Journal of the Association for Information Systems (2019) **20**(11), 1589-1610
doi: 10.17705/1jais.00579

RESEARCH ARTICLE

ISSN 1536-9323

Medical Crowdsourcing: Harnessing the "Wisdom of the Crowd" to Solve Medical Mysteries

Indika Dissanayake[1], Sridhar Nerur[2], Rahul Singh[3], Yang Lee[4]
[1]University of North Carolina at Greensboro, USA, i_dissan@uncg.edu
[2]University of Texas at Arlington, USA, snerur@uta.edu
[3]University of North Carolina at Greensboro, USA, rahul@uncg.edu
[4]Northeastern University, USA, y.lee@northeastern.edu

Abstract

Medical crowdsourcing offers hope to patients who suffer from complex health conditions that are difficult to diagnose. Such crowdsourcing platforms empower patients to harness the "wisdom of the crowd" by providing access to a vast pool of diverse medical knowledge. Greater participation in crowdsourcing increases the likelihood of encountering a correct solution. However, more participation also leads to increased "noise," which makes identifying the most likely solution from a broader pool of recommendations (i.e., diagnostic suggestions) difficult. The challenge for medical crowdsourcing platforms is to increase participation of both patients and solution providers, while simultaneously increasing the efficacy and accuracy of solutions. The primary objectives of this study are: (1) to investigate means to enhance the solution pool by increasing participation of solution providers referred to as "medical detectives" or "detectives," and (2) to explore ways of selecting the most likely diagnosis from a set of alternative possibilities recommended by medical detectives. Our results suggest that our strategy of using multiple methods for evaluating recommendations by detectives leads to better predictions. Furthermore, cases with higher perceived quality and more negative emotional tones (e.g., sadness, fear, and anger) attract more detectives. Our findings have strong implications for research and practice.

Keywords: Crowdsourcing, Online Healthcare Communities, Decision Support System, Solution Evaluation.

Sudha Ram was the accepting senior editor. This research article was submitted on March 12, 2018, and underwent three revisions.

1 Introduction

The use of the Internet as a healthcare information resource has dramatically increased in recent years (Goonawardene & Tan, 2013; Kordzadeh & Warren, 2017). A survey by the Pew Research Center (2012) found that 72% percent of US adult internet users sought health information online (Fox & Duggan, 2013a). Furthermore, one in four adults reported appealing to others who had experienced similar health issues (Fox & Duggan, 2013b). Prior studies have identified two primary types of online health communities, namely, peer-to-peer (P2P) and patient-to-doctor (P2D) (Peng, Sun, Zhao, & Xu, 2015). P2P health communities (e.g., PatientsLikeMe) are peer support groups that facilitate interaction among patients so that they can share their experiences and provide emotional support to one another (Peng et al., 2015), while P2D communities allow patients to access medical advice from healthcare professionals (e.g., Healthtap).

Interestingly, medical crowdsourcing platforms are another type of health community whereby information technology (IT) enables socially connected "crowds" to offer diagnostic suggestions to patients who suffer from chronic and perplexing ailments that are difficult to diagnose. In such platforms, the "wisdom of crowds" effect is facilitated by collaboration among patients and physicians who share similar health interests. Medical crowdsourcing communities, such as CrowdMed, provide emergent solutions[1] to health problems that have long defied diagnosis. Therefore, patients with chronic illnesses turn to such platforms in the hopes that the collective expertise and experience offered will provide an explanation and/or prognosis for medical conditions that have caused them prolonged pain and suffering (Sen & Ghosh, 2017). On these platforms, interactions occur between "detectives," who are typically medical practitioners or experienced patients, and "seekers," who are patients with chronic and puzzling medical conditions. While the body of literature on medical crowdsourcing (Prpić, 2015) is increasing, there is little empirical research on factors that influence detective participation or on the methods used to identify the most likely diagnosis from multiple user suggestions.

Crowdsourcing has been successfully used in clinical and epidemiology research for a variety of purposes, such as finding the structure of a protein molecule (Savage, 2012), examining disease through image analysis (e.g., identifying signs of diabetic retinopathy in eye images), estimating flu prevalence and propagation (Meyer, Longhurst, & Singh, 2016), identifying populations at risk for cancer, and predicting West Nile virus in mosquitos (Kaggle.com). An example is the overnight diagnosis by CrowdMed medical detectives of a young boy's extremely rare medical condition (PANDAS)[2] that had puzzled physicians for months (Arnold, 2014). With a few notable exceptions (e.g., Sen & Ghosh, 2017), there is a dearth of empirical studies on crowdsourced medical platforms in the IS literature.

It has been demonstrated in the crowdsourcing literature that the likelihood of at least one of the solvers finding an extreme value solution increases as the number of solvers grows (Boudreau, Lacetera, & Lakhani, 2011). These extreme values are particularly important when the problem is highly uncertain (Boudreau et al., 2011), as is often the case with rare medical conditions. While a larger pool of solvers (detectives) can yield more potential solutions (diagnoses), it also makes the process of eliminating poor solutions and selecting the correct diagnosis more challenging for both patients and platform providers. Therefore, medical crowdsourcing platform providers are challenged to identify ways to increase participation of solvers and seekers, while at the same time improving the quality of the potential solutions identified by the solvers. Our study addresses this important challenge by answering the following research questions:

RQ1: What factors influence the number of medical detectives who engage with a patient's case?

Specifically, we show that in addition to key factors identified in the crowdsourcing literature such as monetary compensation and duration of case, the detectives' perceptions of the quality and emotional tones (e.g., fear, sadness and anger) of the case also affect the number of potential diagnostic suggestions.

RQ2: How can we improve the process of selecting the correct diagnosis from a list of alternative medical recommendations made by the crowd?

Specifically, we utilize data analytics and clustering techniques to improve upon the existing algorithm for ranking the recommendations of the detectives.

We use data from CrowdMed, an online medical crowdsourcing platform on which patients with rare chronic conditions seek advice from both experienced patients and trained clinicians. Not surprisingly, our findings indicate that monetary "rewards" offered by patients on this site as well as duration of the illness increase participation. Interestingly, we also found that the sentimentality of the case in terms of its emotional tone and the perceived quality of the case has a significant impact on drawing solvers to the case. For example, high-quality cases with negative emotional tones (e.g., anger, fear, and sadness) induce more detectives to participate in solving the case. However, as noted earlier, the efficacy of finding the right diagnosis is negatively impacted by the number of potential solutions, as there is a greater likelihood of noise and/or false positives in the pool of proposed solutions. Currently, the platform uses a prediction market algorithm[3] to rank potential solutions using input from the community. Our results suggest that the accuracy of a diagnosis can be further improved by integrating the current prediction market algorithm with insights from textual analysis of symptoms obtained from both solvers' interactions with the patient (i.e., from discussion forums) and an external crowd such as Wikipedia.

[1] In the context of this paper, we refer to solutions as diagnostic suggestions provided by detectives

[2] See rare diseases list (https://globalgenes.org/rarelist/)

[3] This is a point system where "medical detectives assign points to indicate their confidence in the suggestions offered to a case" (https://www.crowdmed.com/faqs)

Our study makes important contributions to the scarce body of empirical research in medical crowdsourcing. First, it uses a unique dataset to provide an understanding of the various factors that impact the number of medical detectives who participate in a case. Second, our findings offer medical crowdsourcing platform providers insights into how they can improve the process of selecting the correct diagnosis. Thus, our study furthers the understanding of emerging phenomena such as crowdsourcing in the healthcare domain.

The remainder of this paper is organized as follows. The next section reviews the literature related to this study, which is followed by a description of our research model and justification for the hypotheses that emerge from it. Subsequently, we describe our methodology, and then present the findings of our analysis. Finally, the concluding section discusses the theoretical and managerial implications as well as the limitations of our study, followed by directions for future research.

2 Background and Literature Review

In any crowdsourcing model, the crowd plays an important role in the decision-making process by helping to solve problems that are difficult for decision makers. Chiu, Liang, & Turban (2014) adopted Herbert Simon's decision process model to explain how crowds can be involved in different phases of the decision making process, such as intelligence (e.g., information gathering/sharing), design (e.g., generation of alternative solutions), and choice (e.g., evaluation through crowd voting). On the CrowdMed platform we investigated, patients provide much of the information (e.g., the symptoms) pertinent to the medical problem, and the detectives' (i.e., crowd's) involvement in the intelligence phase was limited to possibly gathering additional information or seeking clarifications through the discussion forum. However, the crowd was involved to a greater degree in the design and choice phases of the decision-making process.

As in many crowdsourcing models, three parties are involved in online medical crowdsourcing communities such as CrowdMed: (1) the patient (also referred to as a seeker) who is looking for a solution to a medical problem, (2) medical detectives or solvers who provide plausible solutions to seekers' medical problems, and (3) the platform that facilitates the interaction between seekers and solvers. Seekers may be willing to offer financial compensation to incentivize the detectives to find a solution to their health-related problems. With the recent rise in online medical crowdsourcing communities, the number of patients using these platforms to seek medical information and emotional support has steadily increased. For instance, Yan, Tan, Yan, and Sun (2012) showed that patients are more likely to turn to other patients with similar health issues to learn and understand their problems and to identify effective coping mechanisms.

Crowds in these communities include experts (e.g., physicians, nurses, medical researchers) in the field as well as novices (e.g., patients, regular people). While the "wisdom of the crowd" phenomenon has the potential to expeditiously resolve longstanding medical conditions that have defied explanation, it presents some serious challenges in terms of filtering and evaluating multiple recommendations in order to identify the correct diagnosis. Sen & Ghosh (2017, p. 3294) note that "crowdsourcing systems should embrace tools that provide filtering mechanisms to identify high-quality inputs from the crowd, aggregate them for evaluation, and 'purge' erroneous contributions." This implies that the role of platform providers should extend beyond simple facilitation of interactions among participants and include evaluation and identification of high-quality inputs. In this regard, information technology has the potential to aid platform providers in developing and implementing mechanisms to help seekers find better solutions.

The following subsection provides background information for an algorithm that was developed to assess the quality of detectives' recommendations. This is followed by a review of the literature on emotions that will serve as the conceptual foundation for our research model and hypotheses.

