2012 IEEE Second Conference on Healthcare Informatics, Imaging and Systems Biology

An Approach for Incorporating Context in Building Probabilistic Predictive Models

Juan Anna Wu
Biomedical Engineering IDP
Medical Imaging Informatics Group
University of California, Los Angeles, USA
wujuanzh@mii.ucla.edu

William Hsu, Alex AT Bui
Department of Radiological Sciences
Medical Imaging Informatics Group
University of California, Los Angeles, USA
willhsu,buia@mii.ucla.edu

***Abstract*—With the increasing amount of information collected through clinical practice and scientific experimentation, a growing challenge is how to utilize available resources to construct predictive models to facilitate clinical decision making. Clinicians often have questions related to the treatment and outcome of a medical problem for individual patients; however, few tools exist that leverage the large collection of patient data and scientific knowledge to answer these questions. Without appropriate context, existing data that have been collected for a specific task may not be suitable for creating new models that answer different questions. This paper presents an approach that leverages available structured or unstructured data to build a probabilistic predictive model that assists physicians with answering clinical questions on individual patients. Various challenges related to transforming available data to an end-user application are addressed: problem decomposition, variable selection, context representation, automated extraction of information from unstructured data sources, model generation, and development of an intuitive application to query the model and present the results. We describe our efforts towards building a model that predicts the risk of vasospasm in aneurysm patients.**

Keywords-Electronic health record, predictive modeling, clinical decision support, information extraction

I. INTRODUCTION

Clinicians often face the task of making predictions about individual patients based on available, incomplete data for the purposes of risk assessment and treatment selection. A significant challenge in today's healthcare environment is to improve a physician's ability to harness the large amount of information collected about complex diseases to facilitate personalized medicine. Clinical data collected during routine patient care and captured in the electronic health record (EHR) provide a rich source of information about the evolution of a disease for a given patient; this information has been widely used to support diagnostic and prognostic tasks [1, 2]. Additional information from sources such as published scientific literature, existing models (e.g., ontologies, other predictive models), and established databases from previous research efforts (e.g., clinical trials, national repositories) can be used to guide model development by providing additional cases, identifying relevant variables, specifying relationships, and conveying domain knowledge. In order to utilize these data sources to create an accurate prediction model, several challenges must be addressed:

- Given a clinical question, what is the process to decompose it into component parts (e.g., target variable, evidence variables) and translate it into a query that can be answered using the model?
- How should relevant context be represented to inform the selection of variables and specification of the model structure? How can published literature be used to guide variable selection?
- What processing needs to be done to transform data from structured and unstructured sources into a form that can be used to create the predictive model? How should missing data and inherent biases in the population be accounted for?
- Based on the data and task, what is the optimal modeling approach (e.g., logistic regression, neural network, Bayesian network) to obtain accurate predictions?
- What are the requirements for end-user applications that would enable users to intuitively query the model and understand its results? How can the model results be used to influence clinical practice?

Various approaches for utilizing information in the EHR to facilitate knowledge discovery and disease modeling have been demonstrated. For example, Savova [3] discusses an approach that identifies the drug treatment patterns for endocrine breast cancer therapy by combining information extracted from an electronic prescribing database and unstructured clinical narratives using a natural language processing (NLP) tool called clinical Text Analysis and Knowledge Extraction System (cTAKES). Coden [4] presents a system that extracts relevant cancer-related characteristics from pathology reports, mapping the information to a structured knowledge representation that explicitly standardizes and captures the relationships among concepts. These works demonstrate that current NLP techniques are capable of extracting and representing information in the EHR with a high level of reliability, albeit when tailored for specific tasks and domains.

