Clinical Data-Driven Probabilistic Graph Processing

Travis Goodwin and Sanda Harabagiu
Human Language Technology Research Institute
University of Texas at Dallas
Richardson, TX 75083-0688, USA
{travis,sanda}@hlt.utdallas.edu

Abstract

Electronic Medical Records (EMRs) encode an extraordinary amount of medical knowledge. Collecting and interpreting this knowledge, however, belies a significant level of clinical understanding. Automatically capturing the clinical information is crucial for performing comparative effectiveness research. In this paper, we present a data-driven approach to model semantic dependencies between medical concepts, qualified by the beliefs of physicians. The dependencies, captured in a patient cohort graph of clinical pictures and therapies is further refined into a probabilistic graphical model which enables efficient inference of patient-centered treatment or test recommendations (based on probabilities). To perform inference on the graphical model, we describe a technique of smoothing the conditional likelihood of medical concepts by their semantically-similar belief values. The experimental results, as compared against clinical guidelines are very promising.

Keywords: Information Retrieval, Bioinformatics, Patient Cohort

1. Introduction

An increasing abundance of clinical data is available through massive warehouses of Electronic Medical Records (EMRs). Both within the United States and across the world, hospitals generate millions of EMRs each year. These EMRs include rich clinical information, consisting of detailed notes on patients' medical history, physical exam findings, lab reports, radiology reports, operative reports, and discharge summaries. Clinical information contains multiple mentions of *medical problems*, including observations resulting from a physical exam (known as *signs*), features that the patient observed first-hand (known as *symptoms*), historical and present medical problems (known as *co-morbidities*), in addition to *diagnostic* information. We have used the ontological definitions of medical concepts related to diseases outlined in (Scheuermann et al., 2009) to capture the semantics of clinical information. Hence, we have considered the fact that EMRs also document the medical interventions performed during the patient's hospital stay, including *medical tests* and their results, as well as all the *medical treatments* performed as part of the patient's *therapy*. These forms of clinical information are crucial for performing comparative effectiveness research. As shown in (Ratner et al., 2009), capturing the clinical information from EMRs enables the discovery of alternative methods to prevent, diagnose, treat, or monitor a medical problem.

It has been shown that clinical information – medical concepts (e.g. problems, tests and treatments) – can be automatically identified from clinical texts, as described in (Uzuner et al., 2011). However, because medical science centers around asking hypotheses, experimenting with new methods of care, and evaluating medical evidence, medical concepts are associated with different degrees of belief, or *assertions*. As such, clinical writing entails a large number of speculative statements indicating the physician's belief at the time, rather than strictly quantifying a fact. In order to take into account the physicians' beliefs when automatically processing the clinical information from EMRs, we also recognized the assertions formulated by physicians when discussing any of the medical concepts.

The 2010 i2b2/VA challenge evaluated the task of automatically inferring six types of *assertions*, or belief states, used to qualify medical problems in EMRs (Uzuner et al., 2011). However, those assertions correspond to clinical information found in only one type of EMR: discharge summaries. Because we consider more types of EMRs, we have extended the problem of classifying medical assertions by considering additional types of assertions. The new assertion values were selected based on discussions with practicing clinicians, and by following the guidelines outlined in (Uzuner et al., 2011).

Medical concepts and their assertions were cast as nodes in a graph which encodes a patient's **clinical picture** and **therapy** along with the potential dependencies between them. We called this graph the **clinical graph** (CG). As in (Scheuermann et al., 2009), the clinical picture is defined as the *clinical phenome*[1] which contains the clinical findings (e.g. medical problems, signs, symptoms and tests). Likewise, we use Scheuermann's definition of *therapy* as all the treatments, cures, and preventions included within the management plan for an individual patient. Figure 1 illustrates our representation of the CG for a patient. Given the patient's hospital visit, we automatically discover the medical problems along with the tests and treatments documented during the patient's hospital course. Medical problems, tests, and treatments are qualified by their assertions and connected by their dependencies (e.g. when *cellulitis* was a present diagnostic, a *blood culture* test was conducted).

Moreover, as reported in (Scheuermann et al., 2009), the clinical picture may vary widely between patients with the same disease and even for the same patient during the course of his or her diseases. Therefore, in order to capture the variation in the corresponding clinical graphs (CGs), we have

[1] While the *clinical phenotype* refers to the set of observations related to a medical condition, the *clinical phenome* is the set of observations pertaining to a single patient.

