SF 424 Grant Proposal

“Novel machine learning and disambiguation methods and web application for cohort discovery and from free-text questions”

Principal Investigator: Nicholas Dobbins

University of Washington, Seattle, WA

Project Summary

The proposed work aims to use recent advances in machine learning using recurrent neural networks novel approaches in web scraping and web application development to understand clinical research questions from free-text and automatically generate database queries for researchers. Put another way, this project aims to create a Google-like user interface paired with advanced artificial intelligence methods to understand researcher questions to identify patients meeting arbitrary clinical criteria. All elements of the project code and architecture will be introduced into the public domain and open to use for future research.

Project Narrative

Researchers spend a significant amount of time simply cleaning data and learning tools to extract data before they are able to conduct their research of interest and disseminate it to the public. Recent advanced cohort discovery tools, such as i2b2 and Leaf, make this process simpler and less time-consuming but nonetheless have learning curves and still require significant user training. This project aims to create a suite of tools to understand researcher queries in natural language in order to quickly and efficiently return data and speed up the rate of research discovery.

Budget

Please see attached BudgetWorkbook document.

Biographical Sketch

Please see attached BioSketch document.

A. Significance

The process of cohort discovery and hypothesis generation is often the starting point for a vast number of clinical and biomedical research projects. For example, a researcher may wish to ask the question, “How many patients [at my institution] have had an ejection fraction below 40% with ACS presentation in the past year?”, or “How many patients diagnosed with cardiac sarcoidosis have been seen at my clinic?”. The ability for investigators to quickly ask such questions, iterate upon hypotheses, and explore clinical data of interest also therefore would be expected to improve the output of biomedical research and eventually advance clinical care.

Relational databases are typically the storage and query mechanism used for cohort discovery in large academic medical centers and enterprises, as they conveniently serve as the primary downstream storage system of clinical data for reporting, operational, and research purposes. Data generated in Electronic Health Records [EHR] in the course of patient care “flow” to these clinical databases, and as such are thus able to be queried as secondary uses of data (Botsis, et al. 2010).

A number of hurdles remain before clinical data can be put to use for potential investigators, however. First, many clinicians and researchers have little experience with or knowledge of database programming languages such as SQL (Structured Query Language), which are required for extracting information from a clinical database. While as a query language SQL is largely considered more straightforward to learn than traditional programming languages (Sadiq, et al. 2004), it is nonetheless unrealistic to expect every potential investigator to become facile with SQL in order to conduct research. Second, translating clinical hypotheses into usable database queries is often difficult due to semantic misalignment between clinical phenotypes versus data captured and stored in clinical databases. From the perspective of Boolean logic and programmatic representation, clinical phenotypes are often “loosely” structured and defined by understandings of physiological phenomena not directly captured in the form of coded data stored within Electronical Health Records (EHRs). For example, type 2 diabetes mellitus is understood to result from a combination of resistance to insulin action and inadequate compensatory insulin secretory response (Diabetes Care 2014), but can often be inferred only by clinical database queries of measurement of an associated physiological response (e.g., a hemoglobin A1c laboratory test > 7.0) or clinical or billing-related diagnosis (e.g., an ICD-10 code). Moreover, in addition to forcing the investigator to think in terms of data available in a database rather than phenotypes, diagnoses, vitals, laboratory tests, and other standard elements of clinical care must typically be translated into a standard terminology such as ICD-10 (World Health Organization 2020) or LOINC (Regenstrief Institute 2020).

Before an investigator can execute a cohort discovery query, therefore, she must first 1) translate a clinical phenotype to data believed to be captured within an EHR; 2) map concepts such as diagnoses to a standard terminology; 3) transform the query criteria into a series of AND, OR, and NOT Boolean logic representations; and 4) author and execute the query in SQL or other database query language. While the investigator may have an informatics professional able to run the query available as a departmental or institutional resource, informaticians are often a scarce resource unable to meet the demand of data for all investigators. Even in cases where informaticians are available, it is still nonetheless incumbent on the investigator to accomplish steps 1-3, which require clinical expertise from which to base the query in question.

Modern cohort discovery tools: A number of popular cohort discovery tools exist which significantly simplify the process of constructing and executing database queries for cohort discovery.

Integrating Informatics and Biology at the Bedside, or i2b2 (Murphy, et al. 2010), is a clinical data discovery platform which includes a query interface component accessible from a web browser. Users drag and drop biomedical concepts arranged hierarchically into boxes to create queries, which are translated into SQL and executed on users' behalf. Mappings of clinical concepts into coded terminologies stored in a database are handled by the i2b2 software. Atlas, a part of the OHDSI suite of tools (Hripcsak, et al. 2015), is a web-based platform using the OMOP data model which includes functionality for cohort definition and population analytics. Leaf (Dobbins, et al. 2020), developed by our team at the University of Washington, builds upon successful fundamental components of i2b2 but allows for flexible database structures and offers a more user-friendly user interface. Like i2b2, the Leaf user interfaces includes biomedical concepts arranged hierarchically which can be dragged and dropped to create queries.

While each tool is unique and useful in serving particular purposes, all require investigators to map their knowledge and understanding of clinical concepts to imperfectly modeled electronic representations and perform Boolean logic. Put another way, despite varying levels of user friendliness, potential investigators are still in effect required to learn a new tools and representations rather than focus on their hypotheses and conduct research. This process slows down potential advances in translational biomedical research.

Natural language and machine learning: Humans use natural language nearly effortlessly throughout daily life. In medicine, tasks related to the care of patients, documentation of care in clinical notes, and publication and peer review of research are conducted largely using natural language. While natural language has numerous complexities and challenges in translation to electronic sources and databases, given the efficiency with which humans can produce and understand it there is nevertheless is great potential in tools which allow investigators to query clinical databases using it. We propose to address the gap between natural language and clinical databases for the purpose of cohort discovery using novel methods for query criteria extraction, disambiguation, and visual representation.

