Using the wisdom of the crowds to find critical errors in biomedical ontologies: a study of SNOMED CT

RECEIVED 23 April 2014
REVISED 5 September 2014
ACCEPTED 15 September 2014
PUBLISHED ONLINE FIRST 23 October 2014

Jonathan M Mortensen[1,2], Evan P Minty[2,3], Michael Januszyk[2], Timothy E Sweeney[1], Alan L Rector[4], Natalya F Noy[1,5], Mark A Musen[1,2]

ABSTRACT

Objectives The verification of biomedical ontologies is an arduous process that typically involves peer review by subject-matter experts. This work evaluated the ability of crowdsourcing methods to detect errors in SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms) and to address the challenges of scalable ontology verification.
Methods We developed a methodology to crowdsource ontology verification that uses micro-tasking combined with a Bayesian classifier. We then conducted a prospective study in which both the crowd and domain experts verified a subset of SNOMED CT comprising 200 taxonomic relationships.
Results The crowd identified errors as well as any single expert at about one-quarter of the cost. The inter-rater agreement (κ) between the crowd and the experts was 0.58; the inter-rater agreement between experts themselves was 0.59, suggesting that the crowd is nearly indistinguishable from any one expert. Furthermore, the crowd identified 39 previously undiscovered, critical errors in SNOMED CT (eg, 'septic shock is a soft-tissue infection').
Discussion The results show that the crowd can indeed identify errors in SNOMED CT that experts also find, and the results suggest that our method will likely perform well on similar ontologies. The crowd may be particularly useful in situations where an expert is unavailable, budget is limited, or an ontology is too large for manual error checking. Finally, our results suggest that the online anonymous crowd could successfully complete other domain-specific tasks.
Conclusions We have demonstrated that the crowd can address the challenges of scalable ontology verification, completing not only intuitive, common-sense tasks, but also expert-level, knowledge-intensive tasks.

Key words: crowdsourcing, biomedical ontology, ontology engineering, SNOMED CT

RESEARCH AND APPLICATIONS

OBJECTIVE

Work in biomedicine to retrieve, integrate, and analyze datasets often requires domain-specific ontologies that define the entities and relationships in the discipline. Constructing such ontologies is an arduous task that typically requires substantial expert involvement. Recently, researchers have shown how anonymous online crowds can complete pattern-recognition tasks, such as image identification, in a scalable and inexpensive manner.[1–4] In this work, we evaluated the ability of the crowd to perform the engineering task of ontology verification (ie, finding errors). To that end, we asked both the crowd and domain experts to verify a subset of SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms), an ontology that the US government now mandates for use in the clinic as part of 'Meaningful Use' of electronic health records.

BACKGROUND AND SIGNIFICANCE

Ontology

Ontologies provide a means by which experts can specify entities in some domain, their properties, and their relationships to other entities in a formalized manner. An ontology allows its users to 'speak the same language,' so that they use uniform terminology to refer unambiguously to the same entities. Using this powerful property, its users can refer to the same data elements consistently, and can integrate them readily. Thus, ontologies facilitate data access, integration, and reasoning.[5] Ontologies are vital to many industries, from e-commerce to biomedicine, from education to security to e-science. For example, ontologies support indexing systems such as the Google Knowledge Graph and Medline. Ontologies provide a generalized, portable, and reusable method to apply knowledge

Correspondence to Jonathan Mortensen, Stanford Center for Biomedical Informatics Research, Stanford University, 1265 Welch Rd, MSOB X215, Stanford, CA 94305-5497, USA; jmort@stanford.edu

For numbered affiliations see end of article.

computationally, while abstracting that knowledge from any particular implementation.

In research, the use of ontologies is widespread. Google Scholar reveals over 60 000 publications since 2012 that reference 'ontology'. In biomedicine, ontologies combat the 'data deluge' and play a key role in describing and integrating data related to organisms, anatomy, clinical encounters, genes, and chemicals.[6–9] In a high-impact study, Segal *et al*[10] relied on the Gene Ontology to further the understanding of genetic regulatory modules and to identify novel regulatory roles for previously uncharacterized proteins. Ontologies are central to the early-warning pharmacovigilance methods recently developed by Shah and colleagues.[11] Finally, quantifying the ubiquity and diversity of ontologies, the National Center for Biomedical Ontology's BioPortal repository contains over 380 ontologies that describe various subdomains of biomedicine.[12]

The application of ontology in healthcare dates back to 1893, with the introduction of the Bertillon Classification of Causes of Death, now the International Classification of Diseases (ICD), which is in its 10th revision. Today, ontologies underpin almost all aspects of healthcare including billing, publication indexing, patient care, epidemiology, laboratory testing, and prescribing. For example, a healthcare system may use an ontology to reason about drugs and the classes to which they belong. Representing this relationship is key when a computer alerts a physician about possible drug interactions, given a set of patient prescriptions. In fact, the US Government is mandating the use of such systems through Meaningful Use criteria.[13] These standards require that healthcare providers use electronic healthcare records in meaningful ways. One criterion requires use of standardized vocabularies and ontologies including the ICDs, SNOMED CT, RxNorm, and LOINC (Logical Observation Identifiers Names and Codes). Thus, ontologies are a critical component of healthcare and its technology.

