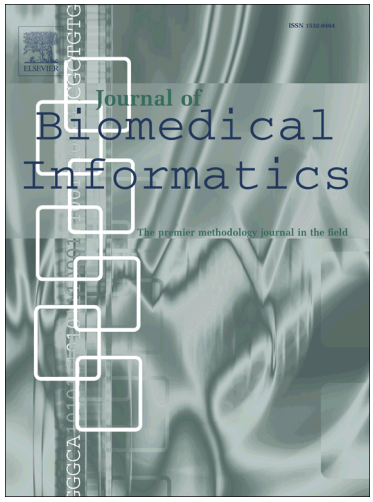


Use of Ontology Structure and Bayesian Models to Aid the Crowdsourcing of ICD-11 Sanctioning Rules

Yun Lou, Samson W. Tu, Csongor Nyulas, Tania Tudorache, Robert J.G. Chalmers, Mark A. Musen

PII: S1532-0464(17)30025-4
DOI: http://dx.doi.org/10.1016/j.jbi.2017.02.004
Reference: YJBIN 2722

To appear in: *Journal of Biomedical Informatics*

Received Date: 17 August 2016
Revised Date: 2 February 2017
Accepted Date: 8 February 2017

Please cite this article as: Lou, Y., Tu, S.W., Nyulas, C., Tudorache, T., Chalmers, R.J.G., Musen, M.A., Use of Ontology Structure and Bayesian Models to Aid the Crowdsourcing of ICD-11 Sanctioning Rules, *Journal of Biomedical Informatics* (2017), doi: http://dx.doi.org/10.1016/j.jbi.2017.02.004

Use of Ontology Structure and Bayesian Models to Aid the Crowdsourcing of ICD-11 Sanctioning Rules

Yun Lou, MS[1*], Samson W. Tu, MS[1*], Csongor Nyulas, MS[1], Tania Tudorache, PhD[1], Robert J. G. Chalmers, MB, FRCP[2], Mark A. Musen, MD, PhD[1]

[1]Stanford University, Stanford, CA, USA [2]University of Manchester, Manchester, UK

*Co-First Authors

Corresponding author:
Samson W. Tu
swt@stanford.edu
+1 650 725 3391
1265 Welch Road, X-259
Stanford, CA 94305-5479, USA

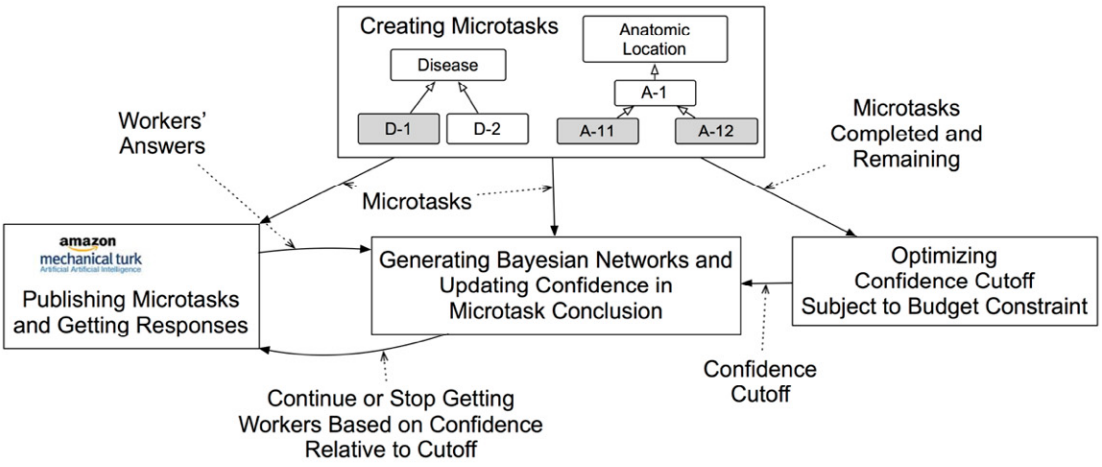
Abstract

The International Classification of Diseases (ICD) is the de facto standard international classification for mortality reporting and for many epidemiological, clinical, and financial use cases. The next version of ICD, ICD-11, will be submitted for approval by the World Health Assembly in 2018. Unlike previous versions of ICD, where coders mostly select single codes from pre-enumerated disease and disorder codes, ICD-11 coding will allow extensive use of multiple codes to give more detailed disease descriptions. For example, "severe malignant neoplasms of left breast" may be coded using the combination of a "stem code" (e.g., code for malignant neoplasms of breast) with a variety of "extension codes" (e.g., codes for laterality and severity). The use of multiple codes (a process called post-coordination), while avoiding the pitfall of having to pre-enumerate vast number of possible disease and qualifier combinations, risks the creation of meaningless expressions that combine stem codes with inappropriate qualifiers. To prevent that from happening, "sanctioning rules" that define legal combinations are necessary. In this work, we developed a crowdsourcing method for obtaining sanctioning rules for the post-coordination of concepts in ICD-11. Our method utilized the hierarchical structures in the domain to improve the accuracy of the sanctioning rules and to lower the crowdsourcing cost. We used Bayesian networks to model crowd workers' skills, the accuracy of their responses, and our confidence in the acquired sanctioning rules. We applied reinforcement learning to develop an agent that constantly adjusted the confidence cutoffs during the crowdsourcing process to maximize the overall quality of sanctioning rules under a fixed budget. Finally, we performed formative evaluations using a skin-disease branch of the draft ICD-11 and demonstrated that the crowd-sourced sanctioning rules replicated those defined by an expert dermatologist with high precision and recall. This work demonstrated that a crowdsourcing approach could offer a reasonably efficient method for generating a first draft of sanctioning rules that subject matter experts could verify and edit, thus relieving them of the tedium and cost of formulating the initial set of rules.

Keywords

Sanctioning Rules; Ontology; ICD; Post-Coordination; Crowdsourcing; Bayesian Network

Graphical Abstract



Highlights

- We defined crowdsourcing microtasks to obtain ICD-11 sanctioning rules.
- We used hierarchical structures to improve the efficiency of crowdsourcing.
- We used Bayesian networks to model our confidence in the acquired sanctioning rules.
- We developed a method to maximize the quality of the rules within a fixed budget.

1 Introduction

The International Classification of Diseases (ICD) is the de facto standard international classification for mortality reporting and for many epidemiological, clinical, and financial use cases. The current 10th edition of ICD was endorsed by the World Health Assembly in 1990 and has been updated periodically over the years. The World Health Organization (WHO) is currently revising ICD to produce a new version, ICD-11 [1].

ICD-10 is a pre-coordinated system in which combinations of diseases and qualifiers, such as *F32.0 mild depressive episode*, *F32.1 moderate depressive episode*, *F32.2 severe depressive episode without psychotic symptoms,* and *F32.3 severe depressive episode with psychotic symptoms*, are enumerated as part of the classification. ICD-11, on the other hand, will provide a mechanism for *post-coordination*, which means that coders will be able to combine disease codes

(*stem codes*) with qualifiers (*extension codes*) from a special ICD-11 chapter to form detailed descriptions of diseases and disorders [2]. In the previous example, *F32.0 mild depressive episode*, a pre-coordinated term, can be represented in a post-coordinated system as a combination of the disease concept *F32 Depressive episode* and a new severity qualifier *mild*. It is easy to see that pre-coordinating all possible combinations of disease categories and their qualifiers will produce a huge classification that is too unwieldy to be usable. Post-coordination systems, on the other hand, can reduce the number of classification entities dramatically at the cost of increasing the complexity of the coding process.

An advantage of post-coordination is that it provides additional dimensions for describing and aggregating diseases beyond the simple parent–child relationships that already exist in a pre-coordinated classification. For instance, *psoriasis affecting the hands and feet* and *palmoplantar keratoderma* are both diseases of the hands and feet that may result in similar disability, such as pain upon walking or difficulty carrying out manual tasks. If the occurrences of psoriasis affecting the hands and feet were coded with the appropriate anatomical extension code in an electronic health record, their incidents could be searchable and aggregated with those of other hand- or foot-related diseases such as palmoplantar keratoderma.

Despite their advantages, post-coordination systems introduce a problem: End-users may create nonsensical combinations of classification entities and qualifiers. To prevent this, we need integrity constraints similar to what the GALEN project, an early effort that pioneered methods for the composition of biomedical concepts, called “sanctioning rules,” in which only combinations permitted by the sanctioning rules can be generated [3]. For example, a sanctioning rule in GALEN may state that fractures sensibly have location bones. In formal terms, sanctioning rules can be seen as description logic axioms used as integrity constraints in an ontology [4]. In the Manchester Syntax for the Web Ontology Language (OWL), if *DiseaseX* and *LocationY* are classes in the Disease and Anatomical Location class hierarchies respectively, and *hasLocation* is an object property, a sanctioning rule can be formulated as *DiseaseX hasLocation* **only** *LocationY*, where **only** is the universal quantifier. Using the OWL integrity constraint semantics suggested by

Tao et al. [4], this rule indicates that *DiseaseX* can be combined with *LocationY* or any of the known specializations of *LocationY*. For any anatomical location *LocationZ* not known to be a descendant of *LocationY*, *DiseaseX hasLocation* **some** *LocationZ* should represent a constraint violation, not permitting the inference that *LocationZ* is a subclass of *LocationY*. For simplicity, we denote the disease/anatomical location sanctioning rules as (*DiseaseX*, *LocationY*). There can be multiple valid sanctioning rules for a disease. For instance, palmoplantar keratoderma can occur in any part of the palm and any part of the sole, so there are two sanctioning rules for palmoplantar keratoderma: *(palmoplantar keratoderma, palm structure)* and *(palmoplantar keratoderma, sole structure)*. In effect, we define a set of sanctioning rules as the disjunction of the individual rules.

High-quality sanctioning rules can improve the quality of post-coordination systems considerably [5]. Suppose we define the sensitivity and specificity of a terminology with respect to the set of concepts the terminology is designed to represent. Sensitivity would be defined in terms of the concepts that can be represented (true positives) and specificity in terms of non-sensible terms that are ruled out (true negatives). High specificity of the sanctioning rules may diminish the sensitivity of the terminology. If excessively aggressive sanctioning rules prevent the formation of some sensible combinations, the usability of post-coordination systems will be affected negatively. Acquiring appropriate sanctioning rules is therefore essential for proper usage of post-coordination systems. For a large classification such as ICD-11, which has tens of thousands of disease and disorder codes and multiple post-coordination axes that can have thousands of terms in their value sets, careful examination of millions of possible combinations of diseases and, for example, anatomical locations is necessary to create high-quality sanctioning rules. Manual definition and verification of all possible sanctioning rules by subject matter experts would be extremely time-consuming. In this work, we sought to develop a more efficient method to generate a first draft of the sanctioning rules that subject matter experts can verify and edit, thus relieving them of the tedium of formulating the initial set of rules. We have chosen to use crowdsourcing—enlisting a crowd of humans—to help obtain sanctioning rules. More specifically, we extended existing crowdsourcing methods to derive constraints between two

hierarchical structures efficiently. In this paper, we focus on crowdsourcing the specification of sanctioning rules between diseases and anatomical locations in ICD-11. However, all the methods that we have developed can be adapted to solve the more general problem of crowdsourcing ontology constraints.

The paper is organized as follows: Section 2 reviews the literature on sanctioning rules, natural language processing (as an alternative approach to acquiring sanctioning rules), crowdsourcing, and the modeling of workers' answers; Section 3 details the methodology proposed in this work—i.e., how to define the microtasks involved in acquiring sanction rules, how to generate and publish the tasks, how to combine workers' answers, and how to define the metrics used to evaluate experimental results; and Section 4 describes the results of two crowdsourcing experiments. The paper ends with a summary of this work's contributions and its limitations.

2 Background and Related Work

2.1 Sanctioning Rules

The GALEN Project [6] was the first system to incorporate sanctioning rules in biomedical terminology development. It distinguished three types of sanctioning rules: (1) those that are conceivable, i.e., statements saying that an attribute exists and can be used; (2) those that are grammatical, i.e., statements that allow the formation of legal queries (e.g., *Anatomical entity whose laterality is Left*), but that also allow grammatically correct but factually incorrect compositions such as *Head whose laterality is Left*; and (3) those that are sensible, i.e., statements that are sufficiently constrained to allow only valid composition of terms. In the ICD-11 revision work, a Content Model [7] formulated in OWL specifies possible post-coordination axes (the conceivable sanctioning). The Web-based iCAT content-editing tool [8] provides facilities for subject matter experts to specify axes that are appropriate for particular branches of ICD-11 entities (the grammatical sanctioning). The ICD-11 sanctioning rules studied in our project correspond to GALEN's sensible sanctioning rules.