This project will leverage work from a number of recent successful projects, including previous work by Dr. Yetisgen of our team using maximum entropy and conditional random fields (CRF) machine learning methods to extract tumor description information from natural text found in radiology notes (Yim, Kwan and Yetisgen 2017). For example, in the sentence, "On the prior study, this *lesion* measured *2.4cm* in maximum diameter", *lesion* is detected as a tumor which has an associated measurement of *2.4 centimeters*. Other work using deep learning and Recurrent Neural Networks (RNNs) by Dernoncourt (Dernoncourt, Lee and Szolovits 2016) found that general-purpose methods using two stacked layers of RNNs for character-enhanced token embeddings and label sequence optimization could produce state-of-the-art results of nearly 99% precision and recall for a wide variety of natural language tasks related to recognition of arbitrary biomedical concepts from text.

We hypothesize that using a two stacked RNN architecture as demonstrated by Dernoncourt, we will be able to extract cohort definition criteria from natural language that can be transformed automatically into a cohort definition query.

***Disambiguation and Named Entity Recognition**:* Data output from the RNN architecture will be subsequently disambiguated into common terminologies and coding systems present in many clinical databases, such as RxNorm (Liu, et al. 2005) and ICD-10, via the Unified Medical Language System (UMLS) (Bodenreider 2004). Disambiguation will be performed by two primary methods: First, the MetaMap (Aronson 2001) natural language processing software will be the first method used for disambiguation of biomedical concepts. For example, if "history of diabetes mellitus" is extracted as a criteria, MetaMap will provide a UMLS concept ID which can then be mapped to one or more ICD-10 codes and used in a final query to a clinical database. Second, for criteria such as "elevated liver function tests", which are themselves not coded values but represent instead a series of coded values, MetaMap and the UMLS often are able to provide possible test types used to test liver function, but we propose to include additional data sources extracted from trusted health websites such as WebMD and mayoclinic.org in order to demonstrate to the user how terms were disambiguated. This will be performed by web-scraping these websites, followed by the use of MetaMap to store extracted data, and finally storage of extracted data in a Semantic Web RDF database (Decker 2000).

Following disambiguation, (1) a SQL query will be compiled and executed on the clinical database, (2) a count of resulting patients found will be reported to the user in a friendly chat-like web interface, and (3) the interface will further explain how the tool was able to disambiguate the query, including URLs and references to sources in order to increase user trust in the results.

The proposed work is highly significant because (1) we will build upon the work of Leaf to allow dynamic query of arbitrary clinical data models, but do so from natural user language, (2) include mappings from web-scraped trusted health websites to corroborate disambiguated results and gain user trust, and (3) display results in a friendly, chat-like interactive user interface to ensure ease-of-use and efficient use of user time.

B. Innovation

The proposed work will incorporate and include a number of novel tools and methodologies which individually will be of significant value to other researchers and as a final deployable application be useful to thousands of potential researchers who use existing cohort discovery tools.

The proposed work also builds upon important recent research conducted by others. Criteria2Query (Yuan, et al. 2019) uses a hybrid machine learning and rule-based extraction pipeline to convert natural language queries from clinical trials into automated SQL queries to the OMOP Common Data Model. Criteria2Query accomplishes this using a process of paragraph and sentence segmentation, negation detection, named entity recognition (using conditional random fields), relation extraction, temporal and numerical normalization, and logic extraction. Criteria2Query also includes a simple user interface which visualizes the translated cohort definition criteria to the user and is able to automated export the extracted cohort to the OHDSI suite of tools. In terms of performance, Criteria2Query was found to achieve respectable F1 scores (the harmonic mean of precision and recall) of .804 for entity recognition and an excellent 0.944 in logic detection. This proposed work builds upon Criteria2Query and other work by:

1) Allowing queries to be mapped and executed to any clinical data model, including but not limited to OMOP. We will leverage and extend our open-source work to-date on Leaf, particularly Leaf's dynamic SQL compiler and abstracted syntax tree transformation functions to accomplish this.

2) Incorporating external data sources and trusted public health websites into results to help the user understand how the query was constructed and where to find further information. Rather than simply displaying the results to the user, the proposed application will build user trust and further be able to inform the user of criteria it could not understand as well, in order to help the user refine their query.
3) Displaying the results in a rich, interactive, natural language-friendly user interface, rather than a traditional drag-and-drop interfaces. Modern tools commonly used for search and communications have made great use of simple chat-like natural language interfaces, and as such we hypothesize that a similar interface can improve user satisfaction and improve efficiency of use.

C. Approach

C.1. Aim 1: Develop and evaluate an annotation guideline for cohort identification criteria using natural language patterns.

In order to extract the query criteria we will first create an annotation guideline for the task of identifying entities and events to be used in the final generated SQL query for the user. The annotation guideline will be developed for use with the BRAT NLP annotation-tool (Stenetorp, et al. 2012) and will follow and build upon the annotation structures of the 2010 i2b2 NLP challenge for extraction of medical problems, tests, and treatments (Uzuner, et al. 2010). The guideline will define *events*, including Conditions, Encounters, Immunizations, and Observations, each of which will include *entities* (similar to properties in object-oriented programming) of the *name*, *domain*, and event-specific entities such as dose and frequency. The annotation guidelines will build upon previous work by Weng and colleagues (Weng, et al. 2011) (Tu 2011) (Kang 2017) but capture greater granularity and specificity of certain elements, for example capturing instances of one event which refer as an example of another using the annotation Example-Of (see Figure 1).

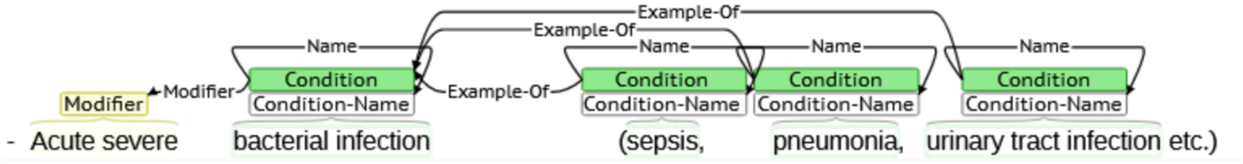


Figure 1: Example annotation incorporating Conditions (an event) and Condition-Names (an entity), with Example-Of Relations annotated.

The annotation guideline will also build upon time-relation annotation guidelines by the same group (Boland, et al. 2012), capturing temporal relations between events and entities (see Figure 2).

