The use of crowdsourcing and the role of game mechanics in identifying erroneous disease burden estimates

Michael K. Freeman

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Public Health

University of Washington

2013

Committee:

Abraham Flaxman

Christopher Adolph

Peter Speyer

Program Authorized to Offer Degree:

School of Public Health – Department of Global Health

University of Washington

Abstract

The use of crowdsourcing and the role of game mechanics in identifying erroneous disease burden estimates

Michael K. Freeman

Chair of Supervisory Committee:

Assistant Professor Abraham Flaxman

Global Health

Objectives: This project evaluates the feasibility of using crowdsourcing techniques for error detection in public health research. Secondarily, it estimates the effect of gamification on the accuracy and volume of disease burden classifications.

Methods: The Global Burden of Disease 2010 estimates served as a database for this project. Algorithms were used to identify potentially erroneous estimates. Two user interfaces (one gamified) were designed to collect non-expert disease burden classifications, which were compared against a GBD expert.

Results: The 43 participants classified 1,114 health trends using the web interface. Of these, 86% of responses matched the classification of a GBD expert, yielding a sensitivity of .71, and a specificity of .89. The presence of a gamified environment increased usage by 70%.

Conclusions: Non-experts were able to accurately classify disease trends. The use of crowdsourcing may be applicable to similarly large databases, and gamification can drive increased use.

Acknowledgements

The success of this project and the completion of my degree could not have happened without the love, support, and guidance of many people in my life. The past three years were an incredible experience, and I would like to extend my gratitude to these people in particular:

To Abie, for challenging me to push myself creatively and technically,

To Peter Speyer and Chris Adolph, for their help in directing and designing the project,

To Emm, for her years of advice, both personal and professional,

To the 2010 PBF Cohort, for navigating the last 3 years alongside me,

To my parents, for their unwavering love and support, and again,

To Emma, for everything.

Background

Data

The Global Burden of Disease (GBD) 2010 study produced estimates of disease burden by age and sex for 291 diseases and 67 risk factors for 187 countries between 1990 and 2010[6,9,16]. Including 4 different metrics for measuring disease burden, there are millions of results to consider. The study required massive computational power, as well as the development of interactive tools to explore the estimates. These innovative tools facilitated the rapid review of the results, but there was no automated and systematic mechanism though which unlikely estimations could be detected. While algorithms could be used to scan the database, the health field is unique in that many dramatic differences in disease burden may be expected due to rapid improvement or deterioration in health conditions. For example, an HIV burden 5,000 times larger in 2010 than 1990 would be expected for some countries (even given limited health knowledge). To improve upon the use of algorithms alone, this project seeks to implement human-based computation (HBC) in the form of crowdsourcing to identify unlikely estimates in the GBD 2010 results.

Crowdsourcing Overview

HBC is a technique through which a machine harnesses the power of human knowledge by outsourcing tasks to individuals[11,14]. There are many tasks which computers are ill-suited to perform because of nuanced, non-systematic knowledge requirements. The particular approach used for this project is crowdsourcing, which solicits opinions from a broad audience on a particular topic. The crowdsourcing function used in this project is what Daren Badham describes as *distributed human intelligence tasking*, in which crowd knowledge is used to

categorize or analyze information[1]. A notable example of this is the use of crowdsourcing for image labeling. Von Ahn and Dabbish introduced a "new interactive system" in which an online gaming environment provided entertainment while generating useful information that described images[15]. By designing an environment in which users attempted to match others' descriptions of images, the authors were able to design a system far more efficient and accurate than alternate techniques such as computer vision.

Distributed human intelligence tasking has also been attempted with more complex tasks. For example, Cooper et. al. explored the ability of using a crowdsourcing system for predicting protein structures with their game Foldit[3]. Because humans still have superior 3D analytical skill compared to machines, this project made many advancements in understanding protein folding. Numerous fields are beginning to realize the strengths and applicability of crowdsourcing, including health and medicine.

Crowdsourcing in Health

In 2013, Ranard et. al. published a systematic review of crowdsourcing approaches that had been applied to health and medicine research. Using the World Health Organization's definition of health as "a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity"[10], the authors identified 21 peer reviewed journal articles that applied crowdsourcing techniques to the field of health[12]. Of these 21 studies, the authors classified 5 of them into the fields of either epidemiology or public health. Two of these studies used crowdsourcing for disease surveillance, the first of which used Healthmap technology to map influenza outbreaks[4]. The second health surveillance study used Amazon's Mechanical Turk application to solicit self-reports of malaria diagnosis in India[2]. Another two studies performed

data collection via crowdsourcing. The first study had participants analyze photographs to determine if the construction of a bike lane had an effect of transportation choices[5]. The second study created a challenge with a $10,000 prize to incentivize individuals to contribute to a comprehensive mapping of automated external defibrillators in Philadelphia[8]. The final public health study was used for content analysis to evaluate the effectiveness of oral health promotional materials[13].

All of these studies harnessed the power of a large volume of participants executing specific tasks that humans are better suited to perform than machines. Often, this approach functions as an excellent complement to using expert opinion. For example, in their research about using lay persons' opinions in making medical diagnoses, Mavandadi et. al. argue that crowdsourcing can help distill the volume of information necessary to be reviewed by experts, which is particularly important in resource poor settings[7]. This study applies a similar philosophy for the identification of erroneous disease burden estimates. While expert opinion is necessary for validation of epidemiological trends, non-expert review can help distill the number of relationships to be reviewed by experts. Given the requisite of human knowledge in evaluating health trends, the set of GBD results is well suited for distributed human intelligence tasking. Moreover, because there is insufficient time for experts to review all results, using algorithms to limit the number of results to review, and parsing out evaluations across many individuals is an efficient way to assess the validity of disease burden estimates.

