Title: Crowdsourcing in Epidemiology: Challenges, Proposed Approaches and Standards

Authors: Katie Pezzort, Rebecca Carter, Lu Wang, Jiayang Sun
Corresponding Author Information: Professor Jiayang Sun

Address: Department of Epidemiology and Biostatistics, Case Western Reserve University,

10900 Euclid Avenue, Cleveland OH 44106

Email: jsun@case.edu

Telephone: 216-368-0630

Commented [KC1]: Update

Manuscript word count: **5338**

KEYWORDS (MeSH terms): Crowdsourcing, quality control, data accuracy

ABSTRACT

Introduction: This paper identifies challenges in practicing quality control (QC) for crowdsourcing in epidemiology (CrowdEpi) and proposes standards for performing sound crowdsourcing in epidemiology.

Methods: We included articles that were peer reviewed and identified to be at the intersection of epidemiology and crowdsourcing, by querying from five databases: Medline, Google Scholar, Pubmed (abstracts and keywords only), Pubmed Central (full text) and MathSciNet. To filter and classify articles, we developed a Java-based text runner/mining tool *TextM*, which used a dictionary we developed to classify articles as CrowdEpi and non-CrowdEpi (bogus). A dataset of N=63 articles was also manually labeled by three annotators for validation of our classification method. A term-document matrix of article terminology reflecting the full corpus was extracted, and a log-odds ratio was calculated for each group term using software from the R programming language. Challenges in QC were identified by systematic reviews of the articles falling into each of the two categories.

Results: Observations included group terms from CrowdEpi articles compared to non-CrowdEpi's. More importantly, although crowdsourcing has an expanding role as a problem solver, its approaches have not been standardized for application to epidemiology and QC has not always been used in CrowdEpi. Therefore, this work also contains important guidelines for designing a responsive and effective protocol for CrowdEpi, especially in terms of QC.

Conclusion: We identified current practices in CrowdEpi. After determining where QC was not consistently used, we concluded with proposed guidelines/standards for crowdsourcing protocols. These guidelines should have applications beyond crowdsourcing in epidemiology.

Abstract word count: **243/250**

INTRODUCTION

The term ‘crowdsourcing’ was coined in 2006 to describe the act of outsourcing a job traditionally performed by a designated agent to an undefined, generally large group of people in the form of an open call for participation [1,2]. Crowdsourcing has expanded since then to involve a wide range of topics and platforms that permit an ever widening range of people with diverse experiences to participate.

In particular, harnessing the power of the crowd in modern medicine has important implications for public health. For example, searches for “flu” via Google trends, a service that shows how often a particular search-term is entered relative to the total search-volume across various regions of the world has been shown to almost mirror CDC estimates of flu prevalence. Similarly, Amazon Mechanical Turk and Twitter have also been used to estimate disease prevalence in real-time [3,4]. Promising new applications include the accurate diagnosis of medical ailments, real-time epidemic disease surveillance, and conducting traditionally expensive studies that characterize rare diseases such as amyotrophic lateral sclerosis (often called ALS or Lou Gehrig's disease) for pennies on the dollar online [5–7]. Thus, crowdsourcing is taking on a rapidly large and informative role in the field of epidemiology.

Our work was motivated by an invited investigation to evaluate the status of crowdsourcing in epidemiology (CrowdEpi). Then we discovered that quality control had not been consistently used in crowdsourcing, such as in CrowdEpi. There are many important and challenging factors to consider in conducting a sound study in crowdsourcing. There are articles discussing some of these vital aspects, such as rating scales, redundant labeling, or the application of machine learning techniques to rate and evaluate participants [8].

Our main objectives in this paper are 1) to identify the presenting characteristics of CrowdEpi and non-CrowdEpi (bogus) articles and 2) to detail the design components of crowdsourcing protocols for

Commented [KC2]: Are you referring to the word usage? Terminology might be a better word than presenting.

both prospective studies and retrospective studies, with regard to data collection. Additionally, we articulate quality control strategies embedded at each stage with the overall intention of promoting reproducible science.

The significance of this work is a framework for designing a responsive and effective crowdsourcing study in Epidemiology and improving data quality from online sources that are prone to large heterogeneities and biases.

MATERIALS AND METHODS

In a three-stage process we used our java-based text classification tool *TextM* to query relevant articles, filter data, mine text and obtain statistics of these articles in epidemiological crowdsourcing.

Data Source and Extraction

In preparation for the statistical analysis of CrowdEpi trends, we conducted a full search of the available literature in crowdsourcing and epidemiology. To obtain as complete a representation of epidemiological crowdsourcing articles as possible, we searched five databases: Medline, Google Scholar, Pubmed (abstracts and keywords only), Pubmed Central (full text) and MathSciNet. The query terms "(crowdsourcing OR crowdsource OR crowdsourced OR macrotask OR microtask) AND (disease OR health OR epidemiology)" were used to retrieve relevant data sources of epidemiological crowdsourcing. We used the query terms "(crowdsourcing OR crowdsource OR crowdsourced OR macrotask OR microtask)" to retrieve more articles that were not epidemiological crowdsourcing for better training of our model.