The clinical terminology SNOMED CT does not have explicit sanctioning rules as GALEN does. Instead, it uses the notions of *defining characteristics* and *qualifiers* and their refinability to specify whether and how a concept can be specialized. Defining characteristics are subtype relationships and other relationships (e.g., *associated finding*) used to express necessary and, in the case of fully defined concepts, sufficient conditions of a SNOMED concept. As stated in the SNOMED CT Technical Implementation Guide [9], qualifiers are attributes "that may have one of several possible values for a particular Concept. If a particular qualifier is applied to a Concept, the resulting expression represents a more tightly defined subtype of that Concept." In Figure 1, the concept *pneumonia* has, as its defining characteristics, *is a*, *associated morphology*, and *finding site*, while *severities*, *episodicities*, and *courses* are its qualifiers. Defining characteristics and qualifiers can be refined if they have subclasses that can be used to define subtypes of a concept. In the *pneumonia* example, the characteristics *associated morphology* of *inflammation* and *finding site* of *lung structure* can be refined (e.g., to *acute inflammation* and to *right lung structure* respectively), but the *associated morphology* of *consolidation* cannot be refined (note the lack of a + symbol in front of the *consolidation* term). As discussed in the SNOMED CT Technical Implementation Guide [9], these refinement specifications can be seen as sanctioning rules that specify legal combinations of SNOMED CT post-coordination. However, some qualifiers, such as *severity*, are never used as a defining characteristic, resulting in concepts such as *severe asthma* non-sensibly having *mild* or *moderate* as allowed values for the *severity* qualifier [10]. Furthermore, SNOMED CT allows post-coordination of concepts that are not explicitly sanctioned by the refinability rules of defining characteristics and qualifiers. The SNOMED CT Technical Implementation Guide gives as an illustrative example where even though the concept *headache* does not have *severity* as a sanctioned attribute, yet the post-coordination (*headache severity severe*) is permitted [9]. The Guide offers some recommendations on such usage of post-coordination. For example, the *causative agent* attribute can be applied to a *clinical finding* concept, even though *causative agent* is not an approved defining characteristic of the *clinical*

finding concept, because *causative agent* is approved for refining *bacterial infectious disease*, which is a subclass of *clinical finding* in SNOMED CT [9].

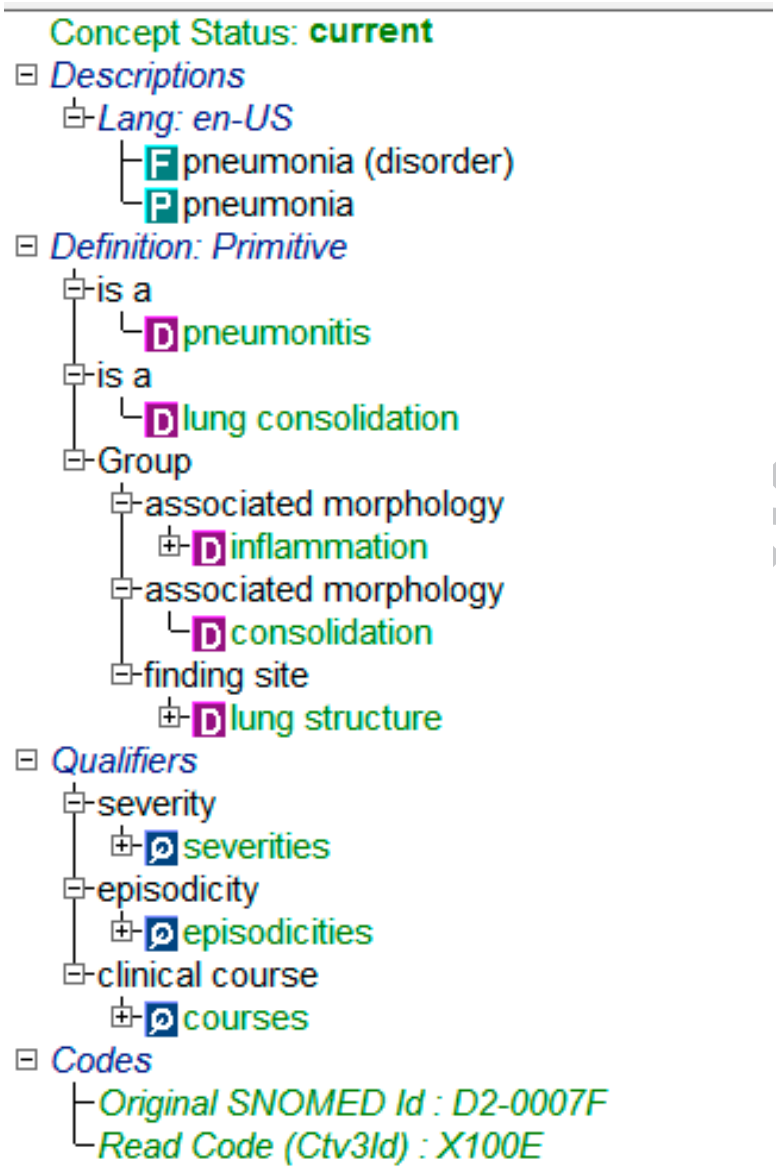


Figure 1 Definition of Pneumonia in SNOMED CT, as displayed in the CliniClue browser.

2.2 Alternative Natural Language Processing and Text Mining Approaches

Natural language processing (NLP) and text mining have become important tools for knowledge discovery in biomedicine [11, 12]. Mining sanctioning rules corresponds to the applied NLP problems of identifying named entities and recognizing relationships between pairs of entities [13,

14]. If we limit ourselves to recognizing relationships between diseases and fixed post-coordination value sets, then a dictionary-based approach for entity recognition may be sufficient [14]. For the extraction of specific types of relationships, pattern-based approaches [15] or more sophisticated machine-learning methods [16] have been used. In bioinformatics, text mining has been used to extract protein-protein interactions and the relationship between gene functions and diseases [17]. Within clinical medicine, text mining has been used to extract causal relationships [18], environmental and phenotypic associations [19], and drug-drug interactions [20] among others. Especially relevant is the work of Coden et al. [21] who built a cancer disease representation model that included characteristics such as anatomic site, histology, and grade value. Using both rule-based and machine-learning techniques, Coden et al. populated the cancer disease representation model with concepts and relationships extracted from free-text pathology reports and obtained adjusted F1 scores that range from 0.65 to 1.00.

Our informal survey of the NLP literature suggests that while there are general techniques to discover relationships in biomedical text [16, 21, 22], application of these techniques to different classes of diseases to find specific relationships may require that we tune the system to a variety of text that have different characteristics (e.g., cancer pathology reports versus medical textbooks or PubMed abstracts). Instead, we chose to use crowdsourcing as a method that has the potential to be more uniformly applicable.

2.3 Crowdsourcing

Doan et al. [23] defined crowdsourcing broadly: A system is using crowdsourcing "if it enlists a crowd of humans to help solve a problem defined by the system owners, and if in doing so, it addresses the following four fundamental challenges: How to recruit and retain users? What contributions can users make? How to combine user contributions to solve the target problem? How to evaluate users and their contributions?" Crowdsourcing has proven useful for many tasks, such as identifying objects in images [24], creating and editing documents (e.g., Wikipedia), and funding new projects by raising contributions from a crowd [25]. Mortensen et al. [26] provided a detailed explanation of the *microtask* variant of crowdsourcing that we have adopted in this

project. A requester, who needs to have a certain task performed, divides the task into microtasks, with each microtask usually requiring a few seconds to a few minutes to complete. The requester publishes the microtasks in an online marketplace. The publication of microtasks is often done automatically through an application programming interface. The workers on the crowdsourcing platform find microtasks they want to perform, and they get paid to do the work. For microtasks that require specialized knowledge, such as answering biomedical questions, a requester usually asks multiple workers to finish the same microtask. Once the workers finish a microtask, the requester collects and assesses the responses, combines the results to draw a conclusion regarding the microtask, and rewards workers according to a pre-defined remuneration scheme. A requester can also pay an additional bonus to workers who give answers that are judged to be "correct."

Recently, researchers reported mixed success with using microtask crowdsourcing for several biomedical-ontology engineering tasks. Mortensen et al. [26, 27] used microtask crowdsourcing for the verification of hierarchical relations in biomedical ontologies. They correctly verified 86% of the hierarchical relations from the CORE subset of SNOMED CT [26]. However, when asked to verify relationships in the Gene Ontology, whose terms are more esoteric and have fewer Internet references, crowd workers performed more poorly [27]. Sarasua et al. [28] developed CrowdMap to identify the alignments of concepts between two ontologies with crowdsourcing. They showed that CrowdMap could improve the accuracy of existing ontology alignment solutions in a fast, scalable, and cost-effective manner.

The results reported by Mortensen et al. and Sarasua et al. suggest that microtask crowdsourcing could be a promising technique for the task of obtaining sanctioning rules in ICD-11, especially since ICD-11 and SNOMED CT have comparable concepts. However, scalability remained a significant challenge. The crowdsourcing of the hierarchical verification investigated in [26, 29] had much smaller search space than that for sanctioning rules. Taking the dermatology chapter of ICD-11 as an example, there are only around 7000 parent–child relationships that need to be verified. However, there are millions of possible combinations of diseases and anatomical locations. A naive method to examine all of them would be extremely expensive. Instead of asking

about all possible alignments, CrowdMap [28] applied an existing automatic algorithm to generate likely potential alignments and had the crowd assess this smaller set. Unfortunately, such automatic algorithms for identifying potential sanctioning rules were not available to us. Instead, we utilized the hierarchical structures of both diseases and anatomical locations to constrain the number of microtasks required for crowdsourcing sanctioning rules (Section 3.1).

2.4 Modeling Workers' Answers

In the crowdsourcing task described in this paper, we asked anonymous workers to select the anatomical locations where a disease can possibly occur. When workers disagreed with one another, we had to combine their collective answers to generate an estimate of the true answer for the question. A simple approach was to ask 10 workers to answer a multiple-choice question, and draw a conclusion based on the "majority vote" rule. However, questions had different levels of difficulty and it was wasteful to elicit 10 answers for each easy question. Furthermore, because crowd workers had different backgrounds, their medical knowledge and their willingness to spend time and effort on each microtask might vary. We should have higher confidence in the answers provided by workers with more medical knowledge and a strong reputation for answering correctly, as revealed in the accuracy rate of their past work. We should block workers who seemed to answer randomly. Modeling these factors helped us draw better conclusions with fewer solicitations for each microtask.

Much work had been done to model workers' revealed ability and to appraise appropriate confidence in crowdsourced answers. Whitehill et al. [30] described a task that used crowdsourcing to classify an image into binary categories (e.g., face/non-face, male/female, smile/non-smile). They presented a probabilistic graphical model and used it to simultaneously infer the categories of an image, the expertise of each worker, and the difficulty of each question. They demonstrated that their model outperformed the "majority vote" rule and was robust to noise. Bachrach et al. [31] developed a similar probabilistic graphical model that dealt with multiple-choice questions with only one correct answer. The model was used to determine which

observations to make next so as to improve the inferred knowledge according to a pre-determined criterion.

The work of Whitehill et al. [30] and Bachrach et al. [31] suggests that using probabilistic graphic models to represent workers' ability and appropriate confidence in the estimated true answer is a promising approach. However, we could not simply adapt their methods to solve our problem because our microtask required a different format. Our microtask involved a multiple-choice question in which each option was an anatomical location, and there might be multiple correct options. There were also three special options: *all*, *other*, and *none*, which we will show in Section 3.3 to be critical for reducing the cost of the crowd work. These special options introduced dependencies among the possible choices, which made assessing workers' answers more complex. Therefore it was necessary for us to develop a novel probabilistic graphical model for our microtasks. The model allowed us to use the results from earlier microtasks to inform our confidence in the answers obtained in subsequent microtasks.

Kittur et al. [32] showed that minor modifications in the configurations of crowdsourcing experiments could impact workers' performance considerably. In the biomedical ontologies domain, Mortensen et al. [33] studied workers' performance in terms of accuracy and speed of responses on many different configurations. They found that providing concept definitions improves workers' performance. Extending their idea, we formulated our microtask to optimize the accuracy of the responses while minimizing the cost (Section 3.3.2).

