Routledge
Taylor & Francis Group

Health Communication

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/hhth20

Does the Gender of Doctors Change a Patient's Perception?

Sonam S. Gupta & Kayden N. Jordan

To cite this article: Sonam S. Gupta & Kayden N. Jordan (26 Apr 2024): Does the Gender of Doctors Change a Patient's Perception?, Health Communication, DOI: 10.1080/10410236.2024.2343467

To link to this article: https://doi.org/10.1080/10410236.2024.2343467

HEALTH COMMUNICATION
https://doi.org/10.1080/10410236.2024.2343467

Routledge
Taylor & Francis Group

Does the Gender of Doctors Change a Patient's Perception?

Sonam S. Gupta and Kayden N. Jordan

Data Science Area, Harrisburg University of Science & Technology

ABSTRACT

In the past, researchers have used various data sources like social media, admission applications, or letters of recommendation to identify gender-based differences in linguistic data. One such avenue in healthcare is the online physician reviews website. Such websites, for example, RateMyDoctors.com, ZocDoc, or even Yelp.com have become a go-to place for patients when choosing their physicians. In the current research, we used two different natural language processing (NLP) approaches: semi-supervised and unsupervised topic modeling to analyze the text of the reviews to identify gender-based linguistic differences from patients' perspectives. We found that female physicians receive more reviews on their personable skills and warmth, aligning with the Stereotype Content Model. We also found other popular topics discussing bedside manners and overall patient experiences, where the reviews suggested that patients were happier with their experience with female physicians and perceived them to have more positive traits than their male counterparts. Although our study did not reflect significant linguistic differences; it highlights the importance for patients and doctors to be more aware of potential gender stereotypes and perceptions.

Women have faced bias and inequality in nearly every profession, either historically or currently. In management, The gender pay gap for full-time female managers was more prominent than the general pay gap. According to estimates, full-time female managers earned 77 cents for every dollar earned by their male counterparts, indicating a pay gap of 23 cents per dollar (GAO, 2023, p. 4). Even though approximately 44% of females make up the total workforce, only 41% of managerial or leadership roles are held by women. In STEM, 43% of women tend to leave their jobs once they are pregnant or choose to go part-time instead of staying in their full-time jobs due to women's underrepresentation, and inequality in workplaces (Cech & Blair-Loy, 2019). Additionally, Casad et al. (2021) found that underrepresentation, stereotyping, and lack of social support led women to think of themselves as less capable or unsuitable to be involved in STEM. These factors ultimately contribute to women not choosing STEM majors in the first place. In medicine, Robinson (2003) discussed the stresses on female physicians, such as not being able to find a mentor at their workplace, facing discrimination, and experiencing slower promotions in academia, which leads to fewer female physicians, especially in male-dominated specialties like cardiovascular and general surgery.

Research into gendered perceptions of physicians has shown that interpersonal skills are more associated with female physicians (Burgoon et al., 1991; Hall et al., 2014; Nicolai & Demmel, 2007). Some research has also suggested that technical skills are more associated with male physicians (Himmelstein & Sanchez, 2016). However, that research did not investigate whether those gendered perceptions impact patient choices. Using experimental methods, Li et al. (2019) showed physician profiles to the participants, including patient reviews that varied in the physician's gender and the emphasized skills (interpersonal or technical). Participants indicated greater willingness to choose a female physician (over the male physician) only when technical skills were emphasized suggesting that interpersonal skills were assumed for female physicians, but the opposite was not true for male physicians. While together these studies paint a robust picture of gendered perceptions of physicians, it remains to be explored how these perceptions play out in everyday life as patients spontaneously review their physicians. Given the abundant text data on online reviews for doctors, language analysis using Natural Language Processing (NLP) is another approach that can help identify how common references to interpersonal versus technical skills are in physician reviews as well as if those references are related to a physician's gender. Using topic modeling we analyzed the language of online reviews from ZocDoc and RateMyDoctors websites using both a top-down approach as well as a bottom-up approach to identify potential types of gendered perceptions. Unlike the other study, we included all specialties from the RateMyDoctors website, positive and negative reviews from both datasets, to keep it more generalized and did not involve any human subjects.

Literature review

From the top-down perspective, we build upon the recently proposed Spontaneous Stereotype Content Model (Nicolas et al., 2022) which is based on Fiske's Stereotype Content Model (Fiske, 2018). Decades of research have shown that stereotypes of groups are common and often automatic

CONTACT Kayden N. Jordan kjordan@harrisburgu.edu Data Science Area, Harrisburg University of Science & Technology, 326 Market St, Harrisburg, PA 17101
Supplemental data for this article can be accessed online at https://doi.org/10.1080/10410236.2024.2343467.

(Casad & Bryant, 2016; Pennington et al., 2016; Yang et al., 2020). The Stereotype Content Model proposed two main dimensions for these stereotypes: warmth and competence. Some groups are perceived as warm but not competent (e.g., elderly people) or competent but not warm (e.g., rich people) or both warm and competent (e.g., middle class) individuals, or neither warm nor competent (e.g., homeless people; Fiske, 2012). However, the weakness of such an approach is its reductionism limiting the content of all stereotypes to two dimensions. To address this limitation, Nicolas et al. (2022) proposed the Spontaneous Stereotype Content Model, which takes the stereotype dimensions from open-ended, free responses.

In the real world, people rarely express their stereotypes directly. Instead, their stereotypes influence their perceptions, judgments, and interactions. Therefore, it is crucial to be able to measure stereotype content in the natural language beyond the experimental context. Measuring stereotypes in natural settings can help identify biases and unfair treatment, which are often caused by stereotypes. Researchers in the past have worked on either building a framework or finding a quantifiable measure to identify such gender stereotypes and biases either that would apply in natural settings or with natural language from settings such as social media (Garcia et al., 2018; Pair et al., 2021; Tran et al., 2019). Here, we build a system where we analyze the language of reviews to identify words that could potentially reflect gender-based differences and quantify the results by calculating the content of reviews using two diverse methods.

Bridging to communication research, decades of research has delved into the numerous aspects of physician-patient communication, and over that time, the context and dynamics of that communication has changed over time. Most relevant to the current work is changes to technology which have greatly improved the access and sharing of information (Timmermans, 2020). Medical information sites, like WebMD, and physician review sites, like ZocDoc, have allowed patients access to information not only on health matters but their doctors as well. These technological changes have shifted the dynamics allowing patients to be more collaborative and engaged in communication with their physician (Bernardi & Wu, 2022; Timmermans, 2020). However, it is important to continue to study how these technological advances impact physician-patient communication interaction. The impact of gender stereotypes and gender-based communication has been widely studied in the context of physician-patient communication broadly. Research from multiple theoretical perspectives such as expectancy theory (Burgoon et al., 1991) and communication accommodation theory (Watson & Gallois, 1998) has shown how communication style of physicians impacts patient perceptions and behavior. Most important to the current study, Li et al. (2019) demonstrated the effect of gender stereotypes in online physician reviews on patient perceptions and choice of physician. While that work shows the impact that gender stereotypes in physician reviews have on patients, what is still unknown is how common those stereotypes are in online physician review sites. In the same vein as Wang et al. (2023), we use qualitative data to expand on knowledge of physician-patient communication by examining the content of the communication. However, instead of using qualitative methods, we make use of linguistic data, and online physicians reviews, with quantitative methods to measure gender bias.

