"Jerk" or "Judgemental"? Patient Perceptions of Male versus Female Physicians in Online Reviews

Byron C. Wallace
College of Computer & Information Science
Northeastern University
byron@ccs.neu.edu

Michael J. Paul
College of Media, Communication & Information
University of Colorado Boulder
mpaul@colorado.edu

Abstract

We analyze patient reviews (posted online) of male and female physicians with respect to numerical ratings and language use. We find that females tend to receive less favorable numerical ratings overall, and that there seems to be higher variance in the sentiment of words used in reviews of female vs. those of male physicians.

Introduction and Motivation

Individuals are increasingly turning to the web to gather information relevant to their healthcare. An important example of this includes online reviews of physicians: a relatively recent survey (Fox and Duggan 2013) found that 72% of internet users have looked online for health information in the past year. And one in five of these users have looked for reviews of either particular treatments or doctors.

Patient-generated reviews are especially interesting as a data source because they provide a direct, unmediated window into the patient experience. Further, these reviews may influence other individuals' opinions of (potential) physicians (Grabner-Kräuter and Waiguny 2015), in turn affecting patient care. Indeed, Li et al. (2015) concluded via a randomized trial that exposure to negative reviews "led to a reduced willingness to use the physician's services." Much of the previous work on examining online reviews of physicians has been qualitative in nature (López et al. 2012; Gao et al. 2012; Kilaru et al. 2016). Following our prior work (Wallace et al. 2014; Paul, Wallace, and Dredze 2013), we adopt a more data-drive approach in this study.

Our focus in this work concerns investigating differences in online reviews of male versus female physicians, both with respect to patient satisfaction (ratings) and language use. We aim to complement the relatively robust body of evidence that strongly suggests that female physicians "don't get the credit they deserve" (Roter and Hall 2015). For instance, Hall et al. (2011) performed a meta-analysis of studies looking at patient satisfaction and concluded that female physicians "are not evaluated as highly by their patients, relative to male physicians, as one would expect based on their practice style and patients' values." Does the same hold in online reviews?

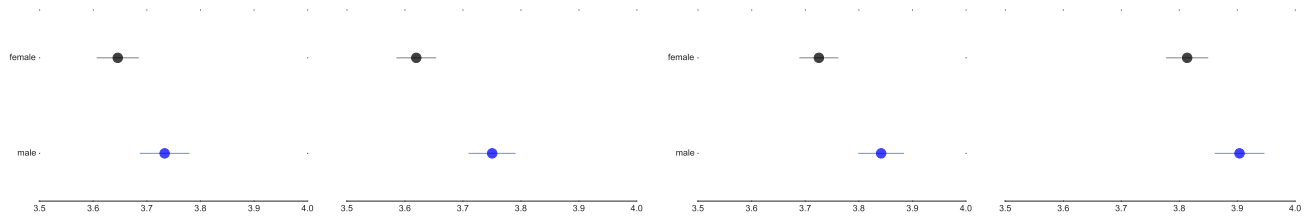
In the remainder of this paper, we (i) describe the dataset we use, (ii) present results from regression analyses of numerical ratings and (iii) results from a lexical analysis of review texts. We then (iv) conclude with a discussion of our results and limitations. As far as we are aware, this is the first work to explicitly explore the relationship between physician gender and attributes of online reviews regarding the care they provide.

Dataset

We use a subset of a corpus of reviews downloaded from `RateMDs.com`, a website of doctor reviews written by patients. Reviews comprise free-text accompanied by numerical ratings ranging from one (most negative) to five (most positive) across four categories: 'knowledge', 'helpfulness' 'punctuality' and 'staff'. The dataset we have assembled comprises 16,488 unique doctors. For additional details, see (Paul, Wallace, and Dredze 2013; Wallace et al. 2014).