3 Materials and Methods

Unlike earlier versions of ICD, ICD-11 has a multi-components architecture consisting of a multi-hierarchical Foundation Component from which multiple use-case–specific mono-hierarchical classifications (e.g., classifications appropriate for mortality-reporting and primary-care use cases) can be derived [34]. Entities in the Foundation Component rely on a model of meaning, the *Common Ontology*, that is shared with SNOMED CT [35]. An *Extension Codes* chapter of ICD-11 enumerates the value sets to be used to refine the meaning of ICD-11 entities. In

the work reported in this paper, we drew subsets of diseases from the *Diseases of the skin* chapter of the ICD-11 Foundation Component as it existed in April 2013. To generate sanctioning rules constraining anatomical locations where the diseases could occur, we used a surface topographical hierarchical classification supplied by WHO staff. The parent/child relationships in the surface topographical hierarchy consisted mostly of part-of relationships (e.g., *Upper trunk* and *Lower trunk* as children of *Trunk*).

As described earlier, the microtask variant of crowdsourcing involves three steps: (1) defining the microtasks; (2) dividing the overall task into microtasks and publishing the microtasks on a crowdsourcing platform; and (3) collecting and combining answers. In Section 3.1 we describe the experiment we conducted to evaluate different ways of formulating questions concerning possible anatomical locations of a disease as microtasks; in Section 3.2 we discuss methods to divide the task into microtasks; and in Section 3.3 we show how to draw conclusions based on the crowdsourcing results. In Section 3.4 we describe a method to compare crowdsourced sanctioning rules with gold-standard rules to measure the quality of the crowdsourced sanctioning rules. Finally, in Section 3.5, we summarize the experiments where we applied and evaluated the methods developed in this section to obtain sanctioning rules through Amazon Mechanical Turk [36], the most popular crowdsourcing Internet marketplace.

3.1 Defining a Microtask

Our goal was to determine the legal combinations of diseases and anatomical locations such that a disease is associated with locations where it may occur. We could formulate the question we posed to crowd workers as a true/false question or as a multiple-choice question. In the true/false question formulation, we asked whether a disease could occur at a specific anatomical location. Figure 2 shows an example of a true/false question. To assist the crowd worker, we provided the definition of the disease if we could find it in the ICD-11 draft. Furthermore, if there was a Wikipedia page related to the disease or anatomical locations, we provided a link to the Wikipedia page; otherwise, we provided a link to the Google search results for the disease and anatomical

locations. We also provided the direct children of the anatomical locations in the hierarchy to help workers better understand the anatomical locations being considered as options.

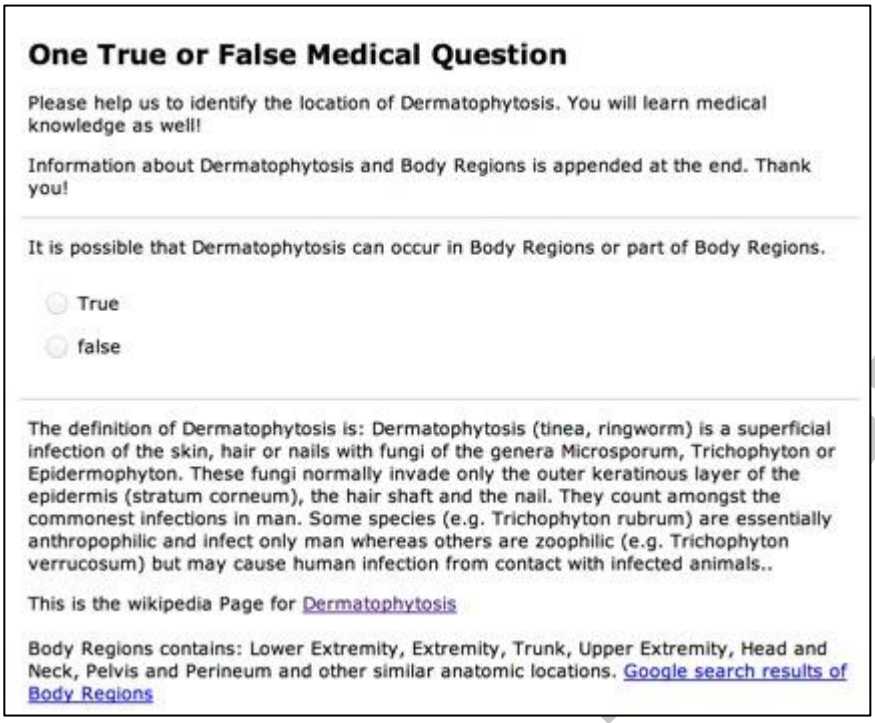


Figure 2 A sample true/false question to verify a possible location of a disease.

In the multiple-choice question formulation, each choice was an anatomical location and choices in one multiple-choice question were siblings in the location hierarchy. We asked workers to identify all the anatomical locations where the disease could possibly occur. Figure 3 shows an example of a multiple-choice question.

To compare the results from true/false questions and multiple-choice questions, we converted the answer to a multiple-choice question into equivalent answers to a set of true/false questions. For a pair of disease and anatomical location in a multiple-choice question, if the worker checked the location, we concluded that this action is equivalent to the worker checking *true* in the true/false question involving the disease and anatomical location. Similarly, if the worker did not check the location in a multiple-choice question, we concluded that this action is equivalent to the worker checking *false* in the corresponding true/false question. The option of *None of the above applies*

gave the worker a way to indicate that a question had been answered, even when no specific anatomical location was selected.

We performed an experiment to compare the results of using true/false questions versus using multiple-choice questions on the same set of diseases and anatomical locations combinations. For rosacea and related disorders in the dermatology chapter of the draft ICD-11, we formulated disease/location combinations as 113 true/false questions and 23 multiple-choice questions. For each question, we obtained 10 responses. For true/false questions, we paid workers 2 cents per response. For multiple-choice questions, we paid workers 4 cents per response. Wang et al. showed that a potential bonus increases the quality of responses [37]. Therefore we advertised that we would pay workers an additional 2 cents if their answers agreed with the reference answers provided by an expert dermatologist. We report the results of this experiment in Section 4.1. In the following sections, we assume the multiple-choice formulation is the preferred format and describe additional refinements of the choices presented to a crowd worker.

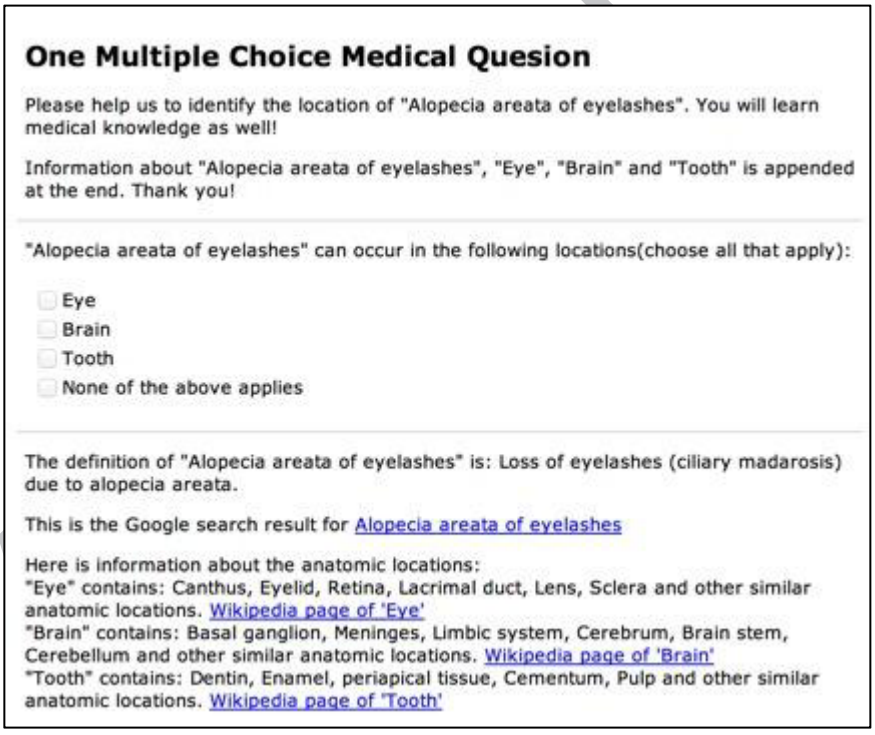


Figure 3 A multiple-choice question to solicit the possible locations of a disease.

3.2 Generating and Publishing Microtasks

To divide the task of finding sanctioning rules for combinations of diseases and anatomical locations into microtasks, we developed microtask-generating software that traversed both the disease hierarchy and anatomical-location hierarchy to generate possible anatomical locations of diseases. In this section, we describe how the software optimized the choices made available to crowd workers during the traversal (Figure 4).

Starting at the root of the disease hierarchy (e.g., *Palmoplantar keratodermas* in Figure 4) whose sanctioning rules we wanted to obtain, we needed to find sensible anatomical locations for that disease. We first displayed all top-level terms of the anatomical locations tree as options in a multiple-choice question, and then asked workers whether the disease could occur in those anatomical locations. To facilitate the traversal of the anatomical-location hierarchy and to reduce the cost, we slightly modified the multiple-choice question. To the anatomical locations being displayed as options, we added another option, 'any part of the *parent location*,' in which *parent location* was the parent of the current sibling anatomical locations (e.g., 'any part of *Body Surface Area*' in Figure 4). If we concluded, based on the answers obtained from the task workers, that the disease could occur in any part of *parent location*, then (disease, *parent location*) was a sanctioning rule and we could stop the traversal without looking at the children of the current options. If *any part of the parent location* was not selected and if workers chose a particular location as a sensible anatomical location, we recursively displayed all of that location's children as options in the next multiple-choice question and asked workers in a similar way. If workers indicated that the disease could not occur in a location, we eliminated all children of that location in the search space. If the traversal didn't stop until the program reached the leaf nodes, the sensible leaf nodes constituted the sanctioning rules.

When the microtask-generating software considered sanctioning rules for a disease in the hierarchy, it took advantage of the sanctioning rules that have been generated previously. We found that in the sample disease branches rooted at *Palmoplantar keratoderma*, 70% of the more specialized diseases in the hierarchy occurred in the same set or in a subset of the anatomical

locations of the diseases represented by their superclasses. Thus, when looking for the sanctioning rules of a child disease, it was wasteful to always start from the top of the locations hierarchy because we knew that the disease was highly likely to occur in exactly or in a subset of its parent's sensible anatomical locations. As a result, for a given disease, when we knew the sensible anatomical locations of the parent disease, the software started searching from the parent's sensible anatomical locations instead of from the top-level locations.

Nevertheless, in the *Palmoplantar keratoderma* sample, 30% of the more specialized diseases in the sample disease branch had anatomical locations other than their parents' locations. For example, *Papillon–Lefèvre syndrome*, also known as *palmoplantar keratoderma with periodontitis*, involves severe destruction of the periodontium. Thus, in addition to hand and foot, mouth is also a possible anatomical location where the disease has manifestation. Consequently, to the initial multiple-choice question on the location of a child disease, we added the option of *some other anatomical location*. If crowd workers thought the disease might occur somewhere other than the anatomical locations of the parent, we started the breadth-first search from the top of the anatomical-location hierarchy; otherwise, we could safely search within its parent's anatomical locations.

