# ABSTRACT

Title of Document: ONLINE DOCTOR REVIEWS: ESSAYS ON THEIR ECONOMIC IMPLICATIONS, INCIDENCE OF FRAUD, AND MOTIVATION OF REVIEWERS

Aishwarya Deep Shukla

Directed By: Ritu Agarwal, Professor,
Decision, Operations and Information Technologies

Guodong (Gordon) Gao, Associate Professor,
Decision, Operations and Information Technologies

The ubiquity of WOM in the business world underscores how instrumental it has been as a consumer engagement lever. Therefore, it is imperative for business to understand the consequential role of WOM in influencing consumer behavior. There is also a great need to improve the quality and quantity of online reviews. I address three overarching questions in my dissertation: (1) What is the effect of WOM on consumer decision making; (2) How to detect fake reviews using the review text; and (3) How to encourage reviewers to reveal their identity and give higher quality reviews.

In the first study, I estimate the causal effect of online WOM on consumer demand and uncover its mechanism in affecting the consumer decision making process. I utilize a natural experiment to examine the causal effect of online WOM on consumer demand. The setup allows me to gain granular understanding of how WOM affects the consideration set size and session

duration. In addition, the availability of provider location in the dataset allows me to estimate the impact of online WOM on the consumer's willingness to travel. In the second study, I detect fake online reviews. To identify fake reviews, I use an incidental honeypot that attracts fraudulent behavior by opening low-cost channels for fraudsters. This allows me to build a large training dataset for the machine learning classifier. Finally, in my third study, I explore how email message framing and assurance of user privacy affect the response rate and response quality of online WOM. I conduct a field experiment to uncover how the propensity of a user to give feedback for their doctor can be influenced by a motivational message and how privacy assurance affects identity revelation. Collectively, these studies advance our knowledge on the antecedents and consequences of online reviews, which helps business and society to better utilize WOM for greater value creation for consumers.

# ONLINE DOCTOR REVIEWS: ESSAYS ON THEIR ECONOMIC IMPLICATIONS, INCIDENCE OF FRAUD, AND MOTIVATION OF REVIEWERS

by

Aishwarya Deep Shukla

Dissertation submitted to the Faculty of Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2019

Advisory Committee:
Professor Ritu Agarwal (Co-Chair)
Professor Guodong (Gordon) Gao (Co-Chair)
Professor Anandasivam Gopal
Professor Kunpeng Zhang
Professor Ping Wang (Dean's representative)

Dedicated to the memory
of my grandparents

**ACKNOWLEDGEMENTS**

I have worked long and hard for the past six years to reach where I am today. Yet none of it would have been possible or mattered if not for the unending support from advisors, collaborators, family, friends, and colleagues.

I would first like to thank my co-advisors, Dr. Ritu Agarwal and Dr. Guodong (Gordon) Gao. Professor Agarwal's dedication and work ethic are extraordinary. She tirelessly trained me, constantly pushed me beyond my comfort zone, and went out of her way to cultivate my growth as a scholar. I joined the PhD program at the University of Maryland with a curiosity about Healthcare Information Systems and she helped me to develop this curiosity into scholarly expertise. I would look forward to her advice as I progress in my academic career.

Professor Gao was always supportive of my questions and ideas – with him by my side, I was able to freely explore and experiment. He spent countless hours – often sacrificing his weekends – to inspire me and help me to refine my work. He was with me every step of the way and made it possible for me to reach this point. I am grateful for our friendship and I look forward to continued research together.

Next, I would like to thank my dissertation committee, comprised of my co-advisors and Dr. Anandasivam Gopal, Dr. Kunpeng Zhang and Dr. Ping Wang. Dr. Anandasivam Gopal, thank you for your encouragement and advice. I have thoroughly enjoyed our conversations and your enthusiasm and insightful suggestions. Dr. Kunpeng Zhang, thank you for insights on machine learning and deep learning methods, and helping me to improve this dissertation. Dr. Ping Wang, thank you for our interaction and insightful questions during my PhD Proposal defense. I thank each of you for taking the time to be part of my dissertation committee.

I would like to offer a special thanks to Abhinav Lal, CTO of Practo.com, for offering to collaborate with the Center of Health Information and Decision Sciences (CHIDS) at the University of Maryland and for making it possible for me to access data and run field experiments.

I would also like to thank the University of Maryland PhD Program for seeing my potential, accepting me into the program and supporting me financially to make this possible. A special thanks to Justina Blanco for her efforts and encouragement.

I would like to thank my seniors, peers and junior PhD students in the Information Systems Department. Thank you to my seniors Dongwon Lee, Vickee Wolcott and Yang Pan for helping me to explore my ideas and for providing support and feedback. Thank you to Che-Wei Liu for being a sounding board for my ideas and helping me to present them better. Thank you to Weiguang Wang, Lanfei Shi, Raveesh Mayya and Sabari Ranjan for encouraging and supporting me.

Thank you to my roommates at College Park – Gokul Chandramouli Iyer and Debdatta Sinha Roy – for supporting me and being with me through the best and worst days of the past six years. I value your friendship immensely.

I would like to thank my colleagues JM Goh and Nilesh Saraf at Simon Fraser University for giving me the opportunity to join their prestigious faculty, welcoming me with warmth and offering me continual support and encouragement.

Finally, I would like to thank my family for their unconditional love and support. I would like to thank my mother for inspiring me to be an academic and supporting me in every way possible. I am grateful to my wife, Sakshi for supporting me immensely during the final year of my PhD.

# TABLE OF CONTENTS

## TABLE OF TABLES

## TABLE OF FIGURES

# INTRODUCTION & OVERVIEW

As the healthcare services industry evolves to serve an increasingly digital world, more and more consumers are consulting online reviews to support their decision-making process when choosing a health service provider. In empirical studies, researchers have found online reviews to be an influential factor in driving sales (Chevalier and Mayzlin 2006, Godes and Mayzlin 2004). Yet despite its influence, limited research has been conducted to determine how online WOM impacts the consumer decision-making process. My research aims to fill this gap.

In this dissertation, I focus on online reviews in the context of healthcare services and ask several pertinent questions, specifically: (1) What is the effect of online word-of-mouth (WOM) on consumer decision making; (2) What are the challenges in detecting fraudulent reviews and how can deep learning address them; and (3) How can platforms motivate users to improve the quantity and quality of online reviews, and how can they increase identity disclosure? These three questions align with the three chapters of my dissertation, which together provide a nuanced overview of the antecedents and consequences of online WOM. The findings are particularly relevant to platforms and service providers that seek to maintain a steady stream of online reviews.

The first chapter utilizes detailed clickstream data on online WOM to uncover mechanisms underlying its influence on consumer decision making. I leverage a feature launch on a major doctor appointment booking platform to study the impacts of online WOM on three dimensions of a consumer's choice process: the consideration set size, the time taken to consider alternatives (web session duration), and the geographic dispersion among the choices considered. I further investigate the effects of online WOM on demand.

This study offers several novel findings. First, the impacts of WOM on consumer decision-making processes are not monotonic, but rather are contingent on the abundance of WOM (number of rated doctors) in a market. When WOM is sparse, the lack of availability of WOM makes patients consider more doctors, browse longer, and consider doctors that are geographically more dispersed. In contrast, when WOM is abundant, the presence of WOM helps patients to filter choices thereby considering fewer doctors, browsing for a shorter duration, and focusing on doctors that are closer to home. Finally, the results show that the presence of WOM can have a cannibalization effect: when doctor ratings are published online, the highly rated doctors reap the benefits (in the form of increased demand) at the expense of unrated doctors. This chapter contributes to the extant literature on online WOM by providing new and highly granular insights into the means by which WOM influences consumer decision making.

My second chapter focuses on fake online reviews, which unsurprisingly are becoming more prevalent and are a significant concern for consumer protection groups and regulatory authorities. Research has shown that up to 17% of all online reviews are fraudulent (Luca and Zervas 2016), which is worrisome given that online reviews have been shown to have a powerful effect on consumer decision making (AMA and WOMMA 2014). It is crucial that we learn how to identify fake reviews, yet this remains a challenge. In this chapter, I examine the two broad categories of characteristics that algorithms utilize to identify fake reviews: behavioral and linguistic (Jindal and Liu 2007a, Mukherjee et al. 2011, Ott et al. 2011). The findings from existing research show that while algorithms that rely on behavioral characteristics are widely used to identify deviations from normal behavior, they can be easily duped by skilled fraudsters. Meanwhile, algorithms relying on linguistic characteristics can be highly effective in identifying deceptive language patterns and are theoretically resilient to gaming. But despite the potential of sophisticated linguistic traits to help us detect fraud, the algorithms that use these traits have not yet demonstrated high accuracy on large real-world datasets (Fei 2013, Ott et al. 2011).

This chapter seeks to fill this gap by applying these linguistic techniques for review detection, utilizing a deep learning approach to analyze linguistic traits of online reviews obtained from India’s largest doctor appointment booking platform. I design a deep learning approach to capture the linguistic traits that differentiate between genuine and fake reviews. The deep learning model is evaluated on a dataset of 181,951 doctor reviews, 8% of which are fake. Since a natural honeypot existed at one point on the platform that hosted these reviews, I am able to label the reviews that exploited the natural honeypot as fraudulent, thus overcoming the major challenge in constructing the ground truth for training the model. The model shows a significant improvement in accuracy when compared to traditional machine learning algorithms such as logistic regression and random forest. Interestingly, I also find that human evaluators perform much worse than machine learning approaches. Compared to 200 human evaluators, the deep learning approach has a true positive rate (14.29% vs. 8.70%) that is twice as high, and it also achieves a much lower false positive rate (0.63% vs. 11.68%). I also observe that these evaluators are susceptible to human bias, as they are more likely to label fake reviews as genuine than they are to label genuine reviews as genuine. This study offers further explanations for the advantages of deep learning and is the first to construct a deep learning model to detect fraudulent online reviews, an approach that can help curb fake reviews and increase information quality and market efficiency. By applying deep neural networks to this challenge, my work substantially improves on the performance of both traditional machine learning algorithms and human evaluation.

Considering the relatively high economic importance of reviews, it is crucial to generate high-quality reviews and to encourage reviewers to share their identity in order to maximize their review's (and thus the platform's) credibility. In the final chapter of my dissertation, I examine this problem through the lens of motivational framing and the effects of privacy assurance. Extant literature has shown the importance of persuading consumers to contribute to online WOM (Hong et al. 2016), although there is no consensus on the best way to do so. In an effort to determine the most effective method to encourage users to contribute, I conduct a field experiment in which 2.4 million users are sent email messages designed to motivate them to review their doctor.

This chapter's surprising finding is that motivational framing backfires: incorporating such messages into the review request actually reduces the likelihood that a patient will respond. This suggests that intrinsic motivation is diminished by such motivational framing. This chapter also yields a useful finding on the effect of increasing privacy assurance on identity disclosure and review quality. The field experiment shows that consumers are more likely to disclose their identity when given privacy assurance through the up-front option of remaining anonymous, indicating that this inspires user trust in the platform.

The work in this dissertation has some key implications for practice. First, I find that variation in WOM abundance is linked to consideration set size and consideration time, which suggests that platforms can fine-tune these factors in order to engage their users differently (i.e., increase the proportion of products displayed with WOM in search results in order to prompt quicker purchase decisions). Second, the deep learning approach for fake review detection shows great promise and is worthy of exploration—both by researchers and by platforms that host online WOM. Finally, the observed backfiring of motivational messaging suggests that platforms should be cautious in their assumptions about user motivation, which may be complex and various. As online word of mouth becomes an increasingly mainstream and influential factor in individuals' healthcare decisions, the essays presented in this dissertation serve as a useful overview of its impact, challenges, and elicitation.

# CHAPTER 1: HOW DIGITAL WORD-OF-MOUTH AFFECTS CONSUMER DECISION MAKING: EVIDENCE FROM DOCTOR APPOINTMENT BOOKING

## 1.1 Introduction

Online word-of-mouth (WOM) in the form of recommendations, ratings, and reviews is widely prevalent across diverse product and service categories today, and it has substantial effects on economic outcomes. Acknowledging the strategic importance of WOM, previous empirical studies have examined its influence on the sales of a range of products such as books, movies, and beverages. While these studies generally conclude that positive online WOM increases[1] sales, the primary focus has been on assessing the impact of WOM and not on exploring its influence on the underlying consumer decision-making process (Gu et al. 2012). Understanding this influence is important because the final purchase of a product or service is the culmination of a complex decision-making process that entails selecting and considering available alternatives (Horowitz and Louviere 1995, Liu et al. 2014, Moe 2003, Wright and Barbour 1977). However, since most research relies on displayed reviews and aggregated sales, the underlying mechanisms by which WOM impacts consumer decision-making processes are seldom explored.

Our work aims to address the above gap. Using detailed data from the largest doctor appointment booking platform in India, we examine how WOM influences three interrelated

---

[1] Examples of related research include Chevalier and Mayzlin (2006), Clemons et al. (2006), Duan et al. (2008), Forman et al. (2008), Godes and Mayzlin (2004), Gu et al. (2012), Li and Hitt (2008b), Sun (2012), Zhu and Zhang (2010).

aspects of the consumer's choice process: the size of the consideration set, the time taken to consider the elements in this set, and the geographic dispersion of the choices considered. In addition to estimating the overall effect of WOM on these consumer choice process variables, we further investigate the heterogeneity in the effects of WOM conditional on the abundance of WOM in a particular market. For stronger causal inference, we leverage a major feature launch in which the platform suddenly displayed the accumulated online WOM data that was obtained from its users over time.

Our research setting of healthcare services also provides a novel context for the study of online WOM on consumer decision making. Choosing a doctor involves much higher stakes than choosing books, movies, beverages, games, or cameras, all products that have been the focus of existing studies. The literature on online WOM in healthcare is relatively nascent (Gao et al. 2015, Lu and Rui 2017, Xu et al. 2016). Our study builds on this literature and provides new empirical evidence in this important domain.

This is the first study in the WOM literature to deploy clickstream data, which affords us better insights into the mechanism by which online WOM affects the consumer choice process and the final purchase decision. Methodologically, we are able to create well-defined markets, enabling us to construct novel WOM measures based on relative ranking and to examine the differential impact of WOM at the market level. We find that the effects of online WOM are non-monotonic and contingent on the abundance of WOM. We observe that when online WOM information is published for a few healthcare providers in a market, it results in an increase in consumers' consideration set size and consideration time, and the choices considered are geographically more dispersed. In contrast, when WOM is available for many healthcare providers in a market, measures of all of the above elements decline. Furthermore, we find that online WOM increases the demand for the doctors who are rated, but this increase is concentrated among those with the highest WOM ranking in a market and comes at a cost to the unrated doctors. Our findings yield important implications for online platforms that offer referral and matching features, as well as for providers of healthcare services.

## 1.2 Review of Literature and Research Questions

### 1.2.1 How Online WOM Affects Consumer Decision-Making Processes

Over 82% of American adults say that they read online reviews when making new purchase decisions (Pew Research Center 2016). Given the prevalence and influence of online WOM, it is important for researchers, businesses, and policy makers to understand how it affects purchase decisions. As summarized in Table A1 in the appendix, a stream of literature has investigated the impact of online WOM. These studies show that a strong and significant relationship exists between online WOM and sales in multiple product categories—ranging from craft beers to books, movies, and games.

The literature predominantly infers the impact of WOM on consumer decision processes via aggregated sales data. As a result, the mechanisms underlying the influence of WOM remain understudied. In the five-stage consumer decision-making process introduced by Engel et al. (1968), the evaluation of alternatives is the most easily influenceable stage in the pre-purchase consumer decision-making process. Therefore, we focus on understanding WOM's influence on the assessment of alternatives in greater depth in this study. We shed light on three facets of this choice process: the number of products considered, the time taken for consideration, and the geographic dispersion among the choices considered.

#### 1.2.1.1 Influence of WOM on Consideration Set Size

Consumer decision making is a complex, multistage process. Consumers must collect information about various alternatives and then face difficult trade-offs (Bettman et al. 1991). In this process, the consumer first identifies a set of alternatives—called the consideration set—and then evaluates products in the consideration set to make the final selection. Thus, the inclusion of a product as part of the consideration set is a necessary pre-condition for it to potentially become the final choice. As might be expected, unraveling how the consideration set is constructed has been a principal concern among researchers in business, and especially so in the study of consumer behavior. More recently, economic modeling work has also elucidated how consumers consistently consider only a small subset of available options (Lleras et al. 2017).

In the present study, we first seek to understand how WOM information affects a consumer's consideration set. Prior research offers plausible competing alternatives for the nature of this effect. One argument derives from the classic satisficing model of choice (Simon 1956), in which decision makers with bounded rationality do not seek to maximize their utility. Rather, when consumers "satisfice," they opt for an incomplete consideration of choices and search only until an acceptable level of utility is achieved. This reasoning is widely accepted and has been experimentally tested in the economics literature (Caplin et al. 2011). Because online WOM makes it easier to screen doctor quality, a satisficing level of utility may be reached even when considering fewer options; thus, it is conceivable that the availability of online WOM could lead to a smaller consideration set.

Alternatively, there are plausible mechanisms for WOM to increase the size of a consideration set. WOM may transform many intangible experience attributes into quantifiable search attributes (Hong et al. 2012) and enable consumers to learn from others' experiences (Liu et al. 2014). This reduction in the cost of evaluation implies that the user can search more extensively and evaluate a larger number of alternatives more efficiently (Bakos 1997). Consistent with this mechanism, in experiments with "recommendation signage" (e.g., "Top rated" or "Best seller") Goodman et al. (2013) find that, rather than reducing the choice set and simplifying decisions, such signals result in consumers expanding their consideration sets.

These opposing mechanisms—one suggesting that online WOM can *increase* and the other that it can *decrease* the number of alternatives a consumer considers—underscores the need for empirical examination. Few prior studies have examined how WOM affects the size of the consideration set. The closest research is Dorner et al. (2013), which uses a laboratory experiment in the context of shopping for digital cameras to show that user-generated reviews reduce consideration set size. Two other related studies, also in a laboratory experiment setting, highlight the importance of WOM on the consumer decision-making process, but they do not examine the change in consideration set size. We extend the above literature by explicitly investigating the impacts of online WOM on the number of alternatives included in the consideration set in a field setting.

#### 1.2.1.2 Influence of WOM on Session Duration

In conjunction with the number of products viewed per session (consideration set size), the average session duration, reflecting how long users stay on the site, is critical for user engagement and stickiness of a platform. Unsurprisingly, longer sessions are preferred by sellers, as they indicate more engaged visits and an increase in the likelihood of purchase (Bucklin and Sismeiro 2003, Moe and Fader 2004). However, longer time also reflects higher search costs (Johnson et al. 2004, Moe 2003). Therefore, a shorter session duration might be more desirable to consumers.

WOM information, by revealing another user's preference for a given product, acts as a quality signal that can influence session duration. There are potentially two mechanisms through which this effect of WOM can manifest. On one hand, WOM can reduce the session duration because it makes high-quality products or services stand out, thereby reducing search costs by virtue of a salience effect.[2] Conversely, WOM may increase consumer awareness of the rated products and their variation in quality.[3] Triggered to be more quality conscious, customers may choose to spend more time considering their options. For example, the patient might need to make the tradeoff between quality and travel distance. This is especially important in our research setting, in which the highly rated doctors might be located far away, leading the patients to scrutinize adjacent unrated doctors more carefully. Therefore, whether WOM extends or shortens session duration critically depends on the relative strength of the two effects described above.[4]

---

[2] To the degree that consumers pay substantial attention to WOM (as documented in the surveys mentioned earlier), the presence of WOM for some doctors and not for others helps the former become more salient. This assertion is consistent with work by Shavitt and Fazio (1990), who demonstrated that salience increases the consistency between the expression of attitude and the likelihood of purchase. In our context, the display of WOM information also introduces user interface elements that are notably distinguishable between doctors with and without WOM, thereby drawing the user's attention and reducing the time needed for cognitive processing.

[3] For instance, research has shown that reviews increase sales via the mechanism of increasing awareness of these products (Berger et al. (2010).

[4] We note that there is a possible confounding factor in the second mechanism. WOM information in the form of consumer produced narratives (reviews) takes time to read and process, which may increase session duration. Indeed, in a lab experiment, Gupta and Harris (2010) found that the availability of WOM in the form of text reviews

#### 1.2.1.3 Influence of WOM on Geographic Dispersion

While making a consumption choice, a key decision factor for the consumer is the location of the product or service provider. Prior empirical research has consistently shown a strong influence of store location on store choice and sales. The distance from a consumer's home to a retail shop is not only influential in driving sales of low-stakes consumer goods and services but also highly influential in driving a high-stakes decision such as the choice of a healthcare provider. Indeed, to the extent that larger travel distances are a major barrier to access to healthcare services (Syed et al. 2013), policy research has underscored the importance of reducing geographic imbalances in the supply of doctors (Robert et al. 1998).

Theoretically, consumers' disinclination to choose a healthcare provider located far away is likely due to two factors. First, consumers may be less aware of and knowledgable about service options that are geographically distant simply because they are less likely to serendipitously encounter these providers or to interact with consumers who have used them. Second, travel represents a cost for the user, which may require both time and monetary resources. Much of the extant research has focused on the first factor (limited awareness) or a combination of the two factors. For example, research demonstrating that transportation is a major barrier for healthcare access (Guidry et al. 1997, Syed et al. 2013) has shown that both the

---

led customers to spend more time on the website. In our context, online WOM is presented in the form of a low cognitive cost metric—the accumulated count of positive recommendations for a provider. The non-availability of text reviews allows us to hold the cognitive cost constant before and after online WOM information is available, thereby enabling us to study its impact via the second mechanism.

lack of awareness and travel costs are potentially influential—patients are less aware of options at a distance and are unwilling to travel far away.

An online marketplace typically publishes identical or similar information on all available options irrespective of the physical distance, thereby effectively eliminating the consumer's lack of awareness of options far away. Introducing WOM information in such a setting provides a signal of quality which may reduce uncertainty and alter the consumer's willingness to travel. However, it is not clear how a consumer makes a tradeoff between the cost of travel versus the benefit of a higher quality provider. Would a consumer be more willing to travel farther if WOM information suggests a net gain in utility? Despite the documented importance of location and travel distance in consumer choice, there is little work on how online metrics of quality such as WOM influence the geographic dispersion of the choices considered and a consumer's willingness to travel, a gap that we address in this study.

#### 1.2.1.4 Moderating Impact of WOM Abundance

Thus far we have argued for the direct effects of WOM on the consumer choice process. We now theorize how the abundance of WOM (i.e., the number of products in a given market that have WOM available) conditions its impact on the consumer choice variables of consideration set size, session duration, and geographic dispersion of choices considered.

In many market interactions where product quality is unobservable, the revelation of others' evaluations of the product in the form of WOM is a critical pre-purchase signal that can reduce the inherent information asymmetry (Chen and Xie 2008). However, since WOM is volitionally contributed by users, there is a variation in the abundance of WOM for products across different markets: some markets have relatively few products with WOM information while others have a greater abundance of WOM. Why would the consumer response be different based on how many products within a given category are rated? We suggest that this response heterogeneity is driven by two factors: the degree of consumer trust in the WOM signal and the salience afforded by WOM.

In market interactions characterized by information asymmetry, consumers prefer to get signals for all products to distinguish between their levels of quality (Boulding and Kirmani 1993). The signal's trustworthiness is likely to increase with its pervasiveness, as the signal associated with one product is meaningful only in comparison to competing products. When only a small number of products in a given market are rated, consumers may not regard the WOM as a systematic and reliable quality signal; indeed, the credibility of a signal has been identified as a major factor in decision making (Boulding and Kirmani 1993). In that case, despite the fact that WOM helps rated options become more salient and more easily identifiable, consumers may not trust this signal enough to limit deliberation to only the rated options Likewise, when WOM information is present for just a handful of alternatives, these alternatives are unlikely to be adjacent to a user. Browsing these rated doctors and local doctors will likely increase the overall geographic dispersion among the choices considered.

In contrast to the setting with low WOM abundance, when WOM information is present for a large proportion of available products, consumers are more likely to trust this quality signal and utilize it as a filtering tool to choose among available options, limiting the consideration set to a size that is deemed reasonable to enable effective choice. Relative to the setting of low WOM abundance, whether such a strategy would result in an increase or decrease in the consideration set size and session duration is unknown, and merits empirical investigation. For the geographic dispersion of choices considered, the abundant presence of WOM could amplify the likelihood of consumers learning about high-quality doctors who are nearby, thereby increasing the chance of choosing a local doctor.

It could be argued that such variation in WOM abundance, especially a situation where a majority of products do not have WOM information, may exist only during the incipient stages of WOM in a particular product category. With the passage of time, more products in a given market are likely to accumulate WOM. However, we note that in high-stakes markets such as healthcare services, WOM is at a relatively nascent stage. In the specific context of doctors and other healthcare providers, user-generated content to date is relatively scarce, and studies show that overall, only one in six doctors in the United States receives consumer-generated online reviews (Gao et al. 2012). Further, a wide variation in WOM abundance across specialties has been documented, ranging from 5% of doctors being rated in some specialties to about 40% being rated in others (Gao et al. 2012). This variation underscores the existence of underlying heterogeneity, which can result in differences in how consumers perceive WOM information and in their subsequent choice processes. The question of how variation in WOM abundance affects consumer decision making has not been previously addressed in the literature. The presence of markets with varying densities of WOM in our dataset gives us the unique opportunity to examine its moderating influence.

In summary, following from the above discussion on the limited understanding of the role of WOM abundance as a moderator and in the light of ambiguity surrounding the effects of WOM on facets of the consumer choice process, we investigate the following research question:

*RQ1: How does online WOM affect a consumer's choice process in terms of (1) the size of the consideration set, (2) the time taken to consider the choices, and (3) the geographic dispersion* *among the choices considered, and how do these effects change with a change in WOM abundance?*

### 1.2.2 How Online WOM Impacts Demand

In addition to extending the literature on the consumer decision process, we also contribute to the study of how WOM affects demand. Our analyses depart from prior research (see Table A1) in several novel ways: we address the endogeneity of WOM, account for the rank effect, and uncover how WOM affects not only the demand for the focal product but also the spillovers on competing and unrated products. Furthermore, we study the moderating impact of WOM abundance on the relationship between WOM and demand.

#### 1.2.2.1 Endogeneity of WOM

Although a majority of the empirical work examining the relationship between online WOM and sales reports positive findings, there are conflicting results as well. For example, whereas Chevalier and Mayzlin (2006) and Dellarocas et al. (2007) find that ratings positively influence sales, Chen et al. (2004) and Duan et al. (2008) find no influence of online WOM information on sales. In this body of literature, a core challenge has been the endogeneity of the online ratings, which could be caused by several factors. First, the provision of online reviews is based on self-selection; customers voluntarily opt in to post a review(Li and Hitt 2008b). This volitional activity is likely subject to a variety of biases and social influences. Second, product quality can be an underlying factor that drives both WOM and sales. Because product quality is often hard to quantify or observe, it is difficult for researchers to determine if high online ratings are leading to higher demand or if product quality alone is driving the high demand (Zhu and Zhang 2010).

The literature has increasingly turned to more sophisticated econometric techniques to address the endogeneity challenge. To illustrate, Dellarocas et al. (2007) use fixed effects to account for unobserved movie quality. Chevalier and Mayzlin (2006) examine reviews and sales of Amazon and Barnes & Noble on different dates and use a difference-in-differences (DID) approach to account for book-specific and website-specific effects. Likewise, Zhu and Zhang (2010) use the DID approach and exploit the difference in reviews of the same video game across two different consoles.

Despite these methodological improvements, two unresolved questions persist regarding endogeneity threats in the empirical estimations. First, while the fixed effects model controls for time-invariant factors related to product quality, it fails to account for the fact that consumer preferences may change over time, which means that perceived quality evolves as well. Second, the studies that exploit variation in ratings across two platforms need to impose the underlying assumption that the differences in WOM across two platforms are exogenous "shocks" uncorrelated with other factors that may affect sales, an assumption that can be difficult to verify.

When investigating a phenomenon that intrinsically raises significant endogeneity concerns, an experiment has been advocated as the preferred approach (Angrist and Krueger 1994, Angrist 1990, Meyer 1995). In our study, we exploit a natural experiment setup in which the doctor appointment booking platform launched online WOM as a new feature on its web portal (more detail regarding this feature launch is provided in Section 3). In this aspect, our research context most closely resembles those of Tucker and Zhang (2011) and de Matos et al. (2016), both of which employ experimental settings. However, our work differs from those studies along the following important dimensions. First, Tucker and Zhang (2011) focus on the effect of the number of clicks rather than WOM. Second, Matos et al. (2016) manipulate the rank of products, whereas in our setting the online WOM was published without changing the relative display order of the doctors.

#### 1.2.2.2 Rank Effect

A second potential confounding factor in evaluating the impact of WOM on sales is the rank or position effect. It is common practice for online retailers to take WOM into consideration when ranking products on their websites. Research has shown that higher-ranked products tend to receive more clicks (Ghose et al. 2013, Ghose et al. 2014, Ursu 2015). Therefore, if the rank of the product is a function of WOM, the impact of WOM on sales could be confounded by the rank effect.

The confounding effect of rank has been a challenging issue in several studies, and researchers have attempted to alleviate the problem in a variety of ways. For instance, in their study of online bookstores, De los Santos et al. (2012) use a control function approach to address the rank effect. In analyzing search advertising, Narayanan and Kalyanam (2015) employ a regression discontinuity design to mitigate the effects of rank. In their study on the impact of ranking on consumer behavior, Ghose et al. (2014) utilize simultaneous equations to account for it. However, although *post hoc* empirical strategies are undoubtedly useful, arguably the best possible approach to account for this endogeneity is to have an empirical setting in which the rank is not endogenous to the effect under study. In our context, the platform held the rankings independent of the WOM information. Due to this decoupling, the rankings are not a function of online WOM.

#### 1.2.2.3 Impact on Competing Products

As described in Section 2.1, consumers make purchase decisions based on a comparison of competing products. Therefore, WOM is most meaningful in a relative rather than absolute sense. Traditional approaches to examining the effect of online WOM on sales tend to treat each product as an isolated and independent unit, with the analysis focusing on how the focal product's online reviews affect its sales without examining its impact on the competing products. This is likely due to the challenge of clearly defining the market where the products compete (Jabr and Zheng 2014). To illustrate, the demand for a fiction book may be substituted by many other fiction books, books in other categories, or even music. Therefore, it is difficult to delineate the exact boundary of the products competing with each other. Our research setting addresses this challenge by defining the city-specialty dyad as the local market. Healthcare services are mostly obtained locally, and doctors are not substitutable across categories because medicine is a highly specialized field. For example, a neurologist cannot fulfill a consumer's requirement for a cardiologist. This clear demarcation in demand helps us unambiguously define delimited markets and competing doctors. Based on the well-defined local market, we are able to construct the relative ranking, i.e., how "well" a doctor fares on consumer-generated assessments, based on WOM specific to local competition, and examine its impact on consumer demand.

In addition, we extend our analysis to products with no ratings. The question of how online ratings affect the sales of non-rated products is clearly important for any new business that is just entering a market. It is also of particular importance in healthcare, where only a modest fraction of the doctors have received online ratings (Gao et al. 2012). Most existing studies limit their samples to products that have online ratings or reviews (Chevalier and Mayzlin 2006, Li and Hitt 2008b); we address this gap.

#### 1.2.2.4 Impact of WOM Abundance on Demand

As described in Section 2.1, the WOM signal's trustworthiness likely increases with its pervasiveness, as consumers evaluate it in comparison to the signals of competing products.

Furthermore, when the abundance of online WOM increases, it is more likely that a known alternative also has this signal, further strengthening its validity. Therefore, it could be plausibly argued that higher WOM abundance would lead to a sharper shift in demand in favor of those with this signal. On the other hand, it is also possible that the presence of WOM for a large number of options may dilute the veracity of the signal or make it difficult for consumers to detect nuanced differences in quality, thereby reducing its impact. Prior research on WOM, for example, Chevalier and Mayzlin (2006) has investigated the role of WOM volume and valence for individual products on demand; however, there is a lack of research on how demand varies with a change in abundance of WOM at the market level. Our context allows us to address this question. Following from the discussion above, the second research question we investigate in our study is:

*RQ2: What is the impact of online WOM on consumer demand for the focal and competing products, and how do relative WOM ranking and WOM abundance moderate this relationship?*

To summarize, we study the impact of online WOM on three important facets of the consumer choice process: consideration set size, the time taken for consideration (session duration), and the geographic dispersion of the considered choices. Furthermore, we examine the impact of online WOM on demand for both rated and unrated physicians. Finally, we investigate the role of important moderators of these relationships—the abundance of WOM and relative WOM ranking. A conceptual model of our research is presented in Figure 1.1.

## 1.3 Research Setting

### 1.3.1 Data Source

Our data source is a leading online doctor appointment booking platform that has listings for over one-third of all the actively practicing private doctors in India. The platform allows patients to search for and book appointments. The search listings provide a concise summary of each doctor's professional details (see Figure A3 and A4 in the appendix). By clicking on the listing, the patient can view a detailed profile page for the doctor and book an appointment.

The website has a market presence in many specialties in the major cities of India. This presence allows us to isolate multiple non-overlapping markets, with each market defined as a combination of a city and a specialty. For example, all the cardiologists in the city of Bangalore constitute one market, and all the dermatologists in Bangalore constitute another. This ability to construct an unambiguous delineation of what constitutes a market allows us to investigate the effect of the abundance of online WOM on the consumer.

### 1.3.2 Exogenous Change in the Visibility of WOM

The website began to collect feedback on doctors from patients on May 7, 2014; the feedback came in the form of a "yes" (positive) or "no" (negative) response to the question of whether the patient would recommend the doctor. A sample email to solicit such recommendations is illustrated in Figure A1 in the online appendix. On June 25, 2014, the website published the count of positive patient recommendations for doctors who had at least three positive recommendations and whose total recommendations were at least 70% positive. This exogenous change in the availability of patient recommendations (WOM) provides a distinctive research opportunity that allows us to examine the impact of online WOM on patients' browsing and appointment-booking behavior. Figure A2 in the online appendix presents snapshots of the search pages before and after the recommendations were published. We construct panel data with an observation window that starts one week before and ends one week after this exogenous change. Figure 1.2 displays the timeline of this exogenous change.

Several strengths of our research setting and design are noteworthy and contribute to the empirical rigor of the study. First, by restricting the time window to one week before and after the publication of recommendations, we can largely rule out any systematic change in doctors' quality arising from the visibility of the WOM information, thereby enabling us to isolate the causal effects associated with the visibility. Second, the rank or position effect is not present in our setting. On the platform, the doctors within the user-specified city are displayed sequentially, with ten results visible on one page. The ranking is determined based on a number of factors, including the percentage of completion of a doctor's profile, prior traffic to that doctor, and proximity to the preferred locality specified by the patient. The ranking of doctors during the period of our study was held constant and was not a function of the number of recommendations received by the doctor. Additionally, unlike some popular e-commerce platforms, our setting does not show products in descending order of ratings or allow users to selectively view five-star rated products only, which might overstate the effect of WOM.

Finally, there have been increasing concerns over practitioners' strategic use of WOM by creating fake reviews (Jabr 2015, Luca and Zervas 2016, Mayzlin et al. 2014). Thus, a legitimate worry is that doctors may act strategically to improve their online WOM rating by proactively adding positive recommendations. In our setting, the platform only solicits feedback from verified patients, thereby lowering the likelihood of such strategic behavior.[5] In addition, no

[5] The website uses a customer verification system to discourage any strategic gaming. Only verified patients can post a recommendation, and even a verified patient cannot post recommendations more often than once a month for any doctor or for a single specialty in a city. Although this largely solves the problem of deliberate gaming, it also lowers the respondent pool. To counteract this issue, the web portal has deployed a feedback collection mechanism

advance notification was given to doctors about when patient feedback would be published, further mitigating the threat that doctors would react strategically.

### 1.3.3 Data

Our dataset is constructed by matching five distinct subsets: clickstream logs, doctor profile data, recommendation logs, appointment logs, and clinic location data. Clickstream logs allow us to reconstruct the browsing behavior of the users, which we use to calculate each user's consideration set. The platform provided us the clickstream logs in the Nginx server format for the duration of the study period. From these logs, we construct clickstream data to investigate our research questions related to consideration sets. The platform replaced each IP address with a unique identifier in each log that was provided to us. This identifier helps us to track the browsing behavior of an individual without compromising their privacy.

To create panel data for our models from the log files, some of which exceeded 50 gigabytes after decompression, we use the cloud computing infrastructure of Amazon Web Services (AWS). After tabulating the cleaned panel, our observation window contains approximately 42,000 web sessions. We consider 30 minutes as the idle timeout for web sessions.

---

that encourages a higher response rate. The mechanism entails soliciting feedback using a "missed phone call" as an additional medium through which the patients can anonymously recommend or decline to recommend the doctor they have consulted. The concept of a "missed phone call," widely used in India, is a relatively novel method of feedback collection whereby users are asked to dial a specified number from their verified number to register a "yes" and to an alternative number to register a "no." The respondent then disconnects the call prior to call completion. This method of feedback ensures that the user does not incur the financial burden of a mobile carrier charge.

In making this determination, we needed to consider that the length of the idle timeout on a given website tends to rely heavily on the security required by certain web platforms or on other site-specific factors. For instance, the Payment Card Industry (PCI) Data Security Standard (DSS) requirements state that if a user has been idle for more than 15 minutes, then the system must require the user to be re-authenticated to re-activate the session. In our context, the platform's own analytics team believes that users may require up to about 30 minutes to browse a doctor's page, sometimes even opening new tabs and searching online for procedures offered by the doctor. Hence, considering the practice followed by the platform for its internal analytics, we allow up to 30 minutes of inactivity before a timeout is in effect. Therefore, if a user opens two doctor profiles within a 30-minute period, these two profile views would be counted in the same session. However, if the user opens two doctor profiles and the inactivity period between these two profile views was longer than 30 minutes, then this episode would be counted as two distinct sessions. This protocol of a user session expiring after 30 minutes of inactivity is also widely used on web data analytics platforms such as Google Analytics. In addition, studies in computer science (Jansen et al. 2007, Nasraoui and Saka 2007) and marketing (Montgomery et al. 2004) have followed a similar inactivity timeout protocol.

The clickstream logs record all requests made to the web server, and this includes requests from all automated clients. To ensure that our analysis measures only the behavior of real users, we conduct extensive data cleansing. We first exclude the traffic from all automated clients (web bots) such as search engine bots, web spiders and scrapers. To do so, we parse the user agent field in our clickstream logs with a user agent library in Python that maintains a repository of signatures of all known bots. This helps us in identifying and removing all web sessions that originate from known web bots. Thereafter, in consultation with the platform, we identify and exclude web traffic initiated by clinic receptionists. Some clinics use the online booking system on the web platform as a convenient online calendar of appointments that are booked by patients through other channels (phone, offline walk-in, etc.). Typically the web traffic of such a clinic's receptionist can be observed in the log as multiple web sessions originating from the same IP address that browse the doctors in the same clinic. Furthermore, in consultation with the platform, we exclude traffic originating from a page refreshed by the patient, which would be done due to internet connectivity issues. This traffic can be identified by observing the web server request status field in the clickstream log files and isolating duplicate requests within a short span of time.

