NextGen Scientific Review - Annual Perspectives on Next Generation Science – Vol. 1
ISSN 2940-0929

Crowd-Powered Medical Diagnosis

The Potential of Crowdsourcing for Patients with Rare Diseases

Josephine Fischer*, Stefan Arnold, Dilara Yesilbas

Abstract

With the recent rise in medical crowdsourcing platforms, patients with chronic illnesses increasingly broadcast their medical records to obtain an explanation for their complex health conditions. By providing access to a vast pool of diverse medical knowledge, crowdsourcing platforms have the potential to change the way patients receive a medical diagnosis. We developed a conceptual model that details a set of variables. To further the understanding of crowdsourcing as an emerging phenomenon in health care, we provide a contextualization of the various factors that drive participants to exert effort. For this purpose, we used CrowdMed.com as a platform from which we gathered and examined a unique dataset that involves tasks of diagnosing rare medical conditions. By promoting crowdsourcing as a robust and non-discriminatory alternative to seeking help from traditional physicians, we contribute to the acceptance and adoption of crowdsourcing services in health economics.

Keywords: machine learning, graph theory, optimization.

1 Introduction

Rare diseases are an emerging public health issue and thus a challenge for medicine, economics, and society. It is estimated that more than 300 million people worldwide are affected by rare diseases[1]. Most rare diseases substantially reduce life expectancy. Dionisi-Vici et al. (2002), for instance, showed that only 11% of newborn children with inborn errors of metabolism (a form of rare disease) reach adulthood. Apart from the reduced life expectancy, patients with rare diseases suffer from severe impairment of their physical and mental abilities, which limits their educational potential, social opportunities, and economic capabilities (Schieppati et al., 2008).

From a fiscal and socio-economic point of view, the impact of rare diseases is of great interest in health economics (Angelis et al., 2015). Meyer et al. (2016) found that patients with difficult-to-diagnose medical conditions need to consult five physicians before obtaining a diagnosis (incurring a median of $10,000 in medical expenses). Due to diffuse disease patterns, these patients also face diagnostic delays. A period of 5-30 years until a correct diagnosis is not unusual (Meyer et al., 2016). To shorten the time to diagnosis, Meyer et al. (2016) suggest the use of second opinions. Due to the lack of adequate medical advice for the patients, online health communities have become the primary source of health information (Sassenberg & Greving, 2016), and are frequently researched before consulting a physician (Kordzadeh & Warren, 2017; Tan & Goonawardene, 2017). In addition to online healthcare communities, (Dissanayake et al., 2019, p. 1590) mention that medical crowdsourcing platforms "provide emergent solutions to health problems that have long defied diagnosis". Howe (2006) introduced crowdsourcing as a way to obtain needed tasks by soliciting contributions from a crowd. It describes a crowd as an undefined and large network of people of varying knowledge, which collaborate to solve a problem in the form of a flexible open contest (Estellés-Arolas & González-Ladrón-de-Guevara, 2012). Apart from the medical diagnosis, crowdsourcing has been employed to accomplish a variety of healthcare tasks, such as medical transcription (Vashistha et al., 2017), estimation of infection prevalence and propagation (K. Sun et al., 2020), identification of malarial infections (Luengo-Oroz et al., 2012; Mavandadi et al., 2012), categorization of tumors (McKenna et al., 2012; Nguyen et al., 2012), examination of diabetic retinopathy (C. J. Brady et al., 2014), localization of pneumonia in chest radiographs (Pan et al., 2019), and segmentation of intracranial hemorrhage (Sen & Gosh, 2017).

With a few notable exceptions (e.g., Dissanayake et al. (2019)), almost no research is concerned with crowdsourcing for medical diagnosis. We attempt to fill this gap in research by conducting empirical research on medical crowdsourcing cases. From the viewpoint of patients, we formulate our guiding research question as follows: Which factors influence the participation effort in crowd¬sourcing involving medical diagnosis? We collected field data from CrowdMed. CrowdMed is an online platform that allows patients to promote medical cases for a monthly fee between $149 and $749. To the field of health economics, we contribute an empirical investigation on the potential of crowdsourcing for tackling challenges of medical diagnosis, e.g., perceived discrimination in medical settings (Benjamins & Middleton, 2019).

The remainder of this study is organized as follows. Following the introduction, we elaborate on the theoretical foundation of crowdsourcing in healthcare. On this basis, we present and justify our conceptual framework with hypotheses and expectations that emerge from it in section 2. In the next step, we operationalize and transfer our conceptual framework into an estimation model in section 3 and section 4. Based on this, we briefly present the findings of our analysis in section 4. As part of the discussion in section 5, we describe implications for both research and practice. We conclude this study by presenting the lim-

[1] In the United States, the Rare Disease Act of 2002 defines rare diseases as populations of less than 200,000 individuals. With that definition, the number of patients suffering from rare diseases is estimated between 25 and 30 million. In the European Union, however, rare diseases are defined by the European Joint Programme on Rare Diseases as any disease affecting fewer than 5 in 10,000 individuals. With that definition, about 5,000 to 8,000 different rare diseases exist that affect an estimated 27 to 36 million people, or 6-8% of the European population).

itations of this study and potential avenues for follow-up research in section 6.

2 Conceptual Framework

2.1 Summary of Related Work

To highlight the research area, we summarize previous research on crowdsourcing with a particular interest in studies dedicated to healthcare economics. For a comprehensive review of theoretical and empirical research on crowdsourcing for general-purposes, we refer to Hossain and Kauranen (2015) or Segev (2020).

