Combining Gamification Techniques and Crowdsourcing to Create a Gold Standard for Medical Text

Master’s Thesis

Anca Dumitrache
Combining Gamification Techniques and Crowdsourcing to Create a Gold Standard for Medical Text

THESIS

submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

ARTIFICIAL INTELLIGENCE

specializing in

KNOWLEDGE TECHNOLOGY AND INTELLIGENT INTERNET APPLICATIONS

by

Anca Dumitrache
born in Bistriţa, Romania

Department of Artificial Intelligence
Faculty FEW, VU University Amsterdam
Amsterdam, Netherlands
http://www.few.vu.nl/

Center for Advanced Studies
IBM Netherlands
Amsterdam, Netherlands
http://www.ibm.nl
Combining Gamification Techniques and Crowdsourcing to Create a Gold Standard for Medical Text

Author: Anca Dumitrache
Student ID: 2205742
Email: anca.dumitrache@student.vu.nl

Abstract

This document proposes a design for a gamified crowdsourcing workflow to extract annotation from medical text. Developed in the context of a general crowdsourcing platform, Dr. Detective is a game with a purpose that engages medical experts into solving annotation tasks on medical case reports, tailored to capture disagreement between annotators. It incorporates incentives such as learning features, to motivate a continuous involvement of the expert crowd. The game was designed to identify expressions valuable for training NLP tools, and interpret their relation in the context of medical diagnosing. In this way, we can resolve the main problem in gathering ground truth from experts – that the low inter-annotator agreement is typically caused by different interpretations of the text. We report on the results of the pilot run of Dr. Detective. We compared the collected crowd data with the results of an NLP parser, performed an analysis of the ambiguity in the answers, and how the gaming elements influenced it, and surveyed the participants for their opinion on the application.

Thesis Committee:

University supervisor: Dr. Lora Aroyo, Faculty FEW, VU University Amsterdam
Company supervisor: Dr. Chris Welty, IBM Watson Research Center, New York
Company supervisor: Robert-Jan Sips, CAS Benelux, IBM Netherlands
I would like to thank my supervisors Dr. Lora Aroyo, Dr. Chris Welty, and Robert-Jan Sips, for their assistance and guidance with the development of this thesis. Furthermore, I would like to thank my colleagues on the Crowd-Watson team, Oana Inel, Guillermo Soberón, Hui Lin, and Manfred Overmeen, who provided their support during this project. I would also like to thank Anthony Levas from the IBM Watson Research Center for his advice and feedback on the Dr. Detective application, as well as his help with setting up the application pilot.

I would like to thank Kathrin Dentler and Dr. Petra Wolffs for their help with finding participants for both the interviews and the application pilot run, as well as the students and medical professionals who were involved in these activities, and who provided their feedback.

Finally, I would like to thank Dr. Miguel Sicart, and Dr. Georgios N. Yannakakis for their advice on gamification and playful interaction topics.

Anca Dumitrache
Amsterdam, Netherlands
August 29, 2013
Contents

Acknowledgments iii

Contents v

List of Figures vii

List of Tables ix

1 Introduction 1
  1.1 Research Questions 2
  1.2 Glossary and Definitions 3

2 Related Work 5
  2.1 Motivating the Crowd 5
  2.2 Design Principles for Crowdsourcing 6
  2.3 Related Applications 7

3 Crowd Incentives 9
  3.1 Interview Setup 9
  3.2 Interview Analysis 10

4 Crowdsourcing Architecture 15
  4.1 Pre-Processing for the Game Platform 16
  4.2 Data Collection: Gaming Platform 17
  4.3 Game Disagreement Analytics 20

5 “Dr. Detective” Experiment Pilot 27
  5.1 Experimental Setup 27
  5.2 Results and Discussion 30

6 Future Work 39
  6.1 Game Design and User Interaction 39
  6.2 Capturing Disagreement 40
  6.3 New Tasks for the Crowd 41
List of Figures

3.1 Overview of the interview results ........................................ 14
4.1 Crowd-Watson Framework Design ..................................... 15
4.2 Game flow ................................................................. 19
5.1 “Dr. Detective” full version ............................................. 28
5.2 Term types in the game .................................................. 28
5.3 “Dr. Detective” simple version ........................................ 29
5.4 Crowd words in UMLS .................................................. 30
5.5 Crowd terms in UMLS .................................................... 30
5.6 Crowd words after each game round ................................ 31
5.7 Total words distributed by game version ............................ 32
5.8 New words distributed by game version ............................ 32
5.9 Term clusters example .................................................... 34
5.10 Cluster clarity for medical observation types ...................... 35
5.11 Cluster clarity for other types ........................................ 35
5.12 Worker-paragraph similarity scores for term clusters .......... 36
5.13 Worker-paragraph similarity scores for term types ............ 36
D.1 Answers to Q1 in the game questionnaire .......................... 107
D.2 Answers to Q2 in the game questionnaire .......................... 108
D.3 Answers to Q3 in the game questionnaire .......................... 109
D.4 Answers to Q4 in the game questionnaire .......................... 109
D.5 Answers to Q5 in the game questionnaire .......................... 110
D.6 Answers to Q6 in the game questionnaire .......................... 111
D.7 Answers to Q7 in the game questionnaire (1) ....................... 111
D.8 Answers to Q7 in the game questionnaire (2) ....................... 112
D.9 Answers to Q8 in the game questionnaire .......................... 112
D.10 Answers to Q9 in the game questionnaire .......................... 113
D.11 Answers to Q10 in the game questionnaire ......................... 113
List of Tables

3.1 Overview of the observations collected from the interviews. . . . . . . . . 14

C.1 Demographics for lecturers (1). . . . . . . . . . . . . . . . . . . . . . . . 101
C.2 Demographics for lecturers (2). . . . . . . . . . . . . . . . . . . . . . . . 101
C.3 Demographics for medical professionals (1). . . . . . . . . . . . . . . . . 101
C.4 Demographics for medical professionals (2). . . . . . . . . . . . . . . . . 101
C.5 Demographics for students (1). . . . . . . . . . . . . . . . . . . . . . . . 102
C.6 Demographics for students (2). . . . . . . . . . . . . . . . . . . . . . . . 102
C.7 Overview of reading content. . . . . . . . . . . . . . . . . . . . . . . . . 103
C.8 Overview of reading habits. . . . . . . . . . . . . . . . . . . . . . . . . . 103
C.9 Studying and learning habits overview (1). . . . . . . . . . . . . . . . . . 104
C.10 Studying and learning habits overview (2). . . . . . . . . . . . . . . . . . 104
C.11 Gaming habits overview . . . . . . . . . . . . . . . . . . . . . . . . . . 104
C.12 Experience with medical competitions overview . . . . . . . . . . . . . . 105
C.13 Interest in a crowdsourced application (1). . . . . . . . . . . . . . . . . . 105
C.14 Interest in a crowdsourced application (2). . . . . . . . . . . . . . . . . . 106
Modern cognitive systems require human annotated data for training and evaluation, especially when adapting to a new domain, and during the first training and setup of such a system. An example of such system is Watson QA [8] developed by IBM, that won the Jeopardy TV quiz show against human competitors. To tune its performance, Watson was trained on a series of databases, taxonomies, and ontologies of publicly available data [12]. Currently, IBM Research aims at adapting the Watson technology for question-answering in the medical domain, which requires large amounts of new training and evaluation data in the form of human-annotated medical text.

Two issues arise in this context: (1) the traditional way of ground-truth annotations is slow, expensive and generates only small amounts of data, and (2) in order to achieve high inter-annotator agreement, the annotation guidelines are highly restrictive. Such practice has proven to create over-generalization and brittleness [1], by losing the sense of diversity in the language, which leads to the fact that natural language processing tools have problems in understanding the ambiguity of expressions in text, which constitutes a critical in the way medical text is processed.

The diversity of interpretation of medical text can be seen at many levels; as a simple example, consider the sentence, “Patients exhibiting acute tailbone pain should be examined for extra bone nodules.” Human experts disagree routinely on whether “acute tailbone pain”, “tailbone pain”, or “pain” is the primary term in this sentence. Proponents of “tailbone pain” argue that there is a medical term for it (Coccydynia) making it primary, others argue that it is pain which is located in the tailbone. Traditional methods of gathering ground truth data for training and evaluation fail to capture such interpretation diversity, leading us to the innovative Crowd Truth approach [2] providing context for this work.

Our analysis led us to believe that the diversity of interpretation occurs at various levels, depending on whether the context is being considered. Term identification, as exemplified in the example above, may be done independently of the clinical context, for example when processing a textbook for background knowledge. However, in the presence of a particular patient, the role of the location and duration modifiers (e.g. tailbone, acute, resp) may or may not be important. We also observe that context-independent tasks tend to require less expertise, allowing us to use a lay crowd more effectively.
1.1 Research Questions

We hypothesize that these different types of annotation tasks can be performed by two different types of crowds in order to optimize the time, effort and the quality of the final result. Given the experience 1,2 with defining micro-tasks for the general crowd via crowdsourcing platforms such as Amazon Mechanical Turk1 or CrowdFlower2, we focus on a method to engage a crowd of medical experts to be able to resolve ambiguity present at the semantic level in medical text. Annotating complex medical text could be a time consuming and mentally taxing endeavor, therefore the monetary incentive might not be sufficient for attracting a crowd of experts. However, providing a tailored experience for medical professionals through features such as e-learning, and competition with peers, could serve as additional motivation for assembling the right crowd for our task. This can be accomplished by incorporating gamification features into our application.

In Chapter 2, we position our work in the context of already existing games with a purpose, crowdsourcing and other niche-sourcing initiatives. Next, in Chapter 3 we analyze possible incentives for the expert crowd, through conducting a series of qualitative interviews with people in the medical field, and analyzing their answers. Then in Chapter 4 we outline our approach by focusing on the gaming elements used as incentives for medical experts, in the context of the overall game application architecture. We show how this gaming platform could fit together with a micro-task platform in a joint workflow combining efforts of both expert and non-expert crowds. Next, in Chapter 5 we describe the experimental setup to explore the feasibility and the usability of such an application. Here we also discuss the results of the pilot run of our application. In Chapter 6 we identify the points of improvement to bring in future versions. Finally, in Chapter 7 we provide the conclusions of our study, as well as a summary of how our research questions have been answered.

1.1 Research Questions

Our theories for crowdsourcing annotations in medical text by using an expert crowd, and capturing the diversity of opinion to build a gold standard can be summarized as the following research questions:

1. How will the expert crowd compare to an automated NLP approach at extracting a gold standard for medical text?

2. How is diversity of language present in crowd annotations?
   a) Does having access to the answers of other users stimulate diversity of opinion?
   b) Is diversity of opinion an inherent trait in medical text, or does it indicate low-quality data (i.e. as a result of a poorly defined task, or low-quality workers)?

1 www.mturk.com
2 www.crowdflower.com
3. Can a gamified crowdsourcing platform be employed to capture annotations of an expert crowd?

This thesis attempts to answer these research questions by proposing a gamified crowdsourcing application for engaging experts in a knowledge acquisition process that involves domain-specific knowledge extraction in medical texts. The goal of such text annotations is to generate a gold standard for training and evaluation of IBM Watson NLP components in the medical domain.

1.2 Glossary and Definitions

- **answer set**: the set of all annotations collected during a game round.
- **crowd incentive**: motivating factor for the crowd to engage in a crowdsourcing activity.
- **crowdsourcing**: the practice of obtaining content by soliciting contributions from a (virtual) community.
- **full term match**: a property of two terms indicating that they are comprised of identical word sets.
- **game round**: a unit in the annotation game pilot, comprised of one user collecting all the annotations for one paragraph.
- **gold standard**: in NLP evaluation, the set of annotations that is considered definitive.
- **NLP**: natural language processing.
- **paragraph saturation point**: the event of having collected all possible annotations for an input paragraph.
- **partial term match**, or **term overlap**: a property of two terms indicating that they are comprised of word sets with at least one word in common.
- **relation**: an annotation object consisting of two terms that are tied together by a medical property.
- **spammer**: a worker that is intentionally providing incorrect answers, for the purpose of cheating the crowdsourcing system into giving rewards.
- **term**: an annotation object consisting of a set of words that together form a medical concept.
- **typing**, or **categorization**: assigning a category from a defined set to an annotation object.
- **worker**: an individual in the crowd involved in a crowdsourcing activity.
Chapter 2

Related Work

In this chapter, we discuss the related work that has been done on the topics relating to our research questions, as defined in Chapter [1]. First we analyze methods for motivating the expert crowd. Then we describe design principles for crowdsourcing in current literature. Finally we discuss related applications.

2.1 Motivating the Crowd

One of the main challenges of our project is motivating the expert medical crowd to complete the annotation tasks. Tokarchuk et al. [21] define a set of crowd incentives that could be employed for such a purpose:

1. *Reciprocity and expectancy* – This motivates the crowd to participate because the participants expect that others will do the same for them when they need it.

2. *Competition* – This motivates participants to contribute to increase their prestige and be well regarded by others. Participants also contribute to show that they do more than others can do.

3. *Altruism* – This motivates the participants to act out of pure sympathy for others, with no expectation of reciprocity.

4. *Learning* – This could motivate the crowd by integrating the crowdsourced application with the mandatory course work and exam studying that medical students need to perform.

5. *Entertainment* – This incentive proposes integrating gaming elements with the crowdsourced application, to make it more attractive to participants.

6. *Money or the implicit promise of future monetary rewards* – This is the primary incentive for general crowdsourcing endeavors, motivating the crowd through direct compensation.

As the aim of this project was to design an alternative environment for crowd labor, the *Money or the implicit promise of future monetary rewards* motivator is not useful
to us. Additionally, as the purpose of this project is to implement an initial prototype of a crowdsourced platform, the community building aspect could not be properly tackled, due to the time required to form such an environment. For this reason, the two community-related incentives, *Reciprocity and expectancy* and *Altruism*, are not as relevant to our project. However, the *Competition*, *Learning*, and *Entertainment* incentives are all applicable to our use case. While Tokarchuk et al. designed these incentives for a general crowd, we aim our project to involve workers with a medical background, therefore these incentives would need to be adapted for the particular needs of the expert crowd.

Corneli et al. [6] also stress the importance of the *Learning* incentive, by studying how crowdsourcing can be employed in E-Learning contexts. By analyzing scientific online learning communities such as PlanetMath[1], where users contribute answers to the questions of their peers, the authors show the role-reversal of peers taking the place of educators in collaborative learning. The contributions of popular peers are widely respected in these communities, and users compete to reach the top standings in the hierarchy. This would seem to indicate that a learning platform could be a useful tool for the medical crowd.

### 2.2 Design Principles for Crowdsourcing

Aside from attracting the crowd to engage with the application, the crowdsourcing task and user interaction should be designed in such a way that the relevant information is extracted from the users, and the users continue using the application. Kittur et al. [13] discuss some of the challenges of building more complex crowdsourcing applications. They argue that targeting user skills is essential, as the users cannot be expected to perform the same menial tasks for a long period of time. This also contributes to getting higher quality output data, as the skills of the workers are better employed. Moreover, the workers need to be given some sense of recognition for high quality work, in order to keep them engaging with the application.

Siorpaes et al. [19] also discuss design principles for crowdsourcing, by introducing a framework of 10 design points for Semantic Web populating games. In the context of our research, of a particular interest are: identifying tasks in semantic-content creation, designing game scenarios, designing an attractive interface, identifying reusable bodies of knowledge, and avoiding typical pitfalls. As not all crowdsourcing tasks are suitable for redesign as part of a gamified platform, identifying which of these tasks could engage successfully medical expert crowd is of a key importance to our research. It is also crucial to involve mechanisms to optimize the ratio of time spent and quality and volume of the output [19]. External knowledge sources for annotations (e.g. vocabularies, NLP parsers) can be used to target the work of the players to problems that are too complex to be handled only by computers [19]. Finally, in order to ensure the quality of the answers, unintentional mistakes of the users need to be avoided through clear instructions in the interface [19]. Based on these tasks, a suitable game scenario needs to be created, one that is able to successfully combine the goals of the crowdsourcing work (retrieving text annotation) with the goals of the

crowd (e.g. learning, entertainment). This scenario needs then to be packaged into an attractive and easy to use gaming interface, that would enable the users to seamlessly solve their tasks.

2.3 Related Applications

In recent years, crowdsourcing has gained a significant amount of exposure as a way for creating solutions for computationally complex problems. By carefully targeting workers with gaming elements and incentives, various crowdsourcing applications were able to garner a significant user base engaged in their tasks. The ESP Game [22] (later renamed Google Image Labeler) pioneered the field by implementing a gamified crowdsourcing approach to generate metadata for images. The reCAPTCHA [24] application combined the CAPTCHA security measure for testing human knowledge with crowdsourcing, in order to perform text extraction from images. The gamified crowdsourcing approach has been employed successfully even in scientific research, with applications such as Galaxy Zoo [15] using crowd knowledge to perform image analysis and extract observations from pictures of galaxies. All of these systems employ mechanisms for a continuous collection of a large amount of human annotated data.

Gamification as applied to text annotation crowdsourcing is an emerging field in different domains. For instance, the Phrase Detective project [5] uses gamified crowdsourcing for building anaphoric annotation ground truth. The input documents are general purpose, and the crowd is not specialized. Two interesting features we considered for Dr. Detective as well, (1) the need for a user training task to improve the usage of the application, and (2) understanding of the user profile (e.g. players can examine a considerable variation in their interaction styles, abilities or background knowledge.

The Sentiment Quiz [17], played through various social networking platforms, employs crowdsourcing to evaluate accuracy of sentiment detecting algorithms over sentences, and to create a lexicon of sentiments in various languages. The requirements for user incentives in Dr. Detective were based on the analysis provided by Sentiment Quiz, e.g. for scoring, high score board, and level-based goals, as well as for enhancing the crowd output through statistical methods applied in the disagreement analytics.

However, neither the Sentiment Quiz, nor the Phrase Detective applications actively seek out to capture the ambiguity in language. Phrase Detective even tries to enforce agreement, by awarding additional points for annotators that agree with the ground truth. Neither do most applications in the domain study the effect of using specialized crowds to perform the information extraction tasks. Our goal is to build an end-to-end gamified crowdsourcing platform that can capture disagreement between annotators, while catering specifically to experts in the medical field.
Chapter 3

Crowd Incentives

In order to collect data from a crowd of medical experts, it is imperative to find the necessary motivators for engaging them into contributing. The literature review in Chapter 2.1 discussed some general crowd motivators. However, in order to engage a specific crowd of medical experts, these incentives need to be discussed from the perspective of these potential users. To this end, we have performed a series of qualitative interviews with medical students and professionals. The purpose was to identify what requirements and features would the medical crowd be interested in seeing in a crowdsourced application, and how this application could be built to help in their work.

3.1 Interview Setup

In order to gauge the interest of people studying and/or working in the medical domain, with regards to participating in a crowdsourced application, a series of interviews was performed. The interviews performed were qualitative, in the sense that interviewees were encouraged to express their ideas on how a crowdsourced application for the medical domain could be implemented. This section describes the setup of these interviews.

For the purpose of this analysis, eleven people were interviewed, each belonging to one of these categories: medical professional, medical student, and lecturer in a medical-related field. These participants were selected as a representative sample of how a possible medical crowd would look like. The questions in the interview questionnaire were meant to capture the reading habits of a medical crowd, as well as their possible interest in taking part in a crowdsourced application. The purpose of the questions on reading habits was to form ideas on how to integrate a crowdsourced application dealing with text annotation, with the the mandatory reading that people in the medical field already need to perform. Similarly, the purpose of the questions on interest in a crowdsourcing application was to collect concrete ideas on how crowdsourced tasks could be integrated with the work that people in the medical field already have to do, from their perspective.

In addition to these questions, we asked the interviewees about specific motivations that could engage them with a crowdsourced application. Through the literature
on motivating factors for crowdsourcing in Chapter 2.1, we identified the Competition, Learning, and Entertainment incentives as being relevant to our project. The interview questionnaire incorporated sections for each of these motivators. In particular, we were interested in how Learning could motivate the crowd by integrating the crowdsourced application with the mandatory course work and exam studying that medical students need to perform. From the side of the lecturers, we designed similar questions to find out how they structure their courses and exams, to find out whether a crowdsourced application could be integrated with their teaching. We also wanted to find out what part does Entertainment, specifically gaming, play in the daily schedule of people in the medical field, and whether a crowdsourced game could be integrated with their gaming habits. Finally, for the Competition incentive, we developed questions to find out about medical competitions, and whether the medical crowd is interested in participating in them.

The interview questionnaire was therefore built following six sections:

1. General demographic information;
2. Reading habits;
3. Learning/teaching in a medical program;
4. Medical knowledge competitions;
5. Gaming;
6. Participating in crowdsourcing activities.

The questionnaires used for each interviewee category can be found in Appendix A.

### 3.2 Interview Analysis

This section describes the result of the interviews from the previous section. The full transcripts of the discussions can be found in the Appendix B. An overview of the answers in tabular format is available in Appendix C. We are interested in getting information based on three categories: the input, the task/mechanics, and the goal.

The main remarks from the interviews are listed below as Observations (abbreviated as Obs.).

The demographic data on each of the interviewees is presented in the following tables: medical professional (tables C.3 and C.4), medical student (tables C.5 and C.6), and lecturer in a medical-related field (tables C.1 and C.2). While most of the interviewees had a background in general medicine, they also displayed a wide array of specializations and related interests in the medical field, which enabled for a wide array of opinions. The interviewed lecturers (three subjects) teach classes aimed at Bachelor’s and Master’s students, in fields adjacent to the medical domain (microbiology, bioinformatics). The interviewed professionals (two subjects) work primarily in general medicine, but also have some experience with non-clinical fields (metabolic diseases, microbiology). Finally, the interviewed students (five subjects) feature at
Crowd Incentives 3.2 Interview Analysis

varying stages during their educational formation (from BSc to PhD), and are involved in both traditional lectures, and research work.

The results of aggregating the answers on reading habits of the medical crowd are documented in tables C.7 and C.8. Overall, the text types that each person is required to be familiar with were consistent across the interviewee categories, with textbooks and publications were referenced by all three. However, there is a clear progression in the type of the contents, that is correlated with the stage the interviewees are in their career - from the answers it appears people in the medical field move from reading textbooks to reading publications, as they become more advanced in their line of work (from interview B.2.1 – “[during the studies] it was not so much scientific publications. Now [during my work] it’s really different, I only read scientific publications.”). At the same time, the content also becomes more specialized according to each specialization, with medical professionals and lecturers appear to rarely read the same publications, as evidenced by the diverse sources cited. This is evidenced by the fact that textbooks are also more commonly used as exam preparation, whereas publications tend to be used to keep track of related work in a field. Based on this trend, an idea for framing the crowdsourced application with the mandatory reading can be formulated:

(Obs.1) Use medical case reports as the primary source for the tasks.

The motivation behind this idea is that, by using case reports as the source of the reading material for the tasks, more advanced students would be incentivized to participate in the crowdsourced activity (as they need to perform the readings anyway). In addition, medical professionals and researcher also report the need of reading case reports.

Medical professionals also report the need to review the comparatively simpler information that medical students need to read, for the purpose of getting quick references to a new field (from interview B.3.2 – “Sometimes you have some projects which are more towards a specific condition, and then you are forced to learn a bit more.”), or preparation from going into ward work (from interview B.2.1 – “I don’t read a lot of medical books anymore, but if I do, it’s because in a while I have to start working in the ward again, so I have to prepare myself.”). The difference seems to be that, due to time constraints, textbooks are impractical to read, therefore professionals look for information in a more concentrated manner (e.g. Wikipedia articles, case books). A different scenario for framing the crowdsourced application can thus be constructed:

(Obs.2) Create an application for getting quick references in a medical field, using Wikipedia as the input text.

This style of application would be useful to medical professionals/lecturers for quickly getting up to date with the information from within a field. At the same time, medical students could also find it useful for reviewing the essential information needed for their classes, as textbooks can often prove overwhelming in terms of
3.2 Interview Analysis

content (from interview B.2.1—“[During my studies] we had to read a lot of textbooks, but many students, including me, skipped them, because most textbooks are very tough to read...”).

Finally, as reflected from the reading habits of the interviewees, most people prefer highlighting parts of the text, as opposed to actually taking notes of what they are reading. This reading technique could be adapted as part of the crowdsourced application:

(Obs.3) Create the tasks such that users have to highlight text, rather than re-write it.

