Is there a Doctor in the Crowd? Diagnosis Needed! (for less than $5)

James Cheng, University of Texas at Austin
Monisha Manoharan, University of Texas at Austin
Matthew Lease, University of Texas at Austin
Yan Zhang, University of Texas at Austin

Abstract

We investigate the feasibility of crowd-based medical diagnosis by posting medical cases on a variety of crowdsourcing platforms: general and specialized volunteer question answering sites, and pay-based Mechanical Turk (MTurk) and oDesk. To assess the crowd's ability to diagnose cases of varying difficulty, three sets of medical cases are considered. While volunteer channels proved ineffective for us, we discuss design limitations and opportunities for improvement. In contrast, Mechanical Turk workers without medical training not only correctly diagnosed easy cases, but also a previously unsolved case from *CrowdMed* involving extensive patient details. Likely due to varying expertise, MTurk workers and oDesk health professionals also differed in their willingness to provide uncertain diagnoses, diagnosis rationales, and reliance on personal experience with a disease to diagnose it.

Keywords: crowdsourcing; human computation; Mechanical Turk; medical and health support

Citation: Cheng, J., Manoharan, M., Zhang, Y., Lease, M. (2015). Is there a Doctor in the Crowd? Diagnosis Needed! (for less than $5). In *iConference 2015 Proceedings*.

Research Data: Available online at: https://www.ischool.utexas.edu/~ml/data

Contact: Matthew Lease <ml@utexas.edu>

1 Introduction

Medical diagnosis is a central component of physician competence, with mastery as a major objective of medical education (Elstein, 1978; Norman, 2005). It is widely recognized that diagnosing is highly individual, dependent on physicians' perceived difficulty of a particular case and their relevant domain knowledge (Elstein, 2002). As a result, health consumers in practice often choose to seek additional opinions from other providers.

Health 2.0 (or *participatory health*) is a growing movement toward use of Internet technologies for personalized health-care (Swan, 2012). Specialized social networks for consumer health have become particularly popular, allowing people to ask questions about their health problems and receive feedback from others with similar symptoms or medical specialists. As a result, the pool from which consumers can seek "second opinions" has been drastically expanded. At present, 35% of adults in the U.S.A. have gone online to self-diagnose a medical condition for themselves or for someone else. Among those seeking online diagnoses, 35% did not visit a clinician to get professional opinions, 41% received confirmations from their doctors, and 19% reported disagreements from or inconclusive conversations with their doctors (Fox & Duggan, 2013).

With the advent of crowdsourcing and human computation (Quinn & Bederson, 2011), a landscape of new and innovative approaches to healthcare are emerging to tackle open problems via crowd-based Internet services. For example, community question answering (CQA), social networks, and specialized medical sites, such as *PatientsLikeMe*, offer new ways to poll the social "village" or *friendsource* one's social network (Horowitz & Kamvar, 2010; Oeldorf-Hirsch, Hecht, Morris, Teevan, & Gergle, 2014; Rzeszotarski & Morris, 2014). Prediction markets (Quinn & Bederson, 2011) such as the *Iowa Electronic Markets* aggregate crowd knowledge to make collective predictions. A study on use of Facebook to diagnose illnesses, the correct diagnosis was found for 5 of 6 cases within 10 minutes (Folkestad, Brodersen, Hallas, & Brabrand, 2011).

A recent startup CrowdMed.com offers an interesting prediction market approach to medical diagnosis. People seeking a diagnosis may fill out a questionnaire detailing their symptoms, history and background for analysis by CrowdMed's "Medical Detectives", who suggest and bet on diagnoses. CrowdMed's proprietary algorithm uses these bets to estimate a probability for each diagnosis. While we are not familiar with any clinical evaluations of CrowdMed's effectiveness, the New Scientist reported that, "In 20 initial test cases, around 700 participants identified each of the mystery diseases as one of their top three suggestions" (Nuwer, 2013). CrowdMed reports finishing over 100 cases with < 10% of patients reporting inaccurate diagnosis.

While such success is seemingly achieved through specialized, proprietary methods, is such complexity truly needed for accurate diagnosis, or might simpler methods suffice in some cases? In fact, due to the above

trade secrets, there is little transparency about the methodology used, its complexity, or whenever it changes, limiting scientific understanding. In contrast, could standard, general purpose crowdsourcing services be used to also correctly diagnose some range of medical cases, and perhaps even more easily or cheaply? While micro-tasking approaches have already been shown effective on tasks such as identifying malaria infected red blood cells (Mavandadi et al., 2012) and detecting melanomas (King, Gehl, Grossman, & Jensen, 2013), a more-general crowdsourcing approach to accurately diagnose a wider range of illness may offer a faster or less expensive alternative, preceding or complementing traditional medical diagnosis.

We investigate the following key research questions in this study:

1. Can we obtain correct diagnoses from general-purpose crowdsourcing platforms, or is it necessary to use healthcare specific Websites and/or professional providers?
2. Can medical cases of varying difficulty be correctly diagnosed online (e.g., rarer or more serious illnesses)?
3. How do laymen vs. contractors differ from each other in qualitative nature and accuracy of diagnosis?

To assess feasibility of diagnosing medical cases of varying difficulty, we include three sets in our evaluation: easy and medium difficulty cases adapted from Folkestad et al. (2011), and more difficult cases taken from CrowdMed's site. We similarly investigate the feasibility of crowd-based medical diagnosis for a variety of platforms: general and specialized volunteer CQA sites, and pay-based Mechanical Turk (MTurk) and oDesk. Because MTurk caters to pseudonymous, low-skilled, and low-paying work, its workforce is heavily-skewed toward laymen rather than experts (Ross, Irani, Silberman, Zaldivar, & Tomlinson, 2010; Ipeirotis, 2010). In contrast, oDesk's workforce of known-identity, skilled contractors enabled us to recruit health-service professionals. Given cost disparities between platforms, we aggregated opinions from 10 MTurk workers and 2 oDesk contractors.

While our experiments with volunteer platforms were largely unsuccessful, we describe our experiences and alternative designs to consider in future work. In contrast, our designs for paid platforms fared better. For easy cases, correct diagnoses were obtained for all questions on both MTurk and oDesk. On medium difficulty cases, however, MTurk (with a two-stage design) yielded largely correct diagnoses, while oDesk did not. For the most difficult cases, MTurk workers remarkably converged on the correct diagnosis for one of the medical cases presented. In a seeming "wisdom of crowds" effect, the ten MTurk workers collectively outperformed our two health professionals from oDesk on both medium and difficult cases. A more nuanced observation we discuss was the qualitative nature of how the two groups differed in their diagnoses. Our analysis of MTurk worker diagnosis rationales identifies several key strategies they employ to diagnose, while in contrast, oDesk professionals both rely much less on personal experience and express less confidence in their willingness to provide uncertain diagnoses and the amount of detail and feedback provided. Overall, we show an exciting potential for crowd-based diagnosis meriting further investigation.

2 Background

Central to medical education (Elstein, 1978; Norman, 2005), diagnosis can been seen as a problem-solving task in which physicians employ a range of strategies to identify a condition: hypothesis testing, pattern recognition, or reference to specific instances or general prototypes. Some view it as a decision-making task in which physicians tacitly follow *Bayes's Theorem* to arrive at a diagnosis based on imperfect information (e.g., patient history and physical examination). However, human biases, such as availability and representativeness, can still contribute to diagnostic errors (Elstein, 2002; Werner, 1995). It is widely recognized that diagnosing is highly individual, dependent on physicians' perceived difficulty and domain knowledge (Elstein, 2002).