Methods

Procedure

This study solicited non-expert opinion on disease trends in order to assess the feasibility of using crowdsourcing to identify errors, which required four distinct tasks. First, a database of comparisons of disease burden either between sexes or over time was designed to highlight epidemiologically significant causes with potentially erroneous differences. Of these trends, 25 were manually selected to be evaluated by a GBD expert and classified as either accurate, or potentially erroneous. Second, a user interface (UI) was built to solicit trend classifications from users in an interactive web environment. Two versions of the interface were designed and randomly distributed to participants to evaluate the effect of gamification elements on user participation. Third, a pilot test was undertaken in which participants were asked to use the UI to classify the 25 trends which had been screened for potential anomalies by a GBD expert. These participants were randomly assigned a version of the UI with or without gamified elements. Finally, the extent of use and accuracy of responses were analyzed and validated against expert opinion.

Comparison Selection

In order to evaluate health trends in the GBD database, comparisons were made both between sexes and over time. A database of trends was constructed in which each observation was a comparison of disease burden as measured by deaths, years of life lost (YLLs), years lived with disability (YLDs), or disability adjusted life years (DALYs). There were two types of comparisons made in the database. First, comparisons were made between sexes, in which the discrepancy between male and female disease burden was evaluated for a given country-year-age-disease. An example comparison between sexes is, the DALY rate of interpersonal violence was 8.5 times higher in men than women in Mexico[9]. The second type of comparison was over time, in which the difference in disease burden was evaluated between 1990 and 2010 for a given

country-age-sex-disease. For example, for males in Canada, the DALY rate of diabetes was 1.5 times higher in 2010 than 1990[9].

In order to select disease burdens to be classified, two dimensions of each comparison were considered. First, the size of a disease's burden (ie, the number of deaths, YLLs, YLDs or DALYS) was ranked against all other burdens within a demographic category (country, age, metric, and year or sex). The purpose of this was to identify diseases which were of high public health significance. The second dimension of each comparison evaluated was the ratio of the two values (ie, diseases burden in 2010 divided by disease burden in 1990, or disease burden in males divided by disease burden in females). This measure was intended to detect dramatic, and therefore potentially unlikely, differences between sexes or over time. To use an example, the death rate of measles in India was 4 times lower in 2010 than 1990 (all ages, both sexes)[6]. The total number of measles deaths, and the ratio of deaths in 2010 and 1990 would each be ranked against other diseases for all ages both sexes in India. The sum of these two ranks was then ranked against other trends to produce a combined measure of importance given both size and ratio. The equation below shows the rank of comparison C, where the final rank is defined as the rank of the sum of the size and ratio ranks.

$$Rank_C = Rank\big(Rank(C_{size}) + Rank(C_{ratio})\big)$$

In order to limit the number of comparisons which were evaluated in this study, only the top 10% of ranks for a given demographic category were included in the database. All relationships were compiled and stored in a MySql database where the unique identifiers were country, year, age, sex, metric, cause, and comparison type (by sex, or over time).

Interface design

A web platform was developed so users could view trends and interactively classify them. Upon opening the page, users were provided with one of two brief sets of instructions. Figure 1 shows the welcome view for both the gamified (top) and standard (bottom) environments.

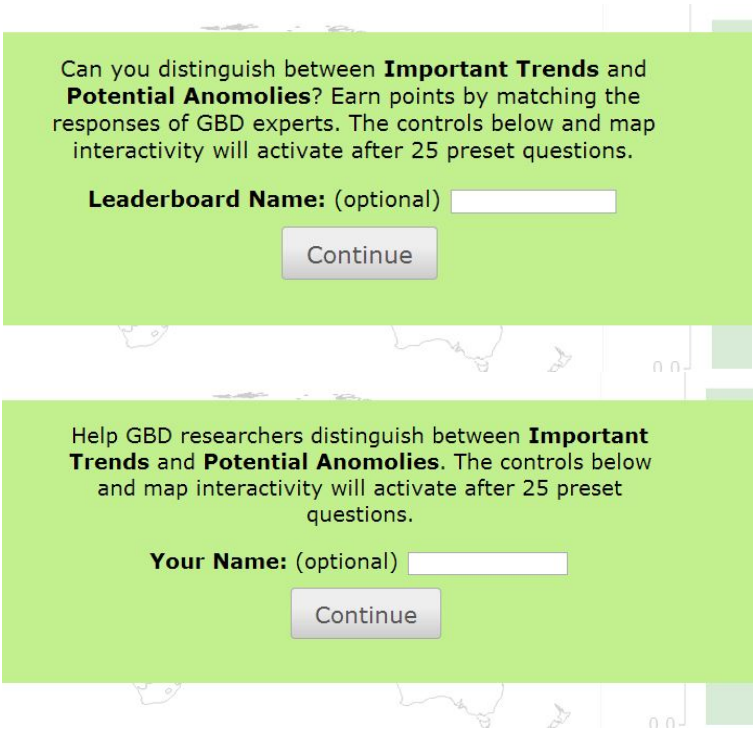

Figure 1: Website greeting (top: gamified, bottom: standard)

The gamified version differs in 3 ways. First, it challenges users to accurately classify trends, indicating a measure of skill. Second, it presents the opportunity for earning points based on performance. Third, it suggests that they enter their name to be tracked on a leaderboard (insinuating competition with other users).