Furthermore, we exclude targeted search sessions by dropping all traffic that requests an individual doctor's profile page without being preceded by a search. We exclude these sessions as they are likely to represent a premeditated choice and do not model the normal consideration process. After this extensive data cleansing, we match the URLs opened in each session to specific doctor profile pages to identify the consideration set. Finally, since it is still possible that the dataset may include traffic from sophisticated human crawlers that do not have a documented signature, we exclude the top 2% outliers[6].

---

[6] The human crawlers can be either web crawlers that do not have a known signature, or competitors/researchers who browse the website extensively to scrape the data. We exclude the top 2% outliers on each of the three constructed measures of consideration set - the consideration set size, the duration of the browsing session and the geographic dispersion. In total, this results in an exclusion of 3.45% of the web sessions. We would like to note that the results of our analysis remain unchanged even when these outlier sessions are included.

The doctor profile data includes information on the city, specialty, and URL, which helps us identify the market for each doctor. The profile data also contains information on the services offered and the facilities at the clinic. A sample profile page is depicted in Figures A2 and A3 in the online appendix.

To determine the doctor's WOM information as displayed to the users, we calculate the count of recommendations visible for each doctor on each day of our study window using the recommendation log files. Similarly, we tabulate the number of appointments for each practice-doctor dyad for each day of our observation period using the appointments log. We note that we use the practice-doctor dyad for tabulation of appointments, whereas the recommendations are tabulated at the doctor level. This is because although recommendations accumulate at the doctor level, the same doctor can practice at multiple clinics. To the extent that a consumer search is likely to be driven by the physical location of the service facility (Bell et al. 1998, Kang et al. 2003), the effects on competing products must be determined at the practice level as that accounts for the location.

On the day online WOM was launched on this website, 209 doctors cleared the threshold for their recommendations to be visible. These constituted 312 practice-doctor dyads. Due to the nature of this feature launch, there was considerable heterogeneity across markets in the number of doctors whose recommendations were published. For example, in one market there may have been fifteen doctors with published recommendations while in another there might be just three. This heterogeneity allows us to investigate how the abundance of WOM in a market affects consumer decision making.

Variables for the consumer decision process: We construct our derived data after merging the datasets described above. To investigate our first research question regarding the impact of

WOM on the dimensions of consumer choice, we construct measures for the three dependent variables—the consideration set size, the duration of the browsing session and the geographic dispersion of the choices considered.

For the first measure, consideration set size, we use the metric suggested by Moe (2003), the count of unique product pages browsed. In our context, this represents the number of unique doctor profiles viewed in a web session. Session duration is calculated as the time elapsed between first and last GET request received by the server from the client's browser. Finally, for geographic dispersion, we obtain the precise latitude and longitude information for the clinics listed on the platform. We match this clinic-level information to doctor details and compute the coordinates of all doctors browsed in a given session. This matching process allows us to calculate the coordinates of all the doctors browsed for approximately a quarter of the web sessions. From this plot of coordinates of all doctors browsed in a web session, we calculate geographic dispersion by measuring *Max_distance*, which is defined as the haversine distance between the two farthest clinics browsed in the session.

<u>WOM abundance:</u> To investigate the role of WOM abundance as moderator, we divide the markets that have at least one doctor with published WOM into three subgroups. These subgroups are based on the number of doctors with WOM in a given market and are constructed such that each subgroup has approximately the same number of sessions in our data sample. The first subgroup is *Spa* (sparse WOM), which includes the markets that have one or two doctors with visible WOM. The *Spa* subgroup has 11,780 sessions in our data sample. The second subgroup is *Mod* (moderate WOM), consisting of all markets that have three to nine doctors with WOM; our data contains 9,339 sessions from these markets. The third subgroup is *Den* (dense

WOM), which is composed of all the markets that have 10 or more doctors with WOM. We have 10,309 user sessions in this subgroup. [7]

Doctor appointments: To investigate the second research question regarding consumer demand, we construct a longitudinal panel of the number of online appointments of each practice-doctor per day. To construct the moderator, which is the relative rank of the doctor among those with WOM in a given market, we divide the rated doctors into three groups in the focal market. The *Top* group is composed of all doctors who are in the top 33% of doctors in their respective city-specialty (market) when ranked by the count of visible positive recommendations. Similarly, the *Middle* group is composed of the doctors who are in the middle 33%, and the *Bottom* group is composed of the remaining doctors. (Note that in the case there is a tie and a doctor can be placed in more than one group, then that doctor is placed in the lower of the two groups. For example, if a doctor can be placed in both the *Top* and the *Middle* groups, then this doctor is placed in the *Middle* group.)

Summary: The variables in our constructed dataset are summarized and described in Table 1.1 to Table 1.9. Table 1.1 and Table 1.2 present the summary statistics for the number of recommendations and appointments. In our data, there are 45 doctors in the *Top* group, 78 in the *Middle* group, and 86 in the *Bottom* group. Table 1.3 summarizes the user sessions. There are

---

[7] In this research context, like most web platforms, the listings are displayed page by page, with each page displaying 10 results. Like most other search engines, even in this context, most users browse the first page only. Therefore, the effective denominator is the same across all markets, making WOM abundance and WOM density interchangeable. In general, the idea of density is a ratio. Therefore, to keep the wording consistent we use the term WOM abundance.

42,229 sessions, with a mean consideration set size of 1.81 and a mean session duration of 632.2 seconds. Table 1.4 presents the count of sessions and markets at the subgroup level. There are a roughly equal number of sessions in the *Spa* (Sparse WOM), *Mod* (Moderate WOM), and *Den* (Dense WOM) markets. Table 1.5 lists the top specialties and cities in our dataset. Dermatology is the most popular specialty, and most traffic comes from Bangalore. Correlation matrices are presented in Tables 1.6 and 1.7. As we can observe, the correlations between the dependent variables studied (*Unique, Duration, Max_dist*) are quite low. Tables 1.8 and 1.9 present the names and descriptions of the variables in the consideration set analysis and doctor appointment analysis respectively.

## 1.4 Main Analyses

### 1.4.1 Impact of Online WOM on Consumer Decision Making

We first analyze how WOM impacts consumer decision making. As described in RQ1, we investigate the impact of WOM on three facets of a consumer's choice process: the consideration set size, the web session duration, and the geographic dispersion among the providers considered. The following subsections present the analysis of the impact of online WOM on each of three facets mentioned above.

#### 1.4.1.1 Impact of WOM on Consumer's Consideration Set Size

To empirically measure the impact of WOM on the consideration set, we implement a simple pre- and post-comparison for markets with at least one provider with visible online WOM during the treatment period (i.e., the treatment group):

$$Y = \alpha + \beta\, Post + Market\ dummies + Day\ of\ the\ week\ dummies + \varepsilon \quad (1)$$

The dependent variable in this analysis is the consideration set size, which is the number of unique doctor profiles viewed in a web session. For the independent variables in this specification, we capture the time effect by including an indicator variable (*Post*) that is set to 1 when WOM is visible and 0 otherwise. Additionally, it is possible that browsing behavior differs on different days of the week. For example, patients may have more time to search for doctors over the weekends; to control for this variation in browsing behavior, we include the *Day of the week* dummies in our model.

Furthermore, we also include dummies for each market, which is defined as a city-specialty dyad, to account for any unobservable market-level factors that may impact consumer browsing behavior. The inclusion of these dummies help us to control for any city- and specialty-specific idiosyncrasies. For example, a patient seeking a cardiologist may have a different mindset, and hence a different browsing behavior, than a patient seeking a dentist. Furthermore, city-specific characteristics—such as population abundance, literacy, IT infrastructure, income levels, and other unobservable factors—may affect browsing behavior. In addition to helping control for idiosyncrasy, the city-specialty dyad dummies are more granular and flexible than city and specialty fixed effects and therefore help us to better control for unobserved heterogeneity[8]. As there may be unique unobservables associated with city-specialty pairs, the inclusion of market dummies adds greater rigor into the analysis than the inclusion of individual city and specialty dummies. Finally, we note that by including the market dummies, our model

[8] For example, if there are 10 cities and 10 specialties, there will be 100 dyad dummies, compared to 20 city and specialty dummies. The 100 dyad dummies supersede the 20 city and specialty dummies.

translates into a within-variation model at the market level; these dummies absorb all the time-invariant market characteristics.

In our specification, the dependent variable is the number of doctors browsed in a session, which is a count; therefore we use a negative binomial estimator. We use robust clustered standard errors in all the regressions, clustered at the level of a market to control for repeated observations from the same market. The key independent variable of interest in this model is *Post*, which captures the effect of visibility of online WOM on the size of the consideration set. The results of the specification in Eq (1) are presented in Column 1 of Table 1.10. The coefficient of *Post* is 0.005 and is statistically insignificant. A potential explanation for the lack of an overall effect can be that some markets may be driving up the consideration set size while others may be driving it down, thereby canceling out the net effect.

To uncover whether this canceling effect is in action and following RQ1, we examine the moderating impact of WOM abundance—the count of providers in a market with WOM information, categorized as *Spa*, *Mod*, and *Den*, as described in Section 1.3. To compare the effects across markets of different WOM densities, we use the following model:

$$Y = \alpha + \beta_1 PostXSpa + \beta_2 PostXMod + \beta_3 PostXDen + Market\ dummies + Day\ of\ the\ week\ dummies + \varepsilon \quad (2)$$

Results of the regression are in Column 2 of Table 1.10. The coefficient of *PostXSpa* is 0.144 and is significant at the $p<0.01$ level, while the coefficient of *PostXMod* is 0.014 and is statistically insignificant. In contrast, the coefficient of *PostXDen* is negative at -0.172 and is significant at the $p<0.01$ level. These results suggest that in the *Spa* markets users browse more options when WOM is presented, while in *Den* markets users browse fewer options when presented with WOM.

The identification in Equation (2) to an extent assumes that there are no other temporal changes in the pre- and post-periods that affect consumer choices. To control for the possible temporal difference, we estimate a difference-in-differences model[9] that includes the markets that do not have doctors with any WOM information. We introduce the *Post* variable to capture the time trend:

$$Y = \alpha + \beta_1 Post + \beta_2 PostXSpa + \beta_3 PostXMod + \beta_4 PostXDen + Market\ dummies + Day\ of\ the\ week\ dummies + \varepsilon \quad (3)$$

Column 3 of Table 1.10 reports the estimates of this model. The results are consistent with estimates from the previous specification (Equation 2). The coefficient of *PostXSpa* is 0.154, statistically significant at the p<0.01 level. The coefficient of *PostXMod* is 0.026, statistically insignificant, and the coefficient of *PostXDen* is -0.162, statistically significant at p<0.01. This further confirms that our findings are unlikely to be driven by the underlying temporal changes.[10]

---

[9] To check the internal validity of the differences in differences model, we also examined the parallel trends assumption. We provide the details in section 1.4.3.10.

[10] In addition to any underlying temporal change driving the results, it is possible the effect of WOM on consideration set size is driven by a mix of two types of consumers, each of which may have significantly different browsing behavior: those that search casually and have a very small consideration set, and those that search exhaustively and have a significantly larger consideration set size. We tested the effect of the presence of WOM on both of these types of users by separating out smaller and larger sessions (by consideration set size). Our analysis shows that there is no significant effect in the smaller sessions (1 unique profile viewed) and the consideration set size remains the same after WOM is revealed in markets of different (*Spa*, *Mod*, *Den*) densities. Further analysis

We plot the bar charts of these effects in Figure 1.3 using marginal effects obtained after running the above-specified regression model. We notice that the magnitude of increase in the size of the consideration set in *Spa* markets is considerable. The size of the consideration set increases from 1.63 to 1.90, an increase of 16.6%. Similarly, when users are presented with WOM information for many alternatives in the market (*Den* markets), the magnitude of the decrease in the size of the consideration set is substantial, from 2.14 to 1.82, a drop of 14.9%. By contrast, in markets that have no doctors with WOM information, the magnitude of change in the size of the consideration set is not substantial, from 1.72 to 1.70, a drop of 1.1%.

We further investigate the impact of WOM on the composition of the consideration set. First, we examine whether the sessions become more likely to include rated doctors when WOM is revealed. *Treated_Viewed*, a dummy outcome variable, indicates whether a web session contains at least one rated doctor. Since it is a dummy, we use the Logit estimation. We present these results in Column 1 of Table 1.11. As we can observe, the coefficient of *PostXSpa* is 0.279, statistically significant at $p<0.01$. The marginal effect analysis shows a change from 0.133 to 0.164 in the likelihood of consideration, a sharp increase of 23.1%. The coefficient of *PostXDen* is 0.176, statistically significant at $p<0.05$. The marginal effect analysis shows a change from 0.304 to 0.340, a relatively moderate increase of 12.0%.

Second, to check whether the change in consideration is leading towards a change in demand, we examine whether users are concluding their sessions more often on a rated doctor, i.e., terminating their search. To do so, we specify *Last_Viewed*, a dummy outcome variable, to

---

confirms that the results are driven by a change in the behavior of consumers who search more exhaustively. We thank the AE for suggesting we explore this heterogeneity in user behavior.

indicate whether a rated doctor is the last viewed doctor in the web session. We present these results in Column 4 of Table 1.11. As we can observe, the coefficient of *PostXSpa* is 0.157, statistically significant at $p<0.05$. The marginal effect analysis shows a change from 0.095 to 0.108, a relatively moderate increase of 13.4%. The coefficient of *PostXDen* is 0.293, statistically significant at $p<0.01$. The marginal effect analysis shows a change from 0.193 to 0.240, a sharp increase of 24.5%.

Taken together, these results suggest that when only a few doctors have WOM, these doctors stand out and are hence more likely to be considered. However, consumers may not trust this signal to a great degree and hence are less likely to conclude their browsing session on these doctors. On the contrary, when many doctors have WOM, consumers may trust the WOM signal and hence are more likely to conclude their browsing session on a doctor with WOM.

#### 1.4.1.2 Effect of Online WOM on Session Duration

To empirically measure the effect of online WOM on session duration, we use specifications similar to those discussed in Section 1.4.1.1, replacing the dependent variable of consideration set size with the session duration measured in seconds. Given the nature of the outcome, we apply an Ordinary Least Squares (OLS) model with clustered robust standard errors to estimate the effects.

Results are in Table 1.12. As seen in Column 1, the presence of WOM on average increases the duration of the session. We observe that the coefficient of *Post* is 33.24 and is significant at the $p<0.01$ level. The marginal effect indicates that the average duration of a web session increases from 638.5 seconds to 671.7 seconds, an increase of 5.2%. In Column 2, we report the results of the moderating influence of WOM abundance. We find that in *Spa* markets, users increase the duration of their browsing session when WOM information is revealed, while in *Den* markets, users decrease the duration of their browsing session when WOM information is revealed. The coefficient for *PostXSpa* is 130.2 and significant ($p<0.01$), whereas for *PostXDen* it is -81.2, significant at the $p<0.05$ level.

Next, we set up the model similar to Equation (3) in Section 1.4.1.1 to account for the time trend by incorporating sessions from markets that have no doctors with visible recommendations. We add the time trend by including the post-period dummy (*Post*) in our model specification and present the results in Column 3 of Table 1.12. The results remain consistent. The coefficient for *PostXSpa* is 128.8 and is significant at the $p<0.01$ level, whereas the coefficient for *PostXDen* is -82.9, significant at $p<0.05$ level.

We plot bar charts in Figure 1.4 to illustrate the marginal effects of WOM information on session duration. In these charts, we compare the average session duration in the pre- and post-periods; the session durations represent marginal effects from the regressions and hence include all controls. We observe that the overall effect is positive, and when WOM is revealed in *Spa* markets, consumers browse longer, while when WOM is revealed in *Den* markets, consumers browse for a shorter duration. The effect size is considerable. In the *Spa* (low WOM abundance) markets, the *Duration* increases from 538.1 seconds to 666.8 seconds, an increase of 23.9%. In the *Den* (high WOM abundance) markets, the *Duration* decreases from 765.1 seconds to 682.3 seconds, a decrease of 10.8%.

Furthermore, to examine the change in user attention to products considered when WOM is available, we examine the influence of WOM on time duration per profile. Time spent by customers on each product's profile can help us understand whether the presence of WOM could be drawing the user's attention to other details mentioned in the profile. To set up an empirical model for this analysis, we specify the outcome variable as *Log_duration_per_profile*, defined as the natural logarithm of web session duration divided by the size of the consideration set. Given the nature of the outcome, we apply an Ordinary Least Squares (OLS) model with clustered robust standard errors to estimate the effects. We present these results in Column 1 of Table A2 in the appendix. As we can observe, the coefficients of *PostXSpa* and *PostXMod* are statistically insignificant, while the coefficient of *PostXDen* is 0.261, statistically significant at $p<0.01$. The marginal effect analysis shows a change from 3.8 seconds to 4.1 seconds, an increase of 7.9%. This indicates that in markets with high WOM abundance, there is a significant increase in browsing time per profile, thus suggesting that in these markets the WOM signal is trustworthy enough to persuade consumers to browse their options more carefully. The results indicate that the presence of online WOM information in dense markets increases users' attention to product details. In our context these are doctor details, such as services offered, sub-specialization(s), certifications and education, past and present hospital affiliation, clinic certifications, etc.

#### 1.4.1.3 Effect of Online WOM on Geographic Dispersion

As noted, in RQ1 we examine how the availability of WOM influences the geographic dispersion of the providers considered. We set up a model similar to the one in Section 1.4.1.1 and estimate the effect of the treatment (i.e., visibility of online WOM) on geographic dispersion. The dependent variable here is the *Max_distance*, measured in kilometers. Results are reported in Table 1.13. As can be seen from Column 1, the coefficient of *Post* is 0.335 and is statistically significant at the $p<0.05$ level, suggesting that on average, when WOM information is revealed, there is an increase in the geographic dispersion of choices considered. Investigating further, we find that this increase is driven by sessions in *Spa* markets, while in contrast, the geographic dispersion decreases in *Den* markets. These estimates are reported in Column 2 of Table 1.13.

The coefficient for *PostXSpa* is 1.307, significant at the p <0.01 level, while the coefficient for *PostXDen* is -0.805, significant at the p<0.05 level.

Together, these results suggest that when there are few rated doctors in a user's city, the user is likely to find a rated doctor farther away and therefore more likely to travel longer distances for a consultation. In other words, when WOM information is scarce, the quality signal embedded in the WOM appears to motivate the consumer to incur some inconvenience to purchase a higher quality product. In comparison, when WOM is available for many providers, consumers may not need to explore doctors far away, as the strong WOM signal persuades them to consider the highly rated options that are close to their home. This shows up as a decrease in the measure of geographic dispersion.

Next, similar to the specification in Equation (2) from Section 1.4.1.1, we include the observations for sessions from markets that do not have any physicians with visible recommendations. We present the results of this model in Column 3 of Table 1.13. The coefficient of *PostXSpa* is 1.466, statistically significant at p<0.01.

Figure 1.5 presents bar chart plots of the marginal effects, with the above specifications. The overall effect is positive. When WOM is revealed in *Spa* markets, consumers browse doctors that are geographically more dispersed while when WOM is revealed in *Den* markets, consumers browse doctors that are geographically closer to each other. The effect size is considerable. In the *Spa* (low WOM abundance) markets, the *Max_distance* increases from 8.73 km to 10.20 km, an increase of 16.8%. In the *Den* (high WOM abundance) markets, the *Max_distance* decreases from 9.09 km to 8.44 km, a change of 7.1%.

These results suggest that online WOM can affect clients' location sensitivity and therefore can also affect the geographic area from which a service provider draws its customers.

In other words, online WOM reputation of a business may be a key variable in determining the importance of proximity to clients.

To summarize, we find that online WOM impacts the three dimensions of a consumer's choice process: the consideration set size, the time taken to consider (web session duration), and the geographic dispersion among the choices considered.[11] Importantly, our results show that the impacts of WOM on the decision-making process are not monotonic. Rather, these impacts are contingent on the abundance of WOM available in a market.

### 1.4.2 Effect of WOM on the Demand for Doctors

To address the second research question, we examine the effects of online WOM on consumer demand. We measure consumer demand as the number of appointments a practice-doctor receives per day. Because appointments booked are a close proxy for "sales" in the context of service providers, they are an important economic outcome.

#### 1.4.2.1 Effect on Rated Doctors

To investigate this effect, we use a negative binomial regression model on doctors with visible WOM, as below:

$$y_{it} = \alpha_i + \beta_1 Post + Dummies\ for\ each\ practice\_doctor + Day\ of\ the\ week\ dummies + \varepsilon_{it} \quad (4)$$

[11] Since we are examining three aspects of the consumer choice process—the consideration set size, the duration, and the geographic distance—it is possible that the error terms in these equations are correlated, as they come from the same search session. To check for this, we apply the seemingly unrelated regression (SUR) estimation. As expected, we gain some statistical significance in certain coefficients; overall results are consistent with the main analyses.

The dependent variable for this model is the count of appointments booked for practice-doctor *i* on day *t*. On the right-hand side, the dummy for each practice-doctor *i* accounts for time-invariant characteristics of the individual practice-doctors. In addition, as before, we incorporate the *Day of the week* dummies to control for variations in browsing behavior arising from any idiosyncrasies of the day of the week, such as the weekday/weekend effect. Our focal independent variable is the *Post* dummy, which captures the treatment effect.

Table 1.14 reports the results of the appointment bookings regression with doctor (practice-doctor) fixed effects. In Column 1, the coefficient of *Post* is 0.144 and is significant at the p<0.01 level, suggesting that doctors' appointments increase after their WOM information is published—an increase of 15.4%, according to the marginal effects.

We next investigate the role of WOM abundance with regards to the demand for doctors. Results are in Column 2 of Table 1.14. As we can observe, the coefficient for *PostXDen* is 0.206 and is statistically significant at p<0.01, whereas the coefficient of *PostXSpa* is statistically insignificant. This shows that the presence of WOM shifts demand only in the dense WOM markets. These results are consistent with our observation that WOM, where it is heterogeneously distributed in abundance across markets, shifts demand only when it is available for many alternatives since it increases consumers' trust in the quality signal.

Furthermore, we examine the role of the market-specific relative rank of doctors. Figure 1.7 displays the pre- and post-period appointment bookings for each measured day of the week[12]. We see that there is a visually discernible change for the *Top* group in the post-period, with a

---

[12] Sunday is excluded as most of the clinics are closed on that day and therefore the appointment count is out of the range compared to other days.

significant increase in appointments after WOM information is revealed. This effect is much subdued in the *Middle* and *Bottom* groups. It is also interesting to notice that unrated physicians see a drop in appointments after the WOM information is revealed for the rated doctors they compete with. Finally, we construct a comparison group of physicians, which consists of doctors who fail to meet the criteria for publication of WOM and have no competitors (physicians in the same city-specialty dyad) with any visible WOM information. These physicians should not be impacted by the availability of WOM in other markets, yet they are still comparable to an extent, as they have some accumulated WOM. As expected, and as observed in Figure 1.7, there is no discernible change in the appointment booking trend for this group between the pre- and post-periods.

To empirically investigate the moderating impact of the relative rank of doctors measured as volume of WOM, we use the following negative binomial model for appointments:

$$y_{it} = \alpha_i + \beta_1 PostXTop + \beta_2 PostXMiddle + \beta_3 PostXBottom + \\ Dummies\ for\ each\ practice\_doctor + Day\ of\ the\ week\ dummies + \varepsilon_{it} \quad (5)$$

The focal independent variables are the terms that interact *Post* with the respective subgroup dummy. From the results reported in Column 3 of Table 1.14, it is evident that there is substantial heterogeneity in the effect of online WOM on appointments for these subgroups. Specifically, the coefficient of *PostXTop* indicates a significant ($p<0.01$) effect of considerable magnitude; the coefficient is 0.255 for *Top* practice-doctors. This suggests that the doctors who have accumulated more positive recommendations than their peers gain substantially more appointments. The marginal effect is substantial: the average count of online appointments booked for the top-rated practice-doctors increases from 0.54 to 0.70, a 29.6% increase. We plot simple bar charts of marginal effects in Figure 1.6 and observe that the overall effect is positive, with the increase being largely driven by *Top* doctors.

The identification in Equation (5) assumes that there are no other temporal changes in the pre- and post-periods that affect consumer choices. To account for possible temporal differences, we estimate a difference-in-differences model by including practice-doctors from markets that do not have any practice-doctor with published WOM. We introduce a variable to capture the time trend (*Post*), as specified in the models below.

$$y_{it} = \alpha_i + \beta_1 \, Post + \beta_2 PostXTreat + Dummies \; for \; each \; practice\_doctor + Day \; of \; the \; week \; dummies + \varepsilon_{it} \quad (6)$$

$$y_{it} = \alpha + \beta_1 \, Post + \beta_2 PostXTop + \beta_3 PostXMiddle + \beta_4 PostXBottom + Dummies \; for \; each \; practice\_doctor + Day \; of \; the \; week \; dummies + \varepsilon_{it} \quad (7)$$

Columns 5 and 6 of Table 1.14 report the estimates of these models. The results are consistent with those in the pre- and post-comparison. As we can observe from Column 4, the coefficient of *PostXTreat* is 0.311, and it is significant at the p<0.01 level. In Column 5, the coefficient for *PostXTop* is 0.421, significant at the p<0.01 level, and the coefficient for *PostXMiddle* is 0.280, significant at the p<0.01 level. These findings are similar to the pre- and post-analysis and indicate results that are robust to any underlying changes in the same time period.

In summary, the results indicate that online WOM shifts demand. This demand shift is more favorable for doctors who have a relatively large number of positive ratings as compared to their competitors in the same market.

#### 1.4.2.2 Effect on Competing Unrated Doctors

We argue in Section 1.2.2.3 that the shifts in demand precipitated by online WOM should be examined by considering the effects on competition. To investigate the impact of online WOM on competing unrated doctors, we select a group of doctors that are the closest in accumulated WOM to those whose recommendation information is visible. As previously mentioned, there is a threshold of at least three positive recommendations with a positive recommendation rate of at least 70% for WOM to be visible. We utilize this threshold to construct our *competing unrated doctor* group, which consists of all practice-doctors who have either fewer positive recommendation than the doctors in the treatment group or who have missed the 70% threshold and therefore do not have their WOM information published. One concern may be that these doctors are systematically different since they do not have enough positive recommendations; to account for all time-invariant characteristics, our empirical model includes dummies for the individual practice-doctor dyads. In effect, the inclusion of dummies for individual practice-doctor translates our model into a within variation model with respect to practice-doctors.

To investigate the impact of WOM on demand for these unrated doctors, we use the same model as in Equation (4) and regress the online appointments of a doctor $i$ on day $t$ on the post-period dummy. We report these results in Column 4 of Table 1.14. As can be observed, there is a substantial and significant negative impact (-0.143 at $p<0.05$) on the appointment count of unrated practice-doctors. The marginal effect size shows that the online appointment bookings decrease from 0.20 to 0.17, a drop of 15%.

We also plot a simple bar chart in Figure 1.6 to depict the marginal effects. The bar chart shows that the overall effect of online WOM is positive on the doctors with ratings, but the competing unrated doctors lose traffic. This suggests that the presence of online WOM increases the demand for rated products, but this increase comes at a cost to those who are not rated.

### 1.4.3 Robustness Tests

We perform several robustness tests to rule out alternative explanations for our findings.

#### 1.4.3.1 Effect of WOM in a Smaller Window

One potential concern in our natural experiment setting is that the doctors may have reacted to their newly visible online WOM immediately, changing their service quality to differentiate themselves from their competitors, and this change in quality may have induced more traffic. While our study considers a narrow window of one week before and after the treatment to avoid confounding the effect of the visibility of WOM with any intrinsic change in quality, to gain further confidence that it is not a change in the quality of service that is driving our results, we repeat the main analyses with an even narrower window of four days from the week prior to the feature launch and four days from the week after the feature launch (selecting the same days of the week). Columns 4 and 5 of Tables 1.10, 1.12, and 1.13 report the impact of online WOM on all the three dependent variables when the window of observation is restricted to four days before and after the treatment. Columns 2 and 5 of Table 1.11 report the impact on consideration set composition in the reduced window size. Results remain consistent.

#### 1.4.3.2 Excluding Repeated IP Addresses

A potential alternative explanation for any influence on consumer choice processes could be that a returning user may remember a few providers from their prior search and this may influence their subsequent search. To ascertain that our findings are not driven by returning users, we exclude all sessions for IP identifiers that have browsed the site in both the pre- and post-periods. We report these results in Columns 6 and 7 of Tables 1.10, 1.12 and 1.13. The results are consistent with our main findings, thereby alleviating this concern to a great degree. However, since some users may be browsing this website from multiple IP addresses/devices or may have some activity on the website outside the observation window, we cannot fully control for the repeat user behavior.

#### 1.4.3.3 Excluding the Top Doctors

In the setting of an early stage doctor appointment booking platform like the one in our study, one concern is that the demand might be disproportionately concentrated on the top few highly recommended doctors, which can drive the results. To rule out this possibility, we re-perform our analyses after excluding the top 10% of rated doctors. Results are in Column 7 of Table 1.14. We find that the results remain consistent.

#### 1.4.3.4 Correlated Outcome Variables

Collinearity of the three outcome measures of the consideration process—the consideration set size, session duration, and geographic dispersion—poses a potential concern. To ensure that our results are not driven by this potential collinearity, we first check the correlation between these dependent variables. As can be observed in Table 1.6, the pairwise correlations between Consideration Set Size (*Unique*), Session Duration (*Duration*), and Geographic Dispersion (*Max_distance*) are low at 0.40, 0.19, and 0.37 respectively.

Furthermore, to ascertain that our results are robust to the correlations between these outcome variables, we set up a seemingly unrelated regression (SUR) model. We report these results in Table A3 in the appendix. The results are consistent with our main findings, thereby alleviating this concern.

#### 1.4.3.5 Including Market Size as a Control

The market size, i.e., the number of options, can have a bearing on consideration set measures such as the consideration set size, session duration, and geographic dispersion. Although in our analyses, we include individual market dummies to account for any time invariant unobservable market characteristics, it is still possible that results for moderating influence of WOM abundance may be driven by the market size. To test for the robustness of our results, we devise two control variables to measure the size of the market. The first variable, *Doctor_count*, represents the number of doctors in a given market. The second variable, *Practice_doctor_count*, represents the number of doctor-clinic combinations. We present the summary statistics of these variables in Table A4 in the appendix. To set up this test, we include these market size controls and drop the market dummies since they are perfectly correlated. Furthermore, to account for time invariant characteristics of each of the cities and specialties in the dataset, we include dummies for cities and specialties. We present these results in Table A5 of the online appendix. As we can observe, the results remain consistent.

#### 1.4.3.6 Including Individual Day Dummies as Control

In all the estimation models, we include *Day of the week* dummies, which control common shocks on individual days of the week but do not control common temporal shocks over the days. To control for these common shocks, we include day dummies in the model, with each dummy indicating a specific day. We present the results of this analysis in Table A6 in the online appendix. As we can observe, the results remain consistent.

#### 1.4.3.7 Excluding Small Consideration Sets from Analysis

Some users may conduct a targeted search for a predetermined choice. The behavior of these users is unlikely to be influenced by the presence of WOM. In preprocessing our dataset, we exclude sessions in which users directly open the doctor's profile page, omitting the search stage. However, it is still possible that some sessions represent the search for a predetermined choice. To validate that our results are not driven by the browsing behavior of these users, we therefore test for robustness of our results by excluding sessions in which the users consider only one doctor. We present these results in Table A7. We find that the results remain consistent.

#### 1.4.3.8 Overlapping Markets

In the main analyses where we estimate the effect of WOM on consumer decision-making and on appointment bookings, the models assume that markets are non-overlapping. This assumption is reasonable because in most parts of India, it is unlikely that patients will travel outside their own cities for clinical appointments. However, it is possible for markets to overlap when they are not so distant from each other. To examine the robustness of our results in such a scenario, we address the possibility of patients traveling to a neighboring city by grouping markets in the neighboring region into a single market cluster. We find two such geographic clusters: the National Capital Region cluster (consists of New Delhi, Gurgaon, Faridabad, Noida, and Ghaziabad) and the Mumbai metropolitan area cluster (consists of Mumbai, Thane, and Navi Mumbai). We re-estimate the models with this modified market definition and find the results to be consistent.

#### 1.4.3.9 Including Only Metropolitan Cities

In the main analyses where we estimate the effect of WOM on consumer decision making, the models assume that sparse, moderate and dense markets are in cities of similar sizes. However, in the data, we observe that sparse markets are more likely to be formed in smaller cities and dense markets are more likely to be formed in larger cities. This variation in the underlying geographic area can be a potential alternative explanation for our results. To rule this out, we set up a robustness check in which we only include the sessions that browse doctors in one of the six metropolitan cities of India. We then re-estimate the models with this revised sample and find the results to be consistent.

#### 1.4.3.10 Excluding Ties

In the demand analysis, we use the field rank when ranking doctors by the number of recommendations. For any given doctor, the field rank is 1 + the number of doctors that have a higher number of recommendations. Subsequently, the allocation of doctors into subgroups is based on percentile cutoffs (0.33 for *Top*, 0.66 for *Middle*). We believe this ranking method to be effective in sorting the doctors into top, middle and bottom subgroups, since this considers the market size and assigns doctors to the top subgroup more selectively. However, these ties can raise a couple of concerns. First, the ties mean that when multiple doctors have the same number of recommendations, they will have the same rank and can then all be placed in the same subgroup. Second, this skews the number of doctors in the respective subgroups (i.e., *Top* has 21%, *Middle* has 37%, and *Bottom* has 41%).

To ensure that our results for demand analysis are not driven by these tied doctors, we conduct a robustness check that excludes doctors from markets in which all doctors in that respective market are tied for the same rank. The results from running the model on this sample are shown in Column 8 of Table 1.14. The results are consistent.

Furthermore, to alleviate the concern that unequal sample sizes across each group may be driving demand, we run additional analysis on a smaller sample comprising a randomly selected equal number of observations from each subgroup. The results are presented in the appendix in Table A17. The results are consistent.

#### 1.4.3.11 Verification of the Parallel Trends Assumption

To ensure the internal validity of the difference-in-differences models, it is essential to assert that the difference between the treatment and control groups is constant over time in the pre-intervention period. To test this validity, we first divide the pre-period into subgroups based on the date of the web session. Next, we encode a dummy for each of these subgroups and interact these dummies with our treatment dummy variables (*Spa, Mod, Den*). Then we include terms for post-period dummy interacted with our treatment dummy variables. Finally, we add the market dummies and day of the week dummies to obtain the full model. We present these results in Columns 1-3 of Table A8 in the online appendix, for consideration set size, session duration, and geographic dispersion respectively. As we can observe, most of the coefficients of pre-period interaction variables with treatment dummies are statistically insignificant. Furthermore, we obtain pairwise differences to ascertain statistical difference between these interaction coefficients. The pairwise comparisons reveal that the difference between the coefficient pairs (of the same group) is statistically insignificant, suggesting that the difference between treatment and control groups in the pre-intervention is broadly constant.

#### 1.4.3.12 Falsification Tests and Alternative Specifications

To further establish whether the effects that we measure are indeed caused by the exogenous shock of the visibility of online WOM, we conduct two falsification tests. First, the online WOM should not have a major influence on the offline appointments during the short period of observation. To test for this, we obtained additional data from the platform on the offline appointment details of these clinics. We then set up our model in the same way as in Section 1.4.2, except that we test for walk-in appointments. As reported in Table A9 of the online appendix, we find that the results are statistically insignificant for all the variables of interest. This helps assuage any concerns regarding the potential for strategic behavior by doctors.

We conduct a second falsification test by defining the pre- and post-periods in a synthetic way. Specifically, we define June 19 and 20 as the pre-period and June 23 and 24 as the post-period, then perform all the major analyses. Since the actual visibility occurred on June 25, if we do detect similar effects in this test, it means that our findings are likely driven by some underlying trends rather than the visibility of the WOM. The results are reported in Table A10 of the online appendix. No significant trends are detected, which increases the confidence that our previous findings are indeed driven by WOM. In Table A10, we notice that the sample size is much smaller, since the falsification test is run on a smaller window. This drop in sample size inherently makes it less likely that we will find statistical significance. In such a scenario, to confirm the validity of falsification test, we re-run the main results on a randomly selected sample of the same size. These results are presented in Table A11. The findings are consistent with our main results, thus reaffirming the validity of the falsification test.

Furthernore, to ensure that the results are not driven by specification of bins (for both the consideration set analysis and demand analysis), we reconduct the analysis by respecifying the bins. In the main analysis that investigates the influence of WOM on the consideration set, we define the *Spa*, *Mod* and *Den* markets to be those with 1-2, 3-9, and 10 or more doctors with visible recommendations, respectively. We reconduct the analysis with an alternate specification in which *Spa*, *Mod* and *Den* markets are defined to be those with 1-3, 4-8, and 9 or more doctors with visible recommendations, respectively. These results are presented in Tables A12-15 in the online appendix. As observed, the results are consistent.

Finally, we set up a linear model to illustrate the influence of WOM on demand. This model specifies the WOM as a linear variable (the number of visible recommendations) instead of a categorical variable. Table A16 in the appendix presents these results. We observe a linear relationship between the continuous meaure of WOM (the number of visible online recommendations) and demand (appointments booked online).

## 1.5 Discussion and Conclusion

In this study, we examine the impacts of online WOM on three key elements of consumers' choice process in the context of healthcare services: the consideration set size, the session duration, and geographic dispersion of the options considered. We find that the effects of online WOM on these elements are contingent on the abundance of online WOM. When the abundance of WOM is low, the availability of online WOM results in an increase in the number of choices considered and the session duration. Furthermore, consumers cast a wider net in regard to the geographic location of choices in the consideration set. These findings are likely driven by the fact that WOM transforms intangible experience attributes into quantifiable search attributes (Hong et al. 2012), reducing the cognitive cost of processing information on product quality. Thus the presence of WOM for a few alternatives (low WOM abundance) may persuade users to pursue the alternatives that she may have otherwise ignored and to consider them even if they are far away.

In contrast, we observe that when WOM is available for many alternatives (high WOM abundance), the availability of online WOM results in a decrease in the size of the consideration set and the duration of the web session and in less dispersion in the geographic area of the considered alternatives. We argue that this happens because when WOM is available for many doctors, patients trust this signal of quality and use this signal to screen choices, therefore reaching a satisficing level of utility sooner (Caplin et al. 2011). Hence, this results in smaller consideration sets and shorter session durations. In addition, our analysis of the average time taken to view a profile suggests that patients trust this signal of quality more when it is available for many alternatives. Furthermore, the availability of online WOM recommendations for many doctors enables consumers to find good options nearby, resulting in geographically condensed consideration sets.