In 2014, Ranard et al. (2014) abstracted peer-reviewed articles to document application scenarios of crowdsourcing in medical research. The authors identified four distinct types of crowd-sourcing tasks in medicine, i.e., problem-solving, data processing, surveillance, and surveying. Ghosh and Sen (2015) examined in their study the role of web-based platforms in promoting the involvement of seekers and solvers in crowdsourcing services for medical diagnosis. Based on existing literature, factors that advance individual participation are developed in the form of a conceptual research model. Later in 2017, Sen and Gosh, in a follow-up study, conceptualize four steps that are necessary to develop an effective crowdsourcing system for medical diagnosis. The authors analyze the existing classification of crowdsourcing and various challenges related to capturing and transferring medical knowledge. As diagnostic suggestions must be discussed from a wide range of medical expertise, Sen and Gosh (2017) recommend involving a multi-disciplinary group of medical experts from around the world. To provide an assessment of medical crowdsourcing platforms, Meyer et al. (2016) collected and evaluated data from CrowdMed. The authors concluded that several patients received helpful hints on their undiagnosed illnesses. Dissanayake et al. (2019) empirically evaluated the participation in medical crowdsourcing on the basis of sentiment analysis. The authors found that cases with higher observed quality and more negative emotions (such as sadness, fear, and anger) yield to more participation. Apart from the empirical analysis, the authors explored ways for selecting the most likely diagnosis from a number of alternative diagnostic suggestions.

Both Yang et al. (2009) and Chen et al. (2014) noticed that crowdsourcing is moderated by the task complexity. Considering that the task of medical diagnosis acts very differently from general-purpose tasks, a contextualization of crowdsourcing for medical diagnosis is suggested.

2.2 Development of Hypotheses

The justification of crowdsourcing for medical diagnosis is predicated on the assumption that a large group with diverse backgrounds is more capable to arrive at a correct diagnosis than a single health practitioner with limited experience in handling certain rare diseases.

To substantiate this assumption, our conceptual framework integrates research issues at the intersection of extreme value theory by Gumbel (1958) and expected value theory by Atkinson (1964). Formally, extreme value theory proceeds from the assumption that each diagnosis can be represented by a random draw, then the chance of getting a correct diagnosis increases as the number of solvers grows (Dahan & Mendelson, 2001). It is documented in the literature that in most crowdsourcing applications at least one of the solvers finds an extreme value solution (e.g., Boudreau et al. (2011)). Note that these extreme values are particularly valuable in situations in which the problem is highly uncertain, such as rare medical conditions. Conversely, expected value theory states that the participation effort is related to the expected value of the payoff and the probability of receiving these payoffs, i.e., when payoff expectations are high, participants are incentivized to exert enthusiasm and participation effort regardless of the financial payoff. From this point of view, many solvers may bring undesired opposing effects to medical cases since each solver exerts a lower equilibrium effort due to the lower expectation of a payout. Boudreau et al. (2011) concluded that the aggregate effect of many solvers depends on whether more diversified diagnoses can mitigate or outweigh the solvers' lower equilibrium effort (which is reflected by less sophisticated medical diagnoses).

The preceding discussion constitutes the conceptual backdrop of our research framework. On the basis of related empirical studies, we developed a conceptual research model that outlines factors affecting the participation effort in crowdsourcing in the context of medical cases. Referring to Yang et al. (2009), we constructed the conceptual model solely with variables that patients may know or control in advance. The factors are grouped into three categories, i.e., (1) *patient-related factors* (e.g., demographic characteristics), (2) *case-related factors*, and (3) *disease-related factors* (e.g., type and count of symptoms). For illustrative purposes, the conceptual model with its effect mechanism on participation effort is illustrated in Figure 1.

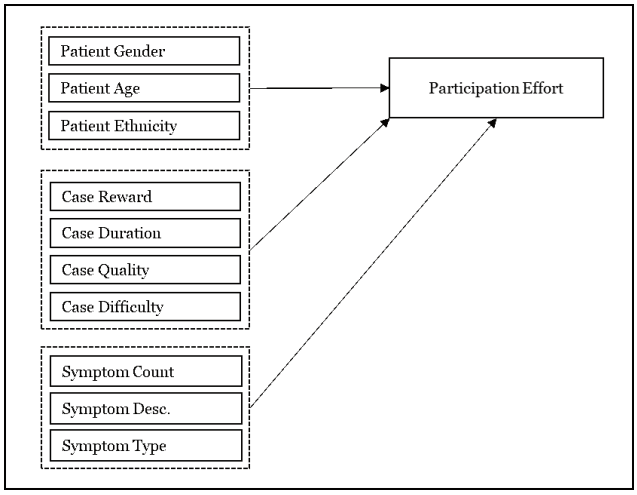

Figure 1: Research model of factors affecting the participation effort in crowdsourcing for medical diagnosis (source: authors own study)

To (1): In the medical field, differences in gender (e.g., Almqvist et al. (2008), K. T. Brady and Randall (1999), and Young et al. (1996)), age (e.g., Klein-Geltink et al. (2006), Rupp et al. (2018), and Zoungas et al. (2014)), and ethnicity (e.g., Corley et al. (2009) and Strakowski et al. (2003)) are widely discussed. Bolin et al. (1968), for instance, identified that the incidence of lactose intolerance is higher in ethnic groups from Asia. From the results of this study can be concluded that many diseases have

a genetic origin (Angelis et al., 2015) and thus the ethnic background of patients can define the boundaries of certain diseases. In contrast to this, prior research indicates that access to healthcare, in general, and the utilization of healthcare services, in particular, varies by sociodemographic characteristics (Casagrande et al., 2007; Hausmann et al., 2008; Kressin et al., 2008; Pascoe & Richman, 2009; Shavers et al., 2012; Sorkin et al., 2010). Sorkin et al. (2010, p. 390), by way of example, identified that "ethnic minorities are more likely to report receipt of lower quality of health care". Casagrande et al. (2007) found that perceived racial discrimination is associated with more delays in medical care and non-adherence to medical care recommendations. Hence, we formulate Hypothesis 1, as follows: *Participation effort is moderated by patient-related characteristics, in particular by the (a) gender, (b) age, and (c) ethnicity of a patient.*