Historically, when sensing a health problem, consumers had to decide whether and when to consult a health professional based upon personal knowledge and that of their close social ties. With the Internet becoming increasingly sophisticated, many people have begun turning to it as an initial diagnostic tool to assist themselves and their loved ones ascertain potential causes of their ailments. In addition to the accessibility of the Internet and the richness of health information it offers, other popular triggers include: lack of access to health care services, lack of health insurance, time delays associated with doctor appointments, desire for a second opinion, privacy concerns, and dissatisfaction with previous encounters with physicians (Eysenbach & Diepgen, 1999; Ybarra & Suman, 2008). Because such concerns can be expected to remain for the foreseeable

future, health consumers will likely continue turning to the Internet for diagnostic purposes. This invites a close examination of how people use the Internet for diagnosis and what Internet technologies can offer, particularly emerging social media and crowd-based computing models (Quinn & Bederson, 2011).

Some research has investigated how people search for diagnostic information. Cline and Haynes (2001) pointed out two major ways that people seek diagnostic information: searching directly and participating in various forms of online health communities (some involve only peers and some also involve health care professionals). Analyzing search engine queries, Cartright, White, and Horvitz (2011) categorized diagnostic searches into two intentional states: evidence-based and hypothesis-directed search. In the former state, consumers pursue details and relevance of signs and symptoms. In the latter state, they pursue content on one or more illnesses (including risk factors, treatments, and therapies), as well as how to discriminate between different diseases. In searching, human psychological biases such as base-rate fallacy and availability bias often lead to irrational escalation of concerns and poor decision-making (Lau & Coiera, 2009; White, 2013).

2.1 Online Health Communities

Particularly among consumers living with chronic conditions, such as high blood pressure, diabetes, and cancer, online health communities are gaining popularity (Fox, 2011). MedHelp.org, founded in 1994, boasts over 12 million visitors each month. PatientsLikeMe.com, founded in 2004 and specializing in amyotrophic lateral sclerosis, specializes in tools for tracking and visualizing health information such as treatment history, weight, mood, and by connecting and by employing an in-house scientific research team.

Most communities only involve peer patients, while a small number are moderated by healthcare providers, such as dietitians, nurses, or medical doctors. In pure peer-to-peer communities, participants provide emotional support to one another, seeking and sharing information, and construct knowledge in a collective manner (Coulson, Buchanan, & Aubeeluck, 2007; Wicks et al., 2010). Patients become intimately knowledgeable about their conditions through direct experience of the conditions and managing those conditions day-to-day, often by trial-and-error (Davison, Pennebaker, & Dickerson, 2000; Hartzler & Pratt, 2011). Consequently, their expertise mainly focuses on coping with highly personal issues arising in an everyday-life context, such as how to control my feelings about diabetes (Thorne, Ternulf Nyhlin, & Paterson, 2000). Such experiential information often cannot be provided by health professionals.

Patients may also serve as peer navigators, interpreters, and sometimes even amateur doctors, helping to fill gaps due to patient-provider communication hurdles: short visit time, too much or too little information from physicians, and difficulties in comprehending the information from providers (Rubenstein, 2012). For example, Wicks et al. (2010) found that by learning from peers, users of PatientsLikeMe were able to know better about symptoms and side effects of treatments, and be better prepared to make changes to medications (including medication type, dosage, or stoppage). More illustrative was that 12% of patients changed their physician as a result of interacting with peer patients. Peer-to-peer communities do not necessarily exclude healthcare providers. As a matter of fact, some health professionals may volunteer to answer questions. Oh (2012) found that, among answerer's in Yahoo! Answers' health forums, many were healthcare professionals, including physicians, surgeons, nurses, nurse assistant, and therapists. In communities that involve healthcare professionals as non-volunteers, professionals serve either as moderators, providing clinical expertise when necessary while facilitating patients exchange of information (e.g., some communities on WebMD.com), or in a traditional provider's role to answer patients' questions (e.g., GoAskAlice and ZocDoc).

2.2 The Rise of Crowdsourcing

Recently crowdsourcing platforms emerged as an addition to more traditional online health communities (i.e., social networking, participation, collaboration, and sometimes free).

Crowd-based sources differ from traditional online health communities in several aspects. Firstly, those paid to answer questions are likely to provide direct answers, helping reduce consumers' effort in filtering out irrelevant information and thereby making information seeking more efficient and effective. This could have significant implications because many consumers value information support over emotional support, and cite getting useful information as the major motivation for visiting online communities (Nambisan, 2011). Secondly, online communities are built upon homophily, particularly in relation to conditions and values, whereas crowdsourcing sites typically are not. Therefore, participants of crowd-sourcing sites may

be a different set of users from those of online communities. This allows consumers to reach to a wider variety of people. According to the weak-tie theory (Granovetter, 1983), consumers would be exposed to more information, which could be particularly valuable for diagnosing more exotic diseases. Finally, on crowdsourcing platforms, user-worker interaction is typically mediated by the crowdsourcing platform, reducing asker-worker interaction. As a result, question askers have a more limited social presence (Short, Williams, & Christie, 1976) on crowdsourcing platforms, and this limited social presence may enable these sites to provide better protection of askers' privacy.

In a human computation approach, with people performing tasks instead of automated algorithms, King et al. (2013) crowdsourced identification of atypical nevi (i.e., moles or beauty marks) and showed it to be more effective than individual self-examination. 500 participants were asked to review techniques in identifying moles, then look at 40 images of nevi and circle those perceived to be abnormal. Participants correctly detected melanomas 58% of the time, with 19 people collectively detecting 90% of melanomas.

In a gamification approach, Mavandadi et al. (2012) designed a game in which 31 non-expert gamers were asked to identify whether or not images of red blood cells exhibited malaria infection. After a tutorial and a test game requiring > 99% accuracy, gamers were shown red blood cell images and asked to diagnose cells by "killing" infected cells or "collecting" healthy ones, achieving 99% accuracy on 6321 images.

Another game, *Dr. Detective*, challenges players to extract information from medical texts in order to create gold standards for training natural language processing tools (Dumitrache, Aroyo, Welty, Sips, & Levas, 2013). The game was designed as clue-finding in which players must annotate information which could lead to a diagnosis. For example, players are presented with a short case history of a person and the game may ask the player to find clues that could help diagnose "mixed germ-cell tumors." Annotations generated by players were deemed comparable to annotations generated by an NLP parser. Similar efforts in generating gold standards through crowdsourcing are being developed for medical imaging (Foncubierta Rodríguez & Müller, 2012) and for classification of colonic polyps (Nguyen et al., 2012).

This past year, a pilot mobile application *DocCHIRP* (Crowdsourcing Health Information Retrieval Protocol for Doctors) was developed to crowdsource medical expertise (Sims, Bigham, Kautz, & Halterman, 2014). Similar to Sermo, it provides near real-time communication from other physicians or health care providers. For a pilot study, the authors recruited 85 participants who used the application over 244 days. Questions typically centered on medication use and complex medical decision making, and most participants reported that the application was potentially helpful, especially for rare diseases and cases difficult to diagnose.

Patients' choice of health information sources is affected by source properties: user-source relationships, social norms, and characteristics of the health problem that prompts the information search (Zhang, 2014). However, little is known about how they perceive and different types of online communities, especially the emerging crowdsourcing technologies, and how they approach these technologies when seeking diagnosis.

3 The Medical Cases

The medical cases used for our study were adapted from different sources and divided into three sets (see **Table 1**). The first and second sets of test cases were adapted from Folkestad et al. (2011), which were in turn drawn from Bain and Gupta (2006). The first set, included in their paper, referred to common and well-known maladies that were diagnosed accurately and quickly on Facebook. We therefore deemed this set to be easiest to diagnose. We constructed the second set of cases from their source materials, intending these to be more difficult in terms of describing rarer and more severe illnesses which we expected fewer people would be acquainted through personal experience. We matched the sentence length, formatting, and style of the first set as much as possible when constructing the cases. Both case sets included short one paragraph long description of a person's background and symptoms and typically did not include any background on family history, treatment, or medication. Complete case details are included in the **Appendix**.

Because CrowdMed advertises itself as being able to solve the "world's most difficult medical cases," we constructed our third set of most difficult cases from CrowdMed itself. In contrast with the short descriptions used in other cases, these cases included extensive medical histories provided by the patients, including descriptions such as symptoms, family history, previously suggested diagnoses, treatments, prescribed medications, and narratives of experiences. These medical histories used technical jargon and required domain knowledge to understand, while the diseases were rare and hard to diagnose. Note that CrowdMed cases

Table 1: The three sets of medical cases considered in our study.