The results of aggregating the answers on learning or teaching medical courses are presented in tables C.9 and C.10. The main exams identified by the interviewees were per-course exams, which usually occur at the end of a lecture block of six to eight weeks, as well as the “Progress” exam for medical students, which occurs two times a year. While the “Progress” exam always consists of multiple choice questions, the content of the regular exams appears to be highly dependent on the type of the course, as well as the educational institute, with not much overlap across fields. In addition, many students indicated they do not prepare for “Progress” exams, as these are intended to measure the overall progression in the field, and not a particular skill (from interview B.1.3—“for the progress exams, there I don’t study, because it’s just to measure your progress, and it should be on the current knowledge you have”). For this reason, tailoring the crowdsourced application to fit a specific exam preparation routine might not be an appropriate solution.

However, several suggestions on how interviewees study have proved to be quite useful. A number of students referred to the “Problem-Based Learning” approach which was employed in their studies (from interview B.1.2—“you have a problem that you discuss in a small group, and everyone contributes what they know about it, then we try to combine all of that into one solution”). In addition, students that have not had experience with this style of teaching have expressed their wish for participating in problem-solving tasks (from interview B.1.1—“We are lacking problem solving from our classes, we have to do a lot of reading, but we do not apply the knowledge.”). Based on these observations, another context to frame the crowdsourced application can be defined:

(Obs.4) Construct application tasks to target problem solving skills.

One of the interviewees indicated a related application, which employs problem solving, as applied to the task of diagnosis making (from interview B.1.4—“You also have the resident bulletins from these journals – you read a thing and they have questions. I think it’s in the New England Journal of Medicine”). The New England Journal of Medicine functions similarly to the case books described in the previous section, in the sense that they describe patient cases, with the added feature of interacting with the

http://www.nejm.org/image-challenge#02162006
information, through the use of dialog boxes, and multiple choice questions. A more in-depth analysis of this application could provide more insight into how to design a successful problem-solving task for the medical domain.

The results of aggregating the answers on gaming incentive are available in table C.11. While most interviewees have reported they do not have much time for playing, most of them engage in playing mobile games, especially while commuting. In addition to this, many interviewees expressly stated that having the crowdsourced application in a mobile format would increase the likelihood they engage with it (from interview B.1.4 – “I would strongly encourage you [to make a mobile app]”, from interview B.1.3 – “[A mobile app] would be very handy, because it does not matter where you are, you can use it.”). Therefore, a proposed task design for the application could be:

(Obs.5) **Develop a flexible crowdsourced application, with short tasks that can be played during free moments.**

The results of aggregating the answers on involvement in medical knowledge competitions can be found in table C.12. When it comes to real life competitions, most had heard of them, but few had participated. Furthermore, some participants were reluctant to participate in competitions that involved facing their opponents in real life (“[I would play] only if it were anonymous.”, from interview B.2.1). However, when looking at the gaming habits of the interviewees, most state they enjoy the challenge, some even going so far to state that they would only compete against their peers if it were part of an online game (from interview B.2.1 – “I don’t like medical competitions, but for gaming, I do like the competitive aspect.”). Based on these findings, we can make the following observation:

(Obs.6) **The application should foster competition between participants.**

The results of aggregating the answers on interest in a crowdsourced application are available in tables C.13 and C.14. The interviewees also enforced their preference for time flexibility of the tasks to solve, in keeping with [Obs.5] as most interviewees prefer to engage for up to two hours per week, in their free moments. A popular domain and task combination suggested was creating a medical diagnosing game, where users would have to look for clues inside textual descriptions of patient cases (from interview B.1.2 – “Certain colleagues of mine would consider House [the TV show] studying, because you had a patient with certain symptoms, and you had to guess the disease.”, from interview B.1.3 – “For medical students, the case problem solving would be the most interesting, because then you have some practical link to the real world.”). This ties into the problem-solving task referenced in [Obs.4].

An overview of the observations collected from the interview, and the way they relate to the application components is available in table 3.2. Furthermore, taking into account how the application components interact with one another, the interview results can be represented graphically as seen in Figure 3.1.
### 3.2 Interview Analysis

<table>
<thead>
<tr>
<th>Application Component</th>
<th>Observation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text input</td>
<td>(Obs.1)</td>
<td>medical case reports</td>
</tr>
<tr>
<td></td>
<td>(Obs.2)</td>
<td>Wikipedia articles</td>
</tr>
<tr>
<td>Task/</td>
<td>(Obs.3)</td>
<td>highlighting annotations</td>
</tr>
<tr>
<td>mechanics</td>
<td>(Obs.5)</td>
<td>flexibility in time for solving tasks</td>
</tr>
<tr>
<td></td>
<td>(Obs.6)</td>
<td>competition with peers</td>
</tr>
<tr>
<td>Goal</td>
<td>(Obs.2)</td>
<td>medical reference guide</td>
</tr>
<tr>
<td></td>
<td>(Obs.4)</td>
<td>medical problem solving</td>
</tr>
</tbody>
</table>

Table 3.1: Overview of the observations collected from the interviews.

Based on these observations, we can formulate two scenarios for our application: a **medical reference guide** for looking up various medical concepts, and a **medical problem solving** application. When factoring in the crowd incentives discussed in the previous section, specifically the competition and entertainment aspect, the reference guide approach appears to be lacking in interactivity. Similarly, by using Wikipedia articles as input, we lose some of the potential of the expert crowd and their ability to understand complex medical text, as the intended audience for these articles is a general public. Therefore, based on the feedback from the interviews, we set on implementing a medical problem solving game, using medical case reports as input.
Chapter 4

Crowdsourcing Architecture

In this chapter, we describe the architecture for *Dr. Detective* – an application for engaging experts in knowledge extraction tasks for creating ground truth annotations in medical texts. We start by framing *Dr. Detective* as part of the general Crowd-Watson framework for crowdsourcing medical text annotation. Then, we tackle the challenge of tailoring the application to a specialized crowd of medical professionals, through a study of possible motivating factors. Finally, we describe how gamification elements were integrated with the crowdsourcing workflow.

The Crowd-Watson framework supports the composition of crowd-truth gathering workflows, where a sequence of micro-annotation-tasks can be executed jointly either by the general crowd on platforms like CrowdFlower, or by specialized crowd of do-

![Crowd-Watson Framework Design](http://crowd-watson.nl/dr-detective-game)

Figure 4.1: Crowd-Watson Framework Design (the highlighted components are the ones related to the Game Platform)
main experts on gaming platform as Dr. Detective. Crowd-Watson framework focuses on micro-tasks for knowledge extraction in medical text. The main steps involved in the Crowd-Watson workflow are: pre-processing of the input, data collection, disagreement analytics for the results, and finally post-processing. These steps are realized as an automatic end-to-end workflow, that can support a continuous collection of high quality gold standard data with feedback loop to all steps of the process. The input consists of medical documents, from various sources such as Wikipedia articles or patient case reports. The output generated through this framework is annotation for medical text, in the form of concepts and the relations between them, together with a collection of visual analytics to explore these results. The architecture of this application, and the way its components interact with each other, can be seen in Figure 4.1. We focus on those aspects of the architecture that relate to the Dr. Detective gaming platform for data collection. A full description of the Crowd-Watson architecture is available at [14].

4.1 Pre-Processing for the Game Platform

Typically, the input is available in an unstructured format (e.g. simple text). As part of the input data filtering step, additional metadata, such as the specialization field in which it was published or, for case reports, the diagnosis of the patient, can be extracted from these documents. In addition, some annotation can also be generated automatically, by mapping the text to the UMLS vocabulary of biomedical terminology, classification, and coding standards [4]. The UMLS parser can be used to identify both concepts and relations, however, as a fully automated approach, it suffers from the typical issues of NLP techniques [9], such as lack of contextual awareness, and limited ambiguity processing capabilities. Nevertheless, UMLS annotations can be employed as a good baseline for measuring the efficiency of the crowdsourced answers.

For the purpose of this project, we have employed the UMLS MetaMap parser [3]. MetaMap was used to extract the terms and relations in the UMLS vocabulary from the input medical text. This data could then be used as a baseline for evaluating the answers of the expert medical crowd.

The workers are asked to perform a series of annotation tasks on the input documents. The purpose of these tasks is creating annotation in the form of concepts and the relations between them. We define these tasks according to four micro-task templates:

1. Term extraction – the task of identifying all the relevant terms in a text, where a term refers to a set of words that forms a coherent medical concept;

2. Term categorization – the task of classifying a medical term into an appropriate category, such as the concepts in the UMLS thesaurus;

3. Relation extraction – the task of identifying whether or not a relation exists between two medical terms;

4. Input data filtering
4. **Relation categorization** – the task of classifying a medical relation into an appropriate category (or set of categories), such as the relations in the UMLS thesaurus.

The workers on Crowd-Watson consist of both an expert crowd, and a general crowd. Each of these crowds interacts with the input documents on a specialized platform – for the general crowd, regular crowdsourcing micro-tasks have been constructed on CrowdFlower, whereas the expert crowd employs the *Dr. Detective* application for solving tasks tailored to their profile. The tasks can be solved by both the general, and the expert crowd. The **target crowd setting** step entails picking the difficulty level of the task according to the level of expertise of the crowd. For instance, when discussing term extraction, we argue that the general crowd can reliably find demographic terms, as they do not require significant medical knowledge, whereas the expert crowd can focus on annotating more difficult terminology.

### 4.2 Data Collection: Gaming Platform

Based on the interviews described in Chapter 3, we analyzed incentives for motivating an expert crowd such as competition, learning, and entertainment in the context of working in the medical field, as well as documents that the medical crowd would be interested in reading. We were thus able to identify several key requirements to incorporate into the gaming platform:

- at the level of the input, the interviewees expressed their interest in **reading medical case reports**;
- **learning** about their field, through targeted micro-tasks and extended feedback on their answers, was the most significant motivator;
- the interviewees expected the tasks to challenge their **problem-solving** skills;
- **competition** with peers emerged as a secondary motivator;
- the tasks need to be fun to solve, making **entertainment** as another secondary motivator;
- medical professionals have difficult schedules, and would prefer to have **flexibility** in the time required to engage with the application;

In order to attract users to the application, a goal that is seen as useful by the players needs to be firmly established. As **learning** proved to be the most relevant incentive from the interviews, we focused the goal of the application on this, while also trying to incorporate the **problem-solving** requirement. We developed the concept of a **clue-finding game**, where the text annotation tasks were put in the context of searching for clues in the history of a patient. For instance, when performing the task of term extraction on a patient case report, the user can annotate any of these three **clue types**:

1. the term is a clue **leading** to the final diagnosis of the case;
2. the term is a false clue that is *irrelevant* to the final diagnosis of the case;

3. the term is a *normal condition* that does not influence the final diagnosis of the case.

The clue types can be used as an incentive, involving users with the task they are solving by redesigning it as a medical puzzle, but it can also be used to generate additional annotation. The annotations retrieved from the general crowdsourcing approach are dependent on the context of the sentence where they were identified, so by asking the expert crowd to find meta-relations at the level of the document, we can generate knowledge that is valid generally for the domain. This kind of task cannot be solved simply with the use of contextual information, and requires background knowledge of the field, therefore making it suitable for an application targeted at experts.

The qualitative interviews helped us identify the extrinsic motivators for engaging the medical crowd. After the goal of the application was established, the final step was translating the user incentives into concrete features for building the *Dr. Detective* gaming platform.

### 4.2.1 Difficulty

In order to support the user learning experience and introduce flexibility in task solving, we define the concept of difficulty. This refers to the combination of skill and time required for reading the document, and then performing the annotation task. While it is difficult to hypothesize on the comparative difficulty of performing annotations, the difficulty of the document can expressed as syntactic and semantic difficulty. The syntactic difficulty expresses the effort need for reading the document in three components: the *number of sentences* in the document ($\text{NoS}$), the *number of words* ($\text{NoW}$), and the *average sentence length* ($\text{ASL}$). The semantic difficulty expresses the effort needed for understanding the text in two components: the *number of UMLS concepts* present in the document ($\text{NoUMLS}$), and the *readability* of the document ($\text{SMOG}$). The SMOG [16] formula for computing readability was employed, as it is often recommended for use in evaluating healthcare documents [7]. Therefore, for every document $D$, its difficulty is defined as the norm of the normalized five-component vector:

$$\text{difficulty}(D) = \| (\text{NoS}, \text{NoW}, \text{ASL}, \text{NoUMLS}, \text{SMOG}) \|.$$

### 4.2.2 Scoring

In order to develop the *competition* incentive, a scoring system was devised, to reward players for their work. Through viewing a high score board, they are also encouraged to compete against each other.

We want to reward users when they perform in a way that is beneficial to us. We want to collect the correct answers to the task, therefore, selecting a high-consensus solution should yield more points. This strategy could, however, make users rely entirely on the answers of others. Therefore, in order to encourage a wider answer set and capture semantic ambiguity, we need to give points for newly discovered answers.
Users should also be penalized for giving wrong answers. We also want to encourage users to return to the application, and keep playing. Finally, in order for users to solve tasks in increasing difficulty, scoring needs to be proportional to the difficulty for solving the task [23]. Based on this, for each user $U$ solving a task $T$ on document $D$, we developed the following scoring components:

- $popular(U, D, T)$: the points users receive if they make annotations that were previously selected by at least one other user; we also want to reward partial answers, in order to capture ambiguity;
- $consecutive(U)$: the points users gain the more consecutive tasks they solve;
- $discovered(U, D, T)$: the points users receive if they are the first to discover an answer, if it is then selected by at least one other user;
- $wrong(U, D, T)$: the points users lose if their answers are not selected by any other user.

Based on this analysis, we developed the following scoring formula:

$$score(U, D, T) = difficulty(D) \cdot \left( popular(U, D, T) + consecutive(U) + discovered(U, D, T) - wrong(U, D, T) \right).$$

### 4.2.3 Immersion

In order to develop the *entertainment* incentive, the crowdsourcing application needs to provide immersion inside the task-solving experience. Immersion is based on the concept of game flow [18], which states that at every point in the game, the difficulty
needs to be proportionate with the skill required to solve the task. Skill at playing is acquired by the user as they solve more tasks. If the difficulty is disproportionately large compared to the skill, it will cause anxiety for the user, whereas if the difficulty is too small, the user will be bored. Immersion is achieved when skill and difficulty are proportionally balanced, as illustrated in Figure 4.2.

Immersion is considered when choosing the next document that the user will be asked to solve as part of the game. When a user solves a task on $D_i$, the document they will be asked to solve next needs to have a higher difficulty in order to avoid boredom, but the increase needs to be low enough to avoid anxiety. Therefore, we define the set of possible documents that occur after $D_i$ as:

$$\text{next}(D_i) = \{D_j | \text{difficulty}(D_j) = \min(\text{difficulty}(D_i) - \text{difficulty}(D_t), \forall t \neq i \text{ where } \text{difficulty}(D_t) \geq \text{difficulty}(D_i))\}$$

### 4.2.4 Levels

Finally, in order to satisfy the constraint for flexibility, game levels were implemented to quantify the skill required for solving the tasks. As skill is proportional with difficulty, we define the game levels by quantifying the difficulty metric previously described into three intervals:

1. easy: $\{D | \text{difficulty}(D) \in [0,2]\}$,
2. normal: $\{D | \text{difficulty}(D) \in [3,4]\}$,
3. hard: $\{D | \text{difficulty}(D) \in [5,6]\}$.

These levels should enable users to plan which task they want to solve in accordance to the time they have at their disposal, while also providing a goal-based incentive of progressing in their skill [23].

### 4.3 Game Disagreement Analytics

After the input data is formatted and filtered appropriately through the pre-processing components, it is sent to the data collection component to gather either expert annotation (through the gaming platform) or lay crowd annotations (through the micro-task platform). Next, the annotation results are analyzed with a set of content and behavior-based metrics, to understand how the disagreement is represented in both cases [3, 20], and to assess the quality of the individual workers, and the quality of the individual and overall crowd truth results. In this section, we focus on metrics for the tasks of term extraction and typing. For relation disagreement metrics, see [20].

Disagreement in user-submitted annotations can be found at three levels:
1. **term existence** – this refers to users disagreeing on whether a term exists in the input text; in the context of *Dr. Detective*, by design we have the additional constraint of whether the term is relevant or not to the diagnosis in the case report;

2. **term overlap** – this refers to users disagreeing on what are the words to make up a term;

3. **term typing** – this refers to users disagreeing on what the type of the term is.

We developed the **term metrics** in order to study the ambiguity at the levels of **term existence** and **term overlap**. The result of this analysis is a set of processed term clusters, used to represent any unique concept as the set of all the user-submitted term representations. This data can then be used further in an analysis of the **term typing** ambiguity, through the use of **type metrics**. This analysis can be used to determine whether the employed types are well defined and understood by the crowd, or whether overlaps exist in their definition. Similarly, an analysis at the level of the input through the use of **paragraph metrics** can be used to determine the clarity of the medical text employed in the tasks, as well as provide an understanding of the term clusters and types in relation to the diagnosis retrieved from the patient case reports. The **term metrics** and **type metrics** can be used to draw these conclusions at the level of the paragraph. Finally, the **worker metrics** were developed to assess the performance of users in the game, and identify low-performing workers. These results can be used to enhance the quality of the term data through outlier removal.

### 4.3.1 Term Metrics

These metrics were developed to analyze how disagreement manifests in the way terms are annotated by the crowd. In order to extract a term, the crowd was asked to first identify all the words that make up the term, then assign a type to the term. The purpose of the term metrics is to measure the disagreement in how a term is built by the crowd, i.e. at the level of choosing the words that make up a term.

**Term clusters:** When solving a term extraction task, each user can pick a term, by selecting all the words that compose it. In order to perform an analysis of their answers, we need to be able to identify their common answers. Aside from the obvious full term matches, when two users select all the same words to make up a term, we are also interested in identifying the partial matches. For example, when annotating the following sentence: “symptoms of *acute persistent fever*”, one user could extract the term *acute fever*, while another could choose *persistent fever*. Both of these annotations refer to the same concept, and therefore should be analyzed as being the same term, even though they are not an exact lexical match. In order to identify these partial term matches that refer to the same concept, we need to perform clustering on the crowd answers. Then, the clarity of the cluster will be determined, to measure the diversity of opinion as to how the concept should be represented in the text.

In order to perform term clustering in the answer graph of a paragraph, we represent each user-submitted term as a node. An edge between two terms means that the
two terms refer to the same concept. To determine whether an edge exists between two terms, we compute the Jaccard’s coefficient of similarity \([11]\) on the sets of words that make up the two terms:

\[
Jaccard(T_1, T_2) = \frac{|T_1 \cap T_2|}{|T_1 \cup T_2|},
\]

where \(T_i\) – word set for term \(i\).

Jaccard’s coefficient gives values in the \([0, 1]\) interval, with 0 values for term pairs that do not have any words in common, and 1 values for term pairs that have all the words in common. We are not concerned with representing identical terms in the graph in a special manner, as they will be connected with edges weighted with 1. Therefore, duplicate nodes are allowed to appear in the paragraph answer graph. In order to eliminate false connections, we first remove the English language stopwords from the word sets. We also set a lower boundary of 0.3 for considering the Jaccard’s coefficient value, in order to disregard loosely couple terms. Therefore, if two terms in the graph have a Jaccard coefficient higher than the boundary value, then an edge is added between them.

Finally, in order to extract the unique concepts found by the crowd, we are left with identifying the connected components in the paragraph graph. A connected component represents a stand-alone concept, with various term representations as submitted by the crowd. From here forward, we will refer to it as a term cluster.

**Cluster word set:** This representation refers to the set of all the words that are annotated as part of a term cluster. We define the cluster word set as \(cws(C)\), for each cluster \(C\) in a paragraph. This is computed according to the following formula:

\[
C = \{T_1, \ldots, T_n\}, \text{ where } \\
T_i - \text{word set for term } i, \\
cws(C) = T_1 \cup \ldots \cup T_n
\]

For example, given

\[
C = \{\text{acute, fever}\}, \{\text{persistent, fever}\},
\]

\[
cws(C) = \{\text{acute, persistent, fever}\}.
\]

**Cluster term word frequency vector:** In order to assess the clarity of a cluster, we need to compute the clarity of each of its representations, i.e. the terms that compose the cluster. The clarity of a term in a cluster is given by the frequency with which each word in the term has been annotated. The length of the vector is equal to the size of the cluster word set, with each element in the vector representing the frequency of its corresponding word in the cluster word set. For the cluster \(C\), and its cluster word set \(cws(C)\) in the example above, the word frequency vectors \(twfv(T, C)\) are:
\[ \text{twfv}\{\text{acute, fever}\}, C) = (1, 0, 2), \]
\[ \text{twfv}\{\text{persistent, fever}\}, C) = (0, 1, 2) \]

**Cluster term clarity:** This measures how clearly a term has been annotated by the users. Ideally, all the users that annotate the paragraph would agree on all the words that compose a term. In this case, the cluster term word frequency vector defined previously would be composed of equal numbers, showing the number of annotators of the paragraph. The clarity for such a term should be equal to 1. The more different representations users submit for the same concept, the lower the clarity for the term should be, showing that this representation has a low agreement between annotators. Therefore, a good measure for the clarity of term \( T \) in cluster \( C \) would be computing the cosine between the word frequency vector of \( T \), and a vector of the same size filled with 1 values, representing universal consensus: \( \cos(\text{twfv}(T, C), \vec{1}) \).

We also want to account for the users that did not select any word in the current cluster. For this reason, we compute an overall popularity metric for a term, by comparing its norm to the norm of a vector of the same size where each value is equal to the number of annotators for the paragraph. The norm was employed to account for the size of the term – a term comprised of more words has a higher probability of being annotated by some users, therefore, if no user has picked it, the clarity should be lower than for a similar term made up of less words. The popularity score for a term \( T \) can therefore be computed accordingly:

\[
\text{given total} \quad \text{– the number of people who annotated the paragraph,} \\
\text{popularity}(T, C) = \frac{\|\text{twfv}(T, C)\|}{\|\text{total}\|}. 
\]

To summarize, the clarity of a term \( T \), as part of the cluster \( C \) is given by the following formula:

\[
\text{term}\_\text{clar}(T, C) = \cos(\text{twfv}(T, C), \vec{1}) \cdot \text{popularity}(T, C). 
\]

**Cluster term clarity vector:** This refers to the clarities of each term in the cluster. The size of the vector is given by the number of terms in the cluster, with each value equal to \( \text{term}\_\text{clar}(T, C) \) for its corresponding term \( T \). The clarity vector for a cluster \( C \) is defined as \( \text{cv}(C) \).

**Cluster clarity:** This metric represents the measure for agreement between annotators when extracting a term from a paragraph. It is computed per cluster, to account for partial representations of the same concept. Its value is equal to the maximum clarity value of all the terms in the cluster:

\[
\text{cluster}\_\text{clar}(C) = \max(\text{term}\_\text{clar}(T, C), \forall T \in C). 
\]
4.3 Game Disagreement Analytics

4.3.2 Type Metrics

Next we want to analyze how disagreement is present at the level of selecting types for the annotated terms. We continue to work with aggregate partial term matches, i.e. term clusters, in order to capture the various lexical representations of the same concept.

Worker cluster type vector: For each user $U$ annotating a term from a cluster $C$, we construct a vector of 0 and 1 values, representing all the types that the user assigns to the term. The size of the vector is equal to the size of the possible types set, with each element in the vector corresponding to a type. At any position in the vector, the value is equal to 1 if the user has assigned the corresponding type to the term, and 0 otherwise. Thus, we define the worker cluster type vector as:

$$wctv(U,C) = (x_1, \ldots, x_n),$$

where $x_i = 1$ if $U$ assigned type $i$ to $\forall T \in C$, $x_i = 0$ otherwise.