To (2): In 2009, Yang et al. (2009) examined crowdsourcing with a particular interest in contest design. As reward is positively related to participation, it seems that the prospect of economic returns encourages participants to continue to invest time, effort, and resources (Y. Sun et al., 2015). In the context of public prosocial activities, however, Ariely et al. (2009) recognized that high financial incentives are more likely to be counterproductive. Considering the reward as compensation for a participant's efforts, we argue that cases that offer substantial rewards receive more (and potentially better) medical diagnoses. To take this into consideration, our conceptual framework accommodates intrinsic motivation in the form of credits (or experience points) and extrinsic motivation in the form of monetary incentives. In addition to the preceding factor, Yang et al. (2009) found that duration is also positively related to participation. We also expect that long-duration medical cases may yield to more (and potentially better) medical diagnoses. This could be explained by the fact that the solvers can use the time to become familiar with the specific conditions of the patient. Besides that, Chen et al. (2014) explained that difficult cases constitute higher barriers to entry for potential solvers. Because of this, our conceptual framework directly accommodates the perceived quality and difficulty of medical cases. Hence, we formulate Hypothesis 2, as follows: *Participation effort is enhanced by case-related settings, in particular by cases with (a) high rewards, (b) long duration, (c) high quality, and (d) low difficulty.*

To (3): In medical crowdsourcing, patients do not engage in face-to-face conversations with potential solvers; instead, patients broadcast their medical case on the platform. For this purpose, the platform provides a form that can be filled with relevant information, such as "demographics, symptom details, current medications, [...], personal medical history, [...], personal lifestyle", and an explanation of partial diagnostic results from the past (Dissanayake et al., 2019, p. 1594). Since patients decide on the way they present their cases, the formulation of the medical case may affect the involvement of a potential solver. To take this into account, we integrate the description length into the conceptual framework. It is rather straightforward that the more clearly and precisely the medical conditions are described, the less the solvers need to guess which diagnosis is the most appropriate (Chen et al., 2014). From a medical point of view, rare diseases can affect any part of the human system (Schieppati et al., 2008). In other words, certain symptoms occur in many disease patterns and thus many simultaneous symptoms aggravate a clear diagnosis. To reflect the variety of medical conditions, we incorporated the number and type of symptoms into the conceptual framework. Hence, we formulate Hypothesis 3, as follows: *Participation effort is moderated by disease-inherent characteristics, in particular by (a) an ambiguous description of the symptoms. (b) the number of simultaneous symptoms, and (c) the frequency or correlation of symptoms.*

3 Research Methodology

3.1 Data Collection

To verify our conceptual framework, we collected unique data from the web-based platform CrowdMed. On CrowdMed, patients with undiagnosed chronic illness describe their symptoms and provide clinical information hoping to receive a potential diagnosis. For a monthly fee, the case is displayed anonymously on the platform. Solvers can self-select cases to which they would like to contribute a potential diagnosis. Note that the community includes experts (e.g., physicians or nurses) as well as non-medical people (e.g., patients). While the cases are open, patients and solvers engaged in the case can use an open discussion forum to "discuss details online about potential diagnoses, further work-up that should be done, and newly obtained test results and/or appointments completed with the patients' [local] physicians" (Meyer et al., 2016, p. 2). Each participant can suggest a diagnosis. Likewise, each participant can use a peer-flagging mechanism to nominate a poor diagnosis of elimination. When a case is closed, the patient receives a detailed report with a list of diagnoses ranked in decreasing order of likelihood. This calculation is based on weighted voting by solvers. Solvers can improve their rating (and thus their weighting factor) by suggesting a correct diagnosis themselves or by assigning points to a likely diagnosis suggested by other solvers. Considering that only highly rated solvers can participate in complex and well-rewarded cases, solvers are emboldened to take part in the assessment of potential diagnoses. Finally, patients have to decide how to divvy up the financial compensation among the engaged solvers.

CrowdMed has already been used as context in other studies (e.g., Bhattacharyya (2015)) as it provides unique access to the study of medical cases from the field. Unfortunately, CrowdMed does not provide access to well-organized archival data. At the time of data collection, the platform comprised 134 active cases in April 2020. To estimate the influencing factor of our variables, we had to exclude certain observations from these 134 medical cases. CrowdMed offers patients a free trial. These cases are only displayed for one week. As the duration is considered a relevant indicator in our conceptual framework, our study only includes non-trial cases. Following this discussion, we eliminated 9 cases immediately. In 6 cases, patients sought treatment for an already diagnosed disease. These cases have also been removed from the sample. To mitigate the impact of outliers, we analyzed the data using a distance metric proposed by Cook (1977). On this basis, we found 4 cases that differed significantly from all other cases. We assumed that these cases were outliers. Otherwise, we imposed no further restrictions on the dataset. After these adjustments, a total of 115 out of 134 medical cases remained.

3.2 Variables Measurement

In this study, crowdsourcing is examined in terms of participation effort. Given that the participation effort is not directly accessible, we use the number of diagnoses submitted by solvers as a proxy. On the assumption that participation effort can be approximated by the number of diagnoses, participation effort is a count variable that captures the aggregated effort of all solvers in the context of a specific medical case.

Besides that, we collected a multitude of independent variables: As part of the demographic characteristics of the patients, we collected the patient's gender as a binary variable to account for gender-specific differences in medicine. In addition to the gender of the patient, we collected the patient's age as a metric variable and the patient's ethnicity as a nominal variable to indicate whether a patient is of Caucasian, Negroid, or Mongolian descent. As part of the case settings, we collected the reward from the platform as a metric variable. On CrowdMed, reward refers to a monetary and non-monetary compensation a patient offers. Each medical case offers, in addition to a reward appointed in US-dollar, a payout in points which is used to increase the rating of a solver. Both types of rewards are directly available. To integrate both into one variable, we aggregated both rewards. We also collected the duration of each case which is determined by the time a case is open on the platform, as measured in days. Hence, duration is a count variable. Since all members of CrowdMed can rate cases according to their perceived quality and difficulty, we collected the average quality and difficulty of each case. To rate a case, members do not need to participate in it. For the quality measure, the platform uses an ordinal scale from 1 to 5, where 1 indicates poor quality while 5 indicates good quality. The perceived level of difficulty is measured in the same way. As such, both quality and difficulty act as a proxy for the complexity which per se cannot be directly measured or verified. In this way, we economize on using more sophisticated measures for complexity, such as incompletion rate. Without controlling the level of complexity, estimation of variables is biased since the effect of complexity may to some extent be picked up by other variables or by the error term (Chen et al., 2014). As part of the disease characteristics, we collected the description length. The description length is a metric variable. We measured the description length by counting the number of characters used to describe the symptoms. Besides that, we collected the number of symptoms described. The number of symptoms is a count variable ranging from one to twelve and is calculated by aggregating all symptoms a patient has selected from the following categories: eyes or vision, head or neck, breathing, heart or cardiovascular, abdominal or digestion, genital or urinary, abnormal bleeding or bruising, neurological, joint or muscular, mental health, skin or hair, and whole body. In addition to the number of symptoms, we stored the symptoms as a multiple response set.