2.1 Solution Evaluation Process

Several crowdsourcing platforms, such as Threadless, use the same crowd to both generate and evaluate solutions (Bao, Sakamoto, & Nickerson, 2011; Malone, Laubacher, & Dellarocas, 2009; O'Leary, 2016). On such platforms, voting is the primary mechanism of evaluation and the crowd votes up high-value solutions and/or votes down low-value solutions (Buettner, 2015). Some platforms have sophisticated algorithms (e.g., CrowdMed's prediction market algorithm) that consider the expertise and qualifications of evaluators in order to identify the potential value of solutions. Bao et al. (2011) compare the effectiveness of two different evaluation mechanisms: prediction voting and Likert-type scale rating. They showed that prediction voting helps to eliminate low-fit solutions in the early stages, while Likert-type scale is more appropriate at later stages when the system is more mature. Some crowdsourcing platforms that specialize in data science projects (e.g., Kaggle) use validation data sets to test the prediction accuracy of solutions provided by the crowd. Walter & Back (2013) presented a text-mining-based approach to identify clusters that help segment and filter out low-value ideas received from the crowd. They argued that smaller clusters that contain fewer submissions (up to three) represent the most innovative ideas. Blohm, Riedl, Füller, and Leimeister (2016) developed an experimental method using rating-scale and preference-markets mechanisms to compare the value of ideas evaluated. The rating scale mechanism of idea evaluation led to higher accuracy when compared with the preference-market mechanism (Blohm, Riedl, Leimeister, & Krcmar, 2011). Graph theoretic, semantic deferential, bootstrapping, and probability are some of the other approaches that have been used to evaluate outcomes of crowd-workers (Buettner, 2015). In this paper, we employ a relatively novel approach to assess the quality of recommendations made by detectives. Specifically, we investigate whether the combination of multiple evaluation methods involving text clustering and prediction market algorithms, as well as the pooling of knowledge from internal and external crowds (i.e., Wikipedia), can improve the effectiveness of the evaluation mechanism.

2.2 Message Characteristics (Emotional Tone and Quality)

The findings of prior studies in crowdsourcing suggest that platform or project design characteristics, such as compensation structure, contest duration, and project complexity, impact the number of solvers who participate in a contest (Yang, Chen, & Pavlou, 2009; Zheng, Li, & Hou, 2011) as well as the quality of the solution (Archak, 2010). However, to the best of our knowledge, none of these studies have paid much attention to the effect of tone or sentiments in problem specifications. In contrast to crowdsourcing platforms in other areas, medical crowdsourcing communities have a distinct sentimental element, since the health and often the lives of patients are at stake. Patients with serious medical conditions (e.g., cancer patients) often experience strong negative emotions such as fear, sadness, and anxiety (Kennifer et al., 2009). Such patients feel a sense of relief in expressing their emotional distress to their healthcare providers, often eliciting an empathic response from them (Alexander et al., 2014).

Emotion refers to "a mental state of readiness that arises from cognitive appraisals of events or thoughts" (Bagozzi, Gopinath, & Nyer, 1999 p. 184). It can be either negative or positive. Researchers have argued that negative information is processed more thoroughly than positive information and that negative emotions have a stronger impact than positive ones (Baumeister et al., 2001). Indeed, (Lazarus, Kanner, & Folkman, 1980, p. 190) note that "negatively toned emotions such as fear, anxiety, anger, guilt, and sadness-depression" overwhelmingly dominate the research on emotions in the psychology literature.[4] In medical crowdsourcing communities, patients' emotions are expressed through the tone of their online messages. Therefore, the way a case is framed (e.g., the quality and tone of the case description) could impact the number of detectives who choose to engage in providing viable solutions to a patient's problems. Beyond the general positive or negative tone of a message, specific affective content can play a role in how medical detectives select cases to engage with. A survey of the extant literature in medical crowdsourcing suggests that the role of emotion in attracting participants is underresearched. In the attempt to fill this gap, our study draws on the longstanding research on emotions to explore the effect of distinct negative emotions on detectives' choice of cases in medical crowdsourcing.

Batson et al. (1989) addressed an important question that is not only pertinent but also central to our hypotheses related to prosocial behaviors associated with negative emotions such as sadness and anger. Specifically, they posed the following question: "Does feeling empathy for a suffering person evoke altruistic motivation?" (Baston et al., 1989, p. 922). Their study as well as several others have affirmed the empathy-altruism hypothesis, which suggests that the altruistic desire to alleviate the suffering of others motivates empathic individuals to render help (e.g., Batson, Duncan, Ackerman, Buckley, & Birch, 1981; Batson, O'Quin, Fultz, Vanderplas, & Isen, 1983; Batson et al., 1988; Toi & Batson, 1982). However, Cialdini et al. (1987) have argued that the willingness to help stems not from altruism but from an egoistic motive to dispel the depressed mood (i.e., negative state) brought about by someone else's emotional distress. This is referred to as negative-state relief hypothesis. Despite their differing views on why empathy elicits helping behavior, both these perspectives are pertinent to our discussions because they support our proposition that negatively toned emotions are likely to be associated with prosocial behaviors.

2.2.1 Theoretical Perspectives on Emotions

Broadly speaking, emotions have been examined using either a dimensional framework or a discrete emotions model (Barrett, 1998). Typically, the former attempts to map all emotions on to a two-dimensional space depending on the extent to which they vary on two qualities that are widely acknowledged to be associated with affect, namely, valence and arousal (Yin, Bond, & Zhang, 2014). According to Barrett (1998, p.580), "Valence is a subjective feeling of

[4] This is reminiscent of the phrase, "bad is stronger than good," which appears to be a recurring theme in the psychology literature (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001)

pleasantness or unpleasantness; arousal is a subjective state of feeling activated or deactivated." The discrete emotions model, on the other hand, views emotions (e.g., anger, sadness, happiness) as "unique experiential states that stem from distinct causes (e.g. Izard, 1977)" (Barrett, 1998, p. 581). One of the shortcomings of the dimensional model is that emotions that exhibit similar valence and arousal may nevertheless elicit different behaviors, as they may "involve distinct phenomenology" (Yin et al., 2014, p. 542).

The cognitive appraisal theory of emotions (see Lazarus et al., 1980) argues that affect and cognition are intertwined, with a cognitive appraisal or evaluation of the stimulus event and the environment in which it occurs being a precursor of the evoked emotion. Thus, an individual's reactions to a stimulus (i.e., the behavior and the effect arising therefrom) are a consequence of the interdependence between affect and underlying cognitive processes. According to (Lazarus et al., 1980, p.189), "cognitive processes shape the quality and intensity of a given emotional response." Plutchik (1980) articulates a framework that attempts to capture the interplay among the stimulus event in the environment, the cognitive processes at work, and the emotion that manifests. In this framework, the emotion that guides individual adaptation and behavior in response to a stimulus is represented as a "complex chain of reactions" (Plutchik, 1980, p. 12). The feeling (e.g., fear, sadness) —what is normally referred to as an emotion—induced in an individual by a stimulus in the environment is mediated by a cognitive evaluation that determines whether the situation is likely to be advantageous or disadvantageous (e.g., perilous or safe, harmful, or beneficial). As a consequence of this feeling, the individual exhibits a suitable behavior (e.g., runs to avoid danger or cries because of bereavement) that presumably has a desired effect (e.g., avoidance of the threat or receiving social support/help to overcome grief). Smith and Pope (1992) have expressed similar views in their review of the appraisal theory of emotions. Table 1 summarizes this sequence of steps.

On CrowdMed, the platform of interest in our study, medical detectives are presented with descriptions of symptoms written by patients who have been suffering from a chronic illness, often for a protracted period. Based on our preceding discussions, it is reasonable to expect the emotional tones derived from linguistics features of the textual account of the symptoms to affect the detective. The stimulus event, in this case, is the written description of symptoms. This may be cognitively interpreted by a detective as "suffering," leading to a feeling (i.e., emotion) of sadness. This feeling of sadness elicits an appropriate response (e.g., attempting to diagnose the ailment) that may eventually result in the desired effect of alleviating or eliminating the pain and suffering that the patient has long endured. The preceding discussions provide the conceptual backdrop for our research model.

Table 1. The Sequence of Events Related to the Development of an Emotion (Plutchik, 1980, p. 11)

Stimulus Event	**→ Cognition →**	**Feeling**	**→ Behavior**	**→ Effect**
Threat by enemy	→ Danger →	Fear	→ Run	→ Protection
Loss of parent	→ Isolation →	Sadness	→ Cry for help	→ Assistance

3 Research Model

The success of our model is predicated on the assumption that a larger and perhaps more diverse pool of knowledge leads to more solutions that have high value for a seeker. This is suggested by previous literature (for example, Boudreau et al., 2011). Inevitably, a larger pool of knowledge implies greater noise in the data because of a larger number of low-value solutions. Therefore, it is useful for medical crowdsourcing platforms (e.g., CrowdMed) to search for ways to improve the filtering and evaluating mechanisms employed, so that they can provide high-value solutions. Our study explores ways to both (1) attract more medical detectives to participate in a case, and (2) evolve an approach to quickly sift through alternatives and identify high-value solutions.

Our research model has two parts. First, we investigate factors that might influence medical detectives' decisions to participate in a patient's case. Second, we explore how to improve the process of evaluating and selecting high-quality diagnostic suggestions provided by medical detectives.

3.1 Number of Detectives

The ability to attract more detectives is paramount for medical crowdsourcing platforms that exist for the express purpose of providing correct diagnoses of rare medical conditions. It is important to note that on CrowdMed, the medical crowdsourcing platform that we analyze, patients do not engage in face-to-face conversations with potential detectives. Instead, their primary communication is posting a description of their symptoms on the platform. Thus, patients' ability to lucidly describe their medical condition, as well as the tone that they use to frame their case, can impact the number of detectives who endeavor to solve the case. Therefore, framing a case in a way that attracts the attention of many detectives is crucial.

