Toward Next Generation
Integrative Semantic Health Information Assistants

Evan W. Patton and Deborah L. McGuinness
Rensselaer Polytechnic Institute
110 8th Street
Troy, NY 12180 USA

Abstract
Managing a complex illness often requires different treatment regimens spread over a long time. The complexity of these potentially life-threatening diagnoses can be daunting to patients while they are most vulnerable. We present a vision for artificial intelligence-enabled tools for assisting patients in managing the complex information given to them over the course of their treatments through the combination of existing and emerging techniques from natural language processing and knowledge representation. We provide examples from an actual breast cancer diagnosis and treatment plan and highlight the development of new combinations of techniques to build tools that can reason about data from a variety of sources and act as intelligence-augmenting agents. We conclude with a discussion of some additional challenges facing artificial intelligence practitioners as applications become patient-centric.

Introduction
Traditionally, artificial intelligence in medical applications has focused on improving the abilities of medical professionals to perform tasks such as diagnosis (e.g., Shortliffe 1986; Wyatt and Spiegelhalter 1991; Garg et al. 2005; Vihinen and Samarghitean 2008) or to aid in managing drug interactions (e.g., Bindoff et al. 2007) or side effects (Edwards and Aronson 2000, p. 1258). These efforts target users who have years of medical experience. In contrast, patients often have limited medical knowledge, and they may be coping with new life-threatening diagnoses that may require a number of time-sensitive decisions involving potentially complex treatment paths. Patients and their support teams may be ill equipped to make good choices, despite doctors providing them with relevant information. Additionally, according to Ong, De Haes, Hoos, and Lammes (1995), patients often fail to recall 50% or more of the important information given to them shortly after leaving a doctor’s office (for additional review, see Williams, Davis, Parker, and Weiss 2002). Simultaneously, many doctors have increasing pressure to be more efficient in their time spent with patients while trans-
tured content (i.e. nanopublications) from which reasoning systems can generate arguments and explanations for patients. We discuss some challenges that artificial intelligence researchers will have to address to make this vision a reality as well as some preliminary steps towards this goal. Lastly, we discuss how self-instrumentation and access to patient contextual information will bring more personalized healthcare and feedback to the patient’s fingertips.

**Motivation**

Cancer is one of the more complex diseases to manage and treat, often requiring chemotherapy, surgery, and drugs to reduce recurrence for years following treatment. We motivate our work using scenarios inspired from a co-author’s recent breast cancer journey along with her surprises at how enhancing patient-doctor interaction. In addition to the motivation for treatment. Mitigation strategies may help patients forgotten or not understood their doctors advice for preparing for treatment. Furthermore, a patient’s care team is composed of many individuals. For example, a breast cancer patient’s team may include some if not all of the following: general practitioner, medical oncologist, radiation oncologist, breast surgeon, plastic surgeon, physical therapist, occupational therapist, radiologist, pathologist, genetic testing counselor, medical nurses, and clinical trial nurses, each of which must uphold standards of care, make judgement calls, and interact with either the patient or other members of the care team. This forms an intricate communication network on which treatment outcome rests. They must work in the best interest of the patient, respect her wishes, and also consult other trained medical professionals to obtain input, such as second opinions on diagnoses or treatment plans. Some of this information must be disseminated to the patient and her support network, such as preparation strategies before chemotherapy or surgery, instructions on when to call a doctor or go to an emergency room, and instructions for taking post-therapy medication or recurrence reduction strategies. Information may be required at many times over the course of a treatment so it is even more important to have tools that address the fact that patients forget about 50% of the information given to them at the doctor’s office (Ong et al. 1995, p. 912).

**Scenario 1. Drug Adverse Effect Mitigation**

A service that provides adverse effect mitigation strategies along with evidence of effectiveness may be useful in helping patients understand treatment impacts along with coping strategies. This can also help patients actively engage in medical discussions. Such services may be particularly useful to patients with limited medical knowledge who may either have forgotten or not understood their doctors advice for preparing for treatment. Mitigation strategies may help patients better tolerate preferred treatment plans and thus improve
Table 1: A summary of search engine results for queries about nail problems related to Paclitaxel.