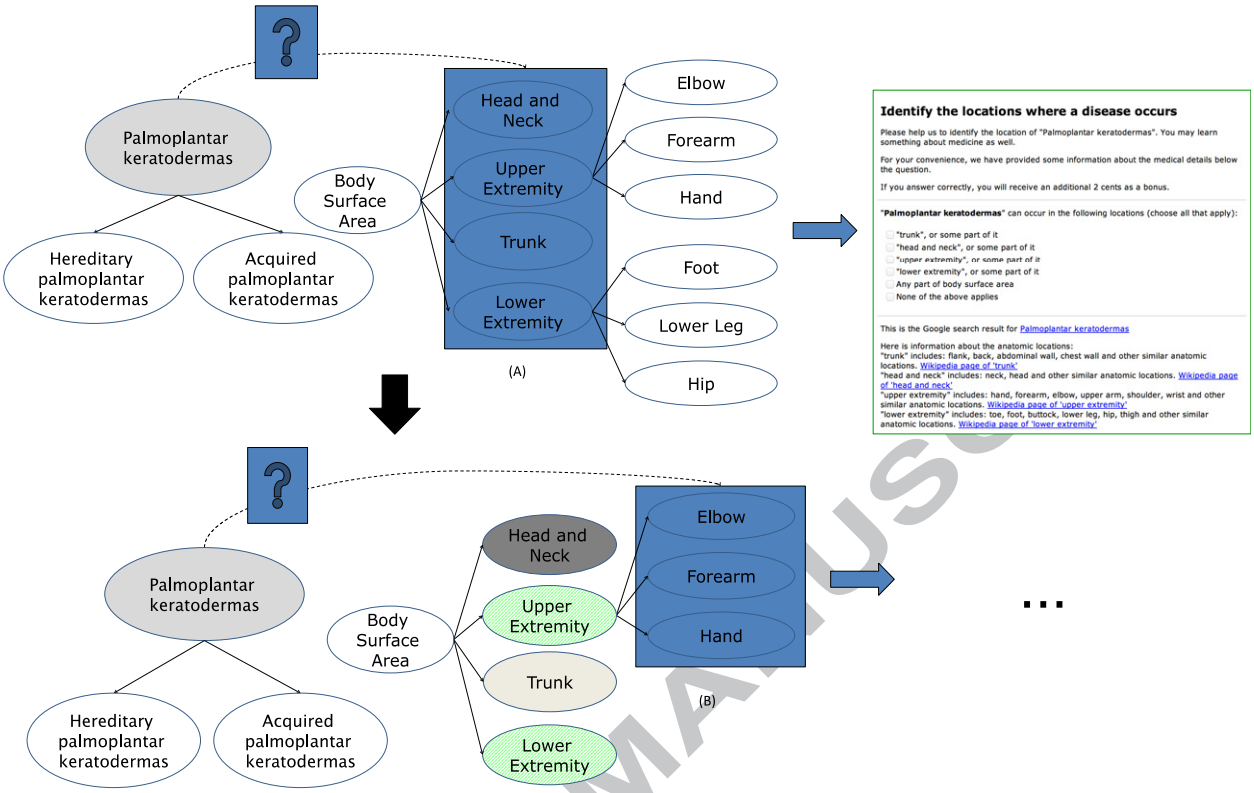


Figure 4 Traversal of hierarchies to generate multiple-choice questions. For a given disease such as *Palmoplantar keratodermas*, the task-generating software generated a multiple-choice question based on the sibling locations of an anatomical-parts hierarchy (A). An anatomical location was selected for further questioning if a majority of crowd workers selected it as a possible location of the disease. The task-generating software generated the next question based on the children of the selected location (B). The steps were repeated recursively until the sanctioning rules for the disease are found. Then the software attempted to find possible anatomical locations of the diseases more specific than the current disease (e.g., *Hereditary palmoplantar keratodermas*).

3.3 Combining Workers' Answers

Using the methods described in Sections 3.1 and 3.2, we conducted a crowdsourcing experiment where we paid 10 workers to answer each multiple-choice question (see Section 3.5). Initially we drew conclusions based on majority votes. As reported in Section 4.2, we were able to derive

high-quality sanctioning rules this way, but the cost was high relative to the number of sanctioning rules we needed to obtain. We observed that, because the difficulty level of each question varied, we did not necessarily require 10 answers for an easy question. We developed a Bayesian network to model the confidence levels of possible choices based on workers' responses. For each microtask, the Bayesian network was updated incrementally as additional answers were provided. Once the confidence of a particular choice passed an adjustable cutoff, we could stop asking more workers to perform the microtask. We further augmented the Bayesian network with a method to dynamically adjust the cutoff so as to derive the best sanctioning rules within a given budget.

The Bayesian model (described in detail in Section 3.3.1) incorporated crowd workers' accuracy rates to help estimate confidence levels in the choices. As described in Section 3.3.1.2, a worker's accuracy rate was updated incrementally each time the system determined the "true" answers to the questions posted in the microtasks. In our initial multiple-choice question exploration, some workers had an accuracy rate as low as 30%. In the evaluation experiments, we blocked a worker once her overall accuracy rate dropped below 55%. Therefore we only use answers from workers whose accuracy rate were 55% or higher to update Bayesian beliefs. The rationale for this threshold was that, when a worker's accuracy rate was below 50%, the confidence in an option she voted for would decrease instead of increase. The intuition is that if a worker was more likely to give wrong answers than right ones, her choice of a particular answer made that answer less likely. Questions that included the *some other anatomical location* option were especially hard to answer correctly. Because a worker's average accuracy rate for that option was often lower than 50% and such answers could not be used for Bayesian updates, we had few usable answers to update Bayesian models that include the *other* option. Because we could not develop alternative Bayesian models without changing the way questions were generated, we experimented with a logistic regression model that did not prove very satisfactory. In the end we left the handling of the *other* option for future work.

Even with heuristics to generate fewer microtasks and Bayesian models to reduce the number of needed responses, the cost would still be a major barrier for crowdsourcing sanctioning rules

because there are tens of thousands of diseases in ICD-11 and similar large number of codes in the *Extension Codes* chapter. One major tuning parameter for the Bayesian network model was the probabilistic cutoff for our confidence about the conclusion for each microtask beyond which we cease asking additional workers to respond to a microtask. A higher probabilistic cutoff allowed us to be more confident about our conclusions, implying higher-quality sanctioning rules. The downside was that collecting more responses for each microtask increased the cost. By adjusting the probabilistic cutoff, we could adjust the trade-off between quality and cost. Thus, we developed methods that, within a fixed budget, chose the highest possible probabilistic cutoffs that generate sanction rules for the largest number of diseases (Section 3.3.2).

The final architecture of our crowdsourcing software is illustrated in **Figure 5**. The microtask-generation component traversed disease and anatomical-location hierarchies to generate microtasks to identify sensible anatomical locations for specific diseases (A). The microtasks were sent to the component that solicited and collected responses on the Amazon Mechanical Turk platform (B). The Bayesian component generated appropriate Bayesian networks for the microtasks and maintained the accuracy profiles of workers (C). Upon receiving a worker's answer, it updated the Bayesian networks to determine our confidence in each possible answer to the question. If the confidence of one option passed a threshold, the system stopped recruiting more workers for that microtask. The budget-monitoring component (D) set the confidence cutoffs to maximize the product of confidence and the number of completed tasks, given a budget constraint.

Using the data obtained from the 10-answer experiment, we simulated a crowdsourcing experiment that applied the budget-constrained Bayesian method to derive sanctioning rules from the answers supplied by the crowd workers.

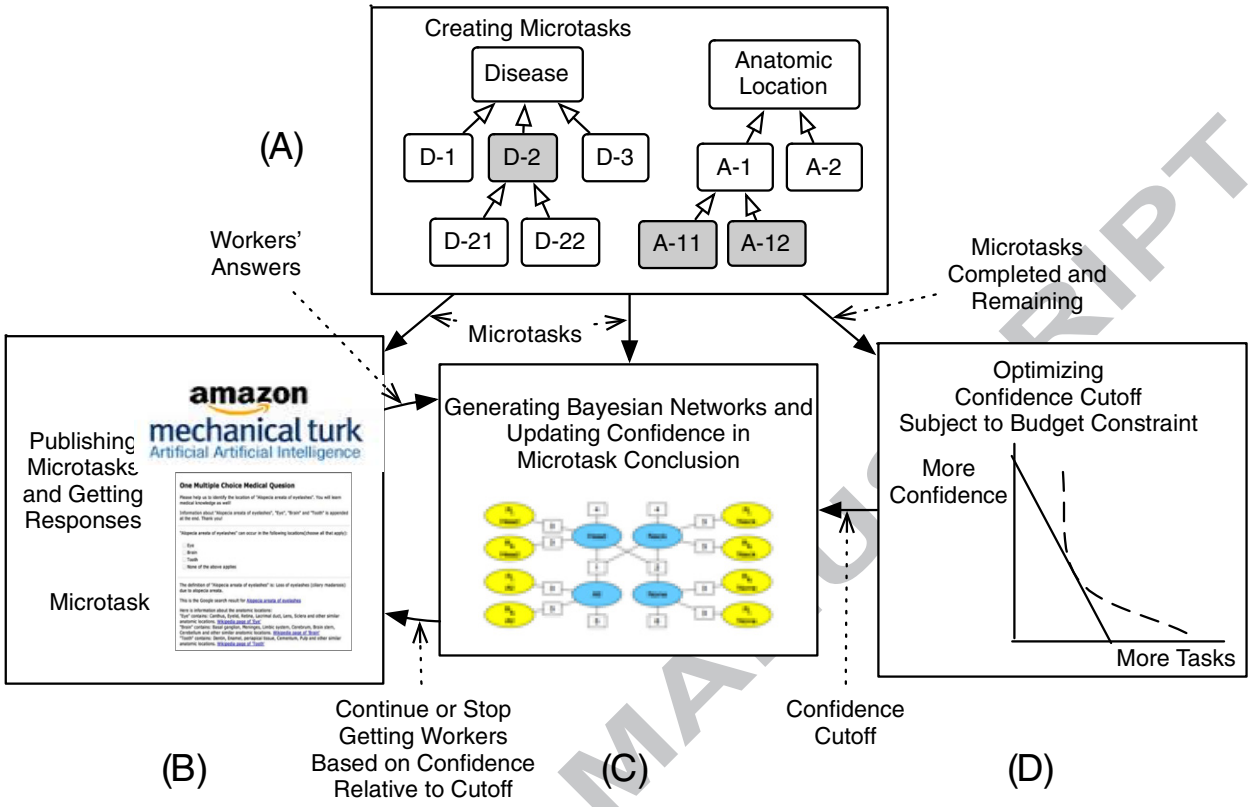


Figure 5 Architecture of the budget-constrained Bayesian crowdsourcing method to acquire sanctioning rules. The system created microtasks by traversing the disease and anatomical-location hierarchies (A); published the microtasks and solicited responses on the Amazon Mechanical Turk platform (B); generated Bayesian networks and updated confidence about conclusions derived from responses to the microtasks (C); and determined the optimal confidence cutoffs, subject to budget constraint, for stopping to solicit more responses to a microtask (D).

3.3.1 Bayesian Model

We divide the description of the Bayesian model used to derive sanctioning rules from crowd workers' responses into three subsections: Section 3.3.1.1 describes the Bayesian graph representation; Section 3.3.1.2 gives an equivalent factor graph representation that is

computationally more efficient to update; and Section 3.3.1.3 describes the run-time operation of the model.

3.3.1.1 Bayesian Graph Representation

Given a simplified version of a multiple-choice question—for example, a disease that could occur in (a) Head, (b) Neck, (c) Any part of Head and Neck, (d) None of the above locations—we denoted the fact that the disease could occur in the head as *head*+, and as *head*- if it could not occur in the head. We also denoted the option "Any part of X" as *all* and "None of the above locations" as *none*. An example of a valid answer was (*head*-, *neck*+, *all*-, *none*-), indicating that *neck* was a sensible anatomical location of the disease in question. There were dependencies among the options: If *all*+ is selected, then the disease could occur in all locations in the question (in this case, *head*+, *neck*+); if *none*+ is selected, then the disease could not occur in any of the locations in the question (in this case, *head*-, *neck*-). It is easy to see that the number of valid answers for this question was limited. We used Bayesian networks to model the probability that each valid answer was the correct answer. Our implementation of Bayesian networks and related factor graphs used standard methods [38].

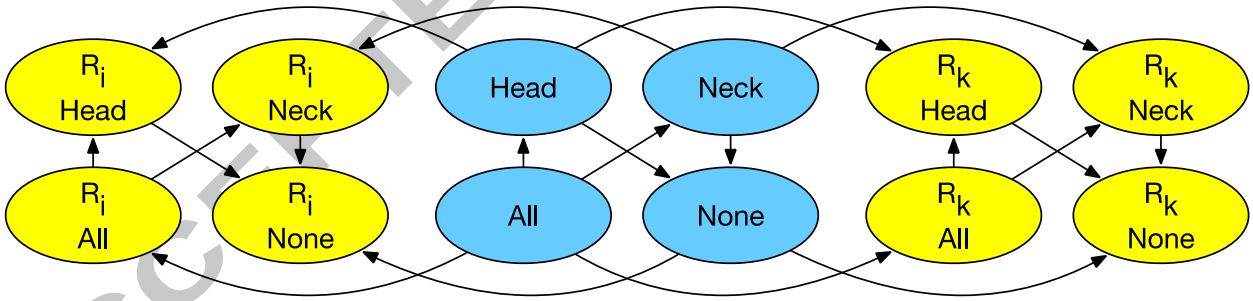


Figure 6 A naïve Bayesian model of the answers to a multiple-choice question (darker-shaded ovals) and the answers from crowd workers i and k (lighter-shaded ovals). R_i and R_k are the responses from workers i and k, respectively.