Finally, our results demonstrate that the online WOM of doctors has a significant impact on their appointments. The findings also point to a cannibalization effect from the platform's perspective: doctors with more WOM gain demand while those with no WOM experience an adverse effect. This could be quite crucial for platforms hosting products with constraints on the supply side, as this concentration of demand for top-rated providers could lead to unfulfilled demand.

Prior to discussing the implications of these findings, we acknowledge the limitations of our study. First, the time duration of our observations is short. Although this helps to control for temporal and other environmental changes, it does not enable a detailed examination of the long-term effects of WOM. Second, the assignment of WOM is non-randomized, and any doctor who accumulates more recommendations than the threshold limit has their recommendations visible. In this regard, the setting resembles the organic growth of WOM. We note the visibility of recommendations is sudden, thus providing a unique opportunity to conduct *ex-ante* and *ex-post* comparisons of consumer decision making and consumer demand and ascertain the changes caused by the availability of WOM. Third, the measure of geographic dispersion (maximum distance) is only available for about ¼ of the sample and can only be computed for sessions in which the users have browsed two or more doctors (in distinct clinics). Therefore, this prevents us from understanding the impact of WOM on travel distance for users who only consider one doctor. Fourth, although the rankings of doctors not take the WOM into account, the rankings do factor in the number of appointments booked in the past one month. Thus, ranking is indirectly affected by WOM. Fifth, patient privacy regulations prohibit us from obtaining any measures of user characteristics, such as gender, age, loyalty to the provider, etc. This limits our ability to explore heterogeneous treatment effects or introduce additional moderators. Sixth, the platform makes the WOM available in the form of positive recommendations, analogous to “likes.” While this creates a clean setting (as any textual reviews would have resulted in a cognitive cost that may have dominated the decision-making process), the information content is parsimonious. It is possible that in other contexts where WOM information is in the form of detailed narratives such as reviews on e-commerce sites like Amazon, the influence of WOM on consumer decision-making variables may be different. Therefore the findings are best applicable to contexts where WOM is presented in the form of simple undetailed metrics such as the count of “votes,” “likes,” and “recommendations,” such as business pages on Facebook.

Seventh, because the platform only makes positive WOM information available (in the form of the count of positive patient recommendations), it may make it difficult for a user to differentiate between a low-rated and an unrated physician. This may overestimate our results regarding the depreciation in demand for the unrated physicians. Finally, as mentioned earlier, the platform only caters to 1/3 of the active practicing doctors in the entire Indian market, and there are other avenues (online and offline), such as hospital websites and phone books, that patients on the platform may also be using to search and book appointments. Since we do not observe these avenues, changes to the bulk of the market are not observed.

Our findings offer important implications for practice from the perspective of both the platform enabling transactions and the service provider (i.e., doctors). The variation in the consideration set size and consideration time with variation in WOM abundance suggests that platforms could adapt the proportion of products with WOM in search results displayed up front to engage the customers differently. If the platform wants to engage customers to view more products and spend more time, then it can keep the WOM abundance low. This can create a long tail effect and increase the chances of selling products with no WOM. On the other hand, if the platform wants users to decide quickly, then it can present the search results such that the WOM abundance is high. This can create a blockbuster effect and lead to a few superstar products. This follows from Brynjolfsson et al. (2011), who conclude that consumers' usage of discovery tools such as recommendation engines can lead to a shift in the distribution of product sales (as well as a shift in the preceding elements of the consumer choice process). As noted, this may potentially be undesirable for platforms with products that have supply side constraints.

Furthermore, the influence of WOM on the geographic dispersion of the choices considered suggests that businesses with strong WOM could choose more cost-effective locations and rely on online reputation to garner positive results. Our findings suggest that presence of online rating platforms can lower the importance of geographic distance and users may willing to travel farther to consult a higher quality provider.

Several significant opportunities exist to build on this research. We examine the impact of one signal of product quality (i.e., WOM) on consumer decision making, among many other signals such as years of experience, affiliation with hospitals, and educational background. The healthcare market is characterized by significant and persistent information asymmetry regarding doctor quality. Much work remains to fully address recent calls for transparency in healthcare in the policy discourse in the United States. As patients increasingly use online information sources about signals of quality—both user-generated signals and information from other sources—it is critical to understand the effects of such evidence on patient decision making (Austin et al. 2015, Duan et al. 2008). To the extent that WOM information may conflict with other quality signals, examining the interaction between complementary and conflicting sources of information represents a fruitful avenue for further work. A second promising direction for future research is to examine how patient characteristics such as demographics interact with the available quality signals of WOM. While we were unable to include these analyses within the scope of this study, it is likely that individual attributes (such as age, gender, nationality, and culture) affect the impact of WOM information on consumer decision making, and this relationship warrants further investigation. Third, future research could examine the moderating impact of WOM quality, reflected, for example, in the length and thoughtfulness of textual reviews, or on the relationship between WOM and aspects of consumer decision making. Finally, future research could utilize detailed clickstream data, as we have done, to link browsing behavior and the final choice of a service provider. Such an analysis would help reveal the granular effects of WOM on the relationship between the consumer consideration process and a “sale.”

# CHAPTER 2: CATCH ME IF YOU CAN—DETECTING FRAUDULENT ONLINE REVIEWS OF DOCTORS USING DEEP LEARNING

## 2.1 Introduction

Online reviews are becoming increasingly prevalent and are exerting a more significant influence on purchase decisions (Clemons et al. 2006, Duan et al. 2008, Forman et al. 2008, Gu et al. 2012, Ott et al. 2011, Sun 2012). Recent consumer reports show that 78% of users look for reviews before making online purchases, and online reviews are now as much as seven times more powerful than advertisements in influencing consumer decisions (AMA and WOMMA 2014). The rising economic influence of these reviews creates strong incentives for businesses to post fraudulent reviews to promote themselves and discredit their competitors (Luca and Zervas 2016). Self-promotion in online reviews was first detected as early as 2004, when Amazon unmasked considerable self-reviewing by book authors (Smith 2004). Further, in 2011, it was reported that businesses were hiring workers on Amazon Mechanical Turk to post fraudulent five-star Yelp reviews on their behalf (Segal 2011).

The issue of fraudulent reviews has not only attracted the attention of researchers and affected platforms; regulatory authorities have also recognized the gravity of the problem and have begun to step in. For example, in 2013, the Attorney General of New York State fined companies $350,000 for faking their online reviews (Attorney General Report 2013). Fraudulent reviews have grown to be a serious concern and have affected the credibility of many review platforms by eroding consumer trust (Charman-Anderson 2012, Streitfeld 2012). Therefore, there is an urgent need to advance the methods of identifying fraudulent reviews. The methods proposed so far in the existing literature on fraudulent review detection can be classified into behavioral and linguistic approaches, as summarized in Table 2.1. Behavioral approaches include finding distributional anomalies of ratings and observing unusually correlated patterns in the review metadata. For example, such patterns could reveal collusion among a group of reviewers by checking whether these reviewers are systematically writing positive reviews for a group of products and/or negative reviews for a group of competing products (Mukherjee et al. 2012, Mukherjee et al. 2011). As another example, behavioral approaches can check whether a product has an abnormally high number of reviews from first-timers (reviewers who have never posted a review before) and singletons (reviewers who have posted just one review) (Lim et al. 2010); all such reviews can then be flagged as suspicious.

So far, behavioral approaches are widely used in industry practice and have reported higher accuracy than linguistic approaches. However, the use of behavioral approaches specifically for the detection of fraud has a fundamental vulnerability. By definition, behavioral approaches attempt to decipher fraudulent behavior by converging on a pattern of behavior that could be considered an outlier. Skilled fraudsters understand that such deviation from normal behavior makes them stand out as suspicious, and over time, they learn to closely emulate the behavior of genuine reviewers to ensure that their reviews are not filtered out, thereby gaming the system.

Alternatively, linguistic approaches detect fraudulent reviews by focusing on the differences in the language patterns of genuine and fraudulent reviews. Research has shown that humans unconsciously use different language patterns when describing real vs. imaginary events, a tendency that can be utilized by machines to detect deception automatically (Mihalcea and Strapparava 2009). For example, studies have shown that when writing fake reviews, people tend to use past tense and superlatives more frequently (Ott et al. 2011). Since the linguistic approaches identify fraudulent reviews by looking for linguistic traits that imposters unconsciously employ, they are more robust against gaming. However, linguistic traits that distinguish a fraudulent review from an authentic review are very subtle. Although word-dictionary-based linguistic approaches appear promising when tested with simulated datasets (Ott et al. 2011), researchers have since observed that dictionary approaches are considerably less effective with real-world data (Mukherjee et al. 2013). This difficulty in detecting specific linguistic traits is the primary limitation to the effectiveness of linguistic approaches.

In this study, we take a significant step forward in using linguistic approaches to detect fraudulent reviews. Rather than explicitly examining a limited set of the linguistic features of reviews, such as by using word dictionaries, we design a deep learning approach that takes into account all the words in the reviews and the relationships among them to detect fraudulent reviews. Recently, machine learning in general and neural network models in particular have yielded promising results in the area of learning semantic representations for tasks related to natural language processing (Le and Mikolov 2014, Tang et al. 2015). We apply deep neural networks to the fraudulent review problem. We believe that deep neural networks can be advantageously applied by using layers of neural networks that automatically combine multiple features to capture the semantic information embedded in the text of reviews. Such representation of multiple combinations is almost impossible to achieve using any linear model that manually selects features.

We tested our algorithm on a dataset of over 181,951 doctor reviews obtained from a platform based in India. We find that our algorithm performs significantly better than traditional Natural Language Processing (NLP) and machine learning algorithms. The traditional machine learning algorithms have comparatively low accuracy in identifying the fraudulent reviews and their Area Under the Curve (AUC) scores are under 0.66, whereas our bidirectional recurrent neural network based algorithm achieves an AUC score of 0.72. We further examined how the deep learning approach compares to 200 human evaluators. Interestingly, humans perform much worse in terms of specificity and sensitivity. In fact, our deep learning model has a true positive rate that is almost twice as high (14.29% vs 8.70%) and also achieves a much lower false positive rate (0.63% vs 11.68%). We find consistent evidence that the chance that a human evaluator will label a fake review as genuine is even higher than the chance that they will label a genuine review as genuine.

Our study is the first to apply deep learning methods to classify online reviews and show a significant improvement in accuracy when compared to traditional machine learning algorithms. Furthermore, unlike studies that use simulated data (Ott et al. 2011), our study utilizes real-world data with labels, which was enabled by inadvertent security loopholes that allowed providers to post reviews for themselves on behalf of their customers. Our labeling technique is exogenous to the techniques used in the classification model, which ensures that the estimated accuracy is close to real-world accuracy. Finally, to the best of our knowledge, ours is the first study that investigates fraudulent reviews in the context of a service of great social consequence: healthcare providers.

## 2.2 Background

### 2.2.1 The Rise of Fraudulent Reviews

Online word-of-mouth (WOM) is essential for businesses in many ways. First, as a quality signal it is highly influential in driving sales and attracting new customers, although it may be biased (Hu et al. 2017). Second, studies have shown that positive online WOM motivates customer loyalty and drives community engagement (Gauri et al. 2008). Third, negative online WOM has an exceptionally detrimental impact on a firm’s cash flows, profitability, and long-term stock prices (Luo 2009). Therefore, given the influence of online WOM on a firm's well-being, it is imperative that firms maintain positive online ratings and reviews (types of online WOM).

However, consistently getting positive ratings and reviews can be very costly as it requires firms to meet or even exceed all or most customer expectations. Unscrupulous businesses instead try to portray their products and services in a positive light by implanting false narratives (online reviews) of customer experiences of those products and services. Typically, such firms either ask their employees to create the fictitious online reviews or employ the services of third-party reputation management firms that create and post such content.

Such implanted product ratings are a major concern for consumer protection groups and regulatory authorities as customers are misled by the false ratings to purchase goods and services that can be of questionable quality, thereby negatively impacting overall customer welfare. Furthermore, such fictitious ratings erode customer trust in platforms that host them. Studies indicate that about 57% of customers are suspicious of online sellers that only have positive reviews, and about 49% of customers believe that sellers use unfair means, such as providing customers monetary incentives for reviews, making the platforms complicit for allowing such practices (Mintel Research 2015). Fraudulent reviews are quite widespread: recent data suggests that about 16% of reviews posted on large online review platforms such as Yelp are fake (Luca and Zervas 2016). These fake reviews are written not just by small firms but also by large, reputable firms. For example, the Taiwan Fair Trade Commission caught Samsung hiring bloggers and students to write reviews that falsely praised that company's products and criticized those of its competitors (FTC Report 2013).

### 2.2.2 Detecting Fraudulent Reviews Using Behavioral Characteristics

Taking cognizance of this emerging issue, a newly evolved stream of research has proposed algorithms to detect fraudulent reviews (Fei 2013, Jindal and Liu 2007a, 2008, Lim et al. 2010, Mukherjee et al. 2012, Mukherjee et al. 2011, Mukherjee et al. 2013). As previously mentioned, the algorithms that detect fraudulent reviews can be classified into two broad categories: those that rely on behavioral characteristics and those that rely on linguistic characteristics.

Algorithms that rely on behavioral characteristics examine the deviation from norms regarding observable characteristics such as the time a review was posted, geolocation of the reviewer's IP address, and the reviewer's past purchases and reviews to detect deception. As summarized in Table B1, the most prevalent methods include tracking abnormal review frequency based on reviewer user ID, browser cookie ID and IP address, deviation in ratings, anomalies in geolocation, suspicious votes that mark the reviews as helpful, and the occurrence of recently received negative reviews as indicators of fraud. Additionally, review filters check whether the reviews for a product are coming from only new users, are outliers in terms of the time taken to write the review, etc. (Jindal and Liu 2007a, 2007b, 2008, Lim et al. 2010, Mukherjee et al. 2012, Mukherjee et al. 2011). All these behavioral features represent a deviation from normal behavior.

While such behavioral deviations may correspond to fraudulent activity, skilled imposters can bypass these review filters by closely emulating the observable behavior of genuine users. A highly-paid, resourceful fraudster can easily spoof an IP address or geolocation, post reviews at irregular time intervals, etc. In other words, he or she can mimic the behavior of a typical user and trick the detection algorithm. This becomes more likely when the stakes are high and the volume of fraudulent reviews required to sway consumer opinion is low, such as in healthcare. Unfortunately, the existing methods of fraud detection that utilize the characteristics of reviewer behavior are not resilient against such gaming. This limits the effectiveness of algorithms that utilize behavioral characteristics.

### 2.2.3 Detecting Fraudulent Reviews Using Linguistic Characteristics

Fraud detection algorithms that rely on linguistic characteristics use the text of reviews to detect deception. Jindal and Liu (2008), for example, discovered fraudulent reviews by identifying similarities among reviews using bigrams and trigrams (groups of two and three words). While the approach of using content similarities as clues to deception is potentially useful, it must be noted that fraudsters also understand that plagiarizing the content of other reviews is a shortcut that is more likely to be flagged by the review filters, especially in markets where the volume of reviews is low. Additionally, customers are likely to doubt the authenticity of reviews when they come across similar content and similar phrases across multiple reviews, which would defeat the very purpose of implanting fake reviews. Therefore, in markets where the incentives are high for obtaining a favorable online reputation, imposters are likely to generate new content instead of plagiarizing existing content, thereby making the fraudulent reviews difficult to detect.

Following a more sophisticated approach, Ott et al. (2011) used a psycholinguistic deception detection methodology based on a linguistic inquiry and word count (LIWC) toolkit to derive features (words) in the reviews to predict fraudulent behavior. Their work synthesized fake reviews by paying freelancers to write about 400 imaginary reviews and collating real reviews by scraping a popular travel site. While their study reported very promising accuracy in classifying authentic and fake reviews, the replication of their approach has shown little promise on real-world data (Mukherjee et al. 2013). Furthermore, their dataset is small. It consists of about 400 fraudulent reviews and 400 genuine reviews, making it difficult to apply any of the supervised machine-learning techniques that might go beyond word dictionaries.

However, theoretically, linguistic traits hold substantial promise and should be resilient to gaming. In their work on memory and perception, Johnson and Raye (1981), noted that truthful opinions are more likely to include sensory information and details about spatial configurations. Recent work by Vrij et al. (2007) also suggests that fraudsters have considerable difficulty in including spatial details.

Conclusively, these linguistic approaches do have promise, as they focus on finding language patterns that can explicitly indicate deception. However, despite the promising results shown by algorithms that exclusively use sophisticated linguistic traits on small synthesized datasets, they have not demonstrated high accuracy within large real-world datasets. Our study aims to fill this knowledge gap by exploring new methodologies and considering text analysis methods that would use all the words in the text of the review along with the relationships between the words, and we are able to achieve high accuracy using real world data.

### 2.2.4 Labeling Fraudulent Reviews

Labeled data refers to samples of reviews that have been categorized as genuine or fraudulent. Labels are required to provide a benchmark that is used not only to train an algorithm to detect fraudulent reviews, but also against which to compare the results of the algorithm to determine its accuracy. Previous studies on fraudulent review detection have mainly employed three methods of labeling the data: using human evaluators, using a pre-labeled dataset that utilizes behavioral filters, or employing humans to fabricate reviews and thus creating a labeled repository of fraudulent reviews.

The first and the most common method of data labeling involves human evaluation. For example, Xie et al. (2012) employed human evaluators to read reviews from 53 stores and make decisions on the suspiciousness of the reviews of each of these stores. They then benchmarked their detection algorithm against the human evaluations. Along similar lines, Mukherjee et al. (2012) employed human judges to label groups of deceptive reviewers. The principal limitation in using human labeling as a proxy for distinguishing between fraudulent and genuine reviews is the low accuracy of human evaluation (Xie et al. 2012).

The second most common method of data labeling is to use available datasets of labeled reviews from leading review platforms such as Yelp (Mukherjee et al. 2013). Although the presence of a large and publicly available labeled dataset—such as the one from Yelp—helps significantly in solving the labeling problem, the underlying filtering techniques used by such websites to label data remain imperfect. In some instances, filtered reviews may not be fraudulent, while in others, fraudulent reviews may not be filtered out by the algorithm. Thus, the filtering algorithm could overestimate or underestimate the number of fraudulent reviews. An underestimation would occur if fraudulent reviews were not filtered. Alternatively, an overestimation would occur if honest reviews were falsely labeled as fraudulent and filtered accordingly (Jabr 2015, Luca and Zervas 2016). Furthermore, platforms that provide pre-labeled datasets (such as Yelp) also use algorithms or heuristics to determine the authenticity of reviews. Therefore, there is a lack of ground truth to ascertain the accuracy of the labels themselves.

The third method used to obtain labeled datasets is to scrape genuine published reviews from major online review portals and employ freelancers to write fraudulent reviews. Ott et al. (2011), for example, acquired labeled genuine reviews by scraping a travel website, assuming all published reviews on that site to be genuine. They then obtained fake reviews by instructing freelancers from Amazon Mechanical Turk (AMT) to pretend to be employees of hotels whose managers wanted them to write fraudulent reviews to post online. In this manner, the study created a labeled dataset of genuine reviews (obtained from the website) and fraudulent reviews (written by freelancers). Although the researchers instructed the freelancers to sound as realistic as possible in their pretense, the reviews generated by the freelancers were likely to be dissimilar to actual fraudulent reviews posted by professional spammers in the real world. Due to the difference in abilities, level of economic affluence, motivations, and psychological states between the hired freelancers and the real fraudsters, the real fraudsters are able to fake customer reviews better.

The most unaddressed limitation across all previous studies, irrespective of the type of technique used to detect fraudulent reviews, is the lack of independence between labeling and detection methods, making it impossible to ascertain the real accuracy of the detection method. Below, we discuss how each of these labeling methods overlaps with the detection method employed in the respective studies.

First, studies that use human evaluators to label the data employ detection algorithms that are similar in principle to the approach used by human evaluators to label the reviews. For example, Lim et al. (2010) labeled their data by showing the individual reviews, along with all reviews provided by each reviewer, to human evaluators; they then asked them to judge the authenticity of the reviews. When human evaluators do this, they look for commonalities across reviews by the same reviewer. They also look for traits such as repeated words and phrases and other signs of deception. The detection algorithm used in their study takes a similar approach by utilizing bigrams and trigrams to detect fraudulent reviews. This apparent similarity in the detection and labeling method makes it difficult to judge the accuracy of the detection algorithm as, in essence, it is an automation of a human process. In another example, Mukherjee et al. (2012) provided human evaluators with review content and asked them to label the reviews as fraudulent or genuine. Their detection algorithm uses features such as group content similarity to detect fraud, which is very similar to the method used by human judges to distinguish between genuine and fraudulent reviews.

Similarly, studies that use pre-labeled datasets from leading websites also suffer from a lack of independence between the labeling method and the detection algorithm. To illustrate, in the Yelp dataset used by researchers, signs of deceptive language provide one basis for distinguishing between fraudulent and genuine reviews (Mukherjee et al. 2013). Studies such as those by Feng et al. (2012) and Mukherjee et al. (2013) use syntactic style information (i.e., grammatical similarity) in their fraudulent review detection algorithm. If the data-labeling method involving human judges also implicitly relies on the grammar of the reviews, then this detection method results in a lack of independence between the labeling method and the detection algorithm, which could create an upward bias to the performance of the model. Similarly, Li (2014) used labeled data from a leading restaurant ratings website in China. The detection algorithm in the study examines metadata patterns (such as the geolocation IP addresses) to detect fraudulent reviews. This detection algorithm overlaps with how the portal itself labels the data and is therefore not independent.

Studies that label data by employing humans to create fraudulent reviews also suffer from the limitation of non-independence between the data-labeling method and the fraud-detection method because there are fundamental differences between the users who are paid to create the fraudulent reviews and genuine reviewers. For example, Ott et al. (2011) created a labeled dataset by assuming that all reviews posted on a travel website were genuine; they then employed freelancers from Amazon Mechanical Turk to write fraudulent reviews. Subsequently, the authors trained their fraudulent review detection algorithm by using a linguistic approach that distinguished between genuine and fraudulent reviews based on differences in vocabulary. However, as explained earlier, the vocabulary in fraudulent reviews generated by freelancers is expected to differ from that of reviews posted by genuine reviewers because of fundamental differences in their abilities, level of economic affluence, motivations, and states of mind. Furthermore, since the freelancers are asked to write a favorable review, those reviews are likely to contain a high number of superlatives. The use of superlatives is, however, a factor utilized by the researchers in differentiating the genuine from fake reviews. Hence, the labeling method and detection algorithm are not entirely independent.

### 2.2.5 Healthcare Market Idiosyncrasies

Our study focuses on reviews that pertain to healthcare providers. Our approach, in addition to addressing the limitations of existing methods as described above, takes into consideration idiosyncrasies of the healthcare market that not only make fake reviews a serious societal concern but make detection an even more challenging problem. The structure of the healthcare market is different from that of typical consumer goods markets, and user privacy in the review content is a more substantial concern.

#### 2.2.5.1 Market Structure

The healthcare market structure differs from that of consumer goods markets in several ways. First, most patients consult their service providers very infrequently. Statistics suggest that the average U.S. citizen makes four doctor visits per year (McCarthy 2014). This low frequency reduces the engagement patients have with the platforms that host doctor reviews, which in turn leads to fewer ratings for each doctor. For example, on RateMDs.com, a popular doctor rating website, the average number of ratings per doctor is 3.2, and about half of the doctors have only one rating. In contrast, on Yelp, the average restaurant has well over 40 ratings. Therefore, the number of fraudulent reviews needed to bias consumer perception is meager—as low as a single review—due to the low average number of total reviews accumulated by a healthcare provider. Also, since patients visit the portals that host doctor reviews infrequently, it is challenging for researchers to use behavioral techniques to detect fraudulent reviewers due to the scarcity of data on browsing behavior.

The second significant difference between the healthcare market and the rest of the consumer goods market is that the healthcare market can be classified as a credence goods market. Therefore, the consumers of healthcare (patients) are never sure of the extent of the service needed (e.g., medication) or about the true competence of the provider. The provider acts as the expert in determining the requirements of the customer, and the customer can only judge healthcare services by the experience component of the service rather than the credence-based component. For example, patients tend to assess a surgeon's surgical skills based on the bedside manner they observe. This lack of certainty about the quality of the service provided to them makes patients less likely to share opinions publicly. Also, there is a non-trivial chance that the doctor might infer which patient writes the review, as the description in the review tends to reflect the detail of the interactions. These factors further reduce the number of reviews obtained by healthcare providers and make it more difficult to ascertain the authenticity of posted reviews.

Third, from the perspective of the provider, the healthcare market offers limited means to signal one's quality. For example, a doctor cannot advertise his or her services or promote his or her quality as easily as a restaurant or any other less regulated provider of goods or services. All over the world, boards of ethics forbid doctors from drawing attention to their professional position, skills, achievements, attainments, appointments, associations, affiliations, or honors in a way that could potentially be self-aggrandizing. For a doctor, the primary signal of quality is word-of-mouth information spread by patients, and as building a reputation in the healthcare industry takes considerable time and effort (due to the nature of the services provided and the relatively low volume), there are greater incentives in this market to create a fake reputation.

Finally, in the healthcare industry, the "ticket" size (the average revenue for a transaction) is very high. In the U.S., a general physician gains about $90 from a single consultation, and surgeons earn thousands of dollars from an individual patient. In comparison with the prices of consumer goods such as hotels and restaurants, the average revenue per customer is much higher; as a result, the incentive to add fraudulent reviews is higher, which has the knock-on effect of increasing the sophistication of the agents employed to create such fraudulent reviews.

Overall, these factors make the context of reviews of credence goods such as healthcare services unique, and it becomes challenging to extend to this context the applicability of existing methods of fraudulent review detection.

#### 2.2.5.2 Differences in Review Content and Concerns About User Privacy

In the healthcare context, the content of reviews differs significantly from that of reviews for consumer goods in two important ways. First, unlike reviews of other products and services, reviews of doctors may contain hyperbole and intense emotion because healthcare is so personal that patients often show immense gratitude or intense frustration toward their healthcare providers. This characteristic renders the application of existing linguistic techniques that utilize intense emotion as a signal of inauthentic reviews less useful.

Second, there are dimensions of information in a healthcare review that are unique to the patient and unverifiable. These aspects correspond to information about the reviewer and information related to the actual consultation. For example, to increase the influence of one's review, a reviewer might explain that he or she is a severely afflicted patient who has a longstanding experience of the described medical condition. The reviewer might go on to present a detailed narration of his or her consultation experience with the doctor. While in the consumer goods market, platforms can take certain steps to verify the credibility of the information and discourage false propaganda, in the healthcare context, any such verification steps are next to impossible due to privacy concerns and regulations.

Third, patient privacy concerns sharply reduce the number of options available to platforms to verify a review's authenticity. For example, patients are not comfortable using any single sign-in service, such as Google, Facebook, or Yahoo, to sign in to patient portals and doctor review sites due to fear of discrimination. Additionally, patients prefer anonymity, reducing the possible means of verification (such as tele-verification of the review content). Finally, regulatory frameworks prohibit the tracking of patients across the internet and forbid platforms that host reviews from requesting documentation from patients, such as medical records, to verify their claims.

These factors make it a challenge not only to apply methods of existing fraudulent review detection but also to develop new methods that would require data about the patient's health condition.

## 2.3 Data

Data source is a leading online doctor search platform based in India. I spent substantial time with the development team of this platform to adequately understand in detail the system architecture and the codebase. The platform has listings for over one-third of all private doctors in the major cities of the country, allowing patients to search and book appointments with doctors, and it subsequently requests that they provide feedback (a yes/no recommendation followed by a text review) about their consultation. Furthermore, it also accepts feedback from patients who have not booked an appointment through the portal.

Listed on the platform are doctors of many specialties present in the major cities of India. Each market is characterized by a combination of a city and specialty. To write a review, one must register on the platform using a cell phone number (mandatory) and an email address (optional). The platform neither allows a patient to post more than one review for a given doctor, nor does it permit a patient to post more than one review for a given market in a month.

### 2.3.1 Labeling of Reviews

The platform has several channels through which patients can send in their reviews: a verified email channel, a verified phone channel, or an unverified open feedback channel. For the verified email channel, the reviews can be sent in by following the link in the feedback request email that is sent to all patients on the day following their appointment. Similarly, for the verified phone response channel, the reviews can be sent in by following the link received via SMS.

For a limited time, some of these channels contained some unintentional security vulnerabilities that allowed fraudsters to abuse the system by posting a review for any doctor on behalf of a patient. The existence of these bugs created a low-cost mechanism for posting a fraudulent review, hence creating a natural honeypot. This natural honeypot remained unnoticed by the platform administrators for almost a year, which allowed many doctors to exploit it to post fraudulent reviews. We identified the loopholes and alerted the platform. With assistance from the platform, we subsequently conducted a detailed investigation of all the reviews using the available digital forensics data, which included the persistent cookie ID, the credentials used for signing in, the session keys used, etc. This allowed accurate identification of all the published reviews that had exploited security vulnerabilities. We then matched the relevant fields to clinic credentials, hospital employee databases, and diagnostic lab employees' contact details in addition to checking for mobile device ID login records to identify imposters.

We labeled the reviews that had exploited security loopholes as fraudulent and the remainder as genuine. To protect the interests of the platform, we cannot reveal the exact security loopholes in great technical detail, but we can state that these bugs allowed fraudsters to subvert the validation process easily. Hence, there remained little incentive to use higher-cost means of fraud (such as creating a new identity by procuring a new cellular connection). Conclusively, the existence of these security loopholes provided a unique opportunity to create a high-quality labeled dataset of fraudulent and genuine reviews.

### 2.3.2 Data Construction

As a part of the data cleaning process, we took additional safeguarding steps. We omitted the reviews that are duplicates of each other, as it is possible that a fraudster may have posted a copy of an existing review. This omission reduced the number of reviews from 231,940 to 228,429, a 1.50% reduction. Additionally, we dropped the reviews that are less than 50 characters long as the platform, as a policy, requires the reviews to be at least 50 characters long to be posted. However some very small reviews may also get collected due to errors in transmission. Since these are never published, we choose to drop them. This reduced our sample by 36,737 reviews. Furthermore, due to prohibitively high computational time on analyzing very long reviews, we also omitted reviews that are longer than 200 characters. This further excludes 9,741 reviews.

After this cleaning, our dataset consists of 181,951 reviews. We then randomly split the dataset into training data (80%, 145,560 reviews), validating data (10%, 18,195 reviews; 18,100 were used), and testing data (10%, 18,196 reviews; 18,100 were used). Furthermore, we conduct the standard preprocessing on the review text for all the subsequent predictive models. The preprocess is mainly text cleaning and tokenization, removing special characters, and word stemming.

### 2.3.3 Summary Statistics

In our cleaned and preprocessed dataset, we have a total of 181,951 reviews, out of which 14,616 reviews, about 8.03% of the total, are labeled as fraudulent. We present the overall summary statistics in Table 2.2. As observed from Table 2.2, an average review has about 113 characters, is viewed by about 34 people, and is given 0.16 positive helpful votes.

We present the individual summary statistics for fraudulent and genuine reviews in Table 2.3. As shown, the fraudulent reviews are relatively longer, at about 125 characters, compared to genuine reviews, which have about 112 characters. Surprisingly, the number of helpful votes received on average by fraudulent reviews is much higher, at 0.56, whereas for genuine reviews, it is 0.13, suggesting that fraudulent reviews are quite convincing. We note that these differences stated above are statistically significant at the $p<0.001$ level.

## 2.4 Methods

Our approach to detecting fraudulent reviews only makes use of the review text and does not use any of the review metadata such as the review posting behavior of the user, ensuring that our detection method is less prone to being gamed. The methodology utilizes deep learning to automatically construct features from the review text. We aim to examine the potential for deep learning to detect fake reviews by utilizing the subtle differences between the writing styles of fraudulent and genuine reviewers, which are harder to game.

There are two major techniques utilized by our method: deep learning-based word embedding and recurrent neural networks for prediction. Prior work in other fields has shown that combining these two techniques often yields the best accuracy.

### 2.4.1 Word Embedding of Doctor Reviews

The word embedding technique begins by finding an optimized vector to represent an entity (usually a word). The vectorization then enables the computability among words and other factors. Traditionally, NLP embeds a word, a sentence, a text paragraph, or even a document to a vector according to the frequency of occurrence of words. This process only focuses on the occurrence of the word, ignoring the positional information. To overcome this limitation, modern deep learning-based NLP utilizes a unique architecture (one hidden layer) of artificial neural networks to optimize the vectorization and preserve positional information.

There are two options for obtaining word embeddings: researchers may use an existing pre-trained word embedding using a super large corpus of text data of similar nature or generate a word embedding using the local text data itself. Because our context consists of healthcare reviews posted by Indian users, it is unique in that the text includes imperfect grammar and spelling, commonly occurring medical terms, and strong sentiments. Therefore, instead of using any known corpus of words to build the word embeddings, we build our own training model from the review text data to find the optimized embeddings of the words in each of our reviews.

Our deep embedding model utilizes the architecture of Word2Vec (Mikolov et al. 2013) and sets the size of the dimensionality of the embedding vectors as 300, the window size as 50, and ignores words with two occurrences or fewer in the entire corpus. We utilize text from all the reviews (training data, validating data, and testing data) to build the word embedding. While the review text in the training data will serve as input for the prediction, leaking is not permitted.

This training data consists of 37,972 words embedded into 300 dimensional vectors. Subsequently, the embeddings are used as input for the predictive model.

Although our local training data is small compared to well-known pre-trained embeddings on large word corpus, there are certain advantages to using self-constructed embeddings in our context. First, locally trained word embeddings incorporate uncommon terms that occur frequently in the texts of doctor reviews. Second, the specific grammar, languages, styles, etc. of the Indian users can be better reflected in self-constructed embeddings.

Given the huge dataset size of text reviews, many rare words (including typos) can explode the vocabulary size but contribute to minimal incremental increase in performance. Hence, we process the review text further and retain the words that ranked in the top 10,000 by frequency.

### 2.4.2 RNN-Based Predictive Model

We then use a deep recurrent neural network (RNN) to process the sequence of the vectorized input reviews. As compared to traditional neural networks, the recurrent neural networks specialize in memorizing the last state and referring to it for the next sequence of input. RNN uses recurrent connections that utilize the context information very effectively, making it especially useful for problems that involve recognizing any user biometric, such as handwriting, speech and electro-cardiogram patterns (Graves et al. 2013, Pham et al. 2014, Salloum and Kuo 2017). Therefore, we apply RNN to the review classification problem as it also involves identifying traits that could uniquely identify the author of the review.

We present the architecture of this bidirectional gated recurrent unit (GRU) neural network in the schematic listed in Figure 2.1. After the standard data preprocessing, the raw review text (list of words) is fed directly to the model. Leveraging the embeddings done in the previous steps, a review text is then converted to a list of vectors (a matrix). After this embedding, the matrix is sent directly as an input to an RNN layer. Here we use a bidirectional GRU layer with 200 forward GRU cells and 200 backward GRU cells. The outputs of both forward and backward cells are then merged to go through an attention layer (attention size is 400). The output data is then flattened to a two-dimensional matrix. The RNN layer is followed by three fully-connected layers with 2,048, 1,024, and 512 nodes respectively. Finally, the output layer (a single node layer) receives input from the last fully-connected layer of 512 nodes and is then activated by a sigmoid function, which outputs the probability of the review being fraudulent.

To train the model, the Adam optimizer is adopted to optimize the cross-entropy loss. To facilitate the training progress, the states of GRU cells are all initialized to be 0. The dropout technique is also incorporated into the training algorithm, setting the keep rate for training at 0.20.

## 2.5 Results

### 2.5.1 Comparison Against Traditional Machine Learning

Consistent with the approach prevalent in evaluating prediction models, we use AUC as the measure of model performance. As previously suggested, different models of one method are selected using the 10% validating data. Then, the best model of each method is further tested on the testing data. This validating-testing evaluation ensures generalizability of the results. To benchmark our performance, we compare the accuracy of our algorithm against two NLP-based traditional machine learning algorithms: logistic classifier and random forest classifier. To provide a basis for fair comparison, we set up the traditional machine learning algorithms such that same preprocessing strategy is used, including adoption of the same vocabulary set of

10,000 words. Then word occurrence frequency is used to convert text into vectors (10,000 input features, one for each word). To fine-tune the benchmarks, we performed a grid search for logistic regression and random forest to obtain the best hyper parameters. For logistic regression, the following two parameters were tuned via grid search: inverse of regularization strength, which was set as 0.001, 0.01, 0.1, 1, 10, 100, or 1,000, and penalty was set as "l1" or "l2."

For random forest, the following six parameters were tuned in the grid search: the function for split quality measure, which was set to "gini" or "entropy;" maximum depth of the tree, which was set as 3 or at "no limit;" minimum number of samples required for split, which was set to 2, 3, or 10; minimum number of samples required for a leaf node, which was set at 1, 3, or 10; number of features to consider for split, which was set as 1, 3, 10, $\log_2$ of total features, or square root of total features; and bootstrap, set to true or false. In a grid search, each of the combinations of the given values was used to build a model. In total, 14 logistic regression models and 360 random forest models were built and tested. The best performing model of each method was then used to represent the performance of the method.

All the models are then trained and tested using the same training, validating, and testing data respectively. The AUC scores of the three models on testing data were then compared. As reported in Table 2.4, both of the traditional methods performed worse than deep learning in distinguishing fraudulent reviews from genuine ones for all similarity thresholds. The AUC scores of both random forest and logistic classifiers are lower than 0.66, whereas the bidirectional GRU model achieves an AUC score of 0.72, significantly higher than traditional methods. The ROC curves of the Logistic, Random Forest and Deep Learning models are presented in Figure 2.2.

To summarize, detecting fraudulent doctor reviews is quite challenging, often making both human readers and traditional machine learning algorithms ineffective. However, our deep learning approach shows significant gains in performance by incorporating two major state-of-the-art techniques: word embedding and RNN-based deep learning. In our approach, we are able to utilize much more information, such as the sequence of words and the co-occurrence of words in both training and testing data, therefore improving the performance significantly.

### 2.5.2 Human vs. Machine

Besides comparing the accuracy of deep learning vis-à-vis traditional machine learning algorithms, we also compare traditional machine learning and deep learning against human evaluation. To do so, first we randomly select a pool of about 10% (1,900) of the reviews from the training dataset. Then, we hire 200 human evaluators[13], who are each given 30 randomly chosen reviews from this pool. We inform the evaluators that about 8-10% of the reviews are fake[14]. The evaluators then use their prior knowledge from their online browsing and shopping

---

[13] We hire these evaluators from Amazon Mechanical Turk. Since the appointment booking platform is based in India, we hire Indian workers, who are more likely to be familiar with the writing style of Indian users.