Following the brief introduction, users are presented with one relationship at a time and asked to categorize it as either an “important trend” or a “potential anomaly”. This terminology

was intended to refer to estimation errors without framing the research as inaccurate. Figure 2 shows the categorization section for both versions of the tool.

Figure 2: Health trend classification (top: standard, middle: gamified, bottom: gamified with bonus)

Here, the top section shows the standard version. The middle section (gamified version) includes a counter showing the number of questions answered, and the number of points earned. There is also button which will display the leaderboard table. The bottom section of the figure shows that when a user's response matched that of a GBD expert, they were flashed a bonus of additional points. Users earned 5 bonus points for matching important trends, and 10 bonus points for matching potential anomalies. As an additional incentive, users were awarded a bonus for streaks of correct answers. These were awarded at streaks which were multiples of five (ie, 5 correct, 10 correct), and were equal to 10 times the streak length.

In the main section of the screen, a map zooms in on the selected country and distinguishes it with dark green shading (Figure 3). Mousing over each country reveals its name, and (following the 25

Figure 3: Interactive map

preset questions), clicking a country allows users to choose the location to investigate. Before 25 questions are complete, both the controls and the map interactivity are disabled.

The right panel of the page shows a chart depicting the trend either between sexes (bar chart), or over time (line graph). While the relationship stated in the top panel is phrased in relative terms (ie, the burden of disease D was N times higher…), the chart is included to communicate absolute numbers. Figure 4 shows the bar chart representing the discrepancy in interpersonal violence in Mexico between men and women. Figure 4 also displays an example of the line chart that shows the number of DALYs attributable to the adverse effects of medical treatment in Cuba over time.

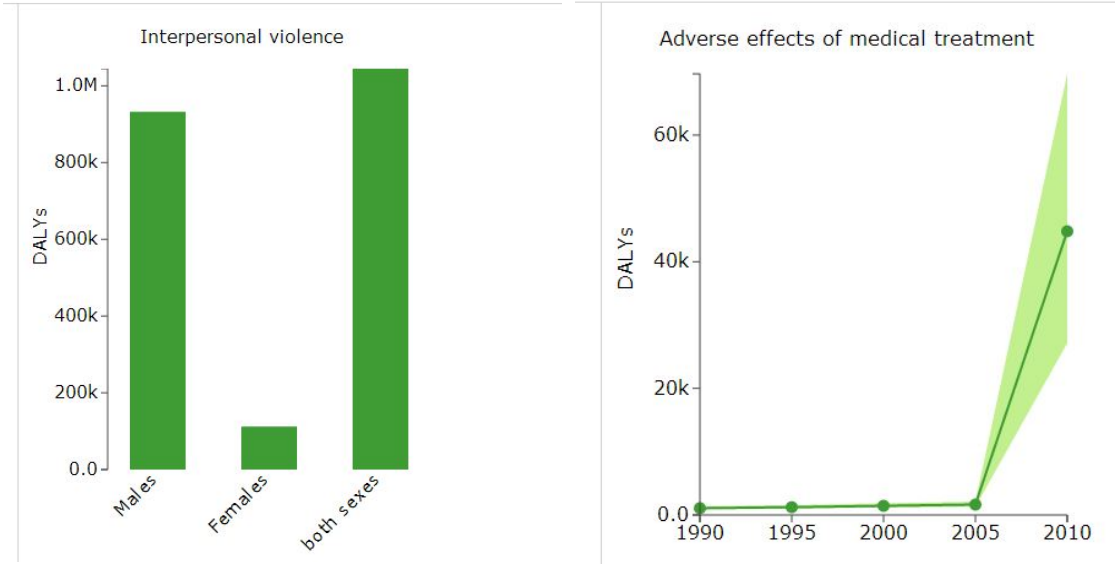

Figure 4: Bar and line charts

At the bottom of the screen were the controls which could be used to select demographic categories of interest. At the bottom right, users selected the type of comparison being shown (by sex, or over time). Depending on the comparison type, the option to select sex or year appears with the other controls. Four dropdown menus allow users to select a country, age,

metric, and year or sex. Each demographic control menu has a lock icon which can be set to focus on the demographic category (locked), or let it vary randomly (unlocked). The Javascript libraries D3 and jQuery were used to build the interactive elements of the page.

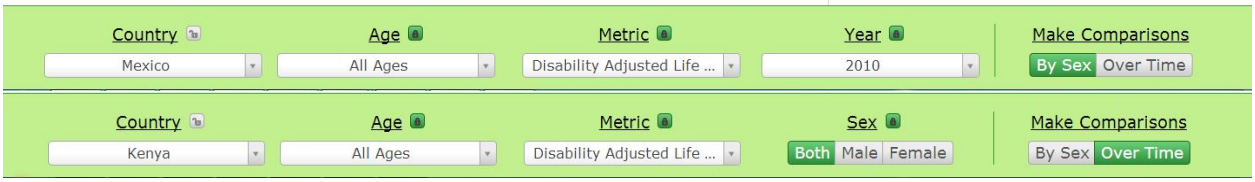

Figure 5: Controls (top: by sex, bottom: over time)

Technical Description

The tool shown above required the use of multiple programming languages to incorporate elements of interactivity, design, transitions, and database querying. Table 1 shows the list of programming languages and libraries that were use used, along with their primary function in building the tool.