In the current research, we identified that linguistic data could be a rich source for analyzing and quantifying gender bias. Some examples from past research include studies where authors used text data from publication abstracts, social media platforms like Reddit, letters of recommendation in medicine or academia, and even college applications. Marjanovic et al. (2022) analyzed Reddit comments to study gender bias against women in leadership roles, such as in politics. They based their analyzes on linguistic aspects (distribution of comments for politicians based on gender, as well as the length of the comments) and non-linguistic aspects (attention given to the politicians and connections among them in the comments). Their analysis led them to conclude that female politicians were commonly referred to using their first names and that their physical appearance was often mentioned. Filippou et al. (2019) analyzed letters of recommendation using a psycholinguistic text analysis tool called Linguistic Inquiry and Word Count (LIWC) to find that letters written for male urology residency applicants contained more references to power, personal drive, and work than letters written for female applicants, which could have contributed to the greater acceptance of male applicants into the department. Rao and Taboada (2021) used topic modeling to investigate gender-based incongruity in news articles for each topic. They found that topics such as sports or business were associated with men, whereas caregiving or entertainment was associated with women.

In this study, we explored and analyzed potential linguistic identifiers by basing our hypotheses and predictions on quantifiable measures of gender bias, especially from various text data available. From previous work (Gupta & Jordan, 2022) on the analysis of online doctors' reviews data using LIWC, we found subtle linguistic gender-based differences such as gendered references were slightly more for female physicians as well as they were reviewed more for their interpersonal skills than technical abilities. To extend past work, we focused our analyzes on online physician reviews to identify linguistic gender differences using two different approaches. The top-down approach is built upon findings from past literature, such as female physicians being reviewed more for their personable communication skills, warmth, and physical appearance. The bottom-up approach explored more topics from the data itself that reflect gender-based differences.

An extension of the original research involves using topic models instead of dictionaries or word embeddings. More specifically, we employ a seeded or semi-supervised topic model to enable a more comprehensive measurement of the original stereotype taxonomies. Semi-supervised topic models allow for specific topics to be captured in line with theoretical constructs like cyberbullying (Zhang & Ramesh, 2018) or with expected clusters of documents (Shanthakumar et al., 2020). Structural topic models (STM), on the other hand, are a type of unsupervised topic model, which unlike traditional unsupervised topic models like LDA allow for topics to be constructed and analyzed alongside numeric or categorical variables. For

example, STM has been used to how the content of reviews relates to ratings (Hu et al., 2019) and the topics economists publish on by the gender of the authors (Conde Ruiz et al., 2022).

Based on past studies, we learned that females in medicine, leadership, or executive roles are often rated and reviewed more for their interpersonal skills, warmth, and physical appearance. Although many text datasets have been used, the online platform for doctor reviews is not commonly used to identify and measure gender bias. In this study, we aimed to explore and analyze online doctors' reviews data further to answer the following questions, based on past literature and the gaps we identified. We took two approaches and the hypotheses for top-down approach are as follows:

Hypothesis 1: Reviews for female physicians will contain more references to their warmth and personable skills.

Hypothesis 2: Reviews for male physicians will focus more on their technical competence.

Hypothesis 3: Reviews for female physicians will contain more gendered references.

Hypothesis 4: Reviews for female physicians will consist of more descriptors of physical appearance.

For the bottom-up approach, the research questions we wanted to answer are:

(1) Do the topics from STM have any connection to theoretical predictions?
(2) What topics tend to be associated with female versus male physicians and vice versa?

Methods

Topic modeling techniques including seeded Latent Dirichlet Allocation (LDA) as a top-down method and Structural Topic Modelling (STM) as a bottom-up method were implemented to explore the words from the online doctors' reviews datasets. The top-down method was used to help find evidence for our hypotheses whereas bottom-up was used to answer our research questions in a more exploratory manner. Seeded LDA was chosen as the top-down approach as it allowed us to specify which topics to model and to test theoretically driven hypotheses. STM was used as the bottom-up approach as it derives topics only from the data allowing us to explore topics not previously found or theorized. We focused on calculating standardized mean differences, standard deviations, and confidence intervals (CI) using Cohen's d from HLM results to interpret the statistical significance. For the bottom-up method, we plotted the effect of the covariate, the gender of the physicians, to help understand any gender-based differences from the words and topics that were returned from STM.

Datasets

Online patient reviews from RateMDs.com and ZocDoc.com were used in this research. The datasets for both RateMDs.com and ZocDoc.com were readily available in a GitHub repository by Shikhar (2020) and Sonawane (2017) respectively. The main analysis was performed on RateMDs data whereas ZocDoc dataset was used for testing the robustness of the findings. Though both datasets are publicly available from open websites with limited identifiable information, IRB approval was sought and granted for the study.

These data consisted of doctor's names, their gender, their specialty, ratings of their staff, punctuality, helpfulness, and knowledge (on a scale of 1–5) along with free text reviews of the doctors. After removing duplicates, there were 46,554 reviews representing 2647 unique female doctors and 3552 unique male doctors from all the specialty. Across doctors, numerical ratings were high: staff ($M = 4.62$, $SD = 0.97$, $Mdn = 5$), punctuality ($M = 4.59$, $SD = 0.98$, $Mdn = 5$), helpfulness ($M = 4.60$, $SD = 1.08$, $Mdn = 5$) and knowledge ($M = 4.65$, $SD = 0.99$, $Mdn = 5$). All reviews were in English, and any non-English words were removed.

The dataset consisted of doctor's names, their location, verified patient names, an overall rating (scale of 1–5), ratings of bedside manner and wait time, doctor's degree, review date, and free text reviews of the doctors. The total number of non-duplicated reviews were 19,372 reviews for 555 unique doctors (214 female and 341 male) from 2008 to 2015 with only primary care doctors from Manhattan, NY. The dimensions of numerical ratings included overall ($M = 4.66$, $SD = 0.89$, $Mdn = 5$), bedside manner of doctor ($M = 4.7$, $SD = 0.77$, $Mdn = 5$) and wait time ($M = 4.34$, $SD = 0.83$, $Mdn = 5$) for the patients. Again, only English language reviews were used.