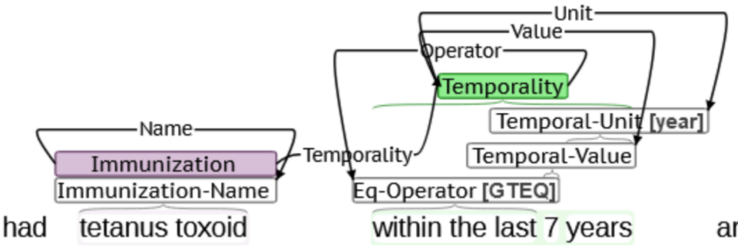
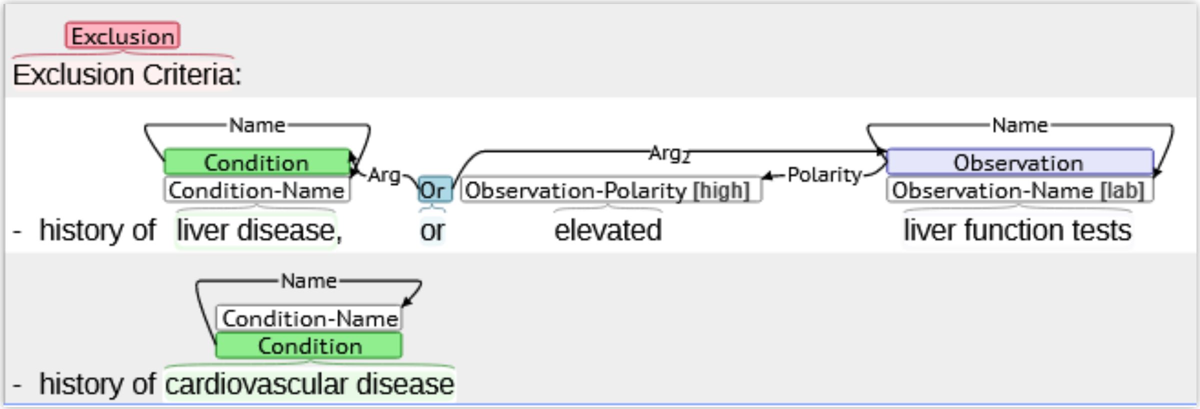


Figure 2: Example time relation annotations for tetanus immunization in the past seven years.

C.1.1. Preliminary Work

To date, we have annotated eligibility criteria uploaded to www.ClinicalTrials.gov (hereafter abbreviated as CT.gov) for over fifty clinical trials conducted throughout the United States in the past decade. The structure of eligibility criteria definitions from CT.gov is expected to closely resemble user cohort definitions from Leaf, as many Leaf users actively lead or participate in clinical trials, and the narrative structure and criteria present used for clinical trials (e.g., BMI > 35 or diagnosis of hypertension) are often seen in cohort definition queries in general in our experience.



Exclusion Criteria:

- history of liver disease, or elevated liver function tests
- history of cardiovascular disease

Figure 3: Example cohort criteria annotation extracted from www.ClinicalTrials.gov.

C.1.2. Proposed Work

In this aim we will generate annotations from two sources of data to be used for training and evaluation of our machine learning algorithms.

Task 1.2.1. Corpus:

With the aim of selecting a diverse group of cohort discovery criteria to use for annotation and training, we will use data from two sources to form our corpus. First, we will conduct a manual review of a subset of user queries of the Leaf cohort discovery tool which will be associated with textual descriptions of what users *intended* to query. These data will be extracted by mining user questions directed to the Leaf support desk alongside user attempted queries immediately before and after asking the Leaf support question. Our goal will be to select 400 Leaf user free-text query descriptions. In order to ensure we have adequate training data, we will augment the Leaf user data with 400 randomly selected free-text clinical trial eligibility criteria documents from CT.gov., for a total of 800 total free-text cohort definitions.

We expect a certain amount of variation in description patterns and word choice between the two corpora, but believe this to be valuable for generalizability. For example, a Leaf user may ask a hypothetical question such as:

> *"How many patients have autoimmune thyroid disease?"*

While a similar query may be phrased on CT.gov as a statement:

> *"Inclusion: history of autoimmune thyroid disease…"*

Based on our preliminary and previous work, we anticipate this variation in phrasing to be a benefit of using these corpora and avoid overfitting our RNN models to data with only a small number of possible phrases.

Task 1.2.2. Information Extraction:

Events, entities, and relations to be extracted will be composed of:

- Events: And, Or, Conditions, Encounters, Equality Comparisons, Immunizations, Medications, Observations, Procedures, Temporal Connections, and Temporalities.
- Entities: Inclusion, Exclusion, Location, Gender, Ethnicity, Life-Stage (i.e., children, adult, elderly), Living-Location.
- Relations: Abbreviation-Of, Caused-By, Treatment-For, Example-Of, If-Then.

We do not anticipate that every criteria will have a corresponding database table and column(s) on which to match on. For example, while a user inquiry may wish to determine cases of "*x* Caused-By *y*", it is often to difficult to determine specific causes of a condition using diagnosis codes alone. Instead, we aim to nonetheless capture this information to be able to inform the user of what information is *not* available to allow her to refine the query.

Task 1.2.3. Annotation:

The goal of the annotation process will be to produce a gold-standard dataset of free-text user inquiries and clinical trials eligibility criteria from which to train our RNN model and evaluate performance.

Annotation plan: We plan to recruit four graduate students from the University of Washington Biomedical Informatics & Medical Education (BIME) Department and four University of Washington medical students for annotation, for a total of eight student annotators. Annotators will be recruited from two separate backgrounds based on our prior experiences with difficulty in finding large numbers of annotators from a single student program. The BIME and medical student annotators will be paired together with one student from each program. The 800 texts will then be divided into subsamples of 200 texts and assigned to each pair, with each student in the pair annotating the same text. Thus each text will be annotated twice, once by each annotator.

Annotation will be performed as an iterative process, beginning with 10 randomly selected notes from the 200 for each pair for training. After each annotator completes their training annotations, we will reconvene the group to review the annotations and further train annotators to ensure they have consistency in annotations. After annotations are complete, we will programmatically determine the intersect of annotations for each text (cases where both annotators agree on the individual labeled annotations) and use the intersected annotations as the gold standard.

We estimate each text document will take approximately 10 minutes to annotate, and as each document will be annotated twice and there will be 800 total, for an estimated time of 66 hours for each annotator and 267 hours total for all. Note that the budget document includes an estimate for this in Year 1 at $15 per hour (current Research Assistant wage).