That being said, we commenced the empirical analysis with a descriptive analysis followed by a correlation analysis. Note that the summary statistics and frequencies are presented in Table 1 and Table 2, respectively; the correlation coefficients are shown in Table 3.

From the descriptive statistics, we can observe that approximately 4 potential diagnoses are suggested, on average. Despite the number of diagnoses fluctuates between 0 and 17, the proportion of cases with zero diagnoses is small in the overall sample. To find these diagnoses, 16 participants are involved per case. Considering the characteristics of the patients, we can see that patients are female with a probability of 64.3% and about 39 years old. This distribution is in line with actual reality. Typically, medical cases are open for participation for about 128 days (which is equivalent to 4.26 months). Considering a monthly fee of at least \$149, 115 patients generate a turnover of \$59,340. Apart from that, the patients show 4.64 symptoms, on average. From the frequency statistics of the multiple response set, we can see clearly that most cases are concerned with the *whole body* (40.9%), followed by symptoms that fall into the categories of *head or neck*, *neurological*, and *abdominal or digestion*. These symptoms represent 17.4%, 14.8%, and 11.3%, respectively, which corresponds to a cumulative value of 84.4%. In addition to these symptoms, patients mentioned *joint or muscular* (6.1%), *breathing* (2.6%), and *heart or cardiovascular* (2.6%). By far the smallest number of cases is concerned with *skin or hair*, *mental health*, and *genital or urinary*, namely 2.7% in combination.

From the correlation results, we can see there is a certain correlation between some variables at a confidence level of $p < .000$. Specifically, the covariate case duration is significantly correlated with the number of solvers and the number of diagnoses following a linear trend of $r = .821$ and $r = .621$, respectively. Likewise, the number of solvers follows a linear relationship with the number of diagnoses according to $r = .749$. For this reason, we tested the extent of multicollinearity between covariates using the variance inflation factor (VIF). VIF was found to be 3.824 and 3.976 for the case duration and the number of solvers, respectively. Craney and Surles (2002, p. 394) mentioned that legitimate cutoff values for the variance inflation factor can be obtained in the range of $[5, 10]$, however, "these cutoff values may be considered extremely lenient in the sense of correlation among the independent variables". Considering that the VIF of the remaining variables was found to be in the range from 1.125 to 1.925, multicollinearity is indicated to be an issue. This being the case, we excluded the number of solvers to make regression analysis feasible.

4 Empirical Analysis and Results

Following the definition and measurement of the variables, we operationalized and transferred our conceptual framework into a generalized linear model. As the response variable in our study is a count measure, we used a Poisson model. All calculations to form an estimate of the participation effort were carried out using IBM SPSS Statistics 25. Our estimation model is in Equation 1. Note that we denote the random error by ξ.

$$\begin{aligned} y = \beta_0 &+ \beta_1 gender + \beta_2 age + \beta_3 ethnicity + \beta_4 reward \\ &+ \beta_5 duration + \beta_6 quality + \beta_7 difficulty \\ &+ \beta_8 desc\_of\_symptoms + \beta_9 num\_of\_symptoms \\ &+ \beta_{10} type\_of\_symptom + \xi \end{aligned} \quad (1)$$

To deal with missing values, we have pre-processed the measurement of the quality and difficulty of a case following the approach put forward by (Dissanayake et al., 2019). Since these variables only had values for 28 and 25 cases, respectively, eliminating all missing cases was not an option.

Table 1: Descriptive Statistics. Sample size $n = 115$, valid sample size $n = 23$. The low number of complete cases is reasoned by the case quality and case difficulty, which are only evaluated by the case solvers in $n = 28$ and $n = 25$ cases, respectively. (Source: Own work.)

Variables	**Min.**	**Max.**	**Mean**	**Std. Dev.**	**Skewness**	**Kurtosis**
Patient's Age	2	80	39.87	14.564	.097	-.083
Reward in Points	3,000	21,000	6,339.13	4,499.292	2.487	5.288
Reward in Dollar	200	2,000	352.15	321.450	3.139	11.609
Case Quality	1	5	4.07	1.245	-.888	-.496
Case Difficulty	1	4	2.76	.831	-.453	.035
Case Duration	9	390	128.36	82.111	1.517	2.047
Symptom Description	53	14,363	2,459.35	2,323.086	1.960	6.105
Symptom Count	1	12	4.64	3.288	.741	-.620
Participants Count	2	43	16.66	7.859	1.097	1.688
Diagnosis Count	0	7	4.23	3.518	1.507	2.458

Table 2: Frequency Statistics. Single Response represents a patients' main symptom, $n = 115$ (100%). Multiple Response Set is calculated using dichotomy groups of a patients' additional symptoms tabulated at value 1, $n = 534$ (464.3%). (Source: Own work.)

