Clinical Relevance of the Doctor's Dilemma Question Set

Daniel R. Schlegel, Sashank Kaushik, Peter L. Elkin
Department of Biomedical Informatics, University at Buffalo, Buffalo, NY

Introduction

Every year at the annual ACP meeting, residents compete in the Doctor's Dilemma (DD) competition [1] — essentially medical Jeopardy. The asked questions cover topics in all of medicine and range from simple trivia to complex diagnosis and treatment decision questions. Questions are split into 26 categories including topic areas such as *History* and *Poisoning*; and entire medical subfields such as *Oncology* and *Cardiology*.

Many computational systems attempt to answer clinical questions. For example, there are a great number of decision support tools which make use of patient data and answer a specific set of questions. There are fewer systems which attempt to answer all types of clinical questions, our eventual goal. In developing such a system, it is important to have a set of questions to train on. Such questions must be: topically diverse, covering many different topics and scenarios in medicine; and *clinically relevant* — questions should be those which could occur to a doctor during clinical practice. Questions may be those a doctor should know the answer to, or they may be ones normally looked up. Even a cursory glance shows that the DD questions cover many topics and scenarios, but it is unclear how many are clinically relevant — the subject of this study. Only the IBM Watson team has previously made use of these questions in developing a question answering system. They only used diagnosis questions, which are clearly clinically relevant [2].

Methodology

The DD question set was obtained through the ACP website and communication with the ACP.[1] As of February 2015 it consists of 1110 questions. Of those, 171 refer to images and were excluded since we are only interested in textual questions. The remaining 939 questions were divided randomly into two datasets — with one to be used in later projects and not examined by our group during this study. From the remaining 465 questions in our dataset, we attempted to extract at least 10 questions from each of the 26 categories for manual review. This was not always possible since at least 10 questions were not available for each category. The result was a sample of 229 questions. Two clinicians, SK and PLE, examined and annotated each question as either clinically relevant, or not. The two annotators were in near–perfect agreement ($\kappa = .87$). The disagreements were mediated and a consensus was formed.

Results

Very little of the DD question set was found to be not clinically relevant — only 8 of the 229 questions (3.5%) in our sample. The primary category where irrelevant questions occur is *History*. Out of the 10 questions from that category, the annotators found that 5 were not relevant clinically. Other irrelevant questions came from the *Ethics* category (1 of 6 found irrelevant), *Biostatistics* (1 of 6), and *Epidemiology* (1 of 10).

Discussion

Once the results were tabulated, we analyzed the question categories which contained questions which were not clinically relevant in more detail. We found that the *History* section tends to contain a large amount of trivia. Questions such as "Nobel awardee for discovery of insulin" were labeled irrelevant since they aren't helpful in clinical practice, while others such as "Previous name for reactive arthritis" may be relevant to a clinician, even though it is historic data. From examining the *Ethics* and *Biostatistics* categories, it was easy to see that they may contain non-clinical questions such as those about IRB approval and certain equation usage. It was surprising to see a question from *Epidemiology* in the irrelevant category, but the question: "Greek letter that represents the probability of making a type I error", would probably have been better categorized as *Biostatistics*. Since the DD questions are mostly clinically relevant, we believe the set is a good tool for the development of clinical question answering systems.

References

[1] Doctor's dilemma competition — ACP. `http://www.acponline.org/residents_fellows/competitions/doctors_dilemma/`. Accessed: February 20, 2015.

[2] D. A. Ferrucci, A. Levas, S. Bagchi, D. Gondek, and E. T. Mueller. Watson: Beyond jeopardy! *Artif. Intell.*, 199:93–105, 2013.

[1] We would like to thank the ACP for allowing us to use the Doctor's Dilemma questions in this research.