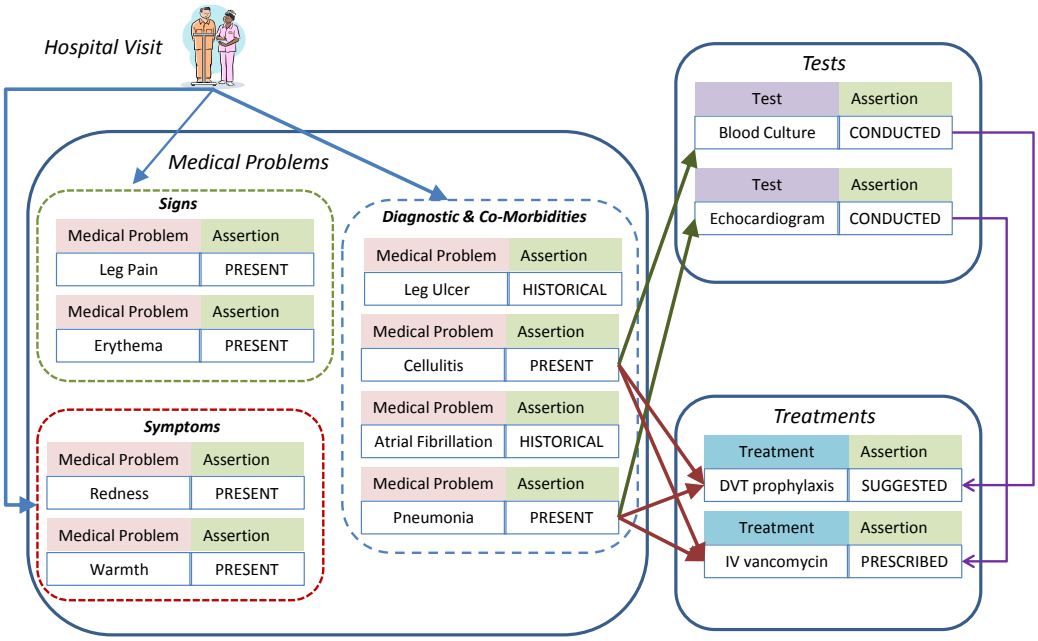

Figure 1: The Clinical picture & therapy Graph (CG).

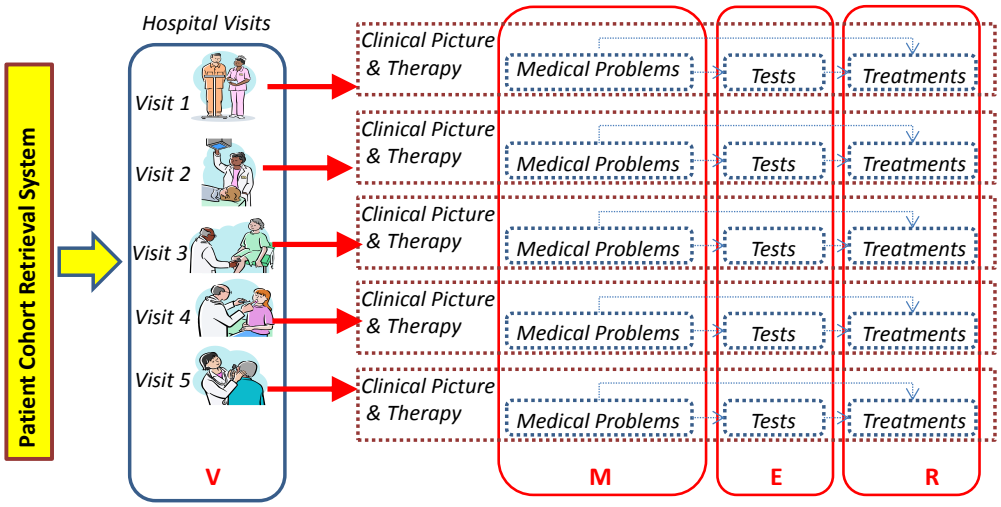

Figure 2: The combined Cohort Clinical Graph (CCG).

considered a **patient cohort** which we obtained by using the system reported in (Goodwin and Harabagiu, 2013). Patient cohort retrieval results in an ordered set of hospital visits which correspond to a cohort of patients sharing the same diagnosis (e.g. *patients with abcess*[2]). As illustrated in Figure 2, this enabled us to access all the clinical pictures and therapies from all the clinical graphs (CGs) of all patients within a cohort. This clinical information regarding a patient cohort constitutes the set of all hospital visits (**V**), the set of all medical problems (**M**), the set of all medical tests (**E**), and the set of all treatments (**R**), across the CGs of all the patients belonging to the cohort. We refer to the graph that combines all CGs as the Cohort Clinical Graph (CCG).

Given a patient cohort, the corresponding CGG was cast as a k-partite graph (where $k = 4$) because there are four types of nodes (**V**, **M**, **E** and **R**), as illustrated in Figure 2. It is to be noted that the edges from the CCG originate from the CGs of patients from the cohort. We also noticed that, crucially, the CCG can also be viewed as a factorization of a Markov network. In this way, we were able to transform the CCPT into a probabilistic graphical model. Probabilistic graphical models (Koller and Friedman, 2009) are known to be a state-of-the-art representation for producing probabilistic inference, which we used for finding recommendations for the most adequate tests or treatments for a patient, given inference on the CCG.

The remainder of this paper is organized as follows. In Section 2, we describe the clinical language processing required for generating the CGs. Section 3 describes the construction of the CCG, as well as how it can be transformed into a prob-

[2]Abscess is an infectious disease of the skin and soft tissue.

abilistic graphical model. Section 4 presents the inference mechanisms we considered and how they may be used for clinical test and treatment recommendation. Section 5 discusses the experimental results, and Section 6 summarizes the conclusions.

