner, whereas female surgeons were more likely praised.

These within-specialty results suggest that in a stereotypically “technical” specialty, like surgery, females were more likely than males to receive praise for their interpersonal manner. In contrast, in a stereotypically “warm” specialty, like primary care, females were more likely to receive criticism for their interpersonal manner. Thus,

whether female physicians were penalized or advantaged for their interpersonal manner relative to males depended on whether they practiced primary care or surgery.

Our results are consistent with prior empirical research, which shows both positive and negative interpersonal manner traits were more likely mentioned in reviews of female physicians than those of males (Dunivin et al., 2020). We also corroborate the results of a study among surgeons, which found female surgeons were more likely to receive positive reviews of social interaction (Marrero et al., 2020). No studies focused on gender differences in judgments among PCPs only.

The Stereotype Content Model and Role Congruity Theory provide a rationale for why female PCPs, as well as female physicians overall, were more likely criticized for their interpersonal manner. Complementing the Stereotype Content Model, Role Congruity Theory suggests people form prescriptive beliefs about gender roles, such that individuals whose behaviors violate their gender stereotype are penalized (Eagly & Karau, 2002). Female PCPs face two compounding interpersonal stereotypes: Both women and PCPs are expected to have high warmth (Cuddy et al., 2008). Thus, a female PCP perceived to have poor interpersonal manner may have been particularly noticeable and judged more harshly than male PCPs (Cuddy et al., 2008; Eagly & Karau, 2002), even if she demonstrated similar interpersonal manner traits as her male counterpart.

An alternative explanation for why female PCPs and female physicians overall were more likely criticized for both interpersonal manner and technical competence is

that they objectively provided poorer patient-centered care than their male counterparts. However, hospital-based research has shown female physicians are more patient-centered than male physicians, providing a reason to doubt the objective differences argument (Hall et al., 2014; Roter & Hall, 2004). In one study, researchers measured patient-centeredness on the following scales: interested, friendly, engaged, sympathetic, dominant (Hall et al., 2014). Although they found females exhibited more patient-centered behavior, patient-centeredness predicted patient satisfaction more strongly in male physicians than in females. In other words, female physicians did not receive as much credit for their patient-centeredness (Hall et al., 2014; Mast & Kadji, 2018).

The Stereotype Content Model and Role Congruity Theory, however, do not provide adequate rationale for why female surgeons were more likely praised for their interpersonal manner. Prior Role Congruity research demonstrates women who exhibited agentic traits and occupied masculinely defined positions received backlash for being insufficiently interpersonal (Eagly & Karau, 2002; Rudman & Glick, 1999). Accordingly, we would have expected women practicing in surgery's "old boys club" would be more likely criticized than male surgeons for their interpersonal manner.

However, the Stereotype Content Model acknowledges that stereotypes for subgroups may differ from those of their broader categories, such that female surgeons and female PCPs may be perceived differently from each other and from female physicians overall (Fiske et al., 2018). Thus, one explanation for why female surgeons

were more likely praised rather than criticized for their interpersonal manner comes from research on double standards for competence. A double standard exists when members of subgroups must meet stricter requirements in evaluations of possessing an attribute (e.g., competence) (Foschi, 2000). Although double standards can hinder women's career advancement, they can also advantage women who perform successfully (Ditonto et al., 2014; Heilman, 2012; Lyness & Thompson, 2000; Rosette & Tost, 2010). Female surgeons, who conformed to their prescriptive interpersonal gender norm while simultaneously excelling in a male-dominated field characterized by high levels of technical competence, may therefore have been evaluated more favorably than their male counterparts.

Although the Stereotype Content Model posits most groups have ambivalent stereotypes, the model suggests that some groups are high in both perceived warmth and competence (Fiske et al., 2018). Female surgeons may be one of these subgroups. Prior research has suggested top women leaders also constitute one of these subgroups (Rosette & Tost, 2010). Thus, warm female surgeons who demonstrated competence despite double standards, may have been particularly noticeable and judged more favorably than male surgeons for their interpersonal manner.

4.4.4 Finding 3: Patients were less likely to reward, more likely to penalize female physicians in ratings based on interpersonal manner and technical competence

We found that when patients praised their PCPs' technical competence, this resulted in asymmetrical receipt of high-star ratings. Female PCPs were not rewarded as much as males when receiving praise for their technical competence. Additionally, patients' criticism of their PCPs' interpersonal manner and technical competence led to higher odds of low-star ratings for female PCPs compared to their male counterparts. Thus, female PCPs were more likely to be penalized in their star ratings when patients commented on their poor interpersonal manner or technical competence.

Patients were equally as likely to highly rate their male and female surgeons when they praised their surgeons' interpersonal manner and technical competence. However, as with female PCPs, female surgeons were more likely than males to be penalized in their star ratings when patients criticized their technical competence.

Previous research has shown that female physicians receive worse ratings than male physicians (Dunivin et al., 2020; Saifee et al., 2022; Thawani et al., 2019). However, few prior studies have investigated gender differences in how review star ratings respond to patient judgments of the physicians' interpersonal manner or technical competence. One study found physicians described with communal language were more likely to receive higher star ratings, but gender did not significantly moderate the

association (Chen et al., 2021). This study, however, only focused on positive communal language and used a limited prebuilt word dictionary.

Other research examining surgeon referrals found asymmetries consistent with our findings, such that female surgeons faced harsher consequences for perceived poor technical competence compared to males. After a bad outcome, female surgeons experienced a 34% drop in referrals, whereas male surgeons experienced only slight stagnation in referrals (Sarsons, 2017). Additionally, after patient death, female surgeons lost about 60% of their Medicare billings from the referring physicians, compared to males, who lost 30%. Thus, female surgeons incurred both substantial referral and pay penalties compared to male surgeons for the same poor outcome.