Symptom Type	**Single Response**			**Multiple Response Set**		
	N	Percent	Cum. Percent	N	Percent	Percent of Cases
Eyes or vision	2	1.7	1.7	37	6.9	32.2
Head or neck	20	17.4	19.1	66	12.4	57.4
Breathing	3	2.6	21.7	27	5.1	23.5
Heart or cardiovascular	3	2.6	24.3	38	7.1	33.0
Abdominal or digestion	13	11.3	35.7	55	10.3	47.8
Genital or urinary	1	.9	36.5	33	6.2	28.7
Bleeding or bruising	0	0	36.5	17	3.2	14.8
Neurological	17	14.8	51.3	58	10.9	50.4
Joint or muscular	7	6.1	57.4	56	10.5	48.7
Mental health	1	.9	58.3	41	7.7	35.7
Skin or hair	1	.9	59.1	42	7.9	36.5
Whole body	47	40.9	100.0	64	12.0	55.7

Instead, we estimated missing values regressing both variables on gender, reward, duration, and the number of symptoms. Apart from that, we encoded nominal variables into dichotomous variables. No further transformations were applied to the variables. The results of the estimation are reported in Table 4. We used the likelihood ratio chi-square as a general basis of assessment for the estimation model, which is recommended for small samples. The estimation model seems reasonable ($\chi^2 = 155.328$) and statistically significant ($p = .000$).

From the regression results, it can be clearly seen that the significance of the estimation coefficients ranges from $p < .000$ to $p < .900$. We present the results of estimation hierarchically with standard errors enclosed in parentheses. The coefficient of $\beta_1 = -.175$ ($p = .157$) is negative and insignificant. This means that participation effort is not moderated by the patients' gender leading to the disconfirmation of Hypothesis 1 (a). Hypothesis 1 (b) predicted that a patients' age is negatively associated with the participation effort. As $\beta_2 = -.012$ ($p = .003$) is negative and statistically significant, we can support Hypothesis 1 (b). All coefficients regarding the patients' ethnicity $\beta_3 = \{-.301, -.261\}$ are insignificant according to $p = .138$, and $p = .224$, respectively. We, therefore, reject Hypothesis 1 (c). Contrary to our expectations, $\beta_4 = -.000$ ($p = .005$) is negative (although this is due to rounding to three decimal places). Since the p-value is .005, the coefficient is statistically significant at a confidence level of 0.01. Following on from this argument, our data suggest that case reward is negatively associated with the number of diagnoses. Hence, Hypothesis 2 (a) cannot be confirmed. In accord with Hypothesis 2 (b), the coefficient of the case duration $\beta_5 = .005$ ($p = .000$) shows a positive and significant relationship. The results suggest that cases with longer durations lead to more participation effort in terms of submitted diagnoses. Thus, Hypothesis 2 (b) is supported with confidence. It can be seen that the coefficients of perceived quality $\beta_6 = -.060$ ($p = .106$) and difficulty $\beta_7 = -.008$ ($p = .882$) are both insignificant. Thus, Hypotheses 2 (c) and (d) are not supported. Consistent with Hypothesis 3 (a), $\beta_8 = .000$ ($p = .001$) is significant and positive (although this is again not evident due to rounding to three decimal places). In line with the actual reality, diffuse diseases with multiple symptoms make it difficult to form a diagnosis and treatment plan. We further assume that the effect of perceived difficulty is picked up to a certain extent by the number of symptoms. Since this effect is marginal, we proceed on the assumption that the community can deal with complex health conditions. Since $\beta_9 = .036$ ($p = .036$) is positive and significant, we support

Table 3: Correlation Matrix. Sample size $n = 115$. Correlation with two-tailed confidence levels (* means correlation significant at 0.05, ** means correlation significant at 0.01 (Source: Own work.)

Variables	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**	**(10)**	**(11)**	**(12)**
P. Gender	1											
P. Age	-.101	1										
P. Ethnicity	-.098	-.024	1									
C. Reward	.345**	-.039	-.016	1								
C. Duration	-.036	-.047	.049	-.145	1							
C. Quality	.356	-.252	-.441*	.195	-.208	1						
C. Difficulty	-.283	-.129	.017	.022	-.156	.321	1					
S. Description	.050	-.015	.078	.260**	.125	.202	.402	1				
S. Number	-.207*	-.089	-.019	-.075	-.074	.590**	.342	.072	1			
S. Type	-.129	-.115	.018	-.058	-.013	.008	.241	.065	.276**	1		
Participants	-.100	-.135	.043	-.274**	.821**	-.259	-.040	.099	-.003	-.020	1	
Diagnoses	.028	-.213	.096	-.175	.621**	-.108	-.014	.232*	.068	-.050	.749**	1

Hypothesis 3 (b). It seems that a clear description of the symptoms helps to find a diagnosis. Finally, Hypothesis 3 (c) posited that the type of symptom is associated with participation effort. The p-values of β_{10} range from .014 to .900. This indicates that certain symptoms significantly affect the number of diagnoses. Using the whole body as the reference category for the main affected area of the body, we can see that cases concerning heart, cardiovascular, abdominal, and digestive symptoms increase the number of diagnoses, significantly. Since some symptoms are present in only a few cases, their insignificance is probably due to the small sample size. With this in mind, we partially confirm Hypothesis 3 (c).

To identify and qualify the influence of patients' demographic characteristics, we conducted an analysis of variance with a significance threshold denoted by $\alpha = .05$. The F-values and p-values are presented in Table 5. It can be seen clearly that for all characteristics the difference in means falls short of being significant. As $p \gg .05$, it appears that the participation effort of solvers is non-discriminatory with regard to the sex, age, and ethnicity of a patient (suggesting the crowd-powered medical diagnosis as a viable alternative for people perceiving discrimination by traditional medical institutions).

5 Discussion

This study developed a conceptual framework to examine factors that encourage participants to exert effort in crowdsourcing involving medical diagnosis. To verify our conceptual framework, we collected a unique dataset from https://www.crowdmed.com. After we operationalized our frameworks into an estimation model, we interpreted a set of variables concerning the direction, intensity, and significance. Most findings fall in with observations demonstrated previously by Dissanayake et al. (2019) on the basis of an empirical analysis of participation. As theorized, the participation effort measured by the number of potential diagnoses is highly correlated with the number of participants, however, some regularities for the participation effort can be derived that might warrant further studies. The results highlight that the number of diagnoses is contingent on the design of a medical case and its difficulty. To determine the difficulty level of a medical case, the number of symptoms seems to be an adequate indicator. Apart from that, we did not find any ethnic discrimination in terms of diagnostic suggestions received from patients belonging to an ethnic minority group; therefore, we can recommend crowdsourcing to all those who perceive discrimination in traditional medical diagnosis. Finally, participants are intrinsically motivated by the opportunity to gain reputation, which is indicated by the fact that the number of diagnoses is positively correlated to the number of points accredited to a correct diagnosis. This correlation does not exist for the financial reward. Since the coefficients of these factors are fairly small, we conclude that the engagement of participants is robust to changes in settings of case design.