The construction of any ontology is a labor-intensive task that requires the involvement of domain experts. In addition, as ontologies become larger and more complex, the challenge of their development increases, as does the likelihood that they contain errors. Many biomedical ontologies contain hundreds of thousands of concepts and relationships among those concepts. At such scale, no single expert can understand the state of an entire ontology or perform quality assurance adequately in an unaided manner. As a result, many large biomedical ontologies, such as the National Cancer Institute Thesaurus (NCIt)[14] and SNOMED CT,[15] have been shown to contain substantial errors in their previous versions. In addition, large ontologies typically use simplified logic, which makes them computationally tractable, but increases the risk that they contain errors that cannot be detected computationally. Thus, finding such errors is an arduous process requiring not only considerable human effort but also a bit of serendipity. Current methods attempt to detect these errors indirectly and automatically focus on finding inconsistencies in ontology syntax and structure.[16,17] The gold standard for finding errors in semantics still requires some form of human peer review.

Crowdsourcing

As the internet has grown, crowdsourcing, the process of 'taking a job traditionally performed by a designated agent and outsourcing it to an undefined large group of people', has emerged.[1,2] With vast numbers of workers available online, crowdsourcing now empowers the scientific community to complete tasks that before were too large, too costly, or too difficult computationally. To crowdsource a problem in practice, a requester decomposes a problem into small subtasks, referred to as micro-tasks, and submits each task to an online community or marketplace, offering compensation or reward (eg, money, enjoyment, or recognition). Multiple workers then complete the task and, in aggregate, produce a final result. This process has enabled the success of popular websites such as Wikipedia, Kickstarter, and reddit. Researchers are now using crowdsourcing as an essential tool. For example, in GalaxyZoo,[3] citizen scientists from the crowd help astronomers to identify and classify galaxies in hundreds of thousands of images; in Foldit,[4] online gamers perform three-dimensional protein folding for fun. The success of these efforts clearly shows the power of crowdsourcing in advancing research.

Crowdsourcing typically solves only tasks that are intuitive (eg, image identification) or that have easily verifiable solutions (eg, a protein conformation that satisfies predefined constraints). Crowdsourcing is generally not applied to tasks that require a trained expert to use domain-specific knowledge. We hypothesized that crowdsourcing can indeed solve such less intuitive problems, and therefore we applied the approach to ontology engineering. Specifically, we focused on the important, challenging task of identifying errors in biomedical ontologies (ie, ontology verification).