Table 1: Programming languages used

Language/Library	**Primary Function**
Cascading Style Sheets (CSS)	Style the page (ie, object size and color schema)
Data Driven Documents (D3)	JS library used to builds charts, execute transitions
HyperText Markup Language (HTML)	Read in script files, assert initial page elements
JavaScript (JS)	Design functionality of page (primary language)
jQuery	JS library used to style controls, handle elements
PHP: Hypertext Preprocessor (PHP)	Used to get connect to servers and execute SQL
Select2	JS library for styling menus
Structured Query Language (SQL)	Used to query the MySql databases

To enable interactive elements of the page, all buttons and menus were constructed as HTML select objects, and then stylized using the select2 library. Figure 6 shows an example of how the year select menu was constructed in JS code (note the use of jQuery ($) for selecting and manipulating elements). Other menus followed a similar approach.

```
// Add a select menu
var yearSelect = $('<select>').appendTo('#controls).attr("id",
                  "ySelect");

// Add each year as an option to the menu
i.metadata.years.map(function(d) {
      yearSelect.append($("<option>").attr('value',d).text(d));
})

// Declare the #yearSelect menu as a select2 menu, provide parameters
$('#ySelect').select2({
      minimumResultsForSearch: 10,
      width: '250px'
});

// Define function to execture when the menu changes
$('#ySelect').change(function() {
      var year = $(this).val()
      i.settings.year = year
      i.update()
})
```

Figure 6: Code for making selection menu

The styles of the page were defined using CSS. A primary challenge of styling was defining a layout where elements could be precisely aligned, such as in the control bar (see Figure 5). To accomplish this, all controls were wrapped in an HTML “div” element (arbitrary grouping of content). As an added layer of organizational layout, each individual control was placed inside of its own HTML div to allow their alignment to be more easily manipulated. By placing these divs inside of a container whose text was centered, these elements were easily

aligned. Figure 7 shows the CSS code which was used to style the controls elements at the bottom of the page.

```
/* Bottom section of the page */
#footer {
      text-align:center;
      background:#c3f08e;
      position:absolute;
      bottom:0;
      width:100%;
      padding:10px;
      border-top:solid #3f9c35 1px;
}

/* Div element that contains all of the controls. */

#optionBox {
      display:inline-block;
}

/* Div elements which contain each controls menu.  */

#metricBox,#ageBox,#sexBox,#yearBox,#dimBox,#countryBox{
      float:left;
      width:264px;
      margin-right:10px;
      margin-top:10px;
}
```

Figure 7: Code for formatting styles

Chart and map transitions were handled by using the D3 JavaScript library. For example, when a new health trend was shown, the heights of the rectangles in the bar chart and the corresponding scale would both update smoothly to reflect the new relationship. In order to do this, the D3 library was first used to build the graphic by appending rectangles to a scalable vector graphic (SVG) object on the page. This is a very powerful ability of the library that binds data to the DOM elements which represent them. In this process, each rectangle is assigned descriptive attributes that describe its representation on the page (ie, height, width, color, etc.). These attribute assignments are executed by handler functions to return the proper value for each

rectangle given an array of data that D3 is passed. The D3 library is then easily able to update these attributes and transition the graphic based on new values. Figure 8 contains two pieces of sample code. The first section of code builds the initial rectangles for the bar chart, and the second section updates the height of the bars. The updating code simply initiates a transition, and then updates the attributes which define the height of each rectangle.

```
// Append rectangles to SVG
i.bars = i.chartSvg.selectAll('rect')
           .data(d3.keys(i.metadata.sexMap))
           .enter().append('rect')
           .attr('x', function(d) {return i.xScale(d.x)})
           .attr('y', function(d){return  i.yScale(d.y)})
           .attr('height', function(d) {return (chart.height –
                i.yScale(d.y})
           .attr('width', i.settings.bar.width)
           .attr('id', function(d) {return d.id})
           .attr('fill', '#3f9c35')
           .attr('class', 'rect')

// Update bar heights
i.bars.transition().duration(500)
           .attr('y', function(d){return  i.yScale(d.y)})
           .attr('height', function(d) {return (chart.height –
                i.yScale(d.y})
```

Figure 8: Code for making, updating charts

As described above, all health trends were stored in a MySql database, each with a unique identifier. Because the first 25 questions were preset, they were accessed from the database using their unique identifying codes that were explicitly stored in the JavaScript code. Following the first 25 questions, participants were shown trends with the highest cause rank for a given demographic category. The database contained over 1.75 million relationships, which required designating country as an index key for rapid querying of the database. To accommodate this, it was required to specify the country for every query. If a user did not select

a location of interest, a country was randomly selected from the list of countries in JavaScript, and this was passed to a PHP script which executed the MySql query. This is shown in Figure 9.

```
// Function for choosing a random country
i.randomCountry = function() {
      var rand = Math.floor(Math.random()*d3.keys(i.locations).length)
      var iso3 = d3.keys(i.locations)[randInt]
      return iso3
}

// Pass a random country if none is selected
var getCountries = i.country != 'none'? i.country:i.randomCountry()

// From JavaScript, use jQuery to call php file to get a trend
$.ajax({
      url: 'php/getQuestion.php',
      dataType: 'json',
      cache:false,
      data:{
            age:getAges,
            sex:getSex,
            metric:getMetrics,
            comp_dim:i.settings.comp_dim,
                        year:getYears,
                        facts:getFacts,
                        country:getCountries,
                  },
      success: function(response){
                  // Process response data here
      },
      error:  function(response,error) {
                  // Process error here
      }
})
```