ZocDoc data did not contain the gender of the doctors. Feminine and masculine pronouns found in the text reviews were used as a basis to label the gender of the doctors. For example, if masculine pronouns like "he," and "him" were found in the review, then the label given was male, and if feminine pronouns like "she" and "her" were found in the review then the physician was labeled female. While not a perfect method for assigning gender, it does capture what the patient perceived the physician's gender to be which is likely most important in this context. When no such pronouns were found, the first author assigned them manually based on the most common gender associated with the first name and photo of the physician if available; this occurred in 99 cases. As further validation, both authors independently coded gender for 99 doctor names and agreed on 96 of those gender labels. The authors' coding agreed with the gender labels based on pronouns for 99 doctors. Given the sample size, this was deemed an acceptable level of error.

For both the datasets, the numerical ratings were very skewed, text analysis of the reviews itself was more useful in understanding patients' attitudes (Gupta & Jordan, 2022). While we do consider low versus high-rated physicians in the top-down approach, the results for the low-rated physicians should be considered very preliminary. For the bottom-up approach, ratings were removed from consideration as the model would not converge with the inclusion of ratings.

Statistical analysis

The analysis involved two parts. At first, two supervised topic modeling techniques, seeded LDA (Latent Dirichlet Allocation) as the top-down approach and Structural Topic Model (STM) as the bottom-up approach were used. Using the topic probabilities from these two topic modeling techniques, bootstrap hierarchical linear models were built to understand the mean differences in topics across reviews written about male versus female physicians.

Based on the past work and theory described in the related work section and detailed in our hypotheses, a seeded LDA topic model was used (Lu et al., 2011; Watanabe & Zhou, 2022). Four separate seed word lists were built for each theoretical concept by the authors. The words in these lists differed for each of the datasets, so the words are more suitable to the respective data. The initial word list was built based on past literature (Chen et al., 2020) that discussed commonly used words to describe physicians. To generate a more comprehensive list of seed words, the Word2Vec model (Mikolov et al., 2013) was used to build word embedding on the entire dataset to find similar words. For each of the hypotheses, the example word lists were H1 warmth ("comfortable," "considerate," "interpersonal," etc.), H2 competence ("superior," "impressive," "competent" etc.), H3 gendered references ("lady," "woman," "guy," etc.), H4 physical appearance descriptors ("attractive," "young," "beautiful," etc.). Table A in Appendix shows the full 4 lists of seed words per dataset to test each of the hypotheses. Since there were four hypotheses to be tested, we chose to have four topics and associated them with the four seed word lists and thus there were no other topics that were generated in this approach. The model was built in R using the seededlda package (Watanabe & Xuan-Hieu, 2022). The means, SDs, and Cohen's d values from the HLM analysis performed using topic probabilities from seeded LDA for both low-rated and high-rated doctors' reviews were calculated separately. The dependent variables in this analysis are the seeded topic labels provided when building topic models whereas the independent variable is the gender of the physician.

STM was another topic model technique implemented in this study. Rather than rely on a seed word list, this technique allows for metadata (in this case the gender of the physician reviewed) to be accounted for in the model. The STM package (Roberts et al., 2019) was used for this topic model. The value of K (number of topics) was chosen to be 10 after running several trials of the STM on the datasets. The list of words grouped in each of the topics usually presents a theme. However, after a certain number of topics, the words may start to repeat, or no theme is detectable. Therefore, we chose K to be 10 in this approach. To evaluate the number of topics and the detectable themes within them, human interpretation is ideal. Therefore, the first author labeled the topics generated by this technique were given based on the words per topic as well as the top reviews.

Results

Hypothesis 1

In line with the Stereotype Content Model (SCM), we predicted that reviews of female physicians would contain more references to warmth. To explore whether this holds true across general ratings of the physician, we examined the differences in low- and high-rated doctors separately. Building on the seed word list (e.g., comfortable, considerate, etc.), the final topic distribution also included words like "care," "disappointed," "patient." The specificity of the final topic words suggests that the model is pinpointing the importance of context where the words relate to the experience of patients with their physicians. These subtle differences suggested that the reviews for female physicians talked slightly more about how personable they are than the male physicians. Somewhat supporting the hypothesis, female physician reviews did contain more warmth though the difference was small (ZocDoc: d = 0.02, 95% CI [−.01, .05]; RateMDs: d = 0.05, 95% CI [.03, .07]) and only for high rated physicians. For low-rated doctors, the effects were inconsistent and not statistically significant (ZocDoc: d = −0.06, 95% CI [−.23, .09], RateMDs: d = 0.03, 95% CI [−.02, .09]) likely due to the sample size.

Hypothesis 2

Following the SCM, we hypothesized that the reviews for male physicians would contain more references to their technical abilities and competence as compared to that of female physicians. We followed a similar pattern of analysis of differences for low- and high-rated doctors as in the first hypothesis. The seed word list for this hypothesis consisted of words like "professional," "skills," "competent," etc., and the final topic distribution as seen in Tables A1 to A4 in Appendix included words such as "information," "extremely," "found." The final topic does seem to capture competence but may also be capturing communication skills which would be outside of the hypothesized construct. The topic distribution supported the hypothesis, but the differences were small (ZocDoc: d = 0.02, 95% CI [−.14, .18]; RateMDs: d = 0.09, 95% CI [.02, .15]) and only for low-rated doctors potentially due to this topic capturing poor communication skills. For high-rated doctors, the effect sizes were inconsistent and not statistically significant (ZocDoc: d = −0.01, 95% CI [−.04, .01], RateMDs: d = −0.13, 95% CI [−.15, −.11]).

Hypothesis 3

Gendered references are where the gender of any individual is explicitly mentioned. Using this concept, we predicted that the reviews for female physicians will have more gendered references than male physicians (Dahlen, 2021). The seed word list consisted of words like "lady," "woman," "man," "guy" etc., and the extended topic distribution included "horrible," "condescending," "results." From the topic distribution, we see that the negative connotations were more associated with gendered

Table 1. Means, SDs, and Cohen's d, the lower and upper level of d, from Hierarchical Linear Modeling (HLM) analysis using SeededLDA topic probabilities for low-rated and high-rated doctors' reviews for top-down method.

Hypothesis	Dependent Variable	M (female)	SD (female)	M (male)	SD (male)	*d*	LL *d*	UL *d*
Low-Rated								
H1	Warmth	0.14	0.21	0.16	0.23	−0.06	−0.23	0.09
H2	Competence	0.39	0.39	0.38	0.37	0.02	−0.14	0.18
H3	Gendered	0.13	0.20	0.16	0.24	−0.14	−0.31	0.01
H4	Appearance	0.16	0.23	0.14	0.22	0.107	−0.05	0.27
High-Rated								
H1	Warmth	0.23	0.40	0.222	0.39	0.02	−0.005	0.05
H2	Competence	0.19	0.36	0.20	0.38	−0.01	−0.04	0.01
H3	Gendered	0.20	0.39	0.21	0.38	−0.02	−0.05	0.006
H4	Appearance	0.18	0.38	0.18	0.40	−0.0002	−0.02	0.02

d = Cohen's *d* to measure effect size; LL *d* = lower level or range of *d*; UL *d* = upper level or range of *d*.

references. However, the effect sizes were inconsistent with our hypothesis, and not statistically significant for both low (ZocDoc: $d = -0.14$, 95% *CI* [−.31, .01], RateMDs: $d = -0.05$, 95% *CI* [−.11, .01]) and high-rated (ZocDoc: $d = -0.02$, 95% *CI* [−.05, .006], RateMDs: $d = 0.01$, 95% *CI* [−.01, .03]) reviews for physicians.