C.1.3. Evaluation

In order to effectively determine inter-annotator agreement and define a gold standard of annotation, each pair will be blinded to their partner's identity and annotations after initial training. Inter annotator agreement will be determined by Cohen's Kappa (McHugh 2012), which can range from -1 to +1. The null hypothesis will be that given annotation guidelines, inter annotator agreement is less than 0.5.

C.1.4. Expected Outcomes

We expect at the conclusion of this aim to have gold standard annotations for 800 cohort criteria definitions. We will use the annotated notes for RNN training and evaluation purposes in Aim 2.

C.1.5. Potential Pitfalls and Alternatives

Potential pitfalls include: (1) inter-annotator Kappa score is below 0.5 and we accept the null hypothesis, and (2): we are unable to recruit sufficient annotators. If inter-annotator agreement is unexpectedly poor, we will identify and further iteratively retrain annotators in successive batches of notes until sufficient agreement is found. If we are unable to recruit sufficient annotators, we will expand the recruitment pool to other departments outside of BIME and the medical students.

C.2. Aim 2: Develop novel concept extraction and semantic web-based approaches for named entity mapping and clinical concept disambiguation.

Natural language is often abbreviated, ambiguous, and contains implicit information which the speaker assumes the reader to understand without explicitly saying so. As an example, the question, “How many patients had an elevated liver function test in the past 6 months?”, does not explicitly state what laboratory tests qualify as liver function tests, nor the threshold above which a given laboratory result should be consider “elevated”. Aim 2 will therefore focus on (1) training and evaluation of our machine learning RNN-based model for detecting criteria, and (2) the use of externally available biomedical information on the internet for disambiguation and translation of those criteria to terminologies such as LOINC and ICD-10.

RNN Training: Utilizing the annotated texts produced by Aim 1, we will train a deep learning algorithm using the two-layered RNN methods described in the *Significance* section. Following previous work by Dr. Yetisgen (Lybarger, Yetisgen and Ostendor 2018), we will first utilize long short-term memory units (LSTMs), a particular type of RNNs, to study the sequence of words in our annotations and learn cues to indicate concept types within each given span of text. To capture the semantics of tokens that appear only a small number of times, we will also utilize a character-embedding layer augment the initial RNN. The character embedding layer has the additional advantage that it allows the LSTM to better learn suffixes and prefixes of input tokens and also detect misspellings. Finally, we will train a second bidirectional LSTM using the token embeddings as features to identify event, entity, and relation types.

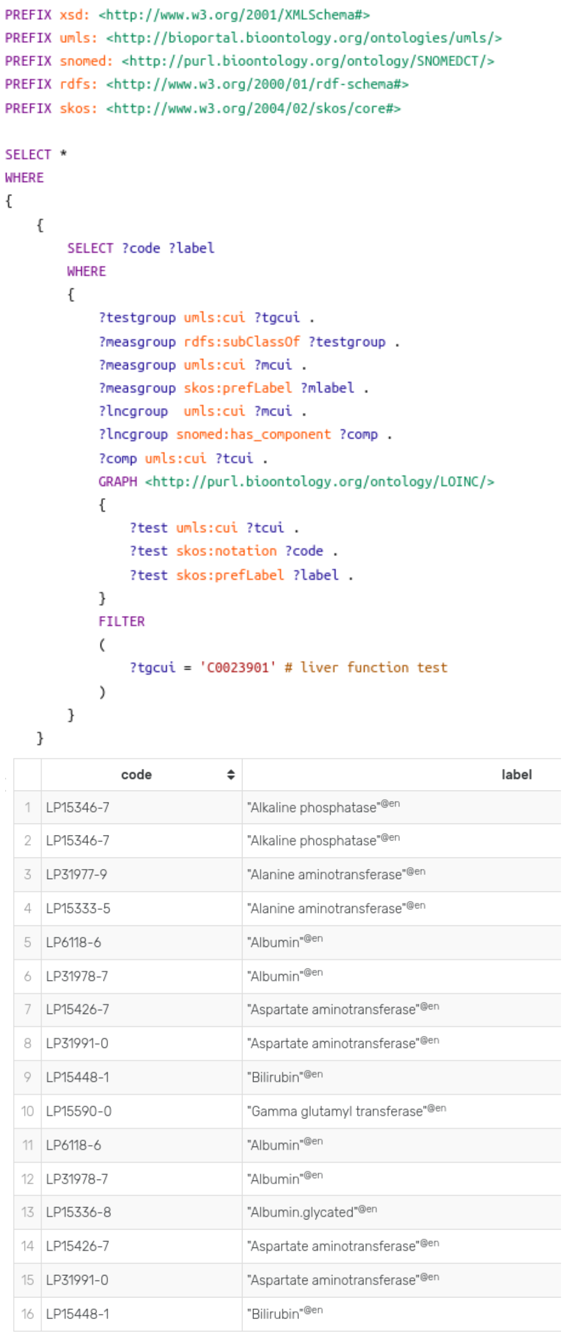
Disambiguation using web-scraped biomedical information and the UMLS: After training, the RNN-based model will be able to read from new, unseen cohort criteria texts and identify events, entities, and relations present based on the annotated training data from Aim 1. For example, we expect that it will be able to identify that “elevated liver function tests” refers to one or more Observations, with an Observation-Polarity of “elevated” (high) and has a name of “liver function test”. These data alone, however, cannot be mapped directly to a clinical database to generate a query. Instead, the text span, “liver function tests” must be mapped to an appropriate set of laboratory test codes in LOINC, because the clinical database to be queried is expected to also code these values and LOINC.

To do this, we will utilize two sources: (1) the Unified Medical Language System (UMLS) (Bodenreider 2004), an electronic metathesaurus which maintains mapping and relations between various biomedical terminologies, and (2) information which is programmatically scraped from trusted health websites such as medlineplus.gov and mayoclinic.org. After each site is scraped, the raw HTML data will be pulled into an intermediate database from which text values will be extracted and analyzed using MetaMap to extract biomedical concepts. The final extracted data and URLs from which they were derived will be stored in an RDF-based graph database.