Figure 9: Code for passing parameters to PHP

PHP was use to get health trends from the database, as well as track the trend classifications made by participants. Upon clicking one of the classification buttons ("Interesting Trend", or "Potential Anomaly"), a PHP script pushed the user's classification into a MySql database of responses. Figure 10 shows the PHP code that is called from JavaScript (as was

done in Figure 9 above). PHP is able to access parameters passed from the JavaScript function, and use those variables in the desired SQL query.

```
<?php
// Get parameters passed from JavaScript
$id = array($_GET['id']);
$vote =  array($_GET['vote']);
$voter = array($_GET['voter']);
$name = array($_GET['name']);
$time = array($_GET['time']);

// SQL code to insert variables into table, represented as ?
$sql = <<<SQL
          INSERT INTO responses(id,vote,voter,name,time)
          values(?,?,?,?,?)
SQL;

// Array of values to pass to sql statement
$values = array();
$values = array_merge($values,$id);
$values = array_merge($values,$vote);
$values = array_merge($values,$voter);
$values = array_merge($values,$name);
$values = array_merge($values,$time);

// Prepare and execute sql code
$query = $dbs['mikefree']['conn']->prepare($sql);
$query->execute(array_values($values));

?>
```

Figure 10: PHP code for pushing into MySql Database

While the integration of multiple languages and libraries presented a technical challenge, it allowed the development of a flexible web-tool that accessed and recorded information in databases.

Pilot test

Staff members at the Institute for Health Metrics and Evaluation (IHME) voluntarily participated in a test of the tool. To assess disease burden classification accuracy of a non-expert audience,

25 pre-selected trends were shown to each user, and user responses were compared to GBD expert opinion. Following the completion of these 25 questions, users were able to continue classifying disease trends, and were also able to choose which demographic categories to learn about by using the controls and interactive map. To assess the impact of gamification on accuracy and duration of use, the gamification elements described above were randomly assigned to users based on the exact time the tool was opened. This data can be used to evaluate the ability of users to correctly classify unlikely relationships, the effects of gamification elements on user engagement, and the overall usability of the tool. Qualitative feedback was also gathered to assess general usability.

Data Analysis

Site use analytics were used to assess the feasibility of such tools in the future. This study looked specifically at the number of trends classified per user and three performance metrics on the 25 pre-selected questions. These metrics included concordance with expert opinion, as well as specificity and sensitivity with respect to identifying potential anomalies. Using this information, Bayes' rule can be implemented to calculate the probability of each categorization based on user feedback. The equation below shows the probability that a relationship is likely given that an individual classifies it as likely. In the analysis section below, various levels of prior likelihood were assessed.

$$P(Likely\,|Vote_L) = \frac{P(Likely)P(Vote_L|Likely)}{P(Likely)P(Vote_L|Likely) + P(Unlikely)P(Vote_L|Unlikely)}$$

In order to evaluate the effect of gamification elements, t-tests were used to measure the statistical difference in the performance metrics mentioned above. This determined if

gamification elements were associated with a different level of accuracy, or an increased number of relationships classified per user.

Results

Descriptive Results

Forty three participants classified 1,114 trends using the web tool. Overall, participants answered an average of 26 questions, and matched the classification of a GBD expert 86% of the time. With respect to identifying errors, users had a sensitivity of .71, and a specificity of .89. Figure 11 shows the distributions of each performance metric across users.