Figure 6 shows a naïve Bayesian network where a disease could occur in two possible anatomical locations. (In the implementation, any number of possible locations was allowed.) The darker-shaded nodes represent the true answers and the lighter-shaded nodes represent workers'

responses. Each node has two possible values, positive or negative. $R_{i_{head}}$ represents worker i's vote on the head option; it could be *head*+ or *head*-. The arrows from darker-shaded nodes to lighter-shaded nodes indicate that the crowd workers' answers are influenced by the true answers. The probability that the lighter-shaded nodes agree with their corresponding darker-shaded nodes is the worker's accuracy rate, while the probability that lighter-shaded nodes disagree with their corresponding darker-shaded nodes is (1- worker's accuracy rate). The arrows from *all* to *head* and *neck* represent the dependencies between anatomical location *all* and the head and neck locations. The arrow from *all* to *head* is interpreted as "if *all*+, then *head*+." Similarly, the arrow from *head* to *none* is interpreted as "if *head*+, then *none*-." Similar dependencies exist within workers' answers as well. There was no dependency between answers of different workers, and disagreements between different workers were allowed. Because each worker's answers should be consistent, we could filter workers' answers and ignore the invalid ones, such as (*head*-, *neck*+, *all*+, *none*-). Because of the filtering, we could assume that all workers' answers were valid and, therefore, that the dependencies between the lighter-shaded nodes could be removed (Figure 7).

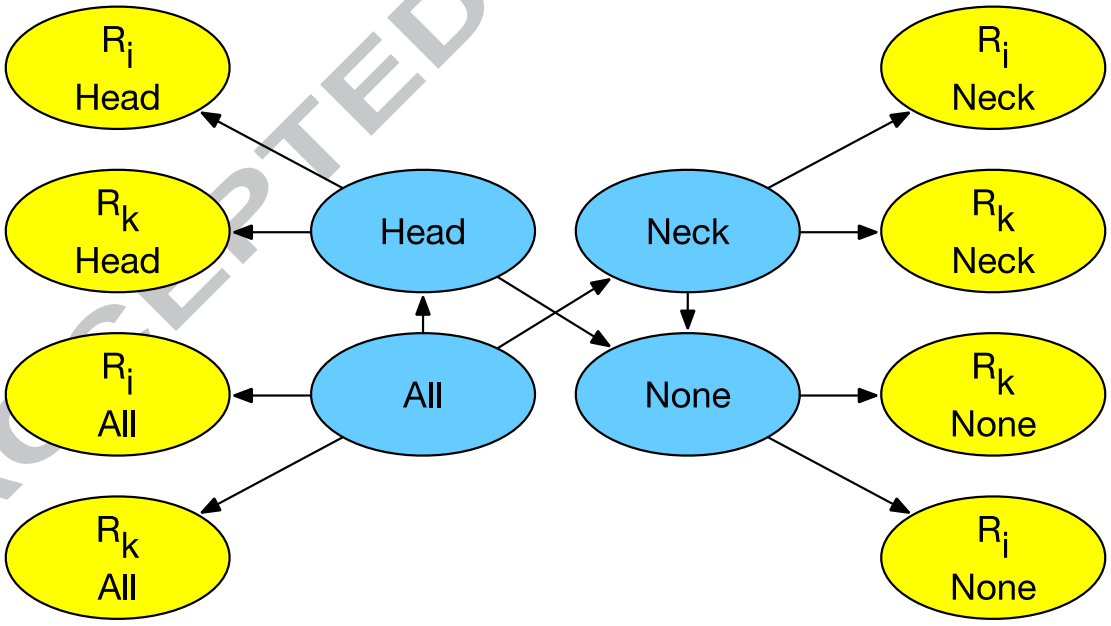


Figure 7 Revised Bayesian model with dependencies among microtask answers removed. It assumes that only valid crowd worker answers are used to update the model and therefore we don't need to model the dependencies in a single worker's responses.

The network shown in Figure 7, with fewer dependencies, is much simpler than the original network in Figure 6.

3.3.1.2 Factor-Graph Representation

Given that the sum-product inference algorithm [38, 39] we used to update the network was based on the *factor graph* formulation of the network, we implemented this network as a factor graph, as shown in Figure 8.

A factor graph formulation of a Bayesian network is a bipartite graph representing the factorization of the joint distribution depicted in the Bayesian network. Given a factorization of a probability density function of n variables into m factors

$$P(X_1, X_2, \dots X_n) = \prod_{j=1}^{m} f_j(S_j)$$

where f_j is a function with domain S_j, a subset of $(X_1, X_2, \dots, X_n)$, the factor graph consists of nodes $(X_1, X_2, \dots, X_n)$ U $(f_1, f_2, \dots, f_m)$ and undirected edges between factor node and variable nodes. An edge between a factor f_j and variable X_i exists if and only if X_i is in the domain of f_j, i.e., X_i is a member of S_j. Graphically each factor is represented as a square connecting to a number of variable nodes. For factor graphs derived from Bayesian networks, the functions incorporate information about the prior probabilities of events represented by the nodes and about the conditional probabilities between nodes.

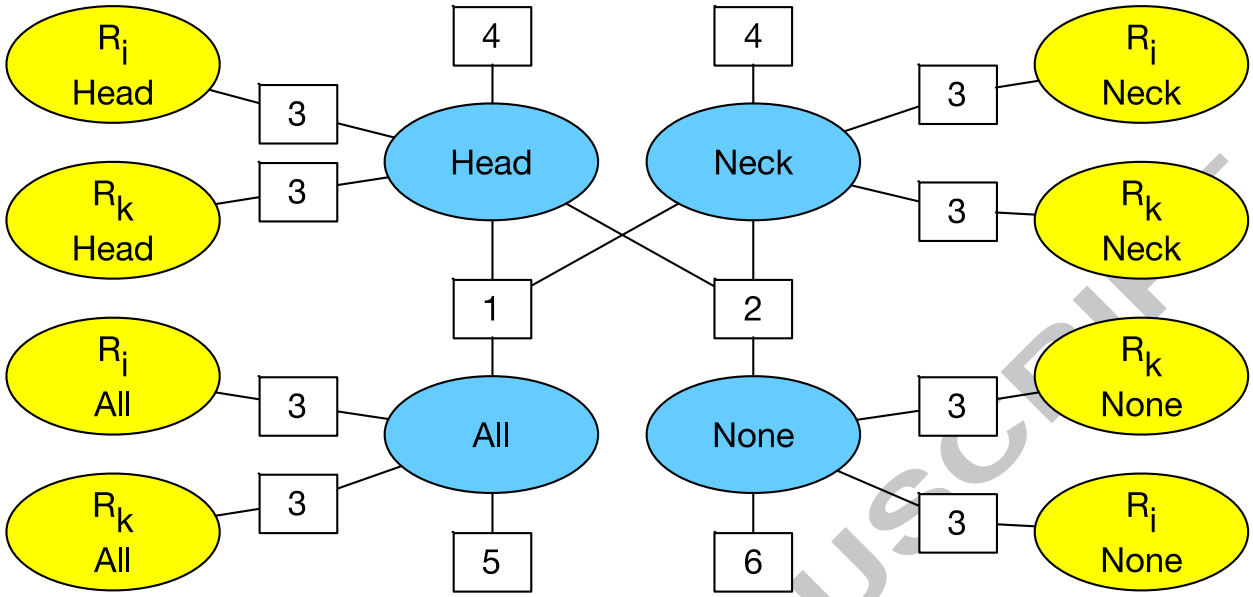


Figure 8 Factor graph implementing valid answers in the Bayesian network. Factors 4, 5 and 6 represent the prior probabilities that the Head, Neck, all, or none of them are in the sanctioning rule. Factors 1 and 2 encode the dependencies among the Head, Neck, All, and None nodes. Factor 3 represents the accuracy of the crowd worker.

The factor graph we derived from the Bayesian network in Figure 8 had three factors (Factor 4, 5, 6) that represented prior probabilities of the variables. We learned the factor functions with a randomly selected training dataset that had 236 questions. To determine the prior probability of a particular answer, we took the ratio of the times each option was selected to the total number of times these options were presented in the questions. For example, Factor 4 represents the chance that any one location—head or neck—is either true or false. The resultant function has the values 0.692 and 0.308 for head+ and neck+, respectively (Table 1). Factor 5 represents the chances that *all* is true or false (Table 2), and Factor 6 indicates the chances that *none*+ is true or false (Table 3).

Table 1. The function associated with Factor 4. It represents a sample prior probability of *Head* or *Neck*.

Head/Neck	Value
+	0.692

-	0.308

Table 2. The function associated with Factor 5. It represents a sample prior probability of *All*.

All	Value
+	0.707
-	0.293

Table 3. The function associated with Factor 6. It represents a sample prior probability of *None*.

None	Value
+	0.0405
-	0.9595

The function of Factor 1, shown in **Table 4**, represents legal values of *head* and *neck*, given possible values of *all*. The function of Factor 2, shown in **Table 5**, represents legal values of *head* and *neck*, given possible values of *none*. The two functions ensure that the factorization evaluates to zero for any illegal combinations of *head*, *neck*, *all*, and *none*. **Table 6** shows the function associated with Factor 3. It represents the definition of the accuracy of a crowd worker's answer.

Table 4. The function represented by Factor 1 maps possible values of *All*, *Head*, and *Neck* to 1 or 0.

All	Head	Neck	Value
+	+	+	1
+	+	-	0
+	-	+	0
+	-	-	0
-	+	+	0

-	+	-	1
-	-	+	1
-	-	-	1

Table 5. The function represented by Factor 2 maps possible values of *None*, *Head*, and *Neck* to 1 or 0.

None	Head	Neck	Value
+	+	+	0
+	+	-	0
+	-	+	0
+	-	-	1
-	+	+	1
-	+	-	1
-	-	+	1
-	-	-	0

Table 6. The function associated with Factor 3. It represents the definition of the accuracy of a task worker's answer. The variable $R_{i_{head}}$ represents the i^{th} worker's choice of whether *head* is a permissible answer and $head_{true}$ represents the true state of *head*.

$R_{i_{head}} = head_{true}$	Value
True	$P_{i,accuracy}$
False	1- $P_{i,accuracy}$

To build the factor graph, we also needed to learn $P_{i,accuracy}$, the accuracy of the i^{th} worker, from the training dataset. A naïve method is simply to measure the accuracy of each worker's response

to each question. However, if one worker answered only a small number of questions in the training dataset, and happens to have a 100% accuracy rate, we could not assume the worker's real accuracy rate is 100% for subsequent questions. An example will illustrate the problem introduced by 100% accuracy rates. Assume the true answer is *head-*, and 5 workers vote for *head-*, so P (*head-*) should be high. But a 6th worker, whose accuracy rate based on the training set is 100%, votes for *head+*. According to Factor 3, because the worker's accuracy rate is 100%, *head+* should be the true answer. Then, P (*head-*) should equal 0, which is the opposite of the true answer. To solve this problem, we assumed a worker's accuracy rate follows a Bernoulli distribution. A conjugate prior to the Bernoulli distribution is the Beta distribution, so we used Beta(6,4) as the prior estimate for all workers' accuracy rates, which simply meant that we assumed each worker voted for 10 options and got 6 right. In this way, no worker gets a 100% accuracy rate, and our Bayesian network analysis is more reliable and robust to noise. Moreover, since workers who answered randomly harm the performance of the Bayesian network, we blocked workers whose accuracy rate was lower than 55%.

With all of the estimated factors in the factor graph, we could calculate the probability of each possible answer—e.g., the probability that a worker would answer (*head-*, *neck+*, *all-*, *none-*)—with standard techniques [38].

3.3.1.3 Run-Time Updates

At runtime, for each microtask, the system generated a Bayesian network similar to that of Figure 7. Every time we got a response from a new worker, we added the corresponding Factor 3 to the factor graph, solved the factor graph, updated the distribution for the true answer, and found the most promising answer (that with the highest probability). When a response came from an existing worker, we similarly updated the distribution for the true answer. To avoid overfitting, we got at least two answers for each microtask. When we were sufficiently confident about our answer (when the probability was higher than a cutoff), we stopped asking for more responses, drew a conclusion, and moved on to the next question.

In Section 3.3.2, we will discuss the derivation of the probabilistic cutoff that defined when we were "sufficiently confident."

3.3.2 Crowdsourcing Within a Budget

Given the astronomical number of disease-location combinations when ICD-11 diseases are tested against a real-life anatomical location hierarchy, a crowdsourcing solution for obtaining sanctioning rules remains expensive even if we pay pennies per question and use the optimization methods described in the earlier sections. We may want to acquire highly accurate crowdsourced rules by getting more answers to satisfy more stringent confidence levels. However, in real life such work is constrained by finite budgets. With the goal of obtaining first-draft sanctioning rules that reduce the burden of manual checking, it is acceptable to have false positive and false negative rules. In a realistic deployment scenario, we would try to obtain the best possible sanctioning rules without exceeding a given budget constraint.

