What Do Patients Care About? Mining Fine-grained Patient Concerns from Online Physician Reviews Through Computer-Assisted Multi-level Qualitative Analysis

Lu He[1], Changyang He[2], Yue Wang, Ph.D.[3], Zhaoxian Hu, MS[1], Kai Zheng, Ph.D.[1], Yunan Chen, Ph.D.[1]
[1]University of California, Irvine, Irvine, CA, USA; [2]Hong Kong University of Science and Technology, Hong Kong, China; [3]University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Abstract

Online physician review (OPR) websites have been increasingly used by healthcare consumers to make informed decisions in selecting healthcare providers. However, consumer-generated online reviews are often unstructured and contain plural topics with varying degrees of granularity, making it challenging to analyze using conventional topic modeling techniques. In this paper, we designed a novel natural language processing pipeline incorporating qualitative coding and supervised and unsupervised machine learning. Using this method, we were able to identify not only coarse-grained topics (e.g., relationship, clinic management), but also fine-grained details such as diagnosis, timing and access, and financial concerns. We discuss how healthcare providers could improve their ratings based on consumer feedback. We also reflect on the inherent challenges of analyzing user-generated online data, and how our novel pipeline may inform future work on mining consumer-generated online data.

Introduction

Choosing the right healthcare provider has been a challenge to many patients due to inherent information asymmetry between the two parties. As a result, patients often seek advice from friends and family who had similar conditions and experiences[1,2]. This pressing information need has given rise to online physician rating (OPR) websites, where millions of patients can share experiences by reviewing and evaluating their physicians. It is estimated that popular OPR websites, such as Vitals.com and RateMDs.com, are consulted by at least 30% Internet users in the U.S. and have significant influence on people's choices of healthcare providers[3].

Data from OPR websites (henceforth called "OPR data") cover a variety of information. This includes physician profiles (specialty, experience, accepted insurance, etc.), overall satisfaction ratings (1-5 stars), break-down ratings (along multiple dimensions such as competence, wait time, bedside manner, etc.), and open-ended reviews written by patients. This data provides a unique lens through which many stakeholders can obtain insights. For example, healthcare providers can better understand patient concerns to improve quality of care; health informatics researchers can gain better understanding of consumer's information needs; healthcare consumers such as patients and caregivers can be empowered through better information access; government agencies can design more comprehensive healthcare quality assessment surveys[4].

OPR has been increasingly studied in the research community. Early studies focused on analyzing consumer ratings as these structured data can be easily processed at scale. One type of work cross-checked consumer ratings against professional surveys and clinical performance, and they discovered inconsistent results. Gao et al. found that ratings on RateMDs and measurements from the official state medical board had significant positive correlation with an increasing support from 2005 to 2010[5]. However, Daskivich et al. found that online ratings failed to correlate with objective ratings of specialists' quality made by other physicians, the Primary Care Physician (PCP) survey, and the Administrator Survey[6]. Such inconsistency is likely due to the fact that OPRs are based more on consumers' subjective experience than objective treatment outcomes, and therefore consumer ratings may reflect different aspects of concerns than those in official surveys.

Compared to ratings, free-text reviews in OPR websites are more nuanced and carry richer information about patient concerns. However, the sheer amount of unstructured reviews makes it infeasible to conduct exhaustive analysis. As a compromise, previous researchers take one of two approaches. The first approach samples a relatively small set of reviews from the big OPR data for a focused qualitative analysis. For example, Lopez et al. analyzed 712 reviews from Yelp and RateMD and identified three major themes: *technical competence*, *interpersonal manner*, and *system*

issues[7]. Kilaru et al. used a grounded theory approach to analyze 1,736 reviews of emergency department (ED) care on Yelp and found that similar topics are shared between Yelp reviews and those in official surveys[8]. These studies, while providing deep insights into patient concerns, only covered a small sample of all reviews. To scale up the analysis, the second approach employs machine learning techniques such as statistical topic modeling to extract topics (each topic consisting of a list of keywords) from large-scale consumer reviews. For example, Wallace et al. adopted the three themes identified in Lopez et al. and applied topic modeling on nearly 60,000 reviews from RateMD[9]. A recent analysis discovered three general topics (*hospital-level services*, *communication skills*, and *professional skills*) from a Chinese OPR website[10]. While these studies demonstrate the potential of computer-assisted qualitative analysis[11], the extracted topics were often coarse-grained and provided only the high-level categories of topics without identifying any detailed aspects under each top topic. Indeed, interpreting topics extracted from consumer-generated reviews can be challenging[12], especially when review texts have short lengths, correlated topics, and nested subtopics[13]. To cope with these challenges, researchers often have to label additional documents and words to "guide" topic models[14].