5.1 Theoretical and Practical Implications

Despite the growing interest in web-based tools for healthcare, almost no empirical research is concerned with crowdsourcing for medical cases. Our results complement the findings of prior literature by presenting determinants of participation effort in the context of medical diagnosis. We demonstrated that participants are more engaged by cases submitted by patients of young age that are afflicted with certain symptoms for a protracted period. We further showed that the perceived discrimination in medical settings (see, e.g., Benjamins and Middleton (2019)) is not prevalent in medical crowdsourcing. By doing this, our study provides a benchmark that may serve as a basis of comparison to other medical tasks, settings, and platforms. In supplement to the theoretical implications, we also provide important implications for practice. By identifying the determinants of participation effort in medical crowdsourcing, we offer guidance to patients on how to design their cases to increase the likelihood of resolving their undiagnosed disease.

5.2 Research Limitations and Extensions

This study has several shortcomings. First, the sample is only subject to the diagnosis of rare diseases. Whether the research results can be generalized to the diagnosis of undi-

Table 4: Regression Results. Sample size $n = 115$. (Source: Own work.)

Variables		**Coeff.**	**β-Value**	**Std. Error**	**Wald χ^2**	**p-Value**	**VIF**
Constant		β_0	1.480	.3232	20.976	.000	.
Patient's Gender	*Male*	β_1	.000	.	.	.	.
	Female		-.175	.1235	2.001	.157	1.759
Patients' Age		β_2	-.012	.0041	8.876	.003	1.442
	Caucasoid		-.261	.1759	2.205	.138	1.759
Patients Ethnicity	Negroid	β_3	-.301	.2480	1.478	.224	1.845
	Mongoloid		.000	.	.	.	.
Case Reward		β_4	-.000	.0000	7.869	.005	1.527
Case Duration		β_5	.005	.0005	88.340	.000	1.301
Case Quality		β_6	-.060	.0373	2.614	.106	1.552
Case Difficulty		β_7	-.008	.0537	.022	.882	1.857
Symptom Count		β_8	.000	.0000	10.945	.001	1.295
Symptom Description		β_9	.036	.0173	4.385	.036	1.434
	Eyes or vision		.125	.3290	.145	.703	1.282
	Head or neck		.262	.1603	2.676	.102	1.642
	Breathing		.131	.2691	.237	.626	1.261
	Heart or cardiovascular		.680	.3079	4.878	.027	1.426
	Abdominal or digestion		.414	.1676	6.099	.014	1.736
Symptom Type	Genital or urinary	β_{10}	-.060	.4754	.016	.900	1.105
	Neurological		.140	.1613	.749	.387	1.867
	Joint or muscular		-.140	.2094	.450	.502	1.569
	Mental health		-.684	1.0262	.444	.505	1.169
	Skin or hair		-.400	.4779	.701	.402	1.210
	Whole body		.000	.	.	.	.

Table 5: Variance Analysis. Sample size $n = 115$. (Source: Own work.)

Variables	**# Participants**		**# Diagnoses**	
	F	p	F	p
Eyes or vision	2	1.7	1.7	37
Head or neck	20	17.4	19.1	66
Breathing	3	2.6	21.7	27

agnosed common diseases needs further investigation. For the medical diagnosis of common diseases, 115 cases may not be an accurate representation. Second, our conceptual framework was tested using aggregated data that is publicly available on the crowdsourcing platform of interest. Prior research demonstrated that the willingness to participate in medical cases is in a large part stimulated by the solvers' intrinsic motivation (Zheng et al., 2011) and the patients' emotional tones (Dissanayake et al., 2019). Third, our study assumes that the medical report, which is eventually accepted by the patient contains the correct diagnosis; however, in reality, it may not include the most accurate diagnosis. To figure this out, patients need consultations with their physician(s). Since it may take years before the first signs of alleviation can be seen, we plan to enrich the validity of our study by interviewing solvers on whether the diagnostic suggestion has turned out to be correct. Such a long-term validation of crowdsourcing for medical diagnosis has already been mentioned in Meyer et al. (2016).

6 Conclusion

Crowdsourcing has the potential to radically change the way patients receive a medical diagnosis or treatment plan. Patients afflicted with chronic, difficult-to-diagnose diseases are already using crowdsourcing as a viable alternative to seeking help from traditional physicians. With no diagnosis or treatment forthcoming, these patients are willing to expend time and money to obtain a potential cure from an unknown. CrowdMed holds considerable promise to find the reason for the ailments of patients with rare diseases. As part of an empirical analysis, we investigated factors affecting the participation effort measured by the number of submitted diagnoses. It seems that the participation effort is fairly robust and non-discriminatory. With this in mind, our study makes an important contribution to the acceptance and adoption of web-based services in healthcare.

References

Almqvist, C., Worm, M., Leynaert, B., & working group of GA2LEN WP 2.5 'Gender'. (2008). Impact of gender on asthma in childhood and adolescence: A ga2len review. *Allergy*, *63*(1), 47–57. https://doi.org/10.1111/j.1398-9995.2007.01524.x

Angelis, A., Tordrup, D., & Kanavos, P. (2015). Socio-economic burden of rare diseases: A systematic review of cost of illness evidence. *Health policy (Amsterdam, Netherlands)*, *119*(7), 964–979. https://doi.org/10.1016/j.healthpol.2014.12.016

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.

Atkinson, J. W. (1964). *An Introduction to Motivation* (D. C. McClelland, Ed.). D. van Nostrand Company, Inc., Princeton, New Jersey.