[14] As noted in Summary Statistics (Table 2.3), 8.03% of the total reviews are fake. In the random draws of 30 reviews each from the evaluation pool, the proportion of fake reviews varies slightly in each draw but remains within the 8-10% window. We believe that specifying the proportion helps in this activity, as otherwise the evaluators are likely to be far more unsure of the proportion and would then label many more reviews as fake. We re-ran the human evaluation without specifying this proportion, and the results indicate that hiding the proportion of fakes lowered the overall accuracy. We present these results in Column 4 of Table 2.5.

experience to distinguish fake from genuine reviews. After collecting the responses from individual evaluators, we collate the responses at the level of each online review. We then compare the labels provided by these human evaluators against the labels in the dataset in order to compute the accuracy of human labeling.

We present the comparison between human and machine labeling in Table 2.5. As observed in Column 1 of Table 2.5, all machine learning algorithms have a much higher specificity (true negative rate) than human evaluation. For instance, the logit, random forest, and deep learning models have true negative rates of 99.71%, 100% and 99.37% respectively, whereas humans have a true negative rate of 88.32%. In addition, as we can observe, deep learning also has a much higher sensitivity (true positive rate) than human evaluation. Deep learning has a true positive rate of 14.29%, whereas humans have a true positive rate of 8.70%.

Furthermore, all machine learning models have very low false positive rates of less than 1%, whereas humans have a very high false positive rate of 11.68%. Finally, humans have a much lower overall accuracy of 81.57%, which is significantly lower than any of the machine learning algorithms, each of which have an accuracy of 90% or more. Interestingly, we also observe that the false positive rate for humans (11.68%) is much higher than the true positive rate (8.70%), suggesting that the fake reviews seem so genuine that they are less likely to be labeled as fake than the actual genuine reviews. This finding goes hand in hand with summary statistics Table 2.3, in which we note that fake reviews are many times more likely to be labeled as helpful (upvotes) than the genuine ones.

In the event of a tie in the human evaluation of a particular review (i.e., when half of the evaluators labeled the review as genuine and the other half labeled it as fake), we considered the overall label to be genuine. In Column 2 of Table 2.5, we present the results after omitting these tied reviews from human evaluation analysis. Furthermore, in Column 3 of Table 2.5, we present the results after considering the tied reviews to be fake. As we can observe from Columns 2 and 3 in Table 2.5, the results are consistent: machine learning algorithms have significantly higher overall accuracy human evaluation, and humans believe that fake reviews are more authentic than the genuine ones.

### 2.5.3 Examples of Reviews Exclusively Detected by Deep Learning

In addition to validating the greater accuracy of the deep learning model, we explore the nuances in reviews that are only detected by deep learning. To do so, we examine the text of the reviews that were classified differently by the multiple algorithms tested, focusing particularly on the reviews that were correctly labeled as fake by deep learning but incorrectly labeled by the other algorithms. We find that these reviews have some underlying patterns (for instance, some of them follow a storyline). We present examples of these patterns in Table 2.6. Anecdotally, we believe that these reviews are quite convincing and are very unlikely to be detected as fake by average online consumers.

## 2.6 Robustness Checks

### 2.6.1 Excluding Reviews of Clinics That Feature in the Training Dataset

It is quite possible that a deep learning algorithm may focus on elements in the writing style of certain specific fraudsters as opposed to subtle traits that may be common among fake reviews. To check for robustness of our results to such a possibility, we create a subset of testing data that is composed of reviews from clinics that do not have any reviews in the training dataset. By creating this subset we are able to isolate the commonality in fake reviews as opposed to commonality in the writing styles of specific fraudsters. This subset of the testing data has 524 reviews, which is about 2.88% of the entire testing data. The accuracy of our deep learning algorithm on this subset is 93.89%, broadly similar to our main results. This suggests that deep learning is able to learn some subtle traits that are common across fake reviews in general and that go beyond the stylistic idiosyncrasies of individual fraudsters.

### 2.6.2 Excluding Similar Reviews

Since some of the fake reviews may emulate the genuine reviews, we omit the near duplicates based on review similarity. We set up five thresholds on the level of similarity ranging from 80% to 100% in intervals of 5%. We present the results of each model with these different thresholds in Table 2.4. As observed, the results are consistent across thresholds.

## 2.7 Discussion and Conclusion

Our work is motivated by the economic significance of online word-of-mouth that creates strong incentives for fraudulent behavior. Currently, even state-of-the-art approaches to detecting fraudulent reviews have limitations; in particular, we described the challenges faced by the two existing algorithmic approaches to fraudulent review detection and the limitations in their applicability to reviews in the healthcare field. Furthermore, prior work has faced difficulties in establishing the ground truth for fraudulent reviews. We presented our solution to the labeling problem using digital forensics. Finally, we demonstrated that deep learning neural networks outperform traditional NLP and machine learning methods in classifying fraudulent and genuine reviews.

Our study advances the literature on linguistic techniques that can be applied to the detection of fraudulent reviews, as our model offers greatly improved performance in this tedious task. However, our study also has limitations that we acknowledge. First, our data is limited to the period during which the platform had a security vulnerability that allowed imposters to publish fraudulent reviews easily. While our fraud-detection model only utilizes the text of the data and has no input from the signals that were used to label the data, it is possible that after the security vulnerabilities are fixed and the cost of writing a fraudulent review then increases, the imposters may also change their writing style substantially, impacting the performance of our detection method. Second, in the healthcare domain, each genuine reviewer provides very few reviews, while each fraudster provides many more reviews using different identities (sock puppet approach). Our algorithm might detect fraudulent reviews by identifying shared words across clusters of fake reviews. Therefore, our approach may be prone to a high false-positive rate in other contexts where each genuine reviewer provides many reviews. Third, we identify fake reviews via the security loophole. While we believe that most of the fraudulent reviewers who posted reviews during our study period were lured by this honeypot, it is possible that there are fake reviews posted by alternate means that may bias our detection. Finally, our human evaluators were recruited from Amazon Mechanical Turk. While we sought out an evaluator population that shares a similar cultural background with the platform's users, the evaluators might be different than actual consumers, a vulnerability shared by any study using AMT.

In conclusion, the proliferation of fake content on online platforms can have severe consequences for consumer welfare. As such, advanced methods for fraud detection are needed across a variety of settings. Advanced machine learning techniques utilizing deep learning offer considerable promise in this context.

# CHAPTER 3: SPEECH IS SILVER, SILENCE IS GOLDEN—MOTIVATIONAL FRAMING BACKFIRES: EVIDENCE FROM A FIELD EXPERIMENT

## 3.1 Introduction

Online consumer reviews and ratings are ubiquitous, from the most inexpensive and low-stakes products and services such as toilet paper and restaurants to the most expensive and high-stakes products and services such as real estate and healthcare(Berger and Iyengar 2013, Gao et al. 2012, Godes and Mayzlin 2004, Moe and Trusov 2011).

Digital marketplaces (for example Amazon, eBay) extensively utilize consumer reviews and ratings to reduce information asymmetry. The ability of a platform to sustain a credible reputation system depends on the representativeness of the platform's word-of-mouth (WOM). However, consumption of WOM is free, whereas creation takes effort; therefore the contributor's motivation becomes a key to continuous flow of WOM. Thus, it remains crucial to understand how to effectively stimulate users into contributing (Fradkin 2015, Fradkin et al. 2016).

Leading digital platforms have tested different means and methods of stimulating the WOM generation, but there is no consensus on which practice is most effective (Fradkin 2015, Fradkin et al. 2016). Primarily, platforms have experimented by increasing the salience of intrinsic motivations. For example, Foursquare provides contributors with badges that recognize the degree of the member's contribution. These badges help to create a warm glow among the contributors, intrinsically making an emotional connection between users and the platform. Dang-Dang, another popular platform, offers monetary rewards to compensate the reviewers for their time. Similarly, Amazon runs a recognized reviewer program (Amazon Vine) that provides reviewers with free products and offers them recognition within the community as well.

Likewise, Stack Exchange provides users with features to help them craft an online social image that subsequently aids them in finding more successful jobs. Each of the incentivization methods mentioned above are suitable for their own contexts and may not work in all cases.

Researchers' conclusions about the effectiveness of these methods are divided. On one hand, studies in prosocial behavior (Hong et al. 2016, Jung et al. 2016) suggest that highlighting the incentive may enhance online contribution. On the other hand, studies in psychology (Deci 1975, Deci and Ryan 1985) suggest that incentives dampen intrinsic motivation and may reduce task performance. Due to these contradicting views, we do not see a clear consensus. Our study probes this question empirically in a socioeconomically driven setting—the choice of a healthcare provider.

Along with the question of incentivization and its efficacy, we also investigate means of increasing user identity revelation. Studies (Emmert et al. 2013, Forman et al. 2008, Lee and Choeh 2016) have shown that reviewer identity disclosure is key to the usefulness of reviews; in other words, attaching an identity to the content can be crucial to its credibility. These studies have further shown that divulgence of identity and personal information increases not only perceived usefulness but also product sales. However, disclosure of identity also reduces user privacy. The desire to withhold information from websites has increased dramatically over the years, as individuals are becoming more aware of how companies are handling their data and how easily their information is being collected online (Goldfarb and Tucker 2012). So on one hand, revealing one's identity is good for the usefulness of their online review; on the other hand, this very revelation makes the user more vulnerable to bad actors. Solving this conflict is crucial.

We ran a two-stage field experiment on a leading platform for doctor appointment booking. In the first stage we randomly allocated users to different email templates. These templates contained different motivational messages. Subsequent to responding to the email, the users were randomly allocated to one of two versions of the form that requested a detailed review. The first version of the form provided an option at the top of the webpage (i.e., before they wrote their review) that allowed users to stay anonymous. The second version also offered reviewers a choice to stay anonymous but did so at the bottom (i.e., after they wrote their review).

Our findings show that increasing the salience of intrinsic motivation by including the motivational message in the review request reduces the propensity to respond, suggesting that such framing crowds out volitional contribution. Furthermore, we find that offering the choice of anonymity up front has a positive effect on identity disclosure, suggesting that it increases user trust on the platform and makes the user feel comfortable revealing their identity.

## 3.2 Background – Motivational Framing

### 3.2.1 Identifying Key Motivations for Online Contribution

Investigating the motivation of users who share by WOM has been an active area of research. Dichter (1966) originally identified that users share by traditional (offline) WOM broadly on the lines of the following four motives: product involvement, self-involvement (warm glow), altruism and message Involvement. Engel et al. (1968) added dissonance reduction as one of the motivations. Sundaram et al. (1998) proposed the additional motivations of helping the company/provider. They also proposed altruism, anxiety reduction, vengeance and advice seeking as key reasons why users may share negative WOM. More recently, Hennig-Thurau et al. (2004) studied WOM motivation in the online context and introduced 11 possible motives, which included expectations of economic incentives. Wang and Fesenmaier (2004) suggested hedonic benefits and reputational gains as other possible motivations. Table 3.1 presents an overview of this research. Reviewing this body of work helped us narrow the potential key motives for online review contribution down to three: self-involvement (warm glow), altruism, and product (or service) involvement.

### 3.2.2 Message Framing

Work by Jung et al. (2016) suggests that nudging the users based on their intrinsic motivation, such as altruism, may increase their likelihood to contribute. Bapna et al. (2014) found that an altruistic component within an incentive enhances the outcome. Similarly, Sun et al. (2018) found that when a single use code is made shareable, users are more willing to help their friends rather than take advantage of the discount themselves. Likewise, Hong et al. (2016) found that pro-socially framed messages had a positive effect on users' content contribution. This stream of work suggests that nudging the users with emphasis on the altruistic dimension would increase their propensity to contribute.

On the contrary, when looking through the lens of the literature on pro-social behavior, studies suggest that extrinsic motivations of any kind may crowd out volunteering behavior or may have no discernible effect. For instance, Bérnabou and Tirole (2006) find that presenting extrinsic rewards and making them visible crowds out volitional contribution. Similarly, Gneezy and Rustichini (2000), Frey and Jegen (1999), Frey and Jegen (2001), Frey and Oberholzer-Gee (1997) also report similar findings. Likewise, Ariely et al. (2009) also find that incentives have no detectable positive effect on a volunteer task. (Refer to Table 3.2 for an overview of the mixed findings).

To summarize, recent experiments by Bapna et al. (2014), Jung et al. (2016), and Sun et al. (2018) all seem to suggest that nudging users by explicitly stating their motivations may have a positive effect, while work on extrinsic incentives for pro-social behavior (Ariely et al. 2009, Bérnabou and Tirole 2006, Gneezy and Rustichini 2000) seems to suggest the contrary. In our study we propose to empirically tease out this question.

*RQ1: Does an email message with either an altruistic, warm glow or product involvement framing stimulate user-generated content creation?*

## 3.3 Background – Identity Disclosure

### 3.3.1 Effect of Perceived Trust on Reviewer Identity Disclosure

Most review platforms empower users to write reviews by confirming an email ID (Lagu et al. 2010), leading to the emergence of fraudulent reviews. Research has shown that about 18% of published reviews are likely to be fake (Luca and Zervas 2016), suggesting an erosion of consumer trust. Although it is an onerous task, review platforms may re-establish user trust by only allowing verified consumers to post reviews and encouraging them to reveal their identity publicly.

Forman et al. (2008) found that transparent review authorship significantly influences consumers' perception of helpfulness. Their results demonstrated that the divulgence of personal information on online reviews is highly and positively correlated with not only the perceived helpfulness of the review but also with increased sales volume. Lee and Choeh (2016) also found that disclosure of reviewers' real names was positively associated with review helpfulness. Similarly, Emmert et al. (2013) underscored the risk of disinformation and seller strategic behavior when online feedback was unverified and anonymous. Therefore, given the benefits of identity revelation, having as many reviewers as possible reveal their identities gives the consumer and platform an advantage.

### 3.3.2 Challenges in User Identity Disclosure

In the present online environment, it has become easier than ever before for users to leave an electronic footprint that can be gleaned by companies to offer personalized services designed to extract the maximum consumer surplus, often to the user's disadvantage. This personalization has increased privacy concern and increased user desire to withhold personal information and identity from being published online (Goldfarb and Tucker 2012). According to a study by the Pew Research Center (Raine 2018), 91% of Americans believe they have lost control over their personal information to firms who have the power to utilize this data as they wish, while up to 86% of online users take precautions with their actions on the internet in hopes of staying anonymous. Therefore, it is becoming increasingly challenging for platforms to obtain users' consent to disclose their identity.

The revelation of identity becomes even harder in a sensitive context, such as posting an online review of one's experience with their healthcare provider. Within the healthcare context, studies have shown that 90% of respondents indicated that they had little to no knowledge about how their health information was being shared between medical organizations and 79% of respondents were not aware of how their National Health Index number was used (Whiddett et al. 2006). Furthermore, many individuals are unaware that their own medical records can be accessed by third parties such as regulatory bodies, insurance companies, and medical researchers without their explicit consent (Park 2013). Personal health information can potentially be used in detrimental ways. For instance, a number of pharmacies sold information from frequent users to several prominent drug companies such as Biogen and Hoffman-LaRoche (Anderson 2000). Patients were then persuaded to select drugs from these companies when they needed a refill on their original prescription. Although many individuals are unaware of how health information can be used and misused, many others, particularly those who tend to be most digitally literate, are aware of the risks associated with online health data and have responded by becoming more wary of what they share online. For this reason, obtaining user consent to reveal their identity along with their online review of their doctor can be challenging. Therefore, confidentiality plays a very significant role in any interaction between a healthcare platform and its user.

### 3.3.3 Increasing Trust to Increase Information Revelation

From the lens of social exchange theory, trust plays a significant role as it can potentially mitigate any costs that may deter individuals from participating (Metzger 2017). In the online environment, there are perceived risks and uncertainties that may impede the user from disclosing their personal information when they believe that the costs associated with identity disclosure outweigh the benefits. Hoffman et al. (1999) expand this notion through their research on how consumers are unlikely to engage with websites when there is a lack of trust between the two. From the user’s perspective, gaining access to certain features of a website through identity disclosure may not be worth overriding privacy concerns that arise when the website disregards their confidentiality (Hoffman et al. 1999).

However, perceived trust can make users disclose information they consider confidential (Cheshire et al. 2010). In this way, it is much like sharing information in person (Gupta and Dhami 2015). In an online setting, it is crucial for platforms to develop trust in order to make their users feel comfortable while browsing the site and sharing their information (Cheshire et al. 2010). As the people behind a platform are hidden from view when users browse the online storefront, it is difficult for users to assess the situation and understand the true intentions of the people they are interacting with. Therefore, any cues that can signal trustworthiness of the platform are useful in establishing trust. The absence of these cues can become an impediment that deters users from participating and creating a relationship with the platform (Cheshire et al. 2010). Research has shown that trust can be facilitated by website design (Cheshire et al. 2010, Corritore et al. 2005, Kim 2014). Professional layout, ease of navigation, ease of carrying out a transaction, and a lack of advertisements are all factors that can improve user perceptions of a website's trustworthiness (Corritore et al. 2005, Kim 2014).

In addition, research has shown that users are more comfortable disclosing information when they understand that a website has strong privacy and security policies (Gupta and Dhami 2015). Individuals will reveal more of their personal information when they have distinguished a decreased risk of a privacy breach and experience increased control over their information. This occurs when users receive confirmation that the online platform will not be distributing their personal information to unauthorized third parties.

Furthermore, the sensitivity of information revealed is dependent on the degree of trust. According to Mesch (2012), online users are reluctant to provide information that they deem sensitive, such as their personal or financial information, due to fears of misuse. However, this does not apply if the website is one they trust, such as their bank's website. For this reason, users often choose to lie to avoid this risk, thus making the requested personal information inaccurate (Mesch 2012). However, when users develop trust in the platform and are convinced that their information will be protected, they are more willing to reveal sensitive information (Tait and Jeske 2015).

To summarize, increasing perceived trust may lead users to reveal their identity and influence the quality of information they disclose, yet this effect may be moderated by sensitivity, or in other words by the personal nature of the patient/provider interaction that may prompt a desire for confidentiality. In our work, we examine this question empirically in the context of online reviews for healthcare providers. Formally, we ask the following question:

*RQ2: How does privacy assurance affect (1) personal identity revelation and (2) quality of information disclosed, and how is this effect moderated by sensitivity of the context?*

## 3.4 Methodology

### 3.4.1 Study Context

A field experiment was conducted on one of the largest healthcare platforms in Asia that enables anyone to book an appointment with a listed practicing physician. The marketplace attracts over 100,000 healthcare practitioners in over 10,000 clinics. Since the platform's inception in 2008, the platform has attracted over ten million users.

The platform allows users to book an appointment with and post a recommendation and a review for any healthcare practitioner. A user can give a positive or negative recommendation (analogous to a like or a dislike) and post a review. The users who book an appointment via online platform or visit the clinic of a practice that subscribes to this platform's electronic health record system and appointment management system are requested on the morning following their appointment to provide a review of their consultation. This request is sent as an email, and the user responds by clicking on the buttons in the form embedded in the email and then filling out a form in a new tab that captures the review.

These emails requesting the recommendation/review are sent to the user one day after their appointment and can be responded to within thirty days of receiving this request. After thirty days, the form in the email and its embedded links expire. A user has no obligation to post the recommendation/review, and not responding to this request has no bearing on their ability to book future appointments.

The feedback received is passed through an algorithm that checks for suspicious behavior, scans the reviews for abusive and indecent language, checks for any embedded spam or solicitation information, and forwards the reviews to a moderating panel. Upon successful machine and human moderation, which on average takes 1-2 business days, the reviews are published on the platform where anyone can view them and vote on their helpfulness (i.e., mark them as helpful or unhelpful). The corresponding healthcare practitioner can even reply to these reviews. Cross-replies are not permitted and the healthcare practitioner's reply to a review is visible to everyone. An example of a review with reply is shown in Figure 3.1.

The user also has an option to submit their review anonymously, in which case it remains anonymous both to the users of the platform and to the healthcare provider for whom the review is posted. However, this option is provided only when the user posts a positive recommendation. When the user gives a negative recommendation, the user is not given the option to keep their review anonymous. This is done to avoid accusations of fraudulent reviews. For instance, if a doctor receives many anonymous negative reviews, they may sue the platform and claim that these negative reviews were posted by their competitors.

Individuals who decide to submit a review can view all their posted reviews by logging in to their account and can even choose to retract any old reviews. However, once a review is posted, the author does not have the right to edit the review or change its anonymity.

### 3.4.2 Experimental Design

We intervene in the recommendation and review collection process in two stages. Our intervention targets patients who have either booked their doctor's appointment online or visited a participating clinic. In the first stage, we intervene by sending the patients an email message that primes them to one of the four motivations: altruistic motivation, altruistic motivation with data on number of potential beneficiaries, warm glow or self-praise motivation, and product involvement. We also have a control group that does not prime them to any specific motivation. We intervene by changing the subject of the message as well as the text of the email to prime the users to a particular motivation. We retain the same user interface across all email templates. Figures 3.2 through 3.6 provide snapshots of these templates.

In the second stage, the intervention involves creating two version of the standard review collection form. The first version (usual privacy assurance form) is the same as the existing form normally used by the platform. The second version (enhanced privacy assurance form) modifies the form to give the user increased privacy assurance. This version differs from the first form with respect to the placement of the checkbox that users can check to keep their review anonymous. This form places the anonymous review checkbox at the very top of the form so that users are aware of the anonymity option from the very beginning, even before they start typing their review. On the contrary, the usual privacy assurance form offers users the choice to be anonymous only at the very end, where they encounter the checkbox just before submitting the form. The design of these two forms is depicted in Figures 3.7 and 3.8.

To summarize, the experiment is set up in a 5x2 design. In the first stage every feedback request sent via email is randomly assigned to one of the five email templates. Following this stage, if the user responds by clicking on the YES button to the question about recommending the doctor to his friends or family, they are randomly selected to be sent to either the usual privacy assurance form or the enhanced privacy assurance form.

## 3.5 Data

### 3.5.1 Overview

Our dataset is constructed by matching six distinct subsets: email template assignment table, privacy form template assignment table, recommendation data, doctor review data, review helpfulness data and doctor specialty information. The email template assignment table gives us information about the email template used for each feedback request sent to patient, and the privacy form template assignment table gives us information about the privacy form used for each doctor review request. The doctor review data gives us the actual review along with the details of the feedback identifier in order to allow us to match it to the corresponding feedback and doctor review request. The review helpfulness data is a log taken at the end of the experiment window. This log records the perceived review helpfulness, as captured by the user response (marked as helpful/not helpful) to these reviews. We combine these data subsets to create a single table that records the details of each review request sent out. In this merged dataset, we capture the email template assigned, recommendation response, privacy form template assigned, doctor review text, and review helpfulness scores for each review request emailed to the users.

### 3.5.2 Summary Statistics

The experiment was initiated on March 17, 2016 and concluded on March 19, 2017, so it was live for a time duration of approximately one year. In this time duration, a total of 2.49 million email requests were sent. Table 3.3 presents the overall descriptive statistics. As observed, out of the 2,493,480 emails sent, 564,893 emails (23%) were opened by the users. These email requests resulted in 178,120 recommendations and 25,429 doctor reviews.

The overall response rate for stage 1 (recommendation) was 7.14% and stage 2 (doctor review) was 1.01%. Conditional on response for stage 1, the response rate in Stage 2 was 13.71%. Table 3.4 presents the summary statistics for recommendations. Out of the 178,120 recommendations posted in total, 163,021 (91%) were positive and 8,751 (5%) were negative. In the remaining 4% of responses (6,348), the users reported that they could not meet the doctor. Tables 3.5 and 3.11 present the summary statistics for doctor reviews. As observed, the average length of a review is 194.50 characters, and in 70% of the reviews, the users decided to reveal their identity. The reviews have 66.61 views on average and 50% of the reviews were captured from the high privacy assurance template.

### 3.5.3 Econometric Specification

Our estimation approach relies primarily on a difference-in-differences approach with a logit model to capture binary outcome variables of interest such as whether the email was opened, whether the user responded to the recommendation request and whether the user successfully completed a review. To estimate the review quality, we use review length, count of views of each review and compound sentiment score obtained after VADER analysis.

We set up models for both stages of the experiment. We begin with the models that estimate the effect of email templates on the open rate, recommendation rate and review response rate. For the second stage, we estimate the impact of privacy assurance on identity revelation and review quality. In the following subsections, we detail the econometric specifications for models for both of these stages.

## 3.6 Main Analyses

### 3.6.1 Impact of Motivational Emails on User Generated Content Creation

We first analyze how adding motivational messages to emails impacts the likelihood of user generated content creation. We investigate this by examining the impact on three facets of a user's response: the likelihood of opening the email, posting a recommendation and posting a physician review. The following subsections present the analysis of the impact of motivational email framing on each of the three facets mentioned above.

#### 3.6.1.1 Impact of Motivational Emails on Email Open Rate

To empirically measure the impact of motivational messaging on their likelihood of the email being opened, we implement a simple comparison across each type of email template. We set up the following model:

Y=α + $\beta_1$ Altruism_basic + $\beta_2$ Altruism_strengthened + $\beta_3$ Product_involvement + $\beta_4$ Warm_glow + ε (1)

The dependent variable in this analysis is the dummy variable that represents the email being opened. For the independent variables in this specification, we include the dummy variables that capture the type of motivational message in the email. To note, both the email body and the email subject include text to invoke the same type of motivation. The email messages along with their subject lines are depicted in Figures 3.11 through 3.15.

In our specification, the dependent variable is the dummy that represents whether the email was opened. This is a binary variable; therefore, we use a logistic estimator. We use robust standard errors in all the regressions. The key variables of interest are the coefficients of each email template, which capture the effect of different motivational framing on the likelihood of the email being opened. The results of the specification in Eq (1) are presented in Column 1 of

Table 3.6. The coefficients of Altruism_basic, Altruism_strengthened, Product_involvement and Warm_glow are -0.17, -0.19, -0.05 and -0.13, respectively, all of them statistically significant at $p<0.01$.

#### 3.6.1.2 Impact of Motivational Emails on Recommendation Rate

To empirically measure the impact of motivational messaging on the likelihood of these emails eliciting a user response in the form of a posted recommendation, we implement a simple comparison across each type of email message, similar to the one in the previous section. We set up the following model:

$$Y=\alpha + \beta_1 \text{ Altruism\_basic} + \beta_2 \text{ Altruism\_strengthened} + \beta_3 \text{ Product\_involvement} + \beta_4 \text{ Warm\_glow} + \varepsilon \qquad (2)$$

The dependent variable in this analysis is the dummy variable that represents whether a recommendation was posted. For the independent variables in this specification, we include the dummy variables that capture the type of motivational message in the email. As mentioned earlier, the email message templates with their text are depicted in Figures 3.11 through 3.15.

In our specification, the dependent variable is the dummy that represents whether a recommendation was posted. This is a binary variable; therefore, we use a logistic estimator. We use robust standard errors in all the regressions. The key variables of interest are the coefficients of each email template, which capture the effect of different motivational framing on the likelihood of posting a recommendation. The results of the specification in Eq (2) are presented in Column 2 of Table 3.6. The coefficients of Altruism_basic, Altruism_strengthened, Product_involvement and Warm_glow are -0.30, -0.31, -0.11 and -0.23, respectively, all of them statistically significant at $p<0.01$.

#### 3.6.1.3 Impact of Motivational Emails on Review Response Rate

To empirically measure the impact of motivational messaging on the likelihood of these emails prompting a user response in the form of posting a doctor review, we implement a simple comparison across each type of email message, similar to the ones in the previous sections.

For the independent variables in this specification, we include the dummy variables that capture the type of motivational message in the email. The dependent variable in this analysis is the dummy variable that represents whether a doctor review was posted. This is a binary variable; therefore, we use a logistic estimator. We use robust standard errors in all the regressions. The key variables of interest are the coefficients of each email template, which capture the effect of different email motivations on the likelihood of posting a recommendation. The results of the specification are presented in Column 3 of Table 3.6. The coefficients of Altruism_basic, Altruism_strengthened, Product_involvement and Warm_glow are -0.31, -0.25, -0.08 and -0.18, respectively, all of them statistically significant at $p<0.01$.

#### 3.6.1.4 Impact of Motivational Emails on Recommendation and Review Valence

In addition to investigating the overall effect of different motivational framing on the response rate, we also investigate the differential effect on the valence of recommendations. The dependent variable in this analysis is the dummy variable that represents whether a doctor recommendation was positive. For the independent variables in this specification, we include the dummy variables that capture the type of motivational message in the email.

The results of this specification are presented in Columns 1-4 of Table 3.7, which shows the effect of motivational emails on users' likelihood of posting a positive recommendation, a negative recommendation, a positive review and a negative review. As observed, the results are aligned in all four columns: the motivational messages reduce the user's likelihood of posting both positive and negative recommendations and reviews.

Additionally, in Table 3.8 we conduct a conditional analysis on the same specification. In this specification, however, we only select those requests that result in a recommendation. This analysis helps us to show the conversion rate. As observed in Column 1, we find that conditional to posting a recommendation, there is broadly speaking little distinction between different email motivations, except for the Altruism_strengthened group. For this group, the coefficients for pos_reco, neg_reco, pos_doc_rev are 0.05, -0.10 and 0.05, statistically significant at $p<0.1$, $p<0.01$ and $p<0.05$ respectively. The Altruism_strengthened email template both lays out an altruistic motivation and supports it with data, and we find that these messages result in an increase in the likelihood of a positive recommendation response and positive reviews. We also observe a decrease in the likelihood of a negative recommendation.

#### 3.6.1.5 Impact of Motivational Emails on Identity Revelation and Review Quality

In addition to investigating the impact of motivational framing on response rate, we investigate its effect on identity revelation and review response quality. We measure the revelation of identity in terms of whether a reviewer chooses to remain anonymous or disclose their first and last name. We measure quality of reviews using two metrics: length in number of characters (*Length*) and number of views (*Review_views*).

$$Y=\alpha + \beta_1 \text{ Altruism\_basic} + \beta_2 \text{ Altruism\_strengthened} + \beta_3 \text{ Product\_involvement} + \beta_4 \text{ Warm\_glow} + \varepsilon \qquad (3)$$

In this specification, the dependent variable is a dummy that specifies whether the user chooses to keep his identity anonymous or not. This dummy, signifying identity revelation, is set to 1 if user does not choose to be anonymous and 0 otherwise. Since the dependent variable is a dummy, we use a logistic estimator.

The key variables of interest are the coefficients of each email template, which capture the effect of different motivational framing on the user's likelihood of posting a recommendation. The results of the specification are presented in Column 1 of Table 3.9. The coefficient of Altruism_basic is -0.09 and is statistically significant at $p<0.05$. The coefficients of Altruism_strengthened, Product_involvement and Warm_glow are statistically insignificant.

Similar to the preceding sections, we set up the specifications for measuring the change using the two metrics of review quality: length in number of characters and number of views. Since the dependent variables are continuous in these specifications, we use OLS models. The results are presented in Columns 2 and 3 of Table 3.9. As observed in Column 2, the coefficients for dummy variables indicating messages with altruistic framing and altruistic framing with supporting data are 5.28 and 6.55, statistically significant at $p<0.1$ and $p<0.05$ respectively. Column 3 of Table 3.9 presents the results for the number of views of these reviews. The coefficients for dummy variables indicating different motivational framing are statistically insignificant.

In Table 3.10, we present the results for negative reviews. As observed, we do not find any statistically significant differences in review quality for reviews received in response to emails with different motivational framing.

### 3.6.2 Impact of Privacy Assurance on Personal Identity Revelation and Review Quality

As mentioned in Research Question 2, we seek to investigate the effect of privacy assurance on personal identity revelation and review quality. To examine this, we set up the following empirical specification:

$Y = \alpha + \beta_1$ Privacy + $\varepsilon$ (4)

Results are presented in Table 3.12. As observed, the coefficient for the dummy privacy assurance (*Privacy*) is 1.20, significant at $p<0.01$, suggesting that privacy assurance increases identity revelation.

To examine the effect on review quality, we use three metrics: review length, number of views and the VADER sentiment score. The results of these analyses are reported in Columns 1, 3, and 5 in Table 3.13. As we can observe, the coefficient of privacy assurance is statistically insignificant for review length and VADER sentiment score, suggesting no effect of privacy assurance on these two metrics of quality. However, we observe that the coefficient of views of the review (Column 3) is statistically significant at $p<0.01$, suggesting that privacy assurance increases along with the number of views that the review acquires.

### 3.6.3 Moderating Impact of Patient Sensitivity on the Influence of Privacy Assurance on Personal Identity Revelation and Review Quality

In addition to the direct effect of privacy assurance on identity revelation and review quality, we also examine the moderating impact of patient sensitivity on this relationship. For our purposes, we describe patient sensitivity as the degree to which the nature and/or content of the patient/provider interaction is deemed personal by the patient. Interactions with higher patient sensitivity are thus more likely to lead to anonymous reviews. To measure patient sensitivity based on the proportion of anonymous reviews, we select two specialty pairs. The first pair of specialties is dermatology and psychiatry. These two specialties have a relatively high proportion of anonymous reviews, and hence we consider reviews for these two specialties to be sensitive. The second pair of specialties is ophthalmology and physiotherapy. These two specialties have a relatively low proportion of anonymous reviews, and hence we consider reviews for these two specialties to be less sensitive. We set up a sensitivity dummy variable that is set to 1 for reviews of dermatologists and psychiatrists and 0 for reviews of ophthalmologists and physiotherapists.

We then set up the following econometric specification:

$$Y = \alpha + \beta_1 \text{ Privacy} + \beta_2 \text{ Sensitivity} + \beta_3 \text{ PrivacyXsensitivity} + \varepsilon \qquad (5)$$

For identity revelation, the results are presented in Column 2 of Table 3.12. As observed, the coefficient is -0.78 for sensitivity, significant at $p<0.01$. The coefficient of *PrivacyXsensitivity* is 0.35, significant at $p<0.1$. This suggests that the effect of privacy assurance is even stronger in sensitive specialties.

For effect on review quality, the results are presented in Columns 2, 4 and 6 of Table 3.14. The coefficients of privacy, sensitivity and the interaction term *PrivacyXsensitivity* are statistically significant for views of the review. As observed in Column 4, the coefficient of privacy is 6.09, significant at $p<0.05$. The coefficient for sensitivity is 76.74, significant at $p<0.01$, suggesting that reviews of doctors in sensitive specialties are viewed more often. The coefficient for *PrivacyXsensitivity* is 27.19, statistically significant at $p<0.01$, suggesting that sensitivity positively moderates the influence of privacy assurance on the number of views of the review, which is one metric of review quality.

## 3.7 Robustness Checks

### 3.7.1 Alternative Measures of Review Quality

In addition to these metrics mentioned for review quality, we also checked review quality using a user-generated measure: the count of “found helpful” votes received by the reviews. The results were statistically insignificant. Probing further, we found that this measure is very sparse, as over 90% of reviews have zero votes. Furthermore, we also checked review quality using different measures of VADER sentiment scores, both positive and negative. The results are consistent with the findings previously presented using the VADER compound score.

### 3.7.2 Rewording the Privacy Assurance Message

For the enhanced privacy version of the review collection form, as illustrated in Figure 3.17, the message shown to the users was “Submit my review anonymously”. For one month during the experiment window, the message was reworded to “Keep my review anonymous.” We reconducted the analysis with this change to the wording and present the results in Tables 3.14 and 3.15. As we can observe, the results remain consistent.

### 3.7.3 Falsification Check

We set up a falsification test for the enhanced privacy experiment group in which we remind the user that they can remain anonymous in text at the top of the collection form, but we move the action button (checkbox) to the bottom of the form. We re-conducted the analysis on this version and present the results in Tables 3.16 and 3.17. We note that the effects are statistically insignificant. This suggests that the option of anonymity only effects identity revelation and review quality when users must make decisions about anonymity at the very beginning of the web form.

## 3.8 Discussion and Conclusion

In this study, we examine the impact of motivational message framing and privacy assurance on quantity and quality of WOM in the context of healthcare services. We find that motivational framing backfires—it reduces the number of recommendations and reviews posted. Furthermore, we find that assuring users of their privacy by offering the option of anonymity up front increases identity disclosure. Finally, our study shows mixed fixings of the impact of motivational framing and privacy assurance on quality of WOM: one of the three metrics of quality illustrate an increase, but the others show no significant change.

We believe that the observed backfiring of motivational messages is likely driven by two mechanisms. First, we believe that stating the motivation up front reduces the volitional drive to contribute—in other words, explicitly stating the reason why the user should contribute turns the implicit motivation into an extrinsic driver. This unintentional removal of the implicit motivation then leads to a drop in contributions for the pro-social cause, as observed in findings by Mellström and Johannesson (2008), Gneezy and Rustichini (2000) and Ariely et al. (2009). Second, we believe that although framing a message with one or the other motivation may positively influence users who have those respective motivations, it may also negatively influence users who have a different intrinsic motivation. We could be observing a net loss in response quantity due to this self-selection, even if the treatment effect is positive for some users. In our context, we do not foresee a way to untangle these two mechanisms.

With respect to our findings on the impact of privacy assurance, we believe that asking users to make their choice up front reassures them that the platform cares about their privacy and helps them place more trust in the platform. This reassurance in turn leads them to be more comfortable disclosing their identity. We also think that this nudges them to post higher quality reviews, as we find that those reviews generated with the enhanced privacy option gather a higher number of views. We further believe, based on the change in the views of the review, that there may be a yet undetected impact on review quality. To uncover these possible additional effects, we explored the VADER sentiment analysis, but so far the findings do not show a significant change between enhanced and usual privacy groups. We believe that there is room for exploration in this area using both the VADER sentiment analysis and other quality metrics.

Finally, prior to discussing the implications of these findings, we acknowledge the limitations of our study. First, emails with motivational subject lines are quite common, especially in the context of online promotions. It is possible that users are accustomed to ignoring messages that ask them to help others, be a hero, etc., and may consider these messages to be unsolicited marketing or spam. In contrast, a message that simply asks them to post their review may be considered more honest and hence trustworthy. Therefore, the mechanism of the observed effect remains unclear.

Second, there is a chance that in the enhanced privacy treatment group some users may not notice that the choice of anonymity is offered up front. In our design, since the default option is identity disclosure, if a number of users are missing out on reading the text, identity disclosure would increase. Therefore, there remains some uncertainty regarding the mechanism of this effect.

Our findings offer important implications for practice from the perspective of the platform. The reduction in response rate in all motivational message treatment groups suggests that it is best to refrain altogether from framing requests with motivational messages. This could be due to any of the mechanisms discussed, but irrespective of the mechanism at play, the results show a decrease. Second, we find that privacy assurance leads to an increase in identity disclosure, thus implying that if platforms looking for an increase in reviewer identity disclosure take steps to assure users' trust, they will be rewarded by an increase in the number of users who are comfortable revealing their identity.