Set #	# Medical Cases	Difficulty	Source
Set 1	6	Easy	Folkestad et al. (2011)
Set 2	3	Medium	Folkestad et al. (2011)
Set 3	5	Difficult	CrowdMed

could not be accessed without registering, thus could not be easily found by the crowd through online search. One of the five cases had no known diagnosis at the time of our study. See the **Appendix** for details.

For each website or platform, we posted the three sets of medical cases for diagnosis a few weeks apart. For diagnoses with accompanying rationales, these were manually coded to identify major categories.

4 Diagnosis with Volunteer-based Platforms

To assess feasibility of using simple, volunteer-based crowdsourcing platforms for medical diagnosis, we posted cases on general and medical CQA forums. For general sites, we tried Yahoo! Answers and Able2Know.org ("medical/health-related" category). For medical forums, which typically let people ask doctors for advice after registering, we tried WebMD.com, MedHelp.org, and eHealthForum.com.

While we also tried posting cases to CrowdMed, we encountered several problems that caused us to ultimately abandon this. Firstly, their "free" option requires a separate $50 deposit for each medical case, to be refunded after the case is closed. While the first case we posted in this way was correctly answered and refunded, the payment and refund process was very tedious and slow. Secondly, their paid option requires a minimum reward of $200 per medical case. Given our focus on exploring simple, fast, and transparent crowdsourcing approaches, we considered the above issues unacceptable. We also made several attempts to contact CrowdMed with invitations to partner with us on the study, but we did not receive a response.

4.1 Methodology

In order to post the cases on these platforms, we created a minimal user profile on each platform. In our first experiment, we posted each case from Sets 1 and 2 as a separate question, using the original text as-is in third-person. Question 1 from Set 1 was posted on all the five forums on the same day. The next day, question 2 was posted, and so on. Once all the six from Set 1 were done, questions from Set 2 were handled similarly. We enabled email notification to to know when any responses were posted.

One must select an appropriate category in which to post questions on each platform, given candidate categories suggested. We chose the most plausible suggestion offered. For example, for the first question in Set 1 (see **Appendix**), the chosen category on Yahoo! Answers was "Infectious Diseases", "Cold, Flu, & Cough" on WebMD, "Allergy" on MedHelp,"Cold, Flu and Viral Infections" on eHealthForum. On AbleToKnow, a general QA forum, all questions were posted in the "Medical/Health" category.

For our second experiment, we modified the questions asked, but not the medical case texts. Specifically, we *gamified* the questions to pose them as a challenge to the audience with a hope that it would elicit quicker and better responses. All the questions from Set 1 were posted together under this header:

> Any Dr. House's out there? - I was playing this medical diagnosis game and I'm stuck on these questions! Can anyone help me figure out what the answer might be?

This second experiment was carried out only on Yahoo! Answers and Able2Know forums because the medical forums discourage posting questions not about real situations.

Our third experiment created new user profiles specific to each medical case in Set 2, entering the appropriate age and suitable name. Medical cases were re-worded to be posted in first-person, with one question per forum posted to three health forums: WebMD, MedHelp and eHealthForum.

4.2 Results and Discussion

Results were generally poor, with most sites failing to yield any response at all. The respondents often expressed concern or sympathy rather than a diagnosis. The most common responses were in the form of

"Go see a doctor". On Yahoo! Answers, someone wrote, "He's relying on someone who's asking a random website for medical advice rather than dragging this old man to see an actual doctor or visit a hospital."

Even though Yahoo! Answers yielded at least three responses to each test case in Set 1, they were mostly irrelevant or not helpful. We were particularly surprised by how few responses and correct diagnoses we received from the medical forums. The best response came for a Set 1 (easy) medical case on eHealthForum in which a medical student suggested possible diagnoses and visiting a doctor. For Set 2 medical cases, our questions to these forums yielded almost no correct diagnoses.

For the second experiment with the gamified questions, the results were slightly more positive. On Yahoo! Answers, we found one person answering two of the three cases correctly for Set 2. Looking closely at his previously answered questions, it seems like this person could have some medical background. However, this technique did not yield many results on Able2Know. A more sophisticated approach to gamification (e.g., a trivia challenge) could also significantly increase participation (Ipeirotis & Gabrilovich, 2014). Specialized applications for medical case study (e.g., prognosisapp.com) are already popular and highly rated mobile games. Like DuoLingo.com, future work might also blend learning and work, where students are assigned diganosis problems, with aggregated answers yielding correct diagnoses that pay for their own education.

Our third experiment posting the test cases in first-person failed to yield any new correct diagnoses. Responses still had the tone of general concern, rather than possible diagnosis.

Given these results, we did not proceed with experiments on Set 3. Instead, we reflect here on our design and what might be improved in future experiments. For example, building relationships in an online community may incentivize both more and better responses. However, one can easily imagine cases where privacy concerns might preclude asking questions of those in one's social network, and reliance upon social ties for effective diagnosis could limit the practical utility of the system in such cases. The ability to obtain medical diagnoses anonymously in volunteer forums will likely be highly-valued by some patients. In addition to the above, each community likely has normative use models that patients learn over time in the manner of language or interaction patterns that are typical for yielding helpful responses for diangosis. Rewording medical cases to use less medical jargon may also make questions more broadly accessible. It may be useful to establish trust that we are not medical students trying to cheat on homework.

5 Medical Diagnosis with Mechanical Turk

Mechanical Turk (MTurk) predominantly caters unskilled, entry-level data processing work. Tasks are posted as *Human Intelligence Tasks* (HITs), created by *requesters* and performed by *workers* (or *turkers*). Each task type is defined by a HIT *group* (i.e., instances of the task to be completed). All HITs in a given group use the same (often simple) template interface designed by the Requester for that group. A worker accepts an *assignment* to answer a particular question, and a requester can specify how many different workers he wants to answer each question (e.g., to aggregate multiple answers). The total cost of task can be calculated by taking the product of # of HITs (questions), the # of assignments per HIT, the payment per assignment, and Amazon's 10% surcharge. One can also specify criteria required to be eligible to work on a particular HIT group (e.g., past *approval rating* or a skill *qualification*).

5.1 Methodology

Our HITs were simply designed following best-practices (Alonso, 2009), with each medical case posted as its own HIT, titled, "Help diagnose a person from reading a short description." Each set was posted to AMT with a time limit of one month. We paid $0.10 per assignment, with 10 assignments per question per set. No qualifications were required. HIT instructions were:

> Read the following description of a person and their symptoms. Enter your best diagnosis for the cause in the text box. You may consult outside information (e.g., Wikipedia) if it would help you form a diagnosis. Optionally, you may provide reasons why you think your diagnosis is correct.

Workers were asked to read each textual case and then enter their response into a free-form text box (which required some text be entered to allow submission). Instructions stated that workers could freely consult outside sources such as Wikipedia if it would help them reach a correct diagnosis.

Workers could also provide an optional rationale for their diagnosis in a separate text box. We did not require this for several reasons. Firstly, we did not want to intimidate workers from accepting our task or make them think that a formal medical background was required. Secondly, a common best-practice for assessing quality of crowdsourced work is to check if optional text is provided (Kazai, Kamps, Koolen, & Milic-Frayling, 2011); requiring such feedback often yields spurious text that makes it harder to automatically recognize higher quality answers. Finally, given time and cost associated with traditional diagnostic practices, we really wanted to push the envelope of speed and affordability by making the task as simple as possible.