Data Preparation and Classification

For the first stage, following aggregation and removal of duplicate articles, 656 articles were filtered via our Java-based tool *TextM* which allowed us to identify articles that met the criterion of both Epidemiology and Crowdsourcing resulting in a final corpus of 187 articles to support dictionary

development. For the second stage, the authors read 94 articles to develop a custom dictionary of 33 group terms that were characterized by terminology and Boolean phrases. For the third stage, each article was downloaded with the full-text scraped with a Ruby script. Each of these documents was scanned for the presence of terms of interest from the custom imported dictionary via *TextM*. The scraped terms were binned into 33 group terms such as "health," "disease," "medical," and "survey" according to the dictionary. Finally, a training sample of 63 articles were assigned a label as either CrowdEpi or Bogus, where a "bogus" article indicated that the article did not fall under the required intersection of epidemiology and crowdsourcing. This labeling was conducted by 3 independent annotators, and all cases in which disagreement was present were reviewed collaboratively by the 3 annotators. Consensus was reached by a final decision moderated by the principal investigator.

Commented [KC3]: A subset of the 187 or the 656.

Commented [KC4]: Each of the 94, 187 or 656?

Commented [KC5]: Subset of which group? If this is a subset of the 187, haven't you already determined that these are NOT bogus?

Metrics

We sought to evaluate which words were most common to CrowdEpi studies relative to Bogus studies, and vice versa. A measure of the log odds ratio was calculated for each group term as:

$$log_2\left(\frac{\frac{Freq.per\ group\ term\ for\ CrowdEpi}{Total\ per\ group\ term\ for\ CrowdEpi}}{\frac{Freq.per\ group\ term\ for\ Bogus}{Total\ per\ group\ term\ forBogus}}\right)$$

For visualization purposes, the absolute negative value of the log odds of Bogus words relative to CrowdEpi words was calculated. All data analyses and visualization were performed with the statistical software R version 3.1.2.

RESULTS

Characteristics of CrowdEpi articles

As a reflection of sparsity in the corpus, 29 out of the 33 group terms were extracted from the corpus of 63 articles. Group terms from CrowdEpi articles compared to Bogus articles were highly distinct. For example, the odds of the word "nonexpert" meaning non-expert participants was 4.34

Commented [KC6]: Why only 29?

times larger for the odds of being used in a CrowdEpi publication compared to the odds of being in a bogus article. The odds of "gold" representing the term "gold standard" in a CrowdEpi article was 3.82 times larger than the odds of being used in a bogus article. Bogus articles were characterized by fewer scientific and more unspecific terms, where the word "media" and "online" had 2.47 and 1.86 times larger odds of being in a bogus article than the odds of being in a CrowdEpi article, respectively.

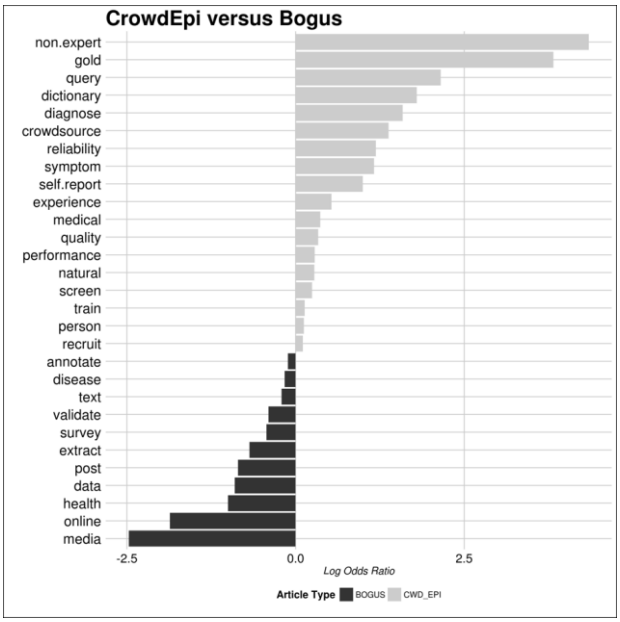

Figure 1. Representative Vocabulary: Crowdsourcing in Epidemiology to Non-Crowdsourcing in Epidemiology Log Ratio of Vocabulary

Commented [KC7]: Does this mean Crowdsourcing but not epidemiology, or Epidemiology but not crowdsourcing, or could it be either one?

In the process of examining these articles, we recognized that although the literature reflected how crowdsourcing was used creatively to tackle difficult problems, the approaches and even terminologies of crowdsourcing had not been standardized with an implementation of quality control for application to epidemiology. This kind of phenomenon is also prevalent in other fields that use crowdsourcing.

Previous publications, notably work from Good & Su (2013)[2] thoroughly addressed the concept of crowdsourcing and analyzed different approaches to crowdsourcing in bioinformatics, by comparing their advantages and disadvantages. Similar works from Khare *et. al.* highlighted the strengths and weaknesses of different approaches, which were classified by data resource used for crowdsourcing and called for innovations in task design and quality control in crowdsourced biomedical research [9]. Brambilla *et. al.* proposed an explorative approach for designing crowdsourcing tasks, but issued no comprehensive guidelines for a crowdsourcing protocol [10]. Here we present important guidelines, written with the intention of being used as a manual by task developers for reproducible task design, especially in regard to crowdsourcing tasks.