Cluster type vector: For each cluster $C$, we construct a vector of all the types assigned to it. It is equal to the sum of all worker cluster type vectors:

$$ctv(C) = \sum_U wctv(U,C).$$

Cluster type clarity: This is a metric to measure the agreement for assigning a type to a cluster. High clarity is defined as a high agreement towards one type, whereas low agreement means different annotators assigned different types to the terms in the cluster. This can be measured using the cosine of the cluster type vector, and a set of unit vectors, one for each type. For instance, a unit vector for type $i$ would have the 1 value in the value at position $i$. The cluster type clarity can then be defined as the maximum value of the cosine with each type unit vector:

$$type\_clar(C) = \max(\cos(ctv(C), uv_i)), \forall uv_i – \text{unit vector for type } i.$$
Paragraph cluster clarity: This metric measures the agreement for all the terms annotated in the paragraph. It is given by the cosine of the paragraph cluster vector, and a vector of equal length formed of 1 values, which signifies universal agreement. Ideally, all the terms in the paragraph are annotated by all the users reading the paragraph, giving a clarity equal to 1. The more disagreement between the annotators, the lower the clarity value. The clarity for a paragraph $P$ is therefore given by the following formula:

$$pcc(P) = \cos(pcv(P), \vec{1}).$$

Paragraph cluster type vector: This vector collects the type clarities $\text{type}\_\text{clar}(C)$ of all clusters in a paragraph. For a paragraph $P$, it is defined as:

$$ptv(P) = (\text{type}\_\text{clar}(C_1), \ldots, \text{type}\_\text{clar}(C_n)), \forall C_i \in P.$$

Paragraph cluster type clarity: This metric measures the agreement for all the term types assigned in the paragraph. It is given by the cosine of the paragraph cluster type vector, and a vector of equal length formed of 1 values, which signifies universal agreement. Ideally, all the terms in the paragraph are given the same types by all the users reading the paragraph, giving a clarity equal to 1. The more disagreement between the annotators, the lower the clarity value. The type clarity for a paragraph $P$ is therefore given by the following formula:

$$ptc(P) = \cos(ptv(P), \vec{1}).$$

### 4.3.4 Worker Metrics

To track the individual performance of a user in the crowd, the expert metrics were developed. For each sentence in the input, the performance of the worker can be measured as a set of vectors, like the worker paragraph clusters vector $wpcv(U, C)$ and the worker paragraph types vector $wctv(U, C)$ defined below. Then the worker metrics are employed to determine how well the workers performed on the tasks.

Worker paragraph clusters vector: This vector aggregates the terms submitted by a user per cluster inside a paragraph. Its size is equal to the number of clusters in the paragraph, and it is formed out of 0 and 1 values, with 1 meaning that the user submitted a term in the corresponding cluster. For user $U$ annotating paragraph $P$, the vector is defined as:

$$wpcv(U, P) = (x_1, \ldots, x_n),$$

$$x_i = 1 \text{ if } \exists T \text{ a term annotated by user } U, T \in C_i, C_i \in P,$$

$$x_i = 0 \text{ otherwise.}$$
**Worker paragraph types vector:** This vector aggregates the types submitted by a user per cluster inside a paragraph. Its size is equal to the number of available types, and it is formed by adding the worker cluster type vectors for this paragraph. For user $U$ annotating paragraph $P$, the vector is defined as:

$$wptv(U, P) = \sum_{C \in P} wctv(U, C).$$

**Worker-paragraph similarity for clusters:** This metric measures how close were the terms annotated by the user to the terms annotated by the rest of the crowd, when choosing the words that form a term. A high similarity means that the worker had a high agreement with the crowd, whereas a low similarity could potentially indicate a spammer, or an unskilled worker. To determine this, we compute the cosine of the worker paragraph clusters vector $wpcv(U, C)$ and the paragraph cluster vector, with the answers of the current user removed:

$$wpsc(U, P) = \cos(wpcv(U, P), pcv(U, P')),$$

where $P'$ – the answer set for paragraph $P$ with the answers of $U$ removed.

**Worker-paragraph similarity for types:** This metric measures how close were the types annotated by the user to the terms annotated by the rest of the crowd. A high similarity means that the worker had a high agreement with the crowd, whereas a low similarity could potentially indicate a spammer, or an unskilled worker. To determine this, we compute the cosine of the worker paragraph types vector $wptv(U, C)$ and the paragraph types vector, with the answers of the current user removed:

$$wpst(U, P) = \cos(wptv(U, P), ptv(U, P')),$$

where $P'$ – the answer set for paragraph $P$ with the answers of $U$ removed.
Chapter 5

“Dr. Detective” Experiment Pilot

5.1 Experimental Setup

In order to test the feasibility of the Dr. Detective setup, we implemented a version of the workflow described in Chapter 4 and set up a pilot run involving a crowd of medical professionals. As part of our pilot run, we performed an initial evaluation of both the quality of the answers, and the user enjoyment as part of this gamified crowdsourcing platform. The goal of this experiment was to answer the research questions described in Chapter 1, which will be discussed as part of our results. An appropriate setup was designed for each research question as follows:

1. How will the expert crowd compare to an automated NLP approach at extracting a gold standard for medical text?
   We compared the annotations extracted by the crowd with data collected from the UMLS thesaurus through the use of the MetaMap parser.

2. Does having access to the answers of other users stimulate diversity of opinion?
   We introduced the feature of accessing the answer of previous users that have extracted annotation on the same input.

3. Is diversity of opinion an inherent trait in medical text, or does it indicate low-quality data (i.e. as a result of a poorly defined task, or low-quality workers)?
   We employed the disagreement analytics introduced in Chapter 4.3 to answer this question.

4. Can a gamified crowdsourcing platform be employed to capture annotations of an expert crowd?
   We surveyed the enjoyment of the users through a questionnaire.

In order to answer these questions, we set up two versions of the game, one in which users had the ability to see the answers of others (Figure 5.1), and one in which they did not (Figure 5.3). In addition, some of the gaming elements that would ensure the users keep in the state of game flow (high scores board, next document selection
5.1 Experimental Setup

5.1.1 Input

Based on a suggestion in the qualitative interviews, the input was selected from clinical cases published in the New England Journal of Medicine. 10 documents were picked out of four of the most popular specialties (Hematology/Oncology, Nephrology, Primary Care/Hospitalist/Clinical Practice, Viral Infections). The diagnosis was extracted from each document, based on a string matching procedure performed on the text marked in “diagnosis” section headings (e.g. clinical diagnosis, pathological diagnosis etc.). The documents were split into paragraphs, to increase the ease of reading, and the difficulty metrics (described in Chapter 4.2.1) were then applied to each paragraph. Finally, we selected a set of 20 paragraphs, with the values in the difficulty vector uniformly distributed to represent a broad range of text types, to use for the
5.1 Experimental Setup

Figure 5.3: Screenshot from the “Dr. Detective” application, the version simplified for the experiment

game, as we wanted to ensure that all of the text would be annotated in the limited time frame of the experiment run.

When creating an account for the game, the users were randomly assigned to one of two groups, to determine which set of paragraphs they would solve in which game setting. Therefore, any user could annotate 10 paragraphs in the full game version, and then 10 other paragraphs in the simple version.

5.1.2 Task

The micro-task templates (described in Chapter 4.1) selected for this pilot were (1) term extraction, and (2) term categorization. Based on how relevant they are at describing patient case reports, 3 meta-types, each with a set of term types taken from UMLS, were selected and implemented in the interface for the categorization task. The type selection menu can be seen in Figure 5.2. In total, 13 term types were available for the users to annotate. As most interviewers expressed their interest in a problem-solving application, we decided to set the clue type user seek as part of the application (described in Chapter 4.2) to (1) the term is a clue leading to the final diagnosis of the case. The users could therefore extract any medical term from the input paragraph, by selecting all the words that form a term, and assign a type to the term. When they were done with the input paragraph, they could submit their answers for that specific paragraph as an answer set. Such a set would contain all the annotated terms, as well as their corresponding types. The submitted terms could not have overlapping words, and they were restricted to a single type.

Finally, in order to encourage the diversity of opinion, and therefore capture ambiguity, we allowed users to look at the answers of others for the task they are solving.
This feature was made available through a button, which the users could choose to press in order to toggle the other answers. The scoring formula (described in Chapter 4.2.2) ensures that users are motivated to find new answers even in this circumstances, through the use of discovery bonus points. The users could access the details of how their score was computed through a hover notification in the menu. An example of how this task was presented to the users as part of the Dr. Detective interface can be seen in Figure 5.1.

5.1.3 Users

The pilot run of the Dr. Detective game had 11 participants in total, with 10 players engaging with the full game version, and 7 engaging with the simple version. In total, 155 user-submitted answer sets were collected for all the paragraphs, with each paragraph solved as part of 2 to 7 different game rounds. Finally, 6 players completed the feedback questionnaire.

5.2 Results and Discussion

5.2.1 Crowd Answers Compared to an NLP Parser

We started by analyzing the first proposed research question: How will the expert crowd compare to an automated NLP approach at extracting a gold standard for medical text? To answer this, we selected the three paragraphs that were played by the most number of users (paragraphs 12, 14 and 15), and compared the answers to the term list generated by the UMLS MetaMap parser (employed for the input pre-processing part referenced in Chapter 4.1) for the same paragraphs.

We first analyzed how well the crowd performed at extracting relevant words in the text, regardless of the term the words were assigned to. Fig. 5.4 shows the crowd was able to identify the majority of the words annotated with UMLS.

Then, we analyzed how the crowd terms (i.e. sets of words identified by the crowd) compare to the terms extracted by the UMLS parser. Fig. 5.5 shows that around one...
third of the terms in UMLS had a full match (i.e. with all the words in common) with terms annotated by the crowd. Factoring in the partial term matches (i.e. with at least one word in common), the crowd was able to identify most of the UMLS terms. This shows the efficiency of the crowd answers is quite high, enough for the crowd to be considered as a viable alternative to automated named-entity recognition, provided that enough users give their input for a paragraph.

### 5.2.2 Game Features for Capturing Diversity of Opinion

In this section, we analyzed the second research question: *Does having access to the answers of other users stimulate diversity of opinion?* To answer this, we looked at how diversity of opinion was expressed by the game users. Specifically, we were interested in finding out whether being able to see the results of other people will stimulate disagreement, or rather make users select each other’s answers. In order to achieve this, we looked at how the answers per paragraph varied according to the version of the game that the user played. In the full version of the game, a user playing the current game round on a paragraph can see all the answers submitted in previous rounds for the same paragraph. We hypothesized that having access to the annotations of others would increase the number of total words annotated, a behavior that should be observed after each game round. We also expected that each paragraph would reach a saturation point when all the relevant terms have been annotated by the crowd, after which no new words should be discovered by the crowd.

Fig. 5.6 shows how the number of new words per paragraph increases after each round of the game, for the top three paragraphs described previously. Each version of the game seems to follow the same progression in the rate of new words identified, with the first users finding most of the words, and then only slight increases as the paragraph is played by other people. However, the simple version of the game seems
to constantly feature a higher total word count, as opposed to the full game version. The same trend was observed both for the number of new types, and the number of distinct terms. This seems to indicate that the full game version was less encouraging for collecting a wide array of terms.

The number of annotators did not seem to influence the total annotations collected for a paragraph, with paragraph 15 collecting a considerably higher number of words than the other two paragraphs for the same number of annotators. There was also no obvious saturation point reached for any of the paragraphs, even though paragraphs 12 and 14 seemed to collect less annotations as the number of users solving it increased. From this we conclude that a higher number of participants is needed to determine where the saturation point for a paragraph can be found, and what are possible factors that influence it.

In order to rule out an issue related to some other feature in the full game version, we looked at how the behavior of pressing the button to view other answers affected the output. Out of 67 game rounds played in the full version, this button was only pressed in 18 of the rounds, so it appears this was not a popular feature to begin with. Fig. 5.7 shows that, actually, users tended to annotate more words in total when they pressed. However, as evidenced in Fig. 5.8, the ratio of new words to total words in this case was much lower than when the button was not pressed. Additionally, it appears there is not much difference between the simple version of the game, and the full version, but where the users chose not to look at the answers of others. Therefore we can infer that having access to all the answers makes the crowd act more conservative, selecting less new words, but rather choosing to validate the answers of others.

When looking at the answers in the questionnaire related to the usefulness of seeing other people’s annotations, we found that most people (67%) were ambivalent to having the option of checking their answers. Some users reported using this feature as a tool for better understanding the task, while others claimed it validated the answers they had already chosen. Overall, it seems that having access to all the other answers makes users less likely to find and annotate new words, which could mean a loss in the ambiguity of the annotation. It also provides an unfair advantage to the first users to...
annotate a paragraph, as their score would likely keep increasing as other people keep selecting their answers.

5.2.3 Disagreement Analytics

In this section, we provide an answer to the third research question: *Is diversity of opinion an inherent trait in medical text, or does it indicate low-quality data (i.e. as a result of a poorly defined task, or low-quality workers)?* We analyzed the way ambiguity of language is represented in the crowd results through the use of the disagreement metrics presented in Chapter 4.3. Ambiguity in the answer set can have multiple meanings – it could be used to confirm our hypothesis that diversity of interpretation is inherent to medical text, and needs an appropriate representation that is different from the traditional gold standard approach, but it could also show that the annotation tasks were poorly defined and/or understood by the medical crowd, or that the medical crowd is not reliable in producing quality annotation. In this section, we will discuss the results of applying these metrics to the crowd data to determine the quality of the answers from the angle of capturing diversity of opinion, but also to assess whether the disagreement metrics are indeed useful at describing this data.

As detailed in Chapter 4.3, disagreement in user-submitted annotations can be found at three levels: *term existence, term overlap*, and *term typing*. Based on them, in this section we discuss the three following issues:

1. **clustering analysis** – its purpose is to study how ambiguity is present at the level of the term composition, by analyzing the disagreement over term existence and term overlap;

2. **type analysis** – its purpose is to study how ambiguity is present at the level of the type assignment, by performing an analysis of the term type disagreement; this is useful to determine the clarity of the term types we employed, i.e. whether they were well defined, or whether the users found them ambiguous;

3. **worker performance analysis** – performed in order to identify possible low-quality users, whose answers would skewer the disagreement statistics.

The first step required extracting the individual concepts from paragraphs, accomplished through clustering the user-submitted terms according to the words they had in common. In order to study inter-user agreement in the game, we removed the paragraphs annotated by just one person. From the 18 remaining paragraphs, we extracted 382 unique terms, and 284 individual term clusters. An example of how a term cluster set for a given paragraph looks like can be seen in Figure 5.9, with each user submission represented as a node, and edges between lexically similar answers. The clustering procedure reduced the total term set with around 25%, indicating that, even though the answer set contained a majority of full term matches, some procedure for identifying overlapping terms might still be needed.
Figure 5.9: Example term clusters for the paragraph: A 60 year-old man was admitted to this hospital because of acute fever.

1. Clustering analysis conclusion: This analysis showed the existence of ambiguity for both term existence and term overlap over the answer set. This indicates the need for multiple representations of the same term. However, we still need to investigate whether the clustering procedure previously outlined produces a good accuracy for analyzing the results.

Next we analyzed the distribution of the cluster clarity values for the user-identified terms. Figure 5.10 shows the value distribution over medical observation types, and Figure 5.11 shows the distribution over demographics and miscellaneous types. This analysis can help us determine which term types are well defined, and which were found ambiguous by the annotators. The clarity of most medical observation types (Factors, Diseases, Medical Tests, Therapeutic Procedures) tends towards a mean of 0.5, seeming to indicate that the categories are somewhat ambiguously defined. This is consistent with answers from the user questionnaire – “I found the term Factors and Medical tests overlapping”, “I thought that the definition for the term factors should be more clear” (from Appendix D). In contrast, the types Allergies and Medications showed relatively high median clarity. Terms referring to Medications (e.g. the names of various drugs) seem to be less ambiguous. However, the high clarity for Allergies could also be the result of a low sample set, as only 12 of the 284 term clusters were attributed to this type by the users.

The clarity scores appeared low for the demographics and miscellaneous types. Low agreement for the Other type was expected, due to its loose definition. Conversely, the Time/Duration type appears relatively well defined, with an agreement score of 0.6. This could be explained by the formulaic representation of time intervals in case reports (e.g. 20 years earlier, previous 5 years), which does not leave much room for disagreement between annotators.

It is not clear whether the low clarity of the rest of the types is due to their ambiguous definitions, or rather the preference of the expert users to focus on medical observations when annotating a case. While every user annotated at least one medical observation per paragraph, they were only 50% likely to extract an other term type, whereas a manual inspection showed that 15 of the 18 analyzed paragraphs contained
such terms. This finding could indicate that either the expert crowd is only suitable for extracting medical observation annotations, or the expert user find demographic terms irrelevant in the context of finding clues for a diagnosis. The low clarity scores could also be influenced by low quality workers submitting unreliable answers. For this reason, an analysis of the user performance in solving the tasks is needed.

2. Type analysis conclusion: When analyzing medical observation types, we found some type definitions (e.g. Factors, Medical Tests) need to be made more clear, as they produced ambiguous data. The demographic types also cored low in user agreement, either as a result of poor crowd performance, or because they were irrelevant in the context of the task. In order to determine the suitability of the expert crowd for this task, future experiments should build the task to be more inclusive, asking users to select terms that are both relevant and irrelevant to the diagnosis. An analysis of the individual performance of the workers is also needed, in order to remove outliers from the answer set.
5.2 Results and Discussion

"Dr. Detective" Experiment Pilot

The worker-paragraph similarity scores are available in Figures 5.12 (for term clusters) and 5.13 (for term types). While the graphs show a cluster of consistently high-performing users (scores above 0.6), the issue of identifying low-performers appears more complicated. Only two users exhibit consistent low scores (under 0.4), while the majority varies between high and low. Furthermore, the graphs do not match the low-performing clusters between the cluster and the type similarity (i.e. the same users have high clarity for term building and low clarity for term typing on the same paragraphs). Low agreement scores also seem to be influenced by the number of people annotating the paragraph – the less annotators, the more diversity of opinion is observed among them. This seems to indicate that there are no obvious spammers amongst the expert crowds. The differences in performance across paragraphs could then be explained by the field of expertise of the user. However, overall, the sample set that we collected appears too small to identify any obvious trend. This is in contrast with the previous analysis of the answers’ quality, which showed that even a small set of users can produce meaningful data.

3. Worker performance conclusion: No obvious low-quality users were detected from the performance analysis. However, this is most likely the result of a small sample set, rather than a valid trait of the expert crowd. Further experiments require a larger user set to assess performance trends, in contrast with the analysis of term extraction and typing data, which produced meaningful results in spite of the small number of annotators.
5.2.4 User Feedback

In this section, we tackled the final research question: *Can a gamified crowdsourcing platform be employed to capture annotations of an expert crowd?* To answer this, we surveyed the enjoyment of the users involved in the pilot. The full results of the user questionnaire are available in Appendix D. We were interested in assessing whether immersion in the game occurred for the users involved, and how each individual game feature (described in Chapter 4.2) was rated:

- **game flow** – It was reported to be good, with 83% of the users saying they were neither too bored, or overwhelmed.

- **levels and difficulty** – Most users found the levels to be a useful addition, with 50% being satisfied with the level progression, and 33% being ambivalent to it. However, some users pointed out that they expected more challenge from the advanced level. As the difficulty is currently computed only based on textual metrics, the game could potentially get boring for users. For this reason, domain difficulty should be incorporated in future versions of the game.

- **scoring** – This part of the game was less well received, with 83% of the users declaring they found the way their score is computed only somewhat clear. Therefore, in future game versions, a more detailed scoring breakdown should be implemented, with users being able to access the history of the cases they solved.

- **overall enjoyment** – Most users reported to have enjoyed the game, and expressed an interest in returning to play, provided they can solve more difficult cases and get more feedback. The full game version was almost universally preferred by the users.
Chapter 6

Future Work

In this chapter, we describe the issues that we discovered after running the initial experiments on the Dr. Detective platform, and how we plan to tackle them in future work. We focus both on issues on the user side, exploring how the setup and task design could be made more user friendly, but also how to improve the gameplay to make it more immersive. We also discuss the quality of the results we gathered in the context of capturing disagreement, and how future experiments can be set up to improve this data. Finally, we explore possible new tasks for the expert crowd.

6.1 Game Design and User Interaction

In this section we propose improvements for both the game design and the user interaction components of our crowdsourcing platform.

**Access to others’ answers:** The experiment results in Chapter 5.2.2 indicate that having access to the answers of others did not encourage users to find more relevant terms in the text. However, answers from the user questionnaire seem to indicate that players found this feature to be useful – “I liked seeing others answers because it validated my responses”, from Appendix D. In future work, the feature of seeing others’ answers needs to be integrated in such a way that it still enables us to capture the ambiguity of language. A possible option would be to transform it into a bonus resource, with users having access to a limited amount of answers, which they can then increase by continuing to play the game. In order to encourage users to find new terms, future versions of the game could also prevent the users from selecting the most popular answers. In this case, a mechanism that ensures that users who come later at annotating a paragraph are not at a disadvantage as compared to the early users, who had the opportunity to annotate the most popular answers.

**Challenging hard levels:** The user questionnaire discussed in Chapter 5.2.4 also provided some ideas for improving the user interaction. A more careful review of the input text should enable us to generate more complex scenarios for the hard level. This includes selecting more difficult case reports (e.g. relating to rare diseases), but also reviewing the difficulty metrics discussed in Chapter 4.2.1 When selecting for a hard...
input, more emphasis should be put on the semantic difficulty components (i.e. medical field, number of UMLS terms), as opposed to the syntactic difficulty components, which are less suitable at describing challenging cases.

Score reports: The users also requested a better breakdown of the way their score is computed. To this end, we plan to implement score reports, which users can employ to review their past submissions in the game, and the way their answers were rated by the rest of the crowd.

Task timer: Finally, a possible feature to add would be a timer for the task solving time. This could be integrated into the scoring mechanism, giving higher points for quicker answers, but also used to set the maximum time during which the users can solve a task. This can help reduce the low quality workers by not registering their answers, while also providing an extra challenge to the users.

6.2 Capturing Disagreement

In this section, we propose improvements in the experiment and task design, that would enable us to collect data of higher quality, while capturing the language ambiguity.

Allowing overlapping terms: The results of the cluster analysis in Chapter 5.2.3 showed a 25% reduce in the number of terms after the clustering procedure was applied. The accuracy of the clustering analysis could be influenced by the strict assignment of words to terms – every word could be assigned by a user to at most one term. In future version, allowing overlapping terms, i.e. where a user can share one or more words between two terms, could potentially influence these results, by increasing the connections between the clusters. The clarity scores for both terms and types could also be influenced by this change.

Term ranking: Our understanding of the extracted terms could also be improved by asking the users to rank their importance to the diagnosis in the input case report. By generating a ranking of importance, we would be able to differentiate between terms that are general with regard to a diagnosis, and specific terms. For instance, the term “fever” would be considered too general to lead to a diagnosis. However, “acute persistent fever” is quite specific – there are considerably less options for diagnoses that contain. Even though by our metrics, both of these terms would be classified in the same cluster, the concepts they define are not fully equivalent. In addition, by generating a term ranking in relation to the diagnosis, we can perform an analysis of conjunctions of terms, indicating which terms are more likely to appear together in the context of the same disease or diagnosis.

Redefining the Factors type: The typing analysis in Chapter 5.2.3 showed some of the medical observation types to be ill defined. After also consulting the user questionnaires, we found that the Factors type was ambiguously defined, being routinely confused with types such as Medical Tests and Therapeutic Procedures. In future, we plan to replace the Factors type with the two more specific Risk Factors and Symptoms
types, in order to improve user understanding of the types. We hypothesize this will result in increased clarity scores for the medical observation type set.

**More clue types:** Also from the type analysis, it is unclear whether expert users are not interested in selecting demographic types, or they find them irrelevant to the diagnosis. In order to solve this issue, the task needs to be redesigned such that users can annotate both terms that are relevant, and irrelevant to the diagnosis, by making all clue types discussed in Chapter 4.2 available. In this manner we would be able to draw a conclusion on the relevance of demographic terms for conducting a diagnosing procedure.

**Alignment to Watson answers:** Finally, the users could benefit from an integration with the Watson system. IBM Watson already contains a vocabulary of terms and relations in the medical domain. These terms can be provided as hints to the users, in the same vein as showing the annotations of others is done, or as a feedback for completing the task, by showing how the confidence of Watson would change for that term after taking into account the user’s answer.

### 6.3 New Tasks for the Crowd

In this section we discuss possible new directions for the *Dr. Detective* crowdsourcing platform, specifically implementing new tasks for the expert crowd to solve.