Several significant opportunities exist to build on this research. In our research setting we changed both the subject and the text of the email to align with the same intrinsic motivation. Future studies can keep the subject consistent across all groups and test the effect of the message text by itself. This may help ensure that users do not deem these requests to be unsolicited promotions. Second, future studies can attempt to personalize messages, tailoring them to user characteristics to make the effect stronger. Finally, future research could mix and match different motivations with appropriate cues in other forms (i.e., images) to come up with an optimal design that maximizes response rate.

## TABLES

### Table 1.1: Summary Statistics – Recommendation Count

| Group | Sub-group | # of doctors | # of practice-doctors | # of positive recommendations (Mean) | S.D. | Min | Max |
|---|---|---|---|---|---|---|---|
| Treatment | *Top* | 45 | 90 | 8.355 | 6.133 | 3 | 44 |
| | *Middle* | 78 | 119 | 5.436 | 5.225 | 3 | 39 |
| | *Bottom* | 86 | 103 | 3.495 | 1.220 | 3 | 10 |
| Control | - | 173 | 233 | 1.605 | 0.706 | 0 | 5* |
| Competition | - | 306 | 425 | 1.687 | 1.472 | 0 | 20 |

*: This indicates the count of positive recommendations. We would like to note that some doctors have more than 3 positive recommendations but may still fail to meet the 70% positive criteria. Therefore, these doctors also constitute the control group since their recommendations remain unpublished.

### Table 1.2: Summary Statistics – Appointment Count

| Group | Sub-group | # of doctors | # of practice-doctors | # of online appointments per day (Mean) | S.D. | Min | Max |
|---|---|---|---|---|---|---|---|
| | *Top* | 45 | 90 | 0.637 | 1.282 | 0 | 15 |
| Treatment | *Middle* | 78 | 119 | 0.457 | 0.920 | 0 | 11 |
| | *Bottom* | 86 | 103 | 0.304 | 0.641 | 0 | 5 |
| Control | - | 173 | 233 | 0.135 | 0.403 | 0 | 4 |
| Competition | - | 306 | 425 | 0.182 | 0.522 | 0 | 7 |

### Table 1.3: Summary Statistics – User Session Level (Overall)

| Variable | Variable Description | Observations | Mean | S.D. | Min | Max |
|---|---|---|---|---|---|---|
| *Unique* | Consideration set size | 42,229 | 1.807 | 1.976 | 1 | 48 |
| *Duration* | Session duration in seconds | 42,229 | 632.2 | 912.1 | 1 | 14,291 |
| *Max_distance* | Distance between the two farthest clinics browsed in a session | 10,479 | 9.022 | 7.359 | 0.10 | 63.22 |
| *Post* | Post period dummy | 42,229 | 0.604 | 0.489 | 0 | 1 |
| *Treat* | Session belongs to treated market | 42,229 | 0.747 | 0.435 | 0 | 1 |
| *Spa* | Sparse_WOM market dummy | 42,229 | 0.290 | 0.454 | 0 | 1 |
| *Mod* | Moderate_WOM market dummy | 42,229 | 0.230 | 0.421 | 0 | 1 |
| *Den* | Dense_WOM market dummy | 42,229 | 0.226 | 0.418 | 0 | 1 |
| *Treated_Viewed* | Dummy to indicate that a rated doctor was browsed | 42,229 | 0.163 | 0.370 | 0 | 1 |
| *Last_Viewed* | Dummy to indicate that a rated doctor was the last doctor browsed | 42,229 | 0.110 | 0.313 | 0 | 1 |
| *Log_duration_per_profile* | Natural logarithm of average time duration per profile (in seconds) | 42,229 | 4.083 | 1.926 | 0.0 | 8.02 |

**Table 1.4: User Session Subgroup Level Categorization**

| Group | Number of practice-doctors (With WOM) | Number of doctors (With WOM) | Number of markets | Number of sessions |
|---|---|---|---|---|
| *Control* | - | - | 58 | 10,801 |
| *Spa* | 73 | 60 | 47 | 11,780 |
| *Mod* | 102 | 60 | 24 | 9,339 |
| *Den* | 137 | 89 | 7 | 10,309 |

**Table 1.5: Top Specialties and Cities in the Dataset**

| | Specialty or City Name | Proportion of doctors | Proportion of sessions |
|---|---|---|---|
| Specialty | Dermatologist/Cosmetologist | 22.0% | 19.5% |
| | Gynecologist/Obstetrician | 13.8% | 15.5% |
| | Dentist | 10.0% | 10.7% |
| | Orthopedist | 6.7% | 7.8% |
| | ENT Specialist | 7.0% | 5.9% |
| | | | |
| City | Bangalore | 44.7% | 30.2% |
| | Delhi | 11.9% | 15.4% |
| | Mumbai | 7.1% | 15.2% |
| | Hyderabad | 12.6% | 13.4% |
| | Chennai | 7.5% | 12.8% |

**Table 1.6: Correlation Matrix – User Session Level**

| Variable | *Unique* | *Duration* | *Max. Distance* | *Treat* | *Spa* | *Mod* | *Den* |
|---|---|---|---|---|---|---|---|
| *Unique* | 1.00 | | | | | | |
| *Duration* | 0.40 | 1.00 | | | | | |
| *Max. Distance* | 0.37 | 0.19 | 1.00 | | | | |
| *Treat* | -0.01 | 0.04 | -0.03 | 1.00 | | | |
| *Spa* | 0.04 | 0.02 | 0.08 | 0.31 | 1.00 | | |
| *Mod* | -0.01 | 0.01 | 0.00 | 0.28 | -0.36 | 1.00 | *Mod* |
| *Den* | -0.04 | 0.00 | -0.05 | 0.30 | -0.39 | -0.35 | 1.00 |

**Table 1.7: Correlation Matrix – Practice_doctor_ID – Day Level**

| **Variable** | *Yes_count* | *Appt* | *Top* | *Middle* | *Bottom* |
|---|---|---|---|---|---|
| *Yes_count* | 1.00 | | | | |
| *Appt* | 0.39 | 1.00 | | | |
| *Top* | 0.49 | 0.17 | 1.00 | | |
| *Middle* | 0.26 | 0.11 | -0.12 | 1.00 | |
| *Bottom* | 0.05 | 0.02 | -0.11 | -0.13 | 1.00 |

## Table 1.8: Variables in the Consumer Decision-Making Analyses

| Variable | Variable Description |
|---|---|
| *Session ID* | The unique identifier for each web session. A web session is defined as a stream of continuous activity received from the same IP identifier. A session lasts until there is thirty minutes of inactivity |
| *Unique* | Number of unique doctors browsed in this web session |
| *Duration* | Session duration measured in seconds (Time difference between first and last GET request received by the server from the client's browser) |
| *Max_distance* | Distance between the two farthest clinics browsed in a session |
| *Market* | Categorical variable to identify the city and specialty dyad of the doctors browsed in this session |
| *Treat* | Dummy to indicate if observation is in control group or treatment group |
| *Post* | Dummy to indicate treatment window |
| *Spa* | Dummy to indicate sessions in markets with 1-2 doctors with visible online WOM |
| *Mod* | Dummy to indicate sessions in markets with 3-9 doctors with visible online WOM |
| *Den* | Dummy to indicate sessions in markets with 10 or more doctors with visible online WOM |

## Table 1.9: Variables in the Doctor Appointment Analyses

| Variable | Variable Description |
|---|---|
| *Practice Doc ID* | Unique identifier for each combination of doctor and clinic |
| *Yes_count* | Count of visible positive recommendations |
| *Post* | Dummy to indicate pre- and post-period |
| *Appt_count* | Number of online appointments booked |
| *Date* | The date for which the appointment bookings are measured |
| *Treat* | Dummy to indicate treatment group |
| *Top* | Dummy to indicate that the doctor is in the top 33% relative to their peers |
| *Middle* | Dummy to indicate that the doctor is in middle 33% relative to their peers |
| *Bottom* | Dummy to indicate that the doctor is among the remaining rated peers |

**Table 1.10: The Impact of Recommendation Visibility on Consumer Consideration Set**

Negative binomial regression. Dependent variable: Number of unique doctors viewed in a search session

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
|---|---|---|---|---|---|---|---|
| *Sample* | *Sessions of treated markets* | *Sessions of treated markets* | *Sessions of all markets* | *Sessions of treated markets (Smaller 4-day window)* | *Sessions of all markets (Smaller 4-day window)* | *Excluding users having a session in both pre- and post-periods* | *Excluding users having a session in both pre- and post-periods* |
| *Post* | 0.005 | | -0.013 | | -0.049 | | -0.024 |
| | (0.015) | | (0.020) | | (0.032) | | (0.023) |
| *PostXSpa* | | 0.144*** | 0.154*** | 0.147*** | 0.196*** | 0.141*** | 0.160*** |
| | | (0.049) | (0.050) | (0.054) | (0.061) | (0.048) | (0.051) |
| *PostXMod* | | 0.014 | 0.026 | -0.011 | 0.038 | 0.028 | 0.048 |
| | | (0.049) | (0.052) | (0.041) | (0.051) | (0.052) | (0.055) |
| *PostXDen* | | -0.172*** | -0.162*** | -0.137*** | -0.088 | -0.170*** | -0.150** |
| | | (0.050) | (0.054) | (0.049) | (0.058) | (0.060) | (0.063) |
| *Constant* | 0.222*** | 0.128*** | 0.130*** | 0.115*** | 0.116*** | 0.146*** | 0.150*** |
| | (0.021) | (0.038) | (0.037) | (0.035) | (0.033) | (0.039) | (0.038) |
| *Observations* | 31,529 | 31,529 | 42,229 | 16,765 | 22,312 | 26,450 | 35,587 |
| *Log Pseudo Likelihood* | -52,177 | -52,063 | -68,950 | -27,460 | -36,129 | -43,424 | -57,936 |
| *Clusters* | 88 | 88 | 137 | 86 | 137 | 86 | 137 |

Clustered robust standard errors in parentheses. Day of the week dummies and market dummies are included in all the above models. *** p<0.01, ** p<0.05, * p<0.1

**Table 1.11: The Impact of Recommendation Visibility on Consideration Set Composition**

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| *Sample* | *Sessions of treated markets* | *Sessions of treated markets (Smaller 4-day window)* | *Excluding users having a session in both pre- and post- periods* | *Sessions of treated markets* | *Sessions of treated markets (Smaller 4-day window)* | *Excluding users having a session in both pre- and post-periods* |
| *Dep. Variable* | *Treated_Viewed* | *Treated_Viewed* | *Treated_Viewed* | *Last_Viewed* | *Last_Viewed* | *Last_Viewed* |
| *Model* | *Logit* | *Logit* | *Logit* | *Logit* | *Logit* | *Logit* |
| *PostXSpa* | 0.279*** | 0.262*** | 0.273*** | 0.157** | 0.148* | 0.110 |
| | (0.085) | (0.096) | (0.085) | (0.063) | (0.077) | (0.073) |
| *PostXMod* | 0.364*** | 0.311*** | 0.361*** | 0.357*** | 0.445*** | 0.354*** |
| | (0.058) | (0.062) | (0.064) | (0.082) | (0.084) | (0.081) |
| *PostXDen* | 0.176** | 0.190** | 0.151* | 0.293*** | 0.262*** | 0.266*** |
| | (0.073) | (0.077) | (0.083) | (0.070) | (0.073) | (0.070) |
| *Constant* | -2.954*** | -2.965*** | -3.301*** | -2.871*** | -2.912*** | -3.204*** |
| | (0.076) | (0.079) | (0.080) | (0.073) | (0.077) | (0.078) |
| *Observations* | 31,393 | 16,696 | 26,356 | 31,010 | 16,129 | 25,728 |
| *Pseudo R-squared* | 0.115 | 0.120 | 0.117 | 0.111 | 0.109 | 0.106 |
| *Clusters* | 81 | 79 | 80 | 75 | 73 | 73 |

Clustered robust standard errors in parentheses. Day of the week dummies and market dummies are included in all the above models. *** p<0.01, ** p<0.05, * p<0.1

Note: The number of clusters change across columns, since there are some markets that do not have any observations when the window size is reduced, or when the sessions of users that have a session in both pre and post period are dropped.

**Table 1.12: The Impact of Recommendation Visibility on User Session Duration**

Ordinary least squares regression. Dependent variable: Duration of a session measured in seconds

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
|---|---|---|---|---|---|---|---|
| *Sample* | *Sessions of treated markets* | *Sessions of treated markets* | *Sessions of all markets* | *Sessions of treated markets (Smaller 4-day window)* | *Sessions of all markets (Smaller 4-day window)* | *Excluding users having a session in both pre- and post-periods* | *Excluding users having a session in both pre- and post-periods* |
| *Post* | 33.24*** | | 4.98 | | 27.54 | | 13.53 |
| | (10.40) | | (19.61) | | (22.81) | | (19.37) |
| *PostXSpa* | | 130.19*** | 128.76*** | 102.50*** | 76.04** | 111.74*** | 98.15*** |
| | | (31.17) | (35.91) | (31.26) | (38.41) | (27.97) | (33.01) |
| *PostXMod* | | 24.41 | 22.95 | -1.70 | -28.71 | 12.11 | 0.020 |
| | | (33.24) | (37.88) | (23.24) | (32.10) | (31.51) | (36.19) |
| *PostXDen* | | -81.20*** | -82.86** | -41.18 | -69.19** | -85.76*** | -98.42*** |
| | | (27.33) | (33.37) | (26.39) | (34.37) | (27.67) | (32.64) |
| *Constant* | 596.75*** | 532.11*** | 541.73*** | 528.59*** | 537.54*** | 518.35*** | 531.92*** |
| | (16.13) | (25.83) | (24.85) | (21.51) | (20.26) | (24.71) | (24.00) |
| *Observations* | 31,529 | 31,529 | 42,229 | 16,765 | 22,312 | 26,450 | 35,587 |
| *Clusters* | 88 | 88 | 137 | 86 | 137 | 86 | 137 |
| *R-squared* | 0.017 | 0.020 | 0.021 | 0.020 | 0.022 | 0.019 | 0.019 |

Clustered robust standard errors in parentheses. Day of the week dummies and market dummies are included in all the above models. *** p<0.01, ** p<0.05, * p<0.1

**Table 1.13: The Impact of Recommendation Visibility on Geographic Dispersion**

Ordinary least squares regression

Dependent variable: Haversine distance between the two farthest doctors browsed in a session measured in km

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
|---|---|---|---|---|---|---|---|
| *Sample* | *Sessions of treated markets* | *Sessions of treated markets* | *Sessions of all markets* | *Sessions of treated markets (Smaller 4-day window)* | *Sessions of all markets (Smaller 4-day window)* | *Excluding users having a session in both pre- and post-periods* | *Excluding users having a session in both pre- and post-periods* |
| *Post* | 0.335** | | -0.202 | | -0.345 | | -0.045 |
| | (0.153) | | (0.376) | | (0.477) | | (0.369) |
| *PostXSpa* | | 1.307*** | 1.466*** | 1.257** | 1.606** | 1.152*** | 1.129** |
| | | (0.415) | (0.524) | (0.502) | (0.692) | (0.292) | (0.459) |
| *PostXMod* | | 0.468 | 0.627 | 0.305 | 0.639 | 0.571* | 0.550 |
| | | (0.348) | (0.494) | (0.270) | (0.545) | (0.311) | (0.474) |
| *PostXDen* | | -0.805** | -0.647 | -0.250 | 0.097 | -0.718*** | -0.748* |
| | | (0.351) | (0.498) | (0.214) | (0.517) | (0.265) | (0.441) |
| *Constant* | 8.226*** | 7.568*** | 7.565*** | 7.504*** | 7.450*** | 7.746*** | 7.763*** |
| | (0.307) | (0.412) | (0.396) | (0.487) | (0.465) | (0.988) | (0.977) |
| *Observations* | 8,434 | 8,434 | 10,479 | 4,432 | 5,487 | 7,087 | 8,864 |
| *Clusters* | 77 | 77 | 135 | 76 | 131 | 77 | 135 |
| *R-squared* | 0.083 | 0.087 | 0.096 | 0.094 | 0.102 | 0.0907 | 0.099 |

Clustered robust standard errors in parentheses. Day of the week dummies and market dummies are included in all models. The sessions used for the results presented above are those that have at least two clinics browsed in the same session, and to exclude outliers, the clinics browsed must be at least 0.1 km apart and up to 70 km apart. Furthermore, these sessions are only those that have location data available for all doctors (their clinics). Finally, the number of clusters (markets) change from Column 3 to Column 5, since there are 4 markets that do not have any observations in the reduced window. *** p<0.01, ** p<0.05, * p<0.1

## Table 1.14: The Impact of Online WOM on Appointment Count of Doctors

Negative binomial regression
Dependent variable: Number of appointments booked online for a practice-doctor in a day

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
|---|---|---|---|---|---|---|---|---|
| *Sample* | *Treatment Group* | *Treatment Group* | *Treatment Group* | *Competition Group* | *Treatment and Comparison Group* | *Treatment and Comparison Group* | *Treatment and Comparison Group excluding top 10% of treated doctors* | *Treatment and Comparison Group, ties excluded* |
| *Comparison* | *Before vs. after for rated doctors* | *Before vs. after for rated doctors* | *Before vs. after for rated doctors* | *Before vs. after for competing unrated doctors* | *Difference-in-differences* | *Difference-in-differences* | *Difference-in-differences* | *Difference-in-differences* |
| *Post* | 0.144*** | | | -0.143** | -0.158* | -0.158* | -0.156* | -0.103 |
| | (0.047) | | | (0.068) | (0.090) | (0.090) | (0.092) | (0.104) |
| *PostXSpa* | | 0.007 | | | | | | |
| | | (0.082) | | | | | | |
| *PostXMod* | | 0.160* | | | | | | |
| | | (0.086) | | | | | | |
| *PostXDen* | | 0.206*** | | | | | | |
| | | (0.075) | | | | | | |
| *PostXTreat* | | | | | 0.311*** | | | |
| | | | | | (0.101) | | | |
| *PostXTop* | | | 0.255*** | | | 0.421*** | 0.359*** | 0.388*** |
| | | | (0.0831) | | | (0.122) | (0.135) | (0.140) |
| *PostXMiddle* | | | 0.115* | | | 0.280*** | 0.280*** | 0.306** |
| | | | (0.0633) | | | (0.108) | (0.108) | (0.140) |
| *PostXBottom* | | | 0.001 | | | 0.167 | 0.167 | 0.110 |
| | | | (0.100) | | | (0.134) | (0.134) | (0.129) |
| *Constant* | -0.27*** | -0.28*** | -0.34*** | -18.45*** | -0.25*** | -0.32*** | -0.30*** | 0.28*** |
| | (0.090) | (0.102) | (0.105) | (1.008) | (0.081) | (0.095) | (0.107) | (0.101) |
| *Pseudo R-squared* | 0.29 | 0.29 | 0.29 | 0.26 | 0.30 | 0.30 | 0.29 | 0.28 |
| *Log Pseudo Likelihood* | -2,630 | -2,626 | -2,628 | -1,657 | -4,289 | -4,287 | -3,768 | -3,457 |
| *Observations* | 4,680 | 4,485 | 4,680 | 6,375 | 8,175 | 8,175 | 7,935 | 7,395 |
| *Clusters* | 312 | 299 | 312 | 425 | 545 | 545 | 529 | 493 |

Clustered robust standard errors in parentheses. Day of the week dummies and practice-doctor dummies are included in all the above models. *** p<0.01, ** p<0.05, * p<0.1

**Table 2.1: Summary of Existing Literature on Review Fraud Detection**

| Study | Approach | Methodology | Dataset | Accuracy reported |
|---|---|---|---|---|
| Jindal and Liu (2008) | Linguistic | Duplicates, extreme sentiment, similarity between review and product description | 5.8 million Amazon reviews | Accuracy is not proposed as ground truth is not established. |
| Gilbert and Karahalios (2010) | Linguistic | Duplicates and high similarity | 100,000 Amazon reviews | Roughly 10-15% of reviews have high similarity. The follow-up qualitative survey reveals that in many cases, reviewers consult existing reviews to draft their own review. |
| Lim et al. (2010) | Behavioral | Identify reviewers who target certain products and product groups and whose ratings have significant deviation from others | 11,038 Amazon reviews | Absence of ground truth labels of users leads them to compare their method to labels generated by experts. They find that their models match experts' labels more closely than "report helpful" metrics. |
| Ott et al. (2011) | Linguistic | SVM and Naive Bayes classifier | 400 synthesized fraudulent reviews from Amazon Mechanical Turk and 400 published reviews from TripAdvisor | They report an 89.8% accuracy with LIWC and SVM with bigrams. |
| Feng et al. (2012) Feng et al. (2012) | Behavioral | Distributional anomalies of ratings to label fraudulent reviews | Reviews scraped from Amazon and TripAdvisor; they subsequently test their algorithm on the dataset of Ott et al. (2011) | They report a 74% accuracy comparing their algorithm on the Ott et al. 2011 dataset. |
| Feng et al. (2012) | Linguistic | Context-free grammar (CFG) parse trees | Same data as Ott et al. (2011) | They report an accuracy of 91.2%. |
| Mukherjee et al. (2012) | Behavioral (At group level) | Detecting groups of fraudulent reviewers by observing behaviors of groups of individuals who | 109,518 Amazon reviews of manufactured product. In addition to temporal signals, they use human judges to label data | They report a 0.95 area under the curve (AUC). |

| | | | | |
|---|---|---|---|---|
| | | provide similar ratings to the same groups of products over time | on a spam index. | |
| Xie et al. (2012) | Behavioral | Observe unusually correlated patterns in the time of review posting and in the content of reviews | 408,469 reviews from resellerratings.com | Their baseline is human evaluation of spam, and they report that their algorithm has 61.11% precision and 75.86% recall on suspicious reviews. Furthermore, they report that human evaluators have a very low concurrence rating, implying that human evaluators often disagree. |
| Fei (2013) | Behavioral | Markov random field model and loopy belief propagation algorithm | 210,761 Amazon reviews | They report an accuracy of 71.2% with all reviews and an accuracy of 77.6% when considering only the burst reviews. |
| Mukherjee et al. (2013) | Behavioral and linguistic | Unigrams and bigrams, review length, reviewer deviation, percentage of positive reviews of a product, etc. | Yelp reviews of 85 hotels and 130 restaurants in the Chicago area | They benchmark the accuracy of their algorithm against the labels on Yelp (filtered reviews and published reviews) and report 86.1% accuracy when using both linguistic and behavioral features. |

**Table 2.2: Summary Statistics (Overall)**

| **Variable** | **Observations (N)** | **Mean** | **Standard Deviation** | **Min** | **Max** |
|---|---|---|---|---|---|
| Number of characters in a review (preprocessed) | 181,951 | 113.55 | 50.00 | 51 | 299 |
| Number of views of the review (preprocessed) | 181,951 | 28.24 | 94.98 | 0 | 2387 |
| Upvote count (review marked helpful) | 181,951 | 0.167 | 1.74 | 0 | 151 |
| Downvote count (review marked not helpful) | 181,951 | 0.019 | 0.23 | 0 | 14 |

**Table 2.3: Summary Statistics (Split by fraudulent vs. genuine reviews)**

| **Variable** | *Characters* | *Views* | *Upvotes* | *Downvotes* |
|---|---|---|---|---|
| **For Genuine Reviews** | | | | |
| *N* | 167,335 | 167,335 | 167,335 | 167,335 |
| *Mean* | 112.51 | 28.90 | 0.13 | 0.019 |
| *S.D.* | 49.45 | 66.67 | 1.12 | 0.22 |
| *Min* | 51 | 0 | 0 | 0 |
| *Max* | 299 | 2387 | 140 | 14 |
| **For Fraudulent Reviews** | | | | |
| *N* | 14,616 | 14,616 | 14,616 | 14,616 |
| *Mean* | 125.49 | 20.67 | 0.56 | 0.026 |
| *S.D.* | 54.5 | 51.3 | 4.81 | 0.31 |
| *Min* | 51 | 0 | 0 | 0 |
| *Max* | 299 | 1382 | 151 | 14 |
| **T-Test** | | | | |
| | Sig. at p<0.001 | Sig. at p<0.001 | Sig. at p<0.001 | Sig. at p<0.001 |

**Table 2.4: Results: Performance by Different Algorithms**

| | | **Thresholds** | | | | |
|---|---|---|---|---|---|---|
| **Metric** | **Model** | **80%** | **85%** | **90%** | **95%** | **100%** |
| *Area under the curve (AUC)* | *Logit* | 0.664 | 0.647 | 0.661 | 0.668 | 0.656 |
| *Area under the curve (AUC)* | *Random Forest* | 0.655 | 0.648 | 0.636 | 0.669 | 0.648 |
| *Area under the curve (AUC)* | *Deep Learning* | 0.724 | 0.709 | 0.714 | 0.722 | 0.723 |
| *Accuracy* | *Logit* | 92.59% | 92.34% | 92.00% | 92.32% | 92.09% |
| *Accuracy* | *Random Forest* | 92.61% | 92.31% | 92.17% | 92.32% | 92.06% |
| *Accuracy* | *Deep Learning* | 92.83% | 92.41% | 92.26% | 92.62% | 92.80% |

## Table 2.5: Results: Human Evaluation vs. Machine Learning

| | | Tie-breaking rule | | | |
|---|---|---|---|---|---|
| | | **(1)** | **(2)** | **(3)** | **(4)** |
| **Metric** | **Model** | *The review with a tie in human evaluation score is considered genuine* | *The review with a tie in human evaluation score is omitted* | *The review with a tie in human evaluation score is considered fake* | *The review with a tie in human evaluation score is omitted* |
| *True Negative* | *Logit* | 99.71% | 99.71% | 99.71% | 99.37% |
| *True Negative* | *Random Forest* | 100% | 100% | 100% | 100% |
| *True Negative* | *Deep Learning* | 99.37% | 99.37% | 99.37% | 99.79% |
| *True Negative* | *Human Evaluation* | 88.32% | 74.28% | 74.28% | 60.29% |
| *True Positive* | *Logit* | 5.59% | 5.59% | 5.59% | 6.38% |
| *True Positive* | *Random Forest* | 1.86% | 1.86% | 1.86% | 4.25% |
| *True Positive* | *Deep Learning* | 14.29% | 14.29% | 14.29% | 10.64% |
| *True Positive* | *Human Evaluation* | 8.70% | 8.70% | 18.63% | 25.53% |
| *False Positive* | *Logit* | 0.29% | 0.29% | 0.29% | 0.06% |
| *False Positive* | *Random Forest* | 0% | 0% | 0% | 0% |
| *False Positive* | *Deep Learning* | 0.63% | 0.63% | 0.63% | 0.02% |
| *False Positive* | *Human Evaluation* | 11.68% | 11.68% | 25.72% | 34.59% |
| *Overall Accuracy* | *Logit* | 91.73% | 91.73% | 91.73% | 91.03% |
| *Overall Accuracy* | *Random Forest* | 91.68% | 91.68% | 91.68% | 91.41% |
| *Overall Accuracy* | *Deep Learning* | 92.15% | 92.15% | 92.15% | 91.79% |
| *Overall Accuracy* | *Human Evaluation* | 81.57% | 68.72% | 69.56% | 57.06% |
| ***Proportion of fakes is revealed*** | | No | No | No | Yes |
| ***Number of evaluators*** *(Each evaluator is shown 30 randomly drawn reviews from a pool of 1,900 reviews)* | | 199 | 199 | 199 | 20 |

**Table 2.6: Examples of Reviews Detected Exclusively by Deep Learning**

| S.No. | Type | Review |
|---|---|---|
| 1 | Modest praise for the doctor | "The hospitality service was good and the medicine he provided was good and the charges are reasonable. Thank you Dr. [Doctor's name]" |
| 2 | | "I am patient of [clinic name] for five months and took medicine for headache. He is polite and kind and treated me well. Now my headache is relieved, and doctor advised to stop medicines" |
| 3 | Storyline | "I am suffering from Neck Pain from the last 10 years and back pain from the last 4 months. I have consulted the local doctors regarding the same but I am not satisfied with their treatment. I have taken the medicine from Dr. [doctor name] and I hope I will recover very soon" |
| 4 | | "I have been suffering from migraine for the past 16 years, visited many hospitals but was not relieved. I took homeopathy from the last four years from a local expert. But my trouble was not cured. There after I came to consult Dr. [Doctor's name] for his expert treatment. Presently the severity and frequency of the attacks have come down very much. I thank the doctor very much for his excellent treatment" |
| 5 | | "I had been suffering from fistula for 2 years and now feeling very much better with Dr. [Doctor's name] treatment. I was told by other doctors that surgery is the only treatment. But Dr. [Doctor's name] took good care and within a span of 8 months I am very much better. Thank you, Dr. [Doctor's name]" |
| 7 | Detailed explanation of a patient's condition | "My wife, [name], was having knee pain, so we consulted doctor [Doctor's name] for the same. After using cold laser device and medicines my wife's pain is reduced now and we would like to recommend the treatment to all who are having pain related problem" |
| 8 | | "I am diabetic and I am having burning sensation and numbness and lack sensation in leg and my slippers falls while walking. Consultation with doctor was good. She explained me how to overcome the illness by life style modification and healthy diet. After speaking to doctor I got confidence that I can regain my health" |
| 9 | | "As I am working in the marketing field, my skin is exposed to sun everyday..due to that brown patches increased on my face. Dr. [Doctor's name] treated them very well and those patches were gone in just 2 months Thank you [Clinic's name]" |

**Table 3.1: Motivations of Reviewers**

| ***Study*** | ***Motivations*** |
|---|---|
| Henning-Thurau et al. (2004) | Platform assistance, venting negative feelings, concern for other consumers, extraversion/positive self-enhancement, social benefits, economic incentives, helping the company, advice seeking |
| Dichter (1966) | Product-involvement, self-involvement, other-involvement, message-involvement |
| Engel et al. (1977) | Involvement, self-enhancement, concern for others, message intrigue, dissonance reduction |
| Sundaram et al. (1998) | Altruism (positive WOM), product involvement, self-enhancement, helping the company, altruism (negative WOM), anxiety reduction, vengeance, advice seeking |
| Wasko and Faraj (2005) | Individual motivations (reputation, enjoy helping), structural capital (centrality), cognitive capital (self-rated expertise, tenure in the field), relational capital (commitment, reciprocity) |

**Table 3.2: Mixed Findings on Impact of Extrinsic Motivation on Prosocial Contribution**

| **Paper** | **Method** | **Context** | **Results** |
|---|---|---|---|
| Mellström and Johannesson (2008) | Field Experiment | Blood Donation | Incentives to complete health exam to become eligible to be blood donors; has negative effect on females and no effect on males |
| Gneezy and Rustichini (2000) | Field Experiment | School Children | Small incentives to collect donation have a negative effect; large incentives have zero effect |
| Ariely et al. (2009) | Lab Experiment | Undergraduate Students | Effort in volunteer task has zero effect when reputations are unknown to experimenter and |

| | | | |
|---|---|---|---|
| | | | students |
| Iajya et al. (2013) | Field Experiment | New Blood Donors | Small supermarket vouchers have zero effect and larger vouchers encourage more donations |
| Carpenter and Myers (2010) | Observational Data | Blood Donation in Firefighters | Small stipends increase turnout rate when reputations are known amongst firefighters |
| Lacetera and Macis (2010) | Observational Data and Field Experiment | Previous Blood Donors | Material incentives increase donations |
| Becker et al. (2014) | Natural Field Experiment | New and Previous Blood Donors | Removing monetary compensation decreases donation rate for frequent donors |
| Lacetera et al. (2014) | Field Experiment | Previous Donors | Gift cards increase the rate of donations amongst previous donors |

## Table 3.3: Summary Statistics Message Framing

| Variable | Description | N | Mean | S.D. | Min | Max | Value = 1 |
|---|---|---|---|---|---|---|---|
| *Mail_open* | Email opened dummy | 2,493,980 | 0.23 | 0.42 | 0 | 1 | 564,893 |
| *Reco* | Recommendation posted dummy | 2,493,980 | 0.07 | 0.26 | 0 | 1 | 178,120 |
| *Doc_rev* | Doctor review posted dummy | 2,493,980 | 0.01 | 0.10 | 0 | 1 | 25,429 |
| *Clin_rev* | Clinic review posted dummy | 2,493,980 | 0.002 | 0.05 | 0 | 1 | 6,575 |

## Table 3.4: Summary Statistics for Recommendation

| Variable | Description | N | Mean | S.D. | Min | Max | Value = 1 |
|---|---|---|---|---|---|---|---|
| *Pos_reco* | Patient recommends for the doctor (thumbs up) | 178,120 | 0.91 | 0.27 | 0 | 1 | 163,021 |
| *Neg_reco* | Patient recommends against the doctor (thumbs down) | 178,120 | 0.05 | 0.21 | 0 | 1 | 8,751 |
| *No_show* | Patient could not meet the doctor | 178,120 | 0.04 | 0.18 | 0 | 1 | 6,348 |

## Table 3.5: Summary Statistics for Doctor Reviews

| Variable | Description | N | Mean | S.D. | Min | Max |
|---|---|---|---|---|---|---|
| *Length* | Review Length | 15107 | 194.50 | 109.05 | 100 | 789 |
| *Identity_revelation* | Dummy variable | 15107 | 0.70 | 0.46 | 0 | 1 |
| *Review_views* | Count Variable | 15107 | 66.61 | 106.85 | 0 | 2971 |
| *Privacy* | Privacy dummy | 15107 | 0.50 | 0.50 | 0 | 1 |
| *Sensitivity* | Sensitivity dummy | 2493 | 0.59 | 0.49 | 0 | 1 |

**Table 3.6: Effect of Email Framing on Opening the Email, Giving a Recommendation and Posting a Review**

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | *Mail_open* | *Reco* | *Doc_rev* | *Clin_rev* |
| *Altruism_basic* | -0.17*** | -0.30*** | -0.31*** | -0.29*** |
| | (0.00) | (0.01) | (0.02) | (0.04) |
| *Altruism_strengthened* | -0.19*** | -0.31*** | -0.25*** | -0.21*** |
| | (0.00) | (0.01) | (0.02) | (0.04) |
| *Product_involvement* | -0.05*** | -0.11*** | -0.08*** | -0.04 |
| | (0.00) | (0.01) | (0.02) | (0.04) |
| *Warm_glow* | -0.13*** | -0.23*** | -0.18*** | -0.17*** |
| | (0.00) | (0.01) | (0.02) | (0.04) |
| *Constant* | -1.12*** | -2.38*** | -4.42*** | -5.80*** |
| | (0.00) | (0.01) | (0.01) | (0.03) |
| *Observations* | 2493980 | 2493980 | 2493980 | 2493980 |
| *Pseudo R-squared* | 0.00 | 0.00 | 0.00 | 0.00 |
| *Log Likelihood* | -1333253.53 | -640499.99 | -141747.07 | -45572.92 |
| *Model* | Logit | Logit | Logit | Logit |

Standard errors are in parentheses. *** p<0.01, ** p<0.05, * p<0.1

**Table 3.7: Effect of Email Framing on Recommendation and Review Valence**

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | *Pos_reco* | *Neg_reco* | *Pos_doc_rev* | *Neg_doc_rev* |
| *Altruism_basic* | -0.30*** | -0.32*** | -0.31*** | -0.34*** |
| | (0.01) | (0.03) | (0.02) | (0.10) |
| *Altruism_strengthened* | -0.31*** | -0.39*** | -0.25*** | -0.25** |
| | (0.01) | (0.03) | (0.02) | (0.10) |
| *Product_involvement* | -0.11*** | -0.12*** | -0.08*** | -0.10 |
| | (0.01) | (0.03) | (0.02) | (0.10) |
| *Warm_glow* | -0.22*** | -0.27*** | -0.18*** | -0.19* |
| | (0.01) | (0.03) | (0.02) | (0.10) |
| *Constant* | -2.48*** | -5.44*** | -4.46*** | -7.68*** |
| | (0.01) | (0.02) | (0.01) | (0.07) |
| *Observations* | 2493980 | 2493980 | 2493980 | 2493980 |
| *Pseudo R-squared* | 0.00 | 0.00 | 0.00 | 0.00 |
| *Log Likelihood* | -601187.00 | -58112.04 | -137264.52 | -8634.39 |
| *Model* | Logit | Logit | Logit | Logit |

Standard errors are in parentheses. *** p<0.01, ** p<0.05, * p<0.1

**Table 3.8: Effect of Email Framing on Valence Conditional on Posting a Recommendation**

| | (1) *Pos_reco* | (2) *Neg_reco* | (3) *Pos_doc_rev* | (4) *Neg_doc_rev* |
|---|---|---|---|---|
| *Altruism_basic* | -0.02 | -0.04 | -0.03 | -0.07 |
| | (0.03) | (0.03) | (0.02) | (0.10) |
| *Altruism_strengthened* | 0.05* | -0.10*** | 0.05** | 0.04 |
| | (0.03) | (0.03) | (0.02) | (0.10) |
| *Product_involvement* | -0.01 | -0.02 | 0.03* | 0.01 |
| | (0.03) | (0.03) | (0.02) | (0.10) |
| *Warm_glow* | 0.01 | -0.06* | 0.03 | 0.02 |
| | (0.03) | (0.03) | (0.02) | (0.10) |
| *Constant* | 2.37*** | -2.92*** | -1.85*** | -5.20*** |
| | (0.02) | (0.02) | (0.01) | (0.07) |
| *Observations* | 178120 | 178120 | 178120 | 178120 |
| *Pseudo R-squared* | 0.00 | 0.00 | 0.00 | 0.00 |
| *Log Likelihood* | -51698.54 | -34896.45 | -71238.16 | -6059.71 |
| *Model* | Logit | Logit | Logit | Logit |

Standard errors are in parentheses. *** p<0.01, ** p<0.05, * p<0.1

**Table 3.9: Effect of Email Framing on Review Quality Conditional on Posting a Positive Review**

| | (1) *Identity_revelation* | (2) *Length* | (3) *Review_views* |
|---|---|---|---|
| *Altruism_basic* | -0.09** | 5.28* | -0.29 |
| | (0.04) | (2.83) | (2.00) |
| *Altruism_strengthened* | -0.06 | 6.55** | -0.93 |
| | (0.04) | (2.81) | (1.85) |
| *Product_involvement* | -0.01 | 2.35 | -1.70 |
| | (0.04) | (2.66) | (1.68) |
| *Warm_glow* | -0.02 | 7.33*** | -0.05 |
| | (0.04) | (2.80) | (1.74) |
| *Constant* | 0.66*** | 197.85*** | 44.78*** |
| | (0.03) | (1.74) | (1.24) |
| *Observations* | 24452 | 24452 | 21582 |
| *R-squared* | N/A | 0.00 | 0.00 |
| *Model* | Logit | OLS | OLS |