Benjamins, M. R., & Middleton, M. (2019). Perceived discrimination in medical settings and perceived quality of care: A population-based study in chicago. *PloS one*, *14*(4), e0215976. https://doi.org/10.1371/journal.pone.0215976

Bhattacharyya, M. (2015). *Studying the reality of crowd-powered healthcare* [Presented as Poster at AAAI HCOMP, San Diego, USA]. Retrieved November 23, 2022, from https://www.humancomputation.com/2015/papers/32_Paper.pdf

Bolin, T. D., Crane, G. G., & Davis, A. E. (1968). Lactose intolerance in various ethnic groups in south–east asia. *Australasian annals of medicine*, *17*(4), 300–306.

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.

Brady, C. J., Villanti, A. C., Pearson, J. L., Kirchner, T. R., Gupta, O. P., & Shah, C. P. (2014). Rapid grading of fundus photographs for diabetic retinopathy using crowdsourcing. *Journal of Medical Internet Research*, *16*(10), e233. https://doi.org/10.2196/jmir.3807

Brady, K. T., & Randall, C. L. (1999). Gender differences in substance use disorders. *Psychiatric Clinics of North America*, *22*(2), 241–252.

Casagrande, S. S., Gary, T. L., LaVeist, T. A., Gaskin, D. J., & Cooper, L. A. (2007). Perceived discrimination and adherence to medical care in a racially integrated community. *Journal of General Internal Medicine*, *22*(3), 389–395.

Chen, P.-Y., Pavlou, P. A., & Yang, Y. (2014). Determinants of open contest participation in online labor markets. *Fox School of Business Research Paper*, (15-074). https://doi.org/10.2139/ssrn.2510114

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, *19*(1), 15–18. https://doi.org/10.2307/1268249

Corley, D. A., Kubo, A., Levin, T. R., Block, G., Habel, L., Rumore, G., Quesenberry, C., & Buffler, P. (2009). Race, ethnicity, sex and temporal differences in barrett's oesophagus diagnosis: A large community-based study, 1994–2006. *Gut*, *58*(2), 182–188. https://doi.org/10.1136/gut.2008.163360

Craney, T. A., & Surles, J. G. (2002). Model-dependent variance inflation factor cutoff values. *Quality Engineering*, *14*(3), 391–403. https://doi.org/10.1081/QEN-120001878

Dahan, E., & Mendelson, H. (2001). An extreme-value model of concept testing. *Management Science*, *47*(1), 102–116. https://doi.org/10.1287/mnsc.47.1.102.10666

Dionisi-Vici, C., Rizzo, C., Burlina, A. B., Caruso, U., Sabetta, G., Uziel, G., & Abeni, D. (2002). Inborn errors of metabolism in the italian pediatric population: A national retrospective survey. *The Journal of Pediatrics*, *140*(3), 321–329. https://doi.org/10.1067/mpd.2002.122394

Dissanayake, I., Nerur, S., Singh, R., & Lee, Y. (2019). Medical crowdsourcing:harnessing the "wisdom of the crowd" to solve medical mysterie. *Journal of the Association for Information Systems*, *20*(11), 1589–1610. https://doi.org/10.17705/1jais.00579

Estellés-Arolas, E., & González-Ladrón-de-Guevara, F. (2012). Towards an integrated crowdsourcing definition. *Journal of Information Science*, *38*(2), 189–200. https://doi.org/10.1177/0165551512437638

Ghosh, K., & Sen, K. (2015). A conceptual model to understand the factors that drive individual participation in crowdsourcing for medical diagnosis. In T. X. Bui & R. H. Sprague (Eds.), *48th Hawaii International Conference on System Sciences (HICSS), 2015* (pp. 2815–2823). IEEE. https://doi.org/10.1109/HICSS.2015.341

Gumbel, E. J. (1958). *Statistics of extremes*. Columbia University Press. https://doi.org/10.7312/gumb92958

Hausmann, L. R. M., Jeong, K., Bost, J. E., & Ibrahim, S. A. (2008). Perceived discrimination in health care and health status in a racially diverse sample. *Medical Care*, *46*(9), 905–914. https://doi.org/10.1097/MLR.0b013e3181792562

Hossain, M., & Kauranen, I. (2015). Crowdsourcing: A comprehensive literature review. *Strategic Outsourcing: An International Journal*, *8*(1), 2–22. https://doi.org/10.1108/SO-12-2014-0029

Howe, J. (2006). The rise of crowdsourcing. *Wired Magazine*, *14*(6), 1–4.

Klein-Geltink, J., Pogany, L., Mery, L. S., Barr, R. D., & Greenberg, M. L. (2006). Impact of age and diagnosis on waiting times between important healthcare events among children 0 to 19 years cared for in pediatric units: The canadian childhood cancer surveillance and control program. *Journal of Pediatric Hematology/Oncology*, *28*(7), 433–439. https://doi.org/10.1097/01.mph.0000212945.20480.26

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. https://doi.org/10.17705/1jais.00446

Kressin, N. R., Raymond, K. L., & Manze, M. (2008). Perceptions of race/ethnicity-based discrimination: A review of measures and evaluation of their usefulness for the health care setting. *Journal of Health Care for the Poor and Underserved*, *19*(3), 697–730. https://doi.org/10.1353/hpu.0.0041

Luengo-Oroz, M. A., Arranz, A., & Frean, J. (2012). Crowdsourcing malaria parasite quantification: An online game for analyzing images of infected thick blood smears. *Journal of Medical Internet Research*, *14*(6), e167. https://doi.org/10.2196/jmir.2338

Mavandadi, S., Dimitrov, S., Feng, S., Yu, F., Sikora, U., Yaglidere, O., Padmanabhan, S., Nielsen, K., & Ozcan, A. (2012). Distributed medical image analysis and diagnosis through crowd-sourced games:

A malaria case study. *PloS one*, *7*(5), e37245. https://doi.org/10.1371/journal.pone.0037245