<table>
<thead>
<tr>
<th>Search engine</th>
<th>&quot;nail pain&quot;</th>
<th>&quot;nail changes&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>86,400</td>
<td>64,800</td>
</tr>
<tr>
<td>Bing</td>
<td>3,330,000</td>
<td>1,040,000</td>
</tr>
<tr>
<td>Yahoo</td>
<td>3,230,000</td>
<td>28,500,000</td>
</tr>
</tbody>
</table>

outcomes. Mitigation strategies may be found in standard medical literature but they may also be found in forums or social media where patients share anecdotal information about treatment journeys. This information can prove useful in scenarios where no best practice standard of care strategy exists in the medical literature for coping with particular side effects, but care must be taken to accumulate evidence and present it in an unbiased manner as well as encourage discussion with the patient’s care team. Consider these two variations on this scenario:

A patient is preparing for a paclitaxel/trastuzumab regimen. Her work requires extensive typing so she wonders if chemo will inhibit her ability to work. She uses an intelligent agent to discover that common side effects of paclitaxel include nail bed death, nail pain, and nail lifting. Using breast cancer forum content, the agent finds treatments shared by other patients and survivors and aggregates supporting evidence passages from those forum postings, which include strategies such as soaking nails in vinegar to reduce the chance of infection and the use of tea tree oil to prevent nail damage along with keeping the nails very cold during infusions to potentially limit the chemotherapy drugs from doing as much damage to the fast growing cells in the nail beds. The agent presents this content along with an explanation drill-down display that she can then present to her oncologist to discuss options that would not limit the effectiveness of chemotherapy yet may potentially improve her experience.

After a successful chemotherapy regimen and surgery to remove the tumor, the oncologist recommends that the patient begin taking Letrozole, which has been shown to decrease the risk of recurrence in hormone sensitive cancer patients by up to 40% (Gross et al. 2004). However, one side effect of Letrozole is decreased bone density. The patient, who is already osteopenic and concerned about developing osteoporosis, thus wishes to learn more about this treatment. The intelligent agent points out that calcium supplementation and in more severe cases, prescription bisphosphonates, may help. Her oncologist also recommends weight-bearing exercise to improve bone density.

In both scenarios, the patient desires more information about potential side effects of a particular drug. The number of side effects may be large. For example, SIDER contains 182 side effects for the drug Paclitaxel. For this patient, the side effects of “nail pain” and “nail changes” are relevant due to her work. To simulate the actions of the patient in our scenario, we performed web searches using popular search engines with the query string “breast cancer paclitaxel <side-effect> treatment.” The number of search results provided by each of the three major US search engines are summarized in Table 1, and are orders of magnitude greater than any human could process. Even if only the top ten results contain useful information, the patient still needs to correctly parse and reason about many arguments to gain insight into how a side effect affects quality of life. Thus, even shallow natural language processing of search results into coherent arguments may significantly reduce the patient’s burden. Additionally, there are often contradictory results where the patient must consider evidence for and against a certain action, as the following two quotations exemplify:

My onc[ology] nurse told me to rub tea tree oil into my cuticles and nails every night. It is a natural anti-septic and for whatever reason can sometimes help prevent nail infections and lifting during taxol. ²

I wouldn’t use tea tree oil. A friend did on some cracked skin and it got worse. ³

The first quotation is a suggestion from one author to the poster of the original question on nail pain, which is that tea tree oil prevents nail infections because “it is a natural anti-septic” and appeals to authority “my onc nurse told me to...”. The second quotation from a different user in the same thread advises against tea tree oil as “a friend [applied tea tree oil] on some cracked skin and it got worse.” Both quotations appear on the same page, so traditional information retrieval methods leave it to the user to appropriately determine the usefulness of information in the page’s context. Alternative ranking mechanisms that take into account the structure of forums, such as the hierarchical model presented by Ganu & Marian (2013), enable the decomposition of web pages to give multiple granularities of scores. The hierarchical model proposed by Ganu et al. is not deep, and thus it is an open question of how richer representation schemes could improve upon their ranking algorithms. Further, these models do not take into account competing views on the same topic as demonstrated by our example.

Natural language processing, on the other hand, might convert these sentences into a tree with entities typed and parts of speech identified, and present this information to the end user. However, this traditional approach forgoes potentially useful additional information, such as whether the patient has a medical history of or currently has dry or cracked skin. Alternatively, she might trust an “onc nurse” differently than a “friend” of an online stranger. Alternatively, if this were on a more structured social network such as Facebook or Google+ where accounts are often tied to real-world identity and if the post authors’ friends are medically literate, the reader may ascribe more trust in the friends of the author of the second quote. Each of these evidence graphs are presented based on an overall trust assessment made by an intelligent agent. How they are shown may vary, and might include a technique of displaying the two highest ranking ev-