According to the limited capacity theory, attention is mainly controlled by two factors: the member (i.e., detective) and the characteristics of the message (i.e., case description) (Bolls et al., 2001; Lang, 2000). In this study, we focus on the characteristics of the message since information about the detectives is not directly available. Specifically, detectives can purposefully select cases based on their interests (e.g., background and "rewards"), or can be persuaded to participate by an affective state evoked by emotional tones latent in a particular case description (Bolls et al., 2001; Lang, 2000). Prior studies have provided overwhelming evidence that negative emotional tones get more attention than positive ones (Bolls et al., 2001; Lang, 1995). This also conforms to the argument that "humans are…hardwired to allocate more attention to negative stimuli" (Bolls et al., 2001, p. 635). Recent literature supports the view that tone is more than simple positivity or negativity; each distinct emotion with the same valence plays a different role (Yin et al., 2014). For example, an experiment conducted in a cancer care setting showed that oncologists are more responsive when patients express intense negative emotions such as sadness (Kennifer et al., 2009). The study also found that oncologists respond with greater empathy to sadness than to fear (Kennifer et al., 2009).

Yin et al. (2014) support the view that emotional tones (e.g., sadness, anger, and fear) are more nuanced than just valence (positivity or negativity) and that assessing their effects independently may be appropriate. The three discrete emotions used in this study—namely, sadness, anger, and fear—belong to a set of basic negative emotions that people experience in everyday life (Shaver, Schwartz, Kirson, & O'connor, 1987). In the context of our study, it is perhaps then reasonable to assume that symptoms described by patients with chronic health problems are laden with these negatively toned emotions

The marketing literature is replete with studies that have examined the role of emotions in predicting consumer behavior (de Hooge, 2014). For example, Bagozzi and Moore (1994) demonstrated that negative emotions such as anger, sadness, fear, and tension elicited by child abuse advertisements are all positively associated with the desire to help abused children. In a similar vein, Burt and Strongman (2005) found that negative emotions embedded in images used in charity advertising were positively associated with greater contributions in terms of money, the number of items, and time. A brain imaging study by FeldmanHall, Dalgleish, Evans, and Mobbs (2015) also suggests that altruistic prosocial behavior is motivated by empathic concern for others rather than by an egoistic desire to alleviate one's own distress or negative state caused by the suffering experienced by others. As mentioned in the literature review section, there are a number of studies that use either the empathy-altruism hypothesis (e.g., Batson et al., 1981; Batson et al., 1983; Batson et al., 1988; Toi & Batson, 1982) or the negative-state relief hypothesis (Cialdini et al., 1987; Schaller & Cialdini, 1988) to provide a rationale for the positive influence that negative emotions have on prosocial behaviors.

Based on the preceding discussions, we hypothesize:

H1: The negative emotional tone expressed in a patient's case is positively associated with the number of medical detectives who participate in the case.

H1a: The sadness expressed in a patient's case is positively associated with the number of medical detectives who participate in the case.

H1b: The anger expressed in a patient's case is positively associated with the number of medical detectives who participate in the case.

H1c: The fear expressed in a patient's case is positively associated with the number of medical detectives who participate in the case.

Based on their interest in solving strangers' rare medical cases, it may be assumed that detectives are likely to engage in critical thinking and have high needs for cognition. Such people may be significantly influenced by the quality of the message presented (Wilson, 2007). In the healthcare domain, effective case presentation plays an important role, and is identified as an essential skill for healthcare practitioners. A case presentation typically includes the history of the relevant illness as well as an explanation of the various diagnostic results. In CrowdMed, patients present their cases to medical detectives by themselves. The platform allows them to populate fields providing relevant information, such as their demographics, symptom details, current medications, problems categorized by specific body systems, personal medical history, family medical history, personal lifestyle, and any available secondary or partial diagnoses. A detective's perception regarding the quality of a case will be affected by how effectively patients present this information. The perceived quality of the case, in turn, is likely to influence whether or not a detective will choose to participate in the case. Thus, we hypothesize:

H2: Perceived quality of a case is positively related to the number of detectives who participate in the case.

In addition to the preceding factors, monetary compensation and duration (i.e., the length of time the case has been on CrowdMed) could also influence how many detectives choose to participate in a case. Yang et al. (2009) found duration and compensation to have a positive relationship with the number of detectives participating in a case in the context of online competitions; however, these relationships have not been investigated in the context of medical crowdsourcing (Meyer et al., 2016). According to Ariely et al. (2009), the propensity to engage in prosocial behavior (e.g., contributing money or donating blood) may be due to intrinsic, extrinsic, or image motivation. For example, Zheng et al. (2011) found extrinsic motivation in the form of monetary compensation to be positively related to solvers' participation intentions. Higher "rewards" not only provide extrinsic motivation but they also compensate detectives for their time. Thus, cases offering substantial rewards are likely to attract more detectives. Similarly, just as long duration auctions attract more bids, prolonged medical cases that remain open and unsolved for longer periods of time are also likely to attract more detectives. (Yang et al., 2009). Thus, we hypothesize:

H3: The monetary compensation offered by a case is positively associated with the number of detectives who participate in the case.

H4: The duration of a case is positively associated with the number of detectives who participate in the case.

$$\begin{aligned} &Number\ of\ Detectives_i \\ &= \alpha_0 + \alpha_1\, Anger_i + \alpha_2\, Fear_i + \alpha_3\, Sadness_i \\ &+ \alpha_4\, Quality_i + \alpha_5\, Reward_i + \alpha_6\, Duration_i \\ &+ \alpha_7\, Description\_Length_i \\ &+ \alpha_8\, Symptoms\_Began_i + \delta_j + \varepsilon_i \qquad (1) \end{aligned}$$

where: i denotes cases; α_k (k=0…8) represents the coefficients of the variables; and δ_j denotes the package dummies, which are included to control for the effect of the package purchased. Details of available packages and other control variables (e.g., Description_Length, Symptoms_Began) are discussed in the variable definitions section.

3.2 Evaluation Process (Selecting the Correct Diagnosis)

Detectives on CrowdMed select cases of interest to them among the cases posted by chronically ill patients and offer their recommendations or potential solutions. CrowdMed uses a patented prediction market algorithm that relies on a weighted voting system to find the most probable solution among detectives' suggestions (Crocker, 2015). This algorithm ranks solutions by assigning points to them based on "relative popularity." Weighted voting by a large pool of experts helps eliminate noisy solutions and makes this prediction market algorithm stable (Mozolyako & Osipov, 2015). As per CrowdMed, detective expertise and information about past performance are taken into account when determining rankings for solutions provided. CrowdMed's prediction market algorithm is proprietary and is not available to the public.

In crowdsourcing, it has been shown that pooling knowledge from multiple individuals enhances the likelihood of finding an unusual solution (Boudreau et al., 2011). Likewise, it may be argued that the integration of diverse knowledge from multiple sources (i.e., wisdom of crowds) increases the likelihood of finding solutions to complex problems. Furthermore, integrating multiple recommendations made by crowds enhances the creativity of solutions, as the pool of knowledge is more diverse and is likely to cover the solution space to a greater degree. Therefore, in this study we combine multiple methods in order to improve outcomes. Specifically, we use a clustering method to identify a group of solutions that are closer to the problem at hand. This helps to identify and filter out low-fit solutions. The prediction market algorithm is then used to re-rank the solutions within the clusters, thus enhancing the effectiveness of the evaluation method (see Figure 1). Thus, we argue that integration of knowledge from multiple crowds and the use of multiple evaluation techniques (text clustering along with CrowdMed's prediction market algorithm) will yield more accurate outcomes than CrowdMed's ranking algorithm alone.

Figure 1 explains the process using a sample case. Column 1 shows the original rankings of potential diagnoses recommended by the crowd. These rankings were based on CrowdMed's prediction market algorithm. As per Column 1 rankings, the best diagnosis (i.e., the one eventually selected by the patient) is ranked sixteenth. Columns 2 and 3 explain the two-step process we used in our study to improve on the rankings.

For each recommended diagnosis, we gathered symptom information (in the form of text) from an external source (i.e., Wikipedia) provided by the solver. This text was then combined with discussions about the diagnosis, if there were any. Thus, there were as many textual descriptions of symptoms as there were detective recommendations (a total of 19 in the example shown in Figure 1). These texts, along with the original description of symptoms provided by the patient, were then subjected to agglomerative hierarchical clustering using cosine distances.

The cluster analysis program yielded two broad clusters, one of which contained the original symptoms submitted by the patient. In our example, Cluster 1 contains the original case. The recommended diagnoses in Cluster 1 are likely to be more accurate than those that appear in Cluster 2 because the symptoms associated with them (i.e., diagnoses in Cluster 1) better mirror the original symptoms provided by the patient. We then used the prediction market algorithm rankings shown in Column 1 to re-rank the recommended diagnoses within each cluster, ensuring that the ones within the same cluster as the original case (i.e., Cluster 1 in our example) were ranked higher than those appearing in the other cluster.

In our example, among all the recommended diagnoses in Cluster 1, Diagnosis 4 was ranked the highest by CrowdMed's prediction market algorithm (see Column 1) and was therefore deemed to be the most likely solution (i.e., ranked first) by our pooled procedure. In addition, the diagnosis that was eventually determined to be the best one (by the patient) moved up from its original rank of sixteen (as per the prediction market algorithm) to ninth when our technique was used. We, therefore, argue that the clustering procedure helps to not only eliminate low-fit recommendations but also identify the most probable solutions. This is similar to the clustering process that Walter and Back (2013) used to identify the most innovative submissions to crowdsourcing contests.