Standard errors are in parentheses. *** p<0.01, ** p<0.05, * p<0.1

**Table 3.10: Effect of Email Framing on Review Quality Conditional on Posting a Negative Review**

| | (1) *Length* | (2) *Review_views* |
|---|---|---|
| *Altruism_basic* | 24.78 | 6.68 |
| | (23.87) | (12.88) |
| *Altruism_strengthened* | -1.97 | -6.40 |
| | (22.12) | (10.55) |
| *Product_involvement* | 11.08 | 4.32 |
| | (20.73) | (11.09) |
| *Warm_glow* | 57.08** | -4.62 |
| | (26.96) | (11.58) |
| *Constant* | 325.69*** | 76.30*** |
| | (13.09) | (7.55) |
| *Observations* | 977 | 926 |
| *R-squared* | 0.01 | 0.00 |
| *Model* | OLS | OLS |

Standard errors are in parentheses. *** p<0.01, ** p<0.05, * p<0.1

**Table 3.11: Summary Statistics: Privacy Assurance**

| | N | Mean | S.D. | Min | Max | P1 | P25 | P75 | P99 |
|---|---|---|---|---|---|---|---|---|---|
| *Length* | 15107.00 | 194.50 | 109.05 | 100.00 | 789.00 | 100.00 | 122.00 | 227.00 | 616.00 |
| *Identity_revelation* | 15107.00 | 0.70 | 0.46 | 0.00 | 1.00 | 0.00 | 0.00 | 1.00 | 1.00 |
| *Review_views* | 15107.00 | 66.61 | 106.85 | 0.00 | 2971.00 | 0.00 | 13.00 | 80.00 | 469.00 |
| *Privacy* | 15107.00 | 0.50 | 0.50 | 0.00 | 1.00 | 0.00 | 0.00 | 1.00 | 1.00 |
| *Sensitivity* | 2493.00 | 0.59 | 0.49 | 0.00 | 1.00 | 0.00 | 0.00 | 1.00 | 1.00 |
| *Altruism_basic* | 15107.00 | 0.17 | 0.38 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| *Altruism_strengthened* | 15107.00 | 0.18 | 0.39 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| *Product_involvement* | 15107.00 | 0.21 | 0.41 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| *Warm_glow* | 15107.00 | 0.20 | 0.40 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| *Compound* | 11672.00 | 0.63 | 0.37 | -0.99 | 1.00 | -0.66 | 0.50 | 0.89 | 0.98 |

**Table 3.12: Effect of Privacy Enhancement on Identity Disclosure**

| | (1) *Identity_revelation* | (2) *Identity_revelation* |
|---|---|---|
| *Privacy* | 1.20*** | 1.04*** |
| | (0.04) | (0.15) |
| *Sensitivity* | | -0.78*** |
| | | (0.12) |
| *PrivacyXsensitivity* | | 0.35* |
| | | (0.19) |
| *Constant* | 0.30*** | 0.65*** |
| | (0.02) | (0.09) |
| *Observations* | 15107 | 2493 |
| *Pseudo R-squared* | 0.06 | 0.08 |
| *Log Likelihood* | -8744.12 | -1451.28 |
| *Model* | Logit | Logit |

Standard errors are in parentheses. *** p<0.01, ** p<0.05, * p<0.1

**Table 3.13: Effect of Privacy Enhancement on Review Quality**

| | (1)<br>*Length* | (2)<br>*Length* | (3)<br>*Review_views* | (4)<br>*Review_views* | (5)<br>*VADER:_Compound_Score* | (6)<br>*VADER:_Compound_Score* |
|---|---|---|---|---|---|---|
| *Privacy* | 1.47 | -0.80 | 9.87*** | 6.09** | -0.01 | 0.01 |
| | (1.77) | (6.71) | (1.74) | (2.59) | (0.01) | (0.03) |
| *Sensitivity* | | -1.16 | | 76.74*** | | 0.03 |
| | | (6.06) | | (4.99) | | (0.03) |
| *PrivacyXsensitivity* | | 11.56 | | 27.19*** | | -0.04 |
| | | (8.80) | | (9.30) | | (0.04) |
| *Constant* | 193.77*** | 191.78*** | 61.69*** | 34.86*** | 0.63*** | 0.59*** |
| | (1.24) | (4.75) | (1.11) | (1.67) | (0.00) | (0.02) |
| *Observations* | 15107 | 2493 | 15107 | 2493 | 11672 | 1951 |
| *R-squared* | 0.00 | 0.00 | 0.00 | 0.11 | 0.00 | 0.00 |
| *Model* | OLS | OLS | OLS | OLS | OLS | OLS |

Standard errors are in parentheses. *** p<0.01, ** p<0.05, * p<0.1

**Table 3.14: Robustness Check 1: Rewording the Message (Effect on Identity Disclosure)**

| | (1)<br>*Identity_revelation* | (2)<br>*Identity_revelation* |
|---|---|---|
| *Privacy* | 1.00*** | 0.60** |
| | (0.06) | (0.24) |
| *Sensitivity* | | -1.12*** |
| | | (0.19) |
| *PrivacyXsensitivity* | | 0.47 |
| | | (0.29) |
| *Constant* | 0.39*** | 1.15*** |
| | (0.03) | (0.16) |
| *Observations* | 6829 | 1172 |
| *Pseudo R-squared* | 0.04 | 0.07 |
| *Log Likelihood* | -4014.81 | -669.36 |
| *Model* | Logit | Logit |

Standard errors are in parentheses. *** p<0.01, ** p<0.05, * p<0.1

**Table 3.15: Robustness Check 1: Rewording the Message (Effect on Quality)**

| | (1)<br>*Length* | (2)<br>*Length* | (3)<br>*Review_views* | (4)<br>*Review_views* | (5)<br>*VADER:_Compound_Score* | (6)<br>*VADER:_Compound_Score* |
|---|---|---|---|---|---|---|
| *Privacy* | 5.37** | 5.40 | 1.49* | -1.43 | 0.00 | 0.01 |
| | (2.66) | (10.16) | (0.81) | (1.10) | (0.01) | (0.04) |
| *Sensitivity* | | -3.72 | | 27.97*** | | -0.03 |
| | | (8.52) | | (2.54) | | (0.04) |
| *PrivacyXsensitivity* | | 13.27 | | 5.57 | | 0.02 |
| | | (12.92) | | (3.58) | | (0.05) |
| *Constant* | 191.04*** | 188.94*** | 20.85*** | 10.84*** | 0.62*** | 0.61*** |
| | (1.83) | (7.04) | (0.56) | (0.85) | (0.01) | (0.03) |
| *Observations* | 6829 | 1172 | 6829 | 1172 | 5331 | 908 |
| *R-squared* | 0.00 | 0.00 | 0.00 | 0.15 | 0.00 | 0.00 |
| *Model* | OLS | OLS | OLS | OLS | OLS | OLS |

Standard errors are in parentheses. *** p<0.01, ** p<0.05, * p<0.1

**Table 3.16: Falsification Test: Repositioning the Checkbox (Effect on Identity Disclosure)**

| | (1) *Identity_revelation* | (2) *Identity_revelation* |
|---|---|---|
| *Privacy* | 0.16* | -0.22 |
| | (0.09) | (0.31) |
| *Sensitivity* | | -0.83*** |
| | | (0.31) |
| *PrivacyXsensitivity* | | 0.54 |
| | | (0.42) |
| *Constant* | 0.37*** | 0.85*** |
| | (0.06) | (0.22) |
| *Observations* | 2241 | 404 |
| *Pseudo R-squared* | 0.00 | 0.02 |
| *Log Likelihood* | -1496.56 | -265.59 |
| *Model* | Logit | Logit |

Standard errors are in parentheses. *** p<0.01, ** p<0.05, * p<0.1

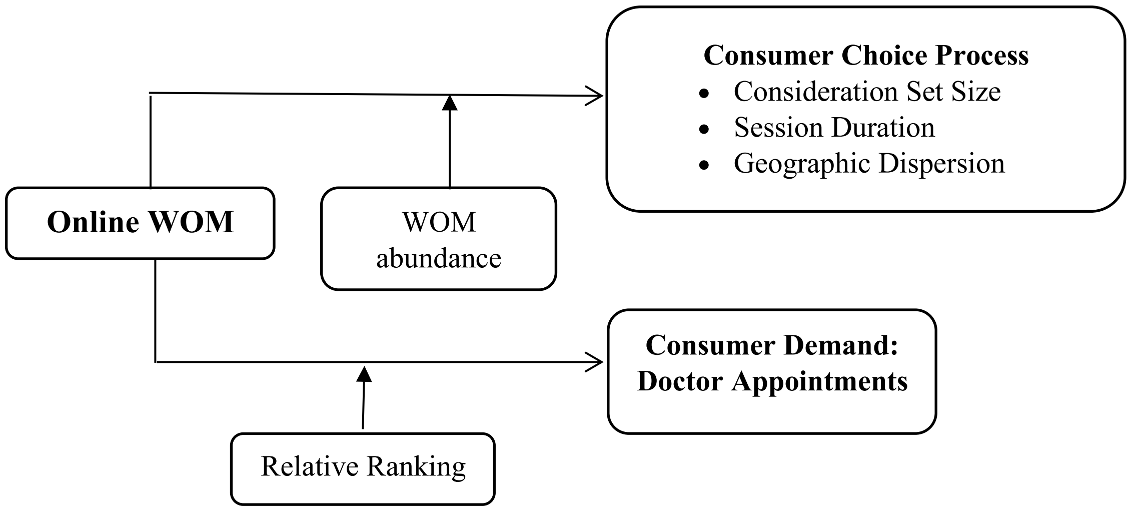
**Table 3.17: Falsification Test: Repositioning the Checkbox (Effect on Review Quality)**

| | (1) *Length* | (2) *Length* | (3) *Review_views* | (4) *Review_views* | (5) *VADER:_Compound_Score* | (6) *VADER:_Compound_Score* |
|---|---|---|---|---|---|---|
| *Privacy* | 9.00* | 7.72 | -0.94 | -0.42 | 0.03 | 0.04 |
| | (4.85) | (15.73) | (0.94) | (1.13) | (0.02) | (0.07) |
| *Sensitivity* | | -16.29 | | 23.31*** | | 0.01 |
| | | (13.30) | | (3.79) | | (0.06) |
| *PrivacyXsensitivity* | | 17.62 | | -4.52 | | -0.01 |
| | | (20.69) | | (4.68) | | (0.09) |
| *Constant* | 192.80*** | 191.54*** | 13.86*** | 6.51*** | 0.61*** | 0.59*** |
| | (3.31) | (10.34) | (0.68) | (0.80) | (0.01) | (0.05) |
| *Observations* | 2241 | 404 | 2241 | 404 | 1739 | 313 |
| *R-squared* | 0.00 | 0.01 | 0.00 | 0.17 | 0.00 | 0.00 |
| *Model* | OLS | OLS | OLS | OLS | OLS | OLS |

Standard errors are in parentheses. *** p<0.01, ** p<0.05, * p<0.1

# FIGURES

## Figure 1.1: Conceptual Model




## Figure 1.2: Event Timeline

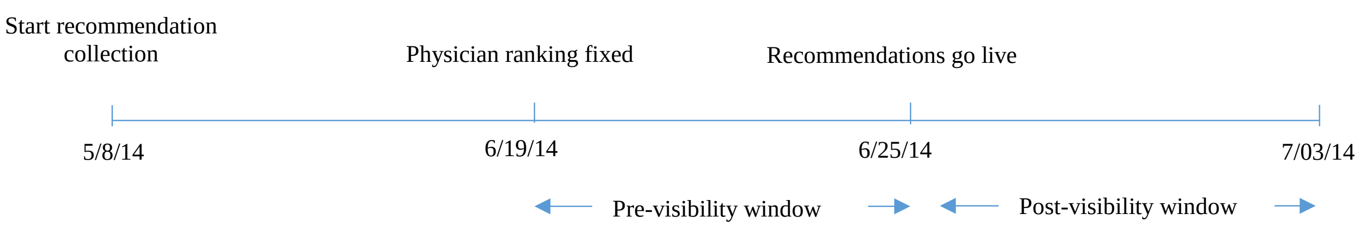


**Figure 1.3: Marginal Effects: Consideration Set**

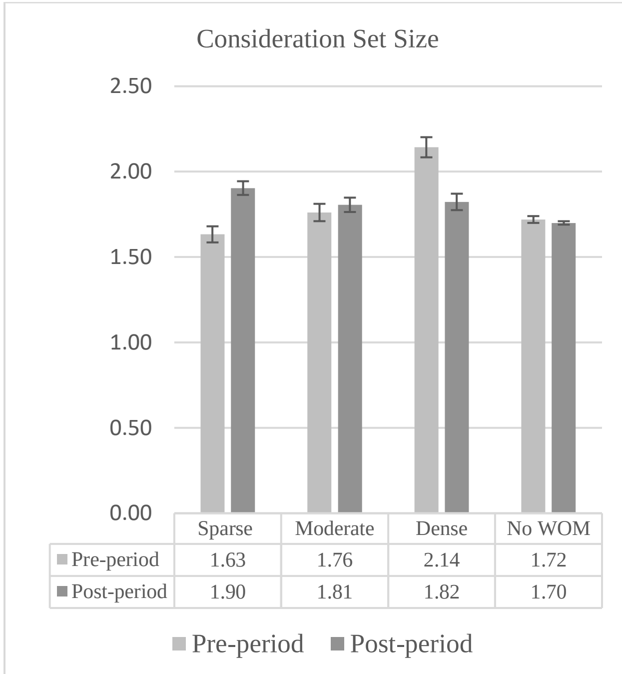
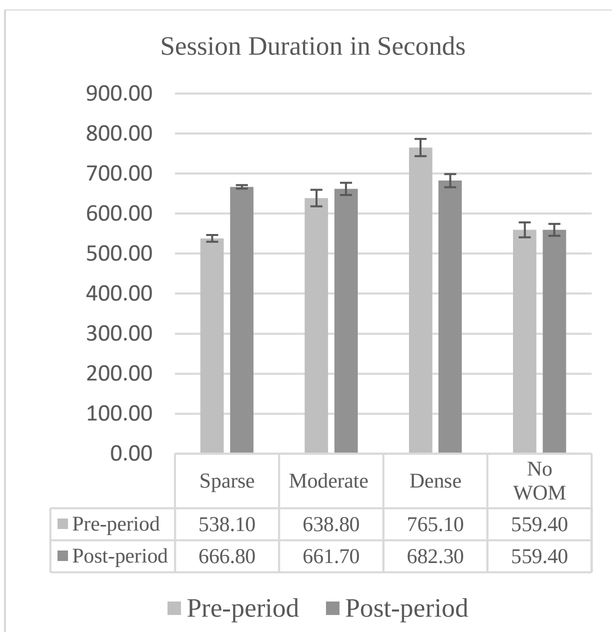
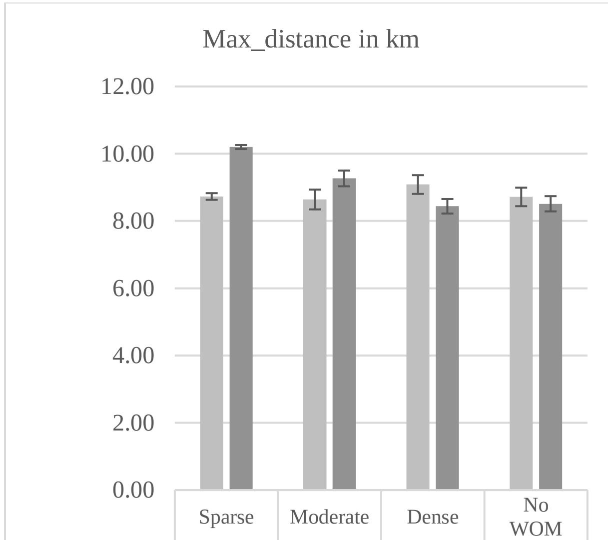



**Figure 1.4: Marginal Effects: Session Duration**

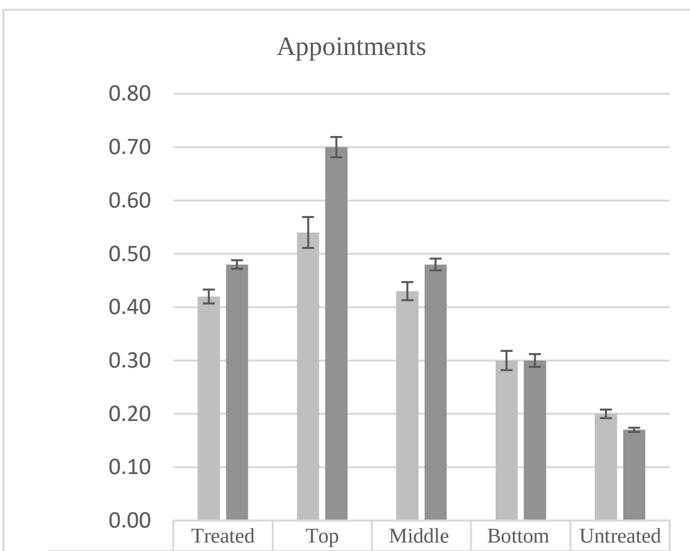


## Figure 1.5: Marginal Effects: Geographic Dispersion




| | Sparse | Moderate | Dense | No WOM |
|---|---|---|---|---|
| Pre-period | 8.73 | 8.64 | 9.09 | 8.72 |
| Post-period | 10.20 | 9.27 | 8.44 | 8.51 |

## Figure 1.6: Marginal Effects: Patient Demand




| | Treated | Top | Middle | Bottom | Untreated |
|---|---|---|---|---|---|
| Pre-period | 0.42 | 0.54 | 0.43 | 0.30 | 0.20 |
| Post-period | 0.48 | 0.70 | 0.48 | 0.30 | 0.17 |

**Figure 1.7: Online Appointments (Trend Line)**

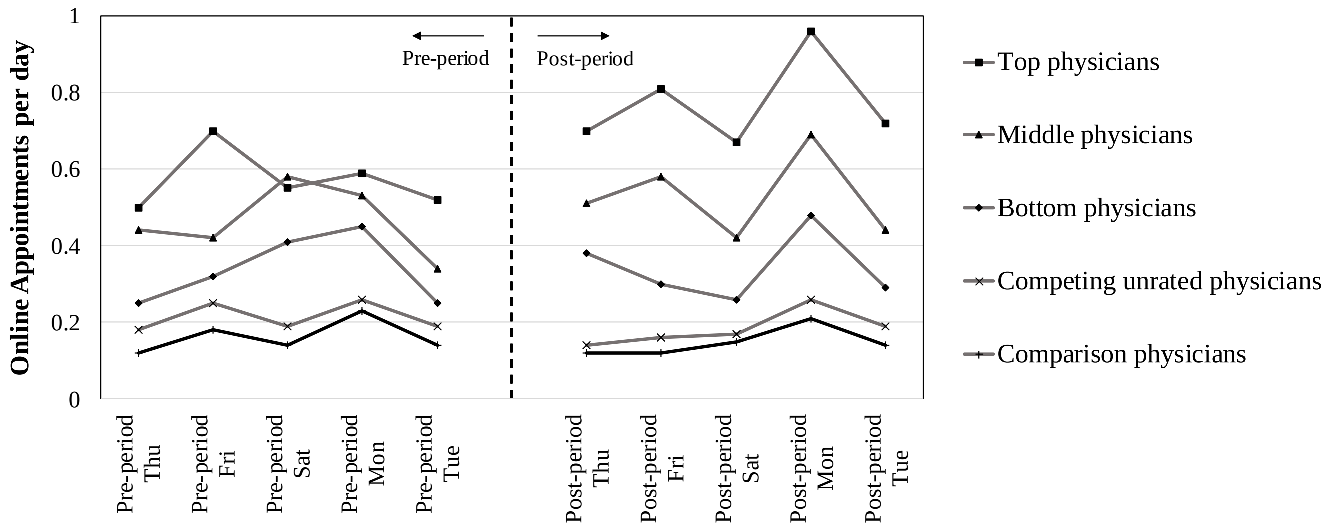



**Figure 2.1: Recurrent Neural Network (RNN) Architecture**

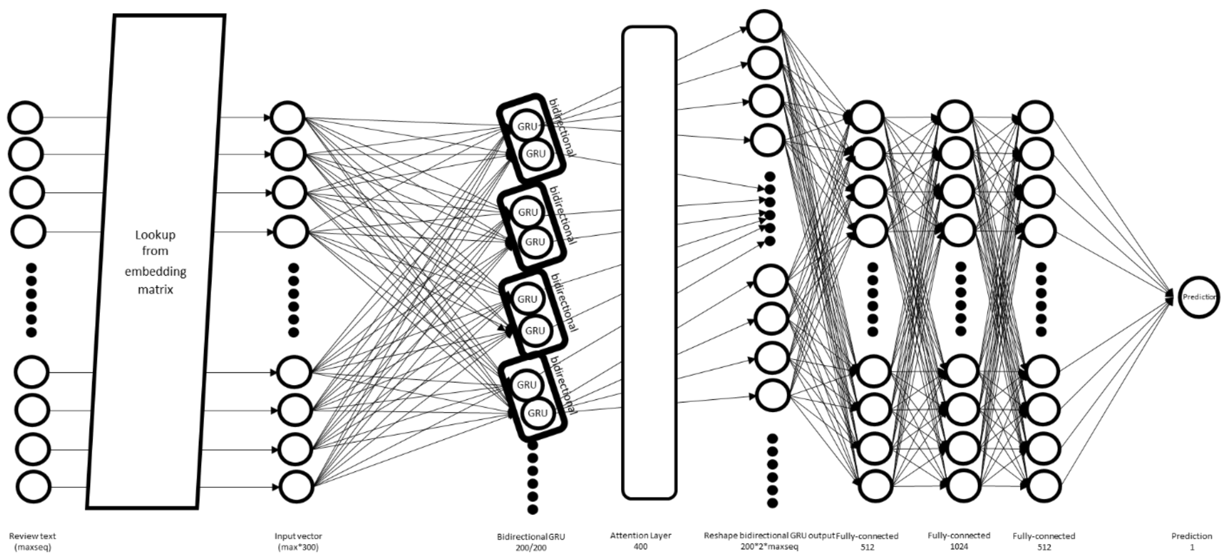


**Figure 2.2: ROC Curve of the Three Models**

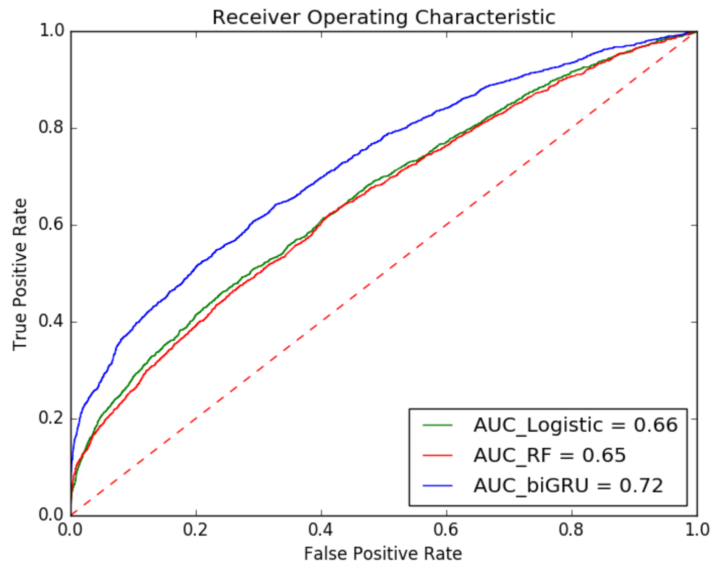



**Figure 3.1: Sample Doctor Review with a Reply from the Healthcare Provider**

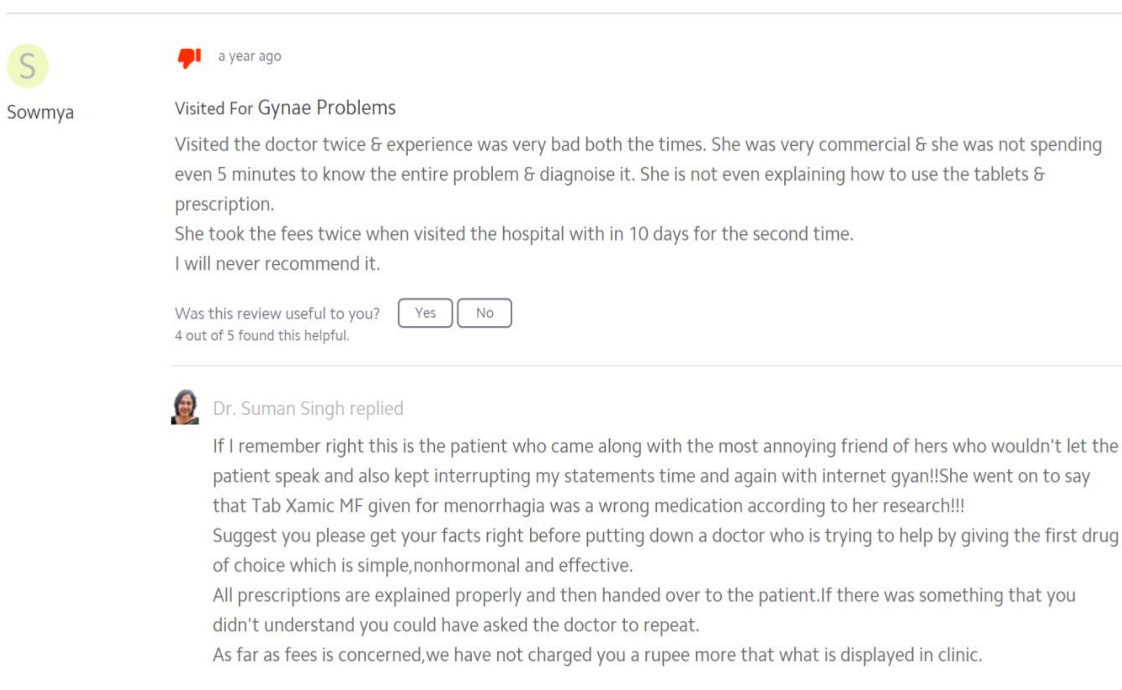
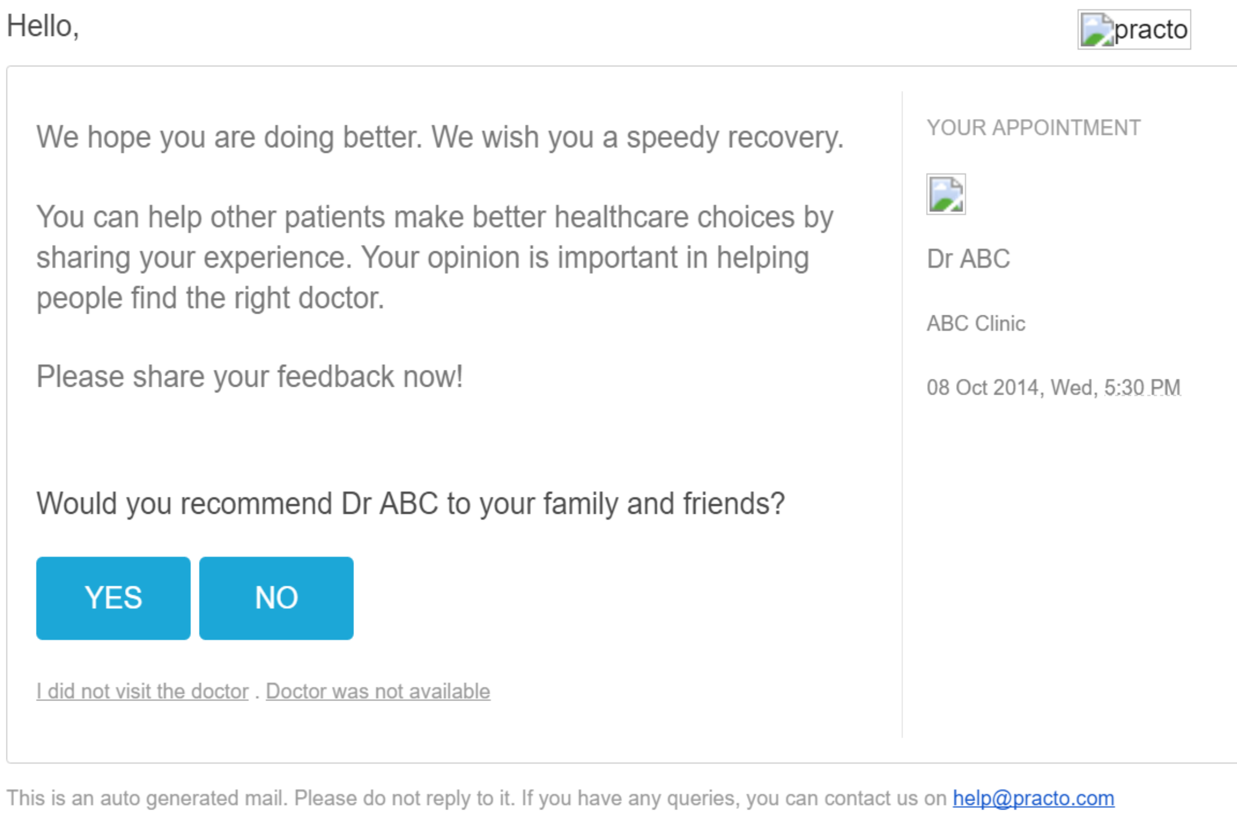


**Figure 3.2: Email Message Framing: Motivation – Altruism**

Email Subject: Help others by sharing your feedback for Dr. ABC



Hello,

practo

We hope you are doing better. We wish you a speedy recovery.

You can help other patients make better healthcare choices by sharing your experience. Your opinion is important in helping people find the right doctor.

Please share your feedback now!

Would you recommend Dr ABC to your family and friends?

YES NO

I did not visit the doctor . Doctor was not available

YOUR APPOINTMENT

Dr ABC

ABC Clinic

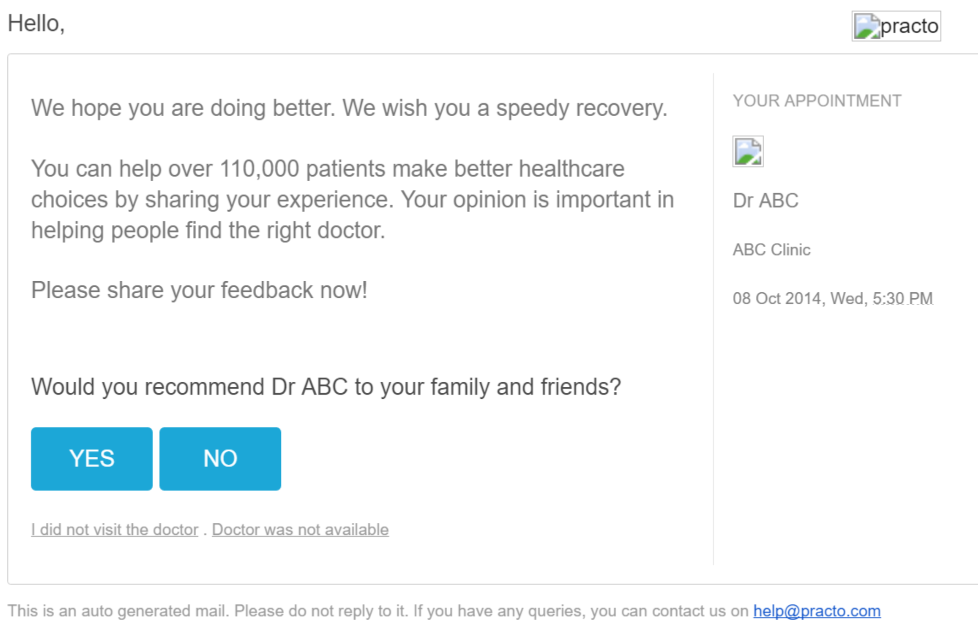
08 Oct 2014, Wed, 5:30 PM

This is an auto generated mail. Please do not reply to it. If you have any queries, you can contact us on help@practo.com

Note: These figures are snapshots of test emails, and the company logo and doctor's photo are hidden. The actual email message seen by the patient is very similar, with a real doctor's name, the company's logo and the doctor's photo. The user interface and the text is exactly the same.

**Figure 3.3: Email Message Framing: Motivation – Altruism without Data**

Email Subject: Help thousands by sharing your feedback for Dr. ABC


Hello,

practo

We hope you are doing better. We wish you a speedy recovery.

You can help over 110,000 patients make better healthcare choices by sharing your experience. Your opinion is important in helping people find the right doctor.

Please share your feedback now!

Would you recommend Dr ABC to your family and friends?

YES NO

I did not visit the doctor . Doctor was not available

YOUR APPOINTMENT

Dr ABC

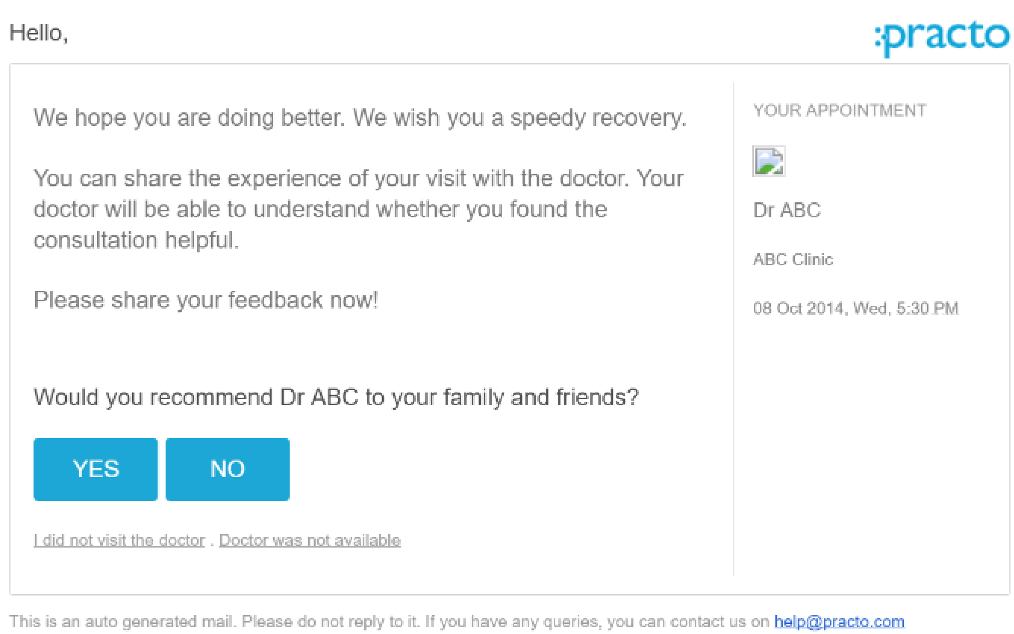
ABC Clinic

08 Oct 2014, Wed, 5:30 PM

This is an auto generated mail. Please do not reply to it. If you have any queries, you can contact us on help@practo.com

**Figure 3.4: Email Message Framing: Product (Healthcare Provider) Involvement**

Email Subject: Please share your feedback for Dr. ABC


Hello,

:practo

We hope you are doing better. We wish you a speedy recovery.

You can share the experience of your visit with the doctor. Your doctor will be able to understand whether you found the consultation helpful.

Please share your feedback now!

Would you recommend Dr ABC to your family and friends?

YES NO

I did not visit the doctor . Doctor was not available

YOUR APPOINTMENT

Dr ABC

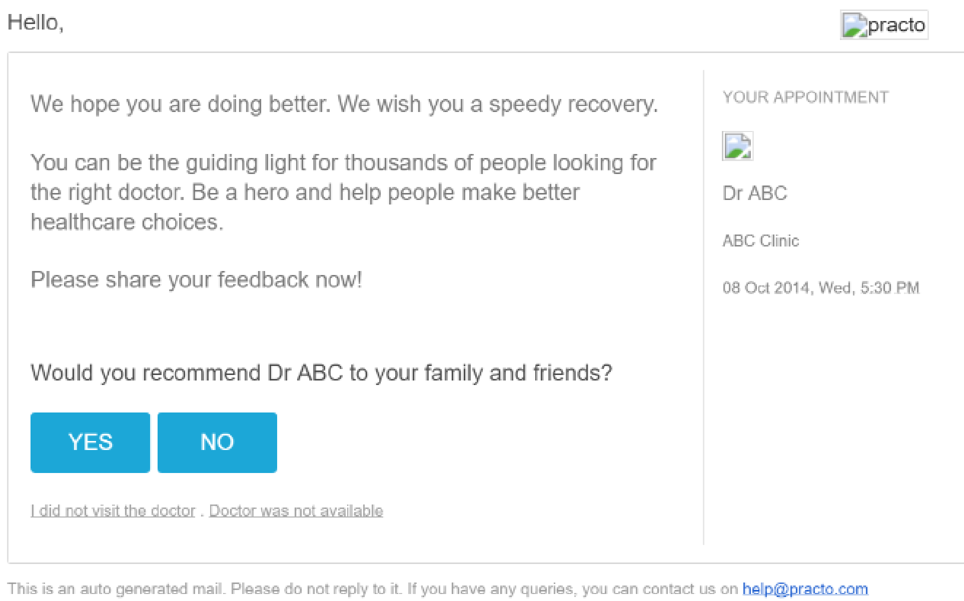
ABC Clinic

08 Oct 2014, Wed, 5:30 PM

This is an auto generated mail. Please do not reply to it. If you have any queries, you can contact us on help@practo.com

**Figure 3.5: Email Message Framing: Self Enhancement or Warm Glow**

Email Subject: People would love to hear your feedback for Dr. ABC



Hello,

practo

We hope you are doing better. We wish you a speedy recovery.

You can be the guiding light for thousands of people looking for the right doctor. Be a hero and help people make better healthcare choices.

Please share your feedback now!

Would you recommend Dr ABC to your family and friends?

YES NO

I did not visit the doctor . Doctor was not available

YOUR APPOINTMENT

Dr ABC

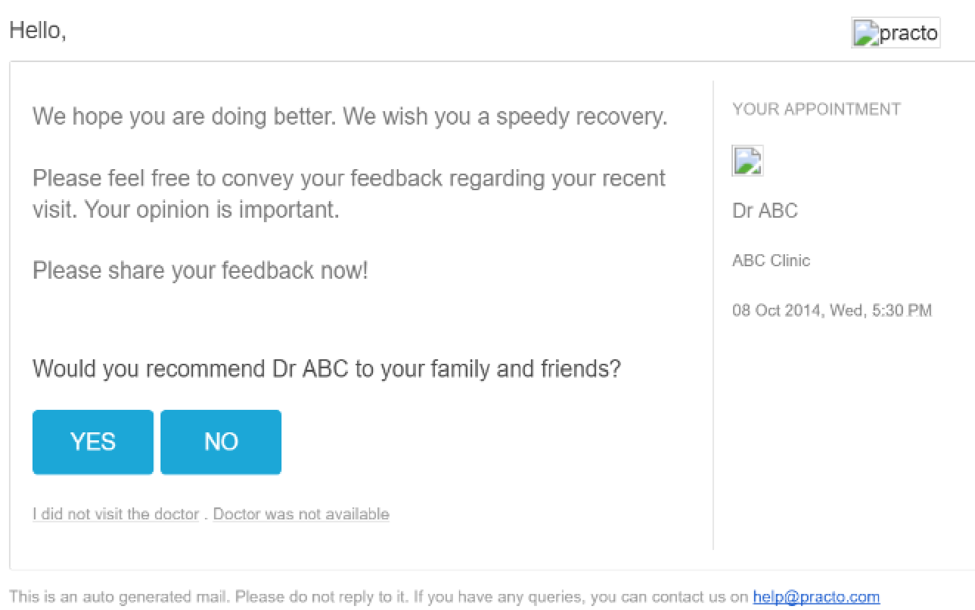
ABC Clinic

08 Oct 2014, Wed, 5:30 PM

This is an auto generated mail. Please do not reply to it. If you have any queries, you can contact us on help@practo.com

**Figure 3.6: Email Message Framing: Non-motivational Framing or Control Group**

Email Subject: Give your feedback for Dr. ABC



Hello,

practo

We hope you are doing better. We wish you a speedy recovery.