To aggregate the 10 free-text responses per medical case, we needed to group minor variants of the same answer (e.g., misspellings and capitalization, as well as recognition of synonymous terms). As a proof of concept, we did this manually. In practice, simple variants would be easy to resolve automatically, while harder cases like synonyms would require more sophisticated text processing or creation of an additional task for workers to recognize and resolve synonyms. One worker actually provided two different answers, and we counted each answer in aggregation. Set 1 (easy) cases involved one HIT per question (6 total), with 10 assignemnts each at $0.10 per assignment, yielding 60 total diagnoses at a cost of $6.60 (including Amazon's 10% surcharge). Set 2 and 3 (medium and hard) cases were run similarly at a cost $3.30 and $5.50.

For the medium difficult cases, we also experimented with adding a second-stage *rating* task to improve diagnosis quality. Workers were asked to read the case description and one of the diagnoses submitted in the first stage. We also included information about the given diagnosis from Wikipedia or WebMD. For proof-of-concept, we manually matched submitted diagnoses to the appropriate Wikipedia or WebMD page manually; in practice, this would again require text analysis and/or a human computation task asking workers to find the page for a given diagnosis. We asked workers whether they thought the given diagnosis was correct. They chose from a 5-degree likert-scale radio-selection: *extremely likely, likely, neutral, unlikely*, and *extremely unlikely*, which we conflated to a 3-degree *likely-neutral-unlikely* scale and aggregated.

We received twenty-three distinct first-stage responses. Each was posted as its own HIT, with 9 assignments per HIT (to avoid ties), and paying $0.05 per assignment (presuming the rating task to be easier than coming up with a diagnosis from scratch). This yielded 189 total responses at a cost of $10.40. Combined with the Set 1-3 diagnosis tasks, we posted 35 total HITs yielding 329 assignments (140 diagnoses and 189 verifications), for a total cost of $25.80.

Because MTurk caters to low-skilled, low-paying work, the platform is heavily-skewed toward laymen rather than experts, as demographic studies have shown (Ross et al., 2010; Ipeirotis, 2010), suggesting our respondents can be safely assumed to be largely laymen, not health professionals. Nevertheless, for Set 3 (difficulty cases) where we were most surprised at the workers' accuracy, we further verified this with a $1 follow-up survey. We asked the workers whether or not they had a background in medicine, and to elaborate in a free-answer text box if so. They were then asked to quickly re-familarize themselves with the description of the case they had diagnosed and to make a radio-selection between one of six options best describing their method of diagnosis: "Searched for answer on the internet (e.g., Google, Bing, WebMD, Wikipedia, etc.)", "Searched for answer from a text", "Knew from previous medical training or knowledge", "Knew from personal experience", "Asked a friend or colleague," or "Other [with free answer box]"). As noted at the end of the Medical Cases section, these categories were derived from our analysis of all rationales submitted.

5.2 Results and Discussion

Whereas many crowdsourcing studies have reported problems with "spammers", manual inspection of the responses suggested all represented good-faith attempts to provide a plausible diagnosis. It is of course still possible that some answers represent little effort, but there is little concern over cheating or collusion since workers were invited to consult other sources of information in forming their diagnoses. Beyond aggregation, the additional second-stage rating task also provided a further means of verifying response quality.

Set 1: Easy. We began by adapting the prior study on Facebook (Folkestad et al., 2011) to MTurk. For all of the six medical cases, workers collectively arrived at the correct diagnosis after aggregating their responses. Workers took from 35-70 seconds per answer. 22 of the 60 workers (36.7%) provided a rationale for their diagnosis. Interestingly, workers who provided the optional justifications for their diagnoses were not correct more often than those who did not. Results appear in **Table 2**. Notably, the accuracy of our results exceeded those obtained by Folkestad et al. (2011) on Facebook, who found the correct answer for only five of the cases at a median latency of 10 minutes. For such easy cases, it seems one could post such a case and

get a reasonable aggregated diagnosis within one minute for $1. Worker rationales suggest familiarity and personal experience were important, suggesting difficult cases could prove more troublesome on MTurk.

Table 2: Results for Set 1 (easy) medical cases on MTurk.

#	Correct Diagnosis	Top Response	Count
1	Tuberculosis	Tuberculosis	7/10
2	Rheumatoid arthritis	Rheumatoid arthritis	10/10
3	Appendicitis	Appendicitis	9/10
4	Gastritis	Ulcer/acid reflux	10/10
5	Thyrotoxicosis	Thyroid issues	9/10
6	Gout	Gout	6/10

Given that MTurk is oriented toward laymen rather than experts (Ross et al., 2010; Ipeirotis, 2010), it was particularly interesting to see how workers arrived at their diagnoses. We recognized three broad categories of diagnosis formation: deductive reasoning, personal experience, and Internet searching. 13 workers (59.09%) used deductive reasoning for their diagnoses, for example, answering *appendicitis* because the appendix is located in the lower right side and pain in this area is a common symptom associated with appendicitis. 6 workers (27.27%) based their judgment on personal experience, such as having had the disease or knowing a close relative who did. While only 3 (13.64%) reported having searched the web or reference websites such as WebMD, workers may have under-reported this.

Set 2: Medium Difficulty. For these three cases, only 1-2 workers per case provided a correct diagnosis. Workers took between 49-69 seconds per answer and provided fewer justifications than with Set 1 cases (8 of 30 workers, or 26.67%). As with Set 1, workers who provided optional justifications were not correct more often than those who did not. 5 (62.5%) used deductive reasoning, 2 (25%) used Internet searches, and only 1 (12.5%) indicated personal experience.

Table 3: Results for Set 2 (medium difficulty) medical cases on MTurk.

#	Correct Diagnosis	Top Response	Count
1	Acromegaly	Menopause	4/10
2	Atrial fibrillation	Anxiety	3/10
3	Colorectal cancer	Constipation/hemorrhoids	2/11

In Set 2, aggregation did not correctly diagnose any case, and one case yielded no consensus at all (see **Table 3)**. For case 2, the correct diagnosis of *atrial fibrillation* was never suggested in any response. The rationales provided by workers, and the diagnoses of all common conditions, suggest that most workers continued to use their own deductive abilities to form a diagnosis. While seeming to lack the appropriate medical background and knowledge to arrive at the correct diagnosis, they also appeared over-confident in their own knowledge and reasoning abilities, or under-motivated to search online.

Limited details of the cases may also contribute to the low accuracy of diagnoses. Because we followed the design from Folkestad et al. (2011), the case descriptions were short and had few details. As a consequence, the descriptions may be ambiguous, allowing for multiple plausible answers. For example, in the case of *atrial fibrillation*, all answers conceivably fit, suggesting the case text was too vague.

Observing that at least one worker proposed the correct diagnosis for each case, we perceived an opportunity for a second-stage rating task (Little, Chilton, Goldman, & Miller, 2010) to recognize this correct diagnosis and promote it to the collective diagnosis. Just as CrowdMed asks their *Medical Detectives* to vote on diagnoses, we asked workers to rate each worker diagnosis collected from the original task. Workers spent an average of 48 seconds (ranging from 5 seconds to 10 minutes) on evaluating each case. This approach was very effective: workers highly rated true answers, and disagreed with most untrue answers (see **Table 4**).

However, while all three correct diagnoses were rated highly, some incorrect diagnoses persisted, such as *hemorrhoids* and *stress*. Because these incorrect top responses were symptoms of the correct diagnosis, further improving textual case descriptions may reduce such ambiguity. Of course, second-stage quality is clearly dependent upon generation of a correct diagnosis in the first stage. Finally, there is also a problem with diseases with very similar symptoms; for example, hemorrhoids and IBS are both similar to colon cancer,

Table 4: Results for Set 2 medical cases on MTurk with 2nd-stage rating. Correct diagnoses are *italicized*.