We develop a novel computer-assisted qualitative analysis methodology to discover coarse- and fine-grained patient concerns from large-scale OPR review texts. We first identify coarse-grained topics by qualitatively coding a small set of reviews. Under each topic, we further apply word clustering to discover fine-grained themes, which are surprisingly easier to interpret than topics directly extracted by topic models. This methodology contributes an empirically effective mechanism to synergize human coding efforts and machine learning capability in extracting fine-grained insights from large-scale text.

Using this novel methodology, we analyzed unstructured reviews from a major OPR website. This large-scale analysis reveals key implications on healthcare service improvement. Specifically, we found that patients primarily evaluated relationship-related aspects in their reviews, highlighting the role of patient-provider relationship in patients' perceived quality of healthcare services. We also note that management related issues could be the triggers for patients to leave unfavorable reviews online. The fine-grained patient concerns greatly complement previous OPR research by providing richer and more granular information of patient narratives online.

Material and Methods

Data Description

Vitals is one of the largest OPR websites for healthcare consumers to provide or access evaluations of physicians in the U.S[15]. The site has 127,300 unique daily visits according to Google Trends. The site provides basic information on physicians, such as their locations, gender, and year of experience, etc. Patients are able to score a doctor on a Likert scale of 1 (poor) to 5 (excellent), write a review and selectively make a detailed quality rating across eight dimensions: *Wait Time, Easy Appointments, Promptness, Friendly Staff, Accurate Diagnosis, Bedside Manner, Spends Time with Patients, Appropriate Follow-up*.

In this study, we collected and analyzed 1,065,631 OPRs posted from January 1, 2008 to November 4, 2018 for 102,540 family physicians in the U.S. on Vitals.

Method Pipeline

We employed a multi-level qualitative analysis method pipeline. The basic idea is to take a top-down approach to mining a large-scale review corpus. We first identified coarse-grained, high-level topics, and then identified fine-grained, low-level subtopics (or detailed patient concerns) under each topic. To scale up the analysis to a large corpus, we combined manual coding with machine learning in both stages. For our text analysis, we only included the 1-star and 5-star reviews that have more than 20 words. We chose this subset because they represent the majority of the reviews and are long enough to be informative. In addition, 1-star and 5-star reviews convey direct negative and positive emotions, while the moderate reviews (2,3&4-star) often convey mixed feelings, which is challenging to disentangle.

Mining coarse-grained topics

To identify coarse-grained topics, we conducted qualitative coding on a sample of reviews, and then used supervised machine learning to generalize the codes to all reviews. We did not use topic modeling to automatically discover coarse-grained topics, because algorithms like latent Dirichlet allocation extracted uninterpretable topics with mixed content in pilot experiments. Indeed, these algorithms work well when topics are well separated[13]. However, themes in OPR reviews are often mingled. For example, dissatisfied consumers often simultaneously complain about lack of

clinical competence and bad interpersonal manners. (Semi-)supervised topic modeling is not pragmatic either as it assumes that qualitative analysis has been done in the first place[9].

To ensure that each review contains enough information for qualitative coding, we only considered reviews with at least 20 words. This resulted in a corpus with 207,029 free-text reviews.

(1) Qualitative coding of reviews: We used concepts from a validated patient complaint taxonomy initially proposed by Reader et al to guide our coding[16]. We chose this taxonomy because it was built through a systematic synthesis of patient complaint literature and has been validated and used in many patient satisfaction studies[17]. Two hundred reviews were randomly selected and coded by two annotators separately. The two annotators discussed to resolve disagreements and reached an agreement ratio above 80%. In this annotation stage, the annotators found that the three concepts in the taxonomy (*management*, *clinical*, and *relationship*) captured all the topics in the reviews and no new topics emerged. The two annotators separately coded another 400 reviews, resulting in a set of 600 annotated reviews. In this training set, 59.3% were labeled as including clinical topics, 34.2% management, and 75.5% relationship.

(2) Supervised review classification: We used the 600 annotated reviews as training data to train text classifiers that assign topics to unannotated reviews. A review was represented as a feature vector by taking the average of its word vectors, known as a continuous bag-of-words representation[18]. Words were represented as 100-dimensional vectors trained by the word2vec algorithm on the review corpus. We trained one classifier for each topic, so that each classifier decided whether a review belongs to a topic. This allows a review to have multiple topics. We chose gradient boosted decision trees as the underlying classification model, as it showed higher accuracy than support vector machine or random forest. Two hyperparameters, maximum depth of trees and minimum sum of instance weights in a leaf, were optimized for each classifier. Under 10-fold cross validation, the classifier achieved 84% F1-score on *management*, 86.7% on *clinical*, and 92.5% on *relationship*. These machine predictions are remarkably accurate since they are about the same as human agreement rate.

(3) Estimating word-topic relatedness. We measured the relatedness between a word and a topic as the probability of a word being classified into a topic, according to the corresponding topic classifier.