The dataset does not explicitly contain the gender of the physicians reviewed, because this is not stored by `RateMDs.com`. However, we observed that this can be readily inferred from gendered pronoun use; review texts practically always refer to the physician, e.g., "he is very nice". Thus, to automatically infer the gender of physicians, we simply count up the number of male and female pronouns in each review text and assume that the gender of the physician agrees with the majority pronoun category. If no majority exists (e.g., when no gendered pronouns are used), we assign the label of "unknown". This simple inference strategy is made more robust by the fact that we usually have multiple reviews per physician. Thus for each review for a given physician we can independently use the strategy just described to infer the physicians' gender (as perceived by the patient), and then generate the final gender assignment by taking a majority vote over the constituent reviews. We discard the (relatively rare) cases in which the majority vote is "unknown".

This resulted in a dataset comprising 53,401 reviews of 16,488 unique physicians, 11,826 (72%) of whom we inferred were male and the remaining 4,662 (28%) female. On average, we have 3.2 reviews per physician.



(a) Helpfulness (b) Punctuality (c) Staff (d) Knowledge

Figure 1: Point estimates and 95% confidence intervals for male (bottom) and female (top) factors across the four targets (see Eq. 1).

Rating Analysis

Recall that each review comes attached with numerical scores assigned by the author with respect to four aspects of care: 'knowledge', 'helpfulness' 'punctuality' and 'staff'. These are provided on a five-point Likert scale where higher implies greater satisfaction. As a simple first analysis, we quantify the correlation between gender and these ratings.

One small complication is that we typically have multiple reviews per physician. We would like to model each individual physicians' score as a function of gender. Thus, we take the mean of multiple reviews for any given physician and treat this as a single target. This aggregation strategy has the advantage of being simple, but it does result in discarding variance across the reviews of individual physicians.

Indexing physicians by i and indexing a specific target (e.g., 'knowledge') by t with mean rating y_i^t, we perform univariate regressions of the following form:

$$y_i^t = \beta_0^t + \beta_{\text{male}}^t \cdot \mathcal{I}(\text{doctor } i \text{ is male}) \quad (1)$$

$\mathcal{I}$ is an indicator function set to 1 if physician i was inferred to be male and 0 otherwise. Thus, β_{male}^t is a coefficient capturing the correlation between being male and the review scores one receives for a given target.

We report results from each of the four independent univariate regressions in Table 1 and visually in Figure 1. One can observe that the male coefficient is significant for all aspects. That is, reviews of male physicians are apparently significantly more favorable than those of female doctors.

Although the result is suggestive, we certainly are *not* endorsing any sort of direct causal claim that ratings are lower *because* these physicians are female. Simpson's paradox may very well be at play here, and we have not attempted to correct for many possible sources of confounding. For example, it could be that female physicians for whatever reason are over-represented in specialties that tend to receive lower favorability ratings in general; we made no attempt to control for this. Another potential source of confounding may be systematic differences in the locations of female and male physicians; we know that ratings distributions vary across states (Wallace et al. 2014). We also note that the outcome variables (aspect ratings) are highly correlated, and hence a multivariate regression may be more appropriate.

Even with these caveats in mind, however, we find the apparent difference in ratings compelling. What accounts for the apparent systematic differences in patient sentiment regarding male and female physicians? To address this question, we next report results from a data-driven linguistic analysis that aims to tease out differences in how patients talk about male and female physicians.

	Coefficient	SE	P>|t|	95.0% CI
		Helpful		
β_0^{helpful}	3.6467	0.020	0.000	(3.608, 3.685)
$\beta_{\text{male}}^{\text{helpful}}$	0.0874	0.023	0.000	(0.042, 0.133)
		Punctuality		
$\beta_0^{\text{punctual}}$	3.6203	0.017	0.000	(3.587, 3.654)
$\beta_{\text{male}}^{\text{punctual}}$	0.1309	0.020	0.000	(0.091, 0.171)
		Staff		
β_0^{staff}	3.7263	0.018	0.000	(3.691, 3.762)
$\beta_{\text{male}}^{\text{staff}}$	0.1161	0.022	0.000	(0.074, 0.158)
		Knowledge		
$\beta_0^{\text{knowledge}}$	3.8143	0.018	0.000	(3.778, 3.850)
$\beta_{\text{male}}^{\text{knowledge}}$	0.0903	0.022	0.000	(0.048, 0.133)

Table 1: Regression analysis of ratings. The 'male' coefficient is significant in all cases, meaning that reviews of male physicians are significantly more favorable than those of women.

