Challenges and Communities of Medical Informatics Research

Vagelis Hristidis
Computer Science and Engineering
UC Riverside
vagelis@cs.ucr.edu

ABSTRACT

This article discusses experiences and lessons learned from working on health informatics research as a computer scientist. In particular, I present challenges faced when conducting research on medical informatics, and explain some of the aspects that make medical data and systems unique. Then, I present the two broad research communities studying medical informatics problems. Finally, I offer advice on how to bridge the gap between these communities and increase their research productivity.

1. CHALLENGES FOR COMPUTER SCIENTISTS WORKING ON HEALTH-RELATED PROBLEMS

My background: I have Computer Science (CS) background, with expertise in Databases and Information Retrieval. I have been regularly attending CS conferences in this area like ACM SIGMOD, VLDB and ACM WSDM. About six years ago I got interested in Medical Informatics (*MedInf*), because I saw that my research could be applied in this area. I started building collaborations with medical, nursing and public health researchers, and attending MedInf conferences like AMIA and the recently founded ACM SIGHIT.

I first want to share my experiences on the barriers for CS researchers who want to get involved with MedInf. First, one has to establish collaborations with medical experts, which often means researchers with MD degree, who have very limited available time. This is challenging because a research topic that sounds intriguing to a CS researcher may be of little value to an MD researcher and vice versa. For example, building a classifier that given an EKG time series decides if a patient is at risk of cardiac arrest, for a specific patient population (e.g., young adults), may sound like an intriguing topic for an MD researcher, but sounds as a simple application of existing data mining algorithms for a CS researcher. As another example, a few years ago I was visiting a hospital clinic and I was discussing with physicians (with excellent PubMed record) on research collaboration opportunities. One of them was excited and said: "I would like to be able to see how many patients in my database are diagnosed with a specific disease grouped by year, race, and so on. " That is, this physician needed OLAP (Online Analytical Processing) functionality on top of his data, which is clearly useful, but a Computer Scientist would not find it interesting. On the other hand, I recently met with another physician and was trying to convince him to join our project on automatically annotating textual clinical notes. His reaction was the opposite from enthusiastic. He said: "I never look back at the text of clinical notes of past patients, but only look at their past vital signs which are numeric structured data. So, why would I care to annotate textual notes?" Obviously, annotating complex text data using rich ontologies sounds like an intriguing CS project. Such interdisciplinary collaborations need patience and compromise, or else they will be short-lived.

Another challenge is that most useful projects require some form of user study of medical experts, or even worse, of patients. It is easy to find hundreds of survey subjects in Amazon Mechanical Turk paying 20 cents each, but finding even 3 MDs for a user study is hard. It is not uncommon for the setup and execution of a user study to take longer than the rest of the research. Interviewing patients or accessing patient data is exponentially harder, due to privacy constraints and Institutional Review Board (IRB) approval requirements. Can a junior CS researcher afford such delays, when the number of publications is critical?

2. WHAT IS UNIQUE ABOUT MEDICAL DATA AND SYSTEMS?

When I discuss about MedInf to colleagues in CS conferences, specifically Database conferences, I often get the reaction that there is nothing unique about medical data, since they can be viewed as dirty, heterogeneous, semi-structured, spatiotemporal and

Table 1: Main Differences between the two Communities.

	CS-MedInf	**Med-MedInf**
Representative Publication Forums	MedInf Tracks or Workshops in CS Conferences, ACM SIGHIT	AMIA, HIMSS, IMIA, BMC Med. Inf. & Dec. Making
Typical Researchers' Background	CS	Healthcare professionals with CS/IT interest or education
Funding agencies	NSF, Computer Industry	NIH, Healthcare Foundations
More prestigious forum	Conference	Journal
Paper content	Equal length describing methods and experiments	About one page describing methods and several pages on experiments
Prototype systems	Public prototypes are uncommon	Robust prototype systems are common
Opinion of other community	Med-MedInf papers are technically shallow	CS-MedInf papers don't understand intricacies of medical requirements
Researchers' Nationality	International	International, but much larger percentage of domestic members
Conference dress code	Jeans	Dress pants or suit

multimodal. Many of the key challenges on medical data like data integration or privacy-preserving querying and mining have been on the agenda of CS conferences for decades. This perspective can be generalized to other medical informatics areas like health systems engineering, architecture of medical devices, or connecting medical devices over networks.

In my opinion, some of the unique challenges and opportunities of working in medical informatics, from the perspective of a CS researcher (with some bias towards data management research), are:

(a) *The rich set of medical ontologies and dictionaries publicly available*, mostly thanks to the US National Institute of Health (*http://www.nlm.nih.gov/research/umls/*). This is also supported by Mussen [2], who identifies research on biomedical ontologies as one of the two key areas where medical informatics research can be viewed as core CS research; the other being problem-solving methods. Yes, there are ontologies in other areas, but they don't come close to the size and richness of the manually curated biomedical ontologies (notice my emphasis on "manually curated", since there has recently been work on automatically generating large Semantic Web ontologies). These ontologies can be leveraged in a wide range of problems, from search to data mining, information extraction, Web services and Natural Language Processing.

(b) *The complex workflows of how medical data and systems are being used must be taken into consideration.* For instance, an algorithm that looks for mistakes in clinical notes must account for the heavy copy-pasting, heavy use of abbreviations, motivation of users to get the billable concepts right, relationships to other elements of the health record of that patient, and the fact that many physicians use transcription to record clinical notes. Understanding these intricacies allows formulating problems that are challenging and interesting for both CS and healthcare researchers.