**Relation extraction:** The term types described previously entail certain relations that could possibly appear in the input paragraphs. For instance, *Risk Factors* and *Diseases* terms can often be grouped in a *causes* relation. Another challenging relation that could potentially appear in patient case reports is *contra-indicates*, as the terms that form it can end up separated by various sentences. For this reason, future versions of *Dr. Detective* could ask the expert crowd to also identify relations between terms as part of the annotation task. An option to extract negative relations should also be provided.

**Passage alignment:** The terms extracted by the expert crowd could be used as input for tasks in future versions of the game. For instance, passage alignment involves giving two passages, each of them contains a term that has been identified by the crowd, and asking the crowd whether those two terms mean the same thing in general, or only in this instance. This data could then be used to expand the way term clusters from the paragraph to the domain level.
Chapter 7

Conclusions

This document proposes a design for Dr. Detective – a gamified crowdsourcing platform to extract annotation from medical text. In order to build a gold standard for the medical domain, we argue that there are two annotation aspects that need to be distinguished in order to improve quality of the annotated data, i.e. (1) diversity of linguistic expressions that a lay crowd with no domain knowledge can identify, and (2) diversity of semantic interpretations that a crowd of domain experts can extract. Dr. Detective was developed in the context of Crowd-Watson, a general crowdsourcing framework for extracting text annotation by engaging both a general crowd, and an domain expert crowd. The gaming platform was designed taking into account motivators to engage a crowd of medical experts. The crowd incentives were collected through a series of qualitative interviews, resulting in the decision to implement the crowdsourcing platform as a clue finding game. Specific gamification elements were incorporated, such as difficulty, scoring, immersion, and levels.

A first version of Dr. Detective was implemented and tested. The pilot run showed that the quality of the results of the crowd are comparable to those of an NLP parser. Allowing users to see the answers of others resulted in increased agreement, and thus decreased the desired diversity in answers. A further analysis of the disagreement showed that some of the types we employed (e.g. Factors) were ambiguously defined, and that more data needs to conduct a proper analysis of user performance. The overall user feedback for the application was positive. However, it was clear that users desire more complex challenges in order to keep them engaged. Furthermore, detailed score reports could be a good candidate to increase the benefits the users get from playing the game, as well as using it as an additional training component for the users to refine their contributions, so that they create optimal ground truth data.

Based on these results, we propose the following answers to the research questions introduced in Chapter 1:

1. How will the expert crowd compare to an automated NLP approach at extracting a gold standard for medical text?
   
   Even with a small number of participants, the expert crowd was able to identify most of the annotations extracted by an NLP parser.

2. How is diversity of language present in crowd annotations?


Conclusions

a) Does having access to the answers of other users stimulate diversity of opinion?

Access to the answers of others increases agreement between annotators, therefore it is not a useful feature for capturing diversity of opinion. However, since users expressed they found the feature to be useful, it should be redesigned, not discarded, in future versions of the application.

b) Is diversity of opinion an inherent trait in medical text, or does it indicate low-quality data (i.e. as a result of a poorly defined task, or low-quality workers)?

The answer dataset collected is too small for any definitive conclusion. However, the crowd results so far are promising, with language ambiguity present at both the term formation, and term typing levels. Future versions should provide more accurate data, by redesigning the tasks we identified as poorly defined, and collecting more user data to identify trends in worker performance.

3. Can a gamified crowdsourcing platform be employed to capture annotations of an expert crowd?

The user group gave a positive feedback to the application, noting their interest in the concept of a clue finding game. Future versions could generate more users by providing more challenging input content and extended user feedback.

In future, we plan to extend Dr. Detective with features that would make it both more engaging (e.g. score reports) and more challenging (e.g. task timer, more difficult levels). In order to better capture diversity of opinion, we will allow users to annotate overlapping terms, introduce term ranking, and redefine the ambiguous term types. We also plan to introduce new annotation tasks for the expert crowd, like relation extraction and passage alignment. Finally, in order to improve the quality of the result set, we aim to increase the number of participants, and run experiments for a longer period of time.
Bibliography


Appendix A

Interview Questionnaires

A.1 Medical Students

1. General questions:
   a) What is your study course/specialization?
   b) How many years have you spent studying medicine?
   c) What kind of courses you have been taking as part of your program (e.g. lecture, laboratory session)?
   d) Describe the contents of your program. How does it progress across the study years?

2. Reading:
   a) What kind of medical texts have you (have to) read? (e.g. textbooks, scientific publications)
   b) Can you name some (online) sources?
   c) Why do you read them? (e.g. preparations for exams, assignments, informative, entertainment etc.)
   d) Do you usually read the whole “resource”, or do you read it in parts?
   e) Do you take notes while reading, and how?
   f) Do you usually need to go back and re-read the same text more than once?

3. Learning:
   a) Describe the way your study program is set up, in terms of schedule, types of exams etc.
   b) Do you study alone or in groups? How big are those groups?
   c) Do you usually use question cards?
   d) How much time do you need to prepare for an exam? For an assignment?
   e) Do you use educational software in your study? (e.g. question card programs, quizzes etc.)
4. Competitions:
   a) Are you aware of any medical (award) competitions? What kind? Have you participated in them?
   b) Which medical areas or topics would you be most comfortable competing in?

5. Gaming:
   a) Do you play video games? Single/multi player? On PC/console/mobile? What kind (e.g. role-playing, social network, puzzle etc.)?
   b) How much time do you spend playing per day? Per week?

6. Participating in crowdsourcing and community activities:
   a) Do you contribute to community projects for sharing medical knowledge (e.g. writing Wikipedia articles)? Can you describe the projects, along with the tasks that you perform?
   b) Would it be possible to incorporate a crowdsourcing activity (like the one described above) with your course work?
   c) What (medical) topics would be best suited for such an activity?
   d) What kind of tasks would you like to solve? (e.g. reading, testing your knowledge etc.)
   e) How much time would you spend on this activity per session? How much time per week?
   f) Would you prefer to complete the activity alone, or in a group?
   g) What would make such an activity useful to you?

7. Further participation:
   a) Do you want to receive a transcript of this interview?
   b) Do you want to stay updated on the progress of this project?
   c) Do you want to try out our application once we have an initial version running?

A.2 Medical Professionals

1. General questions:
   a) What is your study course/specialization?
   b) How many years have you spent studying medicine?
c) What kind of courses you have been taking as part of your program (e.g. lecture, laboratory session)?

d) Describe the contents of your program. How does it progress across the study years?

2. Reading:

a) What kind of medical texts have you (have to) read? (e.g. textbooks, scientific publications)

b) Can you name some (online) sources?

c) Which of those you consider a “regular” reading material, and which “occasional”?

d) Why do you read them? (e.g. preparations for exams, assignments, informative, entertainment etc.)

e) Do you usually read the whole “resource”, or do you read it in parts?

f) Do you take notes while reading, and how?

g) Do you usually need to go back and re-read the same text more than once?

3. Learning:

a) Describe the way your study program is set up, in terms of schedule, types of exams etc.

b) Do you study alone or in groups? How big are those groups?

c) Do you usually use question cards?

d) How much time do you need to prepare for an exam? For an assignment?

e) Do you use educational software in your study? (e.g. question card programs, quizzes etc.)

f) How do you stay in touch with your fellow students to discuss assignments, course recommendations, and so on? Do you use social media (e.g. Facebook groups)?

4. Competitions:

a) Are you aware of any medical (award) competitions? What kind? Have you participated in them?

b) Which medical areas or topics would you be most comfortable competing in?

5. Gaming:

a) Do you play video games? Single/multi player? On PC/console/mobile? What kind (e.g. role-playing, social network, puzzle etc.)?

b) How much time do you spend playing per day? Per week?

6. Participating in crowdsourcing and community activities:
a) Do you contribute to community projects for sharing medical knowledge (e.g. writing Wikipedia articles)? Can you describe the projects, along with the tasks that you perform?

b) Would it be possible to incorporate a crowdsourcing activity (like the one described above) with your course work?

c) What (medical) topics would be best suited for such an activity?

d) What kind of tasks would you like to solve? (e.g. reading, testing your knowledge etc.)

e) How much time would you spend on this activity per session? How much time per week?

f) Would you prefer to complete the activity alone, or in a group?

g) What would make such an activity useful to you?

7. Further participation:

a) Do you want to receive a transcript of this interview?

b) Do you want to stay updated on the progress of this project?

c) Do you want to try out our application once we have an initial version running?

A.3 Lecturers

1. General questions:

a) What is your specialization?

b) What kind of classes do you teach?

2. Reading material and activities (for students):

a) Is there a list of obligatory reading?

b) What kind of (medical) texts do you usually assign for reading?

c) How much time do the students spend on reading?

d) Do you have any other comments about student reading habits?

3. Student homework & assignments:

a) What are typical types of homework? (e.g. summarizing, answering questions etc.)

b) How much time do the students spend on assignments and homework?

c) Do the students usually work individually, or in teams? How big are the teams?

4. Exams:
a) How are the exams usually structured? What types of (how many on the average) questions do the students need to answer?

5. Students participating in crowdsourcing and community activities:
   a) Would it be possible to incorporate (how and where) the student work on homework, assignments or exam preparation as part of such crowdsourcing activities?
   b) What (medical) topics would be best suited for such an activity?
   c) What kind of tasks would be best suited for students to solve as part of this? (e.g. reading, testing their knowledge etc.)
   d) What would make such an activity useful to your students?

6. Reading:
   a) What kind of medical texts have you (have to) read? (e.g. textbooks, scientific publications) Can you name some (online) sources? Why do you read them? (e.g. preparations for exams, assignments, informative, entertainment etc.)
   b) Do you usually read the whole “resource”, or do you read it in parts?
   c) Do you take notes while reading, and how?
   d) Do you usually need to go back and re-read the same text more than once?

7. Gaming:
   a) Do you play video games? Single/multi player? On PC/console/mobile? What kind (e.g. role-playing, social network, puzzle etc.)?
   b) How much time do you spend playing per day? Per week?

8. Professors participating in crowdsourcing activities:
   a) What would make such an activity useful to you or your colleagues?

9. Further participation:
   a) Do you want to receive a transcript of this interview?
   b) Do you want to stay updated on the progress of this project?
   c) Do you want to try out our application once we have an initial version running?
Appendix B

Interview Transcripts

B.1 Medical Students

B.1.1 Interview Medical Student S1

1. General questions:
   a) What is your study course/specialization?
      General medicine, but I did a BSc in biomedical sciences.
   b) How many years have you spent studying medicine?
      I am in my third year.
   c) What kind of courses you have been taking as part of your program (e.g. lecture, laboratory session)?
      Both [lectures and laboratory sessions].
   d) Describe the contents of your program. How does it progress across the study years?
      Each year is centered around a theme. For example, in the one you learn what are the diseases, in the next year you learn how to treat them.

2. Reading:
   a) What kind of medical texts have you (have to) read? (e.g. textbooks, scientific publications)
      Mostly textbooks, in Dutch, but some scientific publications as well. I prefer printed texts, but e-books are better for looking at images.
   b) Can you name some (online) sources?
      The textbooks we get from the university library. The publications are recommended to us by the professors.
   c) Which of those you consider a “regular” reading material, and which “occasional”?
      I read textbooks more often than publications, but both are regular reading.
d) Why do you read them? (e.g. preparations for exams, assignments, informative, entertainment etc.)
   I read mostly to prepare for exams, but I am also part of a reading group, where some of us third year students meet every other week to discuss some publications.

e) Do you usually read the whole “resource”, or do you read it in parts?
   I read whole chapters.

f) Do you take notes while reading, and how?
   No, I just highlight the important parts. I look more at concept definitions, than exact numbers, as those do not appear so often in exams.

g) Do you usually need to go back and re-read the same text more than once?
   I reread the highlighted parts.

3. Learning:

   a) Describe the way your study program is set up, in terms of schedule, types of exams etc.
      We have blocks of courses that last four weeks, at the end of which we have an exam. Four times per year, we also have the progress exams (Vooruitgang). The questions are always multiple-choice.
   
   b) Do you study alone or in groups? How big are those groups?
      I prefer to study alone, but I know many people work in groups.
   
   c) Do you usually use question cards?
      No, but I might use them if there was some software to make them quickly.
   
   d) How much time do you need to prepare for an exam? For an assignment?
      I usually reread the literature one week before the exam.
   
   e) Do you use educational software in your study? (e.g. question card programs, quizzes etc.)
      No.
   
   f) How do you stay in touch with your fellow students to discuss assignments, course recommendations, and so on? Do you use social media (e.g. Facebook groups)?
      We discuss via email.

4. Competitions:

   a) Are you aware of any medical (award) competitions? What kind? Have you participated in them?
      Yes, I know of Olympiad-style competitions, where the questions are also multiple-choice, varying across all domains. There are also practical competitions for first aid. I have participated in first aid competions.
   
   b) Which medical areas or topics would you be most comfortable competing in?
      General medicine.
5. Gaming:

a) Do you play video games? Single/multi player? On PC/console/mobile? What kind (e.g. role-playing, social network, puzzle etc.)?

Yes, I play multi player role-playing games on the PC. I also play games on the mobile. I do not play social network games.

b) How much time do you spend playing per day? Per week?

Two to three hours per week, but I used to play more before studying medicine.

6. Participating in crowdsourcing and community activities:

a) Do you contribute to community projects for sharing medical knowledge (e.g. writing Wikipedia articles)? Can you describe the projects, along with the tasks that you perform?

No.

b) Would it be possible to incorporate a crowdsourcing activity (like the one described above) with your course work?

Yes, but it would have to be a fun activity, something that I can play quickly while commuting, or in the evenings, maybe on the mobile or on the tablet.

c) What (medical) topics would be best suited for such an activity?

General medicine, surgery, and others.

d) What kind of tasks would you like to solve? (e.g. reading, testing your knowledge etc.)

A problem solving task, something challenging, but also entertaining.

e) How much time would you spend on this activity per session? How much time per week?

One to two hours per week.

f) Would you prefer to complete the activity alone, or in a group?

Alone, but I would like to compete against my colleagues.

g) What would make such an activity useful to you?

We are lacking problem solving from our classes, we have to do a lot of reading, but we do not apply the knowledge. Applying the knowledge to a practical task would be useful.

7. Further participation:

a) Do you want to receive a transcript of this interview?

No.

b) Do you want to stay updated on the progress of this project?

Yes.

c) Do you want to try out our application once we have an initial version running?

Yes.
B.1.2 Interview Medical Student S2

1. General questions:

a) What is your study course/specialization?

I did a Bachelor in Life Science at the University College in Middelburg, Zeeland. I had a lot of general courses, like biology, pharmacology, most of them medical, but around that as well, like biochemistry. That was three years. Afterwards I went to study medicine here in Maastricht. Normally, medicine is six years, but because I had already done the Bachelor’s in a similar subject I could do medicine in just four years. In a dual Master’s, part of it is medicine, and you do an internship. The first two years are actually very theoretical, and the last two years are practical, you work in a hospital. The second half of the Master’s is research, you learn more about epidemiology, but also how to write a research proposal. Then when I was done, I started doing my PhD in microbiology, in combination with the GGD. I am doing research about STDs, mostly chlamydia, and why some people exhibit the symptoms while some do not. This is a very clinical topic, I like clinical research, but I did not want to study patients.

b) How many years have you spent studying medicine?

I studied for seven years.

c) What kind of courses you have been taking as part of your program (e.g. lecture, laboratory session)?

I did lectures, and laboratory sessions. Here in Maastricht [University] we also have the Problem-Based Learning (PBL) – you have a problem that you discuss in a small group, and everyone contributes what they know about it, then we try to combine all of that into one solution. In Maastricht [University] there is a department that focuses exclusively on how people study, they came up with this [method]. It is also used in the United States, and is spreading, apparently it’s a big hit with people. Basically, it’s about small groups where you discuss a topic, you try and find an answer, and if you don’t know the answer, you make a hypothesis. It is given as homework, you take it home and you divide [the task], and the next week everyone brings back their part of the solution. [The method] applies to many subjects, we also had lectures and some laboratory sessions, but it was minimal work compared to this [PBL].

d) Describe the contents of your program. How does it progress across the study years?

[My PhD] is quite different. I have patient samples. The people at the STD clinics send the samples to this laboratory to have them tested. I’m trying to determine how many bacteria does the sample have, not just whether they are positive or negative, but the actual amount of bacteria. In the lab, I have to do DNA isolation, for that I have to do a reference curve, PCR, stuff like that. Afterwards, after the computer gives me my data, I have to analyze it with statistics [tools], like SPSS. Then I have to write a paper
about it. I have several studies at the same time. For one, I have to quantify
the bacteria, and for another I have to identify the type of bacteria.

2. Reading:

a) What kind of medical texts have you (have to) read? (e.g. textbooks, scientific publications)
I prefer studying from books. Especially during my Bachelor’s, you had
a book, and you had to study [some] chapters for a test. For each course
you had one book, generally, and at the end of the course you’d have read
almost the entire book. It was the same in Maastricht [University], but you
could choose your own books, in Dutch or English, as long as you found
the answer to your hypothesis/questions. During my studies of medicine,
you had a website called UpToDate, sort of a pool of the relevant medical
literature. But it also includes articles, around a diseases or a certain topic.
The textbooks could be a bit old, so you had [access to] new studies. Now
I have to read more scientific publications, e.g. what did other people in
my field find, how can I use it, [how] can I expand it. I still read textbooks
though, mostly for statistics, to look up what test to use for your data.

b) Can you name some (online) sources?
The textbooks [I got at] the bookstore, and also the library here in Maas-
stricht [University]. Sometimes I buy the books online. I don’t really read
e-books. Sometimes I would use the inter-library borrowing system. Like
I said, I used to use UpToDate. Now I get the articles through Google

c) Which of those you consider a “regular” reading material, and which “oc-
casional”? 
Nowadays, it would be [articles from] PubMed, and other scientific publi-
cations. Occasionally [it] would be the books. I have to say, it has changed.
I used to grab my book, and then if I couldn’t find it, I’d look online, but
now I check online first.

d) Why do you read them? (e.g. preparations for exams, assignments, informative, entertainment etc.)
When I was still studying, I was using [the books] to prepare for exams, or
in PBL. [For] entertainment not so much, it was mostly to inform myself
for assignments. [My] studies were quite busy, so I did not have that much
time to read just for fun. Now [I read] to keep up with literature, maybe
think of new research questions.
[In PBL], usually you had a patient. For example, an old lady, 80 years
old, fell down, and you had to come up with what could it be [wrong
with her]. Could it be something neurological, or was it the blood vessels,
or something [to do] with her brain, or something with her glucose, like
diabetes, or something with her hip? Some in the group would know more
about neurology, and some would know more about the heart.

**e) Do you usually read the whole “resource”, or do you read it in parts?**

I select parts. Depending on my question, I scan [the text] to see if [the answer] is somewhere in there. If I know it has to be somewhere in there, but I can’t find it, then I read the whole thing. Generally I try to find specifically what I’m looking for, but if you’re studying for a test, you have to read the whole paragraph.

**f) Do you take notes while reading, and how?**

I take notes. If it’s a book, I have a notebook next to it, and I write down everything that seems important. I copy down the whole sentences, not just keywords, [because] I remember it in the moment, but then when I’m studying for exams, I can’t remember what the rest of it was. When it’s a research article that I’m reading, I usually have it printed, [because] I prefer reading it on paper instead of on the screen. On the sides [of the paper, I write], summaries and keywords for the paragraph.

**g) Do you usually need to go back and re-read the same text more than once?**

For exams, I just reread my notes, because I know they are quite extensive. [During note taking, however] preferably I like to read [the text] once, underline some things, and then write the summary.

**h) Is there a difference between your reading habits now during your PhD, and when you were a student?**

During my Bachelor’s you had to study certain chapters – read the whole chapter, underline it, make a summary, and then study the summary. Now no one is telling me what to read anymore, it’s more about what I find relevant, and what do I want to read more about. Now I have more freedom about what to study, maybe I even read for fun. But I still like to take notes, and then read from my notes.

---

**3. Learning:**

**a) Describe the way your study program is set up, in terms of schedule, types of exams etc.**

It depends on the time [in the program]. You usually had a block. In Maastricht [University], you usually had a test at the end of a six to eight weeks block. Leading up to it, you had different cases – every week you’d have a different patient. By the end you’d need to know all the patients, and all the diseases associated with the patients. My schedule was split between the [PBL] groups and the lectures. Sometimes you’d have a couple of free periods in between, or you’d study at night more. In the week before the test, I’d start reading all that I had written down in the weeks before that. The exams were usually written. I might have had one or two oral exams though.

Now I’m usually in the lab doing things. Today, I have a day to analyze all my data, look up where I am, what I still need to do, and make up a planning. Sometimes there are a lot of things [to do] in the lab, and sometimes there is a lot [of data] to analyze.
b) Do you study alone or in groups? How big are those groups?

I prefer to study alone. With the PBL though, you’d usually be paired with two other people, and you had to work together to find the answers to your problem. There were usually not big groups, I think the biggest group I ever worked with was three people. We would divide stuff up, and then at the end we would bring it together and make one big document out of [our solution].

Now, I have my supervisor, and my supervisor in the lab, but I am the only one doing research on chlamydia, so I am by myself. We have other people doing other STDs, but most of it is done individually. We have weekly discussions though.

c) Do you usually use question cards?

No, I did not use them. I had a house mate who used them, she was from America, maybe that makes a difference.

d) How much time do you need to prepare for an exam? For an assignment?

In Maastricht [University], you usually had a test at the end of a six to eight weeks block. Leading up to it, you had different cases – every week you’d have a different patient. By the end you’d need to know all the patients, and all the diseases associated with the patients. My schedule was split between the [PBL] groups and the lectures. Sometimes you’d have a couple of free periods in between, or you’d study at night more. In the week before the test, I’d start reading all that I had written down in the weeks before that. The exams were usually written. I might have had one or two oral exams though.

e) Do you use educational software in your study? (e.g. question card programs, quizzes etc.)

I did some quizzes, not many of them though. It was usually for anatomy, because that [subject] was always quite hard, but you could then have a fun game where you would get bonus points for every part of the body that you would guess correctly. They had different ones online, one for the bones, one for the muscles. It was a lot of fun, more fun than studying from the book. I suppose you could try [for other subjects], but I haven’t seen [other examples].

f) How do you stay in touch with your fellow students to discuss assignments, course recommendations, and so on? Do you use social media (e.g. Facebook groups)?

With classmates, we would email each other back and forth. We have a Facebook group now, “find my article”, where you cannot find/open an article, you ask someone to look it up for you. That way you can access journals you would normally not have access to.

4. Competitions:

a) Are you aware of any medical (award) competitions? What kind? Have you participated in them?
During my MSc, we had [a competition]. It started in Germany, between various medical schools, they would have people in the same year, before they started their residencies, in the competition. They copied it into a Dutch version as well. I tried to be in the committee to organize, but it took too long to organize, I had finished my studies in the meantime. You have questions about everything, lab results, diagnosing people, a picture of the disease that you need to name (e.g. a skin disease), pathology (seeing some cells and recognizing their type), practical as well (e.g. how to pipe, how to bandage).

Here in the lab we also have a competition, [organized] by pharmaceutical companies. They say a new mix, for example, and they want us to compare their mix to the one we already use, take a picture of the results, and they pick a winner.

b) Which medical areas or topics would you be most comfortable competing in?

Not microbiology, STDs maybe. Probably a general domain. If you would ask my friends that each specialize in something, like radiology, gynecology, they would probably prefer their field, but I prefer general medicine.

5. Gaming:

a) Do you play video games? Single/multi player? On PC/console/mobile? What kind (e.g. role-playing, social network, puzzle etc.)?

I used to play old-school games, like Mario, on the Nintendo, or GameBoy. I have an old-school Nokia, so I play snake, but nothing more advanced. I know the medical information apps, like for blood pressure. I don’t play Facebook games.

b) How much time do you spend playing per day? Per week?

I play very very little.

6. Participating in crowdsourcing and community activities:

a) Do you contribute to community projects for sharing medical knowledge (e.g. writing Wikipedia articles)? Can you describe the projects, along with the tasks that you perform?

Not really.

b) Would it be possible to incorporate a crowdsourcing activity (like the one described above) with your course work?