Mining fine-grained concerns

To identify fine-grained concerns (or *aspects*) under each topic, we ran a clustering algorithm on topic-related words, and then examined and annotated these word clusters. Here we adopted word clustering instead of manual coding as we found empirically that such an algorithm could already discover interpretable aspects. This is likely because latent aspects are almost uncorrelated under the same topic (i.e., conditionally independent[19]) and give rise to distinct word clusters.

For each topic, we clustered 3,000 words (~10% vocabulary size) with the highest word-topic relatedness computed in (3).

(4) Unsupervised word clustering: Given topic-related words under each topic, we applied k-means algorithm over the word vectors (learned in Step 1). Euclidean distance between two vectors is used as the distance measure. These clusters represented candidate aspects under each topic that expressed fine-grained patient concerns. To avoid omitting aspects, we set $k = 20$ clusters for each topic, which is more than twice the number of aspects in previous work[16].

(5) Qualitative coding of word clusters: Two annotators independently examined 10 words closest to each cluster centroid to determine its meaning. Inspired by the divide-and-merge methodology for clustering[20], we manually merged clusters with similar meaning. If two clusters exhibited opposite attitudes towards the same subject matter, they were also merged. Our manual coding was also guided by Reader et al.'s taxonomy[16].

(6) Estimating review-aspect relatedness: We measured the review-aspect relatedness as the reciprocal of cosine distance between the review document vector to the cluster centroid of aspect. Since a review may talk about more than one aspect, we calculated the review-aspect distance for all aspects under that topic and assigned the normalized relatedness to a review instead of assigning the closest aspect to it.

The overall method pipeline is depicted in Fig. 1.

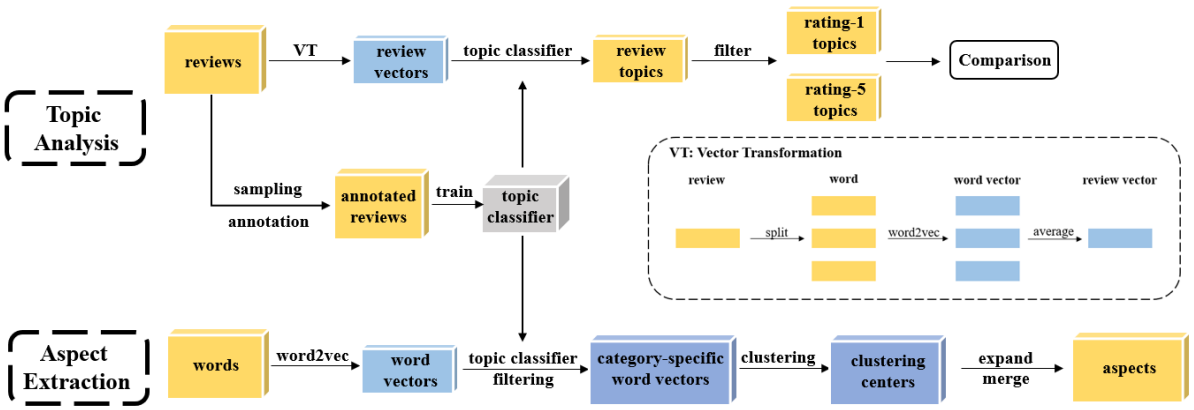

Figure 1. Method Pipeline

Results

This section will first provide an overview of the dataset through consumer rating analysis, and then report the coarse-grained topics and fine-grained aspects identified through our novel computer assisted qualitative coding process.

Consumer Ratings

The rating distribution at the review-level is J-shaped, with 66% being 5-star, 16% 1-star and the rest 18% in the middle, and is consistent with previous findings[15]. At the physician level, the average rating is 4.039 and the standard deviation is 0.926, indicating that physicians tend to receive favorable ratings overall. The average number of reviews a physician received is 10.44 and the standard deviation is 13.2. 24.5% of the physicians only received one or two reviews, suggesting a highly skewed distribution of the number of reviews at the physician level.

While the website allows users to rate physicians on 8 sub-categories listed above, more than half of the reviews did not have any of the 8 categories rated. Among them, *wait time* and *follow-up* have higher unfilled proportions. Moderate reviews (2, 3&4) tend to have more unrated subcategories compared to extreme reviews (1&5). Specifically, 41% of the 1-star reviews have all of the 8 categories unrated, 45% for 5-star reviews, while 72% of the 3-star reviews have all of the 8 categories unrated and 65% for the 4-star reviews.

Coarse-grained Topic Analysis

The reviews were classified using the machine learning model to decide whether they include the three topics: *relationship, clinical, and management*. **Relationship** refers to interaction between patients and physicians. This could include their communication and physicians' empathy toward patients. **Clinical** refers to patients' perceived quality of care. **Management** refers to institutional managerial issues. For example, patients complained about long waiting time and difficulty scheduling appointments.