2. Medical Language Processing

Open-source software, such as MetaMap (Aronson, 2001) or, more recently, cTakes (Savova et al., 2010) can parse EMRs to determine concept unique identifiers (CUIs) which correspond to entries in the Unified Medical Language System (UMLS) (Bodenreider, 2004). However, UMLS includes many concepts that were authored according to ontological principles and, thus, it is too fine-grained for our purpose of data-driven probabilistic processing of EMRs. In selecting a conceptual representation, we also evaluated the more general frameworks developed by the i2b2/VA challenge in 2010 (Uzuner et al., 2011). This framework was designed to detect medical concepts within clinical text and assign one of several distinct assertions indicating the state of the author's belief for each concept. This i2b2 challenge helped popularize the notion that recognizing medical concepts alone is not sufficient for clinical reasoning, because, when medical concepts are used in clinical texts, physicians also express their belief state about such concepts, e.g. that a medical problem is present or absent, that a treatment is conditional on a test. The i2b2 challenge, however, considered assertions only for medical problems. In our aim to build the CCG, we have extended the problem of assertion classification in two ways: (1) we have produced assertions (or belief values) for all medical concepts (including treatments and tests) that we have automatically identified; and (2) we have introduced 6 additional values which are defined in Table 1.

2.1. Medical Concept Recognition

To recognize the nodes of the CCG, we have partitioned medical concepts within three categories: (1) medical problems (e.g. ATRIAL FIBRILLATION – an irregular heart beat); (2) medical treatments (e.g. ABLATION – the removal of undesired tissue); and (3) medical tests (e.g. ECG – an electrocardiogram). We detect these medical concepts using the methods reported in (Roberts and Harabagiu, 2011). Further, we distinguish three sub-classes of medical problems: (a) signs (observations from a physical exam), (b) symptoms (observations by the patient), (c) co-morbidities (diseases or disorders), and (d) the diagnostic. Our method recognizes medical concepts in three steps:

Step 1: Identification of the boundaries within text that refers to a medical concept;

Step 2: Classification of the medical concept into (1) medical problems, (2) medical treatments, or (3) medical tests.

Step 3: Classification of medical problems into (a) signs, (b) symptoms, (c) co-morbidities, or (d) diagnostics.

Medical concepts were recognized both within the narrative (i.e. report text) and structured sections (e.g. CHIEF COMPLAINT) of EMRs. To do this, we used two conditional random fields (CRFs), trained on the i2b2 annotations as well as our own set of 2,349 EMR annotations. As illustrated in Figure 3, we incorporated knowledge from many lexico-semantic resources. In this research, we used the feature set reported in (Roberts and Harabagiu, 2011). Additionally, we have normalized the detected medical concepts by (1) converting the surface string to lowercase, (2) filtering words belonging to closed-class[3] words, and (3) ignoring word order.

2.2. Medical Assertion Classification

In order to encode the medical knowledge from EMRs with the clinical graph (CG) of each patient, we needed to automatically qualify each medical concept with one of the assertions given in Table 1. We performed this automatic classification using an SVM classifier which considers information from: (a) the medical concept to be classified, (b) the section header where the assertion is implied, (c) features available from UMLS (extracted by MetaMap), (d) features reflective of negated statements, disclosed through the NegEx negation detection package, and (e) belief values are available from the Harvard General Inquirer's category information (Stone et al., 1966). Additional details of the automatic assertion identification techniques are provided in (Roberts and Harabagiu, 2011).

3. Generating the Graphical Model

For clinical decision support, it is critical to analyse the relationships between medical problems, medical tests, and associated treatments across patients' hospital visits. As such, we must move beyond merely identifying the textual mentions of medical concepts and their associated belief values. To this end, we present a framework for modelling the data-driven interactions between problems, treatments, and tests. We first create a CG in which connections between medical concepts are not only inferred, but their strength is also quantified by a weight. Because of the economy of language, relations between medical concepts are rarely explicitly stated, but they are rather implied. To capture these implications, we postulate that co-occurrence statistics can inform these relations, and further that they can also inform the strength of these relations.

After we create complete CGs, we can then transform the combined CGs for a cohort of patients (the CCG) into a probabilistic graphical model.

3.1. Inferring Edges in the Cohort Clinical Graph

The nodes of the CCG are automatically discovered by the language processing techniques described in Section 2. In addition, we needed to infer the edges of the CCG and the weights of the edges indicating semantics used in the clinical picture and therapy ontological definition. The observations from the clinical picture of a patient connected hospital visit (or nodes from $\mathbf{V}$) to the observed medical problem (or nodes from $\mathbf{M}$) generating edges of type $T_{\mathbf{VM}}$. In the clinical picture of patients, connections between the observed

[3] In linguistics, a closed-class of words is a class of words for which new words are rarely introduced, for example pronouns, determiners, prepositions, etc.