Moreover, other sources such as published literature also provide a rich source of information. [5] describes an automated method for extracting and generating qualitative graphical summaries of treatment for specified disorders described in clinical trials from PubMed. Furthermore, [6] combines information from PubMed and the EHR to create a weighted Bayesian Network Inference model for pancreatic

978-0-7695-4921-7/12 $26.00 © 2012 IEEE
DOI 10.1109/HISB.2012.30

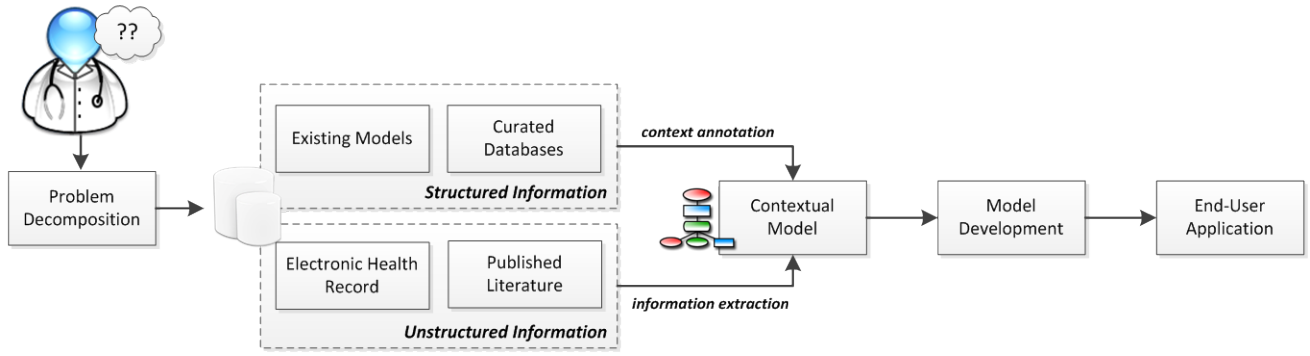

Figure 1. The overall approach for generating probabilistic models from available data sources. Problem decomposition is the process of interpreting complex clinical queries to guide how the predictive model is generated. Two types of data are considered: structured information (curated databases, existing models) and unstructured information (electronic health record, published literature). For unstructured information, information extraction tools are used to extract and standardize the information. For structured information, coxtextual information is captured such as how and why the data was collected. Using information in the structured representation, the probabilistic model can be created and evaluated. Once evaluated, end-user applications can be implemented to interact with the model, assisting the user with interpreting the results.

cancer prediction. Information from PubMed related to each variable is classified as being positive, negative, or neutral and used to compute a weighting factor on the prior probability for that variable. While these works demonstrate the use of structuring and utilizing evidence from literature, the former work does not characterize relationships between variables quantitatively and the latter effort focuses on a high-level characterization of a paper's contents.

This paper presents an approach for developing probabilistic predictive models (e.g., Bayesian belief networks) from structured and unstructured data sources. It includes a standardized representation for incorporating context-related domain knowledge about the disease and the collection of data (e.g., how data are measured, why the information was collected). We demonstrate this work specifically in the context of implementing a predictive model to assess whether an intracranial aneurysm patient is at risk of cerebral vasospasms after endovascular treatment. To build this model, we leveraged sources such as published literature, existing biomedical ontologies, a research database of patient cases from a prior study, and our institution's EHR to identify relevant variables, specify relationships, and compute conditional probabilities. A significant outcome of our work is the development of a structured intermediate representation that integrates available information, capturing the necessary context that may inform the creation of specific models. We discuss our progress towards implementing the overall approach, presenting the performance of our extraction system and initial models and highlighting the challenges encountered.

II. Overall Approach

Figure 1 summarizes the approach for leveraging available data to create predictive models. Each component remains an unsolved problem and much effort from other researchers have been made to address specific challenges in each field. This paper emphasizes the promises and difficulties in putting those components together as a pipeline to facilitate medical decision making.

A. Decomposing Clinical Questions

Modeling tasks are driven by a specific question related to a medical problem. Examples include: "Would my patient be a good candidate for liver transplantation and achieve a positive post-transplant outcome [7]?" and "Does my patient have pneumonia based on findings documented in the admissions note, nursing report, and chest x-ray result [8]?" In each of these cases, the clinical query is decomposed into its component parts such that contributing factors that influence the outcome of interest are modeled. Particularly, questions such as what is the outcome variable, which variables are known factors that influence the outcome variable, and which variables are available in the data need to be addressed. For instance, in [7], the query is decomposed into a model consisting of a binary outcome variable (i.e., 90-day graft survival) comprising two states, yes (alive) or no (death or re-transplantation), and 29 evidence variables identified by domain experts and literature review. Inference on the model is supported by computing the conditional probability tables using data on 12,239 liver transplant events from the United Network for Organ Sharing database.