Among 207,029 1-star and 5-star reviews with at least 20 words, 193,360 (93.4%) were predicted relevant to *relationship*, 146,358 (70.7%) to *clinical*, while only 78,391 (37.9%) were predicted relevant to *management*. This suggests that overall health consumers wrote more about physician-patient relationships and clinical issues than management when evaluating physicians online. Nearly one fifth of the reviews (43,331, 20.9%) were classified to include all the three topics. 126,103 (60.9%) reviews talked about 2 topics, 35,909 (17.3%) mentioned 1 topic, and 1,686 (0.8%) did not belong to any topic. Those reviews that do not include any of the three topics mostly provide general evaluations such as *"His is over all a very good dr. i have been going to him for over 20 years. I have no complaint"*.

Table 1 presents the three topics, words highly related to the topics, selected examples and the proportions. The words have high correlation with the corresponding topics are selected based on word-topic relatedness in Method (3). We replaced real physician names with X to preserve privacy. We kept the misspellings, grammatical errors and capitalization as they appeared in the original dataset.

Table 1. Coarse-grained topics.

Topic	Words	Example	Proportion
Relationship	listening, attentive, respectful, receptive, interrupt, hurry, rush, belittling, empathetic, unconcerned	*Dr. X is one of the nicest Dr's I've met here. He took the time to listen completely without interruption and he explained in a way and could understand.*	93.4%
Clinical	anemia, dangerously, remedy, beneficial, diagnoses, anti-inflammatory, insightful, gallbladder, evaluation, recommendations	*I was initially upset because he wanted to do a lot of workup on my heartburn, but I am glad he did. It turns out it was my heart and not acid reflux. Thank you!*	70.7%
Management	rescheduled, 8am, appointments, follow-ups, billing, insurance, understaffed, chaotic, expired, wednesday	*Once a patient it's becomes increasingly hard to get an appointment or seen in between the "follow-up" visits. It's all about the dollar.*	37.9%

To find the trigger of leaving positive/negative reviews, we made a comparison on the topic distribution of 5-star and 1-star reviews as shown in Fig. 2. Both clinical and relationship related issues appeared slightly more in 5-star reviews than in 1-star reviews. However, management was discussed much less in 5-star reviews as compared to in 1-star reviews. Only around 20% of the 5-star reviews discussed management, while more than 60% of the 1-star reviews discussed management. The proportion of reviews rated as 1-star and 5-star is significantly different across the three topics ($p<0.05$).

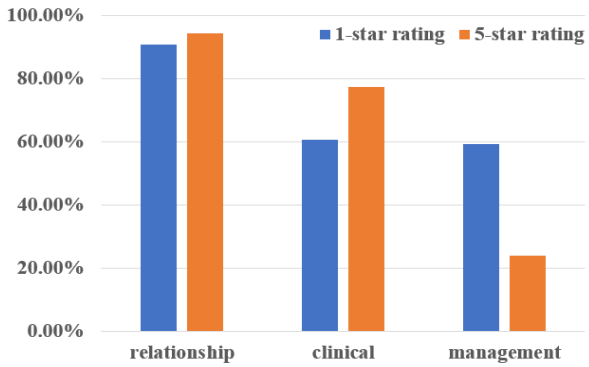

Figure 2. Topic distribution in different ratings

Fine-grained Aspect Analysis

To extract the fine-grained aspects under each topic, we combined unsupervised word vector clustering and qualitative coding. We summarized our findings in Table 2-4. Since a review can include multiple topics and aspects, the review examples we put under one aspect can also be under several other aspects. Note that the sum of aspect proportion under a topic equals 1 because we assigned the normalized relatedness of each aspect to a review.

There are four aspects identified under *relationship*: Patience, Communication, Respect and Compassion as shown in Table 2. **Patience** refers to whether physicians spend time with patients in person. Some patients felt being rushed during clinical encounters and were often ignored or interrupted. **Communication** refers to the quality of patient-provider conversation. Patients commented on whether physicians listened to and addressed their questions. **Respect** refers to whether patients were treated in a respectful manner. For instance, some patients reported that their physicians were arrogant and abrasive. **Compassion** refers to the tenderness, compassion and sympathy toward patients. For example, some patients described their physicians as empathetic and sympathetic.

Our cluster analysis shows that when patients talk about relationship-related aspects, they tend to write using more emotional terms and strong adjectives to express their dissatisfaction or compliment, such as "*He is empathetic, sympathetic and very kind*" in the example for Compassion and "*Very arrogant and patronizing, also quite*

inappropriate and rude at times" in the example for Respect. In addition, though we manually merged the clusters, some of the relationship-related aspects are not exclusive from each other. For instance, in the example for Communication, "*He took time to listen to my concerns and cared about my issues*" also reflects the patience and compassion of the doctor. Besides, we also found when talking about communication issues, patients are more likely to mention whether physicians listen to their concerns instead of whether the doctors express precisely, which echoes the importance of listening in doctor-patient communication as previous work suggested[21].

Table 2. Aspects under Relationship.