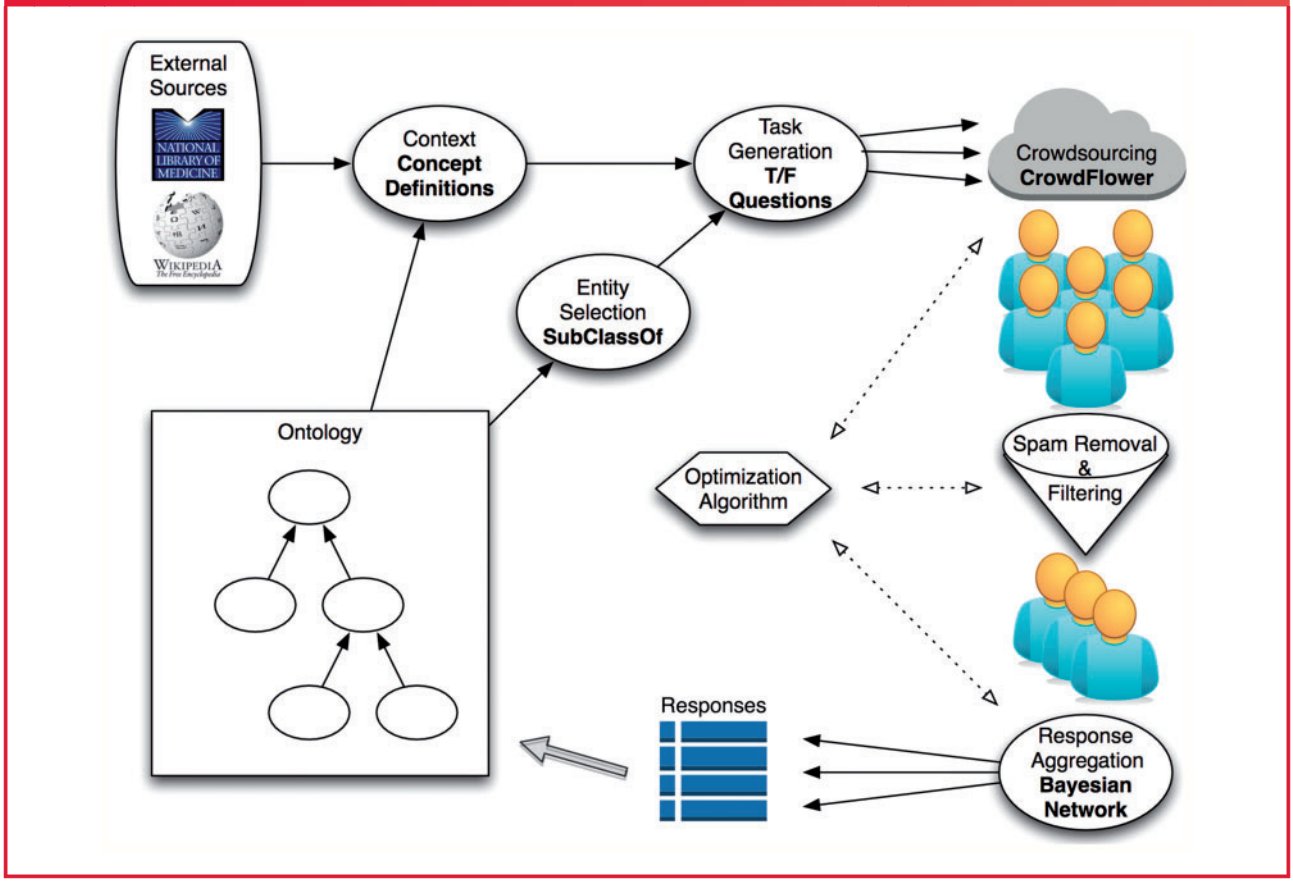
METHODS

We devised a method for scalable ontology verification that integrates several approaches from crowdsourcing research into a unified framework (figure 1). We applied this method as part of a prospective study wherein the crowd verified relationships between entities in SNOMED CT, a large clinical ontology mandated by the US Department of Health and Human Services for Meaningful Use of electronic health records.[9] In addition, we compared the crowd with a panel of five clinical experts who performed the same task, thereby providing a peer-review standard against which to compare the crowd. A description of each step in the methodology and prospective study follows.

Materials

To begin, we selected a portion of the January 2013 version of SNOMED CT upon which to perform the verification. In particular, we performed a basic filtering process (figure 2) to select a random subset of 200 previously unverified, complex, frequently used, entailed hierarchical relationships (eg, 'pneumonia *is a kind of* disease of the lung'). Hierarchical relationships are the dominant type of relationship in biomedical ontologies; thus we only consider hierarchical relationships in this study.[18]

Figure 1: Overview of the method. We devised a standard workflow with which to perform crowdsourcing tasks. We then adapted this workflow to the task of ontology verification, shown above. To note, in this work we combine 'Optimization Algorithm' and 'Spam Removal & Filtering' in the 'Response Aggregation' step. However, we still highlight them because they are integral components in the generalized crowdsourcing workflow. Specific details of each item in the workflow are discussed after the overview section of the Methods.



In addition, Rector *et al*[15] have demonstrated previously that such relationships are particularly prone to error.

Specifically, we began with the SNOMED CT CORE Problem List Subset (http://www.nlm.nih.gov/research/umls/Snomed/core_subset.html), a subset of SNOMED CT. The CORE subset is a selection of the most frequently used terms and concepts across multiple large US healthcare providers. Next, we used Snorocket[19] to find all hierarchical entailments in that subset with the following characteristics:

- non-asserted (ie, not directly stated in the ontology);
- non-trivial (ie, every justification has at least two axioms);
- direct (as described in the OWL API)[20];
- both the parent and child of the entailment is directly listed in the CORE.

To create a manageable study size for the experts, we randomly sampled 200 relationships from the final filtered subset.

The second component necessary for verification is context. In previous work, we showed that the crowd performs best when provided with additional domain information related to the relationship.[21] In this task, we provided users with context by offering English language definitions for each concept (eg, 'pneumonia') in the relationships of interest. Because SNOMED CT did not have definitions available for the concepts in the final relationship set, we selected definitions from either (1) the Medical Subject Headings (MeSH) or (2) the National Cancer Institute Thesaurus (NCIt), both found in the Unified Medical Language System (UMLS). If neither source contained a definition, we did not provide one.

Experiment

We then devised a micro-task with which to perform verification. We presented each worker (either the crowd or expert) with concept definitions and an English language statement of a relationship (figure 3). In previous work, we determined the optimal fashion in which to present this task to a worker.[21] The worker then indicated whether the statement was correct or incorrect, thereby verifying the ontological relationship. We recruited a crowd workforce through CrowdFlower, an online meta-platform with access to a large online labor force. We did not select