(c) *Understand the profile, background and goals of the users of medical informatics systems.* For instance, nurses can process a different set of concepts than physicians, and have generally more time to spend per patient than physicians. As another example, assume one builds a powerful and effective system to annotate and add structure to clinical notes. How can we motivate physicians to use it? Sure, by capturing structured data we enable querying and data mining. But the physician, who wants to see as many patients as possible per day, may not see any direct benefit to spend one extra minute per patient. If the proposed system would also automatically generate the billing codes of a patient's visit, this would potentially motivate a physician to give it a try. As a general rule, anything that may lead to increased healthcare cost is viewed with great skepticism, even if it may potentially improve the quality of care.

3. WHICH ARE THE RESEARCH COMMUNITIES OF BIOMEDICAL INFORMATICS?

One can identify two distinct communities that study MedInf problems. First, the CS-MedInf community consists generally of people like me, who are looking for interesting CS problems in the medical domain. Then, is what I call the Med-MedInf, which generally consists of healthcare professionals (e.g., nurses, MDs) with interest and/or education in CS or IT.

Researchers from the two communities have different mindsets on what constitutes research. CS-MedInf researchers are interested in computationally sophisticated methods that have the potential to improve healthcare, whereas Med-MedInf researchers are looking for evidence that (often simple) computing solutions improve healthcare. Hence, the objectives and writing style of publications is very different, which also means that the learning curve to switch from the one community to the other is steep.

Furthermore, CS-MedInf publications appear in a very wide range of forums, from tracks of CS conferences to specialized CS-MedInf forums like SIGHIT. A query on the ACM Digital Library for publications that contain the word "medical" in their abstract returns 2,460 results as of September 20th 2012. The same query on IEEE Xplore Digital Library returns 24,485 results (14,208 if we exclude the Bioengineering topic). The numbers are much higher if we include articles with this word in their body or if we search for other related keywords. Hence, it is very hard for Med-MedInf researchers to follow this literature. The other direction is less challenging, since Med-MedInf work almost always appears in dedicated MedInf forums like AMIA, and not in other medical journals.

In Table 1, I am trying to summarize the main differences between the two communities.

4. IS THERE ANYTHING WRONG WITH THE COMMUNITIES' SEPARATION?

Yes, in my opinion the fragmentation of the MedInf community may cause decreased research output and impact. In particular, CS-MedInf researchers often spend their time to devise algorithms and evaluate their time performance for medical informatics problems that may sound interesting, but may not be of much practical use. For example, building a classifier to classify patients to male and female based on their clinical notes is of little use since this information is explicitly recorded in all medical records.

On the other hand, Med-MedInf researchers are often unaware of state-of-the-art algorithms or software packages developed by the CS community, and as a result may employ computationally suboptimal solutions or miss software reuse opportunities. For example, CS-MedInf researchers have created several algorithms to query Electronic Health Records (EHRs), e.g., [1], building on top of the rich CS literature on searching semi-structured data, published since 2002 (see [4] for a survey). However, this literature has not been leveraged (or cited) by the Med-MedInf community, who are building EHR search systems based on the much older Information Retrieval literature, which operates on unstructured text document, even though EHRs are semi-structured documents. On the other hand, the CS-MedInf community has not studied what kind of queries health professional use, nor has the excellent Med-MedInf paper on the analysis of clinical queries [3] been adequately cited in the CS-MedInf community.

5. SUGGESTIONS FOR THE FUTURE

Clearly, the creation of an increasing number of Biomedical Informatics departments in universities across the world has greatly helped the two MedInf communities come closer. The main idea of Biomedical Informatics departments is to hire some people with CS background and some with health-related background and make them work together, which has been successful. However, researchers from these departments eventually gravitate to one of the two communities; usually if the department is under the college of medicine then researchers gravitate to Med-MedInf forums, and vice versa. It may be beneficial to establish Biomedical Informatics departments as independent schools, not under any college.

Benchmarks and public datasets are a first step to level the playing field. For example, take the problem of measuring similarity between patients. If a set of EHRs is available, and so is a set of expert judgments on which pairs of patients are most similar, then any researcher can build and evaluate similarity estimation algorithms. A great example of this in the CS community was the Netflix Prize competition. Fortunately, there is a slow increase of EHR datasets that are publicly available, like MIMIC II (http://physionet.org/mimic2/) and i2b2 (https://www.i2b2.org/). However, little progress has been performed in terms of expert relevance judgments on public datasets.

Further, Med-MedInf forums should reach out to the CS-MedInf community, by adding tracks on the execution time performance for well-known health problems, and on new methods to solve benchmarked health problems. The other way is also important, that is, to attract Med-MedInf researchers to application tracks of CS-MedInf forums.

Finally, researchers from both communities must respect the knowledge and experience that the other side brings to the table, and see any interaction with the other side as an opportunity to learn something new, even if this interaction may not lead to successful research collaboration.

6. REFERENCES

[1] F. Farfán, V. Hristidis, A. Ranganathan, M. Weiner. XOntoRank: Ontology-Aware Search of Electronic Medical Records. IEEE International Conference on Data Engineering (ICDE) 2009

[2] M. A. Musen. Medical Informatics: Searching for Underlying Components., Methods Inf Med. 2002; 41 (1): 12-9

[3] K. Natarajan, D. Stein, S. Jain S, N. Elhadad. An Analysis of Clinical Queries in an Electronic Health Record Search Utility. Int J Med Inform. 2010 Jul.;79(7):515–522

[4] J. Xu Yu, L. Qin, L. Chang. Keyword Search in Databases. Morgan & Claypool Publishers 2010