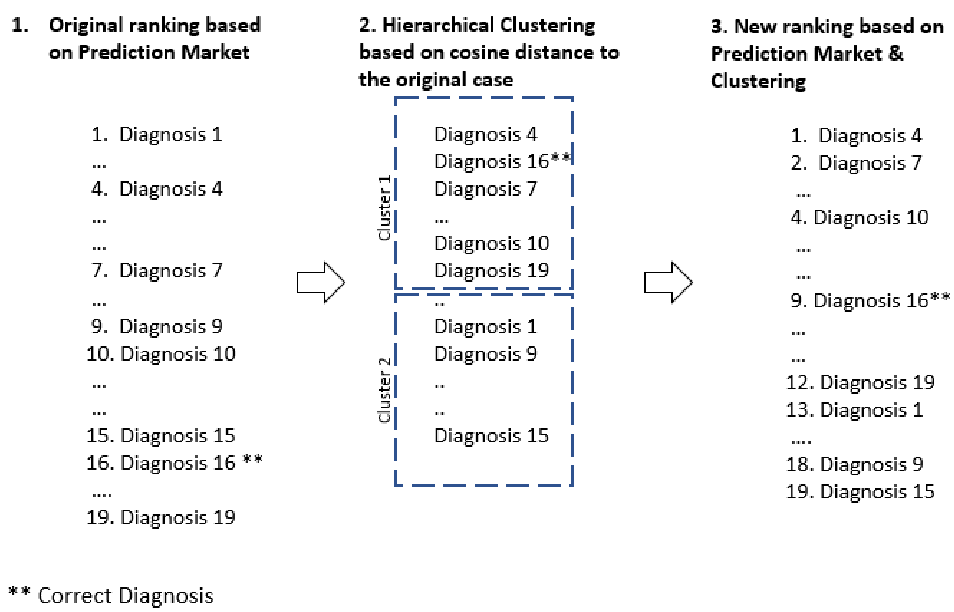

Figure 1. Graphical Representation of Algorithm for Deriving New Rankings of Diagnoses

4 Data Collection and Variable Definitions

4.1 Data Collection

Data for our study was obtained from crowdmed.com, a specialized crowdsourcing platform that focuses on medical cases. As in many crowdsourcing models, there are three parties involved in this business model: seekers (patients), solvers (medical detectives), and the platform provider (crowdmed.com). Medical detectives who participate in this platform come from different backgrounds and geographical locations, and may include credentialed physicians, medical students, nurses, pharmacists, physician assistants, chiropractors, medical researchers, scientists, and patients who have experience with and/or knowledge about similar medical conditions (crowdmed.com). CrowdMed claims that 63% of their medical detectives either work in or study medicine. Patients with a history of chronic health problems can post their medical cases along with clinical information on crowdmed.com (see Figure 2).

Unlike other online patient communities, typical patients on CrowdMed are at their wit's end, having failed to find a solution to their problems over a prolonged period. For example, a survey of patients on CrowdMed indicated that they had visited a median of five physicians, incurred a median of $10,000 in medical expenses, and spent a median of 50 hours researching their illness online (Meyer et al., 2016). In our patient sample, the average number of years since the symptoms first appeared was eight, while the minimum was three months. Thus, all these cases fall within the realm of "chronic disease," as defined by the US National Center for Health Statistics (see http://www.medicinenet.com/script/main/art.asp?articlekey=33490).

As mentioned before, medical detectives can choose the cases they wish to solve. They can suggest diagnoses and/or solutions, or can vote (by allocating points) for diagnoses and/or solutions recommended by other solvers. The platform also features a peer-flagging mechanism that helps eliminate poor recommendations. Furthermore, a credentialed physician moderates every case. CrowdMed also provides an open discussion forum that facilitates the sharing of knowledge and information between the patient and the solvers engaged in the case, while also enabling interactions among medical detectives.

CrowdMed rates medical detectives based on their professional qualifications, as well as on their performance on the CrowdMed platform. Medical detectives can improve their rankings by suggesting a correct diagnosis and/or allocating points to an acceptable recommendation suggested by others. Higher ratings allow them to participate in more complex and high-reward medical cases.

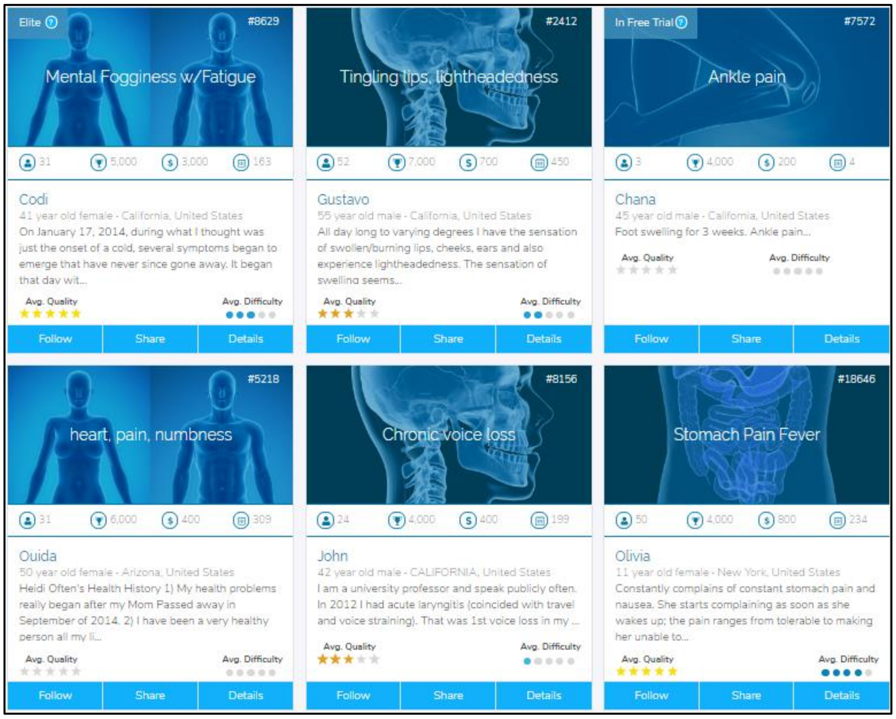

Figure 2. Sample Cases (CrowdMed[5])

4.2 Dependent and Independent Variables

A brief description of the variables used in our study is given below.

Number of detectives: This is the total number of detectives who participated in each case.

Reward: Refers to the monetary compensation that a patient offers for a correct diagnosis.

Duration: Number of days that a case is open to detectives.

Quality: This is the average quality of a case. Detectives (irrespective of whether they participate in the case or not) can rank cases in terms of their perceived quality. For the quality measure, the platform uses a Likert-type scale from 1 to 5, where 1 indicates poor quality and 5 indicates the best quality. The average quality is calculated by taking the average of individuals' rankings for the quality of the case.

Emotional tone: We used IBM's Tone Analyzer API (application programming interface) to perform a linguistics analysis using text information related to patient symptoms, lifestyle, and family background to identify emotional tones (e.g., anger, sadness, fear) associated with the case. Drawing on theoretical and empirical insights from psycholinguistics, the IBM Tone Analyzer uses machine learning to assess emotional tones latent in any written text.[6] According to IBM's benchmarking studies, their ensemble model performs better than other popular models used for

[5]See https://www.crowdmed.com/case-selection/, patients' names are fictitious

[6]https://console.bluemix.net/docs/services/tone-analyzer/science.html#the-science-behind-the-service

deriving emotional tone categories from text. A brief description of the emotional tones is presented in Table 2 below.

4.3 Control Variables

The following control variables were used in our study:

Package Purchased: [7] CrowdMed offers different packages to patients. In our dataset, we had cases with four different types of packages (Elite, Premium, Standard, and Priority), each of which offered different monetary "rewards" to detectives. Thus, the package purchased is likely to impact the number of medical detectives who participate in a case. We considered cases without any package as our base and added dummy variables to control for the effect of package purchased.

Description Length: This is the number of words used to describe symptoms. Studies in crowdsourcing have shown that the description length of a project impacts solvers' decision as to whether to participate in a project or not (Yang et al., 2009). This is used to control for the complexity of the case.

Symptoms Began: This shows when the symptoms began for the first time. This is also used to control for the complexity of the case.

In this study, we collected data related to all 328 completed medical cases that were available online in March 2016. Tables 3 and 4 show the descriptive statistics and correlation matrix, respectively.

Table 2. Emotional Tone (IBM Tone Analyzer[8])

Description
A response to impending danger. It is a survival mechanism that is a reaction to some negative stimulus. It may be a mild caution or an extreme phobia.
Indicates a feeling of loss and disadvantage. When a person can be observed to be quiet, less energetic and withdrawn, it may be inferred that sadness exists.
Evoked due to injustice, conflict, humiliation, negligence or betrayal. If anger is active, the individual attacks the target, verbally or physically. If anger is passive, the person silently sulks and feels tension and hostility.

Table 3. Descriptive Statistics

Variable	Mean	Std. Dev.	Min.	Max.
Number of Detectives	18.81	14.80	1.00	135.00
Anger	0.21	0.14	0.01	0.83
Fear	0.38	0.17	0.04	0.72
Sadness	0.59	0.10	0.09	0.86
Quality	3.99	0.83	1.00	5.00
Reward	159.82	181.06	0.00	1100.00
Case Duration (days)	441.52	186.27	62.00	971.00
Symptoms Began (months)	99.55	116.28	3.00	868.00
Description Length (Symptoms)	278.47	331.30	4	3392

[7] https://www.crowdmed.com/select-package

[8] https://www.ibm.com/watson/developercloud/doc/tone-analyzer/understand-tone.html

Table 4. Correlation Matrix

		1	2	3	4	5	6	7	8	9
1	**No. of detectives**	1.00								
2	**Anger**	0.06	1.00							
3	**Fear**	0.07	-0.27***	1.00						
4	**Sadness**	0.08	-0.42***	0.00	1.00					
5	**Quality**	0.22***	-0.05	0.08	0.05	1.00				
6	**Reward**	0.65***	0.01	0.04	0.08	0.25***	1.00			
7	**Case duration**	0.19***	0.01	-0.05	0.02	-0.03	-0.16***	1.00		
8	**Symptoms began**	0.08	0.06	-0.06	0.02	0.00	0.08	0.03	1.00	
9	**Description length**	0.13**	-0.07	0.15***	0.12**	0.23***	0.22***	-0.17***	0.04	1.00
*** p < 0.01, ** p < 0.05, * p < 0.1										

5 Results

5.1 Number of Detectives

We used a negative binomial regression (NBR) model for our analysis, because it fits well with our data characteristics. Our main dependent variable—number of detectives—is count data, and ordinary least squares (OLS) regression is not appropriate because of the skewness of the data. Poisson and negative binomial models are commonly used for count data. However, our data present overdispersion relative to the poisson distribution. Furthermore, the log likelihood ratio test of alpha suggested that negative binomial distribution is superior to poisson in this case (Cameron & Trivedi, 1998; Martinez-Espineira, 2007). Stata 14.2 was used to test our models.

To test for multicollinearity, we used a linear model to examine the variance inflation factor (VIF). VIF was found to be less than 2, indicating that multicollinearity was not an issue. The quality variable had values for only 167 cases. Four approaches were used to deal with the missing values in the quality measure. First, we dropped all the missing cases and ran the model with only 167 complete cases. However, this reduced the sample size significantly. Furthermore, the 167 cases may not be an accurate representation of the population.