Certain colleagues of mine would consider House [the TV show] studying, because you had a patient with certain symptoms, and you had to guess the disease. It’s usually something exotic, that you do not usually think about. On House, they have very difficult kind of diseases, with very rare complications, so if you knew the diseases, or the complications, you’d be very advanced with your studies.
c) **What (medical) topics would be best suited for such an activity?**

General medicine, or GP (e.g. people come with a symptom, and you have to give an opinion). Maybe you can think of things in gynecology as well. But in microbiology, it’s such a visual domain, you have to see the bacteria, it should be very hard to make something out of words with that. The same with pathology and radiology I think.

d) **What kind of tasks would you like to solve? (e.g. reading, testing your knowledge etc.)**

Diagnosing things mostly. In microbiology you could do it backwards. Say you have [a patient with] pneuomonia, you have to recognize what bacteria are causing it. Testing your knowledge, by making it into a game.

e) **How much time would you spend on this activity per session? How much time per week?**

I would play mostly for fun. I’d like an easy fun way to learn about new fields (e.g. bacteria, microbiology). I would play one hour or so per week.

f) **Would you prefer to complete the activity alone, or in a group?**

Both. [I would play] mostly by myself if there was something to study. But if there was a competition like guess the disease, or guess the bacteria, or medical detective, it might be fun to do it in a group of friends to see if we can outdo each other.

g) **What would make such an activity useful to you?**

It depends on what you develop. If you make a game out of it, it will be a fun way to keep up with your medical knowledge. If you do something like with the House [TV show], even if it’s a topic that is not relevant [to me] now, I would use it to keep up with [the field]. In medicine, you study everything, and now, because I’m just in microbiology, everything else just fades away, so [this] might be a fun way to keep up with other fields as well. Also, I only know a bit about microbiology. This might be an easy and accessible way to learn more about the rest of microbiology.

7. **Further participation:**

a) **Do you want to receive a transcript of this interview?**

You can send it to me, but it’s not necessary.

b) **Do you want to stay updated on the progress of this project?**

Sure.

c) **Do you want to try out our application once we have an initial version running?**

Sure.

**B.1.3 Interview Medical Student S3**

1. **General questions:**
a) **What is your study course/specialization?**

I am now becoming a specialist in microbiology, doing a residency for the next several years. I am not a student anymore, but I do take courses, symposiums and workshops. We have an annual exam. I have just started, so I do not know how the exam looks like.

b) **How many years have you spent studying medicine?**

In total, nine years. First, I started with clinical molecular science. Then, I did a MSc in clinical molecular science, then I started my real medicine studies, together with clinical research.

c) **What kind of courses you have been taking as part of your program (e.g. lecture, laboratory session)?**

We have 30 percent lectures, 30 percent student meetings, and tutor classes where we do Problem-Based Learning (PBL), and one third of the classes were laboratory practice. When I studied molecular science it was more laboratory work, but during my medicine traineeship it was more general skill. We used phantom bodies, or we practiced on each other to learn how to do clinical examinations. We have another small part of the medical traineeship, like a simulation of patient consults.

d) **Describe the contents of your program. How does it progress across the study years?**

In the first year [of the medicine MSc] we have mostly theoretical classes, and PBL. We gather around and we discuss about a problem, then we need to go to the lab review, discuss everything and write two exercises, come back together and resolve the problem. In the second year you have more lessons with applications to the clinic, more about diagnosing, how do you recognize different diseases and how do you treat them. Then you get once a week a practical in the clinic. Every time you move little towards the clinic. Every year you have exams, theoretical and practical exams as well. The third year is a phase of internships (for us [clinical molecular scientists] it’s one and a half years, for regular medicine students it’s two years) that consist of a small theoretical background, and then you go for example, into medical medicine, in a clinic. Most of the time you end with an exam. Sometime it’s on a computer, sometime it’s written. The last phase is a little bit less than a year, you do a real large internship. The part you want to specialize in, you go [and practice] in a clinic. During all the medicine studies, four times in a year, you have a general examination, the Vooruitgang, to see what your progress is. That’s usually on paper.

2. **Reading:**

a) **What kind of medical texts have you (have to) read? (e.g. textbooks, scientific publications)**

Both [textbooks and medical publications]. We use a lot of basic stuff from textbooks. We can use them from the library, and sometimes we use them online. Also, we use publications for projects.
b) Can you name some (online) sources?
We can borrow [textbooks] from the library. Usually we have some really basic textbooks which we use a lot, so maybe you buy two [of these] textbooks online, but all the other stuff we borrow from here. You can do two things: you can borrow them and take them home with you, or you can just use them in the library. For every book, we have ten [copies] from the same one, so you can just go to the library, take a book from the shelf, and then when you’re done, you put it back. [For publications] mostly PubMed. We can use it via the library as well.

c) Which of those you consider a “regular” reading material, and which “occasional”? 50-50. The more you become practical, and experience, the more you use publications, and less textbooks.

d) Why do you read them? (e.g. preparations for exams, assignments, informative, entertainment etc.)
Most of the time, I use them because I have to obtain the knowledge. When I have a direct problem, I look it up in a textbook. That’s what PBL is based on as well. You have a starter problem, and then from there you try to answer the questions, so you look it up in the textbook, and publications. You have to obtain the basic information [from the textbook]. If someone comes with hypertension, you have to understand how blood pressure functions, and that you look up in a textbook. Sometime I also [read] just for fun, because I like to read.
[I don’t read for exams] because you already have to prepare for the PBLs using the textbooks, and then you make notes for yourself. It’s a [summary] from what you have read, so you make notes/schemes, and that is what I used to study for exams. Sometimes, during the internships, if you have the computer exams, we train ourselves by reading a lot of exam questions.

e) Do you usually read the whole “resource”, or do you read it in parts?
Usually, I start from the problem, so I look up where the textbook says something about that problem, and then look at the chapter, which sections are important, which less important, and I start from the important sections. Most of the times I start very enthusiastic, and at the end I don’t have much time left, so usually the first part I have read more profoundly than the last part.

f) Do you take notes while reading, and how?
Usually, when I have a publication I start with a marker, outlining the most important parts of sentences. But I prefer to make schemes of papers. [The concepts] depend on the learning problems we deducted from the issue. From a text, we have some learning objectives, that would be the parts that I line out.

g) Do you usually need to go back and re-read the same text more than once?
I try to line out the parts that, if I take the text again, it would be enough to just read the lined up parts.
3. Learning:

a) Describe the way your study program is set up, in terms of schedule, types of exams etc.

We have a Blackboard system we use a lot to look up our schedules. Four times a year we have the [progress] exam. Usually we would start with a lecture, in the beginning of the week, and then two times a week we have the PBL meetings, so one day you prepare [the problem], and the other day you discuss what you have studied, and you begin with new PBL objectives. Meanwhile, you also have practical skills [courses], like laboratories. Lectures are also in between. In Holland, we don’t study for example physiology and anatomy, it’s all integrated in a block, like for example “infectious diseases”, there you have part physiology, part anatomy, part molecular biology. And then at the end of the module you have an exam, or sometime in between [the block], when it’s very large. [The module] is between four and ten weeks. Most of them are eight weeks.

b) Do you study alone or in groups? How big are those groups?

Both. It depends at the state I’m in during the module. In the beginning, I study sometimes alone, but sometimes you have to work with (small) groups, then you have to go together at the library and you can discuss things. When the exam is coming, we come together with a bigger group to discuss the subjects, and then we can help each other with the difficult stuff, to check if the depth of your information is enough or not. A medium group would be six persons, sometimes with four, sometimes with ten maximum.

c) Do you usually use question cards?

No. But that’s what we basically do for PBL, where afterwards you have to answer the questions from the information you got. We have a tutor who observes the whole group, how they are progressing. The tutor sometimes ask a question to see whether all the people understand the material.

d) How much time do you need to prepare for an exam? For an assignment?

For me, it’s not enough to have the information/knowledge [from the PBLs] to take the exams. For example, for the progress exams, there I don’t study, because it’s just to measure your progress, and it should be on the current knowledge you have. But the other exams, the week before the exam will be [1 study] 2 hours a day, for 4 days. [For the PBLs] we usually have more or less 3 days to prepare, and it takes about 8 hours of preparation [per PBL], sometimes less. Sometimes we publish [the solutions to the learning objectives] on Blackboard. In between you have assignments as well, for example, you know you have different kinds of ways of doing research, randomized controlled trials, clinical trials, and we have to assemble a trial for ourselves. We have a research question, and then we have to make an assignment on how to set the research up. Also, writing papers.

e) Do you use educational software in your study? (e.g. question card programs, quizzes etc.)

66
I use Excel and SPSS, to do my calculations. [I use] EPAS, it’s like an electronic portfolio that we use to demonstrate our progression in studies, and there we usually upload our results from exams, and papers. I know that they are doing some research on how they can use the EPAS system to help people with learning.

f) How do you stay in touch with your fellow students to discuss assignments, course recommendations, and so on? Do you use social media (e.g. Facebook groups)?
When I was studying, it was mostly by text messages. Nowadays it’s Watsapp. Not a lot of Facebook.

4. Competitions:

a) Are you aware of any medical (award) competitions? What kind? Have you participated in them?
There are some. I have not [participated]. They try to make questions about the practical stuff. For example, you have case of a patient, and you try to resolve it. They have a case, and then they ask different question, about what’s your differential diagnosis, how would you like to diagnose it, so what tests would you like to use. Sometimes they use pictures, so they ask you what you think you see here.

b) Which medical areas or topics would you be most comfortable competing in?
For now, I would say general medicine, because I just started with microbiology, so that’s for me the most important to compete in, but I’m not feeling elaborate enough yet to compete with other people who are more experienced with the field.

5. Gaming:

a) Do you play video games? Single/multi player? On PC/console/mobile? What kind (e.g. role-playing, social network, puzzle etc.)?
Not a lot. I have a few games that I play, like SingStar. I like to sometimes raise a city, something like Sims. I’m not a regular game, I think I’m not really representative. I did one game on Facebook, Mafia Wars.

b) How much time do you spend playing per day? Per week?
Between half an hour and three hours. I do a lot of board games/card games though.

6. Participating in crowdsourcing and community activities:

a) Do you contribute to community projects for sharing medical knowledge (e.g. writing Wikipedia articles)? Can you describe the projects, along with the tasks that you perform?
Not at the moment.
b) Would it be possible to incorporate a crowdsourcing activity (like the one described above) with your course work?
   I think it should be possible. But I would use it most of the time at home. It would be via Internet, it would be perfect.

c) What (medical) topics would be best suited for such an activity?
   I think a lot of domains. I think in medicine it would be very applicable. For microbiology, I think I could use it, if you have like a problem, a case study, and ask different questions about it to train, and I could get some feedback, I think it could be very useful.

d) What kind of tasks would you like to solve? (e.g. reading, testing your knowledge etc.)
   For medical students, the case problem solving would be the most interesting, because then you have some practical link to the real world.

e) How much time would you spend on this activity per session? How much time per week?
   Maybe one or two hours a week. In the evenings, because during the day you don’t have much time for that. [A mobile app] would be very handy, because it does not matter where you are, you can use it.

f) Would you prefer to complete the activity alone, or in a group?
   The [activity] I was imagining now was alone, but I could imagine you could do it in a group as well. One time, one of the residents sent [via email] a picture [of a disease] and asked “what is this”, and then we could all talk to each other about what it could be.

g) What would make such an activity useful to you?
   I think it has to have a connection to our line of work. Now I’m doing some bacteriology stuff, so it would be nice if I [could use the app for] bacteriology, so I can use it in my daily work.

7. Further participation:

   a) Do you want to receive a transcript of this interview?
      No.

   b) Do you want to stay updated on the progress of this project?
      No.

   c) Do you want to try out our application once we have an initial version running?
      Yes.
I did a medical degree in Maastricht [University]. I haven’t specialized in a medical domain, but I did study specific fields. I did general surgery for one year and a half, and half a year of EMT. Now I’m doing my PhD here, [at the AMC], non-clinical work. [My specialization now] is epidemiology-nephrology, but it’s non-clinical, so no patient contact.

b) How many years have you spent studying medicine?

Six years is the basic medical training. The PhD is officially also a training. I started about a year ago. The two years of clinical work were not official training, but, of course, they involved training. So I studied medicine for seven to nine years, depending on how you count.

c) What kind of courses have you been taking as part of your program (e.g. lecture, laboratory session)?

I’ve had lectures. We have a lot of practical things in [Maastricht]. We’ve practiced with LOTUS patients, so you would have practical examinations. We had labs, and some main courses, workshops.

d) Describe the contents of your program. How does it progress across the study years?

In Maastricht [University], it’s very mixed. It’s really an incentive for medical students to get them interested. You start out quite practical, with a patient case, and from there you go on, and see what the solution might be. So you start with the patients, then you go into the literature, and you do some practical courses on how to fix it.

Now, there’s a European Project I’m working on. You get a bunch of money for certain deliverables you have to finish. That’s structured. Now I’m doing a comparison study, so contacting lots of people, getting data, and analyzing that data.

2. Reading:

a) What kind of medical texts have you (have to) read? (e.g. textbooks, scientific publications)

I read articles, papers, but also guidelines. [During my studies] we had tests, of course. A lot of things is evidence based, [therefore I use] a lot of PubMed, and systematic reviews. During my studies, [I used textbooks].

b) Can you name some (online) sources?

PubMed, MEDLINE, Cochrane. Textbooks are very diverse, depending on the topic you’re doing. You have the general internal medicine books, for example, general surgery. Then you have the more specific pathology books.

c) Which of those you consider a “regular” reading material, and which “occasional”?  

http://www.vandoesburgbhv.nl/ehbo/lotus  
http://www.cochrane.org/
PubMed and MEDLINE are daily. It’s very rarely that I look something up in a textbook, maybe once a year.

d) Why do you read them? (e.g. preparations for exams, assignments, informative, entertainment etc.)

[I read] to gather information for solving a problem. Entertainment is also combined, you study because you like the subject. During my studies, you had the assignments, and then you had to go find the appropriate reading material. For exams, they didn’t say something like “these are the books you have to learn”. You could choose yourself where you got the information, but you had to discuss it in the weekly groups, so you would be guided in the right direction. I think you’re much more focused reading if you have a question in mind, than if you just read and think like “oh, I have to know all of this information”.

Reading habits are very related to the reason for reading. If someone needs feedback on a paper, then there’s a lot of notes, but if I need to quickly find an answer to a question, then there’s no notes, just scanning.

e) Do you usually read the whole “resource”, or do you read it in parts?

With papers, I usually read the abstract first, to see if it’s something that I want to read in the first place. If it is, then, depending on what I’m looking for, I usually look for the results and the methods first, to see how did they do it. [if I] can trust this, and if I can trust it, then I go look for the results. If you want more basic information, the introduction is quite good, because you have the references to the basic, already known information. [For textbooks], you look for a specific solution, you’re scanning for the answer.

f) Do you take notes while reading, and how?

I underline, or highlight, mostly words or parts of sentences, sometimes sections, if they’re important. Now I have a new app [on my tablet] where I can highlight [what I read]. [I prefer using the app to printing], but that’s more environmentally based. Sometimes I still print out, if I have to take a lot of notes. But that’s usually if I’m reviewing something for a person, that’s different from when I’m reading for my own information. For traveling, you also don’t have to take a whole stack of papers with you [with a tablet], it’s ideal.

g) Do you usually need to go back and re-read the same text more than once?

That depends on my focus at the time. I don’t really take notes, so when I would study for exams, I would go back to the highlighted parts.

3. Learning:

a) Describe the way your study program is set up, in terms of schedule, types of exams etc.

Now, there are no exams, as far as I know. There are some courses which you can voluntarily take. Back when I was doing my [MSc] studies, they had blocks of things. The first one was for ER things, people that could
present to the ER for something. We had six weeks of different cases that you could see, [and for which] you had to find a solution. At the end of the six weeks, we had an exam, usually multiple choice. That’s alternated with a practical exam once a year – you have four cases with four different examiners, and you have to talk and interact with the patient and examiner to find the right information. Four times a year, we had the “Vooruitgang”, the progress exam, which is quite interesting. We had 200 questions, and everyone gets the same exam (if you’re in your sixth year, or your first year, you get the same exam), but depending on how far you are in your studies, [there is a] percentage [of questions] you have to give the right answer to. So you skip the questions if you don’t know the answer to, and if you give a wrong answer, then you get minus points. You have to skip answers you don’t know, or else you lose points. It’s interesting, one of my friends, she’s a brilliant doctor, but this is the one exam she could never do, because she was not able in this [type of] decision making.

b) Do you study alone or in groups? How big are those groups?

We did these groups with two weekly meetings, where we would discuss things. Normally you would look up stuff yourself, or sometimes with a friend, and then [discuss it] in the group. For the exams, we would often study together, we would exam each other. If you have to test yourself, it’s not easy to do it alone, you need a study buddy. [The groups were] three to four persons, not more.

c) Do you usually use question cards?

No. Sometimes you had study tests. For the four times a year test, you had the exams from the previous years, which you could use to prepare.

d) How much time do you need to prepare for an exam? For an assignment?

For the four times a year tests, I did not prepare for. That’s something you cannot prepare for, because it’s such a broad subject, it’s something that you have to learn through your studies. The other exams, it’s a gradual process. In all the six weeks, you study, and then [before the exam] a couple of hours, to re-read certain things. For assignments, it was different per project. You had CATs (critical appraisal of a subject), which would take longer, because you first had to find your article. [On average], it would be a couple of days to a couple of hours, depending on the assignment.

e) Do you use educational software in your study? (e.g. question card programs, quizzes etc.)

I’ve used Blackboard. You also have the resident bulletins from these journals – you read a thing and they have questions. I think it’s in the New England Journal of Medicine. There’s also an anatomy app – they say “pancreas”, and you have to click quickly where it is, then it gets more detailed. I quite like that as well. I would strongly encourage you [to make a mobile app]. You have a Medscape app, and a medical calculator app, so I think it would fit nicely in there.

f) How do you stay in touch with your fellow students to discuss assignments, course recommendations, and so on? Do you use social media (e.g. Facebook groups)?

We had Blackboard at Maastricht [University], and email. We did some MSN messenger. [Here at the AMC], it’s mostly email. We also have a Facebook group, but that’s mostly for entertainment.

4. Competitions:

a) Are you aware of any medical (award) competitions? What kind? Have you participated in them?

They have a quiz night in the Irish pub, with a medical part. There were quiz nights organized by the student associations. I have participated in one quiz. [They ask] questions like how many bones there are in the human body, or what is the most common blood type.

b) Which medical areas or topics would you be most comfortable competing in?

General knowledge is quite interesting.

5. Gaming:

a) Do you play video games? Single/multi player? On PC/console/mobile? What kind (e.g. role-playing, social network, puzzle etc.)?

Yes, I do play, more iPad games. It’s very simple games like Bubble Exploder, puzzles, or card games. Mindless swiping is quite nice for relaxation. [On social networks] we did this massive music game, where you play music, and you have to recognize the song title. It was extremely addictive, because you had to be really fast.

b) How much time do you spend playing per day? Per week?

Half an hour per day. It’s not every day that I play, but in the weekend I play a bit more.

6. Participating in crowdsourcing and community activities:

a) Do you contribute to community projects for sharing medical knowledge (e.g. writing Wikipedia articles)? Can you describe the projects, along with the tasks that you perform?

No.

b) Would it be possible to incorporate a crowdsourcing activity (like the one described above) with your course work?

Yes.

c) What (medical) topics would be best suited for such an activity?

It depends if you’re looking at the students’ perspective, or if you’re looking at the professionals’ perspective, or if you’re looking at the PhD students’ perspective. If you want the broadest [participants], then diverse
general topics are going to attract the most people, but all the specialized people might be less interested. [Radiology and anatomy] are much more about images, but if you can combine [images and text] – if you have an image, you need a text, you need the patient’s characteristics and symptoms – [it could work]. Maybe pathology [would be] even a bit more [suitable].

d) **What kind of tasks would you like to solve? (e.g. reading, testing your knowledge etc.)**

Patient tasks, so if you have a case. Not too long [of a text to read], people get bored easily, especially if it’s an online thing, just a paragraph, and ten sentences maximum, and a couple of questions, or one paragraph and one question.

e) **How much time would you spend on this activity per session? How much time per week?**

If it’s something online that you can do in between [work], then maybe a couple of minutes [per session]. But if it’s something very addictive, then of course people will sit there for even hours maybe.

f) **Would you prefer to complete the activity alone, or in a group?**

Most people would like to see a ranking at least. You don’t necessarily have to play against someone specific, but you have to have some kind of motivation to do better, so a ranking would help. But direct interaction is also fun.

g) **What would make such an activity useful to you?**

If I could learn something from it. If you have a fun way to broaden my knowledge, that would be ideal. The app way is a good way to go, in my opinion.

7. **Further participation:**

   a) **Do you want to receive a transcript of this interview?**

      Yes.

   b) **Do you want to stay updated on the progress of this project?**

      Yes.

   c) **Do you want to try out our application once we have an initial version running?**

      Yes.

**B.1.5 Interview Medical Student S5**

1. **General questions:**

   a) **What is your study course/specialization?**

      What I’m doing right now is a PhD in medical informatics, with a basis in decision support. What I have done previously is veterinary medicine. I was a practicing veterinarian for six years, so my background is medical, but what I have been doing right now is largely technical.
b) **How many years have you spent studying medicine?**

The veterinary medicine degree is a four year program, and then I have done a two year MSc in medical informatics, and I’m now in year four of my PhD.

c) **What kind of courses you have been taking as part of your program (e.g. lecture, laboratory session)?**

I’ve had general medical courses, physiology and so on, as well as technical courses, as a part of the medical informatics program, as well as organizational courses, as part of medical informatics.

d) **Describe the contents of your program. How does it progress across the study years?**

I did two years of undergraduate [studies] in agriculture, which is a common background for veterinarians. I did not finish the undergraduate degree, because I got accepted into the veterinary school, so I did four years of veterinary school. I then practiced for six years, moved to the Netherlands, spent two years in the medical informatics MSc program, and then four years in the PhD program.

2. **Reading:**

   a) **What kind of medical texts have you (have to) read? (e.g. textbooks, scientific publications)**

At this point, I mostly read papers, although we do occasionally refer to textbooks for statistics, and those sorts of things. What I have read, is almost everything, obviously general biology texts, as well as specific things, like pathology diagnostics.

   b) **Can you name some (online) sources?**

I usually start with PubMed, it’s [an] easier interface. If PubMed does not do it, then I’ll use AMED, but the interface on AMED is awful, so I’ll avoid it if I can. For textbooks, it’s mostly asking around. I ask my colleagues what a good textbook on the subject is, and if that fails, I go search on the Internet for it, but usually they know.

   c) **Which of those you consider a “regular” reading material, and which “occasional”?**

I wouldn’t say there’s any regular reading material at this point, I keep meaning to read the articles that are in the main publications in my field, but the truth is that I tend to read them mainly when I hear about an article, rather than actually reading [a journal] regularly.

   d) **Why do you read them? (e.g. preparations for exams, assignments, informative, entertainment etc.)**

Mainly as background material for my research. Anytime you’re starting with a paper, the first thing you need to do is find out anything you can about the subject, which obviously involves a lot of reading. Because our work is so diverse, it tends to be a new batch of articles every time.
During the MSc, you should be aware that many times students don’t read the assigned text [for exams, assignments]. There’s often text assigned in preparation for the class, and if there was an additional motivation for reading it, people might [do it]. [We often had] class discussions, projects in small groups as well [based on the reading]. One thing we do commonly is read published articles critically for various reasons, like as background material for our project, or specifically to critique the article and say what they did right, what they did wrong, that’s pretty common in medical informatics. Sometimes, there’s a fair bit from textbooks, but I think that pertains mostly to the technical side, rather than the medical side. In the medical informatics BSc, I know that they have more medical text that they need to read, but I think those are actually in Dutch. As a veterinary medical student, there are vasts amounts of text that you need to read, so one of the things that you need to do is get good at reading them quickly and efficiently. Something that helps with that could be very helpful for either medical students, or veterinary medicine students.

e) Do you usually read the whole “resource”, or do you read it in parts?

It depends on what I’m reading for. Usually, if I’m reading an article, I will read the abstract, read the last little bit of the introduction, which usually tells you why they’re doing the research, skip to the methods and results, and read the first part of the discussion, which usually tells you the main findings and conclusions. So when I’m reading an article, I tend to skip around a bit, and then if it’s something I’m particularly interested in, I’ll read the whole thing.

f) Do you take notes while reading, and how?