²https://community.breastcancer.org/forum/69/topic/783573
³https://community.breastcancer.org/forum/96/topic/745475
Scenario 2. Context-Sensitive Medical Content Explanation for Non-Experts

Explanation is a key component of transparent systems and user studies have shown that explanations are required if agents are to be trusted and that users prefer explanations that are context-aware (Glass, McGuinness, and Wolverton 2008). We also believe that health assistants must incorporate explanation generation to build patient trust. One way to support explanation is through the tracking of provenance using standards such as the W3C’s PROV (Belhajjame et al. 2013). PROV is a standard for modeling provenance information on the web, which allows tools to integrate distributed provenance information from different systems. This provenance is necessary, but not always sufficient, for generating patient-oriented explanations, as we will discuss.

There are a number of considerations when preparing explanations for non-experts. For example, there is typically a need to take the patient’s knowledge into account particularly when considering how to present complex information. A medically literate patient may benefit from receiving detailed medical language so as to receive precise information while a medically naïve patient may need a more abstract description initially with supporting followup content available on demand. Further, the complexity of statements based on the patient’s background knowledge, the context in which they are receiving the explanation, and highlighting importance of certain pieces of information in the overall explanation, may all affect the patient’s ability to assimilate the explained knowledge. Explanation provides support for transparency by allowing the user to receive more information about how a particular set of evidence was arrived at and what information it depended on, such as can be found in their medical record or in personal records (e.g. a calendar). For accountability purposes, patients may also like to know what personal datasets were accessed to determine a set of potential actions. These records serve a dual use–to generate explanations for non-experts. For example, there is typically a need to take the patient’s knowledge into account particularly when considering how to present complex information. A medically literate patient may benefit from receiving detailed medical language so as to receive precise information while a medically naïve patient may need a more abstract description initially with supporting followup content available on demand. Further, the complexity of statements based on the patient’s background knowledge, the context in which they are receiving the explanation, and highlighting importance of certain pieces of information in the overall explanation, may all affect the patient’s ability to assimilate the explained knowledge. Explanation provides support for transparency by allowing the user to receive more information about how a particular set of evidence was arrived at and what information it depended on, such as can be found in their medical record or in personal records (e.g. a calendar). For accountability purposes, patients may also like to know what personal datasets were accessed to determine a set of potential actions. These records serve a dual use–to generate explanations for patients and evaluate compliance with privacy laws, such as the Health Insurance Portability and Accountability Act (HIPAA). Ongoing research and standardization in the areas of web identification and access control lists will further advance distributed query and explanation capabilities for patient-assisting technologies.

Consider the Letrozole example presented in the motivation. Letrozole is a drug prescribed after chemotherapy of hormone-sensitive cancers to reduce the risk of recurrence, and has been shown to do so up to 40% (Gross et al. 2004; Jakesz et al. 2005). Like many drugs, it comes with a long list of potential drawbacks, including risk of increased cholesterol, increased blood pressure, and decreased bone density, the latter of which can lead to osteoporosis. For patients who already have high cholesterol and high blood pressure, are at risk for heart disease, or who are osteopenic or osteoporotic, this drug may at first sound undesirable. However, helping patients find information and organize information prior to consulting their oncologist will make it easier for them to cope with managing risk-benefit trade-offs in treatments. Therefore, tools that decompose medical arguments into more approachable statements may aid patients in participating in their care. An example explanation may be presented as follows, with the ability for the patient to drill-down and receive explanations for individual statements:

1. Your cancer was tested and found to respond positively to estrogen (link to appropriate medical record segment)
2. Drugs like Letrozole inhibit your body’s natural ability to produce estrogen, thereby reducing something that could stimulate growth or regrowth of cancer. (link to aromatase inhibitor explanation)
3. Certain Letrozole side effects can be reduced by lifestyle changes, such as taking calcium supplements and weight-bearing exercise. (link to strategies for improving bone health in post-menopausal women)

Selecting a particular statement, such as point 2 above may either provide links to literature sources such as Gross et al. (2004) and Jakesz et al. (2005), or lead to more detailed, structured arguments, such as:

1. Estrogen may be synthesized in the body using a conversion process aided by the enzyme aromatase
2. Letrozole is an Aromatase Inhibitor drug
3. Aromatase inhibitors block the ability of aromatase to create estrogen
4. Studies (Gross et al. 2004; Jakesz et al. 2005) have shown that this leads to decreased risk of recurrence

Using explanation drill down options, the patient can obtain as much detail as she feels is necessary to understand a particular decision. Further, if such systems allow feedback on steps within the evidence chain, e.g. that a piece of evidence is difficult to understand or the patient feels irrelevant, then techniques such as relevance feedback (for review see Harman, 92) can be adapted to enable agents to learn and generate more robust explanations in the future.