The overall cost of crowdsourcing sanctioning rules could be reined in by adjusting the probabilistic cutoff for deciding that the current conclusion is good enough. This would require building a model to predict the cost of obtaining crowdsourced sanctioning rules for a range of probabilistic cutoffs. For a given budget, we could then choose the highest cutoff whose overall cost is within the budget. However, predicting costs in this scenario is difficult for two reasons. First, the cost to solicit the sanctioning rules depends on many factors and varies dramatically from disease to disease. For instance, if a disease is a common disease for which most workers know its sensible anatomical locations, workers typically agree with one another and we can gain high confidence with a small number of answers for the microtask. Similarly, if more specific diseases occur in the same anatomical locations as their more general parent classes, then we do not need to traverse much of the anatomical hierarchy, which reduces the number of microtasks that have to be assigned. Conversely, little-known diseases with non-obvious anatomical locations may require us to obtain more answers for more microtasks. Second, we need to estimate a cost model for each of the probabilistic cutoffs. Even if we limit the possible probabilistic cutoffs to one of 0.1, 0.2, …

0.9, we still need to estimate 9 cost distributions. The more cost distributions we have, the more parameters we need to learn, which adds to the time needed to perform the analysis.

Because of these problems, it was difficult to set a probabilistic cutoff for determining the "correct" answers before launching our crowdsourcing effort. Instead, we developed a cost-management approach that would automatically monitor the progress of the crowdsourcing project and incrementally adjust the cutoff based on updated cost distributions for cutoffs, remaining number of diseases, and remaining budget. Before starting to obtain sanctioning rules for a new disease, the monitor would calculate the cost distribution for each cutoff, compare the result with remaining budget, and choose the highest probabilistic cutoff that would satisfy the budget constraint. When all sanctioning rules for a disease were found, the monitor would update the cost distribution and continue the process for other diseases.

It is clearly impractical for a real human monitor to oversee the progress of the crowdsourcing all the time. Instead, we applied reinforcement learning [40] to develop an agent that carried out the strategy described above. The algorithm for the operation of the agent is as follows:

1. Start at the top level of the disease hierarchy, post the microtasks needed to get sanctioning rules for a small number of diseases. Ask for 10 responses to each microtask. For each microtask, use the methods described in Sections 3.2 and 3.3.1 to obtain the number of answers needed at each cutoff in the set [0.1, 0.2, … 0.9]. From the number of needed answers, calculate the cost of getting sanctioning rules of the diseases for each cutoff.
2. For each cutoff, fit the cost of obtaining sanctioning rules for a disease with a normal distribution.
3. Get the remaining budget and the remaining number of diseases.
4. Choose the probabilistic cutoff c from [0.1, 0.2, … 0.9] to maximize the expected reward function E (Reward), where

   Reward =number of finished diseases at the end * cutoff –

   number of unfinished diseases at the end
5. Traverse the disease hierarchy to identity the next disease to work on.
6. Apply the methods in Sections 3.2 and 3.3.1 to solicit sanctioning rules for that disease with cutoff c.

7. Update the cost distribution for cutoff c, and update remaining budget and remaining number of diseases.
8. If number of remaining diseases is 0 or remaining budget is smaller than 40 cents, stop the algorithm; otherwise, go to Step 4.

The intuition for the reward function was that we wanted high cutoffs for most diseases, but we wanted to penalize situations where we run out of budget to solicit sanctioning rules for remaining diseases. Because we could know the real reward only after we finished the task, we had to maximize an expected reward function before completing the task:

E (Reward) = E (number of finished disease at the end) * cutoff –

E (number of unfinished disease at the end)

To estimate the expected number of finished diseases at the end, we assumed a normal distribution for the cost of obtaining sanctioning rules for each disease and modeled the cost distribution for each cutoff. At iteration j of the above algorithm, we know N, the number of remaining diseases, and B, the remaining budget. We can compute the cost distributions for finishing k diseases, where k ranges over 1 to N, as the sum of k identical random variables, each of which is the cost of obtaining the sanctioning rules for one disease at a specific cutoff. The probability p_k of completing k more diseases within the remaining budget B is just the cumulative probability of the cost of completing k diseases being less than B. Thus at iteration j of the algorithm,

E(number of finished disease at the end) =

(the number of finished diseases at iteration j) + $\sum_{k=1}^{k=N} p_k * k$

Similarly, at iteration j, we can compute the expected number of unfinished diseases at the end.

3.4 Evaluation Metric

Accuracy, sensitivity and specificity are standard measures for the performance of binary classification functions. For a given disease, a set of sanctioning rules can be seen as a classification function that classifies each anatomical location into two categories: (1) sensible location or (2) nonsensical location for that disease. Thus, sensitivity and specificity can be used to

compare gold-standard sanctioning rules with sanctioning rules derived from an experimental method.

For a specific disease D, suppose the gold-standard sanctioning rules are [(D, a_1), …(D, a_k)] and the crowdsourced sanctioning rules are [(D, b_1), …(D, b_m)], where a_1, … a_k, and b_1…b_m are anatomic locations. We partition the anatomic location hierarchy H as follows:

1. Let A_i, i=1…k and B_j, j=1…m be the sets of anatomical locations consisting of a_i and b_j and their descendants in H respectively.
2. P, the set of permitted locations according to the gold-standard rules, is $\bigcup_{i=1}^{k}(A_i)$, the union of A_is.
3. N, the set of non-permitted locations according to the gold-standard rules, is H – P
4. Similarly, let P_c and N_c be the permitted and non-permitted locations according to the crowdsourced sanctioning rules.
5. TP, the set of true positive locations of crowdsourced sanctioning rules, is the intersection of P_c and P.
6. FP, the set of false positive locations of crowdsourced sanctioning rules, is the intersection of P_c and N.
7. TN, the set of true negative locations of crowdsourced sanctioning rules, is the intersection of N_c and N.
8. FP, the set of false negative locations of crowdsourced sanctioning rules, is the intersection of N_c and P.

Using these terms, we can use standard definitions for the evaluation metrics accuracy, sensitivity, and specificity.

A toy example illustrates the application of the evaluation metric to the sanctioning rules for one disease. Suppose that, on the one hand, the crowdsourcing method gives us two sanctioning rules for the disease *Palmoplantar keratoderma*:

(Palmoplantar keratoderma, Hand)

(Palmoplantar keratoderma, Lower Leg)

On the other hand, suppose that the gold standard contains these two sanctioning rules about *Palmoplantar keratoderma*:

(Palmoplantar keratoderma, Hand)

(Palmoplantar keratoderma, Foot)

Figure 9 shows a simplified anatomical location graph. Dark rectangles (*Hand* and *Foot*) are those anatomical locations mentioned in the gold-standard sanctioning rules, whereas dark shapes with hatched interiors (*Hand* and *Lower Leg*) are anatomical locations mentioned in the sanctioning rules found through crowdsourcing. The seven rectangles are anatomical locations permitted by the gold-standard sanctioning rules (P). The nine ovals are anatomical locations not permitted by the gold-standard sanctioning rules (N). The three rectangles that have hatched interiors are true positives (TP). The hatched ellipse is a false positive (FP). The eight non-hatched ovals are true negatives (TN), and the four rectangles with no hatched interiors are false negatives (FN). With these numbers, the accuracy ((TP+TN)/(P+N) = (3+8)/16), sensitivity (TP/P = 3/7) and specificity (TN/N = 8/9) of the crowdsourced sanctioning rules are 0.69, 0.43 and 0.89 respectively.

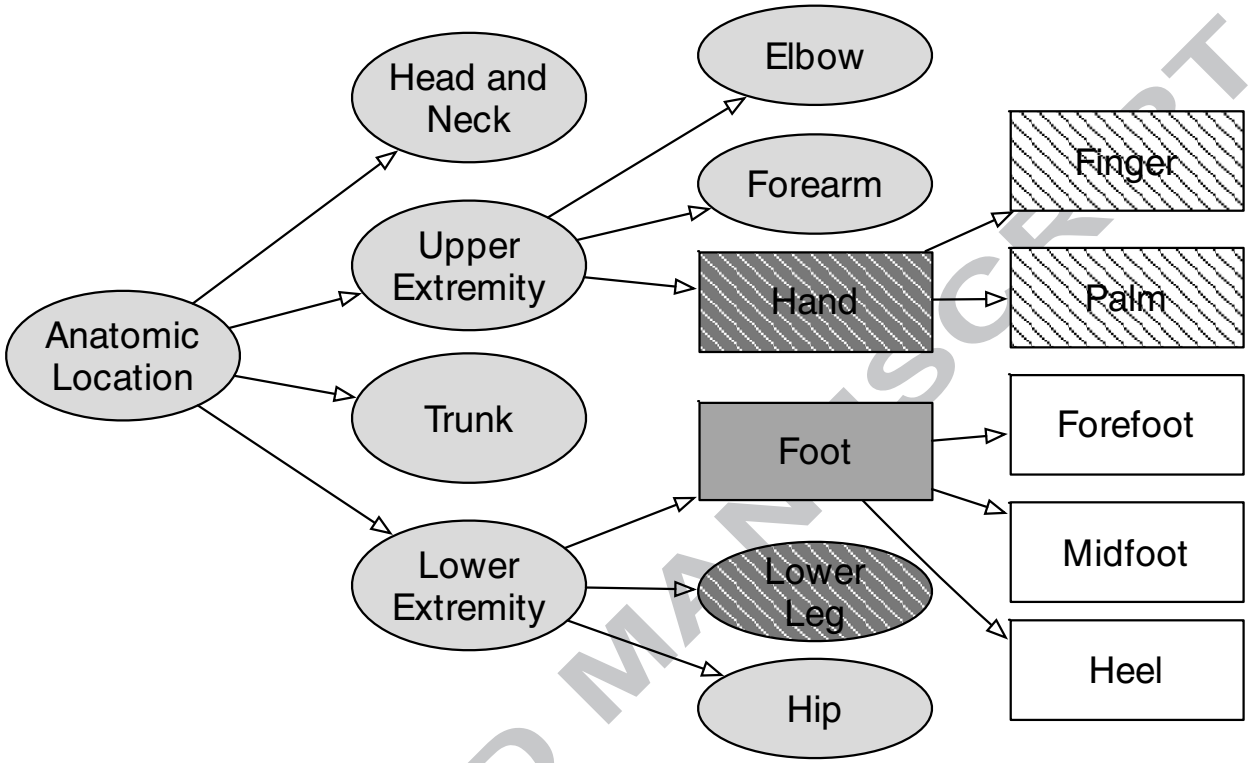


Figure 9 Anatomical locations sanctioned by gold-standard sanctioning rules (rectangles) and by crowdsourced sanctioning rules (hatched shapes). Dark rectangles (*Hand* and *Foot*) are those anatomical locations mentioned in the gold-standard sanctioning rules, whereas dark shapes with hatched interiors (*Hand* and *Lower Leg*) are anatomical locations mentioned in the sanctioning rules found through crowdsourcing. The seven rectangles with white or hatched interiors are anatomical locations permitted by the gold-standard sanctioning rules. The four shapes with hatched interiors are anatomical locations permitted by the crowdsourced sanctioning rules.

3.5 Crowdsourcing Experiments to Obtain Sanctioning Rules

We conducted two experiments to evaluate our crowdsourcing methods for acquiring sanctioning rules. For both experiments, we used ICD-11 disease branches rooted at *Dermatoses of the head, neck and oral cavity*, which contained 201 diseases in the April 2013 draft of ICD-11 Foundation

Component. To obtain crowdsourcing data on multiple samples, we divided the diseases into 11 sub-branches (see Table 8), where each disease branch included 19 diseases on average. For anatomical locations, as mentioned in the introduction to the Materials and Methods section, we used a 335-node tree of topographical locations, supplied by WHO staff.

In the first experiment ("Majority Vote"), we applied the microtask-generation methods described in Section 3.1 and 3.2 to create multiple-choice questions that we posted on the Amazon Mechanical Turk platform. We obtained 10 responses for each multiple choice question, and used the "Majority Vote" method for combining responses from crowd workers to identify sanctioning rules for a disease branch in the hierarchy. We compared the crowdsourced sanctioning rules with a gold standard supplied by one of the authors (RJGC), who was the Chair of the Dermatology Topic Advisory Group for the ICD-11 revision. We computed sensitivity and specificity of the crowdsourced sanctioning rules using the method described in Section 3.4.

In the second experiment ("Budget-Constrained Bayesian Network"), we simulated a crowdsourcing experiment where we applied the budget-constrained Bayesian methods discussed in Section 3.3 to the data obtained in the first experiment. For each disease, instead of using all 10 responses obtained for each question, we stopped looking at additional responses once our confidence in the conclusions reached the threshold computed by the system.