C.2.3. Preliminary Work

Our team has extensive experience in machine learning methods. As the second part of Aim 2, disambiguation of results using UMLS, MetaMap, and RDF is a relatively new and novel area, this is where we have focused our preliminary studies.

To date, we have downloaded the UMLS into a MySQL database, from which we were able to then transform the UMLS into RDF triple format and load into a Semantic Web-based graph database. Using SPARQL queries, Figure 4 demonstrates how the UMLS unique code for liver function tests, “C0023901”, can be queried to find 16 individual LOINC codes. The LOINC codes can then be mapped to a clinical database which contains laboratory results coded in LOINC using a SQL query generated for the user.



```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX umls: <http://bioportal.bioontology.org/ontologies/umls/>
PREFIX snomed: <http://purl.bioontology.org/ontology/SNOMEDCT/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>

SELECT *
WHERE
{
    {
        SELECT ?code ?label
        WHERE
        {
            ?testgroup umls:cui ?tgcui .
            ?measgroup rdfs:subClassOf ?testgroup .
            ?measgroup umls:cui ?mcui .
            ?measgroup skos:prefLabel ?mlabel .
            ?lncgroup  umls:cui ?mcui .
            ?lncgroup snomed:has_component ?comp .
            ?comp umls:cui ?tcui .
            GRAPH <http://purl.bioontology.org/ontology/LOINC/>
            {
                ?test umls:cui ?tcui .
                ?test skos:notation ?code .
                ?test skos:prefLabel ?label .
            }
            FILTER
            (
                ?tgcui = 'C0023901' # liver function test
            )
        }
    }
```

	code ⬍	label
1	LP15346-7	"Alkaline phosphatase"@en
2	LP15346-7	"Alkaline phosphatase"@en
3	LP31977-9	"Alanine aminotransferase"@en
4	LP15333-5	"Alanine aminotransferase"@en
5	LP6118-6	"Albumin"@en
6	LP31978-7	"Albumin"@en
7	LP15426-7	"Aspartate aminotransferase"@en
8	LP31991-0	"Aspartate aminotransferase"@en
9	LP15448-1	"Bilirubin"@en
10	LP15590-0	"Gamma glutamyl transferase"@en
11	LP6118-6	"Albumin"@en
12	LP31978-7	"Albumin"@en
13	LP15336-8	"Albumin.glycated"@en
14	LP15426-7	"Aspartate aminotransferase"@en
15	LP31991-0	"Aspartate aminotransferase"@en
16	LP15448-1	"Bilirubin"@en

C.2.4. Proposed Work

In this aim we will train our machine learning model and generate RDF triples for phrase disambiguation from web-scraped data.

Task 2.4.1. RNN Training:

Using a secure server maintained by the Research IT department of UW Medicine at the University of Washington, we will load 700 of the 800 total cohort criteria texts to be used for training. We anticipate the training process will take approximately two days based on the number of texts and our prior experience.

Task 2.4.1. Web scraping:

Using a popular web-scraping library in the Python language[1], we will create an automated web-scraping script for the websites of medlineplus.gov and mayoclinic.org. The HTML data from each site will be stored in an intermediate SQL database, and subsequently analyzed using MetaMap to extract biomedical information which will be used to disambiguate user phrases and serve as an external corroborating source for the UMLS. For example, the Mayo Clinic website includes a page for liver function tests[2] which will be parsed and cross-referenced downstream when disambiguating user questions. The resulting data will be stored in a Semantic Web graph database in RDF.

Figure 4: SPARQL graph database query on RDF triples to find LOINC codes for lab tests categorized as a type of liver function test.

RDF and the Semantic Web are uniquely well-suited for problems of information or data (nodes) interconnected by relationships (edges) with depths of connections that may not be known ahead of time, and are frequently used in tools for personalized recommendation engines (Carroll 2004) (Sneha 2012).We hypothesize that this will allow us to match the example phrase ‘liver function test’ to a LOINC code using a pattern such as:

UMLS *includes* **LiverFunctionTests** -->
LiverFunctionTests *includes* **Albumin** -->
Albumin *isEquivalentTo* **AlbuminInSerumOrPlasma** -->
LOINC *includes* **AlbuminInSerumOrPlasma** -->
AlbuminInSerumOrPlasma *isCodedAs* **1751-7** -->

Based on the disambiguation results, a query will then be generated and executed against the clinical database with results returned to the user.

C.2.3.3. Evaluation

To evaluate the semantic web disambiguation approach, we will randomly select 50 cohort criteria definition texts which include at least one ambiguous phrase disambiguated using this method. After processing using the semantic web parser, two biomedical informatics students and two medical students will each annotate the same 50 cohort criteria texts. For each ambiguous phrase, each annotator will be instructed to a provide zero (if not mappable) or more associated clinical codes (e.g., ICD-10, LOINC, SNOMED) using the United Medical Language System’s (UMLS) Metathesaurus Browser[3].

The outcome variable to be measured is intersubject disagreement, defined as the average number of ambiguous phrases per cohort definition on which two subjects disagree. The null hypothesis is that each annotator (including the parser) will be no more distant to each other than the physicians are to each other, based on the work of Hripcsak and colleagues (G. F. Hripcsak 1995) with a power of 0.8.

C.2.3.4. Expected Outcomes

We expect that the outcome of Aim 2 will be two state-of-the-art parsers: (1) an RNN based criteria extraction parser able to determine biomedical concepts, phrases, temporal events, and relations from user cohort definition inquires, and (2) a disambiguation parser using MetaMap

[1] https://pypi.org/project/beautifulsoup4/
[2] https://www.mayoclinic.org/tests-procedures/liver-function-tests/about/pac-20394595
[3] https://uts.nlm.nih.gov/metathesaurus.html

and Semantic Web technology to provide mappings to terminologies and supporting evidence from trusted health websites.

C.2.3.5. Potential Pitfalls and Alternatives

There are several potential pitfalls: (1) if the F1 scores from the criteria extraction parser fall below our minimum criteria of 0.7, this will likely be due to inconsistencies in annotations and we will remedy this by further training and review of annotators. (2) if the web-scraped sources prove insufficient or unmatchable to the UMLS sources for disambiguation, we will examine additional possible web sources or move to use only the UMLS as a source for disambiguation.