Assertion Value	Problem	Treatment	Test	Scenario	EMR Excerpt
HISTORICAL*	✓	✓	✓	occurred during a previous hospital visit	the patient's past medical history is significant for CONGESTIVE HEART FAILURE
CONDITIONAL	✓	✓	✓	occurs only during certain conditions	readmit him for REHAB once the WOUND has HEALED
PRESCRIBED*		✓		has been assigned and will occur	she was given ROCEPHIN and ZITHROMAX
ABSENT	✓	✓	✓	is not present	the patient denies any CHEST PAIN at this time
SUGGESTED*		✓		has been advised, but cannot be assumed to occur	was recommended that he be on ALLOPURINOL
PRESENT	✓			is currently happening	there is a moderate PERICARDIAL EFFUSION
HYPOTHETICAL	✓			may occur in the future	she is to return for any WORSENING PAIN
ORDERED*			✓	has been scheduled and will occur in the future	we will do a PULMONARY FUNCTION TEST
ASSOCIATED WITH ANOTHER	✓			not associated with the patient	father died of LUNG CANCER
POSSIBLE	✓			may occur, but there is uncertainty	I believe that this may represent worsening for PULMONARY HYPERTENSION
ONGOING*	✓	✓		currently exists and can be assumed to persist into the future	continue DIALYSIS
CONDUCTED*			✓	has been performed and completed	UNASYN 3 GRAMS IV was given

Table 1: Assertion values for medical concepts (typeset in SMALLCAPS) in each excerpt; "moment" refers to the specific instant when the medical concept was mentioned. Newly defined assertions are marked with an '*'.

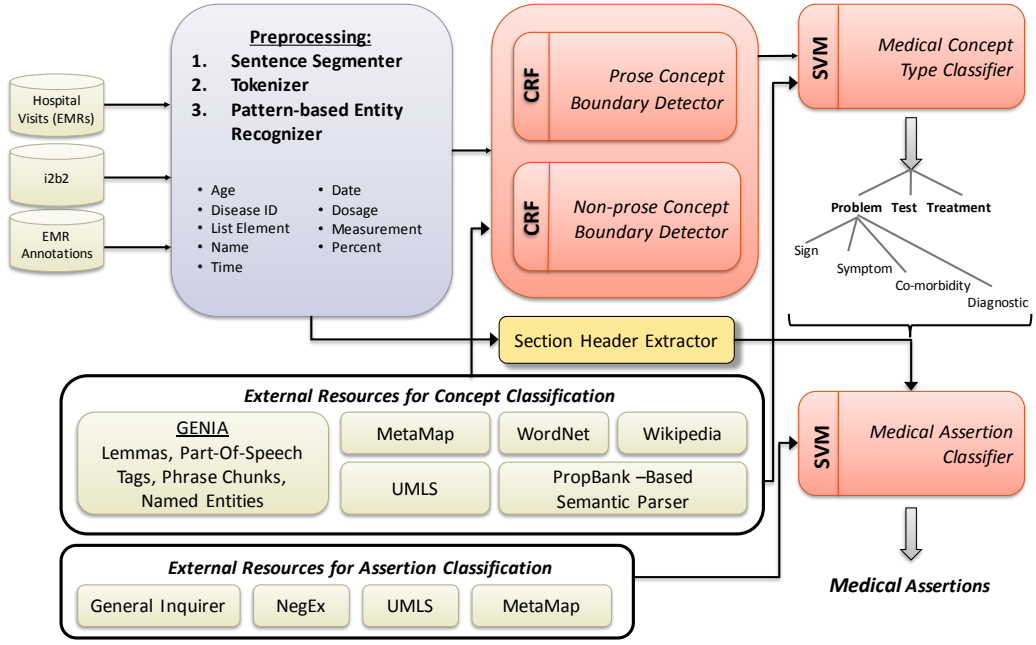

Figure 3: Language processing used for constructing the CGs and CCG.

medical problems (i.e. nodes from **M**) and results of tests (i.e. nodes from **E**) exist as well, giving rise to edges of type $T_{\mathbf{ME}}$ in the CCG. In addition, connection between both types of nodes (medical problems and tests) in the clinical picture and therapies exist. Thus, we shall also have edges in the CCG between medical problems (i.e. nodes from **M**) and treatments (nodes from **R**), generating edges of type $T_{\mathbf{MR}}$. Similarly, we have edges between tests (i.e. nodes from **E**) and treatments (nodes from **R**), generating edges of type $T_{\mathbf{E}}\mathbf{R}$. The weight of edges of each type is computed as

follows:

- The weight of an edge of type $T_{\mathbf{VM}}$ between a visit $v \in \mathbf{V}$ and a medical problem $m \in \mathbf{M}$ is computed as the number of EMRs associated with v which also mention m.
- The weight of an edge of type $T_{\mathbf{ME}}$ between a medical problem $m \in \mathbf{M}$ and test $ein\mathbf{E}$ is computed by the number of EMRs in which both m and e co-occur (regardless of the patient).
- The weight of an edge of type $T_{\mathbf{MR}}$ between a medical problem $m \in \mathbf{M}$ and treatment $rin\mathbf{R}$ is computed by the number of EMRs in which both m and r co-occur (regardless of the patient).
- The weight of an edge of type $T_{\mathbf{ER}}$ between a test $e \in \mathbf{E}$ and treatment $rin\mathbf{R}$ is computed by the number of EMRs in which both e and r co-occur (regardless of the patient).

3.2. The Probabilistic Graphical Model

In Section 3.1 we presented a co-occurrence-based method of building a cohort clinical graph (CCG). The observation that this graph is in fact a k-partite graph (where $k = 4$) enables us to build the factorized Markov network illustrated in Figure 4, which we call the Clinical Markov Network (CMN).