Lexical Analysis

Here we consider variation in review text as a function of the gender of the physician being reviewed. Similar to the open vocabulary approach of (Schwartz et al. 2013) for examining demographic differences of social media users, we seek to identify words in the corpus that are most strongly associated with male or female physicians.

To this end, we use a log-linear regression model, following the approach used in (Paul et al. 2016). Specifically, we model the (log) frequency of a word being used in a review of a physician of a given sex as a function of (i) a background word intercept (capturing overall word frequency); (ii) a general gender intercept (adjusting for differences in the overall volume of text for doctors of that gender), and finally (iii) gender-specific word coefficients. The latter capture deviations from expected word frequencies correlated with physician gender. With y_{gw} denoting the number of doctors of gender g for which word w was used in a review, our model is defined as:

$$\log y_{gw} = \beta_0 + \beta_g + \beta w + \beta g w \quad (2)$$

The word counts y_{gw} from reviews are counted as indicator values by doctor, such that the word count for a particular doctor is at most 1. This is done so that each doctor's reviews contribute roughly evenly to the word counts for their gender. Otherwise, doctors with many reviews could bias the results. We take the log of y_{gw} so that the linear coefficients of the model represent relative, multiplicative differences rather than absolute differences in frequency. We fix the gender-independent word intercepts β_w to the log-frequency of word w in the entire corpus. The other coefficients are learned by fitting a standard least squares model.

We use two additional data pre-processing parameters to constrain the corpus for this model. First, we removed words that appeared fewer than k times in the corpus, to reduce the effects of noise and word associations that do not have much evidence. Second, we only count words that are within a window of $\pm j$ tokens from each mention of a gendered pronoun in the review. We hypothesize that words occurring near pronoun tokens are more likely to be describing the physician, rather than other issues discussed in the review. For example, reviews often provide background on the patient or their family, and we do not want to include tokens specific to these descriptions in our analysis because they are not being used to describe the physician. After some qualitative experimentation, we observed that using a small window results in more gender-specific words that appear to describe people and fewer words that appear to describe more general aspects of healthcare. We use a window of $j = 5$ for our experiments in this section.

Qualitative Examination

Table 2 shows the words with the 50 highest values of β_{gw} for each gender, which can be interpreted as words with the strongest associations with that gender. We show results for three different values of the word frequency threshold k. Some of the top-ranked terms suggest specialties (*orthopedic*, *knee* for males; *pap*, *gyno* for females), while others seem to be more general character traits (*jerk* for males; *unfriendly* for females). Among character traits, words describing arrogance (*arrogant*, *pompous*, *cocky*, *ego*) are associated with males, while both spellings of *judg(e)mental* are associated with females.

Quantitative Comparison

To quantify the differences in language to describe the two genders, we utilized the sentiment lexicon, VADER (Hutto and Gilbert 2014), which is a lexicon of 7,517 words that are associated with a score of sentiment valence on a scale from -4 (extremely negative) to 4 (extremely positive). For each word w that appears in the VADER lexicon, we multiplied each gender's word coefficient β_{gw} with the VADER sentiment value s_w.

The distributions over these scores for each gender are shown in Figure 2. The mean score ($\beta_{gw} \times s_w$) was -0.14 for males and 0.00 for females (difference significant with $p < .01$), which indicates that words associated with males tend to have more negative sentiment than words associated with females. This contrasts with the result that male doctors receive higher ratings on average, but it may be that