McKenna, M. T., Wang, S., Nguyen, T. B., Burns, J. E., Petrick, N., & Summers, R. M. (2012). Strategies for improved interpretation of computer-aided detections for ct colonography utilizing distributed human intelligence. *Medical Image Analysis*, *16*(6), 1280–1292. https://doi.org/10.1016/j.media.2012.04.007

Meyer, A. N. D., Longhurst, C. A., & Singh, H. (2016). Crowdsourcing diagnosis for patients with undiagnosed illnesses: An evaluation of crowdmed. *Journal of Medical Internet Research*, *18*, 1–8. https://doi.org/10.2196/jmir.4887

Nguyen, T. B., Wang, S., Anugu, V., Rose, N., McKenna, M., Petrick, N., Burns, J. E., & Summers, R. M. (2012). Distributed human intelligence for colonic polyp classification in computer-aided detection for ct colonography. *Radiology*, *262*(3), 824–833. https://doi.org/10.1148/radiol.11110938

Pan, I., Cadrin-Chênevert, A., & Cheng, P. M. (2019). Tackling the radiological society of north america pneumonia detection challenge. *American Journal of Roentgenology*, *213*(3), 568–574. https://doi.org/10.2214/AJR.19.21512

Pascoe, E. A., & Richman, L. S. (2009). Perceived discrimination and health: A meta-analytic review. *Psychological Bulletin*, *135*(4), 531–554. https://doi.org/10.1037/a0016059

Ranard, B. L., Ha, Y. P., Meisel, Z. F., Asch, D. A., Hill, S. S., Becker, L. B., Seymour, A. K., & Merchant, R. M. (2014). Crowdsourcing–harnessing the masses to advance health and medicine, a systematic review. *Journal of General Internal Medicine*, *29*(1), 187–203. https://doi.org/10.1007/s11606-013-2536-8

Rupp, C., Rössler, A., Zhou, T., Rauber, C., Friedrich, K., Wannhoff, A., Weiss, K.-H., Sauer, P., Schirmacher, P., & Süsal, C. (2018). Impact of age at diagnosis on disease progression in patients with primary sclerosing cholangitis. *United European Gastroenterology Journal*, *6*(2), 255–262.

Sassenberg, K., & Greving, H. (2016). Internet searching about disease elicits a positive perception of own health when severity of illness is high: A longitudinal questionnaire study. *Journal of Medical Internet Research*, *18*(3), e56. https://doi.org/10.2196/jmir.5140

Schieppati, A., Henter, J.-I., Daina, E., & Aperia, A. (2008). Why rare diseases are an important medical and social issue. *The Lancet*, *371*(9629), 2039–2041. https://doi.org/10.1016/S0140-6736(08)60872-7

Segev, E. (2020). Crowdsourcing contests. *European Journal of Operational Research*, *281*(2), 241–255. https://doi.org/10.1016/j.ejor.2019.02.057

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*, 3289–3296. https://doi.org/10.24251/HICSS.2017.398

Shavers, V. L., Fagan, P., Jones, D., Klein, W. M. P., Boyington, J., Moten, C., & Rorie, E. (2012). The state of research on racial/ethnic discrimination in the receipt of health care. *American Journal of Public Health*, *102*(5), 953–966. https://doi.org/10.2105/AJPH.2012.300773

Sorkin, D. H., Ngo-Metzger, Q., & de Alba, I. (2010). Racial/ethnic discrimination in health care: Impact on perceived quality of care. *Journal of General Internal Medicine*, *25*(5), 390–396. https://doi.org/10.1007/s11606-010-1257-5

Strakowski, S. M., Keck, P. E., Arnold, L. M., Collins, J., Wilson, R. M., Fleck, D. E., Corey, K. B., Amicone, J., & Adebimpe, V. R. (2003). Ethnicity and diagnosis in patients with affective disorders. *The Journal of Clinical Psychiatry*, *64*(7), 747–754. https://doi.org/10.4088/jcp.v64n0702

Sun, K., Chen, J., & Viboud, C. (2020). Early epidemiological analysis of the coronavirus disease 2019 outbreak based on crowdsourced data: A population-level observational study. *The Lancet Digital Health*, *2*(4), e201–e208. https://doi.org/10.1016/S2589-7500(20)30026-1

Sun, Y., Wang, N., Yin, C., & Zhang, J. X. (2015). Understanding the relationships between motivators and effort in crowdsourcing marketplaces: A nonlinear analysis. *International Journal of Information Management*, *35*(3), 267–276. https://doi.org/10.1016/j.ijinfomgt.2015.01.009

Tan, S. S.-L., & Goonawardene, N. (2017). Internet health information seeking and the patient-physician relationship: A systematic review. *Journal of Medical Internet Research*, *19*(1), e9. https://doi.org/10.2196/jmir.5729

Vashistha, A., Sethi, P., & Anderson, R. (2017). Respeak: A voice-based, crowd-powered speech transcription system. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 1855–1866. https://doi.org/10.1145/3025453.3025640

Yang, Y., Chen, P.-y., & Pavlou, P. (2009). Open innovation: An empirical study of online contests. *ICIS 2009 Proceedings - Thirtieth International Conference on Information Systems*, 13.

Young, T., Hutton, R., Finn, L., Badr, S., & Palta, M. (1996). The gender bias in sleep apnea diagnosis: Are women missed because they have different symptoms? *Archives of Internal Medicine*, *156*(21), 2445–2451.

Zheng, H., Li, D., & Hou, W. (2011). Task design, motivation, and participation in crowdsourcing contests. *International Journal of Electronic Commerce*, *15*(4), 57–88. https://doi.org/10.2753/JEC1086-4415150402

Zoungas, S., Woodward, M., Li, Q., Cooper, M. E., Hamet, P., Harrap, S., Heller, S., Marre, M., Patel, A., & Poulter, N. (2014). Impact of age, age at diagnosis and duration of diabetes on the risk of macrovascular and microvascular complications and death in type 2 diabetes. *Diabetologia*, *57*(12), 2465–2474. https://doi.org/10.1007/s00125-014-3369-7