Psychology of Language and Communication 2022, Vol. 26, No. 1

sciendo

DOI: 10.2478/plc-2022-0002

Sonam Gupta, Kayla Jordan
Data Science, Harrisburg University of Science & Technology, United States

Understanding gender bias toward physicians using online doctor reviews

Gender bias continues to be an ongoing issue in the field of medicine. While bias may come in many forms, patients' biases and perceptions have been understudied and may impact adherence to treatment, leading to unequal outcomes. Online reviews for doctors are a naturalistic way to study gender bias. In this study, we leveraged the LIWC psychological linguistic analysis tool to analyze the language styles of ZocDoc and RateMDs reviews and understand the potential role of gender in patients' perceptions of their doctors. Mean differences were calculated using bootstrapped hierarchical linear modeling. We found that reviews for female physicians are generally more informal and emotional than those for male physicians. While our study was exploratory, the results suggest that both patients and physicians need to increase their awareness of how their biases may be affecting how they give and receive vital health information.

Key words: interpersonal communication, LIWC, gender bias, naturalistic language analysis, gendered references.

Address for correspondence: Sonam Gupta, Data Science, Harrisburg University of Science & Technology, Harrisburg, United States.

E-mail: SGupta@my.harrisburgu.edu

Gender inequality and bias have been a continuous problem for several decades. They can ruin professional relationships or affect career growth for individuals. In recent years, various industries have been trying to balance the workforce among men and women in all jobs, but women still seem to struggle to make their place in professional and leadership roles. In many professions, ranging from sports, academia, and professional leadership to health care, women need to work extra hard to be considered equal to their male counterparts.

From sexual harassment to discrimination to professional stereotyping, female physicians often face many more challenges than their male counterparts, which can lead to their expertise being ignored or diminished. Robinson (2003) reviewed Medline articles about the stresses, discrimination, and other difficulties that women physicians go through. Some of the common stresses were: minority status discrimination, no role models and mentors in the field, no promotion to senior positions, or preconceived notions of women leaving their jobs after having children. While there are many avenues to address gender bias in the medical field, in the current research we leveraged advancements in natural language processing (NLP) and communication technologies to investigate potential gender bias in patients' perceptions of their doctors.

In the age of social media, online reviews play an important role in how people choose different services, from restaurants to doctors. Online reviews of physicians are particularly interesting, as many factors beyond a doctor's expertise may influence the review, such as interpersonal skills and the patients' own biases. These nonmedical factors are a potential source of gender bias, both positive and negative. For example, a female physician may be expected to exude more warmth and may be penalized for not conforming to stereotypes, or a male patient's personal biases may lead them to be more willing to accept recommendations from a male physician. The goal of this study was to use online physicians' reviews to examine patients' gender bias towards physicians.

Gender Bias towards Professional Women

Yip (2018) examined gender bias in the media coverage articles of male and female tennis players and found that female players were portrayed more in negative notions over few topics like athletic capabilities and other weaknesses, as well judged over their physical appearance, family, and personal relationships whereas the masculinity of male players was represented more dominantly. Sports journalism is another field where women journalists are under-represented along with the media coverage for women involved in sports. There is a growing number of women participants in both sports and sports journalism but they are still under-represented as compared to male journalists and/or sportsmen (Schmidt, 2017).

Historically, female researchers in academia have been treated unequally due to the societal perceptions and biases, as females were often associated with

family or household responsibilities. Male researchers received more credit for their research or were considered for grants more often in the grant peer review processes. Morgan et al. (2018) discussed findings from various focus groups on gender stereotypes in the peer review process. The findings suggested that gender bias was rooted in the historical beliefs about men and women, resulting in fewer citations for women's academic work as well as in women obtaining fewer leadership roles. Unconscious bias manifests in different industries for female professionals, which results in an unsteady ladder of success in their careers (Kaiser & Wallace, 2016). The reasons for this unconscious bias could be due to societal perceptions and stereotypes.

Gender Bias in Health Care

The preconceived notions and stereotypes regarding gender are quite prominent in the health care industry as well. Keeping women out from important medical decisions or undermining their expertise has been a common practice in academic medicine. Lewiss and Jagsi (2021) observed the lack of recognition of contributions from female doctors during the COVID-19 pandemic and saw a rise in gender bias. Female physicians have been facing this bias in the doctor-patient relationship as well. The doctor-patient relationship can have significant consequences for the patient's health and life, and when such biases or prejudices enter that relationship, unfair treatment and outcomes can result. One doctor recounted on Twitter, "While an Acute Care Surgery senior resident I got a patient complaint that they had not seen a 'doctor' for their entire stay but that the 'head nurse' was great...Sir, not only did I do most of your surgery, I introduced myself as Dr. every morning!" (Salles, 2019). Though these are just anecdotes, substantive research supports the impact that gender has on both female patients and physicians, but as demonstrated by these examples, social media presents an opportunity to expand this research in new directions.

Gender and Doctor-Patient Communication

Such differences in treatment and outcomes may at least be partially explained by differences in doctor-patient communication. Risberg et al. (2009) found that physicians would ask family-related questions more often to their female patients compared to the male ones. Roter et al. (2002) analyzed various physician communications transcript data where they found female physicians to be more involved in interpersonal communication with their patients, which led to longer visits as compared to their male colleagues. Ali et al. (2019) discussed the usage of positive sentiment language, which was associated more frequently with female doctors whereas male physicians were recommended to converse more

about the future with the patients and let them speak more so that it can lead to better patient prognosis.

Sandhu et al. (2009) found that the genders in the doctor-patient dyad impacted their interactions, verbal and nonverbal communication quality, consultation times, and conversation content. For instance, female to female dyads were found to have a longer consultation length with more personable conversations whereas male doctor to female patient dyads had the least patient-centered conversations. Beyond communication quality and content, the language styles used by both physicians and patients may further explain important differences in the doctor-patient relationship. This was the focus of the current study.

Language Analysis in the Health Care Domain

Language styles are a subtler form of potential bias and may reveal information beyond that revealed by traditional measures. For example, self-reports, though commonly used, can be biased in many ways due to the respondents not paying attention to the questions, tending toward positive responses, or not remembering all the relevant information while responding (Whitley et al., 2013). Analyses of language styles can address these limitations of self-reported measures by granting a naturalistic, everyday view of people's thoughts and attitudes through the words they choose to use (Tausczik & Pennebaker, 2010).