Guidelines for a Crowdsourcing Task

We discuss in this section appropriate techniques for developers to consider when designing a crowdsourcing task for epidemiology, although many aspects in CrowdEpi can be translated to general crowdsourcing tasks. We classify these guidelines into two groups, for a) *Crowdsourcing Study Design for Prospective Data Collection* (a.k.a. Prospective Crowdsourcing Study) and b) *Crowdsourcing Study Design for Mining Existing Crowd-sourced Data Sources* (a.k.a. Retrospective Crowdsourcing Study), respectively. A prospective crowdsourced study can include completing non-personal tasks or providing personal information. Existing data sources can include social media (ie: Facebook or Twitter) or mobile applications or websites (ie: MapMyRun or Github).

A. Guidelines Crowdsourcing Study Design for Prospective Data Collection

As with any sound study protocol, it is imperative to have a responsive and effective study design that conforms to the applicable parts of the Institute of Medicine (IOM) standards [11] for

i. Study design/questionnaires
ii. Sample size/endpoints
iii. Study population (Inclusion and exclusion criteria)

iv. Recruitment strategy

v. Data monitoring (Data safety and quality control)

vi. Informed consent

To adhere to the IOM's standards and good statistical practice, we organize key factors specific to crowdsourcing as follows.

A.1. Study design/questionnaires

The most important point in study design is adhering to the particular study objectives while developing a Responsive and Unbiased questionnaire. Here we focus on the guidelines that are specific to designing questionnaires intended for online/crowdsourcing populations with statistical and qualitative considerations, while leaving out the minimal standards that are the same for sampling from the traditional populations, such as, obtaining an IRB approval, storing data in a secure place and having a statistician review your study design to guide the writing of the specific questionnaire.

A.1.1 Having responsive/well-designed questions

Task developers can sandbox the overall functionality of their tasks prior to releasing them to the public. There are several important points to address before releasing a task to the public during the sandbox preparation stage.

Commented [KC8]: Will readers understand this term?

1) Avoid too Many Questions: Developers should refrain from including an excessive number of questions in their task, as participants are more likely to abandon a survey partway through, a phenomenon known as breakoff. Breakoff may be due to a concept known as survey fatigue, when a survey is too long.
2) Consider the Question's Intent: Survey fatigue is also caused because a particular question is ambiguous, lacks an appropriate response, or is considered to be too intrusive given the context of the task at hand [12].
3) Remember BRUSO: The BRUSO acronym meaning, Brief, Relevant, Unambiguous, Specific,

and Objective, is essential [13]. Items should be brief but also be written in complete sentences. Questions should also be related to the subject at hand, as surveys that include items not seemingly relevant to the primary objective are often returned incomplete.

4) Be Specific: Surveys should avoid the use of abbreviations, jargon, and acronyms. When developing questions, center them around a certain time frame: "In the past 6 months, how would you describe your health?" rather than "How would you describe your health?"
5) Frame Positively: Negative questions, e.g., "Do you think parents should not be allowed to be present during their child's non-surgical procedure?" should be avoided as they introduce the linguistic complexity of the double negative. Evaluate your task for stacked questions, i.e., "How satisfied were you with your primary physician and specialist's care?" Questions such as these should be split into two separate questions. In addition, questions should never be leading, evocative, or biased toward a particular point of view.
6) Organize your questions with a purpose: If the questionnaire is for determining knowledge, it is better to start with easier questions and end with the more difficult questions [14].
7) Learn from Design and User Interface: Careful selection of question design and user interface as it pertains to the user experience is an important component as well. Use of checkboxes, sliders, short-answer text boxes, for example, should be appropriate for the particular question and be visually appealing.
8) Test by Focus Group: Most importantly, question design should be tested and evaluated by a focus group of at least 5 individuals [15] prior to release on both a browser format and mobile format. Feedback from the focus group should be integrated where appropriate for the final question choices, including awareness that participants should not be forced to guess on what is expected of them.
9) Anchor vignettes at the beginning of each question: Response outliers and underperformance

may not be due to cheating, but due to misunderstanding or confusion. Validity of your task items can be enhanced by anchoring vignettes at the beginning of each question or task, to ensure interpretability is standardized for all participants [16]. Survey fatigue bias will have to be considered if the anchoring vignette is used.

A.1.2. Minimizing biases in questionnaires

Bias causes results or inferences to deviate from the truth. Although a 100% elimination of possible biases cannot be guaranteed, it is important to understand the causes and strategies that can be used to minimize bias. According to Kleinbaum *et. al.* there are three main classes of biases found in study design: selection bias, information bias, and confounding [17]. These are in addition to the interviewer/observer bias and recall bias from participants. To reduce the three main general biases and the interviewer/observer bias (if there are actual interviewers/observers), we recommend following applicable guidelines, especially the Field Epidemiology Manual by the European Center for Disease Prevention and Control [18]. Hammer *et. al*. [19], Hassan *et. al.* [20] are also good resources and Sun *et. al.* [21] has additional insights for online surveys. Relevant standards from IOM [11] are useful for systematic reviews and clinical practice, if they are applicable to the intent of a questionnaire.