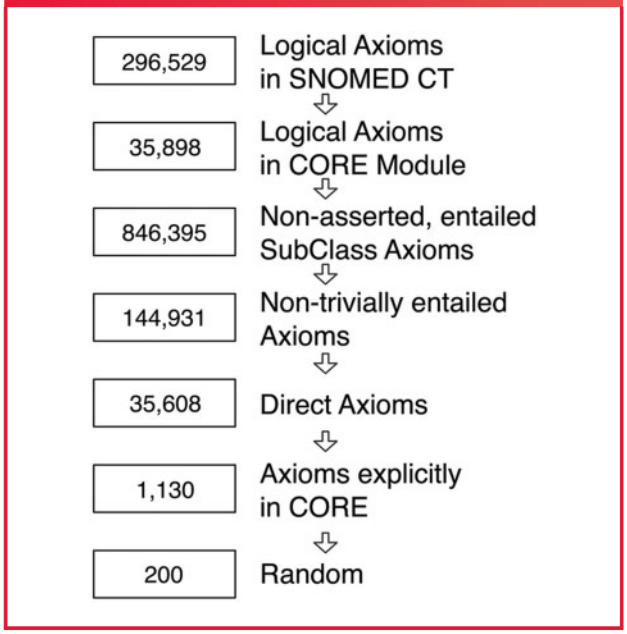


Figure 2: Filtering steps to select relationships for verification. The process of selecting a set of relationships follows a basic filtering strategy. First, we created a syntactic ontology module for the SNOMED CT CORE Problem List. Next, we used Snorocket[19] to find all entailments from the CORE subset. From the entire set of entailments, we then removed all asserted axioms and any trivial entailments (those where the entailment's justification contains only one axiom). From this set, we removed all entailments that were indirect as defined in the OWL API.[20] Finally, we required that both the parent and child of a relationship be contained in CORE, as some entailments contain concepts that are not in CORE but are necessary for the syntactic module. To create a manageable study size for the experts, we subsampled this filtered dataset to 200 relationships.

workers by prespecifying any defining features; thus our workforce represented reasonable worker diversity. In practice, for US$0.02/relationship, workers filled out a basic web form (figure 3) that selected a SNOMED CT relationship at random, reformulated the relationship as an English sentence, provided definitions of the SNOMED CT entities when such definitions were available, and asked whether the sentence was 'True' or 'False'. Twenty-five workers verified each individual relationship for a total cost of US$0.50/relationship.

Asking 25 separate workers to complete a single task may seem gratuitous; however, the power of the crowd lies in its composite response. We aggregated worker responses using a Bayesian method developed by Simpson *et al*,[22] which considers each worker as an imperfect classifier. The method predicts the difficulty of each verification task, the consistency of each worker, and the posterior probability of a relationship being correct or incorrect. Of note, this method performs better than simple majority voting and mitigates the effect of spamming, a common phenomenon in paid online crowdsourcing.

Concurrently, we asked a panel of five experts (MJ, EPM, MAM, ALR, and TES) in both medicine and ontology to perform the same verification task that the crowd performed. These experts are representative of domain experts who assist with the development and maintenance of biomedical ontologies. The experts completed a randomly ordered online survey presented in a fashion identical with the survey that we administered to the crowd, but through a Qualtrics online survey. For purposes of the study, experts also answered questions about the relevancy of the definitions of the terms in the relationship and they provided justifications for their entries. After the experts completed the tasks, we used the Delphi method to assist experts in arriving at a single final judgment.[23] In Delphi, we presented each expert with an anonymized summary of expert responses, including expert-entered comments, and asked the subjects to update their responses, if necessary.

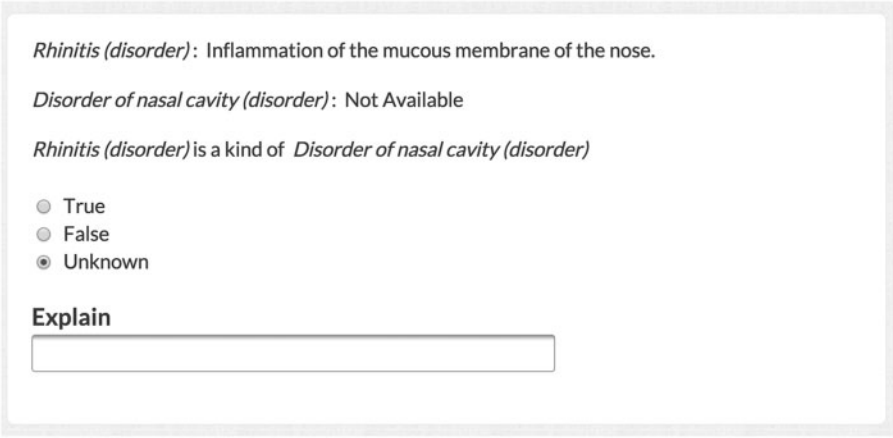
Analysis

We evaluated the votes from the experts and from the crowd in three ways: (1) inter-rater agreement, (2) consensus standard, and (3) cost. For inter-rater agreement, we selected the free marginal κ.[24] The free marginal κ does not assume a fixed number of labels for each rater (expert/crowd) to assign, as is the case for this study—a priori, we do not know the correct labeling nor how many labels there are. Next, we created a consensus standard from the expert majority vote and the result of the Delphi session. Specifically, we included only those relationships in the consensus standard upon which experts reached supermajority (4:1 or 5:0) agreement after Delphi. We then compared the crowd and individual experts against the consensus standard using sensitivity and specificity. In particular, because of the relatively small sample size, we bootstrapped the samples (relationships) to obtain a mean area under the receiver operating characteristic (ROC) curve. By bootstrapping, we were able to estimate how our method would perform, on average, when verifying other relationships not in the experiment. To evaluate whether our method performs better than random, we used permutation testing to compare the null distribution with the bootstrapped distribution.[25] In addition, we bootstrapped the workers to estimate how well our method would perform, on average, when other workers performed the same verification task. Finally, we measured the cost of verifying relationships. To do so, we tracked the number of crowd workers required to complete the task and multiplied by the fixed remuneration of US$0.02/worker/relationship. Likewise, we asked experts to track their time in completing the task. Then, using the mean expert time to completion, we arrived at the approximate cost per relationship using the average hourly salary of a medical expert in California (Bureau of Labor Statistics, http://www.bls.gov/oes/current/oes291069.htm).