The same is true in the health care domain. Analyzing conversations between physicians and their patients, Falkenstein et al. (2016) found evidence that patients were satisfied and likely to follow up if their physicians interacted with them optimistically, as measured by LIWC2015 (Pennebaker, Boyd, Jordan, & Blackburn, 2015). Furthermore, when physicians sensed optimistic language from their patients as well, the overall patient outcome was more successful. With a detailed analysis of patient reviews of their doctors, Sen et al. (2017) found that higher ratings were closely associated with positive communication styles and the quality of words that the doctors used. The higher number of positive and encouraging words a doctor used in their conversation with the patients, the higher and more positively that doctor was rated. According to the authors' findings, patients appreciated when the doctor would equally participate in the conversation with them as well as encourage the patients throughout the treatment to ask questions.

Text analysis has been used increasingly often to analyze unstructured text data in the health care domain. Cammel et al. (2020) analyzed patient experience free-text responses using NLP tools and found that using such data, hospitals and clinicians could improve the health care experience for patients and prioritize the domains for improvement by considering patient feedback/complaints. The authors also found that using such an analysis on patient responses can refine the kind of questions asked in surveys and questionnaires to get a better understanding

of patients' experiences. Zhao et al. (2020) analyzed Twitter posts to investigate mental health signals for gender minority groups. They found that sexual and gender minorities (SGM) had more negative tweets and were found to be much happier to live in places with other SGM groups.

One relatively new source of data on the doctor-patient relationship are online doctor reviews found on websites like ZocDoc.com, RateMDs.com, Vitals.com, and HealthGrades.com. Such reviews provide insights into subtler forms of potential bias by showing how patients viewed their interactions with their doctors and how the gender of the doctor may have influenced that relationship. Marrero et al. (2019) performed a qualitative analysis on online reviews from websites like RateMDs.com and yelp.com. They found that the reviews involving doctors' social and technical skills differed in positive and negative sentiment based on the gender of the physician. They also found that many patients reviewed male physicians more positively and based on their technical skills as compared to female physicians. Approximately 97% of the time, the patients reviewed their female physicians for their social skills.

Current Research

Our study examined how gender bias in the health care industry can be identified from online patient reviews to understand the patients' perceptions of their physicians. The current study used two different online patient reviews datasets, one from RateMDs and the other from ZocDoc. The text analysis on both datasets was done using LIWC2015 (Pennebaker et al, 2015). We used the ZocDoc data to test hypothesized differences. The RateMDs data was used for confirmatory analysis and to test the robustness and generalizability of the findings.

Using psychological text analysis, we investigated the language of reviews to determine if how patients reviewed male and female doctors, revealed gender bias from the patients' perspective. Specifically, we examined the informality, socio-emotional content, and gendered-ness in the text of the reviews of male versus female physicians and compared them to the numerical rating patients gave their physicians. We used LIWC variables to measure these attributes of text. They were described as follows:

Informality

One useful metric of the quality of the doctor-patient interaction is perceived informality. High informality can be an indication of friendliness and warmth whereas low informality can be an indication of professionalism and competence (Kovanović et al., 2018; Dzogang et al., 2018; Pennebaker et al., 2014). LIWC measures informality in two ways: analytic thinking and informality. Lower analytic thinking scores indicate more informal and personal language whereas

higher scores indicate more formal and logical language. Informal language is focused on people and actions that are reflected in the use of pronouns, auxiliary verbs, and adverbs. Formal language is focused on ideas and concepts that are reflected in the use of prepositions and articles.

The informality dictionary includes swear words, netspeak (e.g., “brb,” “lol,” and “thx”), and filler words (e.g., “hm,” “ok,” “you know”). We expected that reviews about male physicians would be higher in analytic and lower in informality, as patients are more likely to focus on their technical skills and competence (Marrero et al., 2019). On the other hand, patients are likely to see interactions with female physicians as more friendly and informal (Derose et al., 2001; Bertakis et al., 1995).

Socioemotional Content

The most obvious metric of the quality of the doctor-patient interactions are the patients’ emotional reactions. We measured the emotional tone of the patients’ reviews in three ways. First, we used the tone variable in LIWC, which is a standardized metric of the difference between the use of terms for positive (e.g., “happy,” “good,” and “enjoy”) and negative emotions (e.g., “bad,” “terrible,” and “dislike”). Additionally, we considered the usage of positive and negative emotion terms on their own.

Finally, we used LIWC’s social dictionary with words like “friend”,” family,” and ”talk” to measure the extent to which patients perceived their physicians caring about their lives outside their specific symptoms. Given the past research showing that female physicians were rated more highly on social and emotional components of interactions (Hall et al., 2011), we expected reviews of female physicians to be higher in Tone, positive emotion, and social references.

Gendered References

To both ensure that our method of assigning physician gender was accurate and to explore the extent to which the reviews were gendered, we used the male and female reference dictionaries from LIWC. Male references included terms like ”boy,” ”man,” and ”his,” while female references included terms like ”lady,” ”girl,” and ”she." Based on the past research and common perceptions, we expected reviews of female doctors to be more gendered than those of male doctors (Bigler & Leaper, 2015).

Based on past research and common stereotypes, we proposed the following hypotheses:

H1: Reviews of male physicians are more formal (measured by LIWC2015’s analytic and informal dimensions) than reviews of female physicians.

H2: Reviews of female physicians contain more social and emotional language than reviews of male physicians.
H3: Reviews of female physicians contain more gendered references than reviews of male physicians.

General Method

Gender differences in the LIWC scores and ratings were analyzed using standardized mean differences, measured with Cohen's *d*, and the associated confidence intervals (CIs) of the mean difference calculated from bootstrapped hierarchical linear modeling (HLM) models to account for multiple reviews about the same doctor. One of the challenges of the data was that the same doctor could be rated by multiple individuals, violating the assumption of independence required for parametric tests like linear regression and analysis of variance (ANOVA). As such, we used HLM in this case to account for nonindependence of reviews about the same doctor. The LIWC variable scores, along with the review ratings (overall, waiting time, and bedside manner), were used as dependent variables, gender of the doctors (male and female) was used as the independent variable, and the doctors' names were used as a random effect. Additionally, we examined the correlations between the linguistic metrics and numeric ratings of doctors separately for male and female doctors. This exploratory analysis was to determine if numerical ratings of male and female doctors may be influenced by different factors. The statistical calculations were conducted with *R* using the *MOTE* (Buchanan et al., 2019), *nlme* (Pinheiro et al., 2019), and *boot* packages (Canty & Ripley 2019).