4.4.5 Limitations

This study has several limitations. First, we used data from a single online physician review website, which may limit the generalizability of our findings. Second, our physician sample may also limit generalizability. The physicians included in our study were different from those excluded for having no reviews. For example, physicians with reviews received nearly four times as many star ratings as physicians without reviews and had lower average ratings than those excluded. Third, although our machine-learning models were 90% accurate, we cannot discern whether our models were more or less accurate for reviews received by PCPs or surgeons, or by female or male physicians. Fourth, we cannot conclude whether the gender differences observed

reflect objective differences in care, differences in patient expectations of care, or differences in patient sensitivity to how male and female physicians exhibited interpersonal manner and technical competence. Further research will need to tease apart these potential explanations.

Although some researchers have attempted to objectively measure patient-centeredness among male and female physicians (Hall et al., 2014), and others have demonstrated male and female physicians have different perceived communication styles depending on the patient's gender (Hall & Roter, 2002; Mast et al., 2007), further research is required to explain the mechanisms behind our findings. Additionally, future research should examine how these differences in physician reviews influence how prospective patients use these reviews to select physicians. Although a larger proportion of reviews in our sample focused on the physician's interpersonal manner, one experimental study found patients preferred technical qualities in their PCPs (Fung et al., 2005), whereas another study concluded patients preferred technical or interpersonal surgeons depending on the type of surgery (Dusch et al., 2014). More research is needed to better understand what prospective patients look for when choosing among physicians, and whether those traits differ by physician characteristics.

4.4.6 Conclusion

How patients judge their physicians in online reviews matters to prospective patients who trust these reviews when choosing a physician, and to physicians who use

these reviews for both feedback and quality improvement. In this study, we found patients were not more likely to judge the technical competence of their male or female physicians, but they homed in on their female physicians' interpersonal manner. Whether patients were relatively more likely to praise or criticize their female physicians' interpersonal manner depended on their physician's practicing specialty. Last, female physicians were more likely to be penalized in their star ratings when receiving reviews criticizing their interpersonal manner or technical competence compared to similarly criticized male physicians.

5. Conclusion

In this dissertation, I used mixed methods to explore patients' judgments of their physicians in online physician reviews and whether those judgments aligned with gender stereotypes. Essay 1 developed an elaborated theoretical framework to describe the range of factors patients use to judge the interpersonal manner and technical competence of their physicians. Essay 2 trained, tested, and validated an advanced machine-learning algorithm to classify the presence and valence of patients' interpersonal manner and technical competence judgments in online physician reviews. Essay 3 analyzed how patients' interpersonal manner and technical competence judgments of their physicians aligned with gender stereotypes and affected review star ratings.

In Essay 1, I developed an elaborated theoretical framework to distinguish the factors comprising patients' interpersonal manner and technical competence judgments of their physicians. To develop this framework, I expanded on prior grounded theory work by Lopez et al. (2012) using a qualitative content analysis of 2,000 reviews received by 2,000 distinct physicians. I built on this theory using a larger, new dataset of physician reviews, purposely sampled to equally represent primary care physicians and surgeons, male and female physicians, and low- and high-rated reviews. I offered thick descriptions and illustrative quotations of the predominant factors within the dimensions of interpersonal manner and technical competence, added new factors

missing from Lopez et al.'s original framework, and removed factors that were too ambiguous to categorize. Patients valued their physicians for demonstrating interpersonal manner qualities such as bedside manner, empathy, availability, and good communication. Patients also appreciated physicians with technical competence qualities such as diagnostic skill, offering treatment and referrals, and successful treatment outcomes.

In Essay 2, I fine-tuned an advanced natural language processing algorithm called RoBERTa to code a large dataset of online physician reviews for the presence and valence of interpersonal manner and technical competence judgments. I used the 2,000 hand-coded reviews from Essay 1 to train and test two classification models, one for interpersonal manner and one for technical competence. Both models were 90% accurate with high precision and recall. These models were as or more accurate than other text coding approaches and comparable to other RoBERTa models. I used these two models to code the entire dataset of 345,053 online reviews written for 167,150 physicians and validated the models using star ratings and findings in prior literature.

In Essay 3, I examined whether patients' judgments of their physicians' interpersonal manner and technical competence aligned with traditional gender stereotypes as described in the Stereotype Content Model. Using the 345,053 machine-coded reviews from Essay 2, I estimated multilevel logistic regressions to analyze physician gender differences in interpersonal manner and technical competence

judgments received by physicians and how those judgments affected review star ratings. Consistent with traditional gender stereotypes, female physicians overall and within primary care and surgery were more likely to receive interpersonal manner reviews than their male counterparts. However, male physicians were not more likely to receive technical competence reviews. Whether female physicians were relatively more likely to receive praise or criticism for their interpersonal manner depended on their practicing specialty. In stereotypically warm specialties, like primary care, females were penalized for seeming cold, whereas in stereotypically technical specialties, like surgery, females were advantaged for appearing warm.

Additionally, female and male physicians who received interpersonal manner and technical competence judgments were asymmetrically rewarded and penalized in their review star ratings. Female primary care physicians were less likely to receive high-star ratings when praised for technical skills and more likely to receive low-star ratings when criticized for interpersonal and technical skills. Likewise, female surgeons were more likely to receive low-star ratings when criticized for technical skills.

In this dissertation, I analyzed online physician reviews using qualitative analysis, machine learning, and logistic regression to better understand not only how patients judge their physicians but also whether those judgments align with gender stereotypes and differentially affect review star ratings. I contribute to the literature with an expanded framework to examine patients' interpersonal manner and technical

competence judgments of their physicians and with an advanced machine-learning algorithm, which enables accurate large-scale classification of the presence and valence of judgments in reviews. Framing this research within psychology theories of social perception and gender stereotypes, I offer nuanced insights on how patients evaluate male and female primary care physicians and surgeons. Continued research is needed to tease apart explanations for why gender differences in physician reviews exist and to better understand how prospective patients use online physician reviews to choose among physicians.