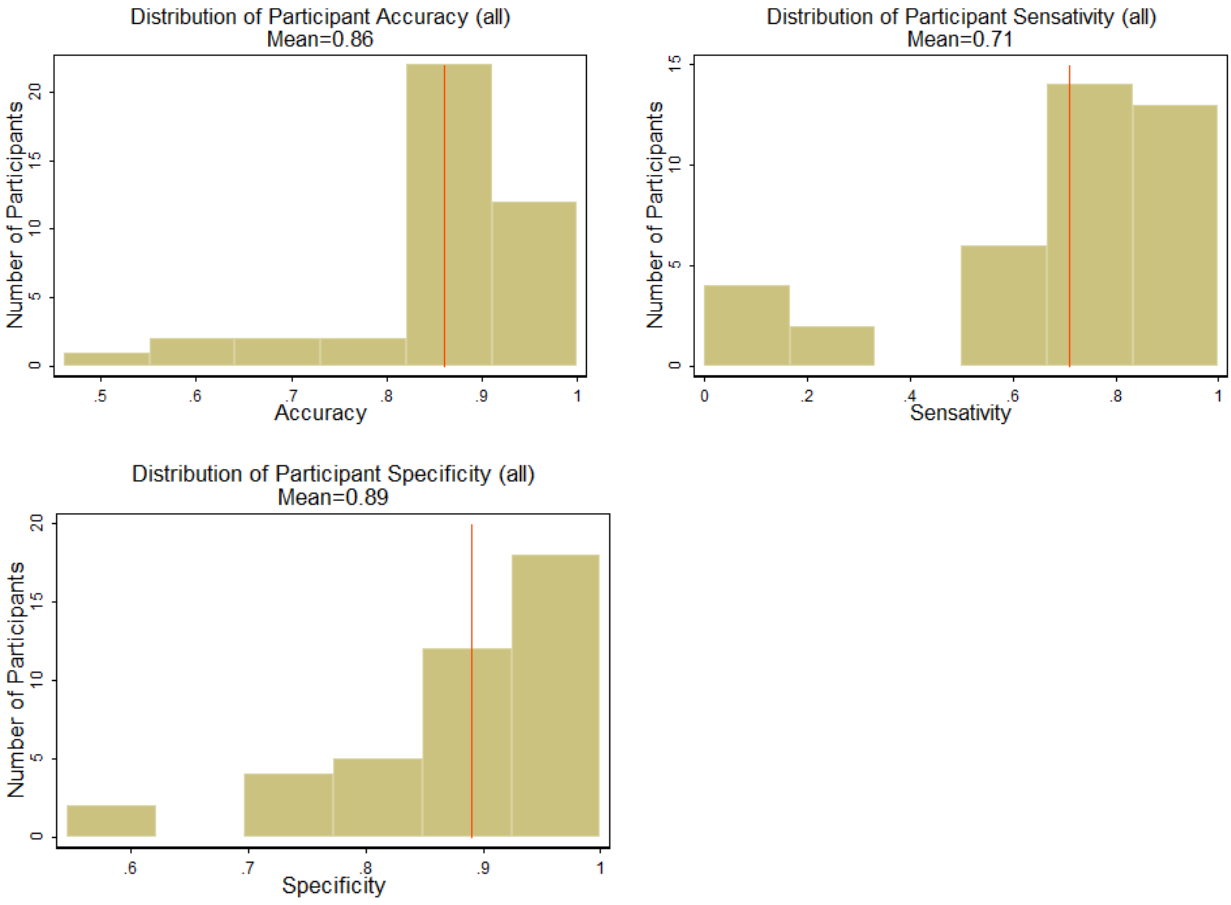

Figure 11: Distribution of performance metrics across users

Bayesian Analysis

On average, the probability that a trend is classified an outlier if it is truly as an outlier was equal to the specificity (.89). The probability that a trend is not classified as an outlier if it is truly not an outlier was equal to the sensitivity (.71). Bayes' rule was used to compute the conditional probabilities of interest. Table 2 shows the full set of results from all users.

Using this table as an average user performance, one can compute the conditional probabilities of interest. Figure 12 shows how the prior probability of being correct determines the posterior likelihood of a relationship being true or false (based on one classification):

	Responses		
Truth	Likely	Unlikely	**Sum**
Likely	630	80	710
Unlikely	41	100	141
Sum	671	180	851

Table 2: Distribution of user responses

The figure shows a horizontal line at 95% to represent a scenario in which a relationship is only

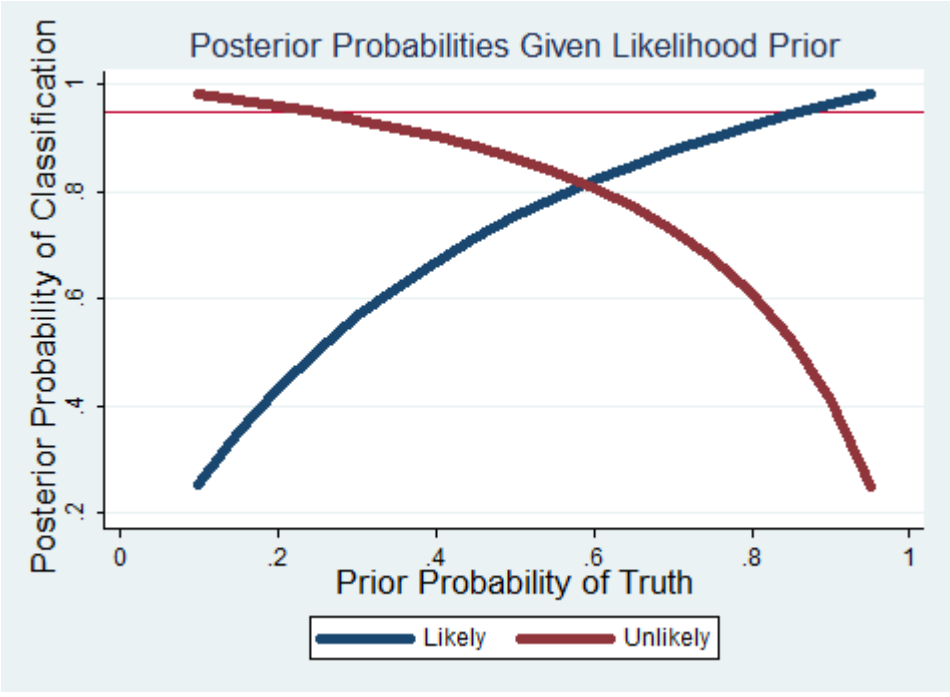

Figure 12: Posterior probabilities of classification given prior likelihoods

reviewed by a researcher when 95% confidence is reached. Based on this analysis, if the prior probability of a trend being likely is greater than .85, the posterior probability of being likely will exceed 95% based on the average user's endorsement. Conversely, if the prior probability of being likely is below 26%, the posterior probability of being unlikely will exceed 95% following one unlikely endorsement from an average user.

Figure 13 expands the Bayesian calculation to incorporate having multiple classifications. Assuming a prior of .9 for each relationship being likely, this figure shows the posterior probability of a trend being potentially anomalous given multiple unlikely classifications. After 3 consecutive unlikely endorsements, a given relationship crosses the 95% threshold, warranting review by an expert.