Table 5. NBR Results for Number of Detectives

	Model 1	**Model 2**	**Model 3**	**Model 4**
Anger	0.5655 **	0.7707 ***	0.7712 ***	0.7688 ***
Fear	0.3481 *	0.4230 ***	0.4231 ***	0.4186 ***
Sadness	0.9142 ***	0.5660 **	0.5668 **	0.5564 **
Quality	0.1146 ***	0.0992 **	0.0992 **	0.1136 ***
Reward	0.0016 ***	0.0021 ***	0.0021 ***	0.0020 ***
Case Duration	0.0020 ***	0.0024 ***	0.0024 ***	0.0024 ***
Quality Missing			-0.0020	
Symptoms Began	0.0002	0.0002	0.0002	0.0002
Description Length	0.0000	0.0001	0.0001	0.0000
Sample Size	167	328	328	328
Log likelihood	-551.10	-1099.36	-1099.36	-1098.57
Pseudo R2	0.1564	0.1237	0.1237	0.1243
Notes: *** p < 0.01, ** p < 0.05, * p < 0.1				

Second, we replaced missing quality values with the average and ran the model. Third, we included a dummy variable in the model to indicate whether the quality values were missing or not. Fourth, we estimated missing values regressing the quality variable on gender, reward, and complexity. The results of these four approaches are shown in Table 5 as Models 1, 2, 3, and 4, respectively.

The coefficients of anger, fear, and sadness are positive and significant in all four models. Thus, Hypotheses 1a, 1b, and 1c are supported. The results suggest that cases with negative emotional tones such as anger, sadness, and fear are more likely to attract detectives. The coefficient of perceived quality is positive and significant in all models, thus supporting Hypothesis 2. This implies that detectives are more likely to select cases with clear descriptions. Consistent with the findings of prior research, both reward and duration showed positive and significant relationships with the number of detectives in all models, thus supporting Hypotheses 3 and 4. In summary, results of all four methods support all the hypotheses, thus confirming the robustness of our model. Table A1 in Appendix A shows standardized coefficients of variables of all four models. Furthermore, Table A2 shows a comparison of model fit results for restricted (controls only) and unrestricted models based on Model 4. The unrestricted model shows lower values for AIC (Akaike's information criterion) and BIC (Bayesian information criterion) statistics compared to the restricted model, suggesting that the unrestricted model has a better overall fit.

We re-ran Model 4 after controlling for the types of problems in the cases (e.g., Neurological, head or cardiovascular, breathing) to see whether they had an impact on the number of detectives who participated. Only 158 cases had specified the main problem area. Owing to the low sample size, this analysis was performed without control variables. As can be seen from the results shown in Table A3 in Appendix A, all the hypotheses are supported.

5.2 Evaluation Process

In order to evaluate our pooled approach, we randomly selected a sample comprising 10% of the archived cases from the data we had collected from CrowdMed, ensuring that the correct diagnosis was among the list of recommendations made by the detectives. Wikipedia was used as the main source of external crowd knowledge. For each case, we extracted symptom information from Wikipedia for all the diagnostic suggestions. The text that was extracted from Wikipedia was restricted to the details of the symptoms. Many of the Wikipedia pages had a separate section for "sign and symptoms" from which we obtained symptom information for the recommendations in our sample cases (see Figure A4).

Subsequently, we combined the description of the symptoms for each diagnosis (i.e., the text from Wikipedia) with relevant text from the discussion forums in which the detectives(s) interacted with the patient (see Figure 3).

We used standard text-mining procedures to process the data. First, we converted the text to lowercase and then preprocessed it to eliminate white spaces, numbers, punctuations, and stop-words. In addition to common stop-words, we also eliminated frequently occurring context-specific words that are not really helpful for understanding symptoms (e.g., patient, symptom). Words that had more than one form were reduced to their root form through a process called stemming. We then created a term document matrix (TDM) based on the frequency of terms occurring in a document. Each diagnostic suggestion and the original case were regarded as separate documents. We then computed cosine distance associated with these documents to create a distance matrix. This distance matrix served as an input to a hierarchical clustering algorithm. Specifically, we used the Ward's method to obtain hierarchical clusters showing the proximities of diagnostic suggestions.

Figure 4 shows the results of a sample case involving CrowdMed's prediction market algorithm. For this particular case, there were 20 documents, which included the19 diagnostic suggestions from the crowd (numbered 1 through 19) and the original description represented by the number 0. Following the procedure outlined in the preceding paragraph, we used these 20 documents to generate the distance matrix. Other numbers represent rankings based on CrowdMed's prediction market algorithm. For example, the number 1 was deemed to be the best diagnostic recommendation by the prediction market algorithm that CrowdMed uses to rank diagnostic suggestions received from detectives. The algorithm relies on the allocation of points by detectives and takes into account detectives' ratings as well as their backgrounds. Once the detectives offer their recommendations, the patients have 30 days to research and/or discuss the suggested diagnoses with their physicians to identify which one of the recommendations was the most accurate/insightful (see https://www.crowdmed.com/faqs). In this example, the patient identified number 16 as the correct diagnosis.

As shown in Figure 3, the clustering algorithm assigned the original case and the correct diagnosis to Cluster 1, and the solution picked by the prediction market algorithm was grouped into Cluster 2. We argue that the correct diagnosis will most likely be in the same cluster as the original case (Cluster 1 in the example), as the cluster would contain diagnostic descriptions that are lexically very similar to the patient's description of symptoms.

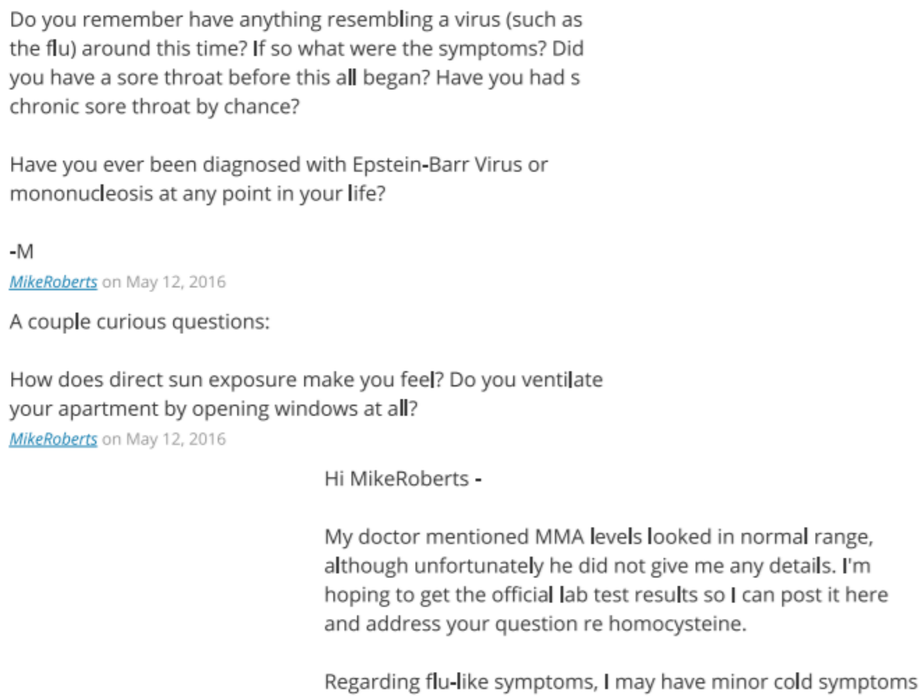

Figure 3. Sample Patient and Detective Chat (CrowdMed[9])

On the other hand, diagnoses that appear in other clusters are likely to be lexically distant from the original case. Thus, this approach helped filter out low-value suggestions that are far from the original case. Subsequently, within each cluster, we followed CrowdMed rankings to re-rank diagnostic suggestions such that diagnostic descriptions in Cluster 1 (in our example) get better rankings than those in Cluster 2. As per the suggested algorithm, the best diagnosis—as determined by the patient (often in consultation with his or her physician)—is now ranked ninth instead of sixteenth.

To validate our ranking algorithm, we used a non-parametric sign test (Snedecor & Cochran, 1989) to compare the rankings of the winning solution with and without clustering. Coffin and Saltzman (2000) note that a comparison of two algorithms can be done by using paired sample t-test, the sign test, or the signed rank test. We computed the differences in rankings of the correct diagnosis as per the original ranking algorithm (prediction market algorithm) and our new ranking algorithm (combination of clustering and prediction market algorithm). These ranking differences were skewed. Hence, we chose the sign test rather than the paired-sample t-test to compare algorithms (Coffin & Saltzman, 2000). A positive difference implies that the new ranking is superior to the original ranking algorithm. Our results show that the difference is significantly greater than 0 (p = 0.0318 < 0.05). Thus, we concluded that the new rankings were significantly better than the original rankings. Our findings suggest that pooling knowledge from multiple sources and combining multiple evaluation methods can significantly improve the likelihood of selecting the correct diagnosis. To further validate our findings, we ran the Wilcoxon signed test (Wilcoxon, 1945), which confirmed that these two algorithms are significantly different (P= 0.05 < 0.10). Figure 5 shows the word cloud of the case that was used to illustrate our algorithm.