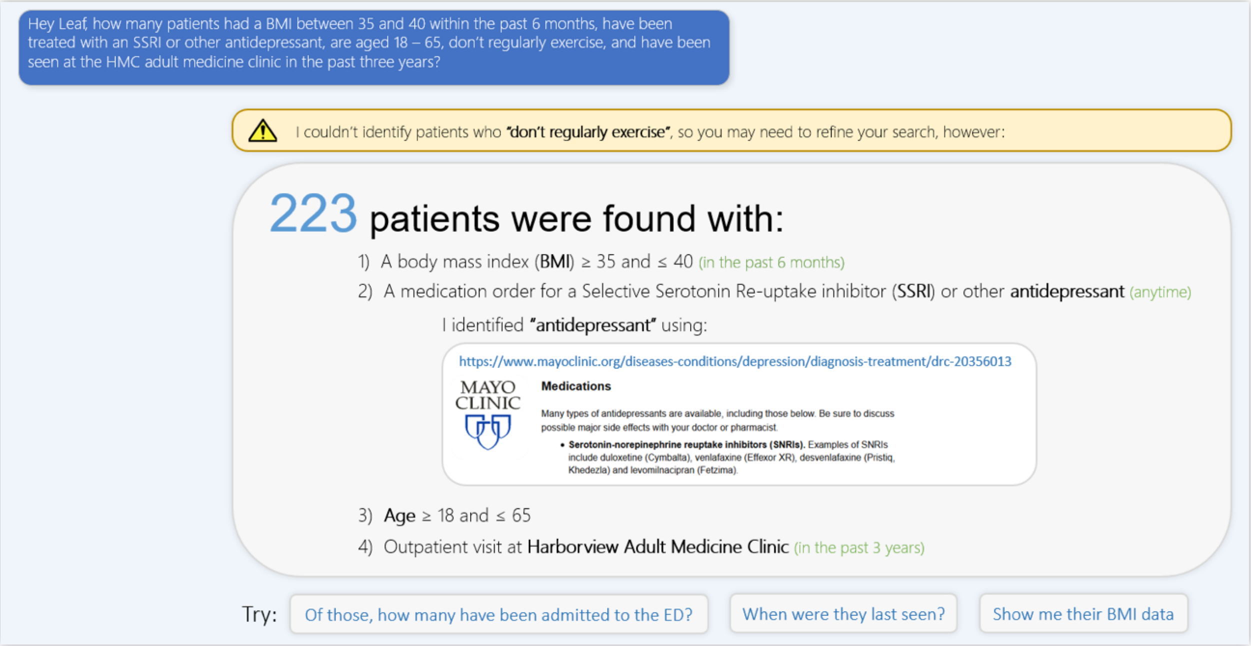
C.3. Aim 3: Evaluate improvements in user satisfaction, productivity, and query accuracy using a blinded experimental user trial.

The final Aim will be to design and implement a user interface and user-facing application which can extract information using annotations from Aim 1 and disambiguate biomedical concepts using the parsing methods from Aim 2. The intervention will be the use of the natural language-based user interface and application, with user satisfaction, productivity, and query accuracy as dependent variables to be measured.

The user interface and client application will build upon our work to-date with Leaf, which has been developed using strong human-centered design techniques and successfully deployed at the University of Washington (Dobbins, et al. 2020).

C.3.3.1. Preliminary Work

Our team has extensive experience developing rich, interactive user interfaces and web applications, particularly Leaf. Preliminary work using mock-ups of the proposed user interface and client application can be seen in Figure 5. While not interactive, the mock-up in Figure 5 demonstrates a hypothetical but realistic free-text query posed by a user.



Hey Leaf, how many patients had a BMI between 35 and 40 within the past 6 months, have been treated with an SSRI or other antidepressant, are aged 18 – 65, don't regularly exercise, and have been seen at the HMC adult medicine clinic in the past three years?

⚠ I couldn't identify patients who **"don't regularly exercise"**, so you may need to refine your search, however:

223 patients were found with:

1) A body mass index (**BMI**) ≥ 35 and ≤ 40 (in the past 6 months)
2) A medication order for a Selective Serotonin Re-uptake inhibitor (**SSRI**) or other **antidepressant** (anytime)

I identified **"antidepressant"** using:

https://www.mayoclinic.org/diseases-conditions/depression/diagnosis-treatment/drc-20356013

MAYO CLINIC

Medications

Many types of antidepressants are available, including those below. Be sure to discuss possible major side effects with your doctor or pharmacist.

- **Serotonin-norepinephrine reuptake inhibitors (SNRIs).** Examples of SNRIs include duloxetine (Cymbalta), venlafaxine (Effexor XR), desvenlafaxine (Pristiq, Khedezla) and levomilnacipran (Fetzima).

3) **Age** ≥ 18 and ≤ 65
4) Outpatient visit at **Harborview Adult Medicine Clinic** (in the past 3 years)

Try:
Of those, how many have been admitted to the ED?
When were they last seen?
Show me their BMI data

Figure 5: Mock-up user interface design for a hypothetical user question and answer.

A number of features of our proposed work are displayed:

- Criteria which could not be mapped to elements available in the clinical database are displayed to the user. This can be seen in the text, "I couldn't identify patients who 'don't regularly exercise'…".
- BMI, SSRI, Antidepressants, Age, and Harborview Adult Medicine Clinic are all disambiguated and mapped to elements in the clinical database. These are transparently displayed to the user.
- Evidence for "antidepressant" disambiguation results are transparently presented. A text snapshot and URL from the mayoclinic.org provide the user context and information with how the application understood their request, building user trust and allowing opportunity for correction if misunderstood.
- Each criteria of the final query is ordinally displayed, highlighted, and summarized. This places the final query result in context and, similar the previous point above, provides the user opportunity to assess and correct if necessary.

C.3.3.2. Proposed Work

In Aim 3 we propose developed of an end-user tool and API to tie the project aims together. The user application will be a web-based application written in TypeScript and React and communicate with the server API over an HTTP RESTful calling interface. The server API will act as an externally-exposed wrapper around the named entity recognition and disambiguation parsers described in Aim 2.

After login, users will be able to enter questions by either typing on their keyboard or using a microphone. If using a microphone, user vocal input will be detected and parsed using Mozilla DeepSpeech open-source software[4]. When questions are detected, their text will be sent to the API for named entity recognition, disambiguation, SQL query compilation, and finally execution. Like the existing Leaf application, results for the executed query will be returned to the user in the form of a count of patients, and will also include the query criteria summary displayed in Figure 5.