Appendix for essay 3

Table 8 Odds ratios and predicted probabilities of female and male physicians receiving any, positive, or negative interpersonal manner judgments and any, positive, or negative technical competence judgments

	<i>Any interpersonal manner</i>		<i>Positive interpersonal manner</i>		<i>Negative interpersonal manner</i>		<i>Any technical competence</i>		<i>Positive technical competence</i>		<i>Negative technical competence</i>	
	OR [95% CI]	PP (%)	OR [95% CI]	PP (%)	OR [95% CI]	PP (%)	OR [95% CI]	PP (%)	OR [95% CI]	PP (%)	OR [95% CI]	PP (%)
Physician Gender												
Female	1.19*** [1.17, 1.22]	76.4	1.02 [1.00, 1.04]	59.9	1.25*** [1.21, 1.29]	16.2	1.00 [0.98, 1.03]	63.1	0.97* [0.95, 0.99]	50.8	1.11*** [1.07, 1.15]	12.5
Male	1.00	73.4	1.00	59.5	1.00	14.0	1.00	63.0	1.00	51.4	1.00	11.6
Observations	345,053											
Clusters	167,150											

Notes: Odds ratios (OR) and 95% Confidence Intervals (CI) are from multilevel logistic regressions, controlling for physician specialty, physician age category, primary office state, and review word count. Predicted probabilities (PP) are obtained using the “margins” command in Stata. *p<.05; **p<.01; ***p<.001

Table 9 Odds ratios and predicted probabilities of female and male PCPs and surgeons receiving any, positive, or negative interpersonal manner judgments and any, positive, or negative technical competence judgments

	<i>Any interpersonal manner</i>		<i>Positive interpersonal manner</i>		<i>Negative interpersonal manner</i>		<i>Any technical competence</i>		<i>Positive technical competence</i>		<i>Negative technical competence</i>	
	OR [95% CI]	PP (%)	OR [95% CI]	PP (%)	OR [95% CI]	PP (%)	OR [95% CI]	PP (%)	OR [95% CI]	PP (%)	OR [95% CI]	PP (%)
PCP Gender												
Female PCP	1.14*** [1.11, 1.17]	79.7	0.98 [0.96, 1.01]	60.6	1.24*** [1.19, 1.28]	19.0	1.00 [0.98, 1.03]	53.2	0.98* [0.95, 1.00]	41.3	1.10*** [1.06, 1.14]	12.1
Male PCP	1.00	77.8	1.00	61.1	1.00	16.6	1.00	53.2	1.00	41.8	1.00	11.3
Observations Clusters						169,267 131,018						
Surgeon Gender												
Female Surgeon	1.35*** [1.29, 1.42]	74.9	1.30*** [1.24, 1.36]	63.5	1.00 [0.92, 1.10]	11.7	1.02 [0.97, 1.07]	72.7	1.04 [0.99, 1.10]	61.3	0.96 [0.88, 1.04]	11.5
Male Surgeon	1.00	69.3	1.00	57.9	1.00	11.7	1.00	72.4	1.00	60.4	1.00	11.9
Observations Clusters						175,786 36,132						

Notes: Odds ratios (OR) and 95% Confidence Intervals (CI) are from multilevel logistic regressions stratified by physician specialty, controlling for physician age category, primary office state, and review word count. Predicted probabilities (PP) are obtained using the “margins” command in Stata. *p<.05; **p<.01; ***p<.001

Table 10 Odds ratios and predicted probabilities of female and male PCPs and surgeons receiving high-star and low-star review ratings when judged for their interpersonal manner and technical competence

<i>High-star review ratings (4–5 stars)</i>		
	OR [95% CI]	PP (%)
Positive interpersonal manner		
Female PCP	1.01 [0.90, 1.13]	91.3
Male PCP	1.00	93.0
Positive technical competence		
Female PCP	0.83* [0.71, 0.96]	84.4
Male PCP	1.00	87.1
Observations		169,267
Clusters		131,018
Positive interpersonal manner		
Female surgeon	1.09 [0.83, 1.44]	94.2
Male surgeon	1.00	95.4
Positive technical competence		
Female surgeon	1.04 [0.75, 1.43]	94.0
Male surgeon	1.00	94.6
Observations		175,786
Clusters		36,132
<i>Low-star review ratings (1–3 stars)</i>		

	OR [95% CI]	PP (%)
Negative interpersonal manner		
Female PCP	1.60*** [1.37, 1.88]	86.9
Male PCP	1.00	79.0
Negative technical competence		
Female PCP	1.67*** [1.38, 2.02]	64.1
Male PCP	1.00	55.3
Observations		169,267
Clusters		131,018
Negative interpersonal manner		
Female surgeon	1.11 [0.78, 1.59]	63.2
Male surgeon	1.00	59.6
Negative technical competence		
Female surgeon	1.52** [1.24, 2.04]	56.9
Male surgeon	1.00	49.0
Observations		175,786
Clusters		36,132

Notes: Odds ratios (OR) and 95% Confidence Intervals (CI) are from multilevel logistic regressions stratified by physician specialty, controlling for physician age category, primary office state, and review word count. Predicted probabilities (PP) are obtained using the “margins” command in Stata. *p<.05; **p<.01; ***p<.001

References

- AAMC. (2021). *Physician Specialty Data Report: Active Physicians by Sex and Specialty, 2021*.
<https://www.aamc.org/data-reports/workforce/interactive-data/active-physicians-sex-specialty-2021>
- Abele, A. E., & Wojciszke, B. (2007). Agency and communion from the perspective of self versus others. *Journal of personality and social psychology, 93*(5), 751.
- Abele, A. E., & Wojciszke, B. (2014). Communal and agentic content in social cognition: A dual perspective model. In *Advances in experimental social psychology* (Vol. 50, pp. 195-255). Elsevier.
- AHRQ. (2015). *An Overview of Version 3.0 of the CAHPS Clinician & Group Survey*.
ahrq.gov: AHRQ Retrieved from
https://www.ahrq.gov/sites/default/files/wysiwyg/cahps/surveys-guidance/cg/about/cg_3-0_overview.pdf
- AHRQ. (2018, December 2018). *CAHPS Adult Hospital Survey*. Agency for Healthcare Research and Quality. Retrieved February 6, 2023 from
https://www.ahrq.gov/cahps/surveys-guidance/hospital/about/adult_hp_survey.html
- AHRQ. (2022a, November 2022). *CAHPS Clinician & Group Survey*. Agency for Healthcare Research and Quality. Retrieved February 6, 2023 from
<https://www.ahrq.gov/cahps/surveys-guidance/cg/index.html>
- AHRQ. (2022b, December 2022). *Six Domains of Healthcare Quality*. Retrieved February 20, 2023 from <https://www.ahrq.gov/talkingquality/measures/six-domains.html>
- Anhang Price, R., Quigley, D. D., Hargraves, J. L., Sorra, J., Becerra-Ornelas, A. U., Hays, R. D., Cleary, P. D., Brown, J., & Elliott, M. N. (2022). A systematic review of strategies to enhance response rates and representativeness of patient experience surveys. *Medical care, 60*(12), 910-918.