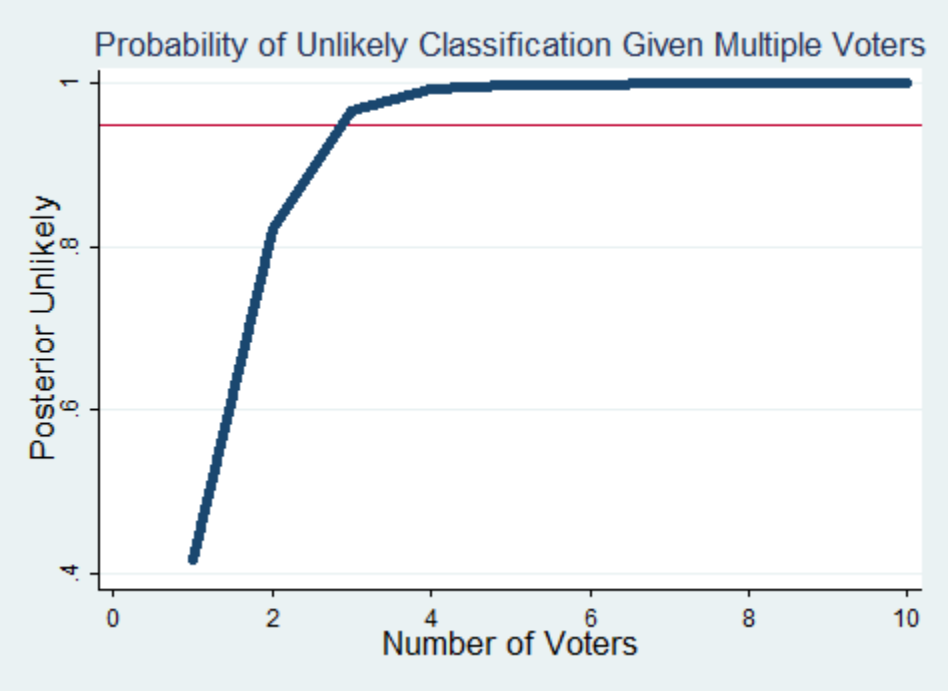

Figure 13: Effect of multiple votes of posterior probability

Effect of Gamification

The performance of users as measured by accuracy, sensitivity, and specificity was remarkably similar for individuals with and without the gamified interface, as shown in Table 3. However, those provided with a gamified interface classified 1.7 times more trends than those with the standard interface. A t-test confirmed that this was a statistically significant difference in the number of trends classified, with a p-value of .01.

Table 3: Performance metrics by gamification

Version	Avg. Num. Classified	Sensitivity	Specificity	Accuracy
Standard	18.39	0.72	0.88	0.86
Game	31.32	0.7	0.89	0.86
All	25.91	0.71	0.89	0.86

Discussion

Conclusions

This study assessed the feasibility of using crowdsourcing techniques for identifying outliers in global health data. There are 3 primary conclusions to draw from this study. First, non-experts were generally able to match expert classifications of disease burden with respect to outlier detection. In a field that relies heavily on expert judgment of feasibility of estimations, crowdsourcing may be a tool to reduce the volume of manual reviews done by experts. Second, the Bayesian analysis demonstrated that the quality of responses can be measured with a small number of pre-selected questions. This is a refined approach to crowdsourcing in which the quality of respondents can vary. Based on these performances, posterior probabilities of a trend

being an outlier can be calculated. This approach provides a quantitative threshold that determines when a relationship needs to be reviewed by an expert. Finally, this study demonstrated that the inclusion of basic gamification elements does not impact the quality of responses, but does increase the number of responses solicited per user. Gamification therefore represents an approach to drive increased use without sacrificing quality responses.

Limitations

There were a number of limitations to this pilot study. First, the relationships shown to users were not randomly selected. As a result, they may not represent the full spectrum of relationships which exist in the database. Additionally, because this study used convenience sampling of individuals in the health field to gather results, the respondents will not be representative of a broader population. Because the participants all worked in the same office, the motivation associated with competition may be overestimated compared to the general public. Furthermore, the crowdsourcing nature of the project was not properly evaluated. A full public release of such a tool is required to assess the plausibility of truly using crowdsourcing to generate public interest and sufficient responses. Finally, validation of results also hinges on expert opinion of GBD researchers. While these experts are very knowledgeable about their fields, their opinions are not necessarily reflective of disease patterns, particularly in data-sparse areas.

Future Research

Despite the limitations of this study, it presents promising evidence that the public health field can harness the power of crowdsourcing to improve research practices. Similarly to image description, measuring the face-validity of global health trends is still a computational challenge

that is not difficult for humans. Additionally, the volume of estimates in global health studies increases the demand for individuals to assess the reasonability of estimates. In the future, more nuanced algorithms can be developed to identify particular types of errors. This study assumed that disease burden levels should be similar both between genders, and over time. Using the GBD results, this type of assumption could be expanded in many ways. For example, one might assume that disease burden levels should be similar between geographically proximal or socio-economically similar countries. Differences between similar countries (or similarities between very different countries) could indicate error. Relationships could also be constructed based on the ratio between death and disability for diseases. These algorithms could have heightened nuance such that different expectations exist for different categories of disease. For example, the deaths of communicable diseases might be expected to be much higher than their disability given short duration of disease and high probability of death. All of these prior expectations and others could be used to design a database of more types of relationships to be reviewed (which would also prove to be more entertaining).

In the future, more research into the specific effects of gamification on classification quality is required. In this study, individuals were rewarded for matching the responses of GBD experts. While this may have incentivized accurate disese classification, it may have also biased results. If participants are rewarded for mirroring expert opinion, many anomalies may be overlooked. Additionally, the types of relationships being shown could be reconsidered. If participants are only being shown large differences in large causes, perhaps they will become insensitive to outliers in the database.

More generally, the combination of algorithms with crowdsourcing represents an incredible opportunity in public health research. This study has shown that non-experts can

provide valuable reviews of the validity of results, decreasing expert workload and improving estimates. Using crowdsourcing will demand a high level of transparency by exposing the least likely conclusions of a study. This heightened level of openness in the research process could facilitate a more collaborative research culture. Moreover, the inclusion of non-expert judgment in the research process allows for larger dissemination of the research. These benefits contribute to higher quality, higher impact research. Researchers should consider the inclusion and value of non-expert review as a part of the research process, and crowdsourcing provides a framework and structure for gathering such information.