RESULTS

Together, the crowd and experts identified 39 critical errors in 200 SNOMED CT relationships (table 1). For instance, the

Figure 3: Online web form for ontology verification. Online workers visited an HTML webpage provided by CrowdFlower (a portion of which is shown here). We required workers to select 'True' or 'False' or explain why they did not know. For each response, we paid a worker US$0.02. Each worker saw the questions in a different, random order. Experts viewed a similar page but with additional fields for comments.



relationship 'short-sleeper *is a kind of* brain disorder' is not true in all cases, and therefore incorrect. Likewise, 'septic shock *is a kind of* soft-tissue infection' mistakes causality with taxonomy. Septic shock may be caused by a soft-tissue infection but itself is not an infection. Generally, these errors are the result of subtly incorrect logical definitions that have unintended effects on computed logical conclusions. We have contacted the International Health Terminology Standards Development Organization (IHTSDO), which develops SNOMED CT, and provided them with the errors and expert justifications. We anticipate that these errors will be corrected in future versions of SNOMED CT. We used results of the Delphi round to produce a consensus standard among the experts against which to compare each individual expert and the crowd. After the Delphi round, the experts reached supermajority agreement (ie, 4:1 or 5:0) on the truth value of 187 of the 200 relationships from SNOMED CT that we studied. We used these 187 relationships as a consensus standard set of relationships. For the remaining 13 relationships, definitions were unavailable in both MeSH and NCIt, explaining why the experts may not have been able to reach agreement.

We then compared the responses of the crowd with the initial expert responses in three ways: inter-rater agreement within groups, performance on the consensus standard, and estimated cost. The average pairwise free marginal κ of the crowd fell within the range of that of the experts (table 2). Furthermore, the crowd identified errors with a bootstrapped mean area under the ROC curve (AUC) of 0.83 ($p<2 \times 10^{-16}$). Generally, each individual expert performed marginally better than the crowd at various points on the ROC curve (figure 4). When bootstrapping the performance of the workers, the crowd performed with a mean AUC of 0.78 ($p<2e-16$). Note that this result provides a lower bound estimate of how other workers may have generally performed on the same task. However, the workers are not independently distributed and therefore the bootstrapped result should only serve as a guide. A real-world replication would likely have a higher AUC. Finally, in comparison with cost of the crowd at US$0.50/relationship, experts cost ~US$2.00/relationship, based on average task completion time (4.5 h) and average salary of a general practice physician in California (~US$182 580).

DISCUSSION

In this work, we used both experts and crowdsourcing to perform quality assurance on a 200-relationship subset of SNOMED CT. We found that the crowd is nearly indistinguishable from any single expert in the ability to identify errors in a random sample of SNOMED CT relationships. This subset contained terms used frequently in many hospitals (as the terms were derived from the SNOMED CT CORE Subset). Moreover, this random sample is representative of SNOMED CT and of many other large complex ontologies in its logical structure. Of note, nearly 20% of the relationships were in error as judged by the experts. While this error rate is likely higher than the overall error rate among relationships in SNOMED CT, it still indicates that further quality assurance of SNOMED CT is essential. The presence of such errors, although not unexpected based on the literature,[14–17] is concerning and elicits some open questions about SNOMED CT and about biomedical ontologies in general.

- At what rate would experts identify errors in all biomedical ontologies or of those ontologies required in electronic health records?
- What is the impact of ontology errors on downstream methods? (For example, could a clinical decision support system misclassify a patient because of ontology errors?)

Table 1: Listing of errors found by both the crowd and experts in a subset of SNOMED CT

Child	Parent
Anterior shin splints (disorder)	Disorder of bone (disorder)
Short-sleeper (disorder)	Disorder of brain (disorder)
Frontal headache (finding)	Pain in face (finding)
Local infection of wound (disorder)	Wound (disorder)
Anal and rectal polyp (disorder)	Rectal polyp (disorder)
Malignant neoplasm of brain (disorder)	Malignant tumor of head and/or neck (disorder)
Diabetic autonomic neuropathy associated with type 1 diabetes mellitus (disorder)	Diabetic peripheral neuropathy (disorder)
Placental abruption (disorder)	Bleeding (finding)
Impairment level: blindness one eye—low vision other eye (disorder)	Disorder of eye proper (disorder)
Gastroenteritis (disorder)	Disorder of intestine (disorder)
Microcephalus (disorder)	Disorder of brain (disorder)
Thrombotic thrombocytopenic purpura (disorder)	Disorder of hematopoietic structure (disorder)
Fibromyositis (disorder)	Myositis (disorder)
Lumbar radiculopathy (disorder)	Spinal cord disorder (disorder)
Vascular dementia (disorder)	Cerebral infarction (disorder)
Chronic tophaceous gout (disorder)	Tophus (disorder)
Full thickness rotator cuff tear (disorder)	Arthropathy (disorder)
Disorder of joint of shoulder region (disorder)	Arthropathy (disorder)
Injury of ulnar nerve (disorder)	Injury of brachial plexus (disorder)
Basal cell carcinoma of ear (disorder)	Basal cell carcinoma of face (disorder)
Bronchiolitis (disorder)	Bronchitis (disorder)
Migraine variants (disorder)	Disorder of brain (disorder)
Gingivitis (disorder)	Inflammatory disorder of jaw (disorder)
Septic shock (disorder)	Soft tissue infection (disorder)
Cellulitis of external ear (disorder)	Otitis externa (disorder)
Inguinal pain (finding)	Pain in pelvis (finding)
Disorder of tendon of biceps (disorder)	Disorder of tendon of shoulder region (disorder)
Pain of breast (finding)	Chest pain (finding)
Injury of ulnar nerve (disorder)	Ulnar neuropathy (disorder)
Injury of back (disorder)	Traumatic injury (disorder)
Achalasia of esophagus (disorder)	Disorder of stomach (disorder)
Pneumonia due to respiratory syncytial virus (disorder)	Interstitial lung disease (disorder)
Sensory hearing loss (disorder)	Labyrinthine disorder (disorder)
Degeneration of intervertebral disc (disorder)	Osteoarthritis (disorder)
Disorder of sacrum (disorder)	Disorder of bone (disorder)
Peptic ulcer without hemorrhage, without perforation AND without obstruction (disorder)	Gastric ulcer (disorder)
Diabetic autonomic neuropathy (disorder)	Peripheral nerve disease (disorder)
Cyst and pseudocyst of pancreas (disorder)	Cyst of pancreas (disorder)
Calculus of kidney and ureter (disorder)	Ureteric stone (disorder)