- Asanad, K., Parameshwar, P. S., Houman, J., Spiegel, B. M., Daskivich, T. J., & Anger, J. T. (2018). Online physician reviews in female pelvic medicine and reconstructive surgery: what do patients really want? *Female pelvic medicine & reconstructive surgery, 24*(2), 109-114.
- Asch, S. E. (1946). Forming impressions of personality. *The Journal of Abnormal and Social Psychology, 41*(3), 258.
- Ashton-James, C. E., Forouzanfar, T., & Costa, D. (2019). The contribution of patients' presurgery perceptions of surgeon attributes to the experience of trust and pain during third molar surgery. *Pain reports, 4*(3).
- Badejo, M. A., Ramtin, S., Rossano, A., Ring, D., Koenig, K., & Crijns, T. J. (2022). Does adjusting for social desirability reduce ceiling effects and increase variation of patient-reported experience measures? *Journal of Patient Experience, 9*, 23743735221079144.
- Bakhsh, W., & Mesfin, A. (2014). Online ratings of orthopedic surgeons: analysis of 2185 reviews. *Am J Orthop, 43*(8), 359-363.
- Ballard, A. O., DeTamble, R., Dorsey, S., Heseltine, M., & Johnson, M. (2022). Dynamics of polarizing rhetoric in congressional tweets. *Legislative Studies Quarterly*.
- Bardach, N. S., Asteria-Peñaloza, R., Boscardin, W. J., & Dudley, R. A. (2013). The relationship between commercial website ratings and traditional hospital performance measures in the USA. *BMJ quality & safety, 22*(3), 194-202.
- Bendapudi, N. M., Berry, L. L., Frey, K. A., Parish, J. T., & Rayburn, W. L. (2006). Patients' perspectives on ideal physician behaviors. *Mayo Clinic Proceedings*,
- Bradley, E. H., Curry, L. A., & Devers, K. J. (2007). Qualitative data analysis for health services research: developing taxonomy, themes, and theory. *Health services research, 42*(4), 1758-1772.

- Burkle, C. M., & Keegan, M. T. (2015a). Popularity of internet physician rating sites and their apparent influence on patients' choices of physicians. *BMC health services research, 15*(1), 416.
- Burkle, C. M., & Keegan, M. T. (2015b). Popularity of internet physician rating sites and their apparent influence on patients' choices of physicians. *BMC health services research, 15*(1), 1-7.
- Carbonell, G., & Brand, M. (2018). Choosing a physician on social media: Comments and ratings of users are more important than the qualification of a physician. *International Journal of Human-Computer Interaction, 34*(2), 117-128.
- Carman, K., Lawrence, W., & Siegel, J. (2019, March 5, 2019). The 'New' Health Care Consumerism. *Health Affairs Forefront*.
<https://www.healthaffairs.org/doi/10.1377/forefront.20190304.69786/>
- Catalyst, N. (2017). What is patient-centered care? *NEJM Catalyst, 3*(1).
- Chen, H., Pierson, E., Schmer-Galunder, S., Altamirano, J., Jurafsky, D., Leskovec, J., Fassiotto, M., & Kothary, N. (2021). Gender differences in patient perceptions of physicians' communal traits and the impact on physician evaluations. *Journal of Women's Health, 30*(4), 551-556.
- Chen, J., Presson, A., Zhang, C., Ray, D., Finlayson, S., & Glasgow, R. (2018). Online physician review websites poorly correlate to a validated metric of patient satisfaction. *Journal of Surgical Research, 227*, 1-6.
- CMS. (2021, December 1, 2021). *HCAHPS: Patients' Perspectives of Care Survey*. CMS. Retrieved February 20, 2023 from <https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/HospitalHCAHPS>
- CMS. (2023, January 25, 2023). *Consumer Assessment of Healthcare Providers & Systems (CAHPS)*. Retrieved February 14, 2023 from <https://www.cms.gov/research-statistics-data-and-systems/research/cahps>

- Correll, S. J., Weisshaar, K. R., Wynn, A. T., & Wehner, J. D. (2020). Inside the black box of organizational life: The gendered language of performance assessment. *American Sociological Review, 85*(6), 1022-1050.
- Cuddy, A. J., Fiske, S. T., & Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. *Advances in experimental social psychology, 40*, 61-149.
- Cuddy, A. J., Glick, P., & Beninger, A. (2011). The dynamics of warmth and competence judgments, and their outcomes in organizations. *Research in organizational behavior, 31*, 73-98.
- Daskivich, T. J., Houman, J., Fuller, G., Black, J. T., Kim, H. L., & Spiegel, B. (2018). Online physician ratings fail to predict actual performance on measures of quality, value, and peer review. *Journal of the American Medical Informatics Association, 25*(4), 401-407.
- Davis, K., Carbone, R., Fredericks, J., Lujan, J., & Basdeo, A. (2021). What makes a good doctor: a qualitative study of patient perspectives.
- De Valck, C., Bensing, J., Bruynooghe, R., & Batenburg, V. (2001). Cure-oriented versus care-oriented attitudes in medicine. *Patient education and counseling, 45*(2), 119-126.
- Detz, A., López, A., & Sarkar, U. (2013). Long-term doctor-patient relationships: patient perspective from online reviews. *Journal of medical Internet research, 15*(7), e131.
- Diekmann, A. B., & Eagly, A. H. (2000). Stereotypes as dynamic constructs: Women and men of the past, present, and future. *Personality and Social Psychology Bulletin, 26*(10), 1171-1188.
- Ditonto, T. M., Hamilton, A. J., & Redlawsk, D. P. (2014). Gender stereotypes, information search, and voting behavior in political campaigns. *Political Behavior, 36*, 335-358.