Combining Natural Language Processing and Knowledge Representation to Create Next Generation Semantic Health Assistants

We are exploring new techniques for combining natural language processing with emerging knowledge representation. This combination has been used for learning (e.g. see Forbes, Riesbeck, Birnbaum, Livingston, Sharma, and Ureel, 2007) and natural language processing has been used to learn class expressions for vocabularies such as SNOMED-CT (e.g. Ma and Distel, 2013). However, our primary focus is on information exchanged on public websites combined with vetted literature. Our example focuses on complex disease settings with the goal of finding coping strategies and supporting and refuting evidence, thereby enabling informed doctor-patient conversations.

The patient on paclitaxel in our scenario is interested in learning about possible adverse effects and mitigation strategies. Using the chemotherapy orders, natural language processing algorithms can extract the drug names that are then
linked to databases such as the Unified Medical Language System (UMLS), DrugBank, and the Side Effect Resource (SIDER) database. Using SIDER, we can discover possible side effects and use this information to perform information retrieval on forum content to identify possible strategies. Knowledge representation techniques allow us to detect contradictory information and to construct an evidence graph that includes supporting and contradictory outcomes. These evidence networks can then be used to provide suggestions and explanations to both medically literate and medically naive users at appropriate levels.

Consider the BioNLP Shared Task (Nédellec et al. 2013), which is a task for natural language tools that revolves around processing documents containing biomedical content. For the 2013 BioNLP task, the best overall performing system was TEES-2.1, which finished first on 6 out of 10 tasks with F-scores ranging from 0.14 to 0.55. Recall on the tasks ranged from 0.12 to 0.49 and precision ranged from 0.18 to 0.82 (Björne and Salakoski 2013). Furthermore, Björne and Salakoski point out that previous versions of TEES obtained comparable F-scores (0.52 in 2009, 0.53 in 2011) on earlier iterations of the gene extraction task used in BioNLP. There are clear opportunities here to introduce new technologies to improve biomedical task performance, such as the use of structured content and other knowledge representation techniques.

Fig. 1 provides an exemplar showing how Natural Language Processing and Knowledge Representation techniques may be used to augment one another. The graph displays information modeled using the Resource Description Framework (RDF). Items on the right are document fragments with provenance and trust modeled using the Proof Markup Language 2 (McGuinness et al. 2007). The first two fragments are from a patient’s chemotherapy plan. UMLS is used to identify drugs such as paclitaxel and trastuzumab. Keyword search using NIH’s
Metamorphosys tool identifies that these two drugs form a Therapeutic or Preventative Procedure, identified as Monoclonal Antibody HER2-Paclitaxel. Using semantic similarity measures adapted to RDF graphs (Zheng 2014), we can identify additional mappings, e.g. Herceptin-Paclitaxel, with which Paclitaxel and Trastuzumab form the relation is_component_of_chemotherapy_regimen. This inference is further corroborated by linking to additional datasets, e.g. DrugBank, that state that Herceptin is a brand name of Trastuzumab. Information-rich, externally curated knowledge bases thus provide additional means of mapping and validating content as it is processed via natural language tools and enable us to reconcile content across documents. After identifying and connecting the chemotherapy drugs, we then extend our knowledge base through other structured content, including the Side Effect Resource (SIDER). Using traditional information retrieval techniques, we find forums containing posts discussing the drugs and their side effects and then use natural language processing to identify coping strategies. Lastly, since different users might experience conflicting results, the knowledge base is capable of modeling conflicting evidence, thus both positive and negative results may be included. By automating these tasks, the patient spends much less energy searching for content and can be more informed through aggregation and presentation of differing viewpoints.

**Beyond General Medical Knowledge**

Sensors provide additional opportunities for novel, collaborative artificial intelligence research. Self instrumenting patients are now collecting large amounts of data with potential applications during complex disease treatment. For example, prior to a diagnosis, a patient may use applications, e.g. MyFitnessPal, combined with sensors, e.g. Fitbit, to monitor activity and sleep while attempting to lose weight. After diagnosis, she is told by her oncologist to maintain her weight as the amount of chemotherapy drugs is dependent on weight. Additionally, weight loss may further weaken a body that is already being bombarded by the combination of the drug therapy and disease. Now, these sensors become a means by which intelligent agents can gather additional data that can inform context-aware lifestyle alterations that can be discussed with her care team to ensure the patient’s safety during chemotherapy and the sensors can help monitor progress toward those goals.