After completion of the software, the work produced by this project will be disseminated as follows:

- Modular components of the software, including the named entity recognition parser and disambiguation parser will be made available as open-source software on GitHub.
- The complete end-user client application and API will also be made available as open-source software on GitHub.
- Submission of journal papers for peer review and/or conference presentations on:
    - Lessons learned related to the named entity recognition and disambiguation parsers.
    - Mixed-methods findings related to the chat-based user interface.
    - Findings and refinements to the annotation guidelines.

C.3.3.3. Evaluation

[4] https://github.com/mozilla/DeepSpeech

Evaluation will be conducted using 30 volunteers from the BIME department and University of Washington School of Medicine. Volunteers will be randomly divided into a control group (using the existing drag-and-drop based Leaf application) of 15 participants and a case group (using the new natural language-based application) of 15 participants. Each user will receive application training from an informatics student and be asked to use their respective cohort discovery application to complete three pre-selected queries from CT.gov. Pre-selection of queries is necessary in order to control for possible differences in cohort criteria complexity or unavailable data if participants were instead asked to create their own cohort criteria, thus confounding our analysis.

While creating the queries, users will be timed by an informatics student to determine they time necessary to complete each query. After the queries are complete, participants will be asked to complete a 15-question usability survey designed to measure metrics suggested by Finstad (Finstad 2010) for Efficiency, Effectiveness, and Satisfaction on a seven-point Likert scale. The study members performing the analysis of results will be blinded as to whether participants received the control or intervention application. The evaluation will consist of 1) a Chi-square analysis of participant survey results; 2) collection of qualitative user quotes and thoughts and coded by the study to detect themes; and 3) a Mann-Whitney rank-sum test of timed elapsed for users to create queries. The null hypotheses are that the intervention and control applications will not show statistically significant differences in time taken to execute queries, nor show significant differences in user satisfaction or perceived application utility.

C.3.3.4. Expected Outcomes

The expected outcome of Aim 3 is a friendly, intuitive end-user-facing open-source tool and API which can be installed and run from any academic medical center or other organization which hosts clinical databases.

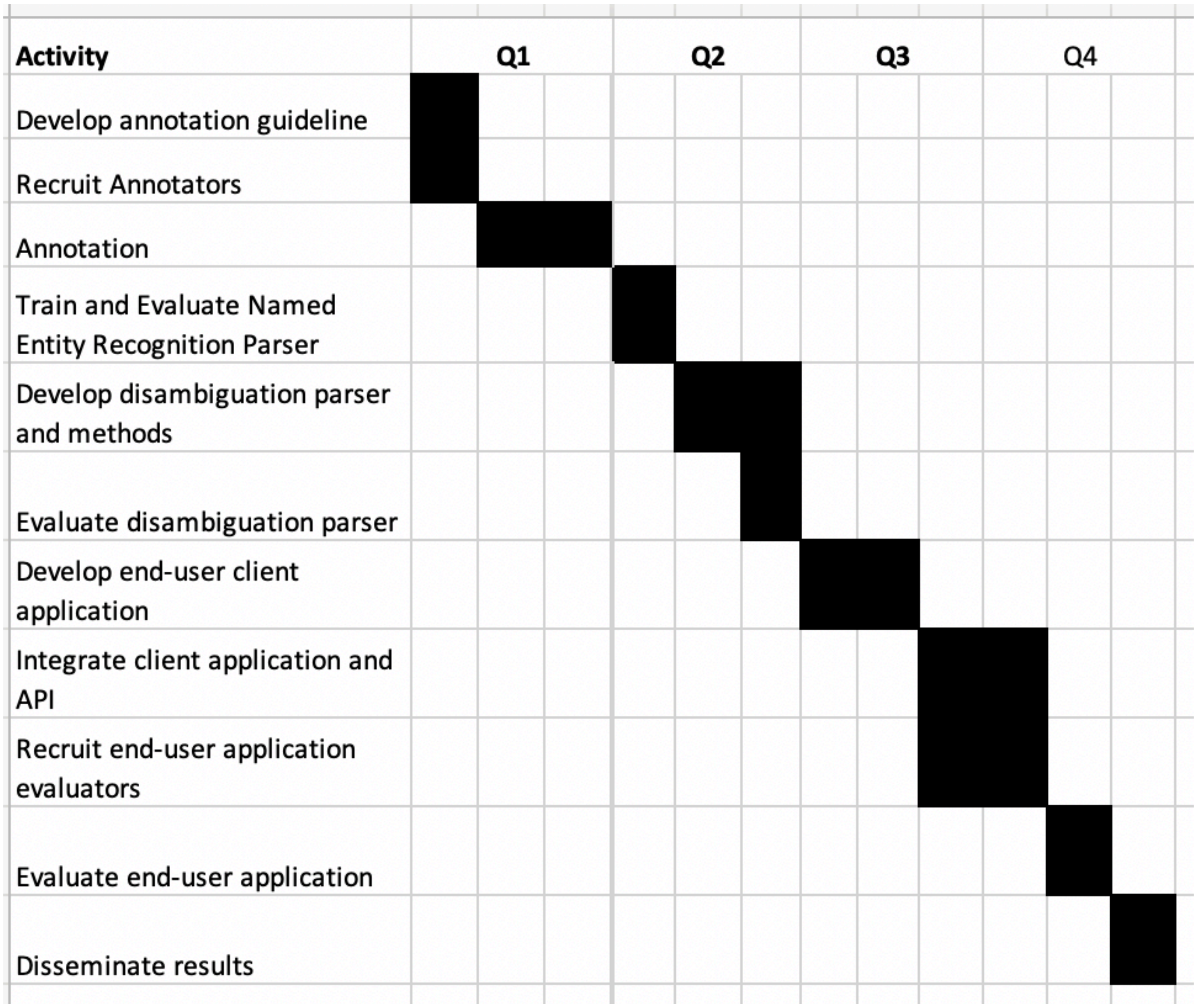
C.3.3.5. Potential Pitfalls and Alternatives

We recognize the following potential pitfalls of Aim 3: (1) The API may not be able to perform named entity recognition and disambiguation within a reasonable time of perhaps 10 seconds or less. We will remedy this by rewriting the API to utilize the same trained model but in a compiled language such as C or C++ rather than Python, which will likely greatly improve speed of execution.