- Dolan, K. (2010). The impact of gender stereotyped evaluations on support for women candidates. *Political Behavior*, 32, 69-88.
- Donnally III, C. J., Roth, E. S., Li, D. J., Maguire Jr, J. A., McCormick, J. R., Barker, G. P., Rivera, S., & Lebowhl, N. H. (2018). Analysis of internet review site comments for spine surgeons: how office staff, physician likeability, and patient outcome are associated with online evaluations. *Spine*, 43(24), 1725-1730.
- DukeHealth. *Duke Health Ratings and Reviews*. DukeHealth. Retrieved October 10, 2020 from <https://www.dukehealth.org/duke-health-ratings-and-reviews>
- Dunivin, Z., Zadunayski, L., Baskota, U., Siek, K., & Mankoff, J. (2020). Gender, soft skills, and patient experience in online physician reviews: a large-scale text analysis. *Journal of medical Internet research*, 22(7), e14455.
- Dusch, M. N., O'Sullivan, P. S., & Ascher, N. L. (2014). Patient perceptions of female surgeons: how surgeon demeanor and type of surgery affect patient preference. *Journal of Surgical Research*, 187(1), 59-64.
- Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological review*, 109(3), 573.
- Eagly, A. H., Wood, W., & Diekmann, A. B. (2000). Social role theory of sex differences and similarities: A current appraisal. *The developmental social psychology of gender*, 12(174), 9781410605245-9781410605212.
- Emmert, M., Meszmer, N., & Sander, U. (2016). Do health care providers use online patient ratings to improve the quality of care? Results from an online-based cross-sectional study. *Journal of medical Internet research*, 18(9), e5889.
- Emmert, M., Sander, U., & Pisch, F. (2013). Eight questions about physician-rating websites: a systematic review. *Journal of medical Internet research*, 15(2), e2360.
- Epstein, R. M., Fiscella, K., Lesser, C. S., & Stange, K. C. (2010). Why the nation needs a policy push on patient-centered health care. *Health Affairs*, 29(8), 1489-1495.

- Epstein, R. M., & Street, R. L. (2011a). The values and value of patient-centered care. In: *Annals Family Med.*
- Epstein, R. M., & Street, R. L. (2011b). The values and value of patient-centered care. In (Vol. 9, pp. 100-103): *Annals Family Med.*
- Findlay, S. D. (2016). Consumers' interest in provider ratings grows, and improved report cards and other steps could accelerate their use. *Health Affairs, 35(4)*, 688-696.
- Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in cognitive sciences, 11(2)*, 77-83.
- Fiske, S. T., Cuddy, A. J., Glick, P., & Xu, J. (2018). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. In *Social cognition* (pp. 162-214). Routledge.
- Foschi, M. (2000). Double standards for competence: Theory and research. *Annual review of Sociology, 26(1)*, 21-42.
- Fox, S., & Duggan, M. (2013). Pew research center. Health online. 2013. *Aralık, 10(2015)*, 53098246-53098242.
- Fox, S., & Jones, S. (2009). *The Social Life of Health Information*.
<https://www.pewresearch.org/internet/2009/06/11/the-social-life-of-health-information/>
- Frequently Asked Questions About CGCAHPS*. Press Ganey. Retrieved August 30, 2020 from
<https://helpandtraining.pressganey.com/researchResources/governmentInitiatives/CGCAHPS/faqs.aspx#payment>
- Frost, M. (2020). How many stars for physician ratings. *ACP Internist, 40(2)*.
<https://acpinternist.org/archives/2020/02/how-many-stars-for-physician-ratings.htm>

- Fung, C. H., Elliott, M. N., Hays, R. D., Kahn, K. L., Kanouse, D. E., McGlynn, E. A., Spranca, M. D., & Shekelle, P. G. (2005). Patients' preferences for technical versus interpersonal quality when selecting a primary care physician. *Health services research, 40*(4), 957-977.
- Furnas, H. J., Korman, J. M., Canales, F. L., & Pence, L. D. (2020). Patient Reviews: Yelp, Google, Healthgrades, Vitals, and RealSelf. *Plastic and Reconstructive Surgery, 146*(6), 1419-1431.
- Gao, G. G., McCullough, J. S., Agarwal, R., & Jha, A. K. (2012). A changing landscape of physician quality reporting: analysis of patients' online ratings of their physicians over a 5-year period. *Journal of medical Internet research, 14*(1), e2003.
- Grant, S., Guthrie, B., Entwistle, V., & Williams, B. (2014). A meta-ethnography of organisational culture in primary care medical practice. *Journal of health organization and management.*
- Greaves, F., Millett, C., & Nuki, P. (2014). England's experience incorporating "anecdotal" reports from consumers into their national reporting system: lessons for the United States of what to do or not to do? *Medical Care Research and Review, 71*(5_suppl), 65S-80S.
- Guo, J. W., Sisler, S. M., Wang, C. Y., & Wallace, A. S. (2021). Exploring experiences of COVID-19-positive individuals from social media posts. *International journal of nursing practice, 27*(5), e12986.
- Gupta, S., & Jordan, K. (2022). Understanding gender bias toward physicians using online doctor reviews. *Psychology of Language and Communication, 26*(1), 18-41.
- Hall, J. A., Gulbrandsen, P., & Dahl, F. A. (2014). Physician gender, physician patient-centered behavior, and patient satisfaction: a study in three practice settings within a hospital. *Patient education and counseling, 95*(3), 313-318.