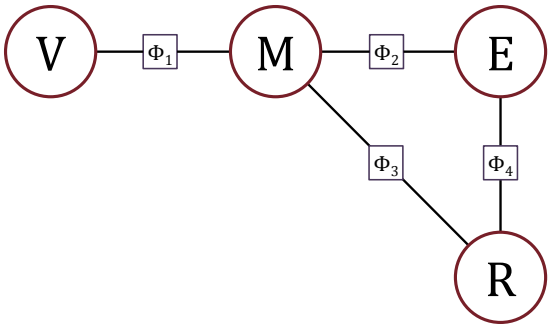

Figure 4: The factorized Clinical Markov Network (CMN).

In the CMN, we assume that each vertex class (**V**, **M**, **E**, or **R**) represents a distinct random variable in the induced Markov network. Similarly, each of the four types of weighted edges ($T_{\mathbf{VM}}, T_{\mathbf{ME}}, T_{\mathbf{MR}}, T_{\mathbf{ER}}$) have associated four different factors to indicated the strength of the edge in the CCG:

- $\Phi_1(v, m) =$ weight of edge $\{v, m\} \in T_{\mathbf{VM}}$
- $\Phi_2(m, e) =$ weight of edge $\{m, e\} \in T_{\mathbf{ME}}$
- $\Phi_3(m, r) =$ weight of edge $\{m, r\} \in T_{\mathbf{MR}}$
- $\Phi_4(e, r) =$ weight of edge $\{e, r\} \in T_{\mathbf{ER}}$

This factorization allows us to perform efficient probabilistic inference by defining the joint probability as the Gibbs distribution given in Equation 1.

$$P(v, m, e, r) = \frac{1}{Z}\Phi_1(v, m)\Phi_2(m, e)\Phi_3(m, r)\Phi_4(e, r) \quad (1)$$

Note that Z is the typical normalization constant equal to the partition function, as given in Equation 2.

$$Z = \sum_{v,m,e,r} \Phi_1(v, m)\Phi_2(m, e)\Phi_3(m, r)\Phi_4(e, r) \quad (2)$$

4. Probabilistic Inference

By modelling the CCG as a probabilistic graphical model, we have gained access to an incredible breadth of probabilistic information through the power of probabilistic inference. We can use this probabilistic information to construct a recommendation engine enumerating the most probable treatments for a given patient given their medical problems and/or their medical tests.

We can use this joint distribution to calculate posterior probability of conducting a medical test during a particular patient's hospital visit (i.e. $P(E = e \mid V = v)$) as shown in Equation 3.

$$P(e \mid v) = \frac{1}{Z}\sum_{m \in M} \Phi_2(e, m)\Phi_1(v, m) \quad (3)$$

Likewise, we can infer the posterior distribution of medical treatments for a given set of N medical problems, $m_0, m_1, \ldots, m_N \in M$, as the conjunction of each problem's posterior distribution, as shown in Equation 4.

$$P(r \mid m_0 \wedge m_1 \wedge \ldots \wedge m_N) = \frac{N}{Z}\sum_{e \in E} \Phi_4(e, r) \prod_{i=1}^{N} \Phi_3(m_i, r)\Phi_2(m_i, r) \quad (4)$$

Although this straightforward approach yields precise results, it suffers from significant sparsity problems induced by our decision to qualify all medical concepts by the physician's belief state. Rather than restricting ourselves to the interactions between concepts exactly matching the specified belief states (e.g. the likelihood that a test is conducted given than a problem is present), we also consider the interaction between the same concepts with semantically similar belief states (e.g. suggested, ordered, prescribed, conditional). For example, consider that assertions ONGOING and CONDUCTED both imply a strong degree of certainty that the medical concept occurred and are likely to have similar semantic relationships despite having different temporal groundings. Thus, they are *semantically coherent*. Based on this observation, we introduce an assertion smoothing factor, S, that encodes the degree to which two assertions are *semantically coherent*, as given in Equation 5.

$$S(a_1,a_2) = \sum_{i=0}^{|C|}\sum_{v \in V} P\big((c_i, a_1), v\big) \sum_{j=0}^{|C|} P\big(v, (c_j, a_2)\big) \quad (5)$$

This smoothing factor, $S(a_1, a_2)$, captures the degree by which occurrences for a certain medical concept labeled with the assertion a_2 may be relevant to probabilistic queries targeting the same medical concept with assertion a_1. We estimate this value as the number of two-step paths in the CMN from any concept with assertion a_1 to any concept with assertion a_2.

This assertion smoothing factor allows us to make recommendations for a *query concept* given an *evidence concept* (e.g. $P\big((q_c, q_a) \mid (e_c, e_a)\big)$), by considering information across all belief values weighted by their semantic similarity to the given belief values. We accomplish this by smoothing the co-occurrence probability as a mixture model of three components as shown in Equation 6: (1) the direct probability, P, that the exact concepts co-occurred; (2) the total probability that the exact query concept co-occurred with the evidence concept qualified by any possible assertion (i.e. $\sum_i P\big((q_c, q_a) \mid (e_c, a_i)\big)$, scaled by the smoothing factor between the encountered evidence assertion and the desired evidence assertion, i.e. $S(q_a, a_i)$; and (3) the total probability that the query concept qualified by any assertion co-occurred with the exact evidence concept (i.e. $\sum_i P\big((q_c, a_i) \mid (e_c, e_a)\big)$, scaled by the smoothing factor between the encountered query assertion and the desired query assertion, i.e. $S(a_i, e_a)$.