Generalizing this process, the challenges of decomposing a query include: 1) identifying the outcome (target) variable; 2) determining the structure and parameterization of the model; and 3) mapping the model to information available in the patient record. Currently, the process of defining the model often involves eliciting the structure from domain experts (e.g., collaborating physicians, basic scientists) or automatically learning the structure from available data (e.g., using learning algorithms such as Taboo search). However, both approaches have associated challenges: eliciting models from domain experts may introduce biases based on a person's experiences and beliefs about a disease and its evolution; learning models from data highly depends on the

number of cases available and whether the data accurately represents the target patient population. Hence, we pursue a strategy that utilizes input from both domain experts and published literature. Rather than solely focusing on the variables that have strongest predictive power, we investigate the utility of creating an intermediate representation—a contextual model that captures all relevant variables, relationships, and associated context—that is used to inform the creation of predictive models using a subset of variables from the contextual model. We hypothesize that the intermediate representation can be used to clearly identify potential sources of error, delineate assumptions made about the domain, and provide contextual information for understanding the results from the model.

B. *Aggregating Data Sources*

Access to electronic data characterizing complex diseases have become increasingly available: EHRs capture detailed clinical observations for individual patients; national data repositories provide convenient access to standardized datasets aggregated across multiple institutions (e.g., Cancer Genome Atlas); and scientific literature reports experimental results that provide context for understanding relationships between variables. We briefly describe the challenges of working with each data source:

- Electronic health record: The patient's medical record contains a rich source of information documented about the disease and its presentation in individual patients. A variety of information can be drawn from sources such as clinical reports, images, pathology, and laboratory values. However, a significant issue is that most of the information is unstructured. Clinical reports, for example, are often written in free-text, and findings are inconsistently documented. Images contain quantitative characterizations of the disease (i.e., size, shape of tumor) but require image analysis to extract this information in a reproducible manner.
- Published literature: Scientific papers provide a rich source of evidence summarizing controlled trials and observational studies. In addition to elucidating the underlying mechanisms that may explain clinical observations (e.g., the biological pathway that explains why a treatment may be more effective in some patients and not others), literature also summarizes the latest evidence that supports the effectiveness of a particular treatment that has yet to be adopted widely in clinical practice. As with clinical records, papers are often unstructured. Furthermore, the results are typically summarized in a set of tables and statements demonstrating statistical significance (confidence intervals, p-values) while access to the analyzed dataset is limited. Ongoing research has examined how to unlock this information, but no formalized approach has been proposed on how to effectively leverage information in literature to inform model construction.
- Curated databases: Researchers have created an abundance of structured datasets by curating unstructured data into research databases, collecting data as part of randomized clinical trials, and constructing large repositories of tissue samples and associated data. These repositories are significant in that they provide a large population of cases for a disease spread across multiple geographical sites, allowing modelers to train on an extensive amount of data. However, depending on how the data is collected and aggregated, the consistency of how information is reported may vary. For example, gene expression data may not be normalized across sites and machines, resulting in values that are not comparable without adequate pre-processing.
- Existing models: Given the increase in available data, the use of machine learning techniques has grown to model and classify available data. In addition, qualitative models such as ontologies that capture the variables, relationships, and attributes can be a source of qualitative information in designing a model. For example, the @neurist ontology [9], developed to accommodate data collection across multiple sites in Europe on intracranial aneurysm patients, provides a means to standardize the representation of variables and their states in a model, facilitating the integration of data from additional sites.