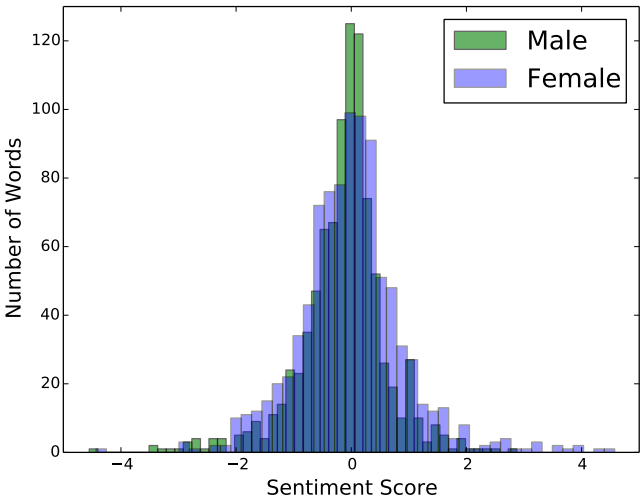


Figure 2: The distribution over sentiment scores for each gender and each word, defined as $\beta_{gw} \times s_w$ where β_{gw} is the association of word w with gender g, and s_w is the sentiment value (positive or negative) given by the sentiment lexicon.

negative reviews of male doctors, though less frequent, use harsher language (e.g., *ass*). This hypothesis is supported by the fact that the distribution over scores has a longer left tail for males: the third moment of the distribution, skewness, is $-.81$ for males (left skewed) compared to 0.57 for females (right skewed).

Another interesting observation is that the sentiment distribution for males has a higher peak near 0 (neutral), while the female distribution is wider (variance of 0.88 for females, compared to 0.60 for males). That is, words associated with female doctors tend to have sentiment values of higher magnitude, both positive and negative. This potentially suggests that reviews of male doctors use more objective language and reviews of female doctors use more opinionated language, irrespective of polarity.

Discussion and Limitations

We have shown that online reviews of physicians tend to be more negative for female than for male doctors, overall. But our lexical analysis tells a more nuanced story. It seems patients tend to use words that have greater variance in terms of sentiment when reviewing female vs. male physicians; reviews of the latter tend to use more neutral language. We believe this warrants additional analysis.

This work has multiple important limitations. First, we have constructed our corpus using a heuristic automated means of inferring physician gender based on pronoun use. While we believe this has resulted in mostly accurate assignments, we cannot be certain of this. Second, we are unable to make any direct causal claims because there are many possible confounders in the data for which we are unable to adjust. Thus we cannot know, e.g., if the observed phenomena are due primarily to physician specialty (e.g., surgeon) or to gender, since the latter likely correlates with the former.

$k = 5$ $(V = 4229)$		$k = 15$ $(V = 2214)$		$k = 25$ $(V = 1581)$	
Male	**Female**	**Male**	**Female**	**Male**	**Female**
guy	colposcopy	guy	shell	guy	shell
jerk	shell	jerk	shes	jerk	shes
prostate	progesterone	prostate	lovely	hes	shed
torn	gynecological	torn	shed	man	woman
hes	ache	hes	physicals	knees	infertility
dads	caution	man	insight	hed	lady
gruff	lessons	knees	woman	fusion	judgemental
ethics	maternity	hed	infertility	knee	pap
screwed	polyp	pompous	welcome	orthopedic	rudely
disk	cope	playing	womens	shouldnt	acne
muscles	confidentiality	fusion	lady	botched	gem
gastric	shes	lying	judgemental	spine	safe
sucked	cookie	cocky	pap	chiropractor	gyn
signed	rescheduled	knee	hormone	injections	enjoy
massive	counselor	fathers	joined	dentists	judgmental
lecture	interactions	orthopedic	assessment	option	gynecologist
pure	lovely	revision	gyno	breath	unfriendly
cat	ectopic	shouldnt	rudely	weve	accepting
man	inexperienced	ego	acne	swelling	emergencies
hips	periods	mris	desired	providing	adore
ass	shed	botched	np	christian	ratings
orthopedist	ideal	chiropractic	gem	walks	annual
creepy	stern	gentleman	safe	wisdom	deeply
insert	accordingly	hospitals	hostile	hurting	refill
fractured	fibroid	bypass	lol	nerve	finding
quoted	tiny	idiot	gyn	mess	finish
ed	carried	lumbar	enjoy	rd	miscarriage
miracles	population	flat	judgmental	absolute	birth
recovered	connected	spine	gynecologist	hero	proactive
fraud	searching	removing	miscarriages	arrogant	therapist
bulging	sonogram	pts	unfriendly	hip	discusses
addicted	absolutley	pointed	tone	management	unsure
income	driving	unsympathetic	relate	retire	signs
numbed	patents	refreshing	smear	boy	stage
joking	accommodate	scheduling	accepting	fixed	favorite
awake	combined	ortho	reach	laid	session
trauma	turning	chiropractor	pcos	brother	plans
united	vbacs	character	closer	repair	handled
knees	dept	loose	nelson	improved	ob
terminology	therapies	injections	star	replacement	highest
someones	obstetrics	dentists	frustrated	pull	stating
slipped	gynecology	hoping	emergencies	shoulder	situations
overcharged	intentionally	implants	adore	deserves	rash
ulcers	healthier	risks	midwife	operated	accepted
mocked	calcium	gi	reschedule	reports	ultrasound
hed	myomectomy	approximately	ratings	local	communication
elbow	kelly	rounds	communicates	hesitate	email
candy	physicals	option	smiling	injuries	cold
contract	ect	entered	researched	attempted	pushes
claiming	thrown	rn	lump	light	pregnancy