- Hall, J. A., & Roter, D. L. (2002). Do patients talk differently to male and female physicians?: A meta-analytic review. *Patient education and counseling*, 48(3), 217-224.
- Hanauer, D. A., Zheng, K., Singer, D. C., Gebremariam, A., & Davis, M. M. (2014). Public awareness, perception, and use of online physician rating sites. *Jama*, 311(7), 734-735.
- Haynes, D., Pampari, A., Topham, C., Schwarzenberger, K., Heath, M., Zou, J., & Greiling, T. M. (2021). Patient Experience Surveys Reveal Gender-Biased Descriptions of Their Care Providers. *Journal of Medical Systems*, 45(10), 1-7.
- Health, W. F. B. *About the Press Ganey Patient Experience Survey*. Retrieved October 10, 2020 from <https://www.wakehealth.edu/About-Us/Quality-and-Awards/Quality-and-Safety-Measures/Patient-Experience/About-the-Press-Ganey-Patient-Experience-Survey>
- Healthgrades. (2022). *Healthgrades Frequently Asked Questions (FAQs)*. Healthgrades. <https://www.healthgrades.com/content/faqs>
- Healthgrades Frequently Asked Questions (FAQs)*. Retrieved September 8, 2022 from <https://www.healthgrades.com/content/faqs>
- Heilman, M. E. (2012). Gender stereotypes and workplace bias. *Research in organizational Behavior*, 32, 113-135.
- Hill, E. J., Bowman, K. A., Stalmeijer, R. E., Solomon, Y., & Dornan, T. (2014). Can I cut it? Medical students' perceptions of surgeons and surgical careers. *The American Journal of Surgery*, 208(5), 860-867.
- Holliday, A. M., Kachalia, A., Meyer, G. S., & Sequist, T. D. (2017). Physician and patient views on public physician rating websites: a cross-sectional study. *Journal of general internal medicine*, 32(6), 626-631.

- Hong, Y. A., Liang, C., Radcliff, T. A., Wigfall, L. T., & Street, R. L. (2019). What do patients say about doctors online? A systematic review of studies on patient online reviews. *Journal of medical Internet research*, 21(4), e12521.
- Howe, L. C., Leibowitz, K. A., & Crum, A. J. (2019). When your doctor “Gets It” and “Gets You”: The critical role of competence and warmth in the patient–provider interaction. *Frontiers in Psychiatry*, 475.
- Jhaveri, Y., Gandhi, T., Naik, T., Nisar, S., & Sonawane, P. (2022). Predicting Doctor Ratings from User Reviews using Deep Learning. 2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC),
- Judd, C. M., James-Hawkins, L., Yzerbyt, V., & Kashima, Y. (2005). Fundamental dimensions of social judgment: understanding the relations between judgments of competence and warmth. *Journal of personality and social psychology*, 89(6), 899.
- Kaba, R., & Sooriakumaran, P. (2007). The evolution of the doctor-patient relationship. *International journal of surgery*, 5(1), 57-65.
- Kadry, B., Chu, L. F., Kadry, B., Gammas, D., & Macario, A. (2011a). Analysis of 4999 online physician ratings indicates that most patients give physicians a favorable rating. *Journal of medical Internet research*, 13(4), e95.
- Kadry, B., Chu, L. F., Kadry, B., Gammas, D., & Macario, A. (2011b). Analysis of 4999 online physician ratings indicates that most patients give physicians a favorable rating. *Journal of medical Internet research*, 13(4), e1960.
- Kilaru, A., Paciotti, B., Ha, Y., Ranard, B., Griffis, H., & Merchant, R. (2014). 377 What Do Patients Say About Emergency Departments in Online Reviews? *Annals of Emergency Medicine*, 64(4), S135.
- Kilaru, A. S., Meisel, Z. F., Paciotti, B., Ha, Y. P., Smith, R. J., Ranard, B. L., & Merchant, R. M. (2016). What do patients say about emergency departments in online reviews? A qualitative study. *BMJ quality & safety*, 25(1), 14-24.

- Ko, D. G., Mai, F., Shan, Z., & Zhang, D. (2019). Operational efficiency and patient-centered health care: A view from online physician reviews. *Journal of Operations Management*, 65(4), 353-379.
- Kordzadeh, N. (2019). Investigating bias in the online physician reviews published on healthcare organizations' websites. *Decision Support Systems*, 118, 70-82.
- Kreitzer, R. J., & Sweet-Cushman, J. (2021). Evaluating student evaluations of teaching: A review of measurement and equity bias in SETs and recommendations for ethical reform. *Journal of Academic Ethics*, 1-12.
- Kullgren, J. T., & Fendrick, A. M. (2021). The price will be right—how to help patients and providers benefit from the new CMS transparency rule. *JAMA Health Forum*,
- Kullgren, J. T., Malani, P. N., Kirch, M., Singer, D., Solway, E., & Hanauer, D. A. (2021). Use of Online Physician Ratings and Reviews by Older US Adults: Results of a National Survey. *Annals of Internal Medicine*, 174(8), 1180-1182.
- Lagu, T., Hannon, N. S., Rothberg, M. B., & Lindenauer, P. K. (2010). Patients' evaluations of health care providers in the era of social networking: an analysis of physician-rating websites. *Journal of general internal medicine*, 25, 942-946.
- Lasek, R. J., Barkley, W., Harper, D. L., & Rosenthal, G. E. (1997). An evaluation of the impact of nonresponse bias on patient satisfaction surveys. *Medical care*, 35(6), 646-652.
- Latimer, T., Roscamp, J., & Papanikitas, A. (2017). Patient-centredness and consumerism in healthcare: an ideological mess. *Journal of the Royal Society of Medicine*, 110(11), 425-427.
- Lee, V. (2017). Transparency and trust—online patient reviews of physicians. *New England Journal of Medicine*, 376(3), 197-199.