We randomly split part of our experimental data obtained in the first experiment into training and test datasets. As discussed in Section 3.2, there were two kinds of multiple-choice questions: questions with the option to select *any part of the parent location*, which we refer to as *all* questions; and questions that included *some other anatomical location*, which we refer to as *other* questions. In the training dataset, there were 236 questions involving 82 diseases, while the test dataset included 148 *all* questions and 62 *other* questions for 71 diseases. As discussed in Section 3.3, crowd workers' accuracy in answering *other* questions often falls below 50%, making Bayesian updates unsuitable. Therefore, we only report the results for the *all* questions.

In this experiment, we compared the cost and accuracy of the majority-vote and budget-constrained Bayesian methods, where the cost was the amount of money paid to crowd workers whose answers we obtained before reaching the cutoff and the accuracy is the proportion of correct responses among the responses used in each method.

4 Results

In this section, we report the results of (1) using true/false questions versus using multiple-choice questions on the same set of diseases and anatomical location combinations (Section 4.1); (2) using the 'Majority Vote' method to derive sanctioning rules (Section 4.2); and (3) using the "Budget-Constrained Bayesian Network" method to derive the sanctioning rules (Section 4.3).

4.1 Results of True/False versus Multiple-Choice Formulation of a Microtask

Table 7 shows the results of the experiment comparing true/false and multiple-choice formulations of our microtasks, as described in Section 3.1. The questions were formulated using a subset of disorders related to hair and the hair follicle in the dermatology chapter of the draft ICD-11.

Table 7. Comparing results of true/false and multiple-choice questions.

	Binary question	Multiple-choice question
Number of questions	113	23
Number of responses	1130	230
Accuracy rate	76%	77%
Cost per response	2.5 cents	4.5 cents
Cost per response per combination of disease and location	2.5 cents	0.9 cents

The accuracy rates for true/false questions and multiple-choice questions were similar, but it was much less costly to get one response to a multiple-choice question about combining a disease with various locations than it was to pay for a series answers to several true/false questions. Thus, in subsequent experiments, we decided to use the multiple-choice format for our microtask.

4.2 Results of the "Majority Vote" Experiment

Using the metrics defined in Section 0, we calculated the sensitivities and specificities of crowdsourcing sanctioned anatomical locations of the dermatological diseases in the 11 hierarchies chosen for the formative experiment (Table 8). **Table 9** shows the gold-standard and crowdsourced sanctioning rules for the computed evaluation metrics for every disease in the *Infectious disorders of the external ear* branch of the draft ICD-11 skin-diseases chapter.[1] Table 10 shows the average sensitivities and specificities of the sanctioning rules across all diseases.

We asked 641 questions; we received 6410 responses from 135 distinct workers; and we paid 3296 bonuses. We paid workers 4 cents per response and 2 cents per bonus. Amazon Mechanical Turk charged 0.5 cents for each transaction. Therefore, the average cost to acquire sanctioning rules per disease is (6410*$0.045+3296*$0.025)/201, or $1.85 USD.

Table 8. Sensitivities and specificities of crowdsourced sanctioning rules for branches of the skin diseases under *Dermatoses of the head, neck, and oral cavity*.

Root concept of skin disease branch	Number of Diseases	Sensitivity		Specificity	
		Average	Standard Deviation	Average	Standard Deviation
Dermatoses of the scalp	34	0.98	0.14	0.97	0.06
Infective disorders of the external ear	20	0.66	0.39	1.0	0.0
Inflammatory disorders of the external ear	11	1.0	0.0	1.0	0.0
Certain specified disorders of external ear	12	0.86	0.33	1.0	0.0
Infectious disorders of eyelid	19	0.45	0.50	0.91	0.23
Disorders of lips	33	1.0	0.0	0.99	0.05
Disturbances of oral epithelium	10	0.4	0.52	0.97	0.02
Lichen planus and lichenoid	12	0.92	0.30	0.92	0.02

[1] The crowdsourced and gold-standard sanctioning rules and the sensitivities and specificities of the crowdsourced rules involved in this experiment can be found at [42].

reactions of oral mucosa					
Non-infective erosive and ulcerative disorders of oral mucosa	30	1.0	0.0	0.95	0.04
Acquired disorders of eyelashes	9	0.58	0.46	0.98	0.02
Inflammatory disorders of eyelid	11	1.0	0.0	1.0	0.0

Table 9. Gold-standard and crowdsourced sanctioning rules and computed evaluation metrics for the *Infectious disorders of the external ear* branch of the draft ICD-11 skin-disease chapter.

Disease	Gold standard	Crowdsourced locations	TP	TN	FP	FN	Sensitivity	Specificity
Infective disorders of the external ear	outer ear	external auditory canal, pinna	20	314	0	1	0.95	1.00
Otomycosis	outer ear	external auditory canal, pinna	20	314	0	1	0.95	1.00
Otomycosis due to Candida	outer ear	external auditory canal	3	314	0	18	0.14	1.00
Aspergillus otomycosis	outer ear	external auditory canal	3	314	0	18	0.14	1.00
Malignant otitis externa	outer ear	external auditory canal, pinna	20	314	0	1	0.95	1.00
Herpes simplex infection of external ear	outer ear	external auditory canal, pinna	20	314	0	1	0.95	1.00
Other infective otitis externa	outer ear	external auditory canal, pinna	20	314	0	1	0.95	1.00
Acute bacterial inflammation of external ear	outer ear	external auditory canal	3	314	0	18	0.14	1.00
Acute diffuse otitis externa	outer ear	external auditory canal	3	314	0	18	0.14	1.00
Tank ear	outer ear	external auditory meatus, tympanic membrane, pinna, apex of pinna	19	314	0	2	0.90	1.00
Acute infective otitis externa	outer ear	external auditory canal, pinna	20	314	0	1	0.95	1.00
Beach ear	outer ear	external auditory canal	3	314	0	18	0.14	1.00
Haemorrhagic otitis externa	outer ear	external auditory canal	3	314	0	18	0.14	1.00
Acute bacterial otitis externa	outer ear	external auditory canal, obverse of pinna, conchal bowl of pinna, pinna	20	314	0	1	0.95	1.00
Diffuse otitis externa	outer ear	external auditory canal	3	314	0	18	0.14	1.00
Infection of external ear	outer ear	external auditory canal, pinna	20	314	0	1	0.95	1.00
Abscess of external ear	outer ear	external auditory canal, pinna	20	314	0	1	0.95	1.00
Erysipelas of external ear	outer ear	external auditory canal, pinna	20	314	0	1	0.95	1.00

Perichondritis of external ear	outer ear	external auditory canal, pinna	20	314	0	1	0.95	1.00
Cellulitis of external ear	outer ear	external auditory meatus, pinna	18	314	0	3	0.86	1.00
Average							**0.66**	**1.00**
Standard Deviation							**0.39**	**0.00**

Table 10. Overall average sensitivity and specificity of crowdsourced sanctioning rules in the *Dermatoses of the head, neck and oral cavity* branches.

Number of Diseases	Sensitivity		Specificity	
201	Average	Standard Deviation	Average	Standard Deviation
	0.85	0.34	0.97	0.08

4.3 Results of the Budget-Constrained Bayesian Network Method

In Section 3.3, we described a budget-constrained crowdsourcing method that uses the Bayesian network and reinforcement learning to combine responses from different crowd workers. As described in Section 3.5, we simulated a crowdsourcing experiment by re-analyzing the answers for the same questions with the proposed method. For our evaluation, we selected a data set for which all questions included the *all* option.

Table 11 shows the comparison between the majority-vote and the budget-constrained Bayesian network crowdsourcing methods at different budget levels for questions involving the *all* option. The *Cost* columns show the dollar amount paid to the workers whose answers were used in each method of combining crowd workers' answers. The accuracy rate for the majority-vote method was 88% and the cost was 126 dollars. The budget-constrained Bayesian network method modeled a crowd worker's accuracy rate and stopped using more responses when the system reached a specific threshold of confidence regarding the results of a microtask question. When the budget constraint allowed a relatively high threshold, the budget-constrained Bayesian method could be more accurate because fewer responses (the denominator) were needed to derive conclusions for each question. When the threshold was low (e.g., when the budget was $40), the system was more likely to draw wrong conclusions based on fewer responses, thus decreasing the accuracy rate.

Table 11. Comparison of results for questions involving the *all* option obtained through majority votes and the budget-constrained Bayesian network methods

Budget-Constrained Bayesian Network Method			'Majority Vote' Method		Percent Improvement by Using the Budget-Constrained Method	
Budget (dollars)	Cost (dollars)	Accuracy Rate	Cost	Accuracy	Cost Savings	Accuracy Change
90	79	94%	126	88%	37%	+6%
60	56	92%	126	88%	55%	+4%
40	38	82%	126	88%	70%	-6%

5 Discussion

This paper reports the development of methods to crowdsource sanctioning rules that constrain possible relationships between concepts from two hierarchies. Given concept definitions and related references, crowdsourcing workers are able to provide reasonably accurate answers to binary/multiple-choice biomedical questions in a short amount of time. Reducing or controlling the cost without loss of accuracy is the major challenge when the hierarchies are large, and we developed several methods to deal with that challenge.

First, we utilized the hierarchical structure of both diseases in ICD-11 and the possible anatomical locations to prune the microtasks presented to the crowd workers. One limitation of this kind of hierarchical pruning is that the overall cost and accuracy of the rules heavily depends on the nature of hierarchies. For example, errors in hierarchies can mislead the crowd workers, as we found in our initial exploratory experiments. We had to manually verify the anatomical-location hierarchy and to correct these errors before performing the final experiments. The structure of the hierarchies also affects the cost of crowdsourcing. The anatomical-location hierarchy that we used has 8 levels, and each parent node has 5 child nodes on average. If the hierarchy had higher fan-out, then

we would have to examine more descendants when traversing the hierarchy, which would of course increase the cost.

Second, because the difficulty level of each question is different, the use of Bayesian networks to model confidence in the conclusions can reduce the number of workers needed for each microtask. Modeling each worker's accuracy rate further reduces the number of responses needed for each microtask. However, the Bayesian model does not work well for questions that include *other anatomical location* as an option, when a worker's average accuracy rate for that option was often lower than 50%. A better method for handling this type of question remains for future work.

Finally, because of the huge size of ICD and its post-coordination axes, even with the above two methods, cost is still the major barrier for crowdsourcing sanctioning rules. In response, we developed methods to trade accuracy against cost and to maximize the accuracy of sanctioning rules within a given budget. By varying the confidence cutoff for halting the solicitation of additional crowd workers, we were able to adjust the total cost. Based on this insight, we could apply reinforcement learning to develop an agent that constantly adjusts the confidence cutoffs during the crowdsourcing progress to maximize the overall quality of sanctioning rules under a budget. One limitation of our approach is that the cost-control method can adjust the cost within a certain range, but it cannot guarantee that the method will be able to find sanctioning rules for all diseases within the budget. In the future, we would like to create a mechanism to estimate the total cost of the experiment. If the budget is lower than the lower bound of the range, we should notify the users.

With all of the methods developed in this project to control the cost of obtaining sanctioning rules, compared with the naïve method to test all possible anatomical locations for one disease, we were able to make tremendous savings in the cost of acquiring sanctioning rules. Nevertheless, there are tens of thousands of ICD diseases and dozens of potential post-coordinator axes are being considered for inclusion in ICD-11. The cost of obtaining sanctioning rules may still be high.

A limitation of the budget-constraint method that we implemented is that it trades accuracy against cost instead of trading specificity against cost. Sanctioning rules are used to specify legal combinations of codes and qualifiers to ensure the integrity of coded data. Coders need to be able to use ICD-11 to express correct combinations. Thus, sensitivity is far more important than specificity. Methods that trade specificity against cost may give even better results.

A limitation of our evaluation experiments was that we used a single expert as the gold standard provider. As described by Rodríguez-González [41], a more rigorous evaluation design would use a panel of assessors who are assigned a random subset of questions and a panel of referees who arbitrate among results returned by the assessors and by the system. Our evaluations were formative in nature, designed to demonstrate the feasibility and plausibility of the methods we developed. Fully assessing the accuracy of crowdsourced sanctioning rules for the entire ICD-11 would involve a large-scale study with more rigorously defined gold standards conducted across multiple diseases classes and post-coordination axes.