It depends. If I need to read it in depth, then I do [take notes], if I’m just trying to get the gist of it, then I do not.

If I’m reading an article, customarily what I’m looking for is what they found, and if their methods justify their conclusions. So the notes that I’m taking [in this case] are specifically to understand what they did, in what order, by whom, as well as what was their objective, and what did they conclude. I’m looking for something specific in that case. If I’m reading more generally, especially if it’s a difficult text, I’ll just summarize paragraph by paragraph. After each paragraph, if there’s something important there, I write it down, if not, I just skip to the next paragraph. I do the same with textbooks. [For exams], I would usually go for the “paragraph by paragraph” approach, and then study my notes from there. I would not go back to the original text, unless at the end I felt that I understood the topic differently, then I would have to go back and do it again.

g) Do you usually need to go back and re-read the same text more than once?

It depends on the purpose. If I’m reviewing the article, which happens a lot as a researcher, then I definitely read it more than once. In almost all other circumstances, I read it once.

3. Learning:
a) Describe the way your study program is set up, in terms of schedule, types of exams etc.

It’s highly variable. Unlike many PhDs in medical informatics, you’re typically working with a project that’s not really set up ahead of time, involves a lot of project management, a lot of meeting with people, a lot of organization. It’s not just measuring blood pressure in rats or something, it tends to be complicated and involve a lot of humans. I don’t think there is such a thing as a typical medical informatics PhD. What I would say is generalizable is that almost all of us do a lot of reading of articles, we do some reviewing of articles. The typical medical informatics PhD has a technical background, so they rely on medical professionals for a lot of the medical information. In my case, I’m the other way around, I have a medical background, so I rely on technical people for the technical stuff.

[During my MSc] we had classes four days a week in most weeks, in the mornings. They were setup mostly in the lecture format. In the afternoons, sometimes we would have labs, projects in small groups. Almost every project was in a small group.

They tried to give us practical assignments in most cases. In the BSc, one of the assignments is to write a small computer program. In the MSc program, we typically had things like [to] imagine that the AMC is going to implement a new patient record system, [and] how would you go organizing that implementation, what meetings would you call, from what stakeholders, and so on. You would have to write out a project plan with your team. We did a lot of PowerPoint presentations on those assignments.

b) Do you study alone or in groups? How big are those groups?

I did [study in groups] in veterinary school, but that was not officially organized. We just did that as a survival strategy. In the medical informatics program, studying was not typically done in groups. The labs and the projects, were done in small groups typically three to four people.

c) Do you usually use question cards?

No.

d) How much time do you need to prepare for an exam? For an assignment?

There’s no standard scope for an exam. Sometimes an exam would be a few questions in five minutes, sometimes it would be for the final exam for the module six to eight weeks of work. Part of the work that I did was a pre-MSc program, which was more independent, and was exam-based. Basically, for each subject, you would be given some material to study, and there would be an exam. I would typically spend three weeks for each topic. That might be a more useful metric.

Again, the scope of the assignments varied a lot. Most of the assignments we were given could be completed in one evening. For the group assignments, we would meet two or three times to work on it for an afternoon.

e) Do you use educational software in your study? (e.g. question card programs, quizzes etc.)
I’ve tried to build educational software. We used Blackboard. I was involved in doing some testing for veterinary software back in the day. There is some interest in our department for doing more with online learning. There’s actually a project being developed right now to develop online learning for medical students that are interested in bio-informatics. I think that doctors too tend to like e-learning, it’s a little more fun. The downside is that the [e-learning software] is a bit more difficult to take on a train, and a lot of the continuing education happens on the train. That can be done with an app, most of the doctors have iPhones.

f) How do you stay in touch with your fellow students to discuss assignments, course recommendations, and so on? Do you use social media (e.g. Facebook groups)?

Mostly email, we have a mailing list and we CC each other. People tend to use Facebook for personal stuff, and email for professional stuff.

4. Competitions:

a) Are you aware of any medical (award) competitions? What kind? Have you participated in them?

I know the PhD association here does quiz nights, but they’re typically in Dutch, so I have not participated.

b) Which medical areas or topics would you be most comfortable competing in?

I mostly wouldn’t [be interested].

5. Gaming:

a) Do you play video games? Single/multi player? On PC/console/mobile? What kind (e.g. role-playing, social network, puzzle etc.)?

I tend to like turn-based strategy games. I play a lot of NetHack. I play Angry Birds, it’s just on the train [though], I don’t play it any other time, [on the train] it’s just handy because it’s on my phone. I don’t play multi-player game, although if there were a multi-player non real-time game, I would probably play it. But we’ve talked a lot about how you can make that work, and it’s difficult, because if someone stops playing for two weeks, [you’re stuck]. The platform is mainly on my phone these days, although when I play NetHack, I don’t play it on my phone [because I need] a keyboard. I also do some non-video role-playing games, and some board games as well. I did used to [play social network games] a little bit, but mostly back when Facebook first was a thing, I did some Facebook games, but it’s taking up too much time. Most of the games were just awful, and they had absolutely no value, other than their addictive value and the fact that you were pressured into playing because your friends benefited when you played. I decided I was tired of that, it would take a little bit extra to expect me to play because of that.
b) How much time do you spend playing per day? Per week?
   [I play] mostly on the train, so mostly about twenty or thirty minutes.

6. Participating in crowdsourcing and community activities:

   a) Do you contribute to community projects for sharing medical knowledge (e.g. writing Wikipedia articles)? Can you describe the projects, along with the tasks that you perform?
   When I find an error in Wikipedia I will fix it, or revert vandalism. There’s a couple of topics that I have actively edited. Mainly I stick to things that I know about, that I think that other people probably don’t know much about. I don’t try and edit the big topics, I just stick to the little corners.

   b) Would it be possible to incorporate a crowdsourcing activity (like the one described above) with your course work?
   The question is one of time. The ideal way to do it would be for it to have an additional benefit to something that you need to do anyway. Ideally, it would make it easier to do something that you have to do anyway. For example, as a student, you have to read a lot of fairly boring text, so having something that makes that less awful is beneficial. I don’t really have any specific idea on how to do that, but that sort of thing would be useful. The same goes for continuing education as a doctor. Often [continuing education] is pretty dry. Of course, you can attend lectures and conferences, but that takes time, it’s expensive, so having ways that do that without all the extra effort would be good. Doctors tend to be interested in things that keep them up to date with the latest research findings, so that could be helpful as well.

   I think, [a mobile app to play while commuting] would be one of the things that would make me more likely to do it. There’s a limited number of things you can do on a train, and my time when I’m not on the train tends to have a lot of demands. Having it be available on mobile platforms would be a big motivator for busy people. Students would be a bit of a different population, in that they have to spend some time reading and studying anyway, so something that integrates with that could be useful even though it’s not on a mobile platform.

   c) What (medical) topics would be best suited for such an activity?
   It would be possible to do almost any [domain]. What [domain] would be interesting is a little bit of a different question. What would be useful for Watson is also a different question. A lot of what I read is more technical than medical, truth be told, and that has to do with the background of the people that write it, they tend to be more technical than medical.

   d) What kind of tasks would you like to solve? (e.g. reading, testing your knowledge etc.)
   I need to read and summarize a fair bit of articles, and retain the basic gist of what is in the article in many cases, so something that helps remember what’s going on [in the article] would be very useful. As a student, when
you're studying something, [an application that] helps you remember the main points of the text would be very helpful.

When I’m playing on the train, [the game] should not be a difficult strategy game. It’s first thing in the morning, or the last thing in the day, and either way my brain is not at 100%. So [the game] should be relatively easy, that nice balance of challenging enough to be interesting, but not too complicated.

There’s a feature like [playing Doctor House game] in most of the big medical journals, “Guess the Diagnosis” feature. It’s kind of interesting, but as a repetitive task, this is what you have to do as a doctor, so I’m not sure if that would work. But something along those lines, or an abstraction of it, maybe playing something like a clue game, not necessarily a realistic depiction of the diagnosing process, could be interesting. But medical students might find it very interesting to do a diagnosing game.

e) How much time would you spend on this activity per session? How much time per week?

Twenty minutes on the train, and twenty minutes before bed. But if it were integrated with other tasks that I am doing already, then I might spend more time.

f) Would you prefer to complete the activity alone, or in a group?

Alone, but the social games do have the added motivation. There’s some benefit to both.

g) What would make such an activity useful to you?

If it is fun, if it helps me keep up to date with what is going on in my field, if it notifies me of the interesting articles in my field. As a student, if it helps me learn the material for exams, and as a medical student, if it could make me a better, more comfortable diagnostician in the end. Knowing whether or not you have the right diagnosis, or that you thought of everything, is difficult as a new doctor, and it’s not something you’re comfortable with for a long time. As an experienced practitioner, keeping up to date with the advances of the field [would be useful]. Once you’re an established professional, the main thing that concerns you is keeping up as things change.

7. Further participation:

a) Do you want to receive a transcript of this interview?

Yes.

b) Do you want to stay updated on the progress of this project?

Yes.

c) Do you want to try out our application once we have an initial version running?

Yes.
B.2 Medical Professionals

B.2.1 Interview Medical Professional P1

1. General questions:

   a) What is your study course/specialization?
      I have studied at the AMC, where you have to do four years at the university, and two years of internships. I have been studying for seven years, because I have also extended my course, I have done research abroad. And then, after that, I immediately started with my PhD project. That was in 2009, so I’m almost finished with my PhD. Then, after this, I will do a specialization for internal medicine. My focus of my research is metabolic diseases, mainly diabetes.

   b) How many years have you spent studying medicine?
      Seven years studying, and four years in my PhD.

   c) What kind of courses you have been taking as part of your program (e.g. lecture, laboratory session)?
      Lectures, smaller sessions where we would discuss medical cases with eight students and one doctor. Furthermore, we also have anatomy lessons, and microscopy lessons, where you learn how to interpret pathology.

   d) Describe the contents of your program. How does it progress across the study years?
      The program has changed a lot since I did the course. When I did the medical course, we still had the doctoral system, so not the BSc-MSc system. In the first year, we had some basic lessons, more the fundamental, biological part of medicine, physics, how the body works in a chemical way. From the second to the fourth year, the whole body was discussed organ to organ. We started with the heart, then the second course would be the lungs, and so on.

2. Reading:

   a) What kind of medical texts have you (have to) read? (e.g. textbooks, scientific publications)
      We had a lot of textbooks when I still was at university, and also some workbooks, with more practical stuff. Those workbooks we would use to discuss the patient cases, they also had questions. Actually, [during the studies] it was not so much scientific publications. Now [during my work] it’s really different, I only read scientific publications. During the four years study at the university, there were only four weeks that we spent on science, and then we suddenly had to read some articles. But it was very difficult for me, because I had not done that before. It was also not enough, a study such as medical informatics is a lot more focused on research than the medical science was. I had to start from the beginning when I started my PhD.
b) Can you name some (online) sources?
I use a lot of UpToDate, PubMed. I also google a lot, the NGE standard for practitioners, there’s also a site NERC, but I haven’t used it in a long time, the pharmaceutical guidelines from Compas. We only get online slides from the lectures, but not really online textbooks, not when I studied.

c) Which of those you consider a “regular” reading material, and which “occasional”?
UpToDate is one that I use very often. [During my studies] we had to read a lot of textbooks, but many students, including me, skipped them, because most textbooks are very tough to read, so mostly students would read the slides and the exams. Then, with the help of the exams, I would look in the book. But it’s not very helpful to read the book from A to Z. I also used the slides and the exams to find the important parts in a textbook, otherwise you read a lot of information that is not very useful.

d) Why do you read them? (e.g. preparations for exams, assignments, informative, entertainment etc.)
In my research, we developed a system that provides assistance to patients with diabetes, where advice would be provided online. To know how to provide this advice in the best way, I have to investigate how information is provided to the patients now, without the online application. So I read a lot of guidelines for doctors, general practitioners, and nurses, and see what kind of information is provided.
I don’t read a lot of medical books anymore, but if I do, it’s because in a while I have to start working in the ward again, so I have to prepare myself. But I don’t read the typical medical books, but the books that work with the cases. You have specific books that would present the case, and then discuss how you can treat the patient, what you have to think of. That’s also something that’s very nice to read.
During the studies, I read because I wanted to pass my exams.

e) Do you usually read the whole “resource”, or do you read it in parts?
In the beginning I read the whole text, because when you are still in high-school, you are used to reading the whole text, you don’t have the big textbooks. But then, after a while, I thought that this is not very efficient, I have to find another way. At that moment I felt like I’m not going to learn by reading textbooks, because I don’t know which part of the textbook I really have to know, so then I stopped [reading the whole book].

f) Do you take notes while reading, and how?
Yes. I took my own notes, but we also had summary books that we could buy at certain stores in Amsterdam, I think some students made them. I would use that as a guideline, and I would add my own notes to that. I underline mostly.

g) Do you usually need to go back and re-read the same text more than once?
Mostly, I have to re-read. But the second time, because of the underlined things, I can read more quickly. When I prepared for an exam, we always
had a week off, so from Monday to Thursday I would read parts of the text, and then Thursday or Friday I would re-read all my notes. I also highlight research papers.

3. Learning:

a) Describe the way your study program is set up, in terms of schedule, types of exams etc.

In one year, we had five or six courses, and each course is focused on a certain organ, and then every part of the organ would be discussed, from anatomy, to more clinical cases, but also the pharmaceutical part of the organ. When the course on the organ was finished, you would get an exam. Now, it can differ a lot from week to week. Things that always repeat themselves is that on Monday I have meetings with my supervisors, and I will discuss the things that I have done in the passed week, and plans for the future. The rest of the week is varied.

b) Do you study alone or in groups? How big are those groups?

I’ve done both. Sometimes I learn completely on my own, and sometimes with other students. I think groups are a very efficient way to study, because you can discuss things, and share notes. I actually like interactive studying more. We also had group assignments.

c) Do you usually use question cards?

No.

d) How much time do you need to prepare for an exam? For an assignment?

The week off before the exam. In AMC they have open questions, [the exams] have different parts – some cases that would be discussed, and always a pathology part, and a pharmaceutical part.

e) Do you use educational software in your study? (e.g. question card programs, quizzes etc.)

I use Blackboard to download material, but not really in an interactive way. But I know that, during my studies, I had a need for [educational software]. If we would have had it, I would have used it, because of the interactive part. Recently I’ve ordered a special DVD, where some kind of teacher will discuss all kinds of cases. I haven’t seen it myself yet, but I know that a lot of my colleagues are using it.

f) How do you stay in touch with your fellow students to discuss assignments, course recommendations, and so on? Do you use social media (e.g. Facebook groups)?

At the time, I didn’t have Facebook yet, also not smartphone. I just used the phone, or we came together in the AMC. Some students from medical informatics have their own chatroom on Facebook, so I think it is being used very often now.

4. Competitions:
a) Are you aware of any medical (award) competitions? What kind? Have you participated in them?
Never.

b) Which medical areas or topics would you be most comfortable competing in?
Only if it were anonymous.

5. Gaming:

a) Do you play video games? Single/multi player? On PC/console/mobile? What kind (e.g. role-playing, social network, puzzle etc.)?
My brother has a Nintendo, that’s something that I like to play, and also on my smartphone. I like multi-player games, mostly on a mobile. I said I don’t like medical competitions, but for gaming, I do like the competitive aspect. I have never played social network games. On public transport is a really good time to play. The video games I play on my smartphone, I always play on the train, almost never on public transport, or at home.

b) How much time do you spend playing per day? Per week?
Once a week, because I don’t have the time.

6. Participating in crowdsourcing and community activities:

a) Do you contribute to community projects for sharing medical knowledge (e.g. writing Wikipedia articles)? Can you describe the projects, along with the tasks that you perform?
One of my supervisors has his own DiaPedia like Wikipedia, but for diabetes. He always asks his PhD students to contribute, and to write some things about our own focus of research. I had to write on the latest knowledge in insulin therapy, for example. It’s also possible that I start writing a certain text, and then other people read it and adapt it. The person that started DiaPedia is also focused on how to provide the information in the most optimal way to, for example, patients that do not have a lot of knowledge, but also to doctors that have a lot of knowledge, so provide more tailored information.

b) Would it be possible to incorporate a crowdsourcing activity (like the one described above) with your course work?
Yes, I think it’s a very good way to learn as a student. It’s also very handy to use on the ward, because there are always things you forget when you’re working, so you could use this [application] as a sort of checklist. I think I would use it in such a way.

c) What (medical) topics would be best suited for such an activity?
Finding the right diagnosis. It can also be used for indentation of laboratory results. It’s also very handy to have a medical tool that will say what kind of evaluations you have to do, a sort of guideline that’s implemented in the software.

d) What kind of tasks would you like to solve? (e.g. reading, testing your knowledge etc.)
I’d like to test my knowledge. I would like to do this in a kind of detective way, that has to find out what the problem of the patient is, what is the diagnosis, and then it’s possible to say in the game that I want to have an X-Ray, or I want to have some laboratory results, and then you have more clues. I would like that kind of application a lot.

e) How much time would you spend on this activity per session? How much time per week?
Once a week, in the weekends, or while going to work.

f) Would you prefer to complete the activity alone, or in a group?
Alone, because then I can choose myself when I want to participate.

g) What would make such an activity useful to you?
For me, it would be useful for preparation before working on the ward. I have not been treating patients outside my studies for three years, so then your knowledge disappears. This would be a very nice way to gain that knowledge again.

7. Further participation:

a) Do you want to receive a transcript of this interview?
No.

b) Do you want to stay updated on the progress of this project?
Yes.

c) Do you want to try out our application once we have an initial version running?
Yes.

B.2.2 Interview Medical Professional P2

1. General questions:

a) What is your study course/specialization?
I studied medicine, I did my internships. In Netherlands we call it “basisarts”, so I’m a basic physician. I do have a registration, but I’m not specialized.

b) How many years have you spent studying medicine?
The study is six years, four years of doctoral studies and two years of internships. I started in September 2000, and I finished in 2008, so I did it in eight years. I started doing research during my studies. I have been doing research since I graduated, so approximately five years.

c) What kind of courses you have been taking as part of your program (e.g. lecture, laboratory session)?
I wrote literature studies and reviews, case studies during my internships, and I have also conducted lab research.
d) *Describe the contents of your program. How does it progress across the study years?*

The program I did is called an MD PhD, I did PhD research. Now I work in a medical microbiology department. Before, I also worked in the neurology department. The field I’m working in is virology – viral infections of the nervous system. I did pathological research, which is research with human and animal tissue sample. I also worked with cells, I conducted lab research [on that].

2. **Reading:**

   a) *What kind of medical texts have you (have to) read? (e.g. textbooks, scientific publications)*

      Research papers, reviews, other types of articles. During my studies, we had medical textbooks, both Dutch ones and English ones, and handouts for studying.

   b) *Can you name some (online) sources?*

      What I use for research is PubMed. There are other ones, but I mainly use PubMed. Sometimes I use the medical library here. I do searches for author names and keywords. Before, when I was studying, I used the textbooks and handouts that I got for studying.

   c) *Which of those you consider a “regular” reading material, and which “occasional”?*

      PubMed is regular.

   d) *Why do you read them? (e.g. preparations for exams, assignments, informative, entertainment etc.)*

      Before, when I was studying, for exams we had certain parts of the textbook [to read]. For the research I use PubMed for the background of research, maybe to set up research, to evaluate research, or to put it in context, and to write articles.

   e) *Do you usually read the whole “resource”, or do you read it in parts?*

      I search for keywords, or an author, and then I look at articles that seem relevant [to my research]. I might also look at the abstract on PubMed. Sometimes I read the whole article, sometimes parts. For textbooks, maybe you’re interested in certain topics, then you look at the summaries, you go to the chapter, and you read that chapter. The search is different, because PubMed is digital, I can just type for keywords.

   f) *Do you take notes while reading, and how?*

      I underline parts in the text. I can write some things next to the text.

   g) *Do you usually need to go back and re-read the same text more than once?*

      Sometimes I do re-read the same text. Sometimes I print articles and I keep them [to re-read].

3. **Learning:**
a) Describe the way your study program is set up, in terms of schedule, types of exams etc.

In Groningen [University] we had a system, Problem-Based Learning (PBL). Sometimes we had teams, we would get a couple of cases, and we had tutor groups with a number of students. We would get a couple of cases, and we would brainstorm about it. For example, a person presenting these symptoms, we would brainstorm about it and search literature. Then we had another meeting where we would present [our results] to the group. For example, on Monday we could have a brainstorm, and on Thursday, a presentation. Those cases were be embedded with a theme for the week, or for the period. Under the umbrella of a course, for several weeks, or a month, we would have different cases. That also had to do with literature search. We also had other things – practical skills, where we did physical examinations, taking histories of simulation patients – people that would register here and we would take their histories.

We did have multiple choice exams, in blocks. We had a period of a couple of weeks, or a month, and then we would have multiple choice exams. Then we also had "Vooruitgang” a couple of times a year, progression exams. These were a vertical thing, you had them in every year. Then we had exams for the physical examinations and taking histories. Most of the times we had simulation patients, they would have a history and present current symptoms, we would question them and take their history, maybe do a physical examination or suggest further testing. We would also get a grade for that. In the last two years of study, we had internships in the hospital. There you would also have other kinds of tests. You followed certain modules, like procedures, and you had exams with clinicians.

b) Do you study alone or in groups? How big are those groups?

Sometimes I did go to the library with friends. But in PBL, we had tutor groups, and also mentor groups.

c) Do you usually use question cards?

When I did my internships, I did have pocket handbooks in my coat. I read about question cards, some people are using them, also apps. I would definitely be interested in [them].

d) How much time do you need to prepare for an exam? For an assignment?

It would be like a process. I would start studying, read the text, maybe for a second time as well. I had different exams, so I don’t know if I can say something on average. [For PBL], the assignments were every week. We also had response classes, where you could file questions about the case that week, and you could get answers.

e) Do you use educational software in your study? (e.g. question card programs, quizzes etc.)

For pharmacy and pharmacology, we had a system called PDrug, “Personalized Drug”. We also had something called “Pharmaceutical Compass”. I think it was a tool to find referenced drugs. We had Nestor, digital learning
environments like Blackboard, Progress for grades and assignments. We also used those during PBL to post our cases.

f) *How do you stay in touch with your fellow students to discuss assignments, course recommendations, and so on? Do you use social media (e.g. Facebook groups)?*

We used those digital learning environments. Sometimes, you were together with your group for a trimester or a semester, and you could share your files with them [online]. Now, I work in a hospital. In the department, we have a mailing list. I’m also a member of an alumni society.

4. **Competitions:**

   a) *Are you aware of any medical (award) competitions? What kind? Have you participated in them?*

   I don’t know if there are any.

   b) *Which medical areas or topics would you be most comfortable competing in?*

   General medical knowledge might be interesting.

5. **Gaming:**

   a) *Do you play video games? Single/multi player? On PC/console/mobile? What kind (e.g. role-playing, social network, puzzle etc.)?*

   I used to play Command and Conquer, Warcraft, FIFA (I really like football games), Sim City. Sometimes I use my mobile phone, they have some pretty funny games.

   b) *How much time do you spend playing per day? Per week?*

   I have some games on my phone, I don’t know how much I use it, it’s just when I have some free moments.

6. **Participating in crowdsourcing and community activities:**

   a) *Do you contribute to community projects for sharing medical knowledge (e.g. writing Wikipedia articles)? Can you describe the projects, along with the tasks that you perform?*

   I use Wikipedia, but I never contributed.

   b) *Would it be possible to incorporate a crowdsourcing activity (like the one described above) with your course work?*

   There might be already some [similar applications] out there, digital learning apps. It might be something interesting, something that could assist with learning. In internships, those apps are also being used.

   c) *What (medical) topics would be best suited for such an activity?*

   It can be applied to a number of topics. Medical decision making is an interesting area. I did study medicine, so I don’t know which systems are already out there, but I think it can be applied to a number of fields.
d) **What kind of tasks would you like to solve? (e.g. reading, testing your knowledge etc.)**

Maybe you could do a game like the PBL, where you would have someone presenting [certain symptoms], and [the app] would guide you through the steps.

e) **How much time would you spend on this activity per session? How much time per week?**

If it would be an app, that’s pretty low-key, that’s pretty nice. You could have it in your pocket, [and use it] whenever you feel like it.

f) **Would you prefer to complete the activity alone, or in a group?**

Both are nice. In a group it’s also nice, because as a physician, you also work in a group, you have group discussions. In a clinic, you have discussions about cases, morning reports that you discuss with other people.

g) **What would make such an activity useful to you?**

I like the fact that it [could be] low-key, to have it in your pocket. If it’s something helpful for medical studying, that might also be nice.