Aspect	Keywords	Examples	Proportion
Patience	hurry, rush, examines, forgets, interrupting, cuts, interrupts, intently, dismisses, patiently	*I have the up most respect for Dr. X. She is kind, patient & her appointments are prompt. She answers all your questions & is not hurried. I believe she schedules patients 30 minutes apart. 1 visit with her is like multiple visits with an Urgent care doctor.*	30.0%
Communication	listening, addressing, dismiss, evaluate, brushed, voiced, hears, brush, receptive, express	*I really felt he had an excellent presence and extremely helpful. He took time to listen to my concerns and cared about my issues. I would highly recommend him to family/friends.*	25.0%
Respect	belittling, patronizing, sarcastic, smug, abrasive, unconcerned, unsympathetic, hostile, combative, argumentative	*Very arrogant and patronizing, also quite inappropriate and rude at times. Did not care to look for a resolution to my ailment. After two years of this my last interaction with him made me switch physicians.*	23.2%
Compassion	empathetic, thoughtful, respectful, approachable, considerate, insightful, informative, personable, conscientious, sympathetic	*Dr X truly makes you feel you are his only patient..He is empathetic sympathetic and very kind...Many days we have cried together...God could not created a better human being to be a Dr to administer care for the sick...I am so grateful to be a patient*	21.9%

For the topic *clinical*, five aspects were identified: Treatment, Diagnose, Medication, Personal Conditions and Professional skills, as shown in Table 3. **Treatment** refers to how physicians treat patients' diseases. For instance, patients described the kinds of treatment plans and whether they turned out to be effective. **Diagnose** refers to the assessment and judgements of clinical symptoms. For example, patients described how the physicians diagnosed them and whether they have been misdiagnosed. **Medication** refers to the prescription and administration of medications. Patients listed the names or types of medications that they were prescribed such as anti-depressant and anti-inflammatory. **Personal conditions** refer to patients' personal health conditions, medical history and symptoms. **Professional skills** refer to physicians' overall clinical competence. Patients generally used adjectives to describe their perceptions of the clinical competence of physicians. For example, they may describe a physician as "meticulous", "well-informed" or "astute". We observed that in clinical-related OPRs, the five aspects tend to be discussed collectively. For example, the following review, *"39 year old male here. I have been dealing with occasional hip pain on and off for years. Dr. X did a physical exam and X-rays. I was diagnosed with bursitis and tendinitis. Some anti inflammatory meds were prescribed which worked. This was good news since I really didn't want to pay for an mri or have surgery. I realize that not everyone may not be so lucky with their diagnosis. He spent a lot of time with me and yet I still feel like I was in and out. His staff was kind and courteous. I rarely write reviews but my experience was just too good to not mention"*, first describes the whole procedure from providing personal medical history (occasional hip pain), being diagnosed (bursitis and tendinitis), and to being prescribed medications (anti-inflammatory). At the end, the review makes an evaluation of the doctor's overall professional skills based on the previous procedures

Table 3. Aspects under Clinical.

Aspect	Keywords	Examples	Proportion
Treatment	possibilities, protocol, appropriately, prognosis, method, symptoms, remedy, pharmaceuticals, determining, effectively	*Dr. X diagnosed and effectively treated a very burdensome problem that many previous physicians could not help with me with*	23.8%
Diagnose	diagnoses, conclusions, direction, prognosis, recommendations, findings, possibilities, assessments, evaluation, judgements	*Dr X did not listen to our needs. She was very full of herself. misdiagnosed sinus infection as a virus. Had to go to another doctor to get treated.*	22.2%
Medication	anti-inflammatory, prednisone, inflammatory, anti-depressant, zoloft, depressants, toxic, temporary, topical, statins	*I went to her throughout my pregnancy. She recommended antidepressants such as Zoloft which cause birth defects. She had no idea of what she was doing. Even the nurses that worked with her told me that I should switch doctors.*	19.2%
Personal conditions	pulmonary, ovarian, gallbladder, colon, cancerous, artery, lymph, cervical, fluid, blockage	*Definetely i do not recommend this dr. to nobody, I had my gallbladder removed last year and this surgery went bad. I had unexpected life threatening complications. She never took the time to figure out what she did wrong in the surgery. Result of this procedure i was admitted to hospital 5 and a half months . weeks of being intubated. i also have permanently health impairments.*	17.6%
Professional skills	diligent, insightful, intuitive, keen, astute, meticulous, forthright, realistic, well-informed, precise	*Dr. X is thorough, insightful, kind and accurate. He quickly diagnosed my case and proposed a plan and solutions. I wish he were available as a primary care doctor--he is a top flight emergency physician!*	17.2%