- Li, W., Jin, B., & Quan, Y. (2020). Review of research on text sentiment analysis based on deep learning. *Open Access Library Journal*, 7(3), 1-8.
- Liao, L., Chung, S., Altamirano, J., Garcia, L., Fassiotto, M., Maldonado, B., Heidenreich, P., & Palaniappan, L. (2020). The association between Asian patient race/ethnicity and lower satisfaction scores. *BMC health services research*, 20(1), 1-11.
- Liao, W., Zeng, B., Yin, X., & Wei, P. (2021). An improved aspect-category sentiment analysis model for text sentiment analysis based on RoBERTa. *Applied Intelligence*, 51(6), 3522-3533.
- López, A., Detz, A., Ratanawongsa, N., & Sarkar, U. (2012). What patients say about their doctors online: a qualitative content analysis. *Journal of general internal medicine*, 27(6), 685-692.
- Lu, S. F., & Rui, H. (2018). Can we trust online physician ratings? Evidence from cardiac surgeons in Florida. *Management science*, 64(6), 2557-2573.
- Lyness, K. S., & Heilman, M. E. (2006). When fit is fundamental: performance evaluations and promotions of upper-level female and male managers. *Journal of Applied Psychology*, 91(4), 777.
- Lyness, K. S., & Thompson, D. E. (2000). Climbing the corporate ladder: do female and male executives follow the same route? *Journal of Applied Psychology*, 85(1), 86.
- Madera, J. M., Hebl, M. R., & Martin, R. C. (2009). Gender and letters of recommendation for academia: agentic and communal differences. *Journal of Applied Psychology*, 94(6), 1591.
- Marrero, K., King, E., & Fingeret, A. L. (2020). Impact of surgeon gender on online physician reviews. *Journal of Surgical Research*, 245, 510-515.
- Mast, M. S., Hall, J. A., & Roter, D. L. (2007). Disentangling physician sex and physician communication style: their effects on patient satisfaction in a virtual medical visit. *Patient education and counseling*, 68(1), 16-22.

- Mast, M. S., & Kadji, K. K. (2018). How female and male physicians' communication is perceived differently. *Patient education and counseling*, 101(9), 1697-1701.
- McGrath, R. J., Priestley, J. L., Zhou, Y., & Culligan, P. J. (2018). The validity of online patient ratings of physicians: analysis of physician peer reviews and patient ratings. *Interactive Journal of Medical Research*, 7(1), e9350.
- McKinstry, B., & Yang, S. Y. (1994). Do patients care about the age of their general practitioner? A questionnaire survey in five practices. *Br J Gen Pract*, 44(385), 349-351.
- Menendez, M. E., Chen, N. C., Mudgal, C. S., Jupiter, J. B., & Ring, D. (2015). Physician empathy as a driver of hand surgery patient satisfaction. *The Journal of hand surgery*, 40(9), 1860-1865. e1862.
- Mitchell, K. M., & Martin, J. (2018). Gender bias in student evaluations. *PS: Political Science & Politics*, 51(3), 648-652.
- Murarka, A., Radhakrishnan, B., & Ravichandran, S. (2021). Classification of mental illnesses on social media using RoBERTa. Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis,
- Murphy, G. P., Radadia, K. D., & Breyer, B. N. (2019). Online physician reviews: is there a place for them? *Risk management and healthcare policy*, 12, 85.
- Nelson, L. K., Burk, D., Knudsen, M., & McCall, L. (2021). The future of coding: A comparison of hand-coding and three types of computer-assisted text analysis methods. *Sociological Methods & Research*, 50(1), 202-237.
- Okike, K., Uhr, N. R., Shin, S. Y., Xie, K. C., Kim, C. Y., Funahashi, T. T., & Kanter, M. H. (2019). A Comparison of Online Physician Ratings and Internal Patient-Submitted Ratings from a Large Healthcare System. *Journal of general internal medicine*, 34(11), 2575-2579.

- Oliveira, F. B., Haque, A., Mougouei, D., Evans, S., Sichman, J. S., & Singh, M. P. (2022). Investigating the Emotional Response to COVID-19 News on Twitter: A Topic Modeling and Emotion Classification Approach. *IEEE Access*, *10*, 16883-16897.
- Park, S.-Y., Yun, G. W., Friedman, S., Hill, K., & Coppes, M. J. (2022). Patient-Centered Care and Healthcare Consumerism in Online Healthcare Service Advertisements: A Positioning Analysis. *Journal of Patient Experience*, *9*, 23743735221133636.
- Paul, M. J., Wallace, B. C., & Dredze, M. (2013). What affects patient (dis) satisfaction? Analyzing online doctor ratings with a joint topic-sentiment model. AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI.
- Peuchaud, S. R. (2020). Respected as a client, cared for as a patient: Evidence of heuristic decision-making from Yelp reviews of Obstetrician-Gynecologists. *Health Communication*, *35*(7), 842-848.
- Pines, J. M., Penninti, P., Alfaraj, S., Carlson, J. N., Colfer, O., Corbit, C. K., & Venkat, A. (2018). Measurement under the microscope: high variability and limited construct validity in emergency department patient-experience scores. *Annals of Emergency Medicine*, *71*(5), 545-554. e546.
- PressGaney. (2021). *Consumer experience trends in healthcare 2021*. PressGaney. <https://info.pressganey.com/e-books-research/press-ganey-consumer-experience-in-healthcare-trends-report-2021#main-content>
- Ranard, B. L., Werner, R. M., Antanavicius, T., Schwartz, H. A., Smith, R. J., Meisel, Z. F., Asch, D. A., Ungar, L. H., & Merchant, R. M. (2016). Yelp reviews of hospital care can supplement and inform traditional surveys of the patient experience of care. *Health Affairs*, *35*(4), 697-705.
- Richmond, R. (2012, September 10, 2012). Yelp Co-Founder Jeremy Stoppelman on Innovating and Staying Relevant. *Entrepreneur*. <https://www.entrepreneur.com/leadership/yelp-co-founder-jeremy-stoppelman-on-innovating-and-staying/224338>