When data is collected from a crowdsourced population, recall bias has been reported to be the largest threat to the internal validity of studies using self-reported data and is of particular concern in regard to the three aforementioned types of bias [20]. Recall bias occurs when participants report past events that are different from the ground truth, which can lead to differential misclassification of the related variable among study subjects with a subsequent distortion of measure of association in any direction from the null, depending on the magnitude and direction of the bias. We suggest minimizing the recall bias by enabling the following techniques:

1) Free-text response questions should be designed with no reminders, hints, anchoring, or other biasing effects. Applying a free-text strategy for some questions can be a significant

improvement over a multiple-choice only strategy, plus it facilitates findings not known before.

2) PROMIS Guidelines: Task developers would benefit from assessing if their task questions conform to standards set by Patient-Reported Outcome Measurement Information System (PROMIS) and the European CDC [18,21]. This is because PROMIS questions were developed to bring attention to various systematic biases (also known as cognitive heuristics), including data collection being sensitive to the manner in which experiences are recollected [22].

A.2. Sample size/endpoints

Study endpoints and desired sample sizes should be determined on a case-by-case basis for the particular study objectives and design. A good reference for sample size determination for various types of studies is Chow, Shao, and Wang (2008) [23]. The guidelines/tips for increasing sample size in CrowdEpi will be addressed in the Recruitment Strategy section.

A.3. Study population

Identifying a data source/data hosting platform for desired study population

Both the task and its intended population are key considerations at the start of any research study, which makes deciding where to host a crowdsourcing task an important decision. In order for a task to reach its desired population, participants from a particular site must be a representative sample of that population outright, or participants must be easily filterable to a representative sample.

1) Determine feasibility of the platform for best accessing your population of choice by posting a brief demographic survey on all potential target sites. Some crowdsourcing sites have already been evaluated on their use for clinical studies [15]. Amazon Mechanical Turk, a site dedicated entirely to crowdsourcing, has been shown to have a very diverse user base [24]. Niche sites, such as PatientsLikeMe, attract individuals from rare disease populations such as multiple sclerosis or ALS [25].
2) Choosing the right platform will influence the success of a task, as task developers can access a

stream of participants who come to the site with the intent of completing tasks. There are also crowdsourcing tasks that cater to games or puzzles. These tasks perform better outside of crowdsourcing-centric platforms [26]. Examples of successful game-based tasks hosted outside of crowdsourcing-centric platforms are Galaxy Zoo, a site where participants can classify objects in space, and Foldit, a game where participants help predict protein structures [27]. These platforms include a participant verification system or a participant reputation system based on the number of tasks a participant completed well. Other advantageous features include qualification tests or participant profiles where participants can list their relevant experience. Designers would best refer to the PCORI methodology standards for formulating research questions to enhance their choice of intervention and target group for their crowdsourced task (PCORI, 2013).

A.4. Recruitment Strategy

A.4.1. Develop an advertising strategy for recruitment

Distribute information about and solicit participants for your study. This can include getting permission to advertise your task on an online forum relevant to your population, or physical fliers that your population of interest may encounter.

A.4.2. Pay participants fairly

Paying participants appropriately is a great way to increase participation [28]. Tasks with higher payouts were found to have a faster study completion time [29]. However, the quality of the responses was not found to increase after a reasonable level of payment [30]. Usually, payment is scaled depending on the time it takes to complete the task. Task developers deciding on per task payment need to consider the complexity of the task, how competitive the pay rate of their task is compared to similar tasks, and how to make their task's payout more attractive to participants given the platform. For example, on a crowdsourcing-centric platform, participants are more likely to choose shorter tasks to

make more money per hour, while non-crowdsourcing websites may not have this pressure on payment fee scales. Our recommended guideline is to avoid the temptation to pay less than 5 cents per task despite the surplus of 'penny tasks' posted online. Out of respect for high-quality participants, we recommended evaluating a task based on its per minute completion time, starting from at least 15 cents per minute. With a complex task such as research or transcription, then the rate should be at least 25 cents per minute. Start at a rate of pay of $9.00 per hour, then increase the rate or supplement with bonuses until the apex of quality and speed is achieved.

A.5. Data Monitoring

A.5.1. Give a good first impression

Participants decide if they are interested in the task based first on the task title and keywords. Therefore, developers need to ensure that the title of their task is concise and direct.

1) Easily Queriable Keywords: Participants should have a good idea of what is expected from them based on the task title and from the task description. Keywords related to the task should make it easily searchable within the host platform. For example, useful keywords for a video transcription are "mp3", "wav", "transcribe", "audio", "transcription", and "video", where 3 to 5 of them could be appended to the task description.
2) Use Descriptive Keywords: At the start of task development, it is necessary not only to consider the study objectives, but also to use the suitable language for the target population. An important exercise would be to evaluate the range of keywords used in tasks similar to the intended task and note the most frequently used keywords, as well as those that are highly-descriptive and understood. For example, although the scientific term for labeling a task is "classify," a more broadly understood word is "categorize".