[9]See https://www.crowdmed.com/sample-report

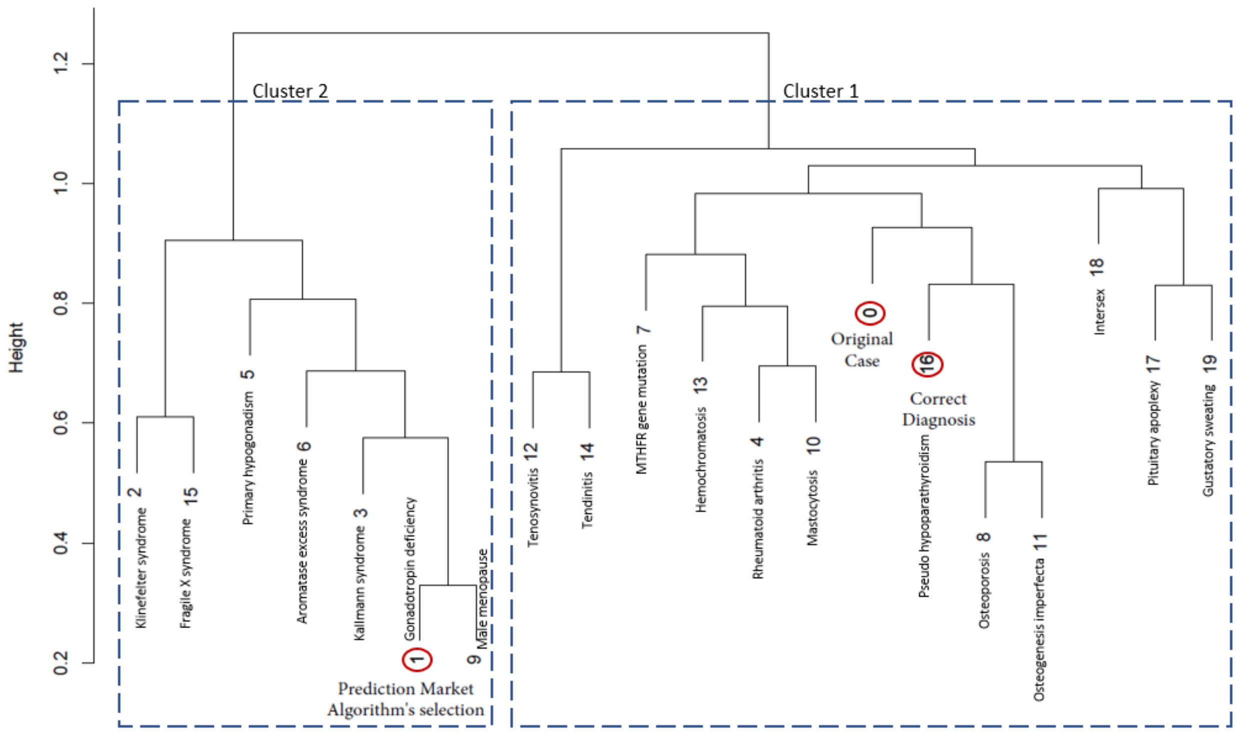

Figure 4. Cluster Dendrogram for a Sample Case

Figure 5. Word Cloud of a Sample Case

6 Discussion

6.1 Key Findings

Despite the growing interest in online medical crowdsourcing platforms, there is little or no empirical research on factors that affect the quality of solutions to medical conditions that have long been unresolved. An understanding of these factors can increase participation of detectives and improve the chances of speedily resolving undiagnosed medical conditions. This, in turn, would lead to greater acceptance and adoption of crowdsourcing platforms as a viable alternative to seeking medical help from one's primary care physician. Our study is a small but important step toward providing insights to platform designers, as well as to patients, on how to increase the likelihood of resolving medical mysteries. Specifically, our results complement the findings of prior studies by showing that not only large rewards and longer case durations, but also presentation quality and negative emotional tones contribute to the number of detectives who participate in a medical case. Patients may use these insights to frame their cases in a way that would attract more detectives. Platform providers could benefit from our findings as well. For example, they may evolve guidelines and platform changes that help patients submit high-quality cases that may induce more solvers to participate. Also, in light of our findings, detectives may alter the way they offer diagnostic suggestions and vote for the recommendations made by their fellow solvers. Furthermore, our study provides insights to platform providers on how to improve the process of selecting the best solution from the alternatives suggested by detectives.

6.2 Theoretical Implications

Crowdsourcing is a fairly recent phenomenon that harnesses the wisdom of crowds to solve problems faced by organizations, individuals, or researchers. Such platforms facilitate collaboration among a large number of people, thereby fostering a climate of collective intelligence that can provide novel solutions to seemingly intractable problems (e.g., Malone, Laubacher, and Dellarocas, 2009). This study focuses on CrowdMed, a medical crowdsourcing site that holds considerable promise for alleviating the pain and suffering of patients who have been chronically ill for a protracted period, with no traditional medical diagnosis or treatment forthcoming. Research on medical crowdsourcing, particularly in terms of attracting more detectives and anticipating the best diagnosis, is still in its infancy. Thus, our study is a step toward building a cumulative tradition in this nascent but rapidly evolving domain.

Our paper has several implications for research. First, it contributes to the emerging literature on crowdsourcing by demonstrating that the number of participants who engage with the case depends not only on design factors (e.g., monetary compensation), but also on the way the question is framed, as well as on the quality of the description. Second, it contributes to research in analytics by showing that combining existing algorithms with text analytics techniques could yield better diagnostic recommendations. Third, it shows that pooling the knowledge of different members of the crowd leads to better outcomes. Fourth, our study is among the first to use IBM's Tone Analyzer to derive emotional tones from text posted by patients on CrowdMed. Given the exponential growth in unstructured data such as text, an understanding of how to extract suitable variables (e.g., emotional and language tones, personality characteristics) from such data can be invaluable to researchers. Finally, to the best of our knowledge, our study is the first to demonstrate the impact of negatively toned emotions on prosocial behavior in the context of medical crowdsourcing. As Yin et al. (2014) observe, there is a paucity of empirical research in the IS domain on the role of affect. Our study is, therefore, a notable contribution to the sparse but growing body of literature in IS that examines the impact of affect, in general, and negatively toned emotions, in particular.

6.3 Implications for practice

Ultimately, crowdsourcing platforms should be designed in such a manner that they attract more problem solvers to engage with their platforms. Furthermore, the credibility of the platform rests on it being able to sift through a potentially large number of recommended solutions in order to identify the most accurate one for the problem at hand. The contributions of this study, therefore, are timely, as these platforms are still in early stages of evolution.

Crowdsourcing providers could use our research to gain insight into how platforms should be designed in order to enhance the engagement of all parties to facilitate the speedy resolution of challenging health problems. This, in turn, would facilitate the creation of knowledge useful to seekers and detectives alike. First, as our results suggest, tone and quality of the case description matter. Platform providers should provide guidance to patients on how to formulate a high-quality case that can increase detective participation. Second, our finding that a combination of text analytics techniques and the pooling of data from an external source can lead to a better ranking of possible solutions should be useful to platform designers. Above all, our results should be of great interest to medical crowdsourcing platforms that aim to expeditiously and effectively resolve undiagnosed chronic medical conditions.

6.4 Conclusion and Future Research

There is growing anecdotal evidence supporting the efficacy of medical crowdsourcing. Patients afflicted with hard-to-diagnose medical conditions are willing to expend time, money, and energy to derive a correct diagnosis that will mitigate their suffering and save them from spending enormous amounts of money in the quest for a cure for their ailments. Medical crowdsourcing platforms provide an environment for collective intelligence to emerge from the interchange of ideas among detectives and amateurs who may have experiential knowledge of the case under advisement. Furthermore, these detectives and amateurs could be from anywhere in the world, thus facilitating the pooling of diverse knowledge that cuts across national, cultural, and professional boundaries. The ability of such platforms to speedily identify diagnoses for rare medical conditions that have puzzled seasoned physicians is contingent upon (1) the number of detectives who undertake a case, and (2) the efficiency with which the platform can filter out alternative recommendations and identify the best diagnosis. Our study expressly addresses these concerns and offers suggestions for improving these platforms based on insight gained from our results.

As with many other empirical studies, our study has some shortcomings. However, we believe that these are minor and do not seriously impact our contribution. Furthermore, these limitations help us look at the problem from multiple perspectives and open up opportunities to pursue further research to generate actionable insights of value to medical crowdsourcing platforms. First, our data only include information that is publicly available on the website. For instance, we did not have access to all the information related to every detective who participated in a case. For example, we did not know the order in which detectives joined cases; it is conceivable that the ratings of detectives who have already joined a case may impact the number of detectives who subsequently choose to participate in the case. Second, we only pooled data from Wikipedia. Additional data sources, such as medical symptom databases, could perhaps lead to better results. Third, hierarchical clusters obtained through cosine similarities provided the sole basis for our results. We believe that a combination of text-mining techniques could be used to improve our findings. Fourth, our study assumes that the correct diagnosis is the one that is eventually accepted by the patient, either subjectively or in consultation with their physician(s). While CrowdMed employs peer evaluations, a point allocation system, and a case moderator to ensure that the recommendations are of top quality, it is possible that the diagnostic suggestion deemed to be the best one by the patient may not actually be the most accurate diagnosis.

Despite these limitations, the findings of our study should provide a good starting point for future research endeavoring to improve the design of medical crowdsourcing platforms in a way that will expeditiously deliver high-quality diagnosis to challenging medical problems.

Acknowledgments

We are grateful to Dr. Narayan Janakiraman of the University of Texas at Arlington and the *Journal of the Association for Information Systems* review team for their insightful comments and suggestions.

References

Alexander, S. C., Ladwig, S., Norton, S. A., Gramling, D., Davis, J. K., Metzger, M., . . . Gramling, R. (2014). Emotional distress and compassionate responses in palliative care decision-making consultations. *Journal of Palliative Medicine*, *17*(5), 579-584.

Archak, N. M. (2010). Glory and cheap talk: Analyzing strategic behavior of contestants in simultaneous crowdsourcing contests on TopCoder.com. *Proceedings of the 19th International Conference on World Wide Web* (pp. 21-30).

Arnold, C. (2014). "Can the crowd solve medical mysteries?" *NOVA Next.* Retrieved from http://www.pbs.org/wgbh/nova/next/body/crowdsourcing-medical-diagnoses/

Bagozzi, R. P., Gopinath, M., & Nyer, P. U. (1999). The role of emotions in marketing. *Journal of the Academy of Marketing Science*, *27*(2), 184-206.

Bagozzi, R. P., & Moore, D. J. (1994). Public service advertisements: Emotions and empathy guide prosocial behavior. *The Journal of Marketing*, *58*(1), 56-70.

Bao, J., Sakamoto, Y., & Nickerson, J. V. (2011). Evaluating design solutions using crowds. *Proceedings of the 17th Americas Conference on Information Systems*.

Barrett, L. F. (1998). Discrete emotions or dimensions? The role of valence focus and arousal focus. *Cognition & Emotion*, *12*(4), 579-599.

Batson, C. D., Batson, J. G., Griffitt, C. A., Barrientos, S., Brandt, J. R., Sprengelmeyer, P., & Bayly, M. J. (1989). Negative-state relief and the empathy-altruism hypothesis. *Journal of Personality and Social Psychology*, *56*(6), 922.

Batson, C. D., Duncan, B. D., Ackerman, P., Buckley, T., & Birch, K. (1981). Is empathic emotion a source of altruistic motivation? *Journal of Personality and Social Psychology*, *40*(2), 290-302.