Rather than using *p* values, results were interpreted based on CIs, which include the same information as a *p* value and also give information on the precision of the estimates (du Prel, 2009). First, results were judged in terms of statistical significance (e.g., the CI did not include 0). Second, results that did not meet statistical significance were judged as to whether the effect was in the hypothesized direction, suggesting that the effect may be present but that the sample size was not sufficient.

The next step was to analyze the text of the reviews using LIWC2015, a psychological language analysis tool that focuses on understanding social, emotional, and cognitive aspects from any given text. LIWC is a dictionary-based method that gives the percentage of a text which falls into any one of over 80 validated categories. For example, a review like "The new doctor was kind" would score 20 on the positive emotion category, as 20% (1 in 5 words) are in the positive emotion dictionary. The LIWC categories used in this analysis have been described in the literature review section.

Study 1

Dataset

The website ZocDoc, an online platform for patients' reviews of their physicians, was used to extract the data for this study. The dataset used in this research was acquired from a GitHub repository by Sonawane (2017) and included only reviews of primary care physicians in Manhattan, New York. The dataset contains doctors' names, location, verified patient names, various ratings (on a scale of 1-5), doctor's degree, review date, and text reviews written by patients.

The dataset had a total of 19,372 reviews (after removing duplicates) for 555 unique doctors from 2008 to 2015 (214 female and 341 male doctors). Patients rated their physicians on three dimensions: the overall rating of the physician (M = 4.66, SD = 0.89, Mdn = 5), a specific rating of the doctor's bedside manner (M = 4.7, SD = 0.77, Mdn = 5) and time spent in the waiting area before seeing their doctor (M = 4.34, SD = 0.83, Mdn = 5).

Text Analysis

The analysis was done in three steps: labeling of doctors' gender, language analysis using LIWC2015 (Pennebaker et al., 2015), and computing means and SDs for each physician gender using a bootstrapped HLM. Reviews containing less than 10 words were excluded from the analysis, leaving 15168 reviews. The gender labeling was based on the number of pronouns used per review. Reviews using masculine pronouns (e.g.," he", "him," and "his") were labeled to be about male physicians whereas those using feminine pronouns (e.g., "she", "her," and "hers") were labeled to be about female physicians. If there were no pronouns used in the reviews, the gender of the doctor was manually assigned by the first author based on the name of the doctors. For further validation of the gender labeling, both authors independently assigned genders to 100 doctor names, and agreed upon 97 of those labels. Additionally, the authors' gender labels corresponded to the gender labels based on the pronouns used in the reviews for 95 of 100 doctors.

Results

The results including means and SDs of the 1000 HLM iterations, the effect sizes (e.g., standardized mean differences between male and female doctors), and their associated CIs for each dependent variable are shown in Table 1. The following forest plots show the statistical significance among the variables as well as the direction of differences. The results are explained in depth, below.

Table 1. *Means, SDs, and Cohen's d, lower and upper level of d, from Hierarchical Linear Modeling (HLM) Analysis*

Hypothesis	Dependent variable	*M* (female)	*SD* (female)	*M* (male)	*SD* (male)	*d*	*LL d*	*UL d*
Ratings	Overall rating	4.59	2.46	4.58	2.62	0.004	−0.02	0.03
	Bedside manner	4.66	2.15	4.63	2.36	0.01	−0.01	0.04
	Wait time	4.22	2.92	4.26	2.90	−0.01	−0.04	0.01
H1	Analytic	39.96	49.32	44.35	48.31	−0.09	−0.12	−0.05
	Informal	0.59	2.45	0.56	2.45	0.01	−0.01	0.04
H2	Tone	82.04	63.65	80.84	61.18	0.01	−0.01	0.05
	Positive emotion	8.19	11.46	7.87	11.08	0.02	−0.003	0.06
	Negative emotion	0.70	3.12	0.71	3.22	−0.004	−0.03	0.02
	Social	12.22	13.66	11.57	12.64	0.04	0.01	0.08
H3	Female	4.85	10.56	0.13	2.62	0.65	0.62	0.68
	Male	0.28	5.56	4.24	8.49	−0.53	−0.56	−0.50

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

Ratings

Numerical ratings are a traditional way for patients to rate the satisfaction level with their doctor's visits. Thus, we analyzed the three different ratings (overall, bedside manner, wait time) from the dataset. Figure 1 shows that no statistically significant differences were found in any of the numerical ratings for male versus female physicians (see Table 1 for specific values). Results further support no evidence of gender bias based on numerical rating, but linguistic analysis may identify subtler, more specific types of bias.

Formality

Two different LIWC dimensions were used to measure formality and analytic language. Analytic language was lower in reviews about female physicians (M = 39.96, SD = 49.32) than that of male physicians (M = 44.35, SD = 48.31), as shown in Figure 2. The difference was small, but statistically significant (d = −0.09, 95% CI = [−0.12, −0.05]). Figure 3 shows that reviews for male physicians (M = 0.56, SD = 2.45) had less informal language compared to the reviews of female physicians (M = 0.59, SD = 2.45). The difference was not statistically significant, but it was in the predicted direction (d = 0.01, 95% CI = [−0.01, 0.04]). These differences supported our predictions derived from past work showing that patients see the interactions with female physicians as more informal and personal and interactions with their male colleagues as more technical and impersonal (Bertakis et al., 1995; Derose et al., 2001). One potential downside to this is patients may also be signaling less respect for female physicians' competence and expertise by using more informal language.