We identified five aspects that fall under *management*: Timing and access, Bureaucracy, Finance and billing, Service issues and Staff and resources, as shown in Table 4. **Timing and access** refer to timely and easy access to healthcare services. For example, patients commented on their waiting time to be seen by doctors, and ease of scheduling and rescheduling appointments. **Bureaucracy** refers to the administrative policies and procedures during patients' interaction with the healthcare organization. For instance, it may involve having a prescription verified and getting a signature or authorization from the office. **Finance and billing** refer to the financial components of healthcare services such as insurance, billing and payment. For example, users shared their experience of being overcharged or having difficulty in their billing processes. **Service issues** refer to hospital services that support patients in their encounters. These include follow-ups and resolving issues. For example, a patient wrote that the billing code was entered incorrectly, and no one has followed up and resolved this problem. **Staff and resources** refer to whether the healthcare organization has adequate and well-trained staff and appropriate resources. Among the five aspects, timing and access, bureaucracy and finance and billing are mentioned most. We also noticed that management-related OPRs are significantly longer than relationship-related OPRs and clinical-related OPRs, which could be attributed to a more detailed description when talking about aspects under management.

Table 4. Aspects under Management.

Aspect	Keywords	Examples	Proportion
Timing and access	noon, wednesday, tuesday, thursday, rescheduled, 8am, app, 10:30, notified, reminder	*Once a patient it's becomes increasingly hard to get an appointment or seen in between the "follow-up" visits*	22.1%

Aspect	Keywords	Examples	Proportion
Finance and billing	charging, co-pays, cards, owed, 250, fees, payments, refund, agency, deductible	*HE IS EXCELLECT, JUST VERY UNAWARE THAT HIS STAFF IS CHARGING FULL ENGORGED OFFICE PRICES FOR CASH PAYMENTS, DESPITE INSURE COMPANIES ONLY PAY ABOUT A THIRD AND ITS ACCEPTABLE FOR THE INSURED!!!!*	21.2%
Service issues	processes, informs, speed, follow-ups, monitors, consultations, adjusts, receptive, conflicting, resolution	*Dr. X is a caring and problem solving doc. she always support and provides her best consultations at par. she and her nurse practitioner provides support even if we had left them a message and they phoned us back providing the refer and consultations.*	17.5%
Staff and resources	inefficient, unwelcoming, sloppy, untrained, understaffed, uncooperative, inattentive, clerical, chaotic, staff's	*Dr. X is professional, engaging and pleasant. The receptionists and other low-level staff are, however, quite unprofessional. They all need training on how they handle people and how to conduct themselves in an office or she will lose patients based purely on her staff's behavior!*	17.3%

Fig. 3 showed the distribution of different aspects mentioned in 1-star and 5-star reviews. Overall, a physician's patience, compassion, professional skills, accuate diagnosis, effective treatment and good services are appreciated by patients in positive (5-star) reviews. In negative (1-star) reviews, patients often refer to their personal conditions and medication to contextualize their complaints, especially on lack of respect and bureaucratic processes.