Table 2: The top 50 words associated with each gender according to the regression model (i.e., the words with the highest β_{gw} coefficients). Results are shown for different values of k, which is the frequency threshold at which a word must appear in the corpus to be included in the experiment. Higher values of k mean we include only relatively frequent words; lower values result in inclusion of rarer words, which may introduce noise. We also report the vocabulary size V for each k.

References

Fox, S., and Duggan, M. 2013. Health online 2013. *Health* 1–55.

Gao, G. G.; McCullough, J. S.; Agarwal, R.; and Jha, A. K. 2012. A changing landscape of physician quality reporting: analysis of patients online ratings of their physicians over a 5-year period. *Journal of medical Internet research* 14(1):e38.

Grabner-Kräuter, S., and Waiguny, M. K. 2015. Insights into the impact of online physician reviews on patients decision making: randomized experiment. *Journal of medical Internet research* 17(4):e93.

Hall, J. A.; Blanch-Hartigan, D.; and Roter, D. L. 2011. Patients' satisfaction with male versus female physicians: a meta-analysis. *Medical care* 49(7):611–617.

Hutto, C., and Gilbert, E. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text.

Kilaru, A. S.; Meisel, Z. F.; Paciotti, B.; Ha, Y. P.; Smith, R. J.; Ranard, B. L.; and Merchant, R. M. 2016. What do patients say about emergency departments in online reviews? a qualitative study. *BMJ quality & safety* 25(1):14–24.

Li, S.; Feng, B.; Chen, M.; and Bell, R. A. 2015. Physician review websites: effects of the proportion and position of negative reviews on readers willingness to choose the doctor. *Journal of health communication* 20(4):453–461.

López, A.; Detz, A.; Ratanawongsa, N.; and Sarkar, U. 2012. What patients say about their doctors online: a qualitative content analysis. *Journal of general internal medicine* 27(6):685–692.

Paul, M. J.; Chisolm, M. S.; Johnson, M. W.; Vandrey, R. G.; and Dredze, M. 2016. Assessing the validity of online drug forums as a source for estimating demographic and temporal trends in drug use. *Journal of Addiction Medicine* 10(5):324–330.

Paul, M. J.; Wallace, B. C.; and Dredze, M. 2013. What affects patient (dis) satisfaction? analyzing online doctor ratings with a joint topic-sentiment model. In *AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI (HIAI)*.

Roter, D. L., and Hall, J. A. 2015. Women doctors dont get the credit they deserve. *Journal of general internal medicine* 30(3):273.

Schwartz, H. A.; Eichstaedt, J. C.; Kern, M. L.; Dziurzynski, L.; Ramones, S. M.; Agrawal, M.; Shah, A.; Kosinski, M.; Stillwell, D.; Seligman, M. E.; and Ungar, L. H. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one* 8(9):e73791.

Wallace, B. C.; Paul, M. J.; Sarkar, U.; Trikalinos, T. A.; and Dredze, M. 2014. A large-scale quantitative analysis of latent factors and sentiment in online doctor reviews. *Journal of the American Medical Informatics Association* 21(6):1098–1103.