$$\hat{P}((c,a)|(d,b);\delta) = \begin{cases} \lambda_0 P\big((c,a)\big|(d,b)\big) \\ \quad + \lambda_1 \sum_{\beta} \hat{P}\big((c,a)\big|(d,\beta);\delta-1\big) S(b,\beta) \\ \quad + \lambda_2 \sum_{\alpha} \hat{P}\big((c,\alpha)\big|(d,b);\delta-1\big) S(\alpha,a) & \text{if } \delta > 0; \\ P\big((c,a)\big|(d,b)\big) & \text{otherwise.} \end{cases} \quad (6)$$

In order to limit the length of transitive paths considered, we introduce a limiting parameter, δ, which limits the recursive depth by which medical concepts will be smoothed (if $\delta = 0$, no smoothing will occur). This smoothing allows us to predict the likelihood of a certain medical test or treatment for a given patient by considering the dependencies encoded in the EMRs across all assertion values without disregarding the semantics of each assertion.

5. Experimental Results

To produce the data-driven Clinical Markov Network (CMN), we used the same EMRs that enabled us to build a patient cohort retrieval system for the medical records track (TRECMed) of the Text REtrieval Conference (TREC) in 2011 and 2012 (Voorhees and Tong, 2011; Voorhees and Hersh, 2012). This dataset includes 95,703 de-identified EMRs which were generated from multiple hospitals during 2007. The EMRs were grouped into hospital visits consisting of one or more medical reports from each patient's hospital stay. Thus, the EMRs were organized into 17,199 different patient hospital visits. Each visit had the patient's admission diagnoses, discharge diagnoses, and related ICD-9 codes. We also used the 826 discharge summaries used during the 2010 i2b2/VA challenge which contained 72,896 medical concepts and their assertions.

As illustrated in Figure 3, in addition to the hospital visits and associated EMRs, we have also used annotations which we produced on the EMRs resulting for three patient cohorts targeted by the queries (Q1) "patients who presented with cellulitis," (Q2) "patients diagnosed with abscess," and (Q2) "patients suffering from both cellulitis and abscess."

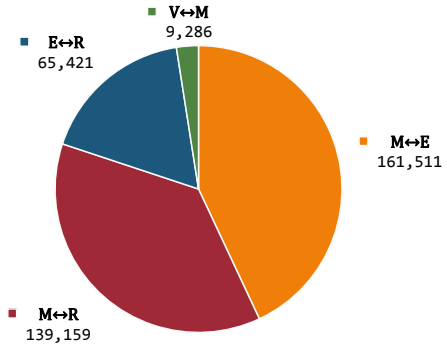

Figure 5: Distribution of edges in the CCG.

	Cellulitis	Cellulitis & Abscess	Abscess
Precision	50%	71%	64%
Accuracy	58%	98%	84%

Table 2: Precision and accuracy for the top 15 treatments for each cohort.

We annotated these EMRs with the medical concepts and assertions described in Section 2.

By automatically processing the medical language in this subset of EMRs, we were able to generate the Clinical Markov Network (CMN) described in Section 4, which corresponds to a cohort of patients with cellulitis or abscess. The distribution of edge classes in the CMN for these cohorts is not uniform, as illustrated in Figure 5.

Figure 5 plots the distribution of edges in the CCG by type. Note that the distribution of edges in the CCG corresponds to the un-normalized probability mass of each factor in the CMN. It is clear from this distribution, that the majority of edges involve medical problems, with a nearly equal number of inferred dependencies between medical problems and tests. In Figure 5, the number of edges between medical problems and tests, $T_{\mathbf{ME}}$ (denoted as $\mathbf{M} \leftrightarrow \mathbf{E}$), and between medical problems and treatments, $T_{\mathbf{MR}}$, denoted as $\mathbf{M} \leftrightarrow \mathbf{R}$, are nearly equal. As such, the number of edges between medical tests and treatments, $T_{\mathbf{ER}}$, denoted as $\mathbf{E} \leftrightarrow \mathbf{R}$, makes up a smaller portion, indicating that there are an abundance of medical problems listed in each EMR. This reinforces to the fact that physicians typically document all the historical, possible, and related or even unrelated medical problems observed during a patient's physical or other examinations.

In order to evaluate the validity of the inference that the CMN enables, we asked two inferential questions: (1) "what are the most probable medical treatments for a certain patient cohort?" and (2) "which tests are most likely to be conducted on patients with the given medical problem(s)?". We answered the first question by computing the conditional probability distribution for all treatments conditioned on the medical problems associated with the cohort retrieved for Q1, Q2, and Q3. These probability distributions are computed according to Equation 4.

The second question was answered by calculating the conditional probability distribution over all tests conditioned on the hospital visits associated with each cohort, as computed with Equation 3.