- Rosenberg, S., Nelson, C., & Vivekananthan, P. (1968). A multidimensional approach to the structure of personality impressions. *Journal of personality and social psychology, 9*(4), 283.
- Rosette, A. S., & Tost, L. P. (2010). Agentic women and communal leadership: How role prescriptions confer advantage to top women leaders. *Journal of Applied Psychology, 95*(2), 221.
- Roter, D. L., & Hall, J. A. (2004). Physician gender and patient-centered communication: a critical review of empirical research. *Annu. Rev. Public Health, 25*, 497-519.
- Rudman, L. A., & Glick, P. (1999). Feminized management and backlash toward agentic women: the hidden costs to women of a kinder, gentler image of middle managers. *Journal of personality and social psychology, 77*(5), 1004.
- Ryan, T., Specht, J., Smith, S., & DelGaudio, J. M. (2016). Does the Press Ganey survey correlate to online health grades for a major academic otolaryngology department? *Otolaryngology--Head and Neck Surgery, 155*(3), 411-415.
- Saifee, D. H., Hudnall, M., & Raja, U. (2022). Physician Gender, Patient Risk, and Web-Based Reviews: Longitudinal Study of the Relationship Between Physicians' Gender and Their Web-Based Reviews. *Journal of medical Internet research, 24*(4), e31659.
- Saliba, H., & Black, N. M. P. (2009). An analysis of healthcare providers' online ratings. *Informatics in primary care, 17*, 249-253.
- Sarsons, H. (2017). Interpreting signals in the labor market: evidence from medical referrals. *Job Market Paper, 141-145*.
- Schlesinger, M., Grob, R., Shaller, D., Martino, S. C., Parker, A. M., Finucane, M. L., Cerully, J. L., & Rybowski, L. (2015). Taking patients' narratives about clinicians from anecdote to science. *N Engl J Med, 373*(7), 675-679.

- Simsekler, M. C. E., Alhashmi, N. H., Azar, E., King, N., Luqman, R. A. M. A., & Al Mulla, A. (2021). Exploring drivers of patient satisfaction using a random forest algorithm. *BMC medical informatics and decision making*, 21(1), 157.
- Sitzia, J. (1999). How valid and reliable are patient satisfaction data? An analysis of 195 studies. *International Journal for Quality in Health Care*, 11(4), 319-328.
- Sommet, N., & Morselli, D. (2017). Keep calm and learn multilevel logistic modeling: A simplified three-step procedure using Stata, R, Mplus, and SPSS. *International Review of Social Psychology*, 30, 203-218.
- Steiner-Hofbauer, V., Schrank, B., & Holzinger, A. (2018). What is a good doctor? *Wiener Medizinische Wochenschrift*, 168(15), 398-405.
- Tak, H., Ruhnke, G. W., & Shih, Y.-C. T. (2015). The association between patient-centered attributes of care and patient satisfaction. *The Patient-Patient-Centered Outcomes Research*, 8(2), 187-197.
- Terlutter, R., Bidmon, S., & Röttl, J. (2014). Who uses physician-rating websites? Differences in sociodemographic variables, psychographic variables, and health status of users and nonusers of physician-rating websites. *Journal of medical Internet research*, 16(3), e97.
- Thawani, A., Paul, M. J., Sarkar, U., & Wallace, B. C. (2019). Are online reviews of physicians biased against female providers? Machine Learning for Healthcare Conference,
- Thompson, M., & Cutler, C. M. (2010). Health care consumerism movement takes a step forward. *Benefits Quarterly*, 26(1).
- Trehan, S. K., DeFrancesco, C. J., Nguyen, J. T., Charalel, R. A., & Daluiski, A. (2016). Online patient ratings of hand surgeons. *The Journal of hand surgery*, 41(1), 98-103.

- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
- Wainwright, D., Harris, M., & Wainwright, E. (2022). Trainee doctors' perceptions of the surgeon stereotype and its impact on professional identification: a qualitative study. *BMC medical education*, 22(1), 1-10.
- Wallace, B. C., Paul, M. J., Sarkar, U., Trikalinos, T. A., & Dredze, M. (2014). A large-scale quantitative analysis of latent factors and sentiment in online doctor reviews. *Journal of the American Medical Informatics Association*, 21(6), 1098-1103.
- Widmer, R. J., Maurer, M. J., Nayar, V. R., Aase, L. A., Wald, J. T., Kotsenas, A. L., Timimi, F. K., Harper, C. M., & Pruthi, S. (2018). Online physician reviews do not reflect patient satisfaction survey responses. *Mayo Clinic Proceedings*,
- Williams, L. E., & Bargh, J. A. (2008). Experiencing physical warmth promotes interpersonal warmth. *Science*, 322(5901), 606-607.
- Wojciszke, B., Abele, A. E., & Baryla, W. (2009). Two dimensions of interpersonal attitudes: Liking depends on communion, respect depends on agency. *European Journal of Social Psychology*, 39(6), 973-990.
- Wojciszke, B., Bazinska, R., & Jaworski, M. (1998). On the dominance of moral categories in impression formation. *Personality and Social Psychology Bulletin*, 24(12), 1251-1263.
- Wolfe, A. (2001). Institute of Medicine report: crossing the quality chasm: a new health care system for the 21st century. *Policy, Politics, & Nursing Practice*, 2(3), 233-235.
- Wong, C. A., Ellsworth, E., Madanay, F., Chandrasekaran, D., Moore, M., Polsky, D., & Ubel, P. A. (2019). The roles of assisters and automated decision support tools in consumers' marketplace choices: room for improvement. *Health Affairs*, 38(3), 473-481.

- Wu, Q. L., & Tang, L. (2022). What satisfies parents of pediatric patients in China: a grounded theory building analysis of online physician reviews. *Health Communication, 37*(10), 1329-1336.
- Xu, Y., Armony, M., & Ghose, A. (2021). The interplay between online reviews and physician demand: An empirical investigation. *Management Science, 67*(12), 7344-7361.
- Yaraghi, N., Wang, W., Gao, G. G., & Agarwal, R. (2018). How online quality ratings influence patients' choice of medical providers: controlled experimental survey study. *Journal of medical Internet research, 20*(3), e99.
- ZocDoc. (2022). *How reviews work on Zocdoc*. ZocDoc, Inc. Retrieved February 6, 2023 from <https://www.zocdoc.com/about/verifiedreviews/>

Biography

Farrah Madanay was born on September 12, 1990, in Honolulu, Hawaii. She received a Bachelor of Arts in Religious Studies and Art History from Rice University in 2013 and a Master of Arts in Modern European Studies from Columbia University in 2016. Her research has been published in journals including *Health Affairs*; *Journal of Health Policy, Politics and Law*; and *Journal of Adolescent Health*. During her PhD studies, she received the Brown-Nagin Graduate Fellowship, the Duke Endowment Fellowship, and the PARISS Fellowship, and was a Margolis Scholar in the Duke Margolis Center for Health Policy.