Figure 1. *Forest Plot for Numerical Ratings*

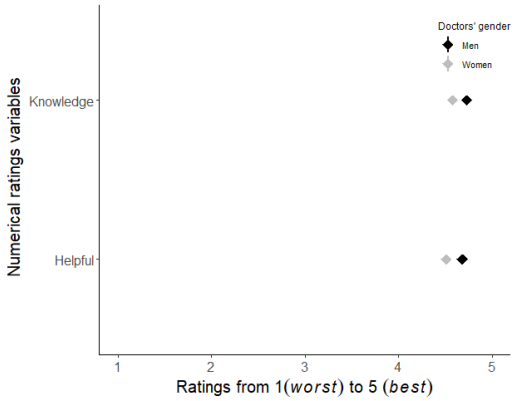
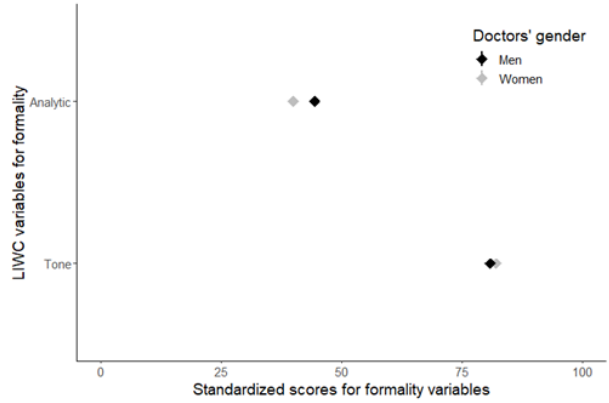


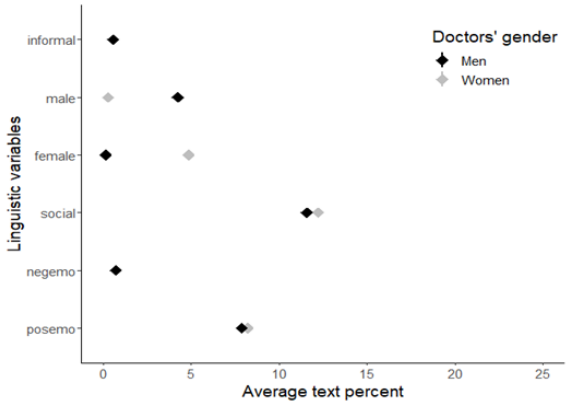
Figure 2. *Forest Plot for LIWC Summary Variables*



Socioemotional Content

The emotional tone was slightly higher in reviews of female physicians (M = 82.04, SD = 63.65) than in those of male physicians (M = 80.84, SD = 61.18), indicating more positive sentiment (measured by the LIWC variable "posemo") for female physicians, as seen in Figure 2. The difference for tone (d = 0.01, 95% CI = [−0.01, 0.05]) was not statistically significant, but was in the predicted direction. Positive emotion language alone was also in the predicted direction, but no difference was found in negative emotion (measured by the LIWC variable "negemo") alone (see Table 1 for detailed values).

Figure 3. *Forest Plot for LIWC Dictionary Measures*



Additionally, we considered the social content of the reviews using LIWC's social dictionary. Figure 3 shows that the reviews of female physicians (M = 12.22, SD = 13.66) contained more social words than those of male physicians (M = 11.57, SD = 12.64). The difference was statistically significant and in the predicted direction (d = 0.04, 95% CI = [0.01, 0.08]). The results of these metrics were in line with the past research suggesting that female physicians were more likely to consider the social and emotional factors with patients as well as exhibit more interpersonal skills in their interactions (Hall et al., 1994).

Gender References

The metric for gendered references was higher in the reviews of female physicians (female references, M = 4.85, SD = 10.56) than of male physicians (male references, M = 4.24, SD = 8.49). As predicted, the difference in gendered language was higher in reviews of female physicians (female references in female vs. male doctors, d = 0.65, 95% CI = [0.62,0.68]) compared to male physicians (male references in female vs. male doctors, d = −0.53, 95% CI = [−0.56, −0.50]), implying that reviews of female physicians were more gendered than reviews of male physicians. When patients talked about their female physicians compared to male physicians, they explicitly mentioned the gender of their physicians more often.

Relationships between Ratings and Linguistic Styles

The correlations between the ratings and linguistic variables for male and female doctors are shown in Table 2 and suggest a few interesting implications. First, the high correlation between the patients' overall rating and bedside manner rating

(r = .87) suggests that patients base their overall perceptions of their physicians primarily on the physicians' bedside manner. Second, the overall ratings were correlated with tone (r = .47), positive emotion (r = .30), and negative emotion (r = −.29), which indicates that text analysis of emotion/sentiment may pick up some of the same variance as numerical evaluative ratings. Finally, the fact that gender differences were found in several of the linguistic categories, but not the numerical ratings suggests that reviews and other sources of text within the doctor-patient relationship can be useful in more clearly understanding sources and types of gender bias.

Differences in these correlations demonstrate that patients may not be considering the same factors to the same extent when rating physicians. Analytic language was significantly more important when rating female doctors (r = −.07, 95% CI = [−.09, −.04]) compared to male doctors (r = −.02, 95% CI [−.04, .001]). Emotional tone was also more important for rating female doctors (r = .51, 95% CI = [.50, .53]) than male doctors (r = .44, 95% CI = [.42, .46]). Along with tone, we found positive emotion to be a more deciding factor for rating female physicians (r = .33, 95% CI = [.30, .35]) than male physicians (r = .29, 95% CI = [.27, .30]). The gendered references for female physicians (r = .13, 95% CI = [.10,.15]) were significantly more important as compared to their male counterparts (r = .07, 95% CI = [.04, .09]). Overall, the factors that patients used to judge female doctors seemed more heavily weighted toward their informality, positivity, and gender conformity.

Discussion

In Study 1, the evidence supported our hypothesis that reviews of male physicians contained more formal language than reviews of female physicians. Furthermore, reviews of male physicians had the personal or informal notion, meaning that the conversations between male physicians and patients tended to be more technical, and with female physicians, the conversations were more personal. Results also

Table 2. *Correlations between Overall Rating and Other Variables for Study 1 Dataset and by Gender*

Hypothesis	LIWC metric	Male doctor	Female doctor
H1	Analytic	−0.02	−0.08*
	Informal	0.03	0.05*
H2	Tone	0.45*	0.52*
	Positive Emotion	0.29*	0.33*
	Negative Emotion	−0.29*	−0.30*
	Social	0.11*	0.15*
H3	Female	−0.06	0.14*
	Male	0.07*	0.01

* $p < 0.001$

suggested that the language of reviews of female physicians contained more social and interpersonal words and gendered references. The higher use of gendered references in reviews for female physicians indicated explicit mention of gender of the physician, which is a cue for gender bias. To understand the robustness and replicability of these findings, we considered another dataset in Study 2. It comprised doctors from different specialties and U.S. geographical regions that gave us more data to find cues of bias in language.