7. **Further participation:**

   a) **Do you want to receive a transcript of this interview?**

      Yes.

   b) **Do you want to stay updated on the progress of this project?**

      Yes.

   c) **Do you want to try out our application once we have an initial version running?**

      Yes.

**B.3 Lecturers**

**B.3.1 Interview Lecturer L1**

1. **General questions:**

   a) **What is your specialization?**

      My specialization is molecular microbiology. I studied in Wageningen, I did a technical education in bioprocessing engineering. Then I went to Sweden for four years, I did my PhD there on food microbiology. I went to Canada for a little over two years, and I did a post-doc, so I was a researcher there. And then I came at the end of 2005 to Maastricht [University], I wanted to use the molecular applications that I was developing in a medical setting. Since 2006 I am an assistant professor here at the university, and at the same time I am the unit leader of the molecular microbiology laboratory.
b) What kind of classes do you teach?
   I am involved in a lot of different classes, mostly for first and second year medical students, as well as for European public health students, and molecular life science students. The way things are organized in Maastricht [University] is that everything is integrated, but I am always involved when part of the course involves microbiology. Mostly the topics are everything that has to do with the human gut. For example, abdominal complaints is one of the topics we lecture about. And then there’s immunology, microbiology, gastroenterology, those are all combined together in a course. We have eight and four weeks courses.

2. Reading material and activities (for students):

a) Is there a list of obligatory reading?
   Yes.

b) What kind of (medical) texts do you usually assign for reading?
   Both [textbooks and papers]. The way the Maastricht system works is that we do not tell the students what they have to read, we have some advice books, but they can get the information from wherever they want. And that’s also because we have the group meetings, the Problem-Based Learning (PBL). So the idea is that they gather their own information. But we do have some advice books, they are usually books or review articles, in some cases, from some very good websites. [We have] all kinds of textbooks. Online, it’s for example the site from the WHO[^8] that’s one that we use quite a lot, or RIVM[^9] for the Dutch students.

c) How much time do the students spend on reading?
   Approximately 23 hours a week they study by themselves, that’s what they report.

d) Do you have any other comments about student reading habits?
   You can tell that they read a lot online. We try to stimulate them using books, but especially the first year students they [do] a lot of Wikipedia-level reading. Especially the first year medical students, they cover a very broad area, and because of that, and since they are just getting into the introductions, a lot of the information that they get is at that level.

3. Student homework & assignments:

a) What are typical types of homework? (e.g. summarizing, answering questions etc.)
   [We have] PBL. The main part is that every week we have a case that we discuss, and there a problems comes out, and we come up with maybe five to ten questions, and from these questions about a specific topic we define our main goals, and then the students will then find information

[^8]: http://www.who.int/en/
[^9]: http://www.rivm.nl/
about these learning goals. That is the main part of the work that they do at home. Aside from that they have practical tutorials. It could be all sorts of things, like technical skills labs – for that they also have to prepare, and sometimes also look up information and write reports about that, or prepare for a presentation. It’s quite a wide variety.

b) *How much time do the students spend on assignments and homework?*

Approximately 23 hours a week they study by themselves, that’s what they report.

c) *Do the students usually work individually, or in teams? How big are the teams?*

Usually, the presentations or the reports that they need to write are team assignments. The teams are as small as two, and as big as ten people. We have to partner up because we have a lot of students.

4. **Exams:**

a) *How are the exams usually structured? What types of (how many on the average) questions do the students need to answer?*

Up until now, especially for the medical students, we had multiple choice questions. If you, for example, have the European public health students, or the molecular life science students, they are the smaller group, they may be 80 students, and there it’s more doable to ask open questions. But we started last year with some experiments doing open questions with medical students. We’re still working on that. We’re also looking at different assignments where they do the exam on the computer, and they get the building blocks, where they have to build the process themselves.

5. **Students participating in crowdsourcing and community activities:**

a) *Would it be possible to incorporate (how and where) the student work on homework, assignments or exam preparation as part of such crowdsourcing activities?*

Yes, I can definitely see it. For example, [with] the ICT solutions that have been looking at how to incorporate into our classes. For example, in immunology, you have different pathways in the immune system. We have pictures that are adapted from textbooks, we take out parts, and we ask the students to fill them in, or we can discuss it. Or for example, you have the intestinal system from mouth to anus, [we ask] what if something is blocked here. I can imagine that you could use something in that area.

b) *What (medical) topics would be best suited for such an activity?*

Immunology is a good area, because it’s quite complex, especially for starting students it’s hard to grasp. It’s very dry when you’re just reading it from a textbook, so if you could use something else, I think that would be a high incentive for the students to use it. Another area maybe could be something like anatomy, there the information is quite structured, there are not really any problems, they could be using little tests to find out things
themselves. There are so many areas in microbiology where I could see a lot of applications.

c) **What kind of tasks would be best suited for students to solve as part of this? (e.g. reading, testing their knowledge etc.)**

   It’s a difficult question, I have to think about that.

d) **What would make such an activity useful to your students?**

   Either to test themselves, or to see things in a different ways. What I said about immunology, they are reading things in the textbook, but when you can build things together in a different way, it’s interactive.

6. **Reading:**

   a) **What kind of medical texts have you (have to) read? (e.g. textbooks, scientific publications) Can you name some (online) sources? Why do you read them? (e.g. preparations for exams, assignments, informative, entertainment etc.)**

   It depends. If I’m teaching in a new area, since I don’t have a medical background myself, I use a lot of the books that the students are using. I also use a lot of databases, for example UpToDate[^10], where you can quickly have comprehensive information.

   b) **Do you usually read the whole “resource”, or do you read it in parts?**

   I read it in parts. I scan through it to find the areas, because you have to be quite efficient.

   c) **Do you take notes while reading, and how?**

   It depends. I prefer to have things in digital or on paper, where I can highlight. If that’s not enough, then I’ll make some small notes.

   d) **Do you usually need to go back and re-read the same text more than once?**

   That’s why I prefer to highlight, because then I can have the summary, but at the same time, if I need more information, or if I can’t grasp it anymore, then I read the text that is around it. I prefer that over making my own notes, because I find that I don’t always know anymore what I wrote down.

7. **Gaming:**

   a) **Do you play video games? Single/multi player? On PC/console/mobile? What kind (e.g. role-playing, social network, puzzle etc.)?**

   Yes, [I play] single and multi player, on the PC, on the PlayStation. Mostly role playing, but also strategy games, puzzle solving. It depends on what time of the day it is, if I’m tired, or if I want to have something quick to relax, or I want to get really into it. At the moment, because I’m so busy, I’m not involved in [social media games], but I have been in the past, I can enjoy it. But you have to have a certain amount of time to get into that. If I had more time, I would enjoy that. I enjoy all kinds of games,
but I don’t play so much on the phone, because it’s smaller, and I enjoy a bigger [screen]. The other thing for me has to do with location – if I’m away/traveling, I use the smartphone quite a lot, but otherwise, I use the phone a lot while working, and while I’m working I’m busy, and then I come home and I throw the phone away somewhere in a corner, and I use other machines.

b) How much time do you spend playing per day? Per week?
Five hours a week.

8. Professors participating in crowdsourcing activities:

a) What would make such an activity useful to you or your colleagues?
It could be [incorporated] in a whole different way. You could use it to summarize knowledge about a topic, or to gather the alternatives. I would probably play it because I find it fun. The difference with the students is that they have to learn something by heart, for us [it’s not needed anymore], if [we] know where to find [the knowledge] quickly. Coming back to courses, to prepare for lecturing, it could be nice to find examples [to use in class]. I try to offer things to students in a way that they can perceive, so [the application] could help me with brainstorming, or for example, help me think for ideas for exam questions. I have just recently been making a lot of multiple choice questions, and it’s very difficult to come up with good new alternatives. If I could use it for inspiration, that would definitely be something I could use. I can definitely see applications – to help with the studying, as a teaching tool. I find it difficult at the moment to think of how to use it during research.

9. Further participation:

a) Do you want to receive a transcript of this interview?
Yes.

b) Do you want to stay updated on the progress of this project?
Yes.

c) Do you want to try out our application once we have an initial version running?
Yes, absolutely.

B.3.2 Interview Lecturer L2

1. General questions:

a) What is your specialization?
I was trained in physics. I did my MSc in medically applied physics, then I started working in medical informatics, so I think that’s my specialization. I have done my PhD in medical terminology, then formal terminology, concept representation in medicine, description logic, things like that.
We started building medical terminology ourselves. We found out that building one is one thing, maintaining it and quality assurance is hard. We tried to do some crowdsourcing in that, getting clinicians to see if they agree with the concept definitions. It turned out to be very hard to get organized, to get the people involved, to reach agreement. We spent quite some time on that, maybe ten years. Then Netherlands has joined the organization that owns SNOMED CT\(^\text{11}\) (medical terminology, 3,000 concepts). Then we thought we can either keep moving on with our own efforts, or we can gear or efforts towards international work.

Since then, we have been working on things related to SNOMED CT, how to assess its quality, how to further improve it, find mistakes, incompleteness in it, questions related to how we can translate it in full or in part from English to Dutch, because we would still need a lot of Dutch to address the needs of clinical users, and also how do you adequately use it in practice. Having a terminology is one thing, but then especially SNOMED you have this compositionality, [for example] I could say I have a complex fracture of the left femur, that would not be a pre-existing concept, but you could say “fracture of the femur”, that is a concept, and then you have the left femur, so you want to compose this concept. Doing that in an adequate way is a pain. The other thing that we are paying attention to is, suppose that we have things expressed using SNOMED CT concepts, how can we use that for clinical support.

b) *How many years of experience do you have?*

I’m in medical informatics for 21 years now.

c) *What kind of classes do you teach?*

We have our program in medical information here, with a three year BSc, two year MSc, and then PhD possibilities. I’ve been teaching in this program for some ten years. I’ve been doing software engineering, databases, and I’m involved in a couple of courses where this formalization takes a role, so where we tell people about these terminologies, what roles do they have, how to use them, what kind of terminologies exist – as reference terminologies, or as input terminologies, classifications for lumping groups of people or diseases together – also stuff pertaining to the information modeling, and reference information models, and archetyping. For example, if I want a formalization of this guideline, according to which standard [should I do it], then you have the terminology standards, but also the information models, and they belong together. This is what we try to explain to our students, telling them about semantic web, and frames, and description logic. I also have a week in our MSc program in knowledge representation and reasoning, where I try to take things a step further. I have also been doing some semantic web with my students.

2. *Reading material and activities (for students):*

   a) *Is there a list of obligatory reading?*
Yes, but more medical informatics readings, with the focus on the informatics part, applied to some medical domain. For example, a paper on what is SNOMED CT will have some medical terminology in it, but it is not about, say, SNOMED applied to cardiology, so the medical content is a bit shallow.

The most medical texts are some guidelines that they have to read. In this knowledge representation course in the MSc, we just moved to a guideline in atrial fibrillation, that [the students] have to read and understand. I think guidelines in general would be good for learning things like treatment and symptoms, especially diagnostic guidelines.

b) **What kind of (medical) texts do you usually assign for reading?**

Either scientific literature, or professional societies. Generally, we always use PubMed for everything. An organization like that would be the American Health Association.

c) **How much time do the students spend on reading?**

Too little. It depends. When I’m teaching databases, there is not much medical stuff in there, we have some medical oriented practical sessions, but it’s mostly just. [In general, it’s] some hours, rather than days, per course.

3. **Student homework & assignments:**

a) **What are typical types of homework? (e.g. summarizing, answering questions etc.)**

The students have to build a rule-based system around the guidelines, which is not my part in the course. My assignment in this course is to develop a computerized patient record in which you can adequately store the information that is needed to support this guideline. For example, if you have a pacemaker, you need to have the procedure for the pacemaker implant. I want students to find, out of all the different ways to say the patient has a pacemaker, what in their opinion is the best way to represent that, and motivate why. It’s a bit of information modeling, a bit of terminology, and putting these together to develop a record, which you could use to trigger the rule-based system.

b) **How much time do the students spend on assignments and homework?**

Three days for the patient record, and then the rule-based system in a couple of weeks. They may spend up to one month working on this.

c) **Do the students usually work individually, or in teams? How big are the teams?**

In small groups, couples.

4. **Exams:**

a) **How are the exams usually structured? What types of (how many on the average) questions do the students need to answer?**

In this case, there are no exams.
5. Students participating in crowdsourcing and community activities:

a) Would it be possible to incorporate (how and where) the student work on homework, assignments or exam preparation as part of such crowdsourcing activities?

That would be great, but hard [because of time constraints]. We have been thinking about things like that for involving our medical students, for example, like I said, for translating SNOMED CT in Dutch.

b) What (medical) topics would be best suited for such an activity?

Both diagnosis and selection of best treatment would be relevant topics.

c) What kind of tasks would be best suited for students to solve as part of this? (e.g. reading, testing their knowledge etc.)

[Tagging concepts] could be one. For example, a student selects thromboembolic complications [from a guideline], and looks up the [appropriate] SNOMED CT concept, if there is a concept for that. Synonym phrases would be hard to do. Relations, such as differential diagnosis, are also found in guidelines.

d) What would make such an activity useful to your students?

There can be many derivative goals. One goal would be understanding medicine a little bit better. My teaching goal would be students understanding terminology better. For example, when you read [a long text], you read on, and you don’t really think about it, there is a lot in it, but what can you learn from it?

6. Reading:

a) What kind of medical texts have you (have to) read? (e.g. textbooks, scientific publications) Can you name some (online) sources? Why do you read them? (e.g. preparations for exams, assignments, informative, entertainment etc.)

For me, it’s generally medical informatics work. It’s not going deep down in the medical domain, but of course it’s all based on medical knowledge. Sometimes you have some projects which are more towards a specific condition, and then you are forced to learn a bit more. I think I would generally first resort to Wikipedia, because that can explain diabetes, for example, in a way that I can understand, rather than reading a scientific article on diabetes.

b) Do you usually read the whole “resource”, or do you read it in parts?

Depends on why I’m reading. Sometimes you focus on the results, sometimes you focus on the quality of the paper, when you’re reviewing papers.

c) Do you take notes while reading, and how?

Not enough. I try to read a lot electronically. I scan and fool myself to think that I will remember, and then I don’t.

d) Do you usually need to go back and re-read the same text more than once?

Yes.
7. Gaming:

a) Do you play video games? Single/multi player? On PC/console/mobile? What kind (e.g. role-playing, social network, puzzle etc.)?
Not many, and if I do, they are really the stupid games. [I play] on mobile, I have been doing WordFeud for a while. If you can really cut down [your app] into small chunks, something you can do on the go for ten minutes, and save your work, then it’s fun.

b) How much time do you spend playing per day? Per week?
Not more than fifteen minutes. Currently, not much at all. Sometimes you discover a nice game, and then you spend more time [playing].

8. Professors participating in crowdsourcing activities:

a) What would make such an activity useful to you or your colleagues?
If I benefit from it. I have never contributed to Wikipedia, because it’s quite some work. Pressure in academia is more and more on production, and Wikipedia articles are not regarded as production, so if there would be a merit or credit system which was accepted, that could help. If you want to get clinicians in, it could be great if you could give them CME (continuing medical education) credits, because then they would know they get something back for their investment. And if it’s fun, that’s easier of course. People play Angry Birds even though they don’t get CME credits. So it should either be fun, or rewarding in some other way, not because people do not support your goals, but because there’s a lack of time and energy. If you could get your application to medical students, then there will be more possibilities, because they have to learn and read anyhow.

9. Further participation:

a) Do you want to receive a transcript of this interview?
No.

b) Do you want to stay updated on the progress of this project?
Yes.

c) Do you want to try out our application once we have an initial version running?
Yes.

B.3.3 Interview Lecturer L3

1. General questions:

a) What is your specialization?
My background is in medical informatics. I am involved in three research lines: one is evaluation of quality of care, another one is developing methodologies for evaluating impact of health IT, and the third is
semantic interoperability, or information standards and terminologies, and especially SNOMED CT.

b) *How many years of experience do you have?*

I graduated in 1994, then started my PhD in 1996, and graduated in 2004. So that is eighteen years.

c) *What kind of classes do you teach?*

The courses are mostly for BSc and MSc students in bioinformatics. I’m also teaching people at the national ICT institute for SNOMED. I teach in medical terminologies, so information standards. I also teach in doing evaluation studies in healthcare, to evaluate the quality of care, or to evaluate the effect of IT implementations to healthcare.

2. **Reading material and activities (for students):**

   a) *Is there a list of obligatory reading?*

   For the BSc and MSc students, there is a list of obligatory reading, it’s mainly scientific papers.

   b) *What kind of (medical) texts do you usually assign for reading?*

   Most of them are medical informatics journals, or medical journals, so you can retrieve them from PubMed.

   c) *How much time do the students spend on reading?*

   Quite a lot of time. Most of their homework, for my lecture, is reading those scientific journal papers, and of course they have to write some reflections on it, and as part of their examination, they need to know the content of the journals.

3. **Student homework & assignments:**

   a) *What are typical types of homework? (e.g. summarizing, answering questions etc.)*

   It’s mostly reading the papers, so we give the lectures and explain the concepts, but [the students] have to read the papers to understand the full context, and get more details. What we did last year was that we asked them to read three papers on a specific topic, and then they have to, as an assignment, write a kind of letter to the editor about those three papers. It’s most of the time, either a critical appraisal of the papers, or doing a presentation on what they learned from the papers, or writing something like a letter to the editor.

   b) *How much time do the students spend on assignments and homework?*

   On average, they have between ten and around fifteen hours of lectures here, so it’s about twenty hours self-study for reading all the text, and writing papers, or doing other assignments. I can imagine combining an assignment of mine with the Watson application, so maybe reformulating an assignment – not writing a letter to the editors, but reading the papers in the context of your goal, and searching for important concepts, answering
questions about that, and then maybe a self-assessment on the part of the student, asking “do I really understand the paper?”, and then that understanding can be used for your project.

c) **Do the students usually work individually, or in teams? How big are the teams?**

It differs. Often they work in couples, but there are also assignments that need to be done individually.

4. **Exams:**

a) **How are the exams usually structured? What types of (how many on the average) questions do the students need to answer?**

At the end of the module there is a three hour exam that covers all the topics of the module. Most of them are open questions, but from this year we have the “bonus points” – students do an intermediate testing to motivate themselves to study regularly during the module, and they can earn bonus points in their grade – we used multiple choice for that. The type of questions is different. I try to avoid questions like “what’s the definition of”. I always try to maybe propose a case study to which they can apply the theory they have learned.

5. **Students participating in crowdsourcing and community activities:**

a) **Would it be possible to incorporate (how and where) the student work on homework, assignments or exam preparation as part of such crowdsourcing activities?**

What the incentive will be for the student, but also for me, as a teacher, is my main concern. To be honest, I cannot imagine that I will just voluntarily read text and solve a task.

There will be a new MSc program developed, “Health Informatics for Healthcare Professionals”, for people working as doctors or nurses, that want to find out more about health informatics. This course will be mainly based on e-learning, because these people have a job, and they maybe want to spend ten hours a week on this program, so we need a lot of self-assessment integrated with e-learning. I can imagine that this question and answering system can be valuable for your program. The questions will be multiple choice though.

b) **What (medical) topics would be best suited for such an activity?**

In my subject (health informatics), it’s all text based. I think most of the medical topics are textbook or publication based.

c) **What kind of tasks would be best suited for students to solve as part of this? (e.g. reading, testing their knowledge etc.)**

Especially for self-assessment, if they have to read a paper and answer some questions on it. Or even more simple, they have to identify which part of the text is more important, that helps them to summarize and understand the paper, and that might help you. But students should get some feedback.
d) **What would make such an activity useful to your students?**

I think they should get the feeling that it helps them to learn, to understand the material, or to make clear what parts are not completely clear to them, so that they can ask the teacher for more explanation.

6. **Reading:**

a) **What kind of medical texts have you (have to) read? (e.g. textbooks, scientific publications) Can you name some (online) sources? Why do you read them? (e.g. preparations for exams, assignments, informative, entertainment etc.)**

You do your literature search in PubMed or Google Scholar, and there’s a paper that might be interesting for your research area, and you start reading it. I can identify that if you find a paper to be relevant, if it comes up with related work, that also uses that method, or compares that method to another one, it would be useful.

b) **Do you usually read the whole “resource”, or do you read it in parts?**

The first thing that I do is look at the end of the introduction to see whether the objective of the paper is really relevant to the work that I’m doing. The abstract can give a general overview, preferably if it’s a structured abstract, to see what the goal of the paper, the methods, the main results, the conclusion.

c) **Do you take notes while reading, and how?**

Sometimes I take notes. I’m quite old fashioned, I use a highlighter. Sometimes I make summaries in a text file. A tool where you highlight, or maybe making an automated summary of the highlighted parts could be useful to you, and it would be also useful to the one reading the paper. But sometimes you read a paper from one perspective, and sometimes it’s a different perspective, and then you highlight different parts. Sometimes I’m only interested in the outcome of the study, so does medication A work or not, and sometimes I want to see if the study design that they used is applicable to my situation. It’s the same paper, but I will highlight completely different things.

d) **Do you usually need to go back and re-read the same text more than once?**

It depends. When the paper is really interesting, I re-read it. I re-read notes for everything else.

7. **Gaming:**

a) **Do you play video games? Single/multi player? On PC/console/mobile? What kind (e.g. role-playing, social network, puzzle etc.)?**

Not at all. My children play a lot, but I really do not like it. But I recently attended a conference on serious gaming, and I was really impressed. I think if I can do a game where I can also learn something, then I would be really interested, because it’s much more fun to play and learn than to just
read. The conference I attended was really on medical topics, like learning anatomy. To translate that to my area, is really difficult. Of course, you can make quizzes, but to make a game that is really attractive, it’s difficult.

b) *How much time do you spend playing per day? Per week?*

N/A.

8. **Professors participating in crowdsourcing activities:**

a) *What would make such an activity useful to you or your colleagues?*

If it saves time, if it motivates our students so that they get better grades, if it gives me incentives like related work. I can imagine, for reading the papers, if it’s a tool like EndNote, and then extending this with something for highlighting, or relating the papers, based on a common method for example. For the students, the gaming idea is very attractive. Maybe also for me, if I’m in the role of a student.

9. **Further participation:**

a) *Do you want to receive a transcript of this interview?*

No.

b) *Do you want to stay updated on the progress of this project?*

Yes.

c) *Do you want to try out our application once we have an initial version running?*

Yes.
Appendix C

Interview Analysis Tables

Table C.1: Demographics for lecturers (1).

<table>
<thead>
<tr>
<th>Intv.</th>
<th>Specialization</th>
<th>Years of Experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>B.3.1</td>
<td>molecular microbiology, bioprocessing engineering, food microbiology</td>
<td>18</td>
</tr>
<tr>
<td>B.3.2</td>
<td>medically applied physics, medical informatics</td>
<td>21</td>
</tr>
<tr>
<td>B.3.3</td>
<td>medical informatics</td>
<td>18</td>
</tr>
</tbody>
</table>

Table C.2: Demographics for lecturers (2).

<table>
<thead>
<tr>
<th>Intv.</th>
<th>Teaching Areas</th>
<th>Courses Taught</th>
</tr>
</thead>
<tbody>
<tr>
<td>B.3.1</td>
<td>medicine, European public health, molecular life science</td>
<td>microbiology, human gut</td>
</tr>
<tr>
<td>B.3.2</td>
<td>medical information</td>
<td>medical terminology formalization</td>
</tr>
<tr>
<td>B.3.3</td>
<td>bioinformatics</td>
<td>medical terminology formalization</td>
</tr>
</tbody>
</table>

Table C.3: Demographics for medical professionals (1).

<table>
<thead>
<tr>
<th>Intv.</th>
<th>Specialization</th>
<th>Years of Experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>B.2.1</td>
<td>medicine, metabolic diseases</td>
<td>11</td>
</tr>
<tr>
<td>B.2.2</td>
<td>medicine, medical microbiology, virology, neurology</td>
<td>13</td>
</tr>
</tbody>
</table>

Table C.4: Demographics for medical professionals (2).