A.5.2. Include questions that circumvent cheating or misunderstanding

A significant component in producing a robust, reproducible task using crowdsourcing is the

consideration of how participants might underperform, take a short-cut, or cheat on a task. Developers may consider questions that can be designed to prevent or detect unusual participant activity in three different ways.

1) Block the copy/paste function: A previous study that required participants to translate text from Urdu to English found that blocking the copy/paste function on the Urdu text significantly reduced online translator usage and the overall translation quality improved [31].
2) Anticipate common sources for responses: Anticipating common cheating answers is also a fast way to filter out biased data. For example, if the task requires participants to write their own definition, ideally the researcher would compare answers to the publicly available Wikipedia definition before analyzing participant responses.
3) During the crowdsourcing task design process, don't begin with a statement for participants not to search the internet for correct answers but possibly include the cue that "valid answers are what you know at the time of the survey." Alternatively, task developers could articulate that the emphasis is learning more about the participant's viewpoints and experiences and that there are no wrong answers.

A.5.3. Screening low quality participants and validating outcomes

When developing a crowdsourcing task, it is important to screen out poorly performing participants early to effectively harness the crowd to obtain accurate conclusions.

1) Gold Standard: One way to validate results of a crowdsourcing study is to compare results to a gold standard data source, which is typically collected from an expert in that task or collected from known reliable resources such as the CDC [32]. If crowdsourcing results are reasonably similar to that of the expert, they can be considered reliable.
2) Majority Rule: One method for assessing if results are reasonably similar is to assume the most popular response for a question is the correct response, validating the data when the ground

truth is unknown [8,33]. Individuals who are far from the gold standard results or do not match majority rule results consistently are not likely to be good participants.

3) Self-Control Method: Another method is to split participants into two groups, a control group and a test group, where one group of participants rates the others' work. If the raters score the task highly enough, then the output is considered valid for reporting [34]. This approach is most practical for tasks that require written answers and are not easily checked with automatic methods. However, this method could be expensive and time-consuming due to the need to hire two groups of participants.
4) Dynamic Screening: Many crowdsourcing platforms have a built-in participant rating system which evaluates participants on previously completed tasks. The administrator of the task can rate participants who completed their task. Participants who consistently perform poorly on tasks can potentially be excluded from participating in certain tasks.
5) Consistent Responses: Incorporating redundant components within a task, i.e. asking the same question differently and with possible answers ordered differently as before, and checking for similar responses is an effective way to assess for attention or care being paid to the task. However, this technique should not be over used as too many redundant components within a task can annoy participants. A successful strategy when using repeated questions with multiple choices is to arrange them in a permuted order. This assesses how sure a participant is of their given answer, independent of his or her self-report of certainty [35].
6) Time the Task Completion: This approach assesses both participant quality and evaluates the task. Participants who take an unusually small amount of time (i.e. less than 60 seconds) are likely not completing the task properly. Participants who require an exceptionally long time are likely confused about the task or are distracted. Given that a participant may have expertise in the subject or may have experience in crowdsourcing tasks, estimating appropriate time to

complete a task can be evaluated by measuring the number of words in each question and approximating how long it takes to read all the words. In 2012 Trauzettel-Klosinski *et. al.* showed that the average reading speed in 17 different languages was 184±29 words per minute or 863±234 characters per minute [36]. If a participant's task completion time is less than it takes to read the words in the task, then the task was likely not completed with care.

7) Provide Training: Including a training component for participants to complete if the question requires some skills or familiarity helps to ensure that the task is more likely to be completed by participants who are competent and who understand the requirements of the task. For example, a study conducted by Mavandai *et. al.* included an informative tutorial and a training stage that all players had to pass with a score of 99% to continue to the main game [7].

A.5.3. Have a well-designed interface that is visually appealing, clean and intuitive

Participants will be more likely to complete tasks containing a clean and intuitive interface.

1) Inspiration from User Experience: Guidelines for developing a clean interface via user experience recommendations include using contemporary speech to make the task more inviting to participants, consideration of a balanced ratio of white space to text and graphics, as well as consistent formatting and size of text and graphics. Above all, alignment is crucial to giving your designs a clean presentation and to convey organization [37]. Slight misalignments are distracting and can give the task an amateur appearance.
2) Design: Overall, the emphasis of the task should be from the perspective of supporting visual thinking so people can meet their informational needs with a minimum of conscious effort. This includes consideration of how a task might look on a mobile phone or tablet. If making a task visually appealing is difficult to do on the host platform of choice, assess if an alternative software or a third-party website can supplement the task design. For example, one can integrate a third-party software site with crowdsourcing platforms such as Amazon Mechanical

Turk, which has the capacity to accept a randomized completion code from the participant if the task is completed using a different site.

A.5.4. Be an active requester

Participants respond well to an active and engaged task developer.