Batson, C. D., Dyck, J. L., Brandt, J. R., Batson, J. G., Powell, A. L., McMaster, M. R., & Griffitt, C. (1988). Five studies testing two new egoistic alternatives to the empathy-altruism hypothesis. *Journal of Personality and Social Psychology*, *55*(1), 52-77.

Batson, C. D., O'Quin, K., Fultz, J., Vanderplas, M., & Isen, A. M. (1983). Influence of self-reported distress and empathy on egoistic versus altruistic motivation to help. *Journal of Personality and Social Psychology*, *45*(3), 706-718.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, *5*(4), 323-370.

Blohm, I., Riedl, C., Füller, J., & Leimeister, J. M. (2016). Rate or trade? Identifying winning ideas in open idea sourcing. *Information Systems Research*, *27*(1), 27-48.

Blohm, I., Riedl, C., Leimeister, J. M., & Krcmar, H. (2011). Idea evaluation mechanisms for collective intelligence in open innovation communities: Do traders outperform raters? *Thirty-Second International Conference on Information Systems*.

Bolls, P. D., Lang, A., & Potter, R. F. (2001). The effects of message valence and listener arousal on attention, memory, and facial muscular responses to radio advertisements. *Communication Research*, *28*(5), 627-651.

Boudreau, K. J., Lacetera, N., & Lakhani, K. R. (2011). Incentives and problem uncertainty in innovation contests: An empirical analysis. *Management Science*, *57*(5), 843-863.

Buettner, R. (2015). A systematic literature review of crowdsourcing research from a human resource management perspective. *Proeceedings of the 48th Hawaii International Conference on System Sciences*.

Burt, C. D., & Strongman, K. (2005). Use of images in charity advertising: Improving donations and compliance rates. *International Journal of Organisational Behaviour*, *8*(8), 571-580.

Cameron, A. C., & Trivedi, P. K. (1998). Regression analysis of count data (Econometric Society Monograph No. 30). Cambridge, UK: Cambridge University Press.

Chiu, C.-M., Liang, T.-P., & Turban, E. (2014). What can crowdsourcing do for decision support? *Decision Support Systems*, *65*(Supplement C), 40-49.

Cialdini, R. B., Schaller, M., Houlihan, D., Arps, K., Fultz, J., & Beaman, A. L. (1987). Empathy-based helping: Is it selflessly or selfishly motivated? *Journal of Personality and Social Psychology*, *52*(4), 749-758.

Coffin, M., & Saltzman, M. J. (2000). Statistical analysis of computational tests of algorithms and heuristics. INFORMS *Journal on Computing*, *12*(1), 24-44.

Crocker, T. (2015). Mining the masses with medical crowdsourcing. *MD News.* Retrieved from http://mdnews.com/mining-masses-medical-crowdsourcing

de Hooge, I. E. (2014). Predicting consumer behavior with two emotion appraisal dimensions: Emotion valence and agency in gift giving. *International Journal of Research in Marketing*, *31*(4), 380-394.

FeldmanHall, O., Dalgleish, T., Evans, D., & Mobbs, D. (2015). Empathic concern drives costly altruism. *Neuroimage, 105*, 347-356.

Fox, S., & Duggan, M. (2013a). For one-third of U.S. adults, the Internet is a diagnostic tool. Retrieved from http://www.pewinternet.org /2013/01/15/information-triage/

Fox, S., & Duggan, M. (2013b). Some seek counsel from fellow patients and caregivers. Retrieved from http://www.pewinternet.org/2013/01/15/ peer-to-peer-health-care/

Goonawardene, N., & Tan, S. (2013). Patients' adherence to health advice on virtual communities: Identity and bond theory perspective. *Proceedings of the 34th International Conference on Information Systems*.

Kennifer, S. L., Alexander, S. C., Pollak, K. I., Jeffreys, A. S., Olsen, M. K., Rodriguez, K. L., … Tulsky, J. A. (2009). Negative emotions in cancer care: Do oncologists' responses depend on severity and type of emotion? *Patient Education and Counseling*, *76*(1), 51-56.

Kordzadeh, N., & Warren, J. (2017). Communicating personal health information in virtual health communities: An integration of privacy calculus model and affective commitment. *Journal of the Association for Information Systems*, 18(1), 45-81.

Lang, A. (1995). What can the heart tell us about thinking? In A. Lang (Ed.), *Measuring psychological responses to media messages* (pp. 99-112). London: Psychology Press.

Lang, A. (2000). The limited capacity model of mediated message processing. *Journal of Communication*, *50*(1), 46-70.

Lazarus, R. S., Kanner, A. D., & Folkman, S. (1980). Emotions: A cognitive-phenomenological analysis. In R. Ptutchik & H. Kellerman (Eds.) *Theories of Emotion* (pp. 189-217). Cambridge, MA: Academic Press.

Malone, T. W., Laubacher, R., & Dellarocas, C. (2009). *Harnessing crowds: Mapping the genome of collective intelligence* (SSRN Scholarly Paper No. ID 1381502). Retrieved from https://papers.ssrn.com/abstract=1381502

Martinez-Espineira, R. (2007). "Adopt a hypothetical pup": A count data approach to the valuation of wildlife. *Environmental and Resource* Economics, 37(2), 335-360.

Meyer, A. N., Longhurst, C. A., & Singh, H. (2016). Crowdsourcing diagnosis for patients with undiagnosed illnesses: An evaluation of CrowdMed. *Journal of Medical Internet Research*, *18*(1), 1-8.

Mozolyako, P. A., & Osipov, N. N. (2015). Virtual prediction markets in medicine. *ERCIM NEWS*. Retrieved from https://ercim-news.ercim.eu /images/stories/EN102/EN102-web.pdf#page=49

O'Leary, D. E. (2016). On the relationship between number of votes and sentiment in crowdsourcing ideas and comments for innovation: A case study of Canada's digital compass. *Decision Support Systems*, *88*(Supplement C), 28-37.

Peng, X., Sun, D., Zhao, Y. C., & Xu, W. (2015). What trigger people use physician-patient interactive OHCs? An empirical research based integration model. *Proceedings of the Pacific Asia Conference on Information Systems*.

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In R. Ptutchik & H. Kellerman (Eds.) *Theories of Emotion* (pp. 3-33). Cambridge, MA: Academic Press.

Emotion: Theory, Research, and Experience (Volume 1: Theories of Emotion) (pp. 3-33). Academic Press.

Prpić, J. (2015). *Health care crowds: Collective intelligence in public health*. Oxford, UK: Center for the Study of Complex Systems.

Savage, S. (2012). Gaining wisdom from crowds. *Communications of the ACM, 55*(3), 13–15.

Schaller, M., & Cialdini, R. B. (1988). The economics of empathic helping: Support for a mood management motive. *Journal of Experimental Social Psychology, 24*(2), 163-181.

Sen, K., & Gosh, K. (2017). Developing effective crowdsourcing systems for medical diagnosis: Challenges and recommendations. *Proceedings of the 50th Hawaii International Conference on System Sciences*.

Shaver, P., Schwartz, J., Kirson, D., & O'Connor, C. (1987). Emotion knowledge: Further exploration of a prototype approach. *Journal of Personality and Social Psychology, 52*(6), 1061-1086.

Smith, C. A., & Pope, L. K. (1992). Appraisal and emption: The interactional contributions of dispositional and situational factors. *Personality and Social Psychology Review*, *14*, 32-62.

Snedecor, G. W., & Cochran, W. G. (1989). *Statistical methods* (8th ed.). Ames, IA: Iowa State University Press.

Toi, M., & Batson, C. D. (1982). More evidence that empathy is a source of altruistic motivation. *Journal of Personality and Social Psychology*, *43*(2), 281-292.

Walter, Thomas P., & Back, A. (2013). A text mining approach to evaluate submissions to crowdsourcing contests. *Proceedings of the 46th Hawaii International Conference on System Sciences*.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*(6), 80-83.

Wilson, B. J. (2007). Designing media messages about health and nutrition: What strategies are most effective? *Journal of Nutrition Education and Behavior*, 39(2), S13-S19.

Yan, L., Tan, Y., Yan, X., & Sun, S. (201p). Shared minds: How patients use collaborative web-based information sharing. *Production and Operations Management*, *28*(1), 9-26.

Yang, Y., Chen, P. Y., & Pavlou, P. (2009). *Open innovation: strategic design of online contests.* Presented at the Workshop on Information Systems and Economics, Phoenix, Arizona. Retrieved from http://pages.stern.nyu.edu/~bakos/wise/papers/wise2009-p16_paper.pdf

Yin, D., Bond, S., & Zhang, H. (2014). Anxious or angry? Effects of discrete emotions on the perceived helpfulness of online reviews. *MIS Quarterly*, *38*(2), 539-560.

Zheng, H., Li, D., & Hou, W. (2011). Task design, motivation, and participation in crowdsourcing contests. *International Journal of Electronic Commerce*, *15*(4), 57-88.

.

Appendix

Table A1. NBR Results with Standardized Coefficients

	Model 1 e^bStdX	**Model 2 e^bStdX**	**Model 3 e^bStdX**	**Model 4 e^bStdX**
Anger	1.0764	1.1113	1.1114	1.1110
Fear	1.0617	1.0738	1.0739	1.0730
Sadness	1.0923	1.0599	1.0600	1.0589
Quality	1.0996	1.0603	1.0603	1.0709
Reward	1.4036	1.4534	1.4531	1.4467
Case duration	1.3204	1.5539	1.5534	1.5536
Symptoms began	1.0279	1.0276	1.0276	1.0292
Description length	1.0000	1.0200	1.0201	1.0140
Notes: e^bStdX = exp(b*SD of X) = change in expected count for SD increase in X				

Table A2. Model Comparison

	BIC	**AIC**
Restricted Model (Controls only)	2471.49	2437.36
Unrestricted Model (Full model)	2284.03	2227.13

Table A3: NBR Results (After Controlling for Case Area)

	Model 1
Anger	0.7608 ***
Fear	0.3000 *
Sadness	0.4880 *
Quality	0.1047 ***
Reward	0.0020 ***
Case Duration	0.0017 ***
Notes: *** p < 0.01, ** p < 0.05, * p < 0.	