MeSH, Medical Subject Headings; NCIt, National Cancer Institute Thesaurus; SNOMED CT, Systematized Nomenclature of Medicine Clinical Terms.

Table 2: Mean free marginal κ between experts themselves and the crowd

Relationship set	Crowd κ	Expert κ
All (n = 187)	0.58 (0.55, 0.61)	0.57 (0.49, 0.66)
Easy (n = 105)	0.9 (0.9, 0.9)	1 (1, 1)
Delphi–Agreement (n = 48)	0.15 (0.08, 0.29)	0.07 (−0.12, 0.25)
Delphi–Near Agreement (n = 34)	0.19 (0.12, 0.29)	−0.05 (−0.35, 0.24)

With the expert votes obtained, we determined the mean free marginal κ between experts (ie, the average agreement of an expert with another expert). On correct relationships, expert κ was ~0.7 before Delphi, and ~0.9 after. On incorrect relationships, expert κ was ~0.0 before Delphi, and ~0.69 after. We then calculated the mean free marginal κ between the final crowd response and each expert (ie, the average agreement of the crowd with each expert). Note that the mean crowd inter-rater agreement (0.58) falls well within the range of the expert agreement (0.49, 0.66). Finally, we stratified by subsets of relationships. Again, note that, on each subset, the mean crowd inter-rater agreement falls well within the range of the experts. Terminology: 'Easy'—relationships for which experts reached immediate consensus; 'All'—entire set of relationships; 'Delphi–Agreement'—relationships for which experts reached complete agreement after Delphi; 'Delphi–Near Agreement'—relationships upon which only a supermajority of experts reached agreement after Delphi.

- What incorrect analytical conclusions could be made because of ontology errors? (For example, might Gene Ontology enrichment analysis mischaracterize microarray data?)[26]
- What are the most important kinds of errors to detect and eliminate?
- What is the best approach to reduce or avoid such errors (eg, crowdsourcing, experts, automated algorithms, best practices)?

Potential impact of an ontology error

The errors the crowd identified are particularly interesting because they involve concepts in the SNOMED CT CORE Subset, indicating (1) that these concepts are used very frequently across many hospitals, and (2) that these concepts and relationships will likely play a role in the clinical decision support systems required by Meaningful Use. To illustrate the significance of the errors identified, we describe two hypothetical situations focused on 'short-sleeper *is a kind of* brain disorder.'

A. A clinical decision support system suggests the immobilization of all persons with a brain disorder. Using the error above, the system would improperly recommend the immobilization of those who experience shortened sleep. This incorrect recommendation would certainly cost practitioner time and trust and may even cause an unwarranted procedure.

B. When querying patient data to extract cohorts (eg, with tools such as i2b2),[27] the query results on persons with brain disorders would entirely mischaracterize the population, classifying short sleepers into the 'cases' instead of the 'controls'. This misclassification could lead to incorrect hypotheses about the population or even an extremely biased retrospective study result.

These two situations show how the errors the crowd identified could affect healthcare today and how the errors are especially important to identify as we move forward with ontology-based health information technology.