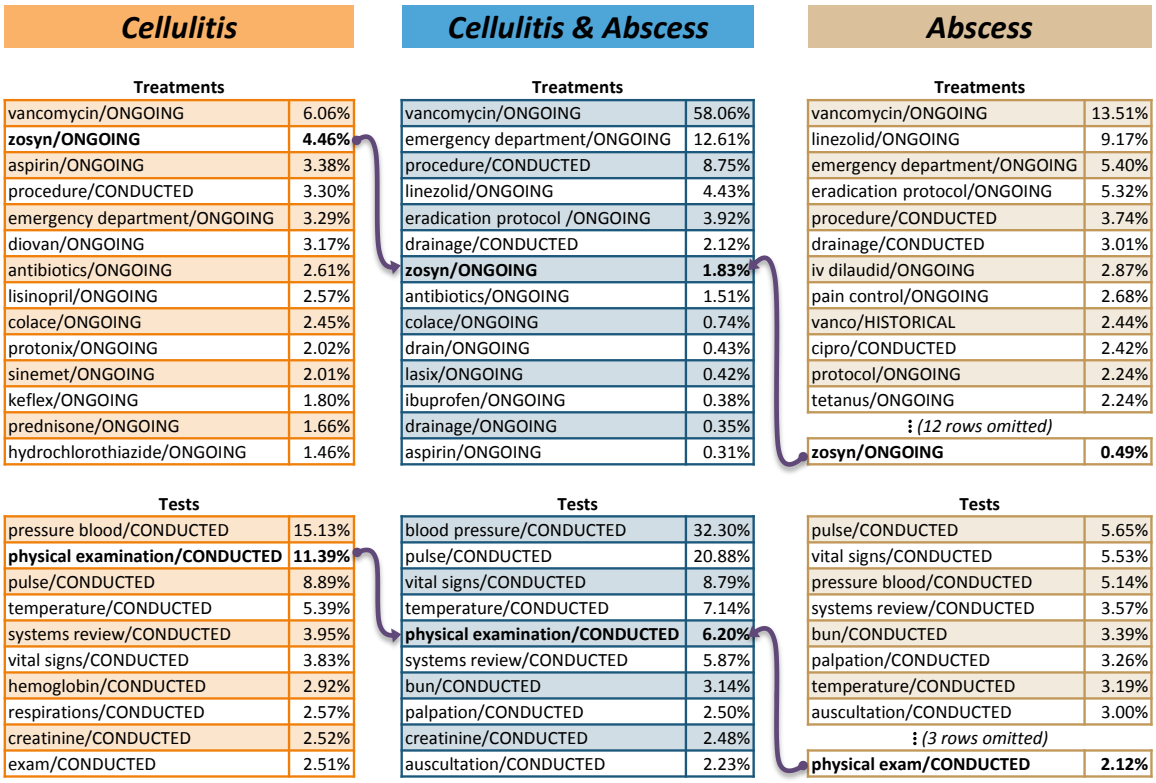

Figure 6: Treatment and test recommendations for present medical problems "cellulitis", "abscess", and both "cellulitis & abscess."

The distributions of the 15 most-likely treatments and 10 most-likely tests for each cohort are illustrated in Figure 6. We have evaluated the recommendations, as shown in Table 2, based on (1) the Infectious Diseases Society of American (IDSA)'s *Practice Guidelines for the Diagnosis and Management of Skin and Soft-Tissue Infectious* (Stevens et al., 2005), (2) Howe and Jones Guidelines for the Management of Periorbital Cellulitis/Abscess (Howe and Jones, 2004), (3) Uzcategui et. al's *Clinical Practice Guidelines for the Management of Orbital Cellulitis* (Uzcategui et al., 1997), and (4) the National Library of Medicine's MEDLINEplus Web Service (Miller et al., 2000).

According to these sources, we achievement a precision within the first 15 treatments of 50% for *cellulitis*, 71% for *cellulitis & abscess*, and 64% for *abscess*. In this measurement, we considered a treatment as relevant if it should be directly associated with the patient cohort. Note: we do not consider treatments for associated symptoms (e.g. pain) as relevant. Additionally, because precision does not take into the probability associated with each item, we have also calculated the *accuracy* of each distribution as the proportion of probability mass assigned to relevant treatments. Using this definition, we achieve an accuracy of 58.2% for *cellulitis*, 98.1% for *cellulitis & abscess*, and 83.6% for *abscess*. Before discussing specific treatments, we list the following abridged definitions from MEDLINEplus:

abscess a pocket of white blood cells, germs, and dead tissues on the skin resulting from an infection.

cellulitis an infection of the skin and underlying tissues caused by bacteria (typically streptococcal).

The most common treatment across all patient cohorts is *Vancomycin* which is the most recommended treatment for methicillin-resistant Staphylococcus aureus (MRSA), the most common cause of cellulitis and abscess. However, after *Vancomycin*, the treatment distributions begin to differ. We have highlighted the treatment *Zosyn* (a mixture of *Piperacillin* and *Tazobactam*) which is an antibiotic approved to treat for infections such as cellulitis and abscess. Despite being commonly given to patients with cellulitis (4.46%, the second highest-ranked treatment), it is ranked twentieth for treating abscess, at only 0.49%. This corresponds to the most typical treatment for abscessing concerning draining the cyst, corresponding to entries four and six. Additionally, more general antibiotics, such as *Linezolid* and *Ciprofloxacin* are more commonly given for abscess, as they treat a variety of underlying infections.

However, for the cohort of patients suffering from both conditions, *Zosyn* rises to position 7 at 1.83% reflecting the fact that it is able to effectively treat both conditions. This shows the ability of the CMN to capture the interaction between treatments for combinations of medical problems.