Study 2

Dataset

Online review data from RateMDs.com was used for Study 2, downloaded from a publicly available GitHub repository maintained by Shikhar (2020). This dataset included doctors from multiple specialties across the U.S. This dataset consisted of 9 different data columns: four separate numerical ratings (on a scale of 1-5, 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: $M = 4.65$, $SD = 0.99$, $Mdn = 5$), doctor ID, doctor's name, gender of the doctor, doctor's specialty, and text reviews. There was a total of 46,554 rows (after removing duplicates) with 3,552 male doctors and 2,647 female doctors.

Text Analysis

Unlike Study 1, the data in Study 2 did not require external labeling of the gender of the doctors, as it was readily available from the source. Study 2 followed the same steps for text analysis as Study 1, but the numerical ratings as well as staff and punctuality ratings were ignored since they were meant for the staff and not just the physicians. Otherwise, the analytic process mirrored Study 1, with the reviews being analyzed by LIWC, followed by calculating of means and SDs for reviews of male and female doctors using bootstrapped HLM to control for multiple reviews of the same doctor.

Results

The same analysis was conducted as in Study 1. Results are further explained in detail and specific values for these results can be found in Table 3 and Table 4.

Ratings

The ratings for helpfulness were higher for male physicians ($M = 4.68$, $SD = 1.78$) than for female physicians ($M = 4.51$, $SD = 2.27$). Similarly, for knowledge, the

Table 3. *Means, SDs, and Cohen's d, lower and upper level of d, from Hierarchical Linear Modeling (HLM) Analysis*

Hypothesis	Dependent variable	*M* (female)	*SD* (female)	*M* (male)	*SD* (male)	*d*	*LL d*	*UL d*
Ratings	Helpfulness	4.51	2.27	4.68	1.78	−0.08	−0.10	−0.06
	Knowledge	4.57	2.02	4.72	1.59	−0.08	−0.10	−0.06
H1	Analytic	46.43	44.93	48.95	43.89	−0.05	−0.07	−0.03
	Informal	0.51	2.17	0.56	2.21	−0.02	−0.03	−0.001
H2	Tone	76.43	56.97	77.90	53.10	−0.02	−0.04	−0.01
	Positive emotion	7.49	9.83	7.67	9.78	−0.01	−0.03	0.001
	Negative emotion	1.30	3.62	1.27	3.56	0.01	−0.01	0.02
	Social	12.56	12.17	11.80	11.99	0.06	0.04	0.08
H3	Female	4.81	7.04	0.31	1.92	0.97	0.95	0.99
	Male	0.27	1.80	4.21	6.79	−0.72	−0.73	−0.70

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

average ratings were higher for male physicians ($M = 4.72$, $SD = 1.5$) than female physicians ($M = 4.57$, $SD = 2.02$). Figure 4 shows that male doctors were rated higher than female doctors (Helpfulness: $d = -0.08$, 95% CI = [−0.10, −0.06], Knowledge: $d = -0.08$, 95% CI = [−0.10, −0.06]). However, the ratings covered only two dimensions and do not convey any other details regarding patient experience.

Formality

Figure 5 shows the standardized means for analytic language, a measure of formality. Reviews of male doctors ($M = 48.95$, $SD = 43.89$) were significantly more analytic and formal than reviews of female doctors ($M = 46.43$, $SD = 44.93$). The difference was small, but similar in magnitude to Study 1 ($d = -0.05$, 95% CI = [−0.07, −0.03]). As shown in Figure 6, the reviews of male physicians ($M = 0.56$, $SD = 2.21$) used slightly more words indicative of informality (e.g., "bc," "abt," "okay," and "dunno") than reviews of female physicians ($M = 0.51$, $SD = 2.17$), $d = -0.02$, 95% CI = [−0.03, −0.001]. While the difference in informality words runs contrary to our hypotheses, the very small base rate of informal language (less than 0.5% of the reviews on average) makes it difficult to draw any conclusions based on this measure. Focusing solely on the analytic measure, the studies taken together indicate a small but significant difference, with male physicians being reviewed more technically and female physicians being reviewed more formally. This difference is in line with the past research indicating that female doctors tend to be evaluated as more personable in communicating with their patients.

Socioemotional Content

Tone, as shown in Figure 5, was slightly higher in reviews of male physicians

Figure 4. *Forest Plot of Numerical Ratings*

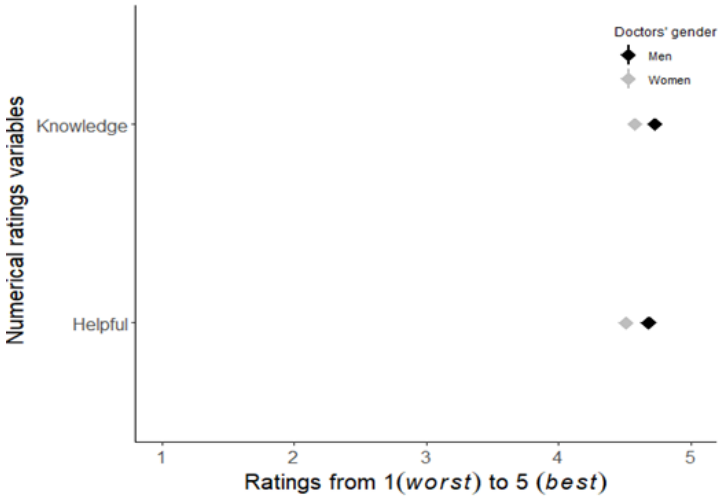
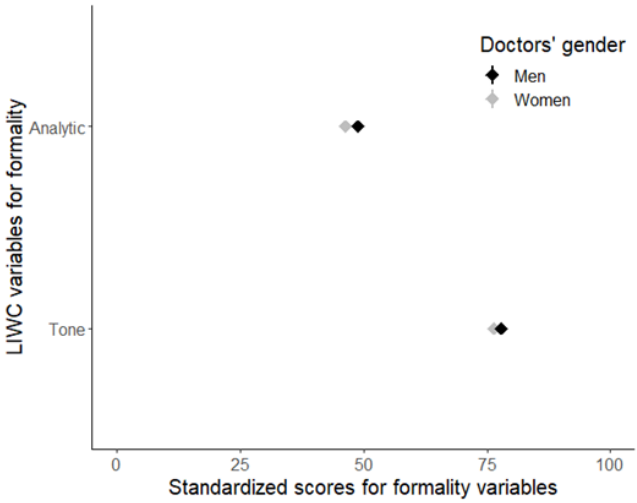