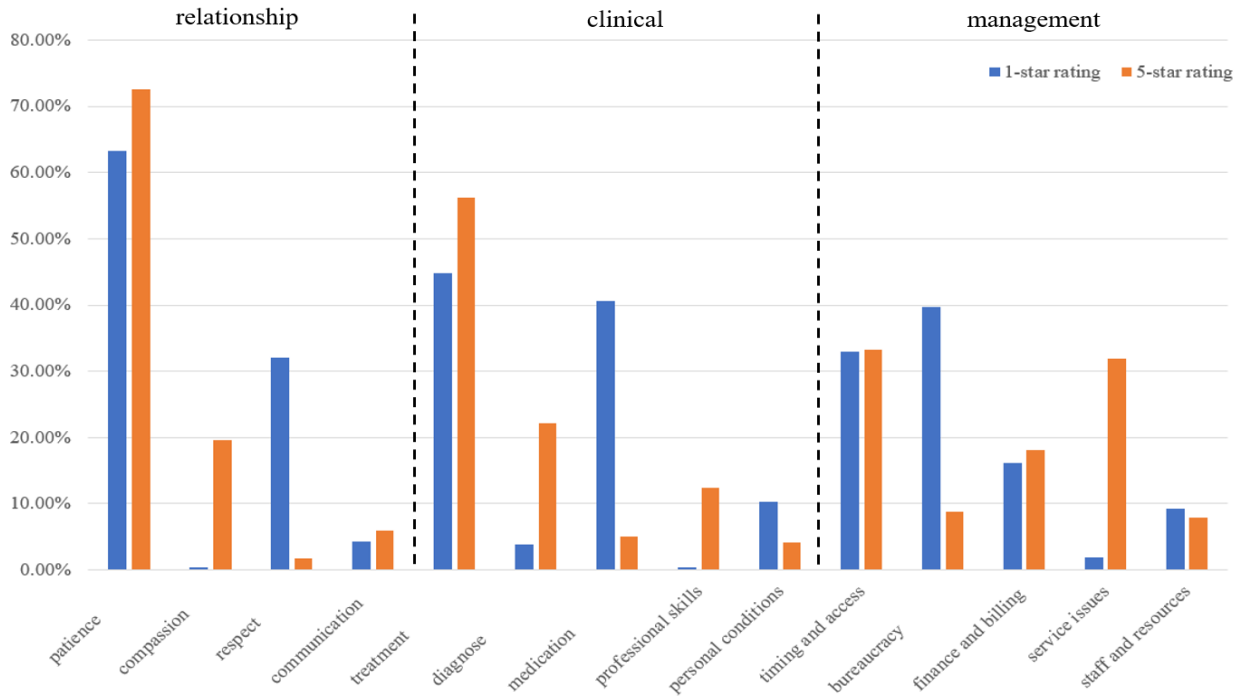

Figure 3. Aspect distribution in different topics and ratings

Discussion

In this paper, we developed a novel computer-assisted qualitative coding methodology to mine coarse-grained topics and fine-grained aspects from consumer-generated OPR data. Through manual coding and supervised machine

learning, we extracted three major topics from OPRs: *management, clinical*, and *relationship*. Through unsupervised word vector clustering and qualitative coding, we further identified fine-grained aspects such as *timing and access*, *diagnosis*, and *communication*. We compared their proportions in the reviews to further understand patients' concerns with healthcare services.

A general methodology for fine-grained analysis of consumer-generated texts. Free-text patient reviews are often mixtures of factual topics intertwined with personal feelings across multiple dimensions and granularities. To fully uncover the fine-grained semantics from texts, it is unrealistic to solely rely on unsupervised algorithms such as topic modeling or their semi-supervised variants that only take one round of human input. Instead, an interleaving of human coding and machine learning is essential to achieve nuanced understanding of these texts. This work introduces a novel analysis methodology that takes a divide-and-conquer approach: it first divides the content into coarse-grained topics, and then zooms in on each topic to locate fine-grained concerns. Human coding is amplified through supervised learning in the first stage and aided by unsupervised learning in the second stage. Together, the methodology effectively interleaves a small but essential amount of human effort with the large-scale processing capability of machine learning in a qualitative analysis task. This general methodology can be useful in a variety of scenarios where fine-grained analysis of consumer-generated texts is needed.

Implications for healthcare service quality improvement. At the coarse-grained topic level, we found that relationship was discussed in 93.4% of the reviews, suggesting that patient-provider relationship is of high-priority for patients. In addition, we found that users discussed management-related topics much more often in 1-star reviews than in 5-star reviews. A hypothesis to explain this phenomenon is that poor management would greatly affect patients' experience with healthcare service, while good management is less noticeable and thus not frequently mentioned in favorable reviews. This finding echoes with previous work which suggests that "[...] 80-94 percent of the damage done by poor service quality is traceable to managerial actions or the system set up by management"[22]. Therefore, though management is not directly related to clinical performance, it could be the triggers for healthcare consumers to leave unfavorable reviews online. These findings also suggest that the inconsistency between online physician ratings and objective clinical performance could be in part due to the fact that they are evaluating very different aspects. Healthcare providers and government agencies should consider better ways of measuring healthcare consumers' satisfaction with their services by gaining insights from consumer-generated online data and including more non-clinical related aspects. Through unsupervised word vector clustering and manual coding, we were able to identify fine-grained aspects that greatly complement OPR literature by providing a granular and richer description of healthcare consumers' narratives on OPR websites, which further shed light on more substantial solutions to improve healthcare service quality. We found that consumer-generated OPR data encompass a wide range of healthcare service aspects, including *timing and access, finance and billing, diagnoses, medication,* and *communication,* etc. In management related reviews, *timing and access, bureaucracy* and *finance and billing* were mentioned more often than *staff and resources* and *service issues*. This indicates that healthcare consumers discussed more about whether they had timely and easy access to healthcare services and whether their interaction with the healthcare organization was smooth.

Limitations and future work. First, we only studied one OPR website and the findings may not generalize to other OPR websites with different designs or target users. Second, we only included family physicians in this study. Patients may value different aspects of family physicians compared to other specialists such as surgeons and dentists. We plan to conduct cross-platform and cross-specialty comparisons in our future work.

Conclusion

We developed a novel computer-assisted qualitative coding method to mine multi-level patient concerns from a large-scale heterogeneous OPR corpus. We identified coarse-grained topics (*management, clinical, relationship*) as well as fine-grained aspects (*e.g., bureaucracy, diagnosis, communication*) which provide more granular and richer information of patients' evaluation of healthcare quality online. Our results complement previous OPR research by contributing the multi-level patient concerns and the novel method for mining large-scale heterogeneous consumer-generated texts.

Acknowledgements. We thank Dr. Xinning Gui, Ruining Tang, Aiden Desai and Chenxi Yang for their contribution at the early stage of this project.