Signs and symptoms [edit]

Gastrointestinal [edit]

Many people with Crohn's disease have symptoms for years before the diagnosis.[21] The usual onset is between 15 and 30 years of age, but can occur at any age.[22] Because of the 'patchy' nature of the gastrointestinal disease and the depth of tissue involvement, initial symptoms can be more subtle than those of ulcerative colitis. People with Crohn's disease experience chronic recurring periods of flare-ups and remission.[23]

Abdominal pain may be the initial symptom of Crohn's disease usually in the lower right area.[24] It is often accompanied by diarrhea, especially in those who have had surgery. The diarrhea may or may not be bloody. The nature of the diarrhea in Crohn's disease depends on the part of the small intestine or colon involved. Ileitis typically results in large-volume, watery feces. Colitis may result in a smaller volume of feces of higher frequency. Fecal consistency may range from solid to watery. In severe cases, an individual may have more than 20 bowel movements per day and may need to awaken at night to defecate.[1][25][26][27] Visible bleeding in the feces is less common in Crohn's disease than in ulcerative colitis, but may be seen in the setting of Crohn's colitis.[1] Bloody bowel movements typically come and go, and may be bright or dark red in color. In the setting of severe Crohn's colitis, bleeding may be copious.[25] Flatulence and bloating may also add to the intestinal discomfort.[25]

Signs and symptoms

	Crohn's disease	Ulcerative colitis
Defecation	Often porridge-like,[19] sometimes steatorrhea	Often mucus-like and with blood[19]
Tenesmus	Less common[19]	More common[19]
Fever	Common[19]	Indicates severe disease[19]
Fistulae	Common[20]	Seldom
Weight loss	Often	More seldom

Symptoms caused by intestinal stenosis are also common in Crohn's disease. Abdominal pain is often most severe in areas of the bowel with stenoses. Persistent vomiting and nausea may indicate stenosis from small bowel obstruction or disease involving the stomach, pylorus, or duodenum.[25] Although the association is greater in the context of ulcerative colitis, Crohn's disease may also be associated with primary sclerosing cholangitis, a type of inflammation of the bile ducts.[28]

Perianal discomfort may also be prominent in Crohn's disease. Itchiness or pain around the anus may be suggestive of inflammation, fistulization or abscess around the anal area[1] or anal fissure. Perianal skin tags are also common in Crohn's disease and may appear with or without the presence of colorectal polyps.[29] Fecal incontinence may accompany perianal Crohn's disease. At the opposite end of the gastrointestinal tract, the mouth may be affected by recurrent sores (aphthous ulcers). Rarely, the esophagus, and stomach may be involved in Crohn's disease. These can cause symptoms including difficulty swallowing (dysphagia), upper abdominal pain, and vomiting.[30]

Systemic [edit]

Crohn's disease, like many other chronic, inflammatory diseases, can cause a variety of systemic symptoms.[1] Among children, growth failure is common. Many children are first diagnosed with Crohn's disease based on inability to maintain growth.[31] As it may manifest at the time of the growth spurt in puberty, up to 30% of children with Crohn's disease may have retardation of growth.[32] Fever may also be present, though fevers greater than 38.5 °C (101.3 °F) are uncommon unless there is a complication such as an abscess.[1] Among older individuals, Crohn's disease may manifest as weight loss, usually related to decreased food intake, since individuals with intestinal symptoms from Crohn's disease often feel better when they do not eat and might lose their appetite.[31] People with extensive small intestine disease may also have malabsorption of carbohydrates or lipids, which can further exacerbate weight loss.[33]

Extraintestinal [edit]

In addition to systemic and gastrointestinal involvement, Crohn's disease can affect many other organ systems.[34] Inflammation of the interior portion of the eye, known as uveitis, can cause blurred vision and eye pain, especially when exposed to light (photophobia).[35] Inflammation may also involve the white part of the eye (sclera), a condition called episcleritis.[35] Both episcleritis and uveitis can lead to loss of vision if untreated.

Crohn's disease that affects the ileum may result in an increased risk of gallstones. This is due to a decrease in bile acid resorption in the ileum and the bile gets excreted in the stool. As a result, the cholesterol/bile ratio increases in the gallbladder, resulting in an increased risk for gallstones.[35]

Crohn's disease is associated with a type of rheumatologic disease known as seronegative spondyloarthropathy.[35] This group of diseases is characterized by inflammation of one or more joints (arthritis) or muscle insertions (enthesitis).[35] The arthritis in Crohn's disease can be divided into two types. The first type affects larger weight-bearing joints such as the knee (most common), hips, shoulders, wrists, or elbows.[35] The second type symmetrically involves five or more of the small joints of the hands and feet.[35] The arthritis may also involve the spine, leading to ankylosing spondylitis if the entire spine is involved or simply sacroiliitis if only the sacroiliac joint is involved.[35] The symptoms of arthritis include painful, warm, swollen, stiff joints, and loss of joint mobility or function.[36]

Crohn's disease may also involve the skin, blood, and endocrine system. The most common type of skin manifestation, erythema nodosum, presents as raised, tender red nodules usually appearing on the shins.[35][37] Erythema nodosum is due to inflammation of the underlying subcutaneous tissue, and is characterized by septal panniculitis.[37] Another skin lesion, pyoderma gangrenosum, is typically a painful ulcerating nodule. Crohn's disease also increases the risk of blood clots;[35] painful swelling of the lower legs can be a sign of deep venous thrombosis, while difficulty breathing may be a result of pulmonary embolism. Autoimmune hemolytic anemia, a condition in which the immune system attacks the red blood cells, is also more common in Crohn's disease and may cause fatigue, a pale appearance, and other symptoms common in anemia. Clubbing, a deformity of the ends of the fingers, may also be a result of Crohn's disease. Finally, Crohn's disease increases the risk of osteoporosis, or thinning of the bones.[35] Individuals with osteoporosis are at increased risk of bone fractures.[38]

People with Crohn's disease may develop anemia due to vitamin B_{12}, folate, iron deficiency, or due to anemia of chronic disease.[39][40] The most common is iron deficiency anemia[39] from chronic blood loss, reduced dietary intake, and persistent inflammation leading to increased hepcidin levels, restricting iron absorption in the duodenum.[40] As Crohn's disease most commonly affects the terminal ileum where the vitamin B12/intrinsic factor complex is absorbed, B12 deficiency may be seen.[40] This is particularly common after surgery to remove the ileum.[39] Involvement of the duodenum and jejunum can impair the absorption of many other nutrients including folate. If Crohn's disease affects the stomach, production of intrinsic factor can be reduced.

Crohn's disease can also cause neurological complications (reportedly in up to 15%).[41] The most common of these are seizures, stroke, myopathy, peripheral neuropathy, headache and depression.[41]

People with Crohn's often also have issues with small bowel bacterial overgrowth syndrome, which has similar symptoms.[42]

In the oral cavity people with Crohn's may develop cheilitis granulomatosa and other forms of orofacial granulomatosis, pyostomatitis vegetans, recurrent aphthous stomatitis, geographic tongue, and migratory stomatitis in higher prevalence than the general population.[43]

Figure A4. Sign and Symptoms of Crohn's Disease (https://en.wikipedia.org/wiki/Crohn%27s_disease)

About the Authors

Indika Dissanayake. is an assistant professor in the Department of Information Systems and Supply Chain Management, Bryan School of Business and Economics at the University of North Carolina at Greensboro. She received her PhD in information systems from the University of Texas at Arlington. Her research interests include crowdsourcing, social media, healthcare, and virtual communities. Her research has appeared in journals and conference proceedings such as *Journal of Management Information Systems*, *Information & Management*, *International Conference on Information Systems*, *Americas Conference on Information Systems*, *Decision Science Institute*, and *Hawaii International Conference on System Sciences*.

Sridhar Nerur. is currently a professor of information systems and the Goolsby-Virginia and Paul Dorman Endowed Chair in Leadership at the University of Texas at Arlington. As chair of the Graduate Studies Committee on Business Analytics, he has been actively involved in updating the curriculum to ensure that it is consistent with industry practices. His research has been published in the *MIS Quarterly*, *Strategic Management Journal*, *Communications of the ACM*, *Communications of the Association for Information Systems*, *The DATA BASE for Advances in Information Systems*, *European Journal of Information Systems*, *Information Systems Management*, *Information & Management*, *IEEE Software*, and *Journal of International Business Studies*. He has served as an associate editor of the *European Journal of Information Systems* and was on the editorial board of the *Journal of Association for Information Systems* until December 2016. His research and teaching interests include social networks, machine learning/AI, text analytics, self-organizing systems, and neuroeconomics.

Rahul Singh. is an associate professor in the Department of Information Systems and Supply Chain Management, Bryan School of Business and Economics at The University of North Carolina at Greensboro. He obtained his PhD in business from Virginia Commonwealth University. His research interests include healthcare, health analytics, and business intelligence. His research has been published in leading IS journals including *Management Information Systems Quarterly Executive*, *The European Journal of Information Systems*, *Journal of Association for Information Systems*, *IEEE Transactions on Systems, Man and Cybernetics*, *Information Systems Frontiers*, *Communications of the ACM*, *AIS Transactions on Human-Computer Interaction* , *Information Systems Management*, *eService Journal*, *International Journal of Semantic Web and Information Systems*, *International Journal of Intelligent Information Technologies*, *Information Resources Management Journal*, and *International Journal of Production Engineering Socio-Economic Planning Sciences*.

Yang Lee. is an associate professor, chair of the Supply Chain and Information Management Group, and faculty director of the MS in Business Analytics at D'Amore-McKim School of Business at Northeastern University. She received her PhD from the Massachusetts Institute of Technology. Her publications have appeared in leading journals such as *Communications of the ACM*, *Journal of Management Information Systems*, *Sloan Management Review*, *Management Information Systems Quarterly Executive*. She is co-author of several books, which have been translated into six languages and are used worldwide: *Journey to Data Quality* (MIT Press, 2006), *Quality Information and Knowledge* (Prentice Hall, 1999), and *Data Quality* (Kluwer Academic, 2000). Her research interests include data quality, chief data officers, healthcare information, collaboration, and medical errors IT-mediated institutional learning, and systems integration. She served as editor in chief for the *ACM Journal on Information and Data Quality* and is a co-founder of multiple conferences including the *International Conference on Information Quality*.