Figure 5. *Forest Plot for LIWC Summary Variables*



(M = 77.90, SD = 53.10) as compared to reviews of female physicians (M = 76.43, SD = 56.97), which did not align with our hypothesis and with past work. The difference was small, but statistically significant (d = −0.02, 95% CI = [−0.04, −0.01]). Similarly, as seen in Figure 6, the positive emotion (measured

by the LIWC variable posemo) in reviews of male physicians (M = 7.67, SD = 9.78) was marginally higher than in reviews of female physicians (M = 7.49, SD = 9.83) with a small, statistically significant difference (d = −0.01, 95% CI = [−0.03, 0.001]), whereas negative emotion (measured by the LIWC variable negemo) in the reviews was higher for female physicians (M = 1.30, SD = 3.62) than for male physicians (M = 1.27, SD = 3.56), but this was not statistically significant (d = 0.007, 95% CI = [−0.01, 0.02]). These findings run counter to the findings from Study 1 and to expectations based on past research, suggesting that the physician's gender does not robustly or consistently impact patients' emotional perceptions of them. However, these results do align with the numerical ratings of helpfulness and knowledge for this dataset which were higher for male physicians, suggesting emotional language in reviews captures the same information as numeric ratings.

Turning to the social content of reviews, as shown in Figure 6, reviews of female physicians (M = 12.56, SD = 12.17) contained significantly more social references than reviews of male physicians (M = 11.80, SD = 11.99). The difference was statistically significant and of similar magnitude to Study 1, (d = 0.06, 95% CI = [0.04. 0.08]). Such a difference supports our hypothesis as well as past research, suggesting the female physicians care more about patients' social wellbeing and are more personable.

Gendered References

As per our predictions, the reviews of female physicians (M = 4.81, SD = 7.04) were found to contain more gendered references as compared to those of male physicians (M = 4.21, SD = 6.79), as shown in Figure 6. Furthermore, as a check on the data, reviews of female physicians used more female words (d = 0.97, 95% CI = [0.95, 0.99]) and reviews of male physicians used more male words (d = −0.72, 95% CI = [−0.73, −0.70]).

Relationships between Ratings and Linguistic Styles

Table 4 shows the correlations between the average of the numerical ratings (helpful and knowledge) and linguistic analysis variables for male and female physicians. The average of the two ratings (helpfulness and knowledge) was correlated with tone (r = .47), positive emotion (r = .28), and negative emotion (r = −.3), showing that the emotional analysis from the reviews may have variance just as the average of the numerical ratings. These correlations showed some important factors that patients considered while rating and reviewing their physicians.

The linguistic variables like analytic language were more important for reviewing female physicians (r = −.01, 95% CI = [−.04, .02]) than male physicians (r = .01, 95% CI = [−.02, .04]). In the same way, for female physician

Figure 6. *Forest Plot for LIWC Dictionary Measures*

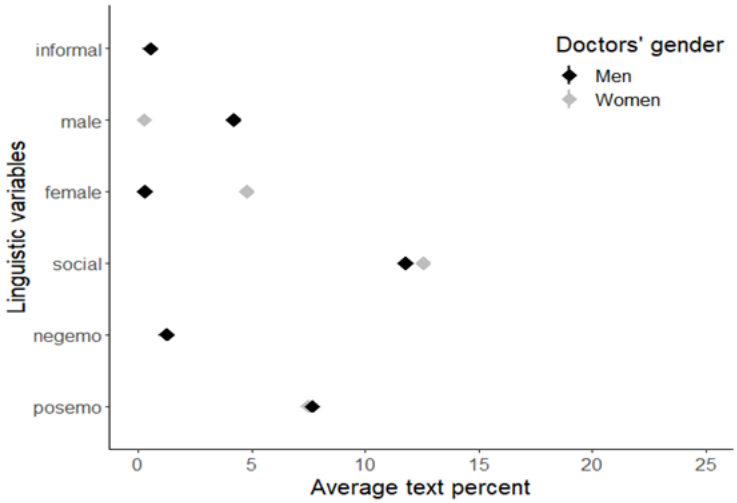


Table 4. *Correlations between Average of Helpful, Knowledge Rating, and Other Variables for Study 2, Split by Gender*

Hypothesis	LIWC metric	Male doctor	Female doctor
H1	Analytic	0.01	−0.01
	Informal	0.02*	0.03*
H2	Tone	0.43*	0.53*
	Positive emotion	0.25*	0.32*
	Negative emotion	−0.27*	−0.34*
	Social	0.001	0.03*
H3	Female	−0.02	0.01
	Male	−0.01	−0.01

* $p < 0.001$

reviews, tone (r = .53, 95% CI = [.50, .55]) and positive emotion (r = .32, 95% CI = [.28, .35]) were more important as compared to male physician reviews (tone: r = .43, 95% CI = [.40, .45], positive emotion: r = .25, 95% CI = [.21, .28]). Another linguistic factor to note is that gendered references for female physicians (r = .01, 95% CI = [−.02, .04]) seemed more significant than those for male physicians (r = −.01, 95% CI = [-.05, .01]). Like in Study 1, emotional tone, gendered references, and informality played an important role in the patients' reviews of their physicians.

Discussion

Study 2 results indicated slightly more use of informal language in reviews for male physicians than for female physicians, which did not align with what we hypothesized. However, reviews of male physicians were found to be more technical and formal, which supported our hypothesis and the past research. Similarly, in Study 2, the emotional tone in the reviews of male physicians was higher than in the reviews of female physicians. The social references in the reviews suggested being more favorable towards female physicians, meaning that female physicians may be more interpersonally involved with their patients. As in Study 1, there was a higher use of gendered references in reviews for female physicians in this dataset, indicating that direct mentions of physician gender that is a cue for gender bias.