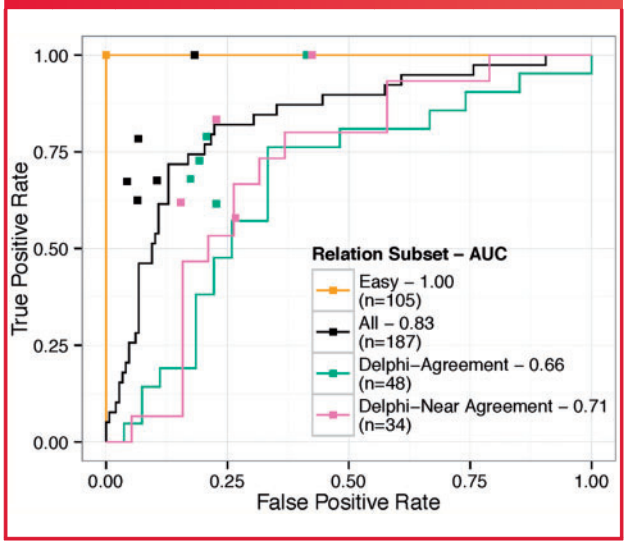
Ontology and ground truth

Evaluating ontology verification methods remains a challenge—there is no ground truth or absolute truth against which to compare methods. No common biomedical ontology verification standard is available outside of peer review. Because ontologies reflect a shared (expert) consensus about a domain, an absolute truth about what is right or wrong cannot exist. Instead, we view an error as a statement in the ontology with which domain experts do not agree (ie, the statement contradicts expert understanding of the domain). In light of this situation, single-expert verification is the most common method to identify an error and to determine if an error is 'real'. In our work, we use a multi-expert consensus to serve as the approximation of ground truth. One should not consider our results (either performance metrics or actual errors) as truth but instead a reflection of how well the crowd compares with experts in what they interpret to be an error.

Crowd-assisted ontology engineering

The cost of ontology engineering and maintenance is considerable—for example, hiring a single physician to perform engineering and verification costs of the order of US$200 000/year. It is unlikely that a single physician working full time could properly verify the entirety of SNOMED CT, let alone have deep, extensive knowledge about all its varied topics. Indeed, the cost to build large artifacts such as SNOMED CT and the ICDs is orders of magnitude greater than that of hiring a single physician. Given these high costs, it is encouraging to see the crowd perform so well at identifying errors at a lower cost. We found that the crowd costs about a quarter of that of an expert, yet performed comparably. These results suggest that the crowd can function as a scalable assistant to ontology engineers. Our crowd-based method is especially appropriate in situations where an expert is unavailable, budget is limited, or an ontology is too large for manual error checking. While this study focused on one particular type of relationship, our methodology is general and thus could be applied readily to other ontology-verification or

Figure 4: Comparison of the crowd and experts. We compared the crowd performance against the expert consensus standard by receiver operating characteristic (ROC) curves (solid lines). In addition, we compared each individual expert against that same consensus standard (square points). In general, the experts performed better than the crowd (above the ROC curve), but not by a marked margin. Note that the key also includes the area under the ROC (AUC). Terminology: 'Easy'—relationships for which experts reached immediate consensus; 'All'—entire set of relationships; 'Delphi–Agreement'—relationships for which experts reached complete agreement after Delphi; 'Delphi–Near Agreement'—relationships upon which only a supermajority of experts reached agreement after Delphi.



ontology-engineering tasks, thereby reducing costs even further. In practice, an ontology-development environment would integrate this crowdsourcing functionality directly, seamlessly allowing crowd-based ontology error checking and engineering with the click of a button.

Crowdsourcing expert-level tasks

Previous work with crowdsourcing has focused primarily on intuitive, pattern-recognition tasks.[1–4] For instance, common sense tasks such image object recognition or text sentiment analysis are readily solved with crowdsourcing. Encouragingly, our results suggest that crowdsourcing can also solve more complex, expert-level tasks. This result is especially relevant for situations where experts are unavailable, expensive, or unable to complete a large task. We identified two factors to consider when one is developing a crowdsourcing solution to knowledge-intensive tasks. First, it is non-trivial to reformulate an expert-level task as one suitable for the crowd. We found that rapid, iterative task design was essential for arriving at a task formulation that the crowd could complete. The second factor is knowledge type. Tasks that require synthesis of knowledge, or that require background knowledge that cannot be provided directly to workers, may not be appropriate. It is likely that tasks that are self-contained (ie, all the necessary information is immediately available) are most appropriate for the crowd. For example, in the ontology-verification task, definitions provided all the necessary information with which to complete the task. We cannot exclude the possibility that crowd workers first accessed web-based resources such as Wikipedia before responding to our online questions, however. There are many expert-level tasks that are similarly 'self-contained', and thus we are excited about the possibility of crowdsourcing other knowledge-intensive tasks.

CONCLUSION

Ontologies, which define for both people and computers the entities that exist in a domain and the relationships between them, support many data-intensive tasks throughout biomedicine and healthcare. The biomedical community, however, faces a challenge in engineering ontologies in a scalable, high-quality fashion, particularly when mature ontologies may include many thousands of concepts and relationships. We have shown that crowdsourcing, which researchers use to provide solutions to intuitive tasks in a scalable way, can address this engineering challenge. We used crowdsourcing methods to solve the difficult task of identifying errors in SNOMED CT, an important, large biomedical ontology. We then compared results from the crowd with those offered by medical experts who performed the same task, and we found that errors that the two groups identified were concordant. The results suggest that crowdsourcing may offer mechanisms to solve problems that require considerable biomedical expertise. Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/amiajnl-2014-002901)

ACKNOWLEDGEMENTS

Special thanks to E Simpson for statistics-related advice on how to use and implement Variational Bayes.

CONTRIBUTORS

JMM, MAM, and NFN developed and executed the study. They also wrote and revised the manuscript. JMM is guarantor. TES, EPM, ALR, MJ, and MAM served as domain experts.