1) Engage in Discussions: Reviewing and contributing to discussions on crowdsourcing-related open forums about improving tasks to benefit the participants can entice participants to complete the tasks submitted by the developer. However, task developers should create an email address specifically for crowdsourcing projects to facilitate communication between the researcher and participants while simultaneously protecting the task developer from potential spam.
2) Be Responsive: Providing your designated crowdsourcing email address for possible questions from participants can boost interest and competency for your task. Furthermore, emailing participants when necessary is a great way to keep them aware and up-to-date with any events of interest concerning your task, and even invite previously well-performing participants to perform their latest tasks [27].
3) Bonuses: Encouraging participants with a small bonus of $1 or less for good performance on a task will oftentimes motivate competent participants to perform more of your tasks.

A.5.5. Provide in task feedback

Ideally, task developers should consider tracking participant progress throughout the course of the task and provide periodic feedback. This technique is most successful for multistage tasks. For example, Hirth *et. al.*'s task design embedded a feedback component after every scoring round in their crowdsourced game. This allowed participants to gauge how well they were performing in the game [34].

A.5.6. Include a comment section

Despite robust development prior to launching the task, some components of the task may still be confusing to participants. Providing a free-text feedback section following the completion of the task allows participants to address concerns, unexpected challenges, design issues, or suggest new ideas. Furthermore, the task developer can use their crowdsourcing-designated email to thank participants for taking the time to provide responses.

B. Crowdsourcing Study Design for Mining Existing Data Sources

Task developers must be cautious when mining from publically available data. If the population being mined from is not representative of the target population, there are likely to be biases in the data. For prevention, two bias minimization techniques for mining real world data should be conducted: 1) Screening sources, 2) Sampling user profiles. This is because the randomness assumption for any crowdsourcing platform may not be valid as participants tend to use applications or websites based on the choices of people in their social network [38,39]. There may be a confounding variable that affects outcome interpretation as user bases of certain applications may be interconnected with latent social structure clusters.

B.1. Screening existing sources

After a target data source is identified, participants need to be evaluated and screened. Evaluating Active Participants: Collecting data from active participants is ideal for obtaining a reasonable sample [40]. This can be accomplished by evaluating the past activity of a user or examining temporal usage patterns and number of posts from the sampled participants in the past year. However, a caveat to choosing active participants is bots. Bots are computer generated profiles that follow many people, but often have few followers and post redundantly [41]. Bots have a variety of uses such as impersonating real participants, providing followers for participants who paid for them, and serving as advertisers for businesses. As a rule of thumb, bots can be identified by long-term hibernation in their temporal patterns, or conversely showing activity only at certain times of the day

Commented [KC10]: Prevention of what?

Commented [KC11]: I don't understand this. Aren't individuals likely to only show activity at certain times of the day?

and by the presence of URLs in tweets [42].

B.2. Targeted sampling

B.2.1. Use keywords and locations

Following screening, task developers may want to access data from a specific population of participants. Using post keywords, group membership, post location, reposting patterns or post tracking can help access a population. Additional user information such as age, location, and gender may be available to help identify an appropriate sample. However, task developers should ensure that accessing information about a population of interest prior to evaluation is within regulatory compliance protocols [43].

B.2.2. Leverage alternative tools to access existing sources.

Most social media websites have a large, publicly accessible database with an application programming interface (API) that allows developers to gather user-specific data. For example, Twitter has several APIs that allow for the collection of tweets relevant to a given search within a specific time frame. This search can also filter by geo-location and can automatically remove duplicate tweets. Twitter gives information related to geographic location, post time, tweet author, and tweet recipient for each tweet, and also provide a large amount of data via Tweet Firehose [44,45].

B.2.3. Consider Existing Trends

Temporal trends determined from data can be verified using federal health data sites such as CDC or SEER [46,47]. Task developers should take caution when conducting analytics with this type of data source given that valid social media data is voluntarily submitted and may also be time sensitive. For example, if a researcher is trying to collect information about a population's nutritional habits, farmers markets and food festivals are more likely to be discussed during the summer months [48].

DISCUSSION

Looking forward, the different uses of the crowdsourcing paradigm open up many new avenues for scientific exploration. Proper epidemiological crowdsourcing and classification is useful and must be continuously updated. Therefore, careful adherence and thoughtfulness to protocol and quality control is required for obtaining valid results. We offered a systematic guideline that promotes effective crowdsourcing as a research tool. Our approach is the first to merge statistical and qualitative considerations, identify new problematic and unique challenges in the approaches being used, and suggest how to embed these approaches with quality control to promote reproducibility.

CONCLUSION

Table 1: Check List for Main Items of Crowdsourcing for Epidemiology with Quality Control

- ✓ Conduct a test run of the questionnaire by volunteers to ensure your participants comprehend the task instructions as originally intended.
- ✓ Consider the platform requirements for the task of interest.
- ✓ Pay participants fairly and in line with similar tasks, if applicable.
- ✓ Sandbox before releasing a task.
- ✓ Follow the recommended strategies to minimize biases.
- ✓ Promote feedback at the development stage and during the execution of the task.
- ✓ Implement a quality control scheme that includes an alert if a keyed-in value is out of range and a measure that can be used to infer the quality of a response.
- ✓ Pre-screen and sample from your data sources.
- ✓ Use stratified sampling to cover as sufficiently as possible the target population, when mining existing crowdsourced data.