General Discussion

Using psychological text analysis, we found that numerical reviews do not tell the whole story about patients' thoughts and feelings about their physicians. Text analysis of patient reviews gave a much more nuanced picture of patients' perceptions. Across both datasets, reviews of male physicians were more formal and logical than reviews of female physicians, partially supporting Hypothesis 1. Despite mixed results on informal language, potentially caused by low word frequencies, the findings are generally in line with past work showing that male physicians tend to be seen as less friendly and more technical whereas female physicians are often seen as more friendly and personable (Marrero et al., 2019; Sandhu et al. 2009).

Hypothesis 2 received only partial support: The use of emotional language in the reviews was not consistently related to physicians' gender. Exploratory correlational analysis with both datasets suggests that rather than reveal aspects of the physicians' style or behavior, emotional language in the reviews captured the patients' overall evaluation of their experience. The social aspect of Hypothesis 2 was supported in both studies. Reviews of female physicians contained significantly more social language, which conforms with studies showing that female physicians tend to spend more time discussing patients' families and social lives (Risberg et al., 2009; Roter et al, 2002).

Hypothesis 3 was supported in both studies: Reviews of female physicians contained significantly more references to their gender than reviews of male physicians. The explicit reference to gender happens not only to female physicians, but to many female professionals. People often default to most professionals being men, so when encountering female professionals, they are more likely to explicitly mention their gender, which may activate stereotypes and prejudices, leading to bias and other negative outcomes (Bigler & Leaper, 2015).

The effects found in this study were small, but that does not necessarily mean

they are insignificant. Multiple researchers have argued that labelling effects as small, medium, or large is limiting and incorrect (Gotz et al., 2022; Primbs et al., 2021). Rather, the significance of effect sizes should be interpreted relative to the research in the field and its potential real world significance. In these terms, the effects found in this study are significant. Text analysis using real world data and dictionary-based methods typically finds comparable effect sizes due to the variation and noise inherent to real world text data (Boyd et al., 2020; Seraj et al., 2021). These effects also potentially have real world consequences. Gender bias is often subtle. Small word choices and slight changes in tone can greatly impact the quality of human interaction. For example, ”lady doctor” versus ”doctor” only has a single word difference, but that single word conveys stereotypes and bias. Therefore, even though the effects found in this study are not traditionally large, they likely have real world impacts.

A few limitations to the current work should be mentioned and potentially examined in future studies. Study 2 had more data across multiple specialties of doctors than just primary care, as was the case in Study 1, which could have impacted the results. The proportions of male versus female physicians in different specialties vary, and the interaction between specialties and gender bias is worthy of further study. Another potential limitation could be the selection of patients writing reviews for physicians on these online platforms. In this study, we did not analyze patient demographics, as that information was not available. However, future work should explore how patient characteristics might mitigate or exacerbate gender bias.

While our work shows how language styles may help identify gender bias in doctor-patient interactions, it remains an open question as to how and to what extent the language styles used by doctors impact the patient outcomes. Analyzing language style in conversation transcripts between doctors and patients can reflect the doctors' stereotypes toward different genders or unconscious bias that can indirectly affect patient outcomes. With that knowledge, intervention work could help male and female physicians communicate more effectively to their male and female patients. Though much work remains to be done in this area, identifying the many forms of gender bias in the medical sphere is the first step to addressing and fixing it.

There can be severe implications of gender bias on the overall patient outcome if patients allow stereotypes and biases to color their interaction with their physicians, potentially leading to an incomplete or improper course of treatment. On the other hand, if doctors project gender bias towards their patients or allow their stereotypes to impact their interactions with patients, it can impact the overall patient care negatively. Our results suggest that patients, as well as physicians, should be aware of their unconscious biases, as they could hamper the doctor-patient relationship.

Acknowledgments

Citations for data availability are included in the manuscript. Code can be found at https://github.com/sonamgupta1105/patient_Reviews_analysis

Conflict of Interest Disclosure

Neither author has conflicting interesting to disclose.

Funding

No funding was used for this project.

Research Ethics Statement

Studies involved existing, publicly available data and were approved by the IRB at Harrisburg University of Science and Technology.

References

Ali, M. R., Sen, T., Nguyen, V. D., Hoque, M. E., Epstein, R. M., Rawassizadeh, R., & Duberstein, P. R. (2019). What computers can teach us about doctor-patient communication: Leveraging gender differences in cancer care. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)* (pp. 1–7). Institute of Electrical and Electronics Engineers.

Bertakis, K. D., Helms, L. J., Callahan, E. J., Azari, R., & Robbins, J. A. (1995). The influence of gender on physician practice style. *Medical Care, 33*(4), 407–416. https://doi.org/10.1097/00005650-199504000-00007

Bigler, R. S., & Leaper, C. (2015). Gendered language: Psychological principles, evolving practices, and inclusive policies. *Policy Insights from the Behavioral and Brain Sciences, 2*(1), 187–194. https://doi.org/10.1177/2372732215600452

Boyd, R. L., Blackburn, K. G., & Pennebaker, J. W. (2020). The narrative arc: Revealing core narrative structures through text analysis. *Science Advances, 6*(32), eaba2196. https://www.science.org/doi/10.1126/sciadv.aba2196

Buchanan E., Gillenwaters A., Scofield J., & Valentine K. (2019). *MOTE: Measure of the Effect: Package to assist in effect size, calculations, and their confidence intervals. R package version 1.0.2.* https://cran.r-project.org/web/packages/MOTE/MOTE.pdf

Cammel, S. A., De Vos, M. S., van Soest, D., Hettne, A. M., Boer, F., Steyerberg, E. W., & Boosman, H. (2020). How to automatically turn patient experience free-text responses into actionable insights: A natural language programming (NLP) approach. *BMC Medical Informatics and Decision Making, 20*, 97. https://doi.org/10.1186/s12911-020-1104-5

Canty, A., Ripley, B. (2019). *Boot: Bootstrap R (S-Plus) Functions. R package version 1.3-22*. https://cran.r-project.org/web/packages/boot/index.html

Derose, K. P., Hays, R. D., McCaffrey, D. F., & Bakerg, D. W. (2001). Does physician gender affect satisfaction of men and women visiting the emergency department? *Journal of General Internal Medicine, 16*(4), 218–226. https://doi.org/10.1046/j.1525-1497.2001.016004218.x

Dzogang, F., Lightman, S., & Cristianini, N. (2018). Diurnal variations of psychometric indicators in Twitter content. *PloS One, 13*(6), e0197002. https://doi.org/10.1371/journal.pone.0197002