Please feel free to convey your feedback regarding your recent visit. Your opinion is important.

Please share your feedback now!

Would you recommend Dr ABC to your family and friends?

YES NO

I did not visit the doctor . Doctor was not available

YOUR APPOINTMENT

Dr ABC

ABC Clinic

08 Oct 2014, Wed, 5:30 PM

This is an auto generated mail. Please do not reply to it. If you have any queries, you can contact us on help@practo.com

**Figure 3.7: Privacy Assurance: Treatment Group – Enhanced Privacy**

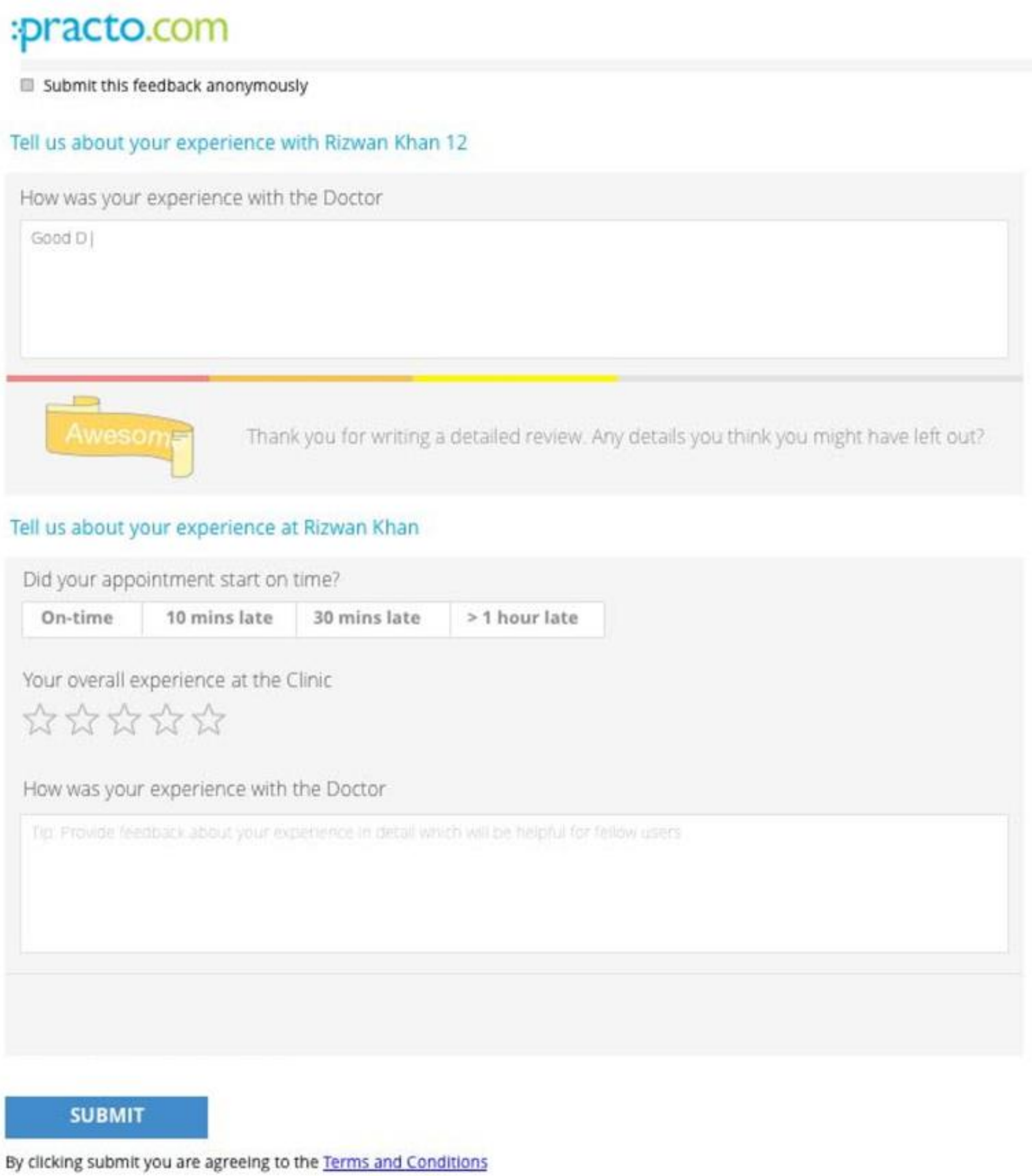
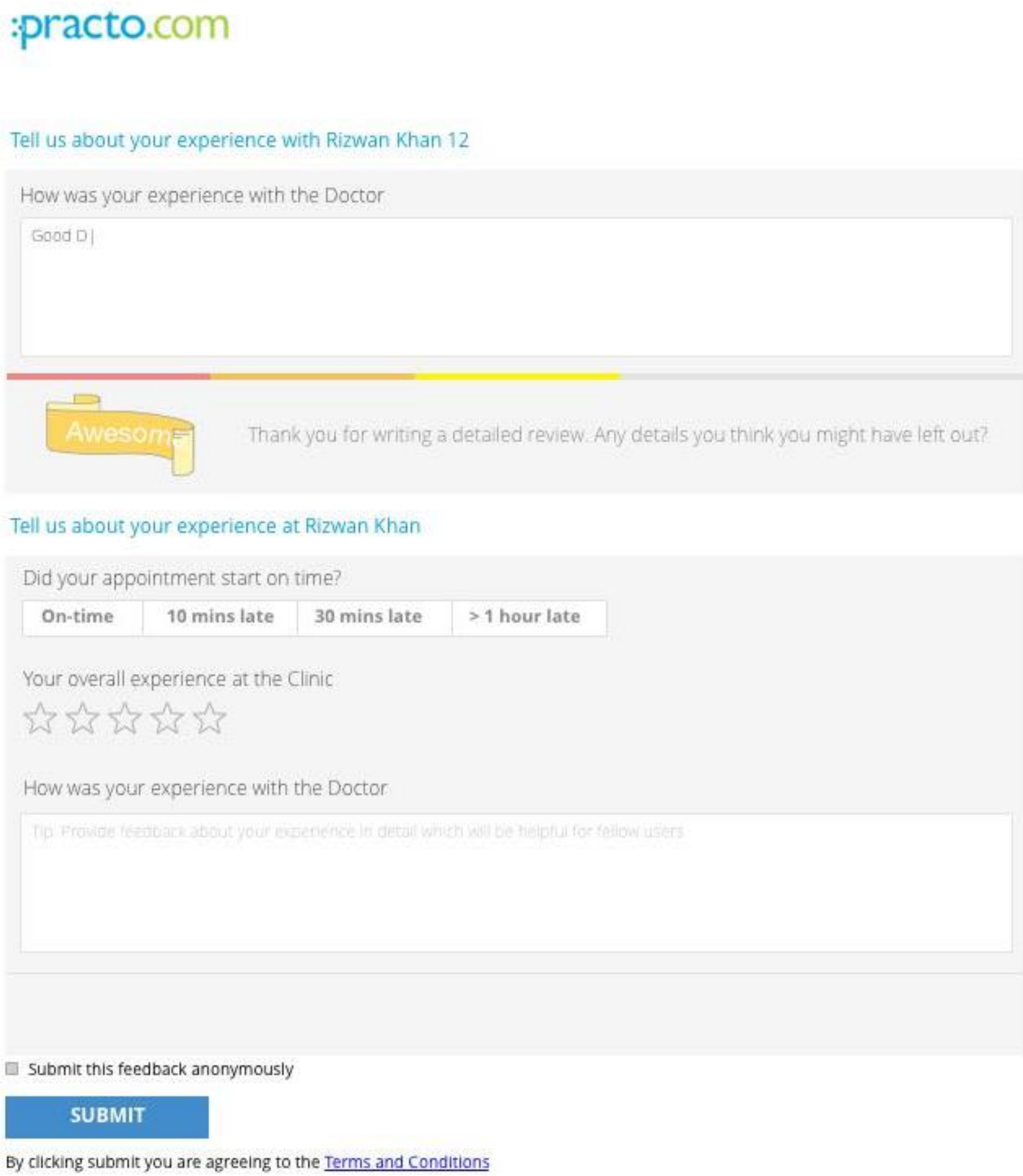


**Figure 3.8: Privacy Assurance: Treatment Group – Usual Privacy**



## APPENDIX

### Figure A1. Sample Email for Solicitation of Recommendation

Date: Tue, Aug 5, 2014 at 2:38 PM
Subject: Feedback for Dr. CD

:practo

**Hi Jesse Zacharias,**

This is regarding the appointment you booked on Practo.
We would like to get your feedback.

Dr. CD
ABC Dental clinic
Aug 4, 2014 9:00 AM

**Would you recommend the doctor to your family and friends?**

**YES** **NO**

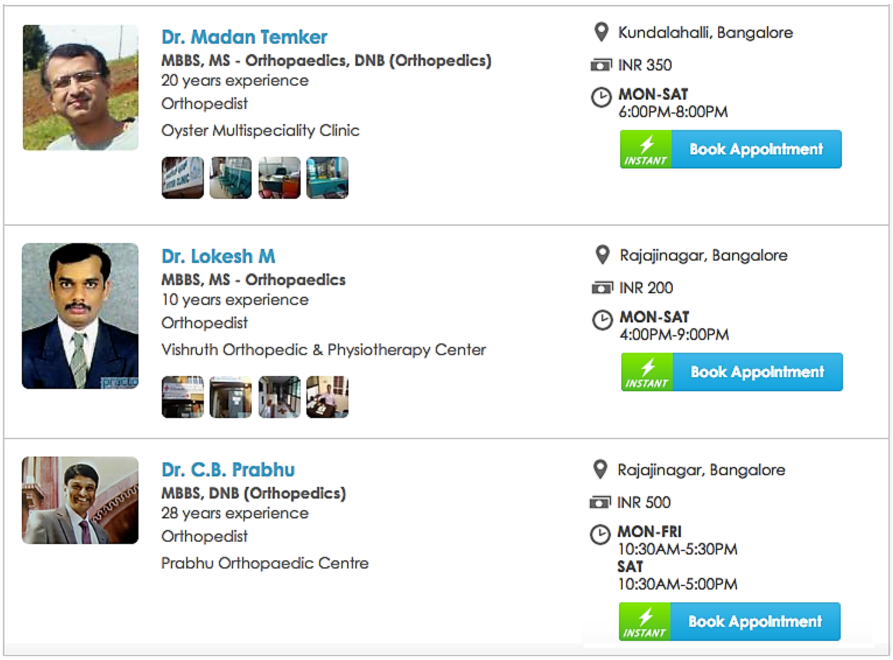
*This recommendation is completely anonymous.*

If you missed your appointment, click here.

This is an auto generated feedback mail. Please don't reply to it. If you want to send us a mail the address is help@practo.com

The Doctor Search Engine
:practo

## Figure A2. Illustrative Snapshots of the Search Page Before and During the Treatment



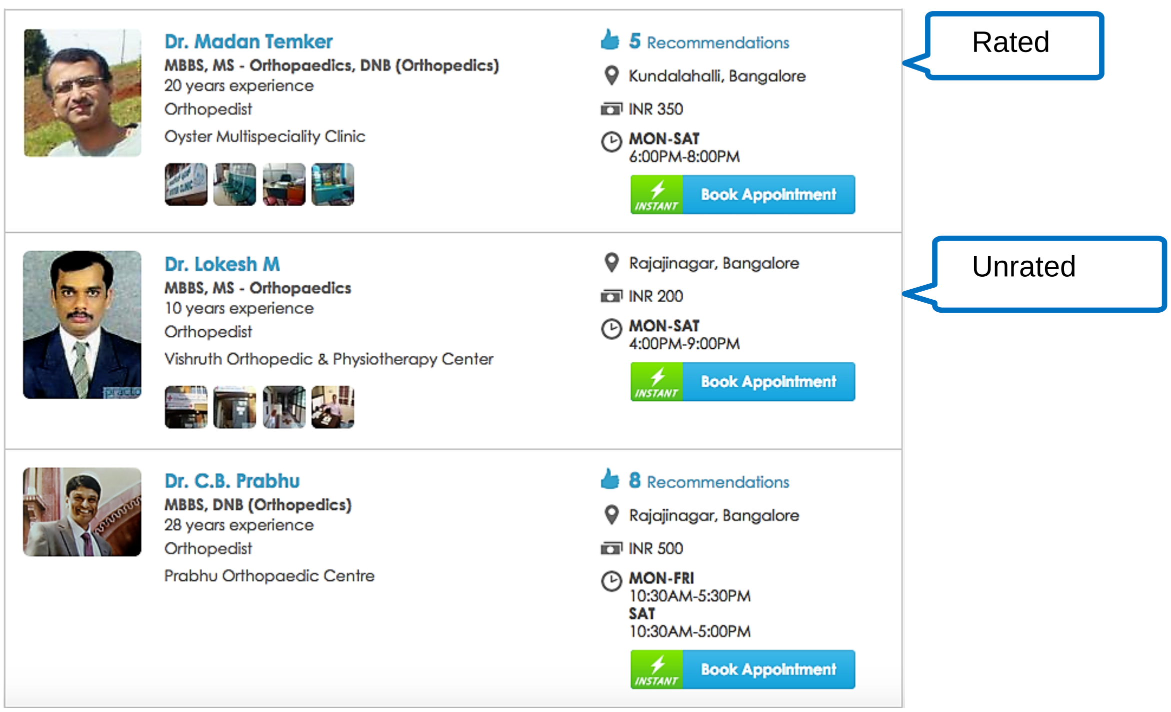



**Figure A3. Snapshot from Doctor Profile Page (Snapshot date: April 30th 2017)**


Note: During the study period, which was the early phase of WOM for this portal, no WOM information was shown on the doctor's profile page.

## Figure A4. An Example Snapshot from a Doctor Profile Page

**Services** View less

- Cervical Cerclage
- In-Vitro Fertilization (IVF)
- Intra-Uterine Insemination (IUI)
- Mirena (Hormonal Iud)
- Artificial Insemination
- Caesarean Section (C Section)
- Hysterectomy (Abdominal/Vaginal)
- Oophorectomy / Ovariectomy / Ovarian Ablation
- Dilatation and Curettage (D & C)
- Tubectomy/Tubal Ligation
- Laparoscopic Surgery
- Normal Vaginal Delivery (NVD)
- Endoscopy
- Intracytoplasmic Sperm Injection (ICSI)
- Sperm Donor Program
- Gynae Problems
- Infertility Evaluation / Treatment
- ART Consultant
- Pap Smear
- Obstetrics / Antenatal Care
- Maternal Care/ Checkup
- Vagina Surgery
- High-Risk Pregnancy Care
- Menopause Clinic
- Diseases in Pregnancy
- Contraception Advice
- Dysmenorrhea Treatment
- Well Woman Healthcheck
- Laparoscopy
- Hysteroscopy
- PCOD/PCOS Treatment
- Ultrasound Scan
- Endoscopic Surgery

**Specializations**

- Gynecologist

**Awards and Recognitions**

- R D Pandit Award for the Best Thesis - 2009

**Education**

- MBBS - PSG Institute of Medical Sciences and Research, Coimbatore, 2005
- MS - Obstetrics & Gynaecology - UCMS (Delhi), 2010
- Fellowship in Minimal Access Surgery - Rajiv Gandhi University of Health Sciences Bangalore, 2011

**Memberships**

- Indian Medical Association (IMA)

**Registrations**

- 96710 Karnataka Medical Council, 2012

**Experience**

- 2010 - 2013 Laparoscopic Surgeon at Dr Ramesh Hospital
- 2007 - 2010 Resident at Gurutegh Bahadur Hospital

**Table A1. Summary of WOM Literature Assessing the Impact of WOM on Product Sales**

| Paper | Context | Nature of Product | Nature of Data Source | Methodology |
|---|---|---|---|---|
| Chen et al. (2004) | Books | Experience Goods | Observational Data | Use linear regression model and do not account for intrinsic quality as a control. They find that online consumer ratings have no association with book sales. |
| Reinstein and Snyder (2005) | Movies | Experience Goods | Observational Data | Exploit the timing difference between critics' ratings and release of the movie and find that the effect is expectedly small but still detectable. |
| Clemons et al. (2006) | Craft Beer | Experience Goods | Observational Data | A within variation model is obtained by regressing sales growth rate on three measures of ratings – mean of ratings, the standard deviation of ratings, and count of ratings. |
| Dellarocas et al. (2007) | Movies | Experience Goods | Observational Data | Used fixed effects to control for unobserved movie quality. |
| Liu (2006) | Films | Experience Goods | Observational Data | Box office revenue for a week is regressed on WOM activity of that week and all previous weeks, therefore predicting revenue based on relative changes in WOM activity. |
| Chevalier and Mayzlin (2006) | Books | Experience Goods | Observational Data | Examine reviews and sales on Amazon and Barnes&Noble on different dates and used a difference-in-differences approach to account for book-specific and website-specific effects. |
| Li and Hitt (2008a) | Books | Experience Goods | Observational Data | A within variation model by regressing sales (obtained by sales rank) on different metrics of online WOM. |
| Duan et al. (2008) | Films | Experience Goods | Observational Data | A within variation model is obtained by regressing daily revenue of a movie on lagged measures of online WOM. They find that reviews have no impact on sales. |

| | | | | |
|---|---|---|---|---|
| Zhu and Zhang (2010) | Games | Experience Goods | Observational Data | Difference-in-differences approach across two video game console platforms to eliminate unobserved game characteristics. |
| Gu et al. (2012) | Cameras | Search Goods | Observational Data | A within variation model is obtained by regressing the daily sales rank of a camera on lagged measures of online WOM. |
| Chen et al. (2011) | Digital Cameras | Search Goods | Observational Data | A natural experiment setting to examine the interaction effect of observational learning and WOM. |
| Li and Wu (2014) | Groupon deals | Experience Goods | Observational Data | Panel data of over 500 deals on Groupon is used to show evidence that Facebook mediated WOM positively affects sales of vouchers after controlling for deal fixed-effects, time-fixed effects, and linear and non-linear time trends. |

**Table A2. The Impact of Recommendation Visibility on Time Duration Per Profile**

| | (1) | (2) | (3) |
|---|---|---|---|
| *Sample* | *Sessions of treated markets* | *Sessions of treated markets (Smaller 4-day window)* | *Excluding users having a session in both Pre- and Post- period* |
| *Dependent Variable* | *Log_duration_per_profile* | *Log_duration_per_profile* | *Log_duration_per_profile* |
| *Model* | *OLS* | *OLS* | *OLS* |
| *PostXSpa* | 0.003 | 0.013 | 0.003 |
| | (0.081) | (0.078) | (0.079) |
| *PostXMod* | 0.009 | 0.073 | 0.021 |
| | (0.090) | (0.080) | (0.090) |
| *PostXDen* | 0.261*** | 0.241*** | 0.250** |
| | (0.079) | (0.076) | (0.099) |
| *Day of the week dummies* | Yes | Yes | Yes |
| *Market dummies* | Yes | Yes | Yes |
| *Constant* | 5.021*** | 4.944*** | 5.054*** |
| | (0.055) | (0.045) | (0.058) |
| *Observations* | 31,529 | 16,765 | 26,450 |
| *R-squared* | 0.041 | 0.040 | 0.035 |
| *Clusters* | 88 | 86 | 86 |

Clustered robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

**Table A3. Seemingly Unrelated Regressions (SUR) on the Three Measures of Consumer Decision Making**

| | (1) | (2) | (3) |
|---|---|---|---|
| *Sample* | *Sessions of treated markets* | *Sessions of treated markets* | *Sessions of treated markets* |
| *Dependent Variable* | *Unique* | *Duration* | *Max_distance* |
| *Model* | *SUR (OLS)* | *SUR (OLS)* | *SUR (OLS)* |
| *PostXSpa* | 0.503*** | 204.8*** | 1.307*** |
| | (0.102) | (41.27) | (0.243) |
| *PostXMod* | 0.093 | 18.50 | 0.468* |
| | (0.112) | (45.18) | (0.266) |
| *PostXDen* | -0.381*** | -92.84** | -0.805*** |
| | (0.105) | (42.35) | (0.250) |
| *Day of the week dummies* | Yes | Yes | Yes |
| *Market dummies* | Yes | Yes | Yes |
| *Constant* | 1.882** | 954.9*** | 7.568*** |
| | (0.756) | (306.3) | (1.805) |
| *Observations* | 8,434 | 8,434 | 8,434 |
| *R-squared* | 0.060 | 0.034 | 0.088 |

Clustered robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

**Table A4. Summary Statistics on Market Size Grouped by Market Type**

| Variable | Subgroup | Obs. | Mean | S.D. | Min | Max |
|---|---|---|---|---|---|---|
| *Doctor_count* | *Spa* | 11,780 | 525.0 | 480.3 | 17 | 1738 |
| *Doctor_count* | *Mod* | 9,339 | 865.2 | 819.2 | 20 | 3000 |
| *Doctor_count* | *Den* | 10,309 | 934.1 | 1077.1 | 264 | 3819 |
| *Practice_doctor_count* | *Spa* | 11,780 | 553.1 | 502.1 | 17 | 1807 |
| *Practice_doctor_count* | *Mod* | 9,339 | 924.2 | 866.2 | 21 | 3184 |
| *Practice_doctor_count* | *Den* | 10,309 | 1032.4 | 1171.5 | 278 | 4164 |

**Table A5. Robustness Check: Regressions After Including Number of Options (market size) as a Control**

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| *Sample* | *Sessions of treated markets* | *Sessions of treated markets* | *Sessions of treated markets* | *Sessions of treated markets* | *Sessions of treated markets* | *Sessions of treated markets* |
| *Dependent Variable* | *Unique (Consideration Set Size)* | *Unique (Consideration Set Size)* | *Duration (Session Duration)* | *Duration (Session Duration)* | *Max_distance (Geographic Dispersion)* | *Max_distance (Geographic Dispersion)* |
| *Model* | *Negative Binomial* | *Negative Binomial* | *OLS* | *OLS* | *OLS* | *OLS* |
| *Practice_doctor_count* | | 0.000*** | | 0.046* | | 0.000 |
| | | (0.000) | | (0.024) | | (0.000) |
| *Doctor_count* | 0.000*** | | 0.049* | | 0.000 | |
| | (0.000) | | (0.027) | | (0.001) | |
| *PostXSpa* | 0.111*** | 0.112*** | 93.32*** | 93.58*** | 1.248*** | 1.250*** |
| | (0.041) | (0.041) | (24.88) | (24.87) | (0.374) | (0.374) |
| *PostXMod* | 0.016 | 0.017 | 45.00* | 45.30* | 0.292 | 0.294 |
| | (0.037) | (0.038) | (25.91) | (25.95) | (0.347) | (0.346) |
| *PostXDen* | -0.138*** | -0.140*** | -58.73*** | -59.35*** | -0.627** | -0.631** |
| | (0.040) | (0.040) | (21.04) | (21.08) | (0.299) | (0.297) |
| *Day of the week dummies* | Yes | Yes | Yes | Yes | Yes | Yes |
| *Market dummies* | No | No | No | No | No | No |
| *City dummies* | Yes | Yes | Yes | Yes | Yes | Yes |
| *Specialty dummies* | Yes | Yes | Yes | Yes | Yes | Yes |
| *Constant* | 0.144*** | 0.143*** | 555.36*** | 555.05*** | 7.604*** | 7.600*** |
| | (0.034) | (0.034) | (21.82) | (21.80) | (0.395) | (0.395) |
| *Observations* | 31,529 | 31,529 | 31,529 | 31,529 | 8,434 | 8,434 |
| *R-squared* | N/A | N/A | 0.016 | 0.016 | 0.072 | 0.072 |
| *Log-likelihood* | -52,167 | -52,166 | N/A | N/A+ | N/A | N/A |
| *Clusters* | 88 | 88 | 88 | 88 | 77 | 77 |

Clustered robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

**Table A6. Robustness Check: Including Individual Day Dummies**

| | (1) | (2) | (3) |
|---|---|---|---|
| *Sample* | *Sessions of all markets* | *Sessions of all markets* | *Sessions of all markets* |
| *Dependent Variable* | *Unique (Consideration Set Size)* | *Duration (Session Duration)* | *Max_distance (Geographic Dispersion)* |
| *Model* | *Negative Binomial* | *OLS* | *OLS* |
| *PostXSpa* | 0.154*** | 129.1*** | 1.460*** |
| | (0.050) | (35.9) | (0.524) |
| *PostXMod* | 0.026 | 23.14 | 0.631 |
| | (0.052) | (37.90) | (0.496) |
| *PostXDen* | -0.162*** | -82.65** | -0.662 |
| | (0.054) | (33.07) | (0.498) |
| *Date dummies* | Yes | Yes | Yes |
| *Market dummies* | Yes | Yes | Yes |
| *Constant* | 0.195*** | 618.4*** | 8.526*** |
| | (0.039) | (29.17) | (0.336) |
| *Observations* | 42,229 | 42,229 | 10,479 |
| *R-squared* | N/A | 0.021 | 0.096 |
| *Log-likelihood* | -68,944 | N/A | N/A |
| *Clusters* | 137 | 137 | 135 |

Clustered robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Note: Please note that in these models, we do not include the *Post* dummy, since the post period effect is captured in the individual date dummies, and *Post* would be collinear with these dummies. The coefficients of the interaction terms (PostXSpa, PostXMod and PostXDen) are not collinear with the individual day dummies.

**Table A7. Robustness Check: Consideration Set Size of More than One**

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| *Sample* | *Sessions of treated markets* | *Sessions of all markets* | *Sessions of treated markets* | *Sessions of all markets* | *Sessions of treated markets* | *Sessions of all markets* |
| *Dependent Variable* | *Unique (Consideration Set Size)* | *Unique (Consideration Set Size)* | *Duration (Session Duration)* | *Duration (Session Duration)* | *Max_distance (Geographic Dispersion)* | *Max_distance (Geographic Dispersion)* |
| *Post* | | 0.025 | | 42.61 | | -0.202 |
| | | (0.031) | | (40.37) | | (0.376) |
| *PostXSpa* | 0.122*** | 0.092** | 191.0*** | 155.1*** | 1.307*** | 1.466*** |
| | (0.034) | (0.042) | (42.85) | (53.35) | (0.415) | (0.524) |
| *PostXMod* | 0.001 | -0.025 | 36.17 | 1.82 | 0.468 | 0.627 |
| | (0.034) | (0.042) | (51.83) | (63.38) | (0.348) | (0.494) |
| *PostXDen* | -0.127*** | -0.153*** | -109.3** | -144.06** | -0.805** | -0.647 |
| | (0.034) | (0.045) | (43.79) | (60.47) | (0.351) | (0.498) |
| *Day of the week dummies* | Yes | Yes | Yes | Yes | Yes | Yes |
| *Market dummies* | Yes | Yes | Yes | Yes | Yes | Yes |
| *Constant* | 0.762*** | 0.770*** | 987.34*** | 999.27*** | 7.568*** | 7.565*** |
| | (0.038) | (0.034) | (45.04) | (41.80) | (0.412) | (0.396) |
| *Observations* | 12,178 | 15,809 | 12,178 | 15,809 | 8,434 | 10,479 |
| *R-squared* | N/A | N/A | 0.051 | 0.055 | 0.087 | 0.096 |
| *Log-likelihood* | -23,966 | -30,985 | N/A | N/A | N/A | N/A |
| *Clusters* | 88 | 137 | 88 | 137 | 77 | 135 |

Clustered robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Note: The results for geographic dispersion are exactly the same as the main results, as geographic dispersion can only be calculated when 2 or more doctors have been browsed.

## Table A8. Formal Verification of Parallel Trends Assumption for D-I-D

| | (1) | (2) | (3) |
|---|---|---|---|
| *Dependent Variable* | *Unique* | *Duration* | *Max_distance* |
| *Model* | *Negative Binomial* | *OLS* | *OLS* |
| *Post* | -0.013 | 3.195 | -0.199 |
| | (0.021) | (20.065) | (0.376) |
| *Pre1XSpa* | 0.032 | 23.876 | 1.195 |
| | (0.078) | (62.318) | (0.824) |
| *Pre1XMod* | 0.032 | -58.074 | -0.881 |
| | (0.104) | (82.437) | (0.852) |
| *Pre1XDen* | -0.085 | -19.853 | -0.515 |
| | (0.075) | (72.152) | (0.467) |
| *Pre2XSpa* | 0.005 | 75.499 | -0.460 |
| | (0.075) | (62.599) | (0.727) |
| *Pre2XMod* | 0.105 | -99.026 | -0.500 |
| | (0.093) | (77.154) | (0.777) |
| *Pre2XDen* | -0.116 | -139.474 | -0.703 |
| | (0.087) | (90.966) | (0.536) |
| *Pre3XSpa* | 0.004 | 76.486 | 0.604 |
| | (0.083) | (67.595) | (0.994) |
| *Pre3XMod* | 0.073 | -80.894 | -0.298 |
| | (0.109) | (82.565) | (0.832) |
| *Pre3XDen* | -0.088 | -64.611 | -1.275 |
| | (0.091) | (73.433) | (0.775) |
| *Pre4XSpa* | 0.017 | 24.559 | 0.765 |
| | (0.077) | (62.108) | (0.685) |
| *Pre4XMod* | -0.022 | -110.759 | -0.766 |
| | (0.092) | (74.024) | (0.785) |
| *Pre4XDen* | -0.077 | -104.747 | -1.158** |
| | (0.081) | (70.954) | (0.568) |
| *Pre5XSpa* | 0.043 | 37.459 | 0.637 |
| | (0.068) | (53.251) | (0.597) |
| *Pre5XMod* | 0.009 | -5.838 | 0.071 |
| | (0.098) | (83.551) | (0.877) |
| *Pre5XDen* | -0.123 | -78.854 | -0.511 |
| | (0.080) | (94.506) | (0.719) |
| *postXSpa* | 0.159** | 147.628** | 1.567*** |
| | (0.078) | (58.362) | (0.597) |
| *postXMod* | 0.041 | -29.090 | 0.233 |
| | (0.098) | (78.391) | (0.729) |
| *postXDen* | -0.229*** | -142.075** | -1.226* |
| | (0.085) | (68.808) | (0.642) |
| *Constant* | 0.116 | 515.199*** | 7.315*** |
| | (0.076) | (58.736) | (0.614) |
| *Observations* | 42,229 | 42,229 | 10,479 |
| *R-squared* | N/A | 0.022 | 0.097 |
| *Log-Likelihood* | -68,933 | N/A | N/A |
| *Clusters* | 137 | 137 | 135 |

Clustered robust standard errors in parentheses. Day of the Week and Market Dummies are included in all models.
*** p<0.01, ** p<0.05, * p<0.1.

## Table A9. Falsification Tests – The Impact of Online Recommendation Visibility on Offline Appointments

Negative binomial regression.
Dependent variable: Number of walk-in appointments booked for a practice doctor in a day

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| *Sample* | *Treatment Group* | *Treatment Group* | *Competition Group* | *Treatment and Comparison Group* | *Treatment and Comparison Group* |
| *Comparison* | *Before vs. after for rated doctors* | *Before vs. after for rated doctors* | *Before vs. after for competing unrated doctors* | *Difference in differences* | *Difference in differences* |
| *Post* | 0.376 | | 0.495 | 0.168 | 0.168 |
| | (0.397) | | (0.372) | (0.425) | (0.425) |
| *PostXTreat* | | | | 0.109 | |
| | | | | (0.56) | |
| *PostXTop* | | -1.055 | | | -1.322 |
| | | (0.844) | | | (0.945) |
| *PostXMiddle* | | 1.143 | | | 0.875 |
| | | (0.965) | | | (1.081) |
| *PostXBottom* | | 0.514 | | | 0.247 |
| | | (0.494) | | | (0.625) |
| *Day of the week dummies* | Yes | Yes | Yes | Yes | Yes |
| *Practice doctor dummies* | Yes | Yes | Yes | Yes | Yes |
| *Observations* | 4,680 | 4,680 | 6,375 | 8,175 | 8,175 |
| *Clusters* | 312 | 312 | 425 | 545 | 545 |

Clustered robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

**Table A10. Falsification Test – Use Artificial Pre and Post Periods**

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| *Sample* | *Sessions of all markets* | *Sessions of all markets* | *Sessions of all markets* | *Online appointments of treatment and comparison group doctors* |
| *Dependent Variable* | *Unique (Consideration set size)* | *Duration* | *Max_distance* | *Online Appointments* |
| *Model* | *Negative Binomial* | *OLS* | *OLS* | *Negative Binomial* |
| *Post* | 0.055 | -69.22** | -0.912** | 0.302** |
| | (0.034) | (30.28) | (0.446) | (0.152) |
| *PostXSpa* | -0.073 | -12.56 | 0.834 | |
| | (0.047) | (42.86) | (0.583) | |
| *PostXMod* | -0.073 | -15.90 | 0.306 | |
| | (0.053) | (51.48) | (0.671) | |
| *PostXDen* | -0.065 | 47.84 | 0.331 | |
| | (0.041) | (69.89) | (0.547) | |
| *PostXTop* | | | | -0.297 |
| | | | | (0.182) |
| *PostXMiddle* | | | | -0.220 |
| | | | | (0.185) |
| *PostXBottom* | | | | -0.0211 |
| | | | | (0.217) |
| *Day of the week dummies* | Yes | Yes | Yes | Yes |
| *Practice doctor dummies* | No | No | No | Yes |
| *Market dummies* | Yes | Yes | Yes | No |
| *Constant* | 0.328*** | 669.4*** | 9.812*** | 1.052*** |
| | (0.0277) | (28.36) | (0.372) | (0.186) |
| *Observations* | 12,489 | 12,489 | 3,094 | 2,180 |
| *R-squared* | N/A | 0.02 | 0.16 | N/A |

Clustered robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Note: In this falsification test, the sessions and appointments of June 19th and June 20th, 2014 are considered to be in Pre-period, and the sessions from June 23rd and 24th 2014 are considered to be in Post-period. We exclude June 21st - 22nd from the regression as they are weekend days. In the main results, as we have two weekends, one in pre-period and the other in post-period, the day of the week dummies adequately captures variation in browsing behavior on the weekends.

## Table A11. Main results on reduced sample size

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| *Sample* | *Sessions of all markets* | *Sessions of all markets* | *Sessions of all markets* | *Online appointments of treatment and comparison group doctors* |
| *Dependent Variable* | *Unique (Consideration set size)* | *Log_Duration* | *Max_distance* | *Online Appointments* |
| *Model* | *Negative Binomial* | *OLS* | *OLS* | *Negative Binomial* |
| *Post* | -0.027 | 0.078 | 0.138 | -0.169 |
| | (0.025) | (0.056) | (0.476) | (0.145) |
| *PostXSpa* | 0.263*** | 0.203** | 1.315** | |
| | (0.065) | (0.082) | (0.656) | |
| *PostXMod* | 0.079 | 0.119 | 0.315 | |
| | (0.079) | (0.090) | (0.740) | |
| *PostXDen* | -0.197** | -0.203** | -0.381 | |
| | (0.086) | (0.098) | (0.721) | |
| *PostXTop* | | | | 0.431** |
| | | | | (0.78) |
| *PostXMiddle* | | | | 0.393* |
| | | | | (0.203) |
| *PostXBottom* | | | | 0.280 |
| | | | | (0.184) |
| *Day of the week dummies* | Yes | Yes | Yes | Yes |
| *Practice doctor dummies* | No | No | No | Yes |
| *Market dummies* | Yes | Yes | Yes | No |
| *Constant* | 0.096* | 5.516*** | 5.597*** | -0.488 |
| | (0.057) | (0.066) | (0.674) | (0.128) |
| *Log-likelihood* | -20,416 | N/A | N/A | -1,341 |
| *Clusters* | 137 | 137 | 128 | 145 |
| *Observations* | 12,480 | 12,480 | 3,099 | 2,175 |
| *R-squared* | N/A | 0.047 | 0.124 | N/A |

Clustered robust standard errors in parentheses
*** $p<0.01$, ** $p<0.05$, * $p<0.1$

Note: The above table presents the main results on a smaller sample. The sample has been chosen randomly, and the sample size has been restricted to the size of the sample for the falsification tests. These results indicate that the main results are valid on a sample size equal to the size of the sample used for carrying out the falsification test.