We are building an intelligent health agent with the ability to reconcile information from a patient’s medical record with detailed lifestyle information from other applications, such as a calendar or food tracking program, available through the patient’s mobile device. For example, by connecting to the calendar application, our agent will attempt to identify potential causes of adverse events in the patient’s data. Figure 2 contains an example where the patient’s blood work indicates blood glucose levels above the normal reference ranges. The agent’s knowledge base contains a rule that steroids can cause an increase in glucose. Looking back at the patient’s calendar, it observes that she had an event called “steroids for chemo” scheduled and that was corroborated by the chemotherapy plan extracted using NL techniques, so it infers a plausible link between the steroid premedication and the glucose increase. In cases where no determination can be found, it can highlight the events so that the patient can discuss them with a medical professional.

As technologies improve, we envision that new ways of tracking detailed information about an individual’s health will make it easier to personalize therapies. Consider our second scenario where taking a new drug is likely to accelerate bone loss. If we had some additional information about the patient’s lifestyle, such as that they are already eating a diet high in calcium (which can be obtained from programs such as MyFitnessPal) then the health assistant may suggest lower supplement amounts of calcium and instead focus more on recommending enhancing other relevant lifestyle modifications, e.g. increasing weight bearing exercises. If records show that the patient is already doing all of the standard of care lifestyle modifications, and bone density continues to decrease, then the health assistant may suggest a discussion with the patient’s doctor about more advanced interventions.
Repurposing Drugs with Semantics

Another project that explores future infrastructures and AI-enhanced health information assistants is our groups Re-purposing Drugs with Semantics (ReDrugS) project (McCusker et al. 2014). ReDrugS is intended for use in academic and pharmaceutical research to identify potential additional uses of previously approved drugs. However, the underlying infrastructure built for ReDrugS, such as its means of modeling massive amounts of medical data as nanopublications (Groth, Gibson, and Velterop 2010) and the probabilistic evaluation framework used to identify interesting potential links, may prove useful in a more general healthcare-oriented setting. ReDrugs ingests drug/protein, protein/protein, and protein/phenotype interactions content from 17 databases and then supports queries about how drugs and diseases may interact (through protein pathways). This allows people to look for diseases that a drug may connect to that may be a potential target for drug repurposing. It also allows the letrozole patient to query for the drug and see what disease conditions it connects to and through which pathways, thus allowing the patient to learn more about how it might be working in her body. This is a direct repurposing of existing content. The infrastructure however may be used to build a nanopublication store of content extracted from a patient’s medical record, along with the drugs in a treatment plan. Additional content is ingested from anecdotal evidence posted on forums can be modelled as nanopublications. If this kind of store is connected then the augmented store may be queried for potential coping strategies for anticipated drug side effects. One of the reasons this kind of infrastructure may be of interest is that it is designed to handle uncertain and potentially conflicting content using trust values and consensus probability support. Further it maintains detailed provenance so that users may always see where information came from, how it was manipulated, and what supports or refutes a particular statement.

Discussion

We argue that artificial intelligence practitioners can benefit from and likely need to cross subfield boundaries to design novel combination strategies from more than a single subfield to improve the impact and adoption of next generation assistants. The idea of computer augmentation of the intellectual capabilities of humans is not new (see Licklider, 1960; Engelbart, 1962), but in light of the limitations of patients to understand and recall medical information related to complex illnesses. We described two different motivating medical information scenarios in the domain of breast cancer, and highlighted how addressing these scenarios will require managing information in different ways and customizing how they are presented to the patient in context-sensitive ways—a challenge that requires manipulating information, both structured and unstructured, in a transparent manner. To achieve this vision, artificial intelligence practitioners must consider novel ways to combine and enhance existing technologies, particularly where existing techniques have plateaued in their effectiveness. We introduced two research efforts aimed at demonstrating future health aids including a prototype health information management assistant that combines best-in-class natural language processing techniques with knowledge representation to provide a tool for patients to manage their health data, better understand it, and to enable browsing and exploration of possible explanations of medical data measurements. We also introduced a novel framework for modeling drug, protein, and condition information as nanopublications, which can be used for probabilistic reasoning about drug interactions and discussed how its methodology could be leveraged and repurposed in broader medical settings.

Acknowledgements

Mr. Patton was supported by a NSF Graduate Research Fellowship. The authors thank Heng Ji and Alex Borgida for their discussions that helped shape this work.

References