Human Subjects Research

Human subjects will be included in this research as participants in usability Aim 3 regarding usability and Aim 1 regarding University of Washington users of Leaf and submitted Leaf support questions to be used as annotated documents. For Aim 1 Leaf users, permission will be sought from the University of Washington Human Subjects Division for a waiver of consent, as all submissions will be used anonymously and identifiers destroyed. If a waiver of consent is not granted, we will contact each user and inform them of their rights and possible benefits and risks and seek permission to use their data, though no prospective interactions with users will be needed. For Aim 3 participants in the usability study, all identifying information will be removed as it is not needed for analysis, though we will also seek IRB approval and protect the privacy of all participants.

Timeline



Activity	Q1			Q2			Q3			Q4		
Develop annotation guideline	■											
Recruit Annotators	■											
Annotation		■	■									
Train and Evaluate Named Entity Recognition Parser				■								
Develop disambiguation parser and methods					■	■						
Evaluate disambiguation parser						■						
Develop end-user client application							■	■				
Integrate client application and API									■	■		
Recruit end-user application evaluators									■	■		
Evaluate end-user application											■	
Disseminate results												■

References

Aronson, A. 2001. "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program." *Proceedings of the AMIA Symposium* 17.

Bodenreider, O. 2004. " The Unified Medical Language System (UMLS): integrating biomedical terminology." *Nucleic Acids Research* D267-D270.

Boland, MR, SW Tu, S Carini, I Sin, and C Weng. 2012. "EliXR-TIME: A Temporal Knowledge Representation for Clinical Research Eligibility Criteria." *Journal of the American Medical Informatics Association* 71-80.

Botsis, T, G Hartvigsen, F Chen, and C Weng. 2010. "Secondary Use of EHR: Data Quality Issues and Informatics Opportunities." *Proceedings - AMIA Join Summits on Translational Science* 1-5.

Carroll, JJ. 2004. "Jena: implementing the semantic web recommendations." *Proceedings of the 13th international World Wide Web conference* 74-83.

Decker, SM. 2000. "The semantic web: The roles of XML and RDF." *IEEE Internet computing* 63-73.

Dernoncourt, F, JY Lee, and P Szolovits. 2016. "NeuroNER: an easy-to-use program for named-entity recognition based on neural networks." *ArXiv.*

Diabetes Care. 2014. "Diagnosis and Classification of Diabetes Mellitus." *Diabetes Care* S81-S90.

Dobbins, NJ, CH Spital, RA Black, JM Morrison, B de Veer, E Zampino, RD Harrington, et al. 2020. "Leaf: an open-source, model-agnostic, data-driven web application for cohort discovery and translational biomedical research." *Journal of the American Medical Informatics Association* 109-118.

Finstad, K. 2010. "The Usability Metric for User Experience." *Interacting with Computers* 323-327.

Hripcsak, G, JD Duke, NH Shah, CG Reich, V Huser, MJ Schuemie, MA Suchard, et al. 2015. "Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers." *Stud Health Technol Inform* 574-578.

Hripcsak, G., Friedman, C., Alderson, P., DuMouchel, W., Johnson, S., & Clayton, P. 1995. "Unlocking Clinical Data from Narrative Reports." *Annals of Internal Medicine* 681-688.

Kang, T., Zhang, S., Tang, Y., Hruby, G., Rusanov, A., Elhadad, N., & Weng, C. 2017. "EliIE: An open-source information extraction system for clinical trial eligibility criteria." *Journal of the American Medical Informatics Association* 1062-1071.

Liu, S, W Ma, R Moore, V Ganesan, and S Nelson. 2005. "RxNorm: Prescription for Electronic Drug Information Exchange." *IEEE Computer Society.*

Lybarger, K, M Yetisgen, and M Ostendor. 2018. "Using Neural Multi-task Learning to Extract Substance Abuse Information from Clinical Notes." *AMIA Annual Symposium Proceedings* 1395-1404.

McHugh, M. 2012. "Interrater reliability: the kappa statistic." *Biochemia Medica* 276-282.

Murphy, SH, G Weber, M Mendis, V Gainer, HC Chueh, S Churchill, and I Kohane. 2010. "Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2)." *Journal of the American Medical Informatics Association* 124-130.

Regenstrief Institute. 2020. "LOINC." *LOINC.* Accessed 3 6, 2020. https://loinc.org/.

Sadiq, S, M Orlowska, W Sadiq, and J Lin. 2004. "SQLator: an online SQL learning workbench." *ITiCSE Proceedings* 223-227.

Sneha, YS. 2012. "A personalized product based recommendation system using web usage mining and semantic web." *International Journal of Computer Theory and Engineering* 202.

Stenetorp, P, S Pyysalo, G Topic, T Ohta, S Ananiadou, and J Tsujii. 2012. "BRAT: a web-based tool for NLP-assisted text annotation." *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* 102-107.

Tu, S., Peleg, M., Carini, S., Bobak, M., Ross, J., Rubin, D., & Sim, I. 2011. "A practical method for transforming free-text eligibility criteria into computable criteria." *Journal of Biomedical Informatics* 239-250.

Uzuner, O, B South, S Shen, and S DuVall. 2010. " 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text." *Journal of the American Medical Informatics Association* 552-556.

Weng, C, X Wu, Z Luo, MR Boland, D Theodoratos, and S Johnson. 2011. " pdfPDF Split View Cite Permissions Icon Permissions Share EliXR: an approach to eligibility criteria extraction and representation." *Journal of the American Medical Informatics Assocation* 116-124.

World Health Organization. 2020. "ICD." *World Health Organization.* Accessed 3 6, 2020. https://www.who.int/classifications/icd/en/.

Yim, W, S Kwan, and M Yetisgen. 2017. "Classifying Tumor Event Attributes in Radiology Reports." *Journal of the Association for Information Science and Technology* 2662-2674.

Yuan, C, P Ryan, C Ta, Y Guo, Z Li, J Harding, R Makadia, et al. 2019. "Criteria2Query: a natural language interface to clinical databases for cohort definition." *Journal of the American Medical Informatics Assocation* 294-305.