Commented [KC12]: You don't discuss this in the paper. How does it help?

Crowdsourcing permits data to be cheaply and rapidly collected. When used in tandem with various recruitment platforms and increasing computational and data storage capabilities, crowdsourcing is even easier to accomplish for many scientists. However, without proper quality control the reliability of crowdsourced data is unknown. The challenge of successfully orchestrating a scientific crowdsourcing initiative should not be underestimated. Therefore, in order for crowdsourcing to become widely accepted, quality control methods need to be well defined and consistently

implemented. The architecture of the crowdsourced task and validation of the outcomes is even more important as the scientific problem-solving capacity of crowdsourcing expands in unpredictable and exciting ways.

References

[1] Howe J. The rise of crowdsourcing. Wired Mag 2006;14:1–4.

[2] Good BM, Su AI. Crowdsourcing for bioinformatics. Bioinformatics 2013;29:1925–33. doi:10.1093/bioinformatics/btt333.

[3] Aramaki E, Maskawa S, Morita M. Twitter catches the flu: detecting influenza epidemics using Twitter. Proc. Conf. Empir. methods Nat. Lang. Process., Association for Computational Linguistics; 2011, p. 1568–76.

[4] Chunara R, Chhaya V, Bane S, Mekaru SR, Chunara R, Chhaya V, et al. Online reporting for malaria surveillance using micro-monetary incentives, in urban India 2010-2011. Malar J 2012;11:1–7.

[5] Ortiz JR, Zhou H, Shay DK, Neuzil KM, Fowlkes AL, Goss CH. Monitoring influenza activity in the United States: a comparison of traditional surveillance systems with Google Flu Trends. PLoS One 2011;6:e18687.

[6] King AJ, Gehl RW, Grossman D, Jensen JD. Skin self-examinations and visual identification of atypical nevi: Comparing individual and crowdsourcing approaches. Cancer Epidemiol 2013;37:979–84.

[7] Mavandadi S, Dimitrov S, Feng S, Yu F, Sikora U, Yaglidere O, et al. Distributed Medical Image Analysis and Diagnosis through Crowd- Sourced Games: A Malaria Case Study. PLoS

One 2012;7:e37245.

[8] Zhang Y, Chen X, Zhou D, Jordan MI. Spectral Methods meet EM: A Provably Optimal Algorithm for Crowdsourcing. arXiv Prepr arXiv14063824 2014:1–37.

[9] Khare R, Good BM, Leaman R, Su AI, Lu Z. Crowdsourcing in biomedicine: Challenges and opportunities. Brief Bioinform 2016;17:23–32. doi:10.1093/bib/bbv021.

[10] Brambilla M, Ceri S, Mauri A, Volonterio R. An Explorative Approach for Crowdsourcing Tasks Design. Www 2015 2015:1125–30. doi:10.1145/2740908.2743972.

[11] Systematic Reviews And Clinical Practice Guidelines Improve Healthcare Decision Making. http://resources.nationalacademies.org/widgets/systematic-review/infographic.html

[12] Tosch E, Berger ED. Surveyman: Programming and automatically debugging surveys. SIGPLAN 2014;49:197– 211.

[13] Peterson R. Contructing Effective Questionnaires. Thousand Oaks, CA: SAGE Publications; 2000.

[14] Tait AR, Voepel-Lewis T. Survey research: it's just a few questions, right? Pediatr Anesth 2015;25:656–62. doi:10.1111/pan.12680.

[15] Shapiro DN, Chandler J, Mueller PA. Using Mechanical Turk to Study Clinical Populations. Clin Psychol Sci 2013:2167702612469015. doi:10.1177/2167702612469015.

[16] King G, Wand J. Comparing incomparable survey responses: Evaluating and selecting anchoring vignettes. Political Analysis. 2007 Jan 1:46-66.

[17] Kleinbaum G, Kupper L, Morgenstern H. Epidemiologic research: Principles and Quantitative Measures. Belmont, CA: Lifetime Learning Publishers; 1982.

[18] ECDPC. Field Epidemiology Manual by European Center for Disease Prevention and Control 2016. https://wiki.ecdc.europa.eu/fem/w/wiki/preventing-bias (accessed August 20, 2016).

[19] Hammer GP, du Prel JB, Blettner M. Avoiding bias in observational studies. Dtsch Arzteblatt

Int. 2009 Oct 9;106:664-8.

[20] Hassan E. Recall bias can be a threat to retrospective and prospective research designs. Internet J Epidemiol. 2005;3:2.

[21] Sun J, Bogie KM, Teagno J, Sun Y-HS, Carter RR, Cui L, et al. Design and Implementation of a Comprehensive Web-based Survey for Ovarian Cancer Survivorship with an Analysis of Prediagnosis Symptoms via Text Mining. Cancer Inform 2014;14:113–23. doi:10.4137/CIN.S18965.Received.