Finally, we observe that crowdsourcing does have limitations in knowledge-intensive tasks, such as obtaining sanctioning rules. As our experiments demonstrated, crowd workers are capable of answering biomedical questions of the type we posed in this study. However, sometimes the right answer is not obvious. For instance, some crowd workers assume that palmoplantar keratoderma cannot occur on the dorsal side of the hand or foot because its name indicates that it occurs only on the palms and soles; yet others think that palmoplantar keratoderma can sometimes extend to the dorsal side of the hands and feet. Furthermore, the topographical anatomical locations where skin diseases occur are probably better known to an average task-worker not trained in human anatomy than they might be for other characteristics of the diseases. Mortensen et al. [27] found that when concepts, such as Gene Ontology terms, are more esoteric or have less freely available knowledge on the Internet, crowd workers do relatively poorly. As we scale up the task of finding sanctioning rules for all chapters of ICD-11 and for more than two dozen proposed post-coordination axes, such as dimensions of etiology, temporality, severity, consciousness, and external causes, we will encounter differing levels of complexity that may make crowdsourcing more successful in some

areas and less so in others. Diseases of some chapters, such as those of the circulatory or nervous system, may occur in anatomical locations that are obscure to an average layman. Some axes, such as severity and temporal pattern, will have simple enumerated values, although detailed information about possible temporal patterns of diseases may be difficult to glean on the web. Other axes, such as histopathology, will have complex hierarchical values that make crowdsourcing sanctioning rules more difficult. Further research is needed to identify where crowdsourcing may be more successful and where we may hit the limits of scalability and task difficulty for crowdsourcing. In the future, instead of relying solely on human intelligence, we will explore methods to combine automated algorithms with human intelligence to generate sanctioning rules. One possibility is to utilize available knowledge sources such as SNOMED CT, which, for example, specifies the finding sites of many diseases. Another possibility is to employ natural-language processing techniques on selected domains, in the spirit of the work of Coden et al. [21]. We believe that medical professionals still need to verify the sanctioning rules from the crowdsourcing workers and make sure that all sanctioning rules are medically valid.

6 Conclusions

The upcoming ICD-11 will allow for the post-coordination of disease concepts based on a set of qualifiers. It is vital that accurate sanctioning rules are available to avoid incorrect combinations of disease concepts and their qualifier values. In this work, we demonstrated that crowdsourcing could be used to obtain a set of quality sanctioning rules. In particular, we defined crowdsourcing microtasks to obtain ICD-11 sanctioning rules; we used hierarchical structures to improve the efficiency of crowdsourcing; we used Bayesian networks to model our confidence in the acquired sanctioning rules; and we developed a method to maximize the quality of the rules within a fixed budget. Our evaluations showed that, even with a cost saving of 70% over majority-vote crowdsourcing, we were able to achieve an accuracy rate of 82%.

References

1. World Health Organization. The International Classification of Diseases 11th Revision is due by 2018 2016 [cited 2016]. Available from: http://www.who.int/classifications/icd/revision/en/.

2. Nyulas C, Tu SW, Tudorache T, Musen MA, editors. Modeling and tools for supporting post-coordination in ICD-11. ICBO; 2015; Lisbon, Portugal.

3. Rogers JE, Rector A. Development of a methodology and an ontological schema for medical terminology: University of Manchester; 2004.

4. Tao J, Sirin E, Bao J, McGuinness DL, editors. Extending OWL with Integrity Constraints. Proc 23rd Int Workshop on Description Logic; 2010; Waterloo, Canada.

5. Navas H, Lopez Osornio A, Gambarte L, Elias Leguizamon G, Wasserman S, Orrego N, et al. Implementing rules to improve the quality of concept post-coordination with SNOMED CT. Stud Health Technol Inform. 2010;160(Pt 2):1045-9. PubMed PMID: 20841843.

6. Rector AL, Bechhofer S, Goble CA, Horrocks I, Nowlan WA, Solomon WD. The GRAIL concept modelling language for medical terminology. Artif Intell Med. 1997 Feb;9(2):139-71. PubMed PMID: 9040895. Epub 1997/02/01. eng.

7. Organization WH. Content Model 2015 [cited 2016]. Available from: http://www.who.int/classifications/icd/revision/contentmodel/en/.

8. Tudorache T, Falconer S, Nyulas C, Storey MA, Ustun TB, Musen MA. Supporting the Collaborative Authoring of ICD-11 with WebProtege. AMIA Annu Symp Proc. 2010;2010:802-6. PubMed PMID: 21347089. PubMed Central PMCID: 3041458. Epub 2011/02/25. eng.

9. International Health Terminology Standards Development Organisation. SNOMED CT® Technical Implementation Guide: January 2015 International Release 2015 [cited 2016]. Available from: http://www.snomed.org/tig.

10. Cornet R. Do SNOMED CT relationships qualify? Stud Health Technol Inform. 2008;136:785-90. PubMed PMID: 18487827.

11. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. J Am Med Inform Assoc. 2011 Sep-Oct;18(5):544-51. PubMed PMID: 21846786. PubMed Central PMCID: 3168328.

12. Chapman WW, Cohen KB. Current issues in biomedical text mining and natural language processing. J Biomed Inform. 2009 Oct;42(5):757-9. PubMed PMID: 19735740. Epub 2009/09/09. eng.

13. Hahn U, Cohen KB, Garten Y, Shah NH. Mining the pharmacogenomics literature—a survey of the state of the art. Briefings in Bioinformatics. 2012 07/24

11/18/received

03/23/accepted;13(4):460-94. PubMed PMID: PMC3404399.

14. Rebholz-Schuhmann D, Oellrich A, Hoehndorf R. Text-mining solutions for biomedical research: enabling integrative biology. Nat Rev Genet. 2012 12//print;13(12):829-39.

15. Hearst MA. Automatic Acquisition of Hyponyms from Large Text Corpora. Proc 14th Conf Comput Ling1992. p. 539–45.

16. Bundschus M, Dejori M, Stetter M, Tresp V, Kriegel HP. Extraction of semantic biomedical relations from text using conditional random fields. BMC Bioinformatics. 2008 Apr 23;9:207. PubMed PMID: 18433469. PubMed Central PMCID: 2386138.

17. Zeng Z, Shi H, Wu Y, Hong Z. Survey of Natural Language Processing Techniques in Bioinformatics. Computational and mathematical methods in medicine. 2015;2015:674296. PubMed PMID: 26525745. PubMed Central PMCID: 4615216.

18. Bui QC, Nuallain BO, Boucher CA, Sloot PM. Extracting causal relations on HIV drug resistance from literature. BMC Bioinformatics. 2010 Feb 23;11:101. PubMed PMID: 20178611. PubMed Central PMCID: 2841207.

19. Wang X, Hripcsak G, Friedman C. Characterizing environmental and phenotypic associations using information theory and electronic health records. BMC Bioinformatics. 2009;10 Suppl 9:S13. PubMed PMID: 19761567. PubMed Central PMCID: 2745684. Epub 2009/09/26. eng.

20. Liu S, Tang B, Chen Q, Wang X. Drug-Drug Interaction Extraction via Convolutional Neural Networks. Computational and mathematical methods in medicine. 2016;2016:6918381. PubMed PMID: 26941831. PubMed Central PMCID: 4752975.

21. Coden A, Savova G, Sominsky I, Tanenblatt M, Masanz J, Schuler K, et al. Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model. J Biomed Inform. 2009 Oct;42(5):937-49. PubMed PMID: 19135551.

22. Leroy G, Chen H, Martinez JD. A shallow parser based on closed-class words to capture relations in biomedical text. J Biomed Inform. 2003 Jun;36(3):145-58. PubMed PMID: 14615225.

23. Doan A, Ramakrishnan R, Halevy AY. Crowdsourcing Systems on the World-Wide Web. Communications of the ACM. 2011;54(4):86-96.

24. von Ahn L, Dabbish L, editors. Labeling images with a computer game. CHI; 2004 2004.

25. Prive T. What Is Crowdfunding And How Does It Benefit The Economy. Forbes. 2012/11/27.

26. Mortensen JM, Musen MA, Noy NF. Crowdsourcing the verification of relationships in biomedical ontologies. AMIA Annu Symp Proc. 2013;2013:1020-9. PubMed PMID: 24551391. PubMed Central PMCID: 3900126. Epub 2014/02/20.

27. Mortensen JM, Telis N, Hughey JJ, Fan-Minogue H, Van Auken K, Dumontier M, et al. Is the crowd better as an assistant or a replacement in ontology engineering? An exploration through the lens of the Gene Ontology. J Biomed Inform. 2016 Apr;60:199-209. PubMed PMID: 26873781. PubMed Central PMCID: 4836980.

28. Sarasua C, Simperl E, Noy N, editors. CrowdMap: Crowdsourcing Ontology Alignment with Microtasks. The 11th International Conference on The Semantic Web; 2012: Springer-Verlag.

29. Mortensen JM. Crowdsourcing Ontology Verification [Ph.D. Dissertation]: Stanford University; 2015.

30. Whitehall J, Ruvolo P, Wu T, Bergsma J, Mavellan J. Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise. Advances in Neural Information Processing Systems. 2009;22.

31. Bachrach Y, Grepel T, Minka T, Guiver J. How To Grade a Test Without Knowing the Answers | A Bayesian Graphical Model for Adaptive Crowdsourcing and Aptitude Testing. Proceedings of the 29th International Conference on Machine Learning (ICML-12)2012. p. 1183-90.

32. Kittur A, Chi EH, Suh B. Crowdsourcing user studies with Mechanical Turk. CHI '08 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems2008. p. 453-6

33. Mortensen JM, Alexander PR, Musen MA, Noy NF. Crowdsourcing ontology verification. ICBO2013.

34. Tudorache T, Falconer, S.M., Nyulas, C.I., Noy, N.F., Musen, M.A. Will Semantic Web Technologies Work for the Development of ICD-11? The 9th International Semantic Web Conference, ISWC 2010; Shanghai, China2010.

35. Schulz S, Rodrigues JM, Rector A, Spackman K, Campbell J, Ustun B, et al. What's in a class? Lessons learnt from the ICD - SNOMED CT harmonisation. Stud Health Technol Inform. 2014;205:1038-42. PubMed PMID: 25160346. Proceedings of MIE2014.

36. Amazon. Amazon Mechanical Turk 2014 [cited 2014]. Available from: http://aws.amazon.com/mturk/.

37. Wang J, Ghose A, Ipeirotis PG. Bonus, Disclosure, and Choice: What Motivates the Creation of High-Quality Paid Reviews? Proceedings of the Thirty-Third International Conference on Information Systems (ICIS 2012); Orlando, FL2012.

38. Koller D, Friedman N. Probabilistic Graphical Models: Principles and Techniques. Cambridge, Mass: The MIT Press; 2009.

39. Frey BJ. Extending factor graphs so as to unify directed and undirected graphical models. Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence; Acapulco, Mexico. 2100615: Morgan Kaufmann Publishers Inc.; 2003. p. 257-64.

40. Sutton RS, Barto AG. Reinforcement learning: An Introduction: The MIT Press; 1998.

41. Rodríguez-González A, Torres-Niño J, Valencia-Garcia R, Mayer MA, Alor-Hernandez G. Using experts feedback in clinical case resolution and arbitration as accuracy diagnosis methodology. Computers in Biology and Medicine. 2013 9/1/;43(8):975-86.

42. Tu SW. ICD-11 Sanctioning Rule Crowdsourcing Results 2016 [cited 2016]. Available from: http://www.stanford.edu/~swt/ICD11SanctioningRuleResultsMajorityVote.xls.

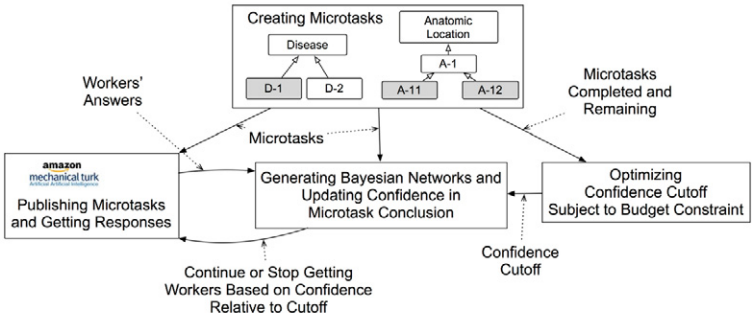
Acknowledgments

This work was supported by Grant GM086587 from the U.S. National Institute of General Medical Sciences (NIGMS) of the National Institutes of Health. The Protégé project is supported by NIGMS grant GM103316. We thank Katharine Miller for copyediting the manuscript.

Conflict of Interest Statement

Conflicts of interest: none.

Graphical abstract



Highlights

- We defined crowdsourcing microtasks to obtain ICD-11 sanctioning rules.
- We used hierarchical structures to improve the efficiency of crowdsourcing.
- We used Bayesian networks to model our confidence in the acquired sanctioning rules.
- We developed a method to maximize the quality of the rules within a fixed budget.