References

1. Harris KM. How Do Patients Choose Physicians? Evidence from a National Survey of Enrollees in Employment-Related Health Plans. *Health Services Research*. 2003;38(2):711-732. doi:10.1111/1475-6773.00141

2. Hoerger TJ, Howard LZ. Search Behavior and Choice of Physician in the Market for Prenatal Care. *Medical Care*. 1995;33(4):332–349.
3. Pew Research Center. The Internet and Health. Pew Research Center: Internet, Science & Tech. https://www.pewresearch.org/internet/2013/02/12/the-internet-and-health/. Published February 12, 2013. Accessed January 9, 2020.
4. CAHPS Patient Experience Surveys and Guidance. http://www.ahrq.gov/cahps/surveys-guidance/index.html. Accessed February 12, 2020.
5. Gao GG, McCullough JS, Agarwal R, Jha AK. 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*. 2012;14(1):e38. doi:10.2196/jmir.2003
6. Daskivich TJ, Houman J, Fuller G, Black JT, Kim HL, Spiegel B. Online physician ratings fail to predict actual performance on measures of quality, value, and peer review. *J Am Med Inform Assoc*. 2018;25(4):401-407. doi:10.1093/jamia/ocx083
7. López A, Detz A, Ratanawongsa N, Sarkar U. What patients say about their doctors online: a qualitative content analysis. *J Gen Intern Med*. 2012;27(6):685-692. doi:10.1007/s11606-011-1958-4
8. Kilaru AS, Meisel ZF, Paciotti B, et al. What do patients say about emergency departments in online reviews? A qualitative study. *BMJ Qual Saf*. 2016;25(1):14-24. doi:10.1136/bmjqs-2015-004035
9. Wallace BC, Paul MJ, Sarkar U, Trikalinos TA, Dredze M. A large-scale quantitative analysis of latent factors and sentiment in online doctor reviews. *J Am Med Inform Assoc*. 2014;21(6):1098-1103. doi:10.1136/amiajnl-2014-002711
10. Pang PC-I, Liu L. Why Do Consumers Review Doctors Online? Topic Modeling Analysis of Positive and Negative Reviews on an Online Health Community in China. In: ; 2020. doi:10.24251/HICSS.2020.087
11. Chuang J, Wilkerson JD, Weiss R, et al. Computer-Assisted Content Analysis: Topic Models for Exploring Multiple Subjective Interpretations. *Advances in Neural Information Processing Systems workshop on human-propelled machine learning*.:9.
12. Chang J, Gerrish S, Wang C, Boyd-graber JL, Blei DM. Reading Tea Leaves: How Humans Interpret Topic Models. In: Bengio Y, Schuurmans D, Lafferty JD, Williams CKI, Culotta A, eds. *Advances in Neural Information Processing Systems 22*. Curran Associates, Inc.; 2009:288–296. Accessed February 4, 2020.
13. Tang J, Meng Z, Nguyen X, Mei Q, Zhang M. Understanding the Limiting Factors of Topic Modeling via Posterior Contraction Analysis. *ICML14'*.:9.
14. Paul MJ, Wallace BC, Dredze M. What Affects Patient (Dis)satisfaction? Analyzing Online Doctor Ratings with a Joint Topic-Sentiment Model. In: *AAAI 2013*. ; 2013.
15. Kadry B, Chu LF, Kadry B, Gammas D, Macario A. Analysis of 4999 Online Physician Ratings Indicates That Most Patients Give Physicians a Favorable Rating. *J Med Internet Res*. 2011;13(4):e95. doi:10.2196/jmir.1960
16. Reader TW, Gillespie A, Roberts J. Patient complaints in healthcare systems: a systematic review and coding taxonomy. *BMJ Qual Saf*. 2014;23(8):678-689. doi:10.1136/bmjqs-2013-002437
17. Harrison R, Walton M, Healy J, Smith-Merry J, Hobbs C. Patient complaints about hospital services: applying a complaint taxonomy to analyse and respond to complaints. *Int J Qual Health Care*. 2016;28(2):240-245. doi:10.1093/intqhc/mzw003
18. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed Representations of Words and Phrases and their Compositionality. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, eds. *Advances in Neural Information Processing Systems 26*. Curran Associates, Inc.; 2013:3111–3119. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.
19. Dawid AP. Conditional Independence in Statistical Theory. *Journal of the Royal Statistical Society Series B (Methodological)*. 1979;41(1):1-31.
20. Cheng D, Kannan R, Vempala S, Wang G. A Divide-and-Merge Methodology for Clustering. *PODS'05*.:26. doi:https://doi.org/10.1145/1189769.1189779
21. Jagosh J, Donald Boudreau J, Steinert Y, MacDonald ME, Ingram L. The importance of physician listening from the patients' perspective: Enhancing diagnosis, healing, and the doctor–patient relationship. *Patient Education and Counseling*. 2011;85(3):369-374. doi:10.1016/j.pec.2011.01.028
22. Ford RC, Bach SA, Fottler MD. Methods of Measuring Patient Satisfaction in Health Care Organizations. *Health Care Management Review*. 1997;22(2):74–89.