[22] DeWalt DA, Rothrock N, Yount S, Stone AA. Evaluation of item candidates: the PROMIS qualitative item review. Med Care 2007;45:S12-21. doi:10.1097/01.mlr.0000254567.79743.e2.

[23] Chow SC, Wang H, Shao J. Sample size calculations in clinical research. CRC press; 2007 Aug

[24] Mason W, Suri S. Conducting behavioral research on Amazon's Mechanical Turk. Behav Res Methods 2012;44:1–23.

[25] Patients Like Me. Patientslikeme and R.A.R.E. Project Unite to Find and Connect One Million Rare Disease Patients. Www.patientslikeme.com 2011:1. https://www.patientslikeme.com/press/20111107/35-patientslikeme-and-rare-project-unite-to-find-and-connect-one-million-rare-disease-patients (accessed August 6, 2016).

[26] Lintott C, Schawinski K, Bamford S, Slosar A, Land K, Thomas D, et al. Galaxy Zoo 1: data release of morphological classifications for nearly 900 000 galaxies. Mon Not R Astron Soc 2011;410:166–78.

[27] Center for Game Science at University of Washington UD of B. Foldit 2015. https://fold.it/portal/ (accessed July 1, 2015).

[28] Rogstadius J, Kostakos V, Kittur A, Smus B, Laredo J, Vukovic M. An Assessment of Intrinsic and Extrinsic Motivation on Task Performance in Crowdsourcing Markets. Fifth Int AAAI Conf Weblogs Soc Media 2011:321–8.

[29] Heer J, Bostock M. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. InProceedings of the SIGCHI Conference on Human Factors in Computing Systems 2010 Apr 10 (pp. 203-212). ACM.

[30] Mason W, Watts D. Financial incentives and the performance of crowds. ACM SigKDD Explor Newsl 2010.

[31] Zaidan OF, Callison-Burch C. Crowdsourcing translation: Professional quality from non-professionals. InProceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1 2011 Jun 19 (pp. 1220-1229). Association for Computational Linguistics.

[32] Warby SC, Wendt SL, Welinder P, Munk EG, Carrillo O, Sorensen HB, et al. Sleep-spindle detection: crowdsourcing and evaluating performance of experts, non-experts and automated methods. Nat Methods 2014;11:385–92.

[33] Felt P, Ringger E, Boyd-Graber J, Seppi K. Making the Most of Crowdsourced Document Annotations: Confused Supervised LDA. Conf Comput Nat Lang Learn 2015:194–203.

[34] Hirth M, Hoßfeld T, Tran-Gia P. Cost-optimal validation mechanisms and cheat-detection for crowdsourcing platforms. InInnovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2011 Fifth International Conference on 2011 Jun 30 (pp. 316-321). IEEE.

[35] Carter RR, DiFeo A, Bogie K, Zhang GQ, Sun J. Crowdsourcing awareness: Exploration of the ovarian cancer knowledge gap through amazon mechanical turk. PLoS One 2014;9. doi:10.1371/journal.pone.0085508.

[36] Trauzettel-Klosinski S, Dietz K. Standardized Assessment of Reading Performance: The New International Reading Speed Texts IReST. Invest Ophthalmol Vis Sci 2012;53:5452–61.

[37] ODonovan P, Agarwala A, Hertzmann A. Learning Layouts for Single-PageGraphic Designs. IEEE Trans Vis Comput Graph 2014;20:1200–13. doi:10.1109/TVCG.2014.48.

[38] Duggan M, Brenner J. The demographics of social media users, 2012. Washington, DC Pew Res Center's Internet Am Life Proj 2013;14.

[39] Haythornthwaite C. Social networks and Internet connectivity effects. Information, Community Soc 2005;8:125–47.

[40] Gundecha P, Liu H. Mining social media: a brief introduction. Tutorials Oper Res 2012;1.

[41] Cresci S, Pietro R Di, Petrocchi M, Spognardi A, Tesconi M. Fame for sale: efficient detection of fake Twitter followers. Decis Support Syst 2015;80:56–71.

[42] Mittal S, Kumaraguru P. Broker Bots: Analyzing automated activity during High Impact Events on Twitter 2014.

[43] Graber MA, Graber A. Internet-based crowdsourcing and research ethics: the case for IRB review. J Med Ethics 2012:Online.

[44] Ghosh D, Guha R. What are we 'tweeting'about obesity? Mapping tweets with topic modeling and Geographic Information System. Cartography and geographic information science. 2013 Mar 1;40(2):90-102.

[45] Twitter Firehose. http://support.gnip.com/apis/firehose/

[46] Rivers C, Lewis B. Ethical research standards in a world of big data [v2; ref 2014.

[47] Young SSD, Rivers C, Lewis B. Methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes. Prev Med (Baltim) 2014;63:112–5. doi:10.1016/j.ypmed.2014.01.024.

[48] Widener MJ, Li W. Using geolocated Twitter data to monitor the prevalence of healthy and unhealthy food references across the US. Appl Geogr 2014;54:189–97. doi:10.1016/j.apgeog.2014.07.017.