As our dataset is represented by primarily hospitalized patients (rather than outpatient procedures), many of the recommended treatments are general purpose medications perscriped during the patients hospital stay, such as pain relievers (e.g. *aspirin, ibuprofen, pain control*), stool softeners (e.g. *colace*), diaretics (e.g. *lasix*) and blood thinners (e.g. *lisinopril*).

We have also evaluated the top 10 tests most likely to be conducted for patients in each cohort, as illustrated in Fig-

ure 6. We observed that the likelihood of conducting a physical examination has a distribution rank which varies across all cohorts. Although it is ranked second for cellulitis (at 11.39% likelihood), it is ranked much lower for abscess at position 12 (at 2.12% likelihood). This reflects the recommendation in the guidelines for cellulitis: because cellulitis leaves a patient vulnerable to secondary conditions, a thorough physical examination should be performed. As such, for patients suffering from both *cellulitis & abscess*, the likelihood of conducting a physical examination moves up to rank 5 (6.20%), reflecting the interaction between the two conditions in EMRs.

We also observed that the first three most-commonly conducted tests (i.e. *blood pressure*, *pulse*, and *vital signs*) constitute the majority of the probability mass. This reflects a critical observation on the utility of medical test annotations: that the mere mention of a medical test is not sufficient for statistical reasoning. EMRs document a wide battery of tests and their results for each patient allowing physicians to ascess not only their primary medical problem, but also any secondary conditions or co-morbidities. In order to improve the capability of clinical reasoning enabled by the CMN, the value of tests should be considered and associated with the identification of the mention of each test.

6. Conclusions

In this paper, we show how medical language processing enables the automatic derivation of clinical pictures and therapies for entire patient cohorts. We explain how this knowledge can inform a data-driven probabilistic graphical model on which inference can be performed in a rigorous way for determining the most probable treatments for a given set of medical conditions. Further, we observe that the utility offered by medical test mentions is limited for probabilistic reasoning. Despite this, we evaluated the most likely treatments against (1) the Infectious Diseases Society of American (IDSA)'s *Practice Guidelines for the Diagnosis and Management of Skin and Soft-Tissue Infectious* (Stevens et al., 2005), (2) Howe and Jones Guidelines for the Management of Periorbital Cellulitis/Abscess (Howe and Jones, 2004), (3) Uzcategui et. al's *Clinical Practice Guidelines for the Management of Orbital Cellulitis* (Uzcategui et al., 1997), and (4) the National Library of Medicine's MEDLINEplus Web Service (Miller et al., 2000) and confirmed the validity the probabilistic information encoded by our model.

7. References

Aronson, A. R. (2001). Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association.

Bodenreider, O. (2004). The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267.

Goodwin, T. and Harabagiu, S. M. (2013). The impact of belief values on the identification of patient cohorts. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pages 155–166. Springer Berlin Heidelberg.

Howe, L. and Jones, N. (2004). Guidelines for the management of periorbital cellulitis/abscess. *Clinical Otolaryngology & Allied Sciences*, 29(6):725–728.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Miller, N., Lacroix, E.-M., and Backus, J. E. (2000). Medlineplus: building and maintaining the national library of medicine's consumer health web service. *Bulletin of the Medical Library Association*, 88(1):11.

Ratner, R., Eden, J., Wolman, D., Greenfield, S., and Sox, H. (2009). *Initial national priorities for comparative effectiveness research*. National Academies Press.

Roberts, K. and Harabagiu, S. (2011). A flexible framework for deriving assertions from electronic medical records. *Journal of the American Medical Informatics Association*, 18(5):568–573.

Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., and Chute, C. G. (2010). Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Scheuermann, R. H., Ceusters, W., and Smith, B. (2009). Toward an ontological treatment of disease and diagnosis. *Proceedings of the 2009 AMIA Summit on Translational Bioinformatics*, 2009:116–120.

Stevens, D. L., Bisno, A. L., Chambers, H. F., Everett, E. D., Dellinger, P., Goldstein, E. J., Gorbach, S. L., Hirschmann, J. V., Kaplan, E. L., Montoya, J. G., et al. (2005). Practice guidelines for the diagnosis and management of skin and soft-tissue infections. *Clinical Infectious Diseases*, 41(10):1373–1406.

Stone, P. J., Dunphy, D. C., and Smith, M. S. (1966). The general inquirer: A computer approach to content analysis.

Uzcategui, N., Warman, R., Smith, A., and Howard, C. (1997). Clinical practice guidelines for the management of orbital cellulitis. *Journal of pediatric ophthalmology and strabismus*, 35(2):73–9.

Uzuner, Ö., South, B. R., Shen, S., and DuVall, S. L. (2011). 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Voorhees, E. and Hersh, W. (2012). Overview of the trec 2012 medical records track. In *The Twenty-First Text REtrieval Conference Proceedings (TREC 2012)*, Gaithersburg, MD. National Institute for Standards and Technology. Unpublished. Draft available at http://trec.nist.gov/.

Voorhees, E. and Tong, R. (2011). Overview of the trec 2011 medical records track. In *The Twentieth Text REtrieval Conference Proceedings (TREC 2011)*, Gaithersburg, MD. National Institute for Standards and Technology.