**Table A12.** **The Impact of Recommendation Visibility on Consumer Consideration Set (Different thresholds for WOM Abundance)**

Negative binomial regression. Dependent variable: Number of unique doctors viewed in a search session

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| *Sample* | *Sessions of treated markets* | *Sessions of all markets* | *Sessions of treated markets (Smaller 4-day window)* | *Sessions of all markets (Smaller 4-day window)* | *Excluding users having a session in both Pre- and Post-period* | *Excluding users having a session in both Pre- and Post-period* |
| *Post* | | -0.014 | | -0.049 | | -0.025 |
| | | (0.020) | | (0.032) | | (0.023) |
| *PostXSpa* | 0.129*** | 0.140*** | 0.122*** | 0.171*** | 0.130*** | 0.149*** |
| | (0.043) | (0.044) | (0.046) | (0.055) | (0.043) | (0.046) |
| *PostXMod* | -0.012 | 0.001 | -0.025 | 0.025 | 0.005 | 0.026 |
| | (0.072) | (0.074) | (0.052) | (0.060) | (0.078) | (0.080) |
| *PostXDen* | -0.163*** | -0.152*** | -0.132*** | -0.083 | -0.162*** | -0.142** |
| | (0.047) | (0.051) | (0.046) | (0.055) | (0.056) | (0.059) |
| *Constant* | 0.137*** | 0.139*** | 0.127*** | 0.129*** | 0.155*** | 0.158*** |
| | (0.034) | (0.033) | (0.031) | (0.030) | (0.035) | (0.034) |
| *Observations* | 31,529 | 42,229 | 16,765 | 22,312 | 26,450 | 35,587 |
| *Log Pseudo Likelihood* | -52,064 | -68,951 | -27,463 | -36,132 | -43,422 | -57,934 |
| *Clusters* | 88 | 137 | 86 | 137 | 86 | 137 |

Clustered robust standard errors in parentheses. Day of the week dummies and Market dummies are included in all the above models. *** p<0.01, ** p<0.05, * p<0.1

**Table A13. The Impact of Recommendation Visibility on Consideration Set Composition (Different thresholds for WOM Abundance)**

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| *Sample* | *Sessions of treated markets* | *Sessions of treated markets (Smaller 4-day window)* | *Excluding users having a session in both Pre- and Post- period* | *Sessions of treated markets* | *Sessions of treated markets (Smaller 4-day window)* | *Excluding users having a session in both Pre- and Post- period* |
| *Dep. Variable* | *Treated_Viewed* | *Treated_Viewed* | *Treated_Viewed* | *Last_Viewed* | *Last_Viewed* | *Last_Viewed* |
| *Model* | *Logit* | *Logit* | *Logit* | *Logit* | *Logit* | *Logit* |
| *PostXSpa* | 0.284*** | 0.243*** | 0.276*** | 0.168*** | 0.158** | 0.124* |
| | (0.085) | (0.088) | (0.082) | (0.060) | (0.071) | (0.068) |
| *PostXMod* | 0.406*** | 0.394*** | 0.408*** | 0.409*** | 0.513*** | 0.412*** |
| | (0.063) | (0.049) | (0.064) | (0.077) | (0.094) | (0.071) |
| *PostXDen* | 0.182*** | 0.186** | 0.159** | 0.295*** | 0.278*** | 0.271*** |
| | (0.068) | (0.072) | (0.077) | (0.065) | (0.068) | (0.066) |
| *Constant* | -2.957*** | -2.955*** | -3.303*** | -2.879*** | -2.917*** | -3.214*** |
| | (0.073) | (0.074) | (0.076) | (0.070) | (0.075) | (0.075) |
| *Observations* | 31,393 | 16,696 | 26,356 | 31,010 | 16,129 | 25,728 |
| *Pseudo R-squared* | 0.115 | 0.120 | 0.117 | 0.111 | 0.109 | 0.106 |
| *Clusters* | 81 | 79 | 80 | 75 | 73 | 73 |

Clustered robust standard errors in parentheses. Day of the week dummies and Market dummies are included in all the above models.*** p<0.01, ** p<0.05, * p<0.1

**Table A14. The Impact of Recommendation Visibility on User Session Duration (Different thresholds for WOM Abundance)**

Ordinary least squares regression. Dependent variable: Duration of a session measured in seconds

| | (2) | (3) | (4) | (5) | (6) | (7) |
|---|---|---|---|---|---|---|
| *Sample* | *Sessions of treated markets* | *Sessions of all markets* | *Sessions of treated markets (Smaller 4-day window)* | *Sessions of all markets (Smaller 4-day window)* | *Excluding users having a session in both Pre- and Post-period* | *Excluding users having a session in both Pre- and Post-period* |
| *Post* | | 4.73 | | 27.41 | | 13.38 |
| | | (19.61) | | (22.81) | | (19.36) |
| *PostXSpa* | 115.59*** | 114.50*** | 80.20*** | 53.73 | 96.79*** | 83.85*** |
| | (28.19) | (33.21) | (26.36) | (34.55) | (26.36) | (31.43) |
| *PostXMod* | 19.11 | 17.13 | 8.37 | -19.16 | 13.31 | 0.73 |
| | (39.51) | (43.78) | (28.87) | (36.28) | (33.57) | (38.21) |
| *PostXDen* | -81.56*** | -82.53** | -43.86* | -71.16** | -88.14*** | -100.32*** |
| | (25.56) | (31.62) | (24.10) | (32.65) | (25.78) | (31.04) |
| *Constant* | 541.42*** | 551.04*** | 539.17*** | 548.24*** | 528.99*** | 542.26*** |
| | (22.71) | (22.04) | (18.57) | (17.46) | (22.20) | (21.88) |
| *Observations* | 31,529 | 42,229 | 16,765 | 22,312 | 26,450 | 35,587 |
| *Clusters* | 88 | 137 | 86 | 137 | 86 | 137 |
| *R-squared* | 0.020 | 0.021 | 0.020 | 0.022 | 0.019 | 0.019 |

Clustered robust standard errors in parentheses. Day of the week dummies and Market dummies are included in all the above models.*** p<0.01, ** p<0.05, * p<0.1

**Table A15. The Impact of Recommendation Visibility on Geographic Dispersion (Different thresholds for WOM Abundance)**

Ordinary least squares regression

Dependent variable: Haversine distance between the two farthest doctors browsed in a session measured in km

| | (2) | (3) | (4) | (5) | (6) | (7) |
|---|---|---|---|---|---|---|
| *Sample* | *Sessions of treated markets* | *Sessions of all markets* | *Sessions of treated markets (Smaller 4-day window)* | *Sessions of all markets (Smaller 4-day window)* | *Excluding users having a session in both Pre- and Post-period* | *Excluding users having a session in both Pre- and Post-period* |
| *Post* | | -0.201 | | -0.345 | | -0.043 |
| | | (0.376) | | (0.477) | | (0.422) |
| *PostXSpa* | 1.074*** | 1.232** | 0.944** | 1.293** | 0.974** | 0.948* |
| | (0.404) | (0.518) | (0.410) | (0.628) | (0.427) | (0.569) |
| *PostXMod* | 0.636** | 0.787* | 0.695** | 1.025* | 0.663* | 0.639 |
| | (0.274) | (0.446) | (0.289) | (0.556) | (0.372) | (0.545) |
| *PostXDen* | -0.735** | -0.574 | -0.272 | 0.075 | -0.617 | -0.648 |
| | (0.330) | (0.488) | (0.195) | (0.510) | (0.442) | (0.597) |
| *Constant* | 7.728*** | 7.724*** | 7.688*** | 7.634*** | 7.911*** | 7.928*** |
| | (0.397) | (0.382) | (0.451) | (0.429) | (0.500) | (0.488) |
| *Observations* | 8,434 | 10,479 | 4,432 | 5,487 | 7,087 | 8,864 |
| *Clusters* | 77 | 135 | 76 | 131 | 77 | 135 |
| *R-squared* | 0.087 | 0.096 | 0.094 | 0.102 | 0.090 | 0.099 |

Clustered robust standard errors in parentheses. Day of the week dummies and Market dummies are included in all the above models.*** p<0.01, ** p<0.05, * p<0.1

**Table A16. Impact of Online WOM on Appointment Count of Doctors (WOM as a continuous variable)**

Negative binomial regression
Dependent variable: Number of appointments booked online for a practice-doctor in a day

| | (1) | (2) |
|---|---|---|
| *Sample* | *Treatment Group* | *Treatment and Control Group* |
| *Comparison* | *Before vs. after for rated doctors* | *Difference in Differences)* |
| *Post* | | 0.039 |
| | | (0.056) |
| *PostXWOM_Count* | 0.010*** | 0.009*** |
| | (0.003) | (0.089) |
| *Constant* | -0.002 | 0.028 |
| | (0.091) | (0.091) |
| *Pseudo-R2* | 0.22 | 0.24 |
| *Log Pseudo Likelihood* | -2,694 | -3,592 |
| *Observations* | 4,485 | 8,175 |
| *Clusters* | 299 | 545 |

Clustered robust standard errors in parentheses. Day of the week dummies and practice doctor dummies are included in all the above models. *** p<0.01, ** p<0.05, * p<0.1

**Table A17. Impact of Online WOM on Appointment Count of Doctors (Equal sub-group size)**

Negative binomial regression
Dependent variable: Number of appointments booked online for a practice-doctor in a day

| | (1) | (2) |
|---|---|---|
| *Sample* | *Treatment Group* | *Treatment and Comparison Group* |
| *Comparison* | *Before vs. after for rated doctors* | *Difference-in-differences* |
| *Post* | | -0.105 |
| | | (0.105) |
| *PostXTop* | 0.263*** | 0.384*** |
| | (0.0933) | (0.141) |
| *PostXMiddle* | 0.116 | 0.235 |
| | (0.109) | (0.147) |
| *PostXBottom* | -0.155 | -0.0351 |
| | (0.158) | (0.187) |
| *Day of the week dummies* | Yes | Yes |
| *Practice doctor dummies* | Yes | Yes |
| *Constant* | -0.329** | -0.302** |
| | (0.139) | (0.127) |
| *Log-likelihood* | -2,327 | -3,224 |
| *Observations* | 4,050 | 7,545 |
| *Clusters* | 270 | 503 |

Clustered robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Note: The above table presents the demand analysis on a reduced sample size. This sample selects observations from randomly selected equal number of practice doctors in the *Top*, *Middle* and *Bottom* sub-groups (90 each).

**Table B1. Fraudulent Review Detection Features Utilized in Prior Literature**

| **Detection Metric Basis** | **Feature Type** | **Feature** | **Study** | **Detail** | **Applicable to Doctor Reviews?** | **Resilience** | **Remarks** |
|---|---|---|---|---|---|---|---|
| Review | Countable feature | Number of feedbacks | Jindal and Liu (2008) | Checking for outliers in the count of feedbacks (votes) a review receives | No; the feedback is very sparsely distributed | Low, a fraudster can also manipulate the votes on the review | A well-written deceptive review will get a large number of positive feedbacks, defeating this rationale. |
| Review | Countable feature | Number of helpful feedbacks | Jindal and Liu (2008) | Checking for how many other users upvote/mark the review as helpful | No; the fraudulent doctor reviews are convincing enough to get many upvotes | Low—same reason as above | Same as above |
| Review | Countable feature | Percentage of helpful feedbacks | Jindal and Liu (2008) | Checking for the proportion of positive/total votes for a review | No; a well-written fake review sounds very convincing and hence receives a high proportion of | Low—same reason as above | Same as above |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | upvotes | | |
| Review | Countable feature | Length of review title | Jindal and Liu (2008) | Checking for outliers in review title | Possible | Low | A skilled fraudster does not write reviews that are outliers in title length and content. |
| Review | Countable feature | Length of review text | Jindal and Liu (2007a) | Checking for outliers in review length | Possible | Low | Same as above |
| Review | Positional feature | Position of review when sorted on date | Jindal and Liu (2008) , Jindal and Liu (2007a) | First few reviews/early reviews are likely to be fraudulent | No; most doctors just have a few reviews, and because most reviewers have just one review | Low, as most review filters account for this | Skilled fraudsters understand that review filters are more sensitive on the first few reviews and hence use trusted accounts (old accounts that have posted several reviews) to post the first few reviews for any product. |
| Review | Positional feature | Binary feature to indicate if this review is the first | Jindal and Liu (2008) | Assumes that the likelihood of the first review being | No; many doctors just have a handful | Low, as most review filters account for | Same as above |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | review | | fraudulent is higher | of reviews | this | |
| Review | Positional feature | Binary feature to indicate if this review is the only review | Jindal and Liu (2007b) | Assumes that if there is just one review, it's likely to be given by a doctor's close confidante | No; many doctors have just one review | Low, as skilled fraudsters are aware of it | Same as above |
| Review | Review metadata | Time duration taken to write the review | Liu (2012) | Tries to catch fraudsters who resort to copying and pasting review texts and hence take a very short duration of time to write the review | Possible | Low, as skilled fraudsters would easily learn this | This also suffers from very high false positive rate as many genuine users write reviews on a word processor and later copy and paste into the web editor. |
| Review | Review text | Percent of positive and negative words | Jindal and Liu (2008) | Points out outliers in terms of proportion of positive (appreciative) and negative (critical) words | Possible, but skilled fraudsters do not write flowery or overly critical reviews | Low, as skilled fraudsters understand this known metric | This also suffers from multiple measurement errors; two negatives can also mean a positive, such as “not bad.” |
| Review | Review text | Similarity between the text | Jindal and Liu (2008) | Checks for cosine similarity between the | Possible, but skilled | Low, as skilled | As most users can identify this type of |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | in the product description and product review | | review text and the product description | fraudsters do not write/copy-paste product descriptions because those reviews are not convincing to the customers | fraudsters understand that users easily understand these types of reviews to be fake | fake easily, skilled fraudsters avoid writing fake reviews that are similar to product descriptions. |
| Review | Review text | Number of times the brand name (doctor name) is used | Jindal and Liu (2008) | Assumes that when the product name/doctor name is used too many times, the review is likely to be an advertisement | Possible, but skilled fraudsters are aware of the customer perception of such reviews and tend to avoid such low-quality writing | Low; fraudsters can easily learn this over time | Similar to above |
| Review | Review text | Percent of characters, words in capital case | Jindal and Liu (2007b) | Assumes that when the use of CAPITALS is high, the review is likely to be fraudulent | Possible, but skilled reviewers avoid it | Low | Similar to above |
| Review | Review | Unigram, bigram, | Ott et al. | Assumes certain words | Possible | Medium | As critiqued in Xie et |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | text | trigram | (2011) | to be more likely used in fraudulent reviews; they draw the word banks from LIWC dictionaries | | | al. (2013), this method does not work as well on real-world data. |
| Review | Review text | Parts of speech distribution | Ott et al. (2011) | Parts of speech distribution is different across fraudulent and genuine reviews | Possible | High | As critiqued in Xie et al. (2013), this method does not work as well on real-world data. |
| Review | Review sentiment | Subjective vs. objective | Mohammadali et al. (2016) | Checks for facts stated vs. opinions given | Unlikely, as healthcare service is a credence good, which makes users give opinions | Moderate | Likely plagued by high false positives in the case of credence good reviews. |
| Review | Review sentiment | Positive vs. negative | Mohammadali et al. (2016) | Checks for any systematic deviation in sentiment from the mean sentiment | No; people give extreme opinions when they are relieved from pain | Low | Skilled fraudsters do not write overly critical or overly positive reviews, as users then discount the review’s |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | | credibility. |
| Review | Rating | The numeric rating given along with the review | Jindal and Liu (2008) | Considers 5-star and 1-star reviews to be more likely fraudulent than reviews with more moderate ratings | No | Moderate | This feature is not effective due to the bi-modal nature of distribution of customer ratings. |
| Review | Rating | The deviation of the numeric rating of the review from the mean rating | Jindal and Liu (2008) | Calculates the deviation of individual ratings from the mean of ratings and marks the ones that are the most off as suspicious | No | NA | The ratings have a bimodal distribution, and hence this metric has limited applicability. |
| Review | Rating | Binary feature indicating if a good review is written right after the first bad review was received | Jindal and Liu (2008) | This metric checks the time order of when the reviews are received and checks whether a negative review is immediately followed by a positive review | No | Low | The reliability of this metric becomes questionable when the review volumes are low. |
| Author | Countable feature | Ratio of number of first reviews to total reviews | Jindal and Liu (2008) | The metric checks whether said author is the first reviewer for an abnormally high | No; most authors only write a few reviews; as a | Low | The reliability of this metric is questionable in a low-volume, all first-timer market in |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | proportion of products | result, most products (doctors) have very few reviews | | which even fraudsters do not write multiple reviews from the same author ID. |
| Author | Countable feature | Ratio of number of products in which this reviewer was the only reviewer/ total products reviewed by this reviewer | Jindal and Liu (2008) | This metric calculates the ratio of products for which this reviewer is the only reviewer of the number of total products for which this reviewer has given a review | Moderately useful | Moderate | This metric can be extended to low-volume markets, but as fraudsters indulge in creating multiple accounts (sock puppets), the applicability is limited. |
| Author | Author metadata | Geolocation of reviewer's IP | Liu (2012) | This metric uses the geolocation of the reviewer IP to decipher whether the review is genuine or fraudulent; for example, reviews posted from California for a restaurant in Boston are deemed | Moderately useful | Moderate | Skilled fraudsters use proxies to ensure that their location as recorded by the server is close to an average customer's location. |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | fraudulent | | | |
| Author | Author metadata | IP address of the reviewer | Liu (2012) | This metric uses the IP address to determine whether a given review is likely to be fake; it checks for multiple account logins from the same IP address | Moderately useful | Low | Most skilled fraudsters are aware of this well-publicized metric used by review filters; they spoof their IP or use dynamic IPs instead of static ones to avoid detection. |
| Author | Author metadata | User ID of the reviewer | Liu (2012) | This metric uses the user ID to identify sock-puppet behavior; for example, if a product is reviewed by users with user IDs such as user9915, user9916, and user9917, they are considered fraudulent | Moderately useful | Low | Most skilled fraudsters will not create user IDs in a sequence and thus avoid this obvious detection. |
| Author | Author metadata | Whether reviewer gave non-extreme | Jindal and Liu (2007b) | This metric checks whether the reviewer | Limited use, as the volume of | Low | The applicability of this feature is very |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | ratings at all | | also gives non-extreme ratings; for example, among all the products rated, has reviewer ever given 2-star to 4-star ratings on a scale of 5? | reviews per reviewer is too low | | limited as online ratings have a bi-modal distribution, with most of the ratings concentrated on the extremes. |
| Author | Author metadata | Reviewer's trustworthiness based on browsing, reviewing history | Wang et al. (2011) | This metric assigns a trustworthiness score to reviewers based on their activity levels on the platform | Limited use, as reviewers have very limited browsing history on need-based platforms such as doctor search platforms | High | Leading platforms, such as Google and Yelp, use this as one of the primary metrics in their review filters, as they have extensive user browsing data across multiple platforms. |
| Author | Author metadata | Reviewer helpfulness score | Liu (2012) | This metric calculates the total upvotes given to all the reviews posted by the reviewer | Limited, as review volume per reviewer is very low | Low | This metric can be abused, as fraudulent reviewer groups can collude. |

## REFERENCES


AMA, WOMMA (2014) Return on word of mouth study. Accessed in Dec 2014. Retreived from https://womma.org/free-womm-resources/.

Anderson JG (2000) Security of the distributed electronic patient record: A case-based approach to identifying policy issues. *International Journal of Medical Informatics.* 60(2):111-118.

Angrist J, Krueger AB (1994) Why do world war ii veterans earn more than nonveterans? *Journal of Labor Economics.* 12(1):74-97.

Angrist JD (1990) Lifetime earnings and the vietnam era draft lottery: Evidence from social security administrative records. *The American Economic Review.* 80(3):313-336.

Ariely D, Bracha A, Meier S (2009) Doing good or doing well? Image motivation and monetary incentives in behaving prosocially. *American Economic Review.* 99(1):544-555.

Attorney General Report (2013). Accessed in July 2017. Retreived from http://www.ag.ny.gov/press-release/ag-schneiderman-announces-agreement-19-companies-stop-writing-fake-online-reviews-and.

Austin JM, Jha AK, Romano PS, Singer SJ, Vogus TJ, Wachter RM, Pronovost PJ (2015) National hospital ratings systems share few common scores and may generate confusion instead of clarity. *Health Affairs.* 34(3):423-430.

Bakos JY (1997) Reducing buyer search costs: Implications for electronic marketplaces. *Management Science.* 43(12):1676-1692.

Bapna R, Gupta A, Jung J, Sen S (2014) Analyzing the impact of incentive structure on the diffusion of mobile social games: A randomized field experiment. *Workshop on Information Systems and Economics (WISE).*

Becker DM, Niessen-Ruenzi A, Ruenzi S (2014) To pay or not to pay-evidence from whole blood donations in germany.

Bell DR, Ho T-H, Tang CS (1998) Determining where to shop: Fixed and variable costs of shopping. *Journal of Marketing Research.* 35(3):352-369.

Berger J, Iyengar R (2013) Communication channels and word of mouth: How the medium shapes the message. *Journal of Consumer Research.* 40(3):567-579.

Berger J, Sorensen AT, Rasmussen SJ (2010) Positive effects of negative publicity: When negative reviews increase sales. *Marketing Science.* 29(5):815-827.

Bérnabou R, Tirole J (2006) Incentives and prosocial behavior. *The American Economic Review.* 96(5):1652-1678.

Bettman J, Johnson E, Payne J (1991) Consumer decision making. *Handbook of Consumer Behaviour.* 50-84.

Boulding W, Kirmani A (1993) A consumer-side experimental examination of signaling theory: Do consumers perceive warranties as signals of quality? *Journal of consumer research.* 20(1):111-123.

Brynjolfsson E, Hu Y, Simester D (2011) Goodbye pareto principle, hello long tail: The effect of search costs on the concentration of product sales. *Management Science.* 57(8):1373-1386.

Bucklin RE, Sismeiro C (2003) A model of web site browsing behavior estimated on clickstream data. *Journal of Marketing Research.* 40(3):249-267.

Caplin A, Dean M, Martin D (2011) Search and satisficing. *The American Economic Review.* 101(7):2899-2922.

Carpenter J, Myers CK (2010) Why volunteer? Evidence on the role of altruism, image, and incentives. *Journal of Public Economics.* 94(11-12):911-920.

Charman-Anderson S (2012) Fake reviews: Amazon's rotten core. Accessed in August 2017. Retreived from https://www.forbes.com/sites/suwcharmananderson/2012/08/28/fake-reviews-amazons-rotten-core/#5984d42a7428.

Chen P-Y, Wu S-Y, Yoon J (2004) The impact of online recommendations and consumer feedback on sales. *International Conference on Information Systems 2004*, Washington DC.

Chen Y, Wang Q, Xie J (2011) Online social interactions: A natural experiment on word of mouth versus observational learning. *Journal of Marketing Research.* 48(2):238-254.

Chen Y, Xie J (2008) Online consumer review: Word-of-mouth as a new element of marketing communication mix. *Management Science.* 54(3):477-491.

Cheshire C, Antin J, Cook KS, Churchill E (2010) General and familiar trust in websites. *Knowledge, Technology, & Policy.* 23(3-4):311-331.

Chevalier JA, Mayzlin D (2006) The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research.* 43(3):345-354.

Clemons EK, Gao GG, Hitt LM (2006) When online reviews meet hyperdifferentiation: A study of the craft beer industry. *Journal of Management Information Systems.* 23(2):149-171.

Corritore CL, Marble RP, Wiedenbeck S, Kracher B, Chandran A (2005) Measuring online trust of websites: Credibility, perceived ease of use, and risk. *Americas Conference on Information Systems*.

De los Santos B, Hortaçsu A, Wildenbeest MR (2012) Testing models of consumer search using data on web browsing and purchasing behavior. *The American Economic Review*. 102(6):2955-2980.

Deci EL (1975) Intrinsic motivation. *Plenum Press*.

Deci EL, Ryan RM (1985) Intrinsic motivation and self motivation in human behaviour. *New York Plenum Press*.

Dellarocas C, Zhang XQ, Awad NF (2007) Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *Journal of Interactive Marketing.* 21(4):23-45.

Dichter E (1966) How word-of-mouth advertising works. *Harvard Business Review*. 44(6):147-160.

Dorner V, Ivanova O, Scholz M (2013) Think twice before you buy! How recommendations affect three-stage purchase decision processes. *International Conference on Information Systems, ICIS 2013*, Milan.

Duan W, Gu B, Whinston AB (2008) The dynamics of online word-of-mouth and product sales—an empirical investigation of the movie industry. *Journal of Retailing.* 84(2):233-242.

Emmert M, Sander U, Pisch F (2013) Eight questions about physician-rating websites: A systematic review. *Journal of Medical Internet Research.* 15(2).

Engel J, Kollat D, Blackwell R (1968) Consumer behavior.

Fei G, Mukherjee, A., Liu, B., Hsu, M., Castellanos, M., & Ghosh, R. (2013) Exploiting burstiness in reviews for review spammer detection. *ICWSM*.

Feng S, Banerjee R, Choi Y (2012) Syntactic stylometry for deception detection. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, Jeju Island, Korea.

Forman C, Ghose A, Wiesenfeld B (2008) Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Information Systems Research.* 19(3):291-313.

Fradkin A (2015) Search frictions and the design of online marketplaces. Working Paper.

Fradkin A, Grewal E, Holtz D (2016) The determinants of online review informativeness: Evidence from field experiments on airbnb. Working Paper.

Frey BS, Jegen R (1999) Motivation crowding theory: A survey of empirical evidence. *Institute for Empirical Research in Economics, University of Zurich.*

Frey BS, Jegen R (2001) Motivation crowding theory. *Journal of Economic Surveys.* 15(5):589-611.

Frey BS, Oberholzer-Gee F (1997) The cost of price incentives: An empirical analysis of motivation crowding- out. *The American Economic Review.* 87(4):746-755.

Gao G, Greenwood BN, Agarwal R, McCullough JS (2015) Vocal minority and silent majority: How do online ratings reflect population perceptions of quality. *MIS Quarterly.* 39(3):565-590.

Gao GG, McCullough JS, Agarwal R, Jha AK (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.

Gauri DK, Bhatnagar A, Rao R (2008) Role of word of mouth in online store loyalty. *Communications of the ACM.* 51(3):89-91.

Ghose A, Ipeirotis PG, Li B (2013) Surviving social media overload: Predicting consumer footprints on product search engines. Working Paper, Carnegie Mellon University, Pittsburg.

Ghose A, Ipeirotis PG, Li B (2014) Examining the impact of ranking on consumer behavior and search engine revenue. *Management Science.* 60(7):1632-1654.

Gilbert E, Karahalios K (2010) Understanding deja reviewers. *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, Savannah, Georgia, USA.

Gneezy U, Rustichini A (2000) Pay enough or don't pay at all*. *The Quarterly Journal of Economics.* 115(3):791-810.

Godes D, Mayzlin D (2004) Using online conversations to study word-of-mouth communication. *Marketing Science.* 23(4):545-560.

Goldfarb A, Tucker C (2012) Shifts in privacy concerns. *The American Economic Review.* 102(3):349-353.

Goodman JK, Broniarczyk SM, Griffin JG, McAlister L (2013) Help or hinder? When recommendation signage expands consideration sets and heightens decision difficulty. *Journal of Consumer Psychology.* 23(2):165-174.

Graves A, Mohamed Ar, Hinton G (2013) Speech recognition with deep recurrent neural networks. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*.

Gu B, Park J, Konana P (2012) Research note—the impact of external word-of-mouth sources on retailer sales of high-involvement products. *Information Systems Research.* 23(1):182-196.

Guidry J, Aday L, Zhang D, Winn R (1997) Transportation as a barrier to cancer treatment. *Cancer practice.* 5(6):361-366.

Gupta A, Dhami A (2015) Measuring the impact of security, trust and privacy in information sharing: A study on social networking sites. *Journal of Direct, Data, and Digital Marketing Practice.* 17(1):43-53.

Gupta P, Harris J (2010) How e-wom recommendations influence product consideration and quality of choice: A motivation to process information perspective. *Journal of Business Research.* 63(9-10):1041-1049.

Hennig-Thurau T, Gwinner KP, Walsh G, Gremler DDJJoim (2004) Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the internet?18(1):38-52.

Hoffman DL, Novak TP, Peralta M (1999) Building consumer trust online. *Communications of the ACM.* 42(4).

Hong Y, Chen P-y, Hitt LM (2012) Measuring product type with dynamics of online product review variance. In international conference on information systems. *International Conference on Information Systems, ICIS 2012*.

Hong Y, Gu B, Burtch G, Huang N, Liang C, Wang K, Fu D, Yang B (2016) Effectiveness of performance feedback in stimulating user-generated content. *Thirty Seventh International Conference on Information System*, Dublin 2016.

Horowitz JL, Louviere JJ (1995) What is the role of consideration sets in choice modeling? *International Journal of Research in Marketing.* 12(1):39-54.

Hu N, Pavlou PA, Zhang JJ (2017) On self-selection biases in online product reviews. *MIS Quarterly*. 41(2):449-471.

Iajya V, Lacetera N, Macis M, Slonim R (2013) The effects of information, social and financial incentives on voluntary undirected blood donations: Evidence from a field experiment in argentina. *Social Science & Medicine.* 98214-223.

Jabr W (2015) The fallacy of spam reviews. *Workshop on Information Systems and Economics*, University of Texas at Dallas.

Jabr W, Zheng Z (2014) Know yourself and know your enemy: An analysis of firm recommendations and consumer reviews in a competitive environment. *MIS Quarterly*. 38(3):635-A610.

Jansen BJ, Spink A, Kathuria V (2007) How to define searching sessions on web search engines. *Springer Berlin Heidelberg*. Berlin, Heidelberg.

Jindal N, Liu B (2007a) Analyzing and detecting review spam. *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on Data Mining*.

Jindal N, Liu B (2007b) Review spam detection. *Proceedings of the 16th international conference on World Wide Web*, Banff, Alberta, Canada.

Jindal N, Liu B (2008) Opinion spam and analysis. *Proceedings of the 2008 International Conference on Web Search and Data Mining*, Palo Alto, California, USA.

Johnson EJ, Moe WW, Fader PS, Bellman S, Lohse GL (2004) On the depth and dynamics of online search behavior. *Management Science.* 50(3):299-308.

Johnson MK, Raye CL (1981) Reality monitoring. *Psychological Review.* 88(1):67-85.

Jung J, Bapna R, Golden J, Sun T (2016) Altruism pays! Towards optimal call-to-action for online referral: A randomized field experiment. *Conference on Information Systems and Technology (CIST 2016)*, Nashville, US.

Kang Y-S, Herr PM, Page CM (2003) Time and distance: Asymmetries in consumer trip knowledge and judgments. *Journal of Consumer Research.* 30(3):420-429.

Kim Y (2014) Trust in health information websites: A systematic literature review on the antecedents of trust. *Health Informatics Journal.*

Lacetera N, Macis M (2010) Do all material incentives for pro-social activities backfire? The response to cash and non-cash incentives for blood donations. *Journal of Economic Psychology.* 31(4):738-748.

Lacetera N, Macis M, Slonim R (2014) Rewarding volunteers: A field experiment. *Management Science.* 60(5):1107-1129.

Lagu T, Hannon NS, Rothberg MB, Lindenauer PK (2010) Patients' evaluations of health care providers in the era of social networking: An analysis of physician-rating websites. *J Gen Intern Med.* 25(9):942-946.

Le Q, Mikolov T (2014) Distributed representations of sentences and documents. *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*.

Lee S, Choeh JY (2016) The determinants of helpfulness of online reviews. *Behaviour & Information Technology.* 35(10):853-863.

Li F, Huang M, Yang Y, Zhu X (2011) Learning to identify review spam. *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*.

Li J, Ott, M., Cardie, C., & Hovy, E. H. (2014) Towards a general rule for identifying deceptive opinion spam. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*

Li X, Hitt LM (2008a) Self-selection and information role of online product reviews. *Information Systems Research.* 19(4):456-474.

Li X, Wu L (2014) Herding and social media word-of-mouth: Evidence from groupon. Working Paper.

Li XX, Hitt LM (2008b) Self-selection and information role of online product reviews. *Information Systems Research.* 19(4):456-474.

Lim E-P, Nguyen V-A, Jindal N, Liu B, Lauw HW (2010) Detecting product review spammers using rating behaviors. *Proceedings of the 19th ACM international conference on Information and knowledge management*, Toronto, ON, Canada.

Liu B (2012) Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies.* 5(1):1-167.

Liu Y (2006) Word of mouth for movies: Its dynamics and impact on box office revenue. *Journal of Marketing.* 70(3):74-89.

Liu Y, Chen P-y, Hong Y (2014) Value of multi-dimensional rating systems: An information transfer view. *International Conference on Information Systems, ICIS 2014*, Auckland.

Lleras JS, Masatlioglu Y, Nakajima D, Ozbay EY (2017) When more is less: Limited consideration. *Journal of Economic Theory.* 17070-85.

Lu SF, Rui H (2017) Can we trust online physician ratings? Evidence from cardiac surgeons in florida. *Management Science.* 2876-2885.

Luca M, Zervas G (2016) Fake it till you make it: Reputation, competition, and yelp review fraud. *Management Science.* 62(12):3412-3427.

Luo X (2009) Quantifying the long-term impact of negative word of mouth on cash flows and stock prices. *Marketing Science.* 28(1):148-165.

Matos MGd, Ferreira P, Smith MD, Telang R (2016) Culling the herd: Using real-world randomized experiments to measure social bias with known costly goods. *Management Science.* 62(9):2563-2580.

Mayzlin D, Dover Y, Chevalier J (2014) Promotional reviews: An empirical investigation of online review manipulation. *The American Economic Review.* 104(8):2421-2455.

McCarthy N (2014) Americans visit their doctor 4 times a year. People in japan visit 13 times a year.

Mellström C, Johannesson M (2008) Crowding out in blood donation: Was titmuss right? *Journal of the European Economic Association.* 6(4):845-863.

Mesch GS (2012) Is online trust and trust in social institutions associated with online disclosure of identifiable information online? *Computers in Human Behavior.* 28(4):1471-1477.

Metzger MJ (2017) Privacy, trust, and disclosure: Exploring barriers to electronic commerce. *Journal of Computer-Mediated Communication.* 9(4).

Meyer BD (1995) Natural and quasi-experiments in economics. *Journal of Business & Economic Statistics.* 13(2):151-161.

Mihalcea R, Strapparava C (2009) The lie detector: Explorations in the automatic recognition of deceptive language. *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, Suntec, Singapore.

Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*.

Mintel Research (2015). Accessed in July 2017. Retreived from http://www.mintel.com/blog/mintel-market-news/mintel-in-the-media-this-weeks-highlights-23-october-2015.

Moe WW (2003) Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream. *Journal of Consumer Psychology*. 13(1-2):29-39.

Moe WW, Fader PS (2004) Dynamic conversion behavior at e-commerce sites. *Management Science.* 50(3):326-335.

Moe WW, Trusov M (2011) The value of social dynamics in online product ratings forums. *Journal of Marketing Research.* 48(3):444-456.

Mohammadali T, Atefeh H, Zuriati I, Naomie S (2016) A framework for review spam detection research. *International Journal of Computer, Electrical, Automation, Control and Information Engineering.* 1066-71.

Montgomery AL, Li SB, Srinivasan K, Liechty JC (2004) Modeling online browsing and path analysis using clickstream data. *Marketing Science.* 23(4):579-595.

Mukherjee A, Liu B, Glance N (2012) Spotting fake reviewer groups in consumer reviews. *Proceedings of the 21st international conference on World Wide Web*, Lyon, France.

Mukherjee A, Liu B, Wang J, Glance N, Jindal N (2011) Detecting group review spam. *Proceedings of the 20th international conference companion on World wide web*, Hyderabad, India.

Mukherjee A, Venkataraman V, Liu B, Glance N (2013) What yelp fake review filter might be doing? *Association for the Advancement of Artificial Intelligence (AAAI).* 409-418.

Narayanan S, Kalyanam K (2015) Position effects in search advertising and their moderators: A regression discontinuity approach. *Marketing Science.* 34(3):388-407.

Nasraoui O, Saka E (2007) Web usage mining in noisy and ambiguous environments: Exploring the role of concept hierarchies, compression, and robust user profiles. *Springer Berlin Heidelberg*. Berlin, Heidelberg.

Ott M, Choi Y, Cardie C, Hancock JT (2011) Finding deceptive opinion spam by any stretch of the imagination. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, Portland, Oregon.

Park M (2013) Using patient medical records for medical research. *Korean J Fam Med.* 34(3).

Pew Research Center (2016) Online shopping and e-commerce.

Pham V, Bluche T, Kermorvant C, Louradour J (2014) Dropout improves recurrent neural networks for handwriting recognition. *2014 14th International Conference on Frontiers in Handwriting Recognition*.

Raine L (2018) Americans' complicated feelings about social media in an era of privacy concerns. *The Pew Research Center.*

Reinstein DA, Snyder CM (2005) The influence of expert reviews on consumer demand for experience goods: A case study of movie critics. *The Journal of Industrial Economics.* 53(1):27-51.

Robert MP, James MC, Alfred JM (1998) The geographic distribution of physicians in the united states and the contribution of international medical graduates. *Medical Care Research and Review.* 55(1):116-130.

Salloum R, Kuo CCJ (2017) Ecg-based biometrics using recurrent neural networks. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

Segal D (2011) A rave, a pan, or just a fake? Accessed in August 2017. Retreived from http://www.nytimes.com/2011/05/22/your-money/22haggler.html?mcubz=0.

Shavitt S, Fazio RH (1990) Effects of attribute salience on the consistency of product evaluations and purchase predictions. *ACR North American Advances.*

Simon HA (1956) Rational choice and the structure of the environment. *Psychological review, 63(2), 129.*

Smith D (2004) Amazon reviewers brought to book. Accessed in August 2017. Retreived from http://www.guardian.co.uk/technology/2004/feb/15/books.booksnews.

Streitfeld D (2012) For $2 a star, an online retailer gets 5-star product reviews. Accessed in August 2017. Retrieved from http://www.nytimes.com/2012/01/27/technology/for-2-a-star-a-retailer-gets-5-star-reviews.html?_r=2&ref=business.

Sun M (2012) How does the variance of product ratings matter? *Management Science*. 58(4):696-707.

Sun T, Viswanathan S, Huang N, Zheleva EJAaS (2018) Monetizing sharing traffic via incentive design: Evidence from a randomized field experiment.

Sundaram DS, Mitra K, Webster CJANAA (1998) Word-of-mouth communications: A motivational analysis.

Syed ST, Gerber BS, Sharp LK (2013) Traveling towards disease: Transportation barriers to health care access. *Journal of Community Health.* 38(5):976-993.

Tait SE, Jeske D (2015) Hello stranger! Trust and self-disclosure effects on online information sharing. *International Journal of Cyber Behavior, Psychology and Learning.* 5(1):42-55.

Tang D, Qin B, Liu T (2015) Document modeling with gated recurrent neural network for sentiment classification. *EMNLP*.

Tucker C, Zhang J (2011) How does popularity information affect choices? A field experiment. *Management Science.* 57(5):828-842.

Ursu RM (2015) The power of rankings: Quantifying the effect of rankings on online consumer search and purchase decisions. *Searle Conference on Internet Search and Innovation*.

Vrij A, Mann S, Kristen S, Fisher RPJL, behavior h (2007) Cues to deception and ability to detect lies as a function of police interview styles. *Law and Human Behavior.* 31(5):499-518.

Wang G, Xie S, Liu B, Yu PS (2011) Review graph based online store review spammer detection. *Proceedings of the 2011 IEEE 11th International Conference on Data Mining*, Vancouver, Canada.

Wang Y, Fesenmaier DRJTm (2004) Towards understanding members' general participation in and active contribution to an online travel community25(6):709-722.

Whiddett R, Hunter I, Engelbrecht J, Handy J (2006) Patients' attitudes towards sharing their health information. *International Journal of Medical Informatics.* 75530-541.

Wright P, Barbour F (1977) Phased decision strategies: Sequels to an initial screening. *Graduate School of Business, Stanford University*.

Xie S, Wang G, Lin S, Yu PS (2012) Review spam detection via temporal pattern discovery. *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, Beijing, China.

Xu Y, Armony M, Ghose A (2016) The effect of online reviews on physician demand: A structural model of patient choice. Working Paper.

Zhu F, Zhang X (2010) Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of Marketing.* 74(2):133-148.