PRACTICE-BASED EVIDENCE IN MEDICINE: WHERE INFORMATION RETRIEVAL MEETS DATA MINING

KARIN M. VERSPOOR

1. Introduction

A new approach in medical practice is emerging thanks to the increasing availability of large-scale clinical data in electronic form. In *practice-based evidence* [5, 6], the clinical record is mined to identify patterns of health characteristics, such as diseases that co-occur, side-effects of treatments, or more subtle combinations of patient attributes that might explain a particular health outcome. This approach contrasts with what has been the standard of care in medicine, *evidence-based practice*, in which treatment decisions are based on (quantitative) evidence derived from targeted research studies, specifically, randomised controlled trials. Advantages of consulting the clinical record for evidence rather than relying solely on structured research include avoiding the selection bias of the inclusion criteria for a clinical trial and monitoring of longer-term outcomes and effects [5]. The two approaches are, of course, complementary — a hypothesis derived from large-scale data mining could in turn form the starting point for the design of a clinical trial to rigorously investigate that hypothesis.

Information retrieval can play an important role in both approaches to collecting medical evidence. However, the use of information retrieval methods in collecting practice-based evidence requires moving away from traditional document-oriented retrieval as the end goal in itself, to viewing that retrieval as an intermediate step towards knowledge discovery and population-scale data mining. Furthermore, it may require the development of more context-specific retrieval strategies, designed to identify specific characteristics of interest and support particular tasks in the medical context.

2. IR and Evidence-based Practice

In evidence-based medicine, collection and meta-analysis of the published literature of clinical trials form the foundation of *systematic reviews* (e.g., Cochrane Reviews [1]). The production of such reviews has traditionally been done using painstaking exhaustive searches of the literature and human synthesis of published experimental results. It has been argued that automation is both necessary and possible [2, 7]. There is a clear role for information retrieval in this process, to identify publications relevant to a given review, although further structuring of the information within the documents retrieved is also needed [3].

A number of targeted search engines for the published biomedical literature have been developed that aim to improve search effectiveness for biomedical researchers [4]. Several incorporate the results of information extraction, such as named entity recognition for specific relevant entity types (e.g., drugs and diseases), with the objective of enabling concept-based indexing of the literature.

3. IR and Practice-based Evidence

Data mining of electronic health records for medical evidence demands processing of the wealth of clinical data now recorded in natural language text. Transformation of this unstructured data into a structured representation is needed for incorporation of the information it contains into broader data mining. Many transformations can be cast as information retrieval tasks: for instance, identifying patients satisfying particular profiles (e.g., for recruitment into clinical trials or registries), or retrieval of case histories corresponding to specific treatment protocols. Development of general approaches to such tasks will likely require a mix of information retrieval and domain-specific information extraction.

4. Conclusion

The boundaries between information retrieval, information extraction, and data mining are blurring; bringing them together, in an activity commonly referred to as *text mining*, can result in heterogeneous methods that will enable sifting through the entirety of the clinical record, including both its unstructured and structured components. This in turn will enable clinical decision making based on data derived from large populations in the "laboratory" of the natural world.

References

[1] Cochrane Collaboration. `http://www.cochrane.org`.
[2] T. Guy et al. The automation of systematic reviews. *BMJ*, 346, 2013.
[3] S. Kim et al. Automatic classification of sentences to support evidence based medicine. *BMC Bioinformatics*, 12(Suppl 2):S5, 2011.
[4] Z. Lu. Pubmed and beyond: a survey of web tools for searching biomedical literature. *Database*, baq036, 2011.
[5] T. Pincus and T. Sokka. Evidence-based practice and practice-based evidence. *Nat Clin Pract Rheum*, 2(3):114–115, 2006.
[6] N. H. Shah. Mining the ultimate phenome repository. *Nat Biotech*, 31(12):1095–1097, 2013.
[7] I. Shemilt et al. Pinpointing needles in giant haystacks: Use of text mining to reduce impractical screening workload in extremely large scoping reviews. *Research Synthesis Methods*, 2013. online preprint.