Hypothesis 4

We hypothesized that the language of reviews for female physicians would contain more descriptors of physical appearance (Tran-Harding, 2021). Thus, the seed word list consisted of words like "beautiful," "pretty," "handsome," and the topic distribution extended to contain words such as "disappointed," "good," and "experience." The full topic distribution suggests we were not fully able to construct perhaps due to low frequencies of these types of words with patients simply not mentioning their physician's appearance often. The differences were not strong enough to align with our hypothesis (see Table 1 and Table 2). The effect sizes for both low (ZocDoc: $d = 0.107$, 95% *CI* [−.05, .27], RateMDs: $d = -0.06$, 95% *CI* [−.13, −.004]) and high-rated (ZocDoc: $d = -0.0002$, 95% *CI* [−.02, .02], RateMDs: $d = -0.006$, 95% *CI* [−.02, .01]) reviews for doctors were not consistent or statistically significant.

Results

Bottom-up method

From the above results, theoretical differences in patient reviews seem unclear. If we want to fully understand gender differences and biases in patient reviews, a bottom-up approach to discover the actual trends/topics in the data is needed. Hence, here we present an unsupervised topic model, Structural Topic Model (STM), to measure the topics in patient reviews and potential differences in reviews of male and female physicians without imposing preexisting expectations or theories.

STM on ZocDoc data

Figures 1 and B1–B10 to Figure B10 in Appendix are words per topic distribution and top 3 reviews per topic, respectively, which were used to label the topics. Figure 2 shows the differences in the topic distributions of reviews written about male versus female physicians with physician gender being entered in the model as a covariate. Topic 1 captures patients' experience with scheduling appointments and getting their physical or blood tests done in the hospitals, for example, one of the reviews said, "Nope, you can't have a physical after 4 there are no nurses here" and the differences suggest the reviews were more for male physicians. Topics 2, 3, and 4 covered various aspects of positive hospital and overall patient experience. Topic 2, "Positive Hospital Exp." had reviews like "Overall best medical office I have ever encountered" and words like "medic," "profession" suggesting a positive experience for patients with doctors and their clinics, reflecting slightly more reviews for male physicians. Topic 3, "Positive Overall Exp." included reviews like "best dr. visit I've had a long time," "They made me feel I should and could ask any questions that I might have had…" and words "comfort," and "inform," "quick" expressed patients' positive overall experience during their doctor visit. The reviews, just like in topic 2, were more for male physicians than female. Topic 4, "Friendly Office Exp." incorporated reviews "I HIGHLY recommend

Table 2. Means, SDs and Cohen's d, lower and upper level of d, from Hierarchical Linear Modeling (HLM) Analysis using SeededLDA topic probabilities for low-rated and high-rated doctors' reviews for RMD data.

Hypothesis	Dependent Variable	M (female)	SD (female)	M (male)	SD (male)	*d*	LL *d*	UL *d*
Low-rated								
H1	Warmth	0.24	0.32	0.23	0.30	0.03	−0.02	0.09
H2	Competence	0.26	0.33	0.23	0.30	0.09	0.02	0.15
H3	Gendered	0.15	0.25	0.17	0.25	−0.05	−0.11	0.01
H4	Appearance	0.16	0.28	0.18	0.30	−0.06	−0.13	−0.004
High-rated								
H1	Warmth	0.27	0.41	0.25	0.40	0.05	0.03	0.07
H2	Competence	0.13	0.30	0.19	0.42	−0.13	−0.15	−0.11
H3	Gendered	0.19	0.34	0.19	0.34	0.01	−0.007	0.03
H4	Appearance	0.17	0.40	0.18	0.43	−0.006	−0.02	0.01

d = Cohen's *d* to measure effect size; LL *d* = lower level or range of *d*; UL *d* = upper level or range of *d*.

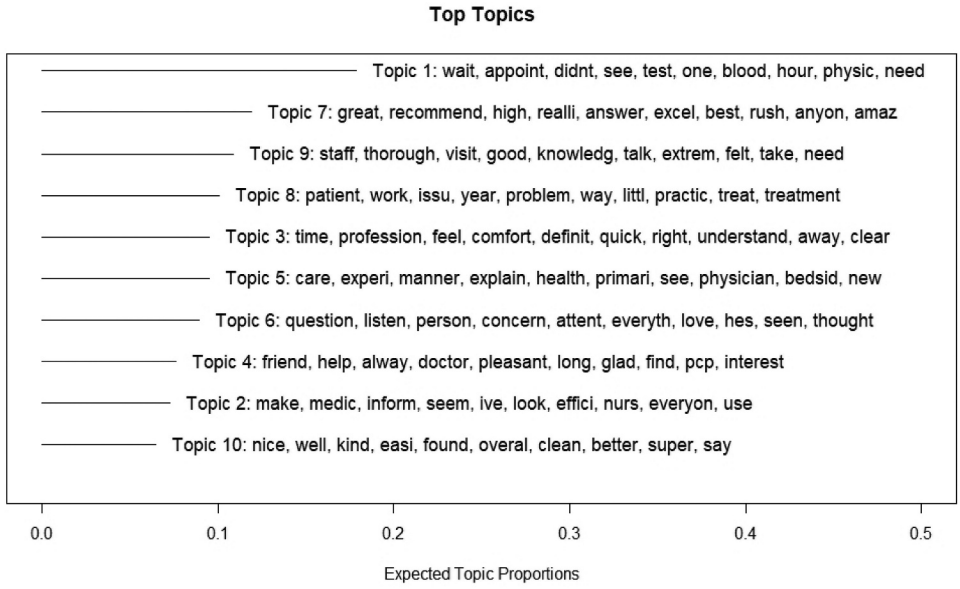

Figure 1. Words per top 10 topics with their expected topic proportions for ZocDoc doctors.