#	Response	Likely	Neutral	Unlikely
1	*Acromegaly*	*5*	*3*	*1*
2	Hypothyroidism	2	2	5
3	Menopause	1	2	6
4	Cherubism	1	1	7
5	Diabetes	1	1	7
6	Bone cancer	0	1	8
7	Congestive heart failure	0	1	8
1	Stress	6	2	1
2	*Atrial fibrillation*	*6*	*1*	*2*
3	Anxiety	5	3	1
4	Caffeine intoxication	5	2	2
5	Bad drug reaction	2	0	7
6	Lung cancer	2	0	7
7	Food poisoning	0	1	8
8	Legionnaires' disease	0	0	9
1	Hemorrhoid	5	2	2
2	*Colon cancer*	*4*	*2*	*2*
3	IBS	4	3	2
4	Colitis	3	2	4
5	Celiac	2	2	5
6	Constipation	1	4	4
7	Diverticulosis	1	3	5

resulting in workers agreeing all three were likely diagnoses. Despite this, workers were able to filter out dissimilar diagnoses and narrowed potential diagnoses for later investigations. Implementing this second-stage task was overall successful and may be further improved with greater refinement.

Set 3: CrowdMed Cases. Diagnoses for Set 3 were inconsistent (see **Table 5**). In some cases aggregation found no consensus or majority, while in two cases, we found 50% of workers agreeing on a diagnosis. It took considerably more time to collect all of the data, ranging from one to three days, with workers taking from 8 seconds up to approximately 40 minutes. 21 of 50 workers (42%) provided reasons for their diagnosis; the majority, 15 people (71.42%), cited deductive reasoning for their diagnoses, while 4 (19.05%) used Internet searching and 2 (9.52%) relied upon personal experience.

Intriguingly, 5 of the 10 workers converged on the same diagnosis of *rheumatoid arthritis* for a case which had no diagnosis at the time of our study. To evaluate these responses, we consulted a local doctor (of internal medicine) to review the case details and provide a diagnosis (without seeing the crowd responses). While the doctor noted that speculations for academic purposes should not be construed as actual medical evaluation, his extensive analysis centered on the same diagnosis by the crowd. See the **Appendix** for details.

Workers providing correct diagnoses took significantly more time than those who did not, approximately 20 minutes. We also followed up with these 5 with a paid survey about their medical background and method of diagnosis (see Methodology sub-section above). Of the 3 workers who responded, none reported a medical background or training, having made their diagnoses primarily via Internet search.

Another CrowdMed case showed similar convergence of workers, although the collective diagnosis of *narcolepsy* was not a diagnosis provided by CrowdMed. On another case, workers proposed a similar diagnosis as CrowdMed of *bradycardia*, but without convergence. Thus, even if collective wisdom cannot agree upon a correct diagnosis, individual members of the crowd might still generate useful hypotheses, perhaps validated via a later second-stage or expert-escalation.

Because these cases were long and detailed, one would have to synthesize several different aspects of case history, such as the symptoms and medications. Since this is difficult and time consuming, it may lead to some cheating, laziness, or fatigue, potentially regardless of payment offered. Each case had at least one worker spend < 1 minute, with a minimum of 8 seconds, on completing the task, indicating low effort. As with other complex tasks, exploring *task decomposition* (Bernstein et al., 2010), iterative-refinement (Little et

al., 2010), or other workflow-design strategies may further improve answer quality.

Table 5: Results for Set 3 (CrowdMed difficult) medical cases on MTurk.

#	Top Diagnosis on CrowdMed	Top Response	Count
1	Diabetic gastropathy	All different	n/a
2	No diagnosis	Arthritis	5/10
3	Hypertrophic cardiomyopathy	Bradycardia	2/10
4	Nystagmus	Narcolepsy	5/10
5	Candidiasis	Sinus/abdominal	2/10

6 Medical Diagnosis with oDesk

oDesk (odesk.com), which recently merged with Elance (elance.com), provides a marketplace for skilled online contracting. An employer posts an open-call describing the task for which contractors may apply. To apply, interested contractors submit their resumes, and employers screen these to select workers to form contracts with. Payment, negotiated by the parties, can be project-based or hourly, and parties may easily interact via rich communication channels as work progresses. In stark contrast with MTurk's microtask, oDesk targets longer duration and/or more complex working relationships between known parties. This design avoids a range of work quality problems typical of MTurk (e.g., due to pseudonymous workers having unknown skill sets and only minimally-established reputations). While this reduces typical management effort required with MTurk for quality assurance, oDesk shifts this effort to other aspects of task management, such as screening applicant resumes, negotiating wages, and establishing a contract. Also, whereas MTurk's model is typified by micro-tasks offering minuscule pay, oDesk's model encourages batching work into larger units, with a $5 minimum. Overall, MTurk and oDesk represent strikingly different models of paid crowd work (Ipeirotis, 2012), with oDesk speifically enabling us to recruit healthcare professionals for diagnosis.

oDesk tasks must be posted under one of several fixed categories. Given oDesk's intent to provide a broad platform for online contracting, we were surprised that the most appropriate category we could find was the "Other" sub-category under "Writing and Translation". Similarly, oDesk asks each task also be posted with a set of predefined contractor skills, for which the best-matching skill we found and used was "medical-translation". While these metadata were not ideal for advertising our task, the detailed text describing the task is both searchable by contractors seeking work, as well as used by oDesk's automated recommendation engine to suggest 10 suitable contractors. We sent invitations to all 10 of those recommended, and 13 other contractors also applied to our open-call. In contrast with MTurk, we spent time communicating with contractors and answering their questions, with a corresponding greater expectation of work quality given a greater understanding of and commitment to our task. We sent each contractor a Word document containing all of the medical cases, and they responded by sending a document with diagnoses.

6.1 Results and Discussion

Four of the ten contractors we invited to apply did so, and from those four and the additional 13 contractors who responded to our open-call, we hired two. One was a registered nurse, and the other a medical transcriptionist with seven years experience on clinical reports ranging from radiology, discharge summaries, clinic notes, physician's letters, medico-legal reports, operative notes, and treatment reports. We selected them due to their healthcare professions, their good ratings from past work, their interactions with us (responsive, positive attitude), and their willingness to accept the $5 payment offered.

Both contractors correctly diagnosed all five Set 1 cases. On Set 2, however, each provided only one correct diagnosis, and each for a different case, though their diagnoses were all close to being correct. For each case they answered correctly, we found that a Turker also provided a correct diagnosis as well. For the case which neither contractors correctly diangosed, no Turker correctly diagnosed it either.

Based on these results, we decided to proceed with Set 3 cases. We invited the same two contractors to the task, but they declined, indicating lack of sufficient expertise. We thus posted another open-call to seek out contractors with more extensive medical knowledge. We ultimately hired a registered nurse from

the Philippines and a medical doctor from Bosnia-Herzegovina, following similar criteria as earlier. The nurse and doctor each provided different diagnoses for each case, explicitly noting that these cases were difficult and that they were unsure of their answers. Comparing their answers along with the ones given by Turkers, we saw that for one particular case, there was a consensus, which strengthens the possibility of that diagnosis being most probable one. In other cases too, we saw an overlap of answers between the Turkers and oDesk workers, but with the latter being more technical with their terms to describe the diagnosis. With regard to the difficult case we had diagnosed by a local doctor of internal medicine (see the **Appendix**), the doctor's diagnosis was in alignment with majority of the Turkers, but surprisingly, not with either of the oDesk workers. Intelligent aggregation across platforms would be interesting in future work.

With regard to diagnosis uncertainty, we also note a parallel with the local doctor of internal medicine. In comparison to laymen, knowledgeable experts tend to provide more nuanced answers to complex questions, often with less confidence, likely due to knowledge of more possible explanations. In addition, medical diagnosis brings both the high-stakes of potential harm to the patient due to incorrect diagnosis (and potential associated malpractice liability), as well as a tremendous gap between reading a written dossier for "academic speculations" vs. conducting an in-person, in-depth medical evaluation per established standards of practice for professional health-care providers. This suggests that determining appropriate best-practices for any mainstream online diagnosis process, and the conditions under which such an online diagnostic tool might be used appropriately, will clearly be an important challenge for future work. Health professionals may also be naturally reluctant to even participate in such a radical departure from established practices.