FUNDING

This work was supported by the National Institute of General Medical Sciences grant number GM086587, by the National Center for Biomedical Ontology, supported by the National Human Genome Research Institute, the National Heart, Lung, and Blood Institute, and the National Institutes of Health Common Fund grant number HG004028, and by the National Library of Medicine Informatics Training grant number LM007033.

COMPETING INTERESTS

None.

PROVENANCE AND PEER REVIEW

Not commissioned; externally peer reviewed.

REFERENCES

1. Howe J. The rise of crowdsourcing. *Wired Mag*. 2006;14:1–4.
2. Quinn AJ, Bederson BB. Human computation: a survey and taxonomy of a growing field. *Proceedings of the 2011 annual conference on Human factors in computing systems—CHI'11*. Vancouver, BC: ACM, 2011:1403–1412.
3. Lintott CJ, Schawinski K, Slosar A, *et al*. Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Mon Not R Astron Soc*. 2008; 389:1179–1189.
4. Cooper S, Khatib F, Treuille A, *et al*. Predicting protein structures with a multiplayer online game. *Nature*. 2010;466: 756–760.
5. Staab S, Studer R. *Handbook on ontologies*. 2nd edn. Springer-Verlag New York Inc, 2009.
6. Bodenreider O, Stevens R. Bio-ontologies: current trends and future directions. *Brief Bioinform*. 2006;7:256–274.
7. Rubin DL, Shah NH, Noy NF. Biomedical ontologies: a functional perspective. *Brief Bioinform*. 2008;9:75–90.
8. Hunter L, Lu Z, Firby J, *et al*. OpenDMAP: an open source, ontology-driven concept analysis engine, with applications to capturing knowledge regarding protein transport, protein interactions and cell-type-specific gene expression. *BMC Bioinformatics*. 2008;9:78.
9. Hoehndorf R, Dumontier M, Gennari JH, *et al*. Integrating systems biology models and biomedical ontologies. *BMC Syst Biol*. 2011;5:124.
10. Segal E, Shapira M, Regev A, *et al*. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet*. 2003;34:166–176.
11. LePendu P, Iyer S V, Bauer-Mehren A, *et al*. Pharmacovigilance using clinical notes. *Clin Pharmacol Ther*. 2013;93: 547–555.
12. Whetzel PL, Noy NF, Shah NH, *et al*. BioPortal: enhanced functionality via new web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Res*. 2011;39:W541–545.
13. Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records. *N Engl J Med*. 2010;363: 501–504.
14. Ceusters W, Smith B, Goldberg L. A terminological and ontological analysis of the NCI Thesaurus. *Methods Inf Med*. 2005;44:498.
15. Rector AL, Brandt S, Schneider T. Getting the foot out of the pelvis: modeling problems affecting use of SNOMED CT hierarchies in practical applications. *J Am Med Informatics Assoc*. 2011;18:432–440.
16. Zhu X, Fan JW, Baorto DM, *et al*. A review of auditing methods applied to the content of controlled biomedical terminologies. *J Biomed Inform*. 2009;42:413–425.
17. Ochs C, Perl Y, Geller J, *et al*. Scalability of abstraction-network-based quality assurance to large SNOMED hierarchies. *AMIA Annu Symp Proc*. 2013;2013:1071–1080.
18. Noy NF, Mortensen JM, Alexander PR, *et al*. Mechanical Turk as an ontology engineer? Using microtasks as a component of an ontology engineering workflow. Web Science, 2013.
19. Lawley MJ, Bousquet C. Fast classification in Protégé: Snorocket as an OWL 2 EL reasoner. Proceedings of the 6th Australasian Ontology Workshop (IAOA'10). Conferences in Research and Practice in Information Technology 2010; 45–49.
20. Horridge M, Bechhofer S. The OWL API: a Java API for working with OWL 2 ontologies. OWLED 2009;11–21.
21. Mortensen JM, Noy NF, Musen MA, *et al*. Crowdsourcing ontology verification. *International Conference on Biomedical Ontologies*. 2013.
22. Simpson E, Roberts S, Psorakis I, *et al*. Dynamic Bayesian combination of multiple imperfect classifiers, 2012. http://arxiv.org/abs/1206.1831
23. Linstone HA, Turoff M. *The Delphi method: techniques and applications*. Addison-Wesley, 1975.
24. Randolph JJ. Free-marginal multirater kappa (multirater κfree): an alternative to Fleiss' fixed-marginal multirater kappa. Joensuu learning and instruction symposium, 2005.
25. Efron B. *The Jackknife, the Bootstrap and other resampling plans*. SIAM, 1982.
26. Khatri P, Druaghici S. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*. 2005;21:3587–3595.
27. Murphy SN, Weber G, Mendis M, *et al*. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J Am Med Informatics Assoc*. 2010; 17:124–130.

AUTHOR AFFILIATIONS

[1]Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, California, USA

[2]Biomedical Informatics Training Program, Stanford University, Stanford, California, USA

[3]Faculty of Medicine, University of Calgary, Calgary, Canada

[4]School of Computer Science, University of Manchester, Manchester, UK

[5]Google Inc., Mountain View, California, USA