Falkenstein, A., Tran, B., Ludi, D., Molkara, A., Nguyen, H., Tabuenca, A., & Sweeny, K. (2016). Characteristics and correlates of word use in physician-patient communication. *Annals of Behavioral Medicine, 50*(5), 664–677. https://doi.org/10.1007/s12160-016-9792-x

Götz, F., Gosling, S., & Rentfrow, J. (2022). Small effects: The indispensable foundation for a cumulative psychological science. *Perspectives on Psychological Science, 17*(1), 205–215. https://doi.org/10.1177/1745691620984483

Hall, J. A., Blanch-Hartigan, D., & Roter, D. L. (2011). Patients' satisfaction with male versus female physicians: A meta-analysis. *Medical Care, 49*(7), 611–617. https://doi.org/10.1097/MLR.0b013e318213c03f

Hall, J. A., Irish, J. T., Roter, D. L., Ehrlich, C. M., & Miller, L. H. (1994). Gender in medical encounters: An analysis of physician and patient communication in a primary care setting. *Health Psychology, 13*(5), 384–392. https://doi.org/10.1037/0278-6133.13.5.384

Hamberg, K., Risberg, G., Johansson, E. E., & Westman, G. (2002). Gender bias in physicians' management of neck pain: A study of the answers in a Swedish national examination. *Journal of Eomen's Health & Gender-Based Medicine, 11*(7), 653–666. https://doi.org/10.1089/152460902760360595

Kaiser, R. B., & Wallace, W. T. (2016). Gender bias and substantive differences in ratings of leadership behavior: Toward a new narrative. *Consulting Psychology Journal: Practice and Research, 68*(1), 72–98. https://doi.org/10.1037/cpb0000059

Kovanović, V., Joksimović, S., Mirriahi, N., Blaine, E., Gašević, D., Siemens, G., & Dawson, S. (2018, March). Understand students' self-reflections through learning analytics. In *Proceedings of the 8th international conference on learning analytics and knowledge* (pp. 389–398). https://doi.org/10.1145/3170358.3170374

Lewiss, R. E., & Jagsi, R. (2021). Gender bias: Another rising curve to flatten? *Academic Medicine, 96*(6), 792–794. https://doi.org/10.1097/ACM.0000000000003987

Marrero, K., King, E., & Fingeret, A. L. (2020). Impact of surgeon gender on online physician reviews. *The Journal of Surgical Research, 245*, 510–515. https://doi.org/10.1016/j.jss.2019.07.047

Morgan, R., Hawkins, K., & Lundine, J. (2018). The foundation and consequences of gender bias in grant peer review processes. *Canadian Medical Association Journal, 190*(16), E487–E488. https://doi.org/10.1503/cmaj.180188

Pennebaker, J. W., Boyd, R., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. University of Texas at Austin. https://repositories.lib.utexas.edu/handle/2152/31333

Pennebaker, J. W., Chung, C. K., Frazee, J., Lavergne, G. M., & Beaver, D. I. (2014). When small words foretell academic success: The case of college admissions essays. *PloS One, 9*(12), e115844. https://doi.org/10.1371/journal.pone.0115844

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team (2019). *nlme: Linear and nonlinear mixed effects models. R package version 3.1-140.* https://cran.r-project.org/web/packages/nlme/nlme.pdf

du Prel, J. B., Hommel, G., Röhrig, B., & Blettner, M. (2009). Confidence interval or p-value? Part 4 of a series on evaluation of scientific publications. *Deutsches Arzteblatt International, 106*(19), 335–339. https://doi.org/10.3238/arztebl.2009.0335

Primbs, M., Pennington, C. R., Lakens, D., Silan, M. A., Lieck, D., Forscher, P. S., ... & Westwood, S. (2021). There are no 'small'or 'large'effects: A Reply to Götz et al. (2021). *PsyArXiv*. https://doi.org/10.31234/osf.io/6s8bj

Risberg, G., Johansson, E. E. & Hamberg, K. (2009). A theoretical model for analysing gender bias in medicine. *International Journal for Equity in Health, 8*, 28. https://doi.org/10.1186/1475-9276-8-28

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

Salles, A. (2019, July 23). *Gender bias narratives in medicine.* Physician's Weekly. https://www.physiciansweekly.com/gender-bias-narratives-in-medicine

Sandhu, H., Adams, A., Singleton, L., Clark-Carter, D., & Kidd, J. (2009). The impact of gender dyads on doctor-patient communication: A systematic review. *Patient Education and Counseling, 76*(3), 348–355. https://doi.org/10.1016/j.pec.2009.07.010

Schmidt, H. C. (2017). Forgotten athletes and token reporters: Analyzing the gender bias in sports journalism. *Atlantic Journal of Communication, 26*(1), 59–74. https://doi.org/10.1080/15456870.2018.1398014

Seraj, S., Blackburn, K. G., & Pennebaker, J. W. (2021). Language left behind on social media exposes the emotional and cognitive costs of a romantic breakup. *Proceedings of the National Academy of Sciences, 118*(7), e2017154118. https://doi.org/10.1073/pnas.2017154118

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

Sonawane, S. (2017). *Ratings predictor* [Data set]. GitHub. https://github.com/turbosantosh/MLReviewsToRatings

Sen, T., Ali, M. R., Hoque, M. E., Epstein, R., & Duberstein, P. (2017). Modeling doctor-patient communication with affective text analysis. In *2017 seventh international conference on affective computing and intelligent interaction (ACII)* (pp. 170–177). IEEE. https://doi.org/10.1109/ACII.2017.8273596.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology, 29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Whitley, B. E., Kite, M. E., & Adams, H. L. (2013). *Principles of research in behavioral science (3rd ed.)*. Routledge. https://doi.org/10.4324/9780203085219

Yip, A. (2018). Deuce or advantage? Examining gender bias in online coverage of professional tennis. *International Review for the Sociology of Sport, 5*3(5), 517–532. https://doi.org/10.1177/1012690216671020

Zhao, Y., Guo, Y., He, X., Wu, Y., Yang, X., Prosperi, M., Jin, Y., & Bian, J. (2020). Assessing mental health signals among sexual and gender minorities using Twitter data. *Health Informatics Journal, 26*(2), 765–786. https://doi.org/10.1177/1460458219839621