7 Discussion and Conclusion

The rise of participatory health involves people increasingly turning toward Internet technologies for personalized health care. Given new sites like CrowdMed, offering to diagnose difficult medical cases through non-traditional "medical detectives", we wondered how accurately more typical crowdsourcing sites could be used to diagnose such medical cases.

With regard to our specific research questions, we were able to obtain correct diagnoses from the paid platforms we considered (MTurk and oDesk), but our over-simplistic designs for volunteer sites largely failed to yield correct diagnoses. While we tried simple variants to post questions first-person or gamify questions, we typically received few responses, and those were often oriented towards expressions of personal concern or recommendations rather than a definitive answer. This may be due to the nature of the platforms themselves, which are often focused on advice and social support, rather than giving definitive answers. Members of such sites may also be hesitant in giving a diagnosis because they may feel they are under-qualified or may not want to feel personally responsible for their answers. Calling upon social ties rather than strangers may yield more responses of higher quality, but may not be possible given a patient's desire for privacy.

Regarding case difficulty, MTurk workers (in aggregate) and oDesk health professionals correctly diagnosed all easy cases (more accurately than a prior method using Facebook (Folkestad et al., 2011), with MTurk being faster as well). MTurk workers were also largely correct (in aggregate) on medium cases (though a two-stage design was required). Finally, while MTurk workers generally struggled on the most difficult cases, on one they remarkably converged upon a correct collective diagnosis, despite lack of any medical training.

As for different behaviors of MTurk and oDesk workers, beyond varying diagnosis accuracy, we also noted the qualitative nature of how the two groups differed in their diagnoses. oDesk professionals relied much less on personal experience and answered mostly based on their medical knowledge and expertise. They also express less confidence in both their willingness to provide uncertain diagnoses and the amount of detail and feedback provided. While hiring more contractors or paying more could natually be expected improve oDesk diagnosis, our current design for using oDesk already exceeded that of MTurk.

More generally, studying how amount of pay impacts work products (cf. (Mason & Watts, 2010)) was beyond the scope of our study. While turning to "Dr. Turk" for a medical opinion probed the lower boundary of affordable diagnosis-by-Internet, we expect many people would be happy to pay far more for reliable diagnoses (in proportion to their means and the severity of the case). Our ultra-budget approach was largely unreliable for difficult cases, and in comparison, we note CrowdMed requires that cash rewards offered by patients be *at least $200* (40x the cost our most expensive method). The "sweet spot" for viable crowd-diagnosis may even exceed costs of conventional medicine, in exchange for providing a wider diversity

of opinions beyond the typical 1st or 2nd opinion of traditional practice. Knowing collective wisdom was brought to bear in reaching a diagnosis may offer patients greater confidence at an acceptable premium.

7.1 Future Crowdsourcing and Healthcare

With over a million articles published each year in the biomedical literature, it is tremendously difficult for physicians to keep up with the most recent development in knowledge about diseases and treatments. At the same time, the U.S. healthcare system is calling for patient-centered care, a model actively involving patients in the decision-making about individuals' options for treatment. The overload of medical information, as well as the healthcare movement's focus on patient values, forces decentralization of responsibilities, with increasing responsibility passing to individuals (Johnson, 2014). As such, crowd-based platforms provide an innovative potential channel for consumers to access personalized health information.

Crowdsourcing solutions also pose new risks to consumers, such as quality. Participants on popular crowdsourcing platforms, such as Yahoo! Answers and MTurk, consist primarily of laymen who may provide poor answers (Oh & Worrall, 2013). When workers answer questions for monetary motivations, particularly low payment, they may naturally respond without thoroughly researching a health problem. Moreover, because of MTurk's platform design, workers cannot view each other's answers (unless the requester specifically engineers such). As a result, askers are not able to rely on others' comments to evaluate answers, like they could on other online communities. At the extreme, one could even imagine organized malicious action by others online intentionally seeking to cause harm (Lasecki, Teevan, & Kamar, 2014).

Looking forward, we envision a tiered approach to diagnosis which integrates crowdsourcing platforms, automated algorithms, and healthcare professionals. Like ideation platforms, the crowd might be responsible for generating candidate diagnoses to be further scored/pruned by the crowd and/or an intelligent systems (e.g., an evidence-based knowledge engine), and finally presented to healthcare professionals to interpret. A variety of such hybrid crowd-expert systems can be imagined to combine the best of both worlds.

References

Alonso, O. (2009). Guidelines for designing crowdsourcing-based relevance experiments. In *Proceedings of the 32nd international acm sigir conference on research and development in information retrieval.*

Bain, S., & Gupta, J. K. (2006). *Core clinical cases in medicine and surgery*. London, Great Britain: Hodder Education.

Bernstein, M. S., Little, G., Miller, R. C., Hartmann, B., Ackerman, M. S., Karger, D. R., ... Panovich, K. (2010). Soylent: A word processor with a crowd inside. In *Proceedings of the 23nd annual acm symposium on user interface software and technology* (pp. 313–322). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/1866029.1866078 doi: 10.1145/1866029.1866078

Cartright, M.-A., White, R. W., & Horvitz, E. (2011). Intentions and attention in exploratory health search. In *Proceedings of the 34th international acm sigir conference on research and development in information retrieval* (pp. 65–74). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2009916.2009929 doi: 10.1145/2009916.2009929

Cline, R. J. W., & Haynes, K. M. (2001). Consumer health information seeking on the internet: the state of the art. *Health Education Research*, *16*(6), 671-692. doi: 10.1093/her/16.6.671

Coulson, N. S., Buchanan, H., & Aubeeluck, A. (2007). Social support in cyberspace: a content analysis of communication within a huntington's disease online support group. *Patient Education and Counseling*, *68*(2), 173-178. doi: 10.1016/j.pec.2007.06.002

Davison, K. P., Pennebaker, J. W., & Dickerson, S. S. (2000). Who talks? the social psychology of illness support groups. *The American Psychologist*, *55*(2), 205-217.

Dumitrache, A., Aroyo, L., Welty, C., Sips, R.-J., & Levas, A. (2013). Dr. detective: combining gamification techniques and crowdsourcing to create a gold standard in medical text. In *1st international workshop on crowdsourcing the semantic web at the 12th international semantic web conference* (pp. 16–31).

Elstein, A. S. (1978). *Medical problem solving: an analysis of clinical reasoning*. Cambridge, Mass: Harvard University Press.

Elstein, A. S. (2002). Evidence base of clinical diagnosis: Clinical problem solving and diagnostic decision making: selective review of the cognitive literature. *BMJ*, *324*(7339), 729-732. doi: 10.1136/bmj.324.7339.729

Eysenbach, G., & Diepgen, T. L. (1999). Patients looking for information on the internet and seeking teleadvice: Motivation, expectations, and misconceptions as expressed in e-mails sent to physicians. *Archives of Dermatology*, *135*(2), 151-156. doi: 10.1001/archderm.135.2.151

Folkestad, L., Brodersen, J. B., Hallas, P., & Brabrand, M. (2011, December). [laypersons can seek help from their facebook friends regarding medical diagnosis]. *Ugeskrift for laeger*, *173* (49), 3174–3177. (PMID: 22142603)

Foncubierta Rodríguez, A., & Müller, H. (2012). Ground truth generation in medical imaging: A crowdsourcing-based iterative approach. In *Proceedings of the ACM multimedia 2012 workshop on crowdsourcing for multimedia* (pp. 9–14). New York, NY, USA: ACM. Retrieved 2014-05-07, from http://doi.acm.org.ezproxy.lib.utexas.edu/10.1145/2390803.2390808 doi: 10.1145/2390803.2390808

Fox, S. (2011). *Peer-to-peer healthcare.* (http://www.pewinternet.org/media//Files/Reports/2011/Pew_P2PHealthcare_2011.pdf)

Fox, S., & Duggan, M. (2013). *Health online 2013.* (http://www.pewinternet.org/files/old-media/Files/Reports/PIP_HealthOnline.pdf)

Granovetter, M. (1983). The strength of weak ties: A network theory revisited. *Sociological theory*, *1* (1), 201–233.