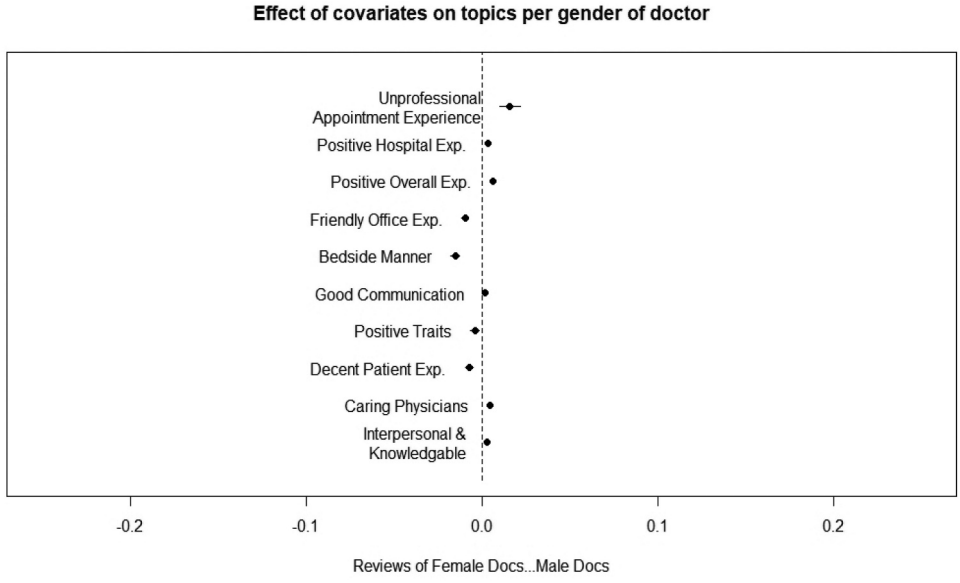

Figure 2. Effect of covariates on topics per gender of the doctors from ZocDoc.

Dr. X and his team," and words such as "help," "pleasant," "interest" that explained patients were satisfied with their experience with the doctor as well as the staff and the reviews in this topic were more for female physicians than male.

Topic 5 is "Bedside Manner" where reviews like "her bedside manner is outstanding…" or "facilitates a great eat doctor–patient relationship" were classified as telling us about the interaction between doctors and patients that talked more about bedside manner in their review more likely to be reviewing female physicians as seen in Figure 2. Topics 6, 7, 9, and 10 capture various aspects of warmth. Topic 6 is "Good Communication"; reviews classified in this topic say the physician "really listened," "was easy to talk to," and "answer[ed] all my questions with clear answers," and from the differences, we saw the reviews were moderately more for male physicians. Topic 7 is "Positive Traits" with patients describing physicians as "caring," "honest," "understanding," and "sweet." Even though the difference was not big, the reviews were more for female physicians. Topic 9, "Caring Physicians" covered how doctors made their patients comfortable during their visit including reviews like "Very easy to talk to a great listener…" whereas Topic 10 is "Interpersonal & Knowledgeable" which covers the professional side of the physicians as well as their interpersonal skills. For both topics 9 and 10, the reviews were slightly more for male physicians but weren't off by a large difference compared to that for female physicians. Topic 8 is "Decent Patient Experience" including reviews that mentioned "The tools to cure many chronic conditions are there in Western medicine, but it takes a genius doctor to unlock their potentials…" or "altogether a very satisfactory treatment experience" suggests patients having a good experience. The patients who had decent patient experience tended to review it more for female physicians.

STM on RateMyDoctors

To determine the robustness of the topics across different review websites, a separate STM model was fitted with the RateMyDoctors data. Like the analysis done on ZocDoc data, Figures 3, B1, and B2 in Appendix are words per topic distribution and top 3 reviews per topic, respectively, collectively which used to label the topics. We analyzed the effect of gender as the covariate from RMD data as seen in Figure 4. Topics 1 and 4 covered various aspects of how the doctors relayed or interpreted the results from treatments on their patients. For example, topic 1, "detailed results," consisted of reviews like "identified my problem areas and performed the treatment. . ." or "she listened to my concerns and was open and honest about what results to expect. . .." In Topic 4, "Healthy Prescriptions," the reviews were about the doctors who considered the results from the treatment and prescribed the potential next steps toward improvement of the patient's health. It consisted of reviews like "he has helped my health turn around. . .," and words such as "life," "medic," "save," "change," and others. Both topics included reviews that were given for male physicians.

Topics 2 and 7 talked about different views on patients' overall experiences during their doctor visits. Topic 2, "Positive Overall Experience" included words such as "feel," "love," and "comfort" and reviews such as "don't make you feel overwhelmed with the procedure and make sure you understand everything" that reflects on patients having a good experience during their treatment. This topic consisted of reviews that were more likely to be about male physicians. Topic 7, "Decent Patient Exp.," on

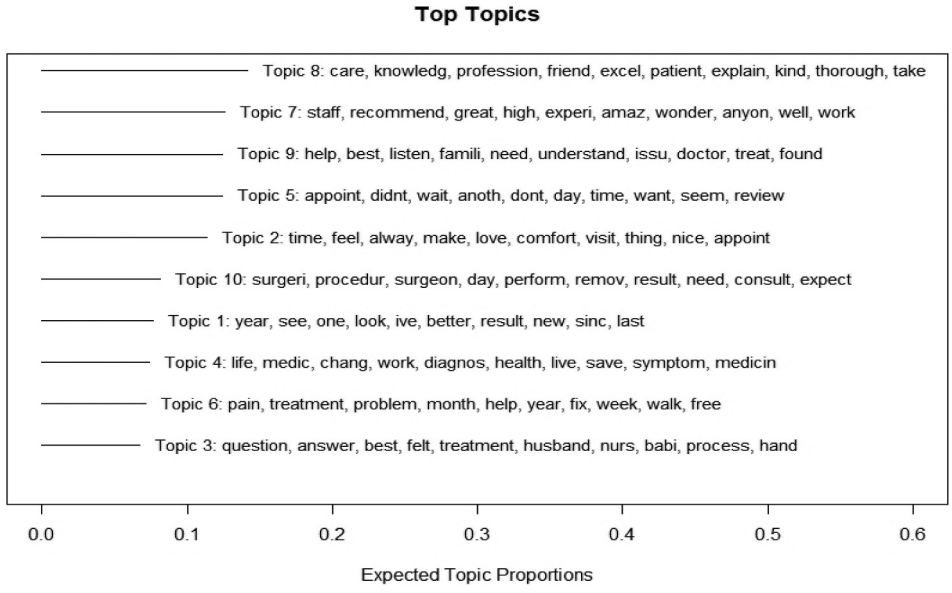

Figure 3. Words per top 10 topics with their expected topic proportions for RMD doctors.