References

1. Brabham, Daren C. *Crowdsourcing*,Cambridge: MIT Press, 2013.

2. Chunara, Rumi, Vina Chhaya, Sunetra Bane, Sumiko R. Mekaru, Emily H. Chan, Clark C. Freifeld, and John S. Brownstein. "Online Reporting for Malaria Surveillance Using Micro-monetary Incentives, in Urban India 2010-2011." *Malaria Journal* 11, no. 1 (February 13, 2012): 43. doi:10.1186/1475-2875-11-43.

3. Cooper, Seth, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, and Zoran Popovic. "Predicting Protein Structures with a Multiplayer Online Game." *Nature* 466, no. 7307 (August 5, 2010): 756–760. doi:10.1038/nature09304.

4. Freifeld, Clark C., Rumi Chunara, Sumiko R. Mekaru, Emily H. Chan, Taha Kass-Hout, Anahi Ayala Iacucci, and John S. Brownstein. "Participatory Epidemiology: Use of Mobile Phones for Community-Based Health Reporting." *PLoS Medicine* 7, no. 12 (December 2010). doi:10.1371/journal.pmed.1000376.

5. Hipp, J. Aaron, Deepti Adlakha, Amy A. Eyler, Bill Chang, and Robert Pless. "Emerging Technologies: Webcams and Crowd-Sourcing to Identify Active Transportation." *American Journal of Preventive Medicine* 44, no. 1 (January 2013): 96–97. doi:10.1016/j.amepre.2012.09.051.

6. Lozano, Rafael, Mohsen Naghavi, Kyle Foreman, Stephen Lim, Kenji Shibuya, Victor Aboyans, Jerry Abraham, et al. "Global and Regional Mortality from 235 Causes of Death for 20 Age Groups in 1990 and 2010: a Systematic Analysis for the Global Burden of Disease Study 2010." *The Lancet* 380, no. 9859 (December 2012): 2095–2128. doi:10.1016/S0140-6736(12)61728-0.

7. Mavandadi, Sam, Stoyan Dimitrov, Steve Feng, Frank Yu, Uzair Sikora, Oguzhan Yaglidere, Swati Padmanabhan, Karin Nielsen, and Aydogan Ozcan. "Distributed Medical Image Analysis and Diagnosis through Crowd-Sourced Games: A Malaria Case Study." *PLoS ONE* 7, no. 5 (May 11, 2012). doi:10.1371/journal.pone.0037245.

8. Merchant, Raina M, David A Asch, John C Hershey, Heather M Griffis, Shawndra Hill, Olivia Saynisch, Alison C Leung, et al. "A Crowdsourcing Innovation Challenge to Locate and Map Automated External Defibrillators." *Circulation. Cardiovascular Quality and Outcomes* 6, no. 2 (March 1, 2013): 229–236. doi:10.1161/CIRCOUTCOMES.113.000140.

9. Murray, Christopher J L, Theo Vos, Rafael Lozano, Mohsen Naghavi, Abraham D Flaxman, Catherine Michaud, Majid Ezzati, et al. "Disability-adjusted Life Years (DALYs) for 291 Diseases and Injuries in 21 Regions, 1990–2010: a Systematic Analysis for the Global Burden of Disease Study 2010." *The Lancet* 380, no. 9859 (December 2012): 2197–2223. doi:10.1016/S0140-6736(12)61689-4.

10. Preamble to the Constitution of the World Health Organization as adopted by the International Health Conference, New York, 19-22 June, 1946; signed on 22 July 1946 by the representatives of 61 States (Official Records of the World Health Organization, no. 2, p. 100) and entered into force on 7 April 1948.

11. Quinn, Alexander J., and Benjamin B. Bederson. "Human Computation: a Survey and Taxonomy of a Growing Field." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1403–1412. CHI '11. New York, NY, USA: ACM, 2011. doi:10.1145/1978942.1979148.

12. Ranard, Benjamin L, Yoonhee P Ha, Zachary F Meisel, David A Asch, Shawndra S Hill, Lance B Becker, Anne K Seymour, and Raina M Merchant. "Crowdsourcing-Harnessing the Masses to Advance Health and Medicine, a Systematic Review." *Journal of General Internal Medicine* (July 11, 2013). doi:10.1007/s11606-013-2536-8.

13. Turner, Anne M, Katrin Kirchhoff, and Daniel Capurro. "Using Crowdsourcing Technology for Testing Multilingual Public Health Promotion Materials." *Journal of Medical Internet Research* 14, no. 3 (2012): e79. doi:10.2196/jmir.2063.

14. Von Ahn, L. "Human Computation." In *46th ACM/IEEE Design Automation Conference, 2009. DAC '09*, 418–419, 2009.

15. Von Ahn, Luis, and Laura Dabbish. "Labeling Images with a Computer Game." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 319–326. CHI '04. New York, NY, USA: ACM, 2004. doi:10.1145/985692.985733.

16. Vos, Theo, Abraham D Flaxman, Mohsen Naghavi, Rafael Lozano, Catherine Michaud, Majid Ezzati, Kenji Shibuya, et al. "Years Lived with Disability (YLDs) for 1160 Sequelae of 289 Diseases and Injuries 1990–2010: a Systematic Analysis for the Global Burden of Disease Study 2010." *The Lancet* 380, no. 9859 (December 2012): 2163–2196. doi:10.1016/S0140-6736(12)61729-2.