Hartzler, A., & Pratt, W. (2011). Managing the personal side of health: how patient expertise differs from the expertise of clinicians. *Journal of Medical Internet Research*, *13* (3). doi: 10.2196/jmir.1728

Horowitz, D., & Kamvar, S. D. (2010). The anatomy of a large-scale social search engine. In *Proceedings of the 19th international conference on world wide web* (pp. 431–440). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/1772690.1772735 doi: 10.1145/1772690.1772735

Ipeirotis, P. G. (2010). Demographics of Mechanical Turk. (CeDER-10-01).

Ipeirotis, P. G. (2012, Feb. 18). Mechanical Turk vs oDesk: My experience. (www.behind-the-enemy-lines.com/2012/02/mturk-vs-odesk-my-experiences.html)

Ipeirotis, P. G., & Gabrilovich, E. (2014). Quizz: Targeted crowdsourcing with a billion (potential) users. In *Proceedings of the 23rd international conference on world wide web* (pp. 143–154). Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee. Retrieved from http://dx.doi.org/10.1145/2566486.2567988 doi: 10.1145/2566486.2567988

Johnson, J. (2014, September). Health-related information seeking: Is it worth it? *Information Processing & Management*, *50* (5), 708–717. Retrieved from http://dx.doi.org/10.1016/j.ipm.2014.06.001 doi: 10.1016/j.ipm.2014.06.001

Kazai, G., Kamps, J., Koolen, M., & Milic-Frayling, N. (2011). Crowdsourcing for book search evaluation: Impact of hit design on comparative system ranking. In *Proceedings of the 34th international acm sigir conference on research and development in information retrieval* (pp. 205–214). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2009916.2009947 doi: 10.1145/2009916.2009947

King, A. J., Gehl, R. W., Grossman, D., & Jensen, J. D. (2013, December). Skin self-examinations and visual identification of atypical nevi: Comparing individual and crowdsourcing approaches. *Cancer Epidemiology*, *37* (6), 979–984. Retrieved 2014-02-20, from http://www.cancerepidemiology.net/article/S1877-7821(13)00145-8/abstract doi: 10.1016/j.canep.2013.09.004

Lasecki, W. S., Teevan, J., & Kamar, E. (2014). Information extraction and manipulation threats in crowd-powered systems. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 248–256). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2531602.2531733 doi: 10.1145/2531602.2531733

Lau, A. Y. S., & Coiera, E. W. (2009). Can cognitive biases during consumer health information searches be reduced to improve decision making? *Journal of the American Medical Informatics Association*, *16* (1), 54-65. doi: 10.1197/jamia.M2557

Little, G., Chilton, L. B., Goldman, M., & Miller, R. C. (2010). Turkit: Human computation algorithms on mechanical turk. In *Proceedings of the 23nd annual acm symposium on user interface software and technology* (pp. 57–66). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/1866029.1866040 doi: 10.1145/1866029.1866040

Mason, W., & Watts, D. J. (2010, May). Financial incentives and the "performance of crowds". *SIGKDD Explor. Newsl.*, *11* (2), 100–108. Retrieved from http://doi.acm.org/10.1145/1809400.1809422 doi: 10.1145/1809400.1809422

Mavandadi, S., Dimitrov, S., Feng, S., Yu, F., Sikora, U., Yaglidere, O., . . . Ozcan, A. (2012, May). Distributed medical image analysis and diagnosis through crowd-sourced games: A malaria case study. *PLoS ONE*, *7* (5), e37245. Retrieved 2014-02-20, from http://dx.doi.org/10.1371/journal.pone.0037245 doi: 10.1371/journal.pone.0037245

Nambisan, P. (2011). Information seeking and social support in online health communities: impact on patients' perceived empathy. *Journal of the American Medical Informatics Association*, *18* (3), 298-304. Retrieved from http://jamia.bmj.com/content/18/3/298.abstract doi: 10.1136/amiajnl-2010-000058

Nguyen, T. B., Wang, S., Anugu, V., Rose, N., McKenna, M., Petrick, N., . . . Summers, R. M. (2012, March). Distributed human intelligence for colonic polyp classification in computer-aided detection for CT colonography. *Radiology*, *262* (3), 824–833. Retrieved 2014-05-07, from http://pubs.rsna.org/doi/full/10.1148/radiol.11110938 doi: 10.1148/radiol.11110938

Norman, G. (2005). Research in clinical reasoning: past history and current trends. *Medical Education*, *39*(4), 418-427. doi: 10.1111/j.1365-2929.2005.02127.x

Nuwer, R. (2013, April). Crowd diagnosis could spot rare diseases doctors miss. *New Scientist*. Retrieved 2014-

02-20, from http://www.newscientist.com/article/dn23392-crowd-diagnosis-could-spot-rare-diseases-doctors-miss.html#.UwV29IXtuHs

Oeldorf-Hirsch, A., Hecht, B., Morris, M. R., Teevan, J., & Gergle, D. (2014). To search or to ask: The routing of information needs between traditional search engines and social networks. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 16–27). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2531602.2531706 doi: 10.1145/2531602.2531706

Oh, S. (2012). The characteristics and motivations of health answerers for sharing information, knowledge, and experiences in online environments. *Journal of the American Society for Information Science and Technology*, *63*(3), 543–557. Retrieved from http://dx.doi.org/10.1002/asi.21676 doi: 10.1002/asi.21676

Oh, S., & Worrall, A. (2013). Health answer quality evaluation by librarians, nurses, and users in social q&a. *Library & Information Science Research*, *35*(4), 288–298. doi: 10.1016/j.lisr.2013.04.007

Quinn, A. J., & Bederson, B. B. (2011). Human computation: A survey and taxonomy of a growing field. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 1403–1412). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/1978942.1979148 doi: 10.1145/1978942.1979148

Ross, J., Irani, L., Silberman, M. S., Zaldivar, A., & Tomlinson, B. (2010). Who are the crowdworkers?: Shifting demographics in mechanical turk. In *Chi '10 extended abstracts on human factors in computing systems* (pp. 2863–2872). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/1753846.1753873 doi: 10.1145/1753846.1753873

Rubenstein, E. L. (2012). "Things my doctor never told me": Bridging information gaps in an online community. *Proceedings of the American Society for Information Science and Technology*, *49*(1), 1–10. Retrieved from http://dx.doi.org/10.1002/meet.14504901126 doi: 10.1002/meet.14504901126

Rzeszotarski, J. M., & Morris, M. R. (2014). Estimating the social costs of friendsourcing. In *Proceedings of the 32nd annual acm conference on human factors in computing systems* (pp. 2735–2744). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2556288.2557181 doi: 10.1145/2556288.2557181

Short, J., Williams, E., & Christie, B. (1976). *Social psychology of telecommunications*. London : New York: Wiley.

Sims, M. H., Bigham, J., Kautz, H., & Halterman, M. W. (2014). Crowdsourcing medical expertise in near real time. *Journal of Hospital Medicine*, *9*(7), 451–456. Retrieved from http://dx.doi.org/10.1002/jhm.2204 doi: 10.1002/jhm.2204

Swan, M. (2012, March). Crowdsourced health research studies: An important emerging complement to clinical trials in the public health research ecosystem. *Journal of Medical Internet Research*, *14*(2). Retrieved 2014-05-07, from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3376509/ (PMID: 22397809 PMCID: PMC3376509) doi: 10.2196/jmir.1988

Thorne, S. E., Ternulf Nyhlin, K., & Paterson, B. L. (2000). Attitudes toward patient expertise in chronic illness. *International Journal of Nursing Studies*, *37*(4), 303–311. doi: 10.1016/S0020-7489(00)00007-9

Werner, M. (1995). A model for medical decision making and problem solving. *Clinical Chemistry*, *41*(8), 1215-1222.