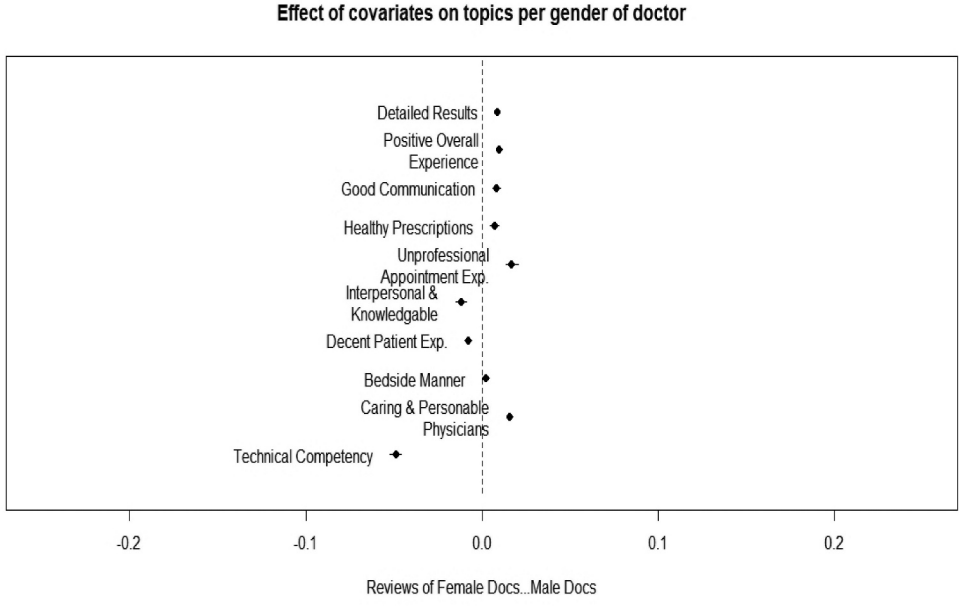

Figure 4. Effect of covariates on topics per gender of the RMD doctors.

the other hand, discussed patients' experience with their physician as well as the rest of the team as seen in reviews like, "The whole experience was very good" or "I had a fantastic experience." Unlike topic 2, the reviews in topic 7 tended to be about female physicians. Topic 3 grouped reviews like "She presented the information objectively – giving no false hope or false promises," or "He also answered our questions, no matter how simple or complicated they were" suggesting "Good Communication." In terms of differences in the covariate plot, topic 3 reviews were more likely to be about male physicians.

Topic 5 "Unprofessional Appointment Exp." heavily discussed reviews about the doctors' staff and their services. For example, the words in topic 5 included "didnt," "wait," etc., whereas the reviews were "I was told by the very nasty and condescending business manager that you'll pay this amount, or it will go to collection." Patients' disappointment with this group of reviews was more likely to be about male physicians as seen in Figure 4. Topics 6 and 10 covered reviews talking about the interpersonal skills and technical knowledge of doctors. For instance, in topic 6, "Interpersonal & Knowledgeable," the reviews mentioned "Dr. X suggested that my sciatic issue was a side effect from my neck injury. . .I finally have some relief from all my problems." and words such as "help," "fix" etc. Topic 10, "Technical Competence" included reviews like "I cannot say enough about her professionalism and knowledge. . ." or "The procedure went very smoothly, and recovery was faster than expected." The reviews in topics 6 and 10 were significantly more likely to be about female physicians. Topics 8, "Bedside Manner" and 9, "Caring & Personable Physicians" mentioned the interpersonal side of the doctors through reviews such as "She takes the time to listen to your concerns and genuinely cares for your well-being," "He is a very compassionate and insightful therapist." The reviews on both these topics were more likely about male physicians.

We identified some similarities in the topics using both the top-down and bottom-up approaches. Reviews that covered the interpersonal and warmth skills of physicians aligned with the SCM. The findings from both methods indicate that patients value interpersonal and warmth skills in their physicians. Reviews mentioned how these skills helped patients feel more welcome and safer. For instance, reviews like "Kind up to date & his upbeat spirit & charm make everything much better. I always leave this office knowing more & feeling better.," and "She is very attentive a great listener and detailed with questions about my health history and was sure to answer all of my questions with clear answers"; words like "nice," "kind," and "super."

In the bottom-up approach, common topics were identified across both datasets, including "Positive Overall Experience," "Decent Patient Exp.," and "Good Communication." This approach also uncovered additional topics that were not present in the top-down approach, such as bedside manner, positive traits, technical knowledge, and overall patient experience. Reviews included specific mentions of physicians' traits and patients' experiences during interactions with doctors and hospital staff, with words like "staff," "recommend," "answer," and "best." Examples of reviews included, "His approach is holistic, different, and, most importantly, caring," "So sweet and straight to the point. I would highly recommend!" and "caring and very understanding. . .I love his staff also."

Discussion

The contribution of this work is threefold. First, we extend upon the long-standing work seeking to understand how gender stereotypes impact patient perceptions of physicians. Bringing the work into a more naturalistic setting, we find that, while the data was obviously noisier, reviews of female physicians did contain somewhat more references to interpersonal skills, in line with past research (Roter et al., 2002). From the top-down topic model, the results showed that generally female physicians were perceived as warmer and more personable in high-rated doctor reviews and male physicians were reviewed more for their technical competence, but these differences were small and somewhat inconsistent across datasets. Additionally, interpersonal skills were more commonly mentioned in reviews with most of the topics related to interpersonal skills and only one or two related to technical competency. In the bottom-up topic approach, the results suggest that patients perceive female physicians to be more interpersonal and tend to have a much better experience with them and their staff.

The second contribution of the current work is methodological. We show how theoretical and atheoretical approaches can be used in conjunction with each other to develop a more holistic picture of a dataset and to investigate theoretical models, like the Stereotype Content Model, in naturalistic settings using unstructured data. In this case, both approaches presented some advantages; for instance, the top-down approach allowed us to generate guided topics, and we found some linguistic differences that aligned with the literature review through the grouping of words into four categories. Similarly, using the bottom-up approach, we learned about several different topics and observed various categories in which patients write reviews for their physicians.

Finally, the work when combined with the work of Li et al. (2019) and others, the results can contribute to a more comprehensive theoretical model of the impact and prevalence of gendered perceptions in physician-patient communication. While Li et al. (2019) found that not only is a patient's choice of physician from reviews impacted by perceptions of interpersonal skills but that those perceptions interact with the physician's gender. Here, we show that differences between how male and female physicians are reviewed are relatively small, potentially minimizing, though not eliminating, the impact physician gender has on patient choices.

While naturalistic data like online reviews sites open a new way to study social processes like bias and stereotypes, they are not without their limitations. As mentioned above, naturalistic data, especially unstructured data like text, tend to be noisy making it more difficult to uncover patterns and differences. Hence, the effects in the current study are small. Selection bias is also a limitation both in terms of what data is available and the types of people which contribute to it. For example, the

ZocDoc data only consisted of reviews from the NYC area. Additionally, patients who wrote reviews in the datasets we used tended to be positive in their reviews suggesting that patients who have bad to average experiences and perceptions may not be contributing to these online review sites.

For future research, linguistic data such as the ones used in this study, could prove to be useful as an initial point for analysis of gender stereotypes and bias. As an extension of the current study, the proportion of reviews could be more balanced between male and female physicians across all specialties. Another potential future direction could involve analyzing other forms of doctor–patient interaction data to enhance patient outcomes and treatment courses. As seen in the literature review, enabling doctors and patients to be more cognizant of implicit and explicit biases is of utmost importance.