<table>
<thead>
<tr>
<th>Intv.</th>
<th>Courses Taken</th>
<th>Current Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>B.2.1</td>
<td>lectures, patient case seminars, laboratory sessions (anatomy, microscopy, pathology)</td>
<td>PhD student, resident</td>
</tr>
<tr>
<td>B.2.2</td>
<td>literature studies, patient case seminars, problem-based learning seminars</td>
<td>PhD, basic physician</td>
</tr>
</tbody>
</table>
### Table C.5: Demographics for students (1).

<table>
<thead>
<tr>
<th>Intv.</th>
<th>Specialization</th>
<th>Years of Experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>B.1.1</td>
<td>medicine, biomedical sciences</td>
<td>3</td>
</tr>
<tr>
<td>B.1.2</td>
<td>life science, medicine, microbiology</td>
<td>7</td>
</tr>
<tr>
<td>B.1.3</td>
<td>clinical molecular science, medicine, microbiology</td>
<td>9</td>
</tr>
<tr>
<td>B.1.4</td>
<td>medicine, epidemiology, nephrology</td>
<td>9</td>
</tr>
<tr>
<td>B.1.5</td>
<td>veterinary medicine, medical informatics</td>
<td>10</td>
</tr>
</tbody>
</table>

### Table C.6: Demographics for students (2).

<table>
<thead>
<tr>
<th>Intv.</th>
<th>Courses Taken</th>
<th>Current Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>B.1.1</td>
<td>lectures, laboratory sessions</td>
<td>BSc student</td>
</tr>
<tr>
<td>B.1.2</td>
<td>lectures, laboratory sessions, problem-based learning seminars</td>
<td>PhD student</td>
</tr>
<tr>
<td>B.1.3</td>
<td>lectures, laboratory sessions, problem-based learning seminars</td>
<td>PhD student</td>
</tr>
<tr>
<td>B.1.4</td>
<td>lectures, laboratory sessions, patient case seminars</td>
<td>PhD student</td>
</tr>
<tr>
<td>B.1.5</td>
<td>lectures, technical courses, organizational courses</td>
<td>PhD student</td>
</tr>
</tbody>
</table>
### Table C.7: Overview of reading content.

<table>
<thead>
<tr>
<th>Intv.</th>
<th>Types</th>
<th>Sources</th>
<th>Reasons</th>
</tr>
</thead>
<tbody>
<tr>
<td>B.3.1</td>
<td>textbooks</td>
<td></td>
<td>teaching</td>
</tr>
<tr>
<td></td>
<td>publications</td>
<td>UpToDate</td>
<td>quick references</td>
</tr>
<tr>
<td>B.3.2</td>
<td>publications</td>
<td></td>
<td>research</td>
</tr>
<tr>
<td></td>
<td>short reviews</td>
<td>Wikipedia</td>
<td>quick references</td>
</tr>
<tr>
<td>B.3.3</td>
<td>publications</td>
<td>PubMed, Google Scholar</td>
<td>research</td>
</tr>
<tr>
<td>B.2.1</td>
<td>publications</td>
<td>UpToDate, PubMed</td>
<td>research</td>
</tr>
<tr>
<td></td>
<td>workbooks</td>
<td>professors</td>
<td>exam preparation</td>
</tr>
<tr>
<td></td>
<td>slides</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>guidelines</td>
<td>NGE, NERC, Compas</td>
<td>research</td>
</tr>
<tr>
<td></td>
<td>course summary books</td>
<td>Student-written</td>
<td>exam preparation</td>
</tr>
<tr>
<td></td>
<td>case books</td>
<td></td>
<td>ward work preparation</td>
</tr>
<tr>
<td>B.2.2</td>
<td>publications</td>
<td>PubMed, university library</td>
<td>research</td>
</tr>
<tr>
<td></td>
<td>textbooks</td>
<td></td>
<td>exam preparation, problem solving</td>
</tr>
<tr>
<td></td>
<td>handouts</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B.1.1</td>
<td>textbooks</td>
<td>university library</td>
<td>exam preparation</td>
</tr>
<tr>
<td></td>
<td>publications</td>
<td>professors</td>
<td>reading group</td>
</tr>
<tr>
<td>B.1.2</td>
<td>textbooks</td>
<td>university library</td>
<td>exam preparation, problem solving</td>
</tr>
<tr>
<td></td>
<td>publications</td>
<td>UpToDate, PubMed, Google Scholar</td>
<td>research</td>
</tr>
<tr>
<td>B.1.3</td>
<td>publications</td>
<td>PubMed, university library</td>
<td>research</td>
</tr>
<tr>
<td></td>
<td>textbooks</td>
<td>university library</td>
<td>problem solving</td>
</tr>
<tr>
<td></td>
<td>exam questions</td>
<td></td>
<td>exam preparation</td>
</tr>
<tr>
<td>B.1.4</td>
<td>publications</td>
<td>PubMed, MEDLINE, Cochrane</td>
<td>research, entertainment problem solving</td>
</tr>
<tr>
<td></td>
<td>guidelines</td>
<td></td>
<td>entertainment problem solving</td>
</tr>
<tr>
<td></td>
<td>textbooks</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B.1.5</td>
<td>publications</td>
<td>PubMed, AMED</td>
<td>research</td>
</tr>
<tr>
<td></td>
<td>textbooks</td>
<td></td>
<td>exam preparation, class preparation</td>
</tr>
</tbody>
</table>

### Table C.8: Overview of reading habits.

<table>
<thead>
<tr>
<th>Intv.</th>
<th>Whole/Parts</th>
<th>Notes</th>
<th>Re-read</th>
</tr>
</thead>
<tbody>
<tr>
<td>B.3.1</td>
<td>parts</td>
<td>highlighting</td>
<td>just the notes</td>
</tr>
<tr>
<td>B.3.2</td>
<td>both</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>B.3.3</td>
<td>parts</td>
<td>highlighting</td>
<td>yes/yes</td>
</tr>
<tr>
<td>B.2.1</td>
<td>parts</td>
<td>highlighting, consulting summary books</td>
<td>just the notes</td>
</tr>
<tr>
<td>B.2.2</td>
<td>parts</td>
<td>highlighting, comments, keywords</td>
<td>yes</td>
</tr>
<tr>
<td>B.1.1</td>
<td>chapters</td>
<td>highlighting</td>
<td>just the notes</td>
</tr>
<tr>
<td>B.1.2</td>
<td>parts</td>
<td>copying paragraphs, keywords</td>
<td>just the notes</td>
</tr>
<tr>
<td>B.1.3</td>
<td>parts</td>
<td>highlighting, making schemas</td>
<td>just the notes</td>
</tr>
<tr>
<td>B.1.4</td>
<td>parts</td>
<td>highlighting</td>
<td>just the notes</td>
</tr>
<tr>
<td>B.1.5</td>
<td>both</td>
<td>summarizing paragraphs</td>
<td>yes/yes</td>
</tr>
</tbody>
</table>
Table C.9: Studying and learning habits overview (1).

<table>
<thead>
<tr>
<th>Intv.</th>
<th>Avg. Course Duration</th>
<th>Exams</th>
<th>Work Alone/Groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>B.2.1</td>
<td>5-6 courses/year</td>
<td>per course</td>
<td>both</td>
</tr>
<tr>
<td>B.2.2</td>
<td>1 month</td>
<td>per course, progress,</td>
<td>groups</td>
</tr>
<tr>
<td></td>
<td></td>
<td>patient examinations</td>
<td></td>
</tr>
<tr>
<td>B.1.1</td>
<td>4 weeks</td>
<td>per course, progress</td>
<td>alone</td>
</tr>
<tr>
<td>B.1.2</td>
<td>6-8 weeks</td>
<td>per course</td>
<td>both</td>
</tr>
<tr>
<td>B.1.3</td>
<td></td>
<td>per course, progress</td>
<td>both</td>
</tr>
<tr>
<td>B.1.4</td>
<td>6 weeks</td>
<td>per course, progress</td>
<td>groups</td>
</tr>
<tr>
<td>B.1.5</td>
<td>projects</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table C.10: Studying and learning habits overview (2).

<table>
<thead>
<tr>
<th>Intv.</th>
<th>Group Size</th>
<th>Educational Software</th>
<th>Peer Contact</th>
</tr>
</thead>
<tbody>
<tr>
<td>B.2.1</td>
<td>N/A</td>
<td>Blackboard, lecture DVDs</td>
<td>phone, Facebook</td>
</tr>
<tr>
<td>B.2.2</td>
<td>N/A</td>
<td>Pdrug, Compas, Nestor, Progress</td>
<td>E-learning platforms, mailing list</td>
</tr>
<tr>
<td>B.1.1</td>
<td>N/A</td>
<td>N/A</td>
<td>email</td>
</tr>
<tr>
<td>B.1.2</td>
<td>2-3</td>
<td>anatomy quizzes</td>
<td>email, Facebook</td>
</tr>
<tr>
<td>B.1.3</td>
<td>4-8</td>
<td>EPAS</td>
<td>Whatsapp, Facebook</td>
</tr>
<tr>
<td>B.1.4</td>
<td>3-4</td>
<td>Blackboard, NEJoM quizzes</td>
<td>Blackboard, MSN, Facebook</td>
</tr>
<tr>
<td>B.1.5</td>
<td>3-4</td>
<td>Blackboard</td>
<td>email</td>
</tr>
</tbody>
</table>

Table C.11: Gaming habits overview.

<table>
<thead>
<tr>
<th>Intv.</th>
<th>Type</th>
<th>Medium</th>
<th>Single/ Multi Player</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>B.3.1</td>
<td>role-playing, strategy, puzzle, social media</td>
<td>PC, PlayStation mobile</td>
<td>both</td>
<td>5h/week</td>
</tr>
<tr>
<td>B.3.2</td>
<td>puzzle</td>
<td>mobile</td>
<td>N/A</td>
<td>very little</td>
</tr>
<tr>
<td>B.3.3</td>
<td>no</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B.2.1</td>
<td>mixed</td>
<td>Nintendo, mobile</td>
<td>multi</td>
<td>once a week</td>
</tr>
<tr>
<td>B.2.2</td>
<td>mixed, mobile games</td>
<td>PC, mobile</td>
<td>both</td>
<td>in free moments</td>
</tr>
<tr>
<td>B.1.1</td>
<td>role-playing, mobile games</td>
<td>PC, mobile</td>
<td>multi</td>
<td>2-3h/week</td>
</tr>
<tr>
<td>B.1.2</td>
<td>no</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B.1.3</td>
<td>social games</td>
<td>console, PC</td>
<td>multi</td>
<td>0.5-3h/day</td>
</tr>
<tr>
<td>B.1.4</td>
<td>puzzle, card games</td>
<td>iPad</td>
<td>single</td>
<td>0.5h/day</td>
</tr>
<tr>
<td>B.1.5</td>
<td>strategy, mobile games</td>
<td>PC, mobile</td>
<td>single</td>
<td>20-30min/day</td>
</tr>
</tbody>
</table>
### Table C.12: Experience with medical competitions overview

<table>
<thead>
<tr>
<th>Intv.</th>
<th>Known Competitions</th>
<th>Competitive Interests</th>
</tr>
</thead>
<tbody>
<tr>
<td>B.2.1</td>
<td>no</td>
<td>anonymous competitions</td>
</tr>
<tr>
<td>B.2.2</td>
<td>no</td>
<td>general medicine</td>
</tr>
<tr>
<td>B.1.1</td>
<td>Medical Olympiad, practicing first aid</td>
<td>general medicine</td>
</tr>
<tr>
<td>B.1.2</td>
<td>inter-university competition, pharmaceutical competition</td>
<td>general medicine, STDs</td>
</tr>
<tr>
<td>B.1.3</td>
<td>patient case solving</td>
<td>general medicine</td>
</tr>
<tr>
<td>B.1.4</td>
<td>quiz night</td>
<td>general medicine</td>
</tr>
<tr>
<td>B.1.5</td>
<td>quiz night</td>
<td>no</td>
</tr>
</tbody>
</table>

### Table C.13: Interest in a crowdsourced application (1).

<table>
<thead>
<tr>
<th>Intv.</th>
<th>Topics to Incorporate</th>
<th>Tasks to Solve</th>
<th>Time to Participate</th>
<th>Alone/Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>B.2.1</td>
<td>diagnosis making, indentation of lab results, guideline for patient evaluations</td>
<td>testing knowledge, clue finding</td>
<td>once a week</td>
<td>alone</td>
</tr>
<tr>
<td>B.2.2</td>
<td>medical decision making</td>
<td>problem solving, clue finding</td>
<td>in free moments/as a mobile app</td>
<td>both</td>
</tr>
<tr>
<td>B.1.1</td>
<td>general medicine, surgery, others</td>
<td>problem solving</td>
<td>1-2 h/week</td>
<td>alone</td>
</tr>
<tr>
<td>B.1.2</td>
<td>general medicine, gynecology</td>
<td>diagnosis making</td>
<td>1 h/week</td>
<td>both</td>
</tr>
<tr>
<td>B.1.3</td>
<td>medicine, microbiology</td>
<td>problem solving</td>
<td>1-2 h/week</td>
<td>both</td>
</tr>
<tr>
<td>B.1.4</td>
<td>general topics</td>
<td>diagnosis making</td>
<td>a couple of minutes per session</td>
<td>both</td>
</tr>
<tr>
<td>B.1.5</td>
<td>almost any domain</td>
<td>reading/summarizing text, clue finding</td>
<td>40 minutes/day</td>
<td>alone</td>
</tr>
<tr>
<td>Intv.</td>
<td>What would make this useful to you?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>------------------------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B.2.1</td>
<td>preparation before working on the ward</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| B.2.2 | low-key mobile app  
helpful for studying |
| B.1.1 | problem solving  
practical application of mandatory readings |
| B.1.2 | a fun way to keep up with your medical knowledge |
| B.1.3 | connection to studies/research |
| B.1.4 | a fun way to broaden my knowledge |
| B.1.5 | fun  
keep up to date with what is going on in my field  
find interesting articles in my field |
| B.3.1 | summarize knowledge about a topic  
fun |
| B.3.2 | getting CME credits  
fun |
| B.3.3 | find related work in my field  
paper annotation tool |
Appendix D

“Dr. Detective” Game Evaluation Form Answers

Q1. Were the instructions of the game clear enough?

Please, give us some details on what was not clear in the instructions. What should we change?

A1: Maybe some of the categories can be more specified.... Which clues can be selected for diseases for instance? Is that all clues that can aid in the “diagnostic process”, including ” both negative” (e.g. “no lymphadenopathy”) as well as positive (e.g. “lymphadenopathy”) clues or test results etc etc??

A2: At first, I thought that we should be annotating the whole passage for every term, then I realized that the passage should be annotated for the disease that was being presented before the passage. This should be made more clear–should every term possible be annotated regardless of whether it relates to the disease or case presented? Also, I thought that the definition for the term factors should be more clear–does it include results of lab tests? is it anything that is important in the case? I do not see an option to choose the difficulty of the game.

A3: In the first game the task was to highlight the terms relevant to the specified diagnosis, however it appeared that all medical terms should be categorized. Therefore i would suggest rephrasing the task to indicating all factors that should be considered
to come to a diagnosis and exclude differential diagnosis. In Step 1: I found the term Factors and Medical tests overlapping: e.g. normally one would include Blood pressure and temperature in the physical examination and one would expect it to belong to Factors (for these were all observations) However one needs a thermometer or blood pressure device to measure these. Therefore I would suggest to specify the following categories: 1. symptoms 2. history (including intoxication and previous diseases) 3. medical test specified as the actual test and not the result

A4: Directions a pretty clear. It is hard to understand exactly how the points are earned. Is it the exact span of words? Are point earned when we select words such as “the”? If we select words that do not contribute to determining the diagnosis do we lose points? What’s to stop the player from just selecting nearly every word in the case to maximize the chance of earning the most points?

Q2. Was the overall speed/tempo of the game OK for you?

Figure D.2: Answers to Q2 in the game questionnaire

Please tell us, why it was OK or not OK for you. What should we change?

A1: The speed was fine.

A2: Each case report does not take that long to read and answer. Some case reports involve diseases which I was not familiar with, so I would take 2 min to read about it before searching for factors. This slowed down the game a little.

A3: The cases are short enough and the points are earned quick enough to keep it moving at a good pace.

A4: The first game was more interesting for one saw the points immediately after submitting the terms. In the second game it was a bit boring for there was no feedback on submitting the terms.

A5: I guess the speed was okay!
Q3. Did you find the game levels in alignment with your expectations?

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>17%</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>33%</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>33%</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>17%</td>
</tr>
</tbody>
</table>

Figure D.3: Answers to Q3 in the game questionnaire

Please, tell us why do you think so. Do you have suggestions how would you like it to be?

A1: The game was not too challenging, however it would have been better if the passages were more in line with the disease/cases. Some of the passages were vague and I wasn’t sure how they fit exactly with the case/disease presented.

A2: I could not find the ability to change the game level.

A3: Very interesting game... Can imagine this can also be used for training purposes as part of medical studies etc etc

A4: I did not perceive much difference between the levels apart from the length of the case description. Correlate the diagnosis to the case description and use more rare cases/diagnosis for higher levels

Q4. Were the game levels helpful at estimating the effort you needed to solve a case report?

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>33%</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>33%</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>33%</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0%</td>
</tr>
</tbody>
</table>

Figure D.4: Answers to Q4 in the game questionnaire

What (if anything) would you change about the game levels?

A1: I didn’t get the difference between a Quick game and a Normal game. I thought the harder game just had a longer passage than the Normal game.

A2: if effort is time duration yes. see 3
A3: The game levels did not impact the game all that much. The hard games just seem to have longer passages, but aren’t necessarily harder. Also, the same medical topic was used for multiple game levels, so that made each of the levels appear to have the same difficulty. Some medical topics are more rare or are harder to understand. These should appear in only the hard section, whereas a disease such as the flu or pneumonia should be included in the quick or normal games.

A4: See above

Q5. When solving a task, was it useful to look at the answers of other players?

Can you explain why? What else would be helpful here?

A1: I liked seeing others answers because it validated my responses. I didn’t like that you saw the answers at the end after you had submitted all of your answers—I wished I would have seen others responses one by one as I clicked mine

A2: Helped somewhat, but still unclear how points are earned and whether choosing most of a span or words has the same impact of choosing a longer span of words.

A3: at the beginning it helped to see what was meant by the category term (step 1) Assuming that other players correctly interpreted this.

A4: Did not always use these.... I wonder whether the idea is to first fill out all your answers and then compare them to other players, or compare them “whilst playing” .... I guess the later might also be useful because you can also have staff meetings etc etc in practice??! (but then when you’re doing consultations you might see someone by yourself: maybe different modules, e.g. “staff meetings” and “(individual) consultations” or anything could be developed?? I guess the current set-up where some features are “optional” is also useful!

A5: I did not see any other players answers while playing the game. This would be helpful to see in order to understand how I was awarded points and what I can do to earn more points on on future cases.
Q6. When you saw the answers of other players, were you more inclined to select the same answers or look for others?

![Pie chart showing survey responses]

- Yes, I only selected other people’s answers: 0 (0%)
- Sometimes, I changed my answer to somebody’s else answer: 0 (0%)
- Sometimes, I added also the answers of others: 2 (40%)
- No, I only selected my own answers: 0 (0%)
- It stimulated me to search for new (other) answers: 2 (40%)
- Other: 1 (20%)

Figure D.6: Answers to Q6 in the game questionnaire

What (if anything) would you change about the interaction with the answers of other players?

A1: I don’t know how to add other answers after I saw the answers of other players? I didn’t know that was possible. It would be nice if I saw others answers before I submitted mine.

A2: See the above.... I guess both looking as well as not looking at other answers might be interesting.... Especially when it comes to final diagnostic/therapeutic decision making where you might have group/staff meetings as well as (individual) consultations...

Q7. Did seeing the High Scores board motivate you to perform better in the game?

![Bar chart showing survey responses]

- 1: 0 (0%)
- 2: 1 (17%)
- 3: 2 (33%)
- 4: 1 (17%)
- 5: 2 (33%)

Figure D.7: Answers to Q7 in the game questionnaire (1)

Can you explain why? What would you like to change/add here?

A1: I’m a competitive person, so seeing other people’s scores motivated me to play more.

A2: I didn’t really see the high score board I guess.... I was, occasionally, looking at the score, however... I was, however, curious about the “scoring process”... Since you might get points if others choose the same answer or agree with yours, might that mean that if you’re the “first” to work on a case you might potentially automatically get more points (because of relatively more people playing “after” you than in case of other players...)

A3: I wanted to do more to raise my score when I saw my rank.
**“Dr. Detective” Game Evaluation Form Answers**

**A4:** I did not see a high scores board while playing the game

When others voted on your answers, was this motivating for you to perform better in the game?

![Graph showing responses to Q7](image)

Figure D.8: Answers to Q7 in the game questionnaire (2)

Can you explain why? What would you like to change/add here?

**A1:** It validated my answers.

**A2:** I did not see this - however I’m sure it would have motivated me to perform better.

**A3:** This did not happen as far as I am aware.

**Q8. Was it clear how the scoring works?**

![Graph showing responses to Q8](image)

Figure D.9: Answers to Q8 in the game questionnaire

What (if anything) would you change about the score and scoring methods?

**A1:** Please see the above, at question 7....

**A2:** I assume that having a better correlation to other players improved the scoring, but I am not sure.

**A3:** I’d like more info about the scoring methodology and how the software calculates the scores.

**A4:** I didn’t know how much my answers were worth exactly and how the score was calculated
Q9. How much did you enjoy this game?

Can you tell us which features did you like most and why?

A1: It seem like I did not see many of the features that would make this game more enjoyable.

A2: I liked it a lot... I can potentially see something like this being used in medical training.... Maybe “positive” as well as “negative” clues (pro or contra a certain diagnosis) could be added then as well??

A3: I liked that it felt like the annotation had a meaning for it— that it wasn’t just annotating everything, it was trying to annotate to solve a case.

A4: it is new and probably due to the immediate scoring you want to perform better.

Q10. Would you play this game again?

Can you tell us why?

A1: 1. to improve the score 2. playing the game is not wasting time for you know it will improve future systems 3. ?

A2: Ver interesting.... It’s reading medical texts and then picking out certain features....
A3: I don’t know how beneficial it is yet for learning clinical scenarios because I didn’t feel like the passages were specific enough. I would play again to help Watson and to further Watson’ s learning, but for my own sake I don’t think it was that beneficial.

Q11. Please, let us know if you have any other comments, suggestions, or bug reports that would help us improve the game further?

A1: in the first game in the quick level, domain Primary care hospitalist etc: step 1 has no drop down list and therefore cannot be played also the button to next diagnosis does not work second game when skipping step 3 and proceeding immediately to step 4 an error page comes up (i.e. page not found)

A2: It would be more challenging for doctors if they would be presented the whole medical case presentation. In the end, players should be shown a detailed scoring report and perhaps the final resolution (outcome) and some comments/discussion on the case study.

A3: I guess the most important things deal with the categories/clues and their definitions (e.g. what might fall under a clue and what not? Are *diseases* indeed definite diagnoses or “positive” tests or can “negative” information (test result normal, clean X-ray) be included as well.... These also aid in the diagnostic process because they show that something is “absent” (making potential diagnoses less likely) Typically, I guess there might be a “differential diagnosis” of several potential diagnoses, which you aim to narrow down in the diagnostic process.... Related to this, would medical tests for instance point to the test alone (e.g. chest X-ray) or the complete test result, e.g. “the physical exam was normal”.... A diagnosis (e.g. viral infection, diabetes) could then be a disease and the complete “testing process” (plus result) the *medical test*.... I guess currently it’s (also) very good though and really interesting!!!

A4: There were some bugs in the game portion–for a few other cases it wouldn’t let me select any clues when I opened the case. The drop-down menu for clue type was empty. I also didn’t like how I couldn’t change my answers once I submitted my clues. It would be better if the clues/annotations could be saved before submitting so that you can move back and forth between the types of clues before submitting for your final answer. I found that sometimes I missed something and then I couldn’t go back to add it.

A4: Many of the elements of the game were missing when I played the game as I described above.

A5: I completed the simple game very quickly and did not enjoy it as much as the other game. I will continue to play and provide additional reviews.