White, R. (2013). Beliefs and biases in web search. In *Proceedings of the 36th international acm sigir conference on research and development in information retrieval* (pp. 3–12). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2484028.2484053 doi: 10.1145/2484028.2484053

Wicks, P., Massagli, M., Frost, J., Brownstein, C., Okun, S., Vaughan, T., ... Heywood, J. (2010). Sharing health data for better outcomes on patientslikeme. *Journal of medical Internet research*, *12*(2). doi: 10.2196/jmir.1549

Ybarra, M., & Suman, M. (2008). Reasons, assessments and actions taken: sex and age differences in uses of internet health information. *Health Education Research*, *23*(3), 512-521. doi: 10.1093/her/cyl062

Zhang, Y. (2014). Beyond quality and accessibility: Source selection in consumer health information searching. *Journal of the Association for Information Science and Technology*, *65*(5), 911–927. Retrieved from http://dx.doi.org/10.1002/asi.23023 doi: 10.1002/asi.23023

Table of Tables

Table 1 The three sets of medical cases considered in our study. 5
Table 2 Results for Set 1 (easy) medical cases on MTurk. 8
Table 3 Results for Set 2 (medium difficulty) medical cases on MTurk. 8
Table 4 Results for Set 2 medical cases on MTurk with 2nd-stage rating. Correct diagnoses are *italicized*. 9
Table 5 Results for Set 3 (CrowdMed difficult) medical cases on MTurk. 10

8 Appendix

8.1 MEDICAL CASE DESCRIPTIONS

Set 1: The first set consisting of six easy cases were drawn from Folkestad et al. (2011).

1. A 62- year-old man had a cough and a fever since he came back from India two months ago. Now he is starting to get a little blood when he coughs. What could be wrong? (Tuberculosis)
2. What disease do you think of when you read: A 38-year-old man has swollen finger joints, swollen wrists and ankles and the joints are sore and swollen and stiff for over an hour every morning? (Rheumatoid arthritis)
3. If you have pain in the lower right side of the abdomen, all the way down below the navel, what might be wrong? (Appendicitis)
4. A 35-year-old woman has a burning sensation in her stomach after eating, even if she only eats very little. She can no longer eat spicy food, drink coffee, or chew gum. What's wrong? (Gastritis)
5. What do you think is wrong? A woman of 26 years has lost 6 kg (13 lbs), feels restless, and sometimes has heart palpitations. She also has a slight swelling on the neck. (Thyrotoxicosis)
6. An elderly gentleman has terrible pain in the big toe phalangeal, it is all white, and he cannot even have a quilt resting on his foot. What do you think is wrong with him? (Gout)

Set 2: The second set consisting of medium cases were constructed from Bain and Gupta (2006).

1. A 44 year old woman says that she sweats excessively, has to get her hats replaced because her old ones are now too small, and says that her face and hands seem bigger than before. What might be wrong with her? (Acromegaly)
2. A 34 year old woman has been feeling anxious and suffering from insomnia for the last few weeks. She also says that she also has diarrhea for the last 10 days, shortness of breath, and some chest pain. In the last few hours, her heart has been racing. What might she have? (Atrial Fibrillation)
3. A 48 year old man mentions that his bowel movements have become infrequent and that it sometimes seems looser. He also once noticed bright red blood on his toilet paper. What could be wrong? (Colorectal Cancer)

Set 3: The third set of difficult cases were taken from CrowdMed. Because the cases are long, only a summary is provided for each case.

1. A 24 year old Caucasian/White woman from the United Kingdom began experiencing her symptoms 8 months ago. Some symptoms include significant abdominal bloating, fatigue, lethargy, nausea, and chronic stomach pain. She has been hospitalized several times previously, and is currently taking up to twelve medications daily, such as insulin (20 units/day), ursodeoxycholic acid (500mg/3 day), and Cetirizine (10mg/day). Her parents are type 2 diabetic. Some previous diagnostics and imaging tests include gastroscopy, CT and X-ray scans, endoscopic ultrasound, and blood tests.
2. A 48 year old Caucasian/White woman from Canada began experiencing her symptoms 1 year and 7 months ago. Some symptoms include severe swelling of all body parts, limited mobility, heart palpitations, and compressed joint integrity, all of which increase in severity with precipitation. She is taking up to 17 medications such as Actonel (35mg/week), folic acid (5mg/day), and Leflunomide (10mg/day). In the past, she has had 3 miscarriages, sepsis due to breast infections, and hospitalized from a motor vehicle accident. Her family has a history of heart disease, diabetes, arthritis, and autism. Beside white blood cell and inflammatory cell activity, her blood work and scans have not detected any abnormalities.
3. A 31 year old Caucasian/White man from Kentucky, United States began experiencing his symptoms 5 months ago. He had early signs of a cold/sinus infection, but chronic pain gradually developed in the abdominal and hip area, sometimes moving to areas. The only medication he is currently taking is Levaquin (500mg/day). Besides glaucoma on his father's side, there are no other family diseases he is aware of. His previous diagnostic and imaging tests include physical exams, blood work, and urinalysis, which were all normal.

4. A 54 year old Jewish/Ashkenazi man from Massachusetts, United States began experiencing his symptoms 4 months ago. His primary symptom is a low pulse of 40 which does not exceed 60 even with serious exertion. He is currently taking simvastatin (10mg/day) and aspirin (81mg/day). There are no significant problems with his background or family history.

5. A 17 year old Caucasian/White woman from Nebraska, United States began experiencing her symptoms 1 year and 8 months ago. Primary symptoms include lethargy, difficulty walking, episodes of total body collapse, eye and facial twitching, and stiffness in the limbs. She is currently taking Baclofen and Valium (no doses listed). She has previously had an adenoidectomy, tonsillectomy, and worms, as well as been diagnosed with dyslexia. Her family has a history of diabetes, high blood pressure, Guillian Barre Syndrome, arthritis, and gout. Her tests included CAT scans, EEGs, EGKs, MRIs, and various drug tests, all of which were normal. Tests for specific diseases such as Lyme's Disease, West Nile virus, toxicara, and toxiplasmosis have been negative as well. Some other excluded diagnoses are narcolepsy with cataplexy, epilepsy, and migranes.

8.2 M.D. DIAGNOSIS OF UNSOLVED CROWDMED CASE

Variable swelling. The patient describes some generalized swelling and abdominal distension. Well described serious causes of generalized edema include heart failure, renal failure and liver failure. Behavioral changes, for example, variation in dietary sodium leading to differences in water retention may underlie some of the variability in generalized swelling. The patient describes some axillary and inguinal lesions that sounds like she was treated for hiradenitis suppurativa. She may have impaired lymphatic drainage from those surgeries which predisposes to edema in the part of the body that lymph drains. Venous insufficiency is a fairly common cause of especially lower extremity edema. Abdominal distension may be due to bloating (gas) that would be more common and more probable explanation in variability in abominal distention. Heart, renal and liver failure can certainly cause fluid to accumulate in the abdomen, but would be unusual to come on abruptly and go away fairly quickly spontaneously. Angioedema can cause episodic abdominal distension due to edema. Cold can be provoking factor.

However, this is not a common diagnosis. *It sounds like the patient is being treated for seronegative rheumatoid arthritis*; this would generally cause distal symmetrical joint swelling rather than more generalized swelling. This patient reports she is smoking, that tends to make RA symptoms worse; obviously she should stop smoking for her overall health. There is a syndrome called remitting seronegative synovitis with pitting edema, that could cause rheumatoid arthritis like symptoms with edema. People have one autoimmune disease are more likely to have a second autoimmune condition. Inflammatory bowel disease could potentially cause episodic abdominal distension. Patients usually report other symptoms so this seems less likely.

As far as the patient's pains, some other thoughts besides RA, she is on the young side to have polymyalgia rheumatica. Vitamin D deficiency sometimes can cause generalized aches. Inflammatory muscle conditions, e.g. polymyositis, can cause body aches. Fibromyalgia is fairly common cause of generalized body pain, she is already being treated for this. There may also be a non-organic component to her pain, it is common for people with depression to have body pains associated. On the other hand, it is common for people with chronic pain to become depressed, hard to tell which came first here.