Implications

The current work supports three major practical applications for the area. First are changes that could be made to how reviews are written. Currently, most open-ended text reviews are elicited broadly with patients just asked to write about their experience with a provider. Based on the current findings, more specific questions could be developed to elicit feedback on specific aspects of their experience such as "What comments do you have on your physician's bedside manner?," "What comments do you have about your physician's technical skills?," "What comments do you have about your physician's office space?," etc. This more structured approach to reviews would prompt patients to consider the totality of their experience rather than the parts that readily come to mind potentially due to bias.

Secondly, the work has potential applications to physician education. Knowing the content of patient reviews could help physicians better understand and center patient concerns and priorities. Finally, there are applications to better understanding bias more generally in online forums. Identifying and flagging biased content continues to be an issue on social media and other online forums, and the methods and findings in the current work could be used to develop new or better algorithms for bias detection online.

Conclusion

From this study and past research, there are several implications of gender bias on the overall patient experience. Patients who are oblivious to gender stereotypes or have preconceived notions about their physicians could pose an obstacle to achieving a positive patient outcome. For example, if patients believe that female physicians are less technically capable, they may avoid being treated by a female doctor, which could result in a negative treatment outcome, as well as preventing physicians from receiving full credit for their work (Blanch-Hartigan et al., 2010). It's not just the treatment outcome; if patients exhibit bias toward their physicians, it could lead to physician burnout, negatively impacting their mental health and job satisfaction (Sheridan, 2021). Additionally, if physicians themselves hold gender stereotypes about their patients, it could result in inappropriate or ineffective treatment courses. Physicians having implicit biases toward their patients could affect patient-doctor interactions (Fitzgerald & Hurst, 2017). Understanding that there are pain differences between men and women is important, but if physicians don't consider these differences due to such implicit biases, the course of treatment or pain management may not be effective (Samulowitz et al., 2018). Therefore, it is essential for healthcare providers to be aware of potential gender stereotypes and biases in their interactions with patients and strive to provide equal and unbiased care to all patients.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Data availability statement

Citations for data availability are included in the manuscript. Code can be found at https://github.com/sonamgupta1105/patient_Reviews_analysis. Studies involved existing, publicly available data and were approved by the IRB at Harrisburg University of Science and Technology.

References

Bernardi, R., & Wu, P. F. (2022). Online health communities and the patient-doctor relationship: An institutional logics perspective. *Social Science & Medicine*, *314*, 1–9. https://doi.org/10.1016/j.socscimed.2022.115494

Blanch-Hartigan, D., Hall, J. A., Roter, D. L., & Frankel, R. M. (2010). Gender bias in patients' perceptions of patient-centered behaviors. *Patient Education and Counseling*, *80*(3), 315–320. https://doi.org/10.1016/j.pec.2010.06.014

Burgoon, M., Birk, T. S., & Hall, J. R. (1991). Compliance and satisfaction with physician-patient communication: An expectancy theory interpretation of gender differences. *Human Communication Research*, *18*(2), 177–208. https://doi.org/10.1111/j.1468-2958.1991.tb00543.x

Casad, B. J., & Bryant, W. J. (2016). Addressing stereotype threat is critical to diversity and inclusion in organizational psychology. *Frontiers in Psychology*, *7*, 8. https://doi.org/10.3389/fpsyg.2016.00008

Casad, B. J., Franks, J. E., Garasky, C. E., Kittleman, M. M., Roesler, A. C., Hall, D. Y., & Petzel, Z. W. (2021). Gender inequality in academia: Problems and solutions for women faculty in STEM. *Journal of Neuroscience Research*, *99*(1), 13–23. https://doi.org/10.1002/jnr.24631

Cech, E. A., & Blair-Loy, M. (2019). The changing career trajectories of new parents in STEM. *Proceedings of the National Academy of Sciences*, *116*(10), 4182–4187. https://doi.org/10.1073/pnas.1810862116

Chen, H., Pierson, E., Schmer-Galunder, S., Altamirano, J., Jurafsky, D., Leskovec, J., Fassiotto, M. A., & Kothary, N. (2020). Gender differences in patient perceptions of physicians' communal traits and the impact on physician evaluations. *Journal of Women's Health*, *30*(4), 551–556. https://doi.org/10.1089/jwh.2019.8233

Conde Ruiz, J. I., Ganuza, J. J., García, M., & Puch, L. A. (2022). Gender distribution across topics in the top five economics journals: A machine learning approach. *SERIEs*, *13*(1–2), 269–308. https://doi.org/10.1007/s13209-021-00256-2

Dahlen, S. (2021). Do we need the word 'woman' in healthcare? *Postgraduate Medical Journal*, *97*(1150), 483–484. https://doi.org/10.1136/postgradmedj-2021-140193

Filippou, P., Mahajan, S., Deal, A., Wallen, E. M., Tan, H. J., Pruthi, R. S., & Smith, A. B. (2019). The presence of gender bias in letters of recommendations written for urology residency applicants. *Urology*, *134*, 56–61. https://doi.org/10.1016/j.urology.2019.05.065

Fiske, S. T. (2012). Warmth and competence: Stereotype content issues for clinicians and researchers. *Canadian Psychology = Psychologie Canadienne*, *53*(1), 14–20. https://doi.org/10.1037/a0026054

Fiske, S. T. (2018). Stereotype content: Warmth and competence endure. *Current Directions in Psychological Science*, *27*(2), 67–73. https://doi.org/10.1177/0963721417738825

Fitzgerald, C., & Hurst, S. (2017). Implicit bias in healthcare professionals: A systematic review. *BMC Medical Ethics*, *18*(1), 19. https://doi.org/10.1186/s12910-017-0179-8

Garcia, D., Mitike Kassa, Y., Cuevas, A., Cebrian, M., Moro, E., Rahwan, I., & Cuevas, R. (2018). Analyzing gender inequality through large-scale Facebook advertising data. *Proceedings of the National Academy of Sciences*, *115*(27), 6958–6963. https://doi.org/10.1073/pnas.1717781115

Government Accountability Office. (2023). *Medicaid managed care plans: States' actions to address managed care plans that have a large share of enrollees*. (Publication No. GAO-23-106041). U.S. Government Publishing Office. https://www.gao.gov/assets/gao-23-106041.pdf

Gupta, S., & Jordan, K. (2022). Understanding gender bias toward physicians using online doctor reviews. *Psychology of Language & Communication*, *26*(1), 18–41. https://doi.org/10.2478/plc-2022-0002

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 & Counseling*, *95*(3), 313–318. https://doi.org/10.1016/j.pec.2014.03.015

Himmelstein, M. S., & Sanchez, D. T. (2016). Masculinity in the doctor's office: Masculinity, gendered doctor preference and doctor–patient communication. *Preventive Medicine*, *84*, 34–40. https://doi.org/10.1016/j.ypmed.2015.12.008

Hu, N., Zhang, T., Gao, B., & Bose, I. (2019). What do hotel customers complain about? Text analysis using structural topic model. *Tourism Management*, *72*, 417–426. https://doi.org/10.1016/j.tourman.2019.01.002

Li, S., Lee-Won, R. J., & McKnight, J. (2019). Effects of online physician reviews and physician gender on perceptions of physician skills and primary care physician (PCP) selection. *Health Communication*, *34*(11), 1250–1258. https://doi.org/10.1080/10410236.2018.1475192

Lu, B., Ott, M., Cardie, C., & Tsou, B. K. (2011). Multi-aspect sentiment analysis with topic models. In M. Spiliopoulou, H. Wang, D. Cook, J. Pei, W. Wang, O. Zaiane, & X. Wu (Eds.), *2011 IEEE 11th International Conference on Data Mining Workshops* (pp. 81–88). IEEE. https://doi.org/10.1109/ICDMW.2011.125.

Marjanovic, S., Stańczak, K., Augenstein, I., & Shook, N. J. (2022). Quantifying gender biases towards politicians on Reddit. *PLOS ONE*, *17*(10), e0274317. https://doi.org/10.1371/journal.pone.0274317

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*. https://doi.org/10.48550/arXiv.1301.3781

Nicolai, J., & Demmel, R. (2007). The impact of gender stereotypes on the evaluation of general practitioners' communication skills: An experimental study using transcripts of physician-patient encounters. *Patient Education & Counseling*, *69*(1–3), 200–205. https://doi.org/10.1016/j.pec.2007.08.013

Nicolas, G., Bai, X., & Fiske, S. T. (2022). A spontaneous stereotype content model: Taxonomy, properties, and prediction. *Journal of Personality and Social Psychology*, *123*(6), 1243–1263. https://doi.org/10.1037/pspa0000312

Pair, E., Vicas, N., Weber, A. M., Meausoone, V., Zou, J., Njuguna, A., & Darmstadt, G. L. (2021). Quantification of gender bias and sentiment toward political leaders over 20 years of Kenyan news using natural language processing. *Frontiers in Psychology*, *12*, 712646. https://doi.org/10.3389/fpsyg.2021.712646

Pennington, C. R., Heim, D., Levy, A. R., Larkin, D. T., & Pavlova, M. A. (2016). Twenty years of stereotype threat research: A review of psychological mediators. *PLOS ONE*, *11*(1), e0146487. https://doi.org/10.1371/journal.pone.0146487

Rao, P., & Taboada, M. (2021). Gender bias in the news: A scalable topic modelling and visualization framework. *Frontiers in Artificial Intelligence*, *4*, 664737. https://doi.org/10.3389/frai.2021.664737

Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). stm: An R package for structural topic models. *Journal of Statistical Software*, *91*(2), 1–40. https://doi.org/10.18637/jss.v091.i02

Robinson, G. E. (2003). Stresses on women physicians: Consequences and coping techniques. *Depression and Anxiety*, *17*(3), 180–189. https://doi.org/10.1002/da.10069

Roter, D. L., Hall, J. A., & Aoki, Y. (2002). Physician gender effects in medical communication: A meta-analytic review. *JAMA*, *288*(6), 756–764. https://doi.org/10.1001/jama.288.6.756

Samulowitz, A., Gremyr, I., Eriksson, E., & Hensing, G. (2018). "Brave men" and "emotional women": A theory-guided literature review on gender bias in health care and gendered norms towards patients with chronic pain. *Pain Research & Management*, *2018*, 6358624. https://doi.org/10.1155/2018/6358624

Shanthakumar, S. G., Seetharam, A., & Ramesh, A. (2020). Analyzing societal impact of Covid-19: A study during the early days of the pandemic. In J. Hu, G. Min, N. Georgalas, Z. Zhao, F. Hao, & W. Maio (Eds.), *2020 IEEE International Conference on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (Ispa/bdcloud/SocialCom/SustainCom)* (pp. 852–859). IEEE. https://doi.org/10.1109/ISPA-BDCloud-SocialCom-SustainCom51426.2020.00132.

Sheridan, M. (2021). *The effects of patient-physician interaction time on overall visit satisfaction and prevention of physician burnout* [Unpublished Honors Thesis]. Liberty University.

Shikhar, S. (2020). *Multi-aspect-sentiment-classification-for-online-medical-reviews*. https://github.com/Shikhar-S/Multi-Aspect-Sentiment-Classification-for-Online-Medical-Reviews/

Sonawane, S. (2017). *Ratings predictor*. https://github.com/turbosantosh/MLReviewsToRatings

Timmermans, S. (2020). The engaged patient: The relevance of patient–physician communication for twenty-first-century health. *Journal of Health and Social Behavior*, *61*(3), 259–273. https://doi.org/10.1177/0022146520943514

Tran-Harding, K. (2021, December 19). *The physician gender bias: What every female has faced*. sheMD. https://www.shemd.org/post/the-physician-gender-bias-what-every-female-has-faced

Tran, N., Hayes, R. B., Ho, I. K., Crawford, S. L., Chen, J., Ockene, J. K., Bond, M., Rayman, P., Dean, B., Smith, S., Thorndyke, L., Frankin, P., Plummer, D., & Pbert, L. (2019). Perceived subtle gender bias index: Development and validation for use in academia. *Psychology of Women Quarterly*, *43*(4), 509–525. https://doi.org/10.1177/0361684319877199

Wang, Y. F., Lee, Y. H., Lee, C. W., Lu, J. Y., Shih, Y. Z., & Lee, Y. K. (2023). The physician-patient communication behaviors among medical specialists in a hospital setting. *Health Communication*, 1–11. https://doi.org/10.1080/10410236.2023.2210379

Watanabe, K., & Xuan-Hieu, P. (2022, April 10). *Seededlda: Seeded-LDA for topic modeling*. The Comprehensive R Archive Network. https://cran.r-project.org/web/packages/seededlda/seededlda.pdf

Watanabe, K., & Zhou, Y. (2022). Theory-driven analysis of large corpora: Semisupervised topic classification of the UN speeches. *Social Science Computer Review*, *40*(2), 346–366. https://doi.org/10.1177/0894439320907027

Watson, B., & Gallois, C. (1998). Nurturing communication by health professionals toward patients: A communication accommodation theory approach. *Health Communication*, *10*(4), 343–355. https://doi.org/10.1207/s15327027hc1004_3

Yang, Y., White, K. R. G., Fan, X., Xu, Q., & Chen, Q. W. (2020). Differences in explicit stereotype activation among social groups based on the stereotype content model: Behavioral and electrophysiological evidence in Chinese sample. *Brain Sciences*, *10*(12), 1001. https://doi.org/10.3390/brainsci10121001

Zhang, Y., & Ramesh, A. (2018). Fine-grained analysis of cyberbullying using weakly-supervised topic models. In F. Bonchi, F. Provost, T. Eliassi-Rad, W. Wang, C. Cattuto, & R. Ghani (Eds.), *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, (pp. 504–513). IEEE. https://doi.org/10.1109/DSAA.2018.00065