While the number of structured data sources is large, there are important considerations that need to be addressed prior to integrating the data into a model. First, existing datasets have typically been collected to answer a specific hypothesis or clinical question and could introduce bias when utilized naively to answer secondary questions. For instance, given a database primarily developed to determine the safety and effectiveness of endovascular coiling, the patient population will not be representative to answer other questions associated with aneurysms when treatment is a risk factor. Therefore, context of original data collection (e.g., purpose of original study, patient eligibility) is needed to identify potential sources of bias. A second consideration is the issue of missing data. Missing data remains a common problem for medical data and may be due to a variety of causes (information intentionally unreported, measurement error, variable added after data collection started). While missing data may be addressed through elimination, imputation or sampling techniques [10], the appropriate technique should take into consideration whether or not the data is missing at random. A potential solution to address the issues of bias and missing data is to leverage information that is captured in unstructured patient records. We discuss the importance and challenges of structuring information in the following section.

C. *Information Extraction*

To incorporate unstructured information from the EHR, we adapted existing open-source NLP tools to automatically identify and extract information from relevant clinical reports. A patient's EHR, a collection of observational data from clinical practice over time, contains a variety of pertinent information such as a patient's medical and family history, clinical presentation, active medications, radiology findings, interventions, and follow-up exams. Information extraction, though, faces significant challenges. How does

one identify the clinical findings when they were not consistently reported? How can individual mentions of terms be mapped to a standard representation? Can attributes associated with each variable be extracted reliably? The task of manually annotating patient records is time-consuming and labor-intensive work. Hence, we are pursuing the development of an NLP-assisted annotation tool that leverages machine-learning techniques to automatically identify and extract values from free-text reports. We are implementing the tool based on the Apache Unstructured Information Management Architecture (UIMA), which is an open source project that provides a flexible architecture for creating annotators to analyze unstructured textual information [11]. Users can combine individual components such as parts-of-speech taggers, regular expressions, and named-entity detectors into a unified processing pipeline.

D. Contextual Model

The central part of our approach is the contextual model, designed to capture and organize relevant variables, relationships, attributes, and associated context that are extracted from available data sources. The contextual model acts as a blueprint for generating the predictive model, providing a comprehensive knowledge source for identifying relevant variables that can be modeled and their relationships. For example, a variable in the model derived from clinical data is associated with a set of attributes describing the real-world entity that the variable represents, how it is measured, and its potential sources of error. For variables that can be linked to the patient record, conditional probability tables can be computed based on available data. For relationships, the model encodes information such as the source of the relationship's identification and any partial statistics (e.g., prior probabilities) that may have been reported. The contextual information encoded in the model may be useful to understanding potential sources of bias from the existing dataset, determine the appropriate approach to address missing data, and guide the construction of a quantitative model based on evidence from literature through a meta-analysis of reported statistics.

Development of the contextual model is performed using a combined top-down, bottom-up approach. The top-down approach incorporates information that may be provided by interviewing domain experts and performing systematic reviews of literature. Domain experts are able to provide a list of variables based on their personal experience. Literature serves as a source of evidence variables and potential (causal) relationships as elucidated through experimentation. Biomedical ontologies, formal representations of domain knowledge, provide an easy way to learn or validate the structure. Alternatively, a bottom-up approach learns relations and estimates probability distributions from large data with data mining techniques. The context along with the relation is of great importance because a change in the context can result in a different relation or no relation at all. For example, a medication may be effective for some patients but can also cause adverse events for others. The contextual model serves to integrate contextual knowledge as annotations for relationships between the treatment variable and other aspects of the model. The model does not attempt to comprehensively model a given domain, but rather, it formalizes what is known about a disease process based on available information to explain the relations and possible confounders between available predictors and outcome variable.

E. Model Development

Given the contextual model, we can utilize it to guide the creation of probabilistic models. Model development proceeds as follows: 1) variable selection, which identifies the significant predictors of the target variable (e.g., using the forward or backward approach to select covariates for a logistic regression model based on the likelihood); 2) modeling approach, which specifies the type of modeling framework to use (e.g., naïve Bayes, support vector machine); 3) network topology, which explicates the relationships between variables; and 4) parameter estimation, which computes the weights or conditional probability tables associated with each variable. Different modeling approaches hold different assumptions of the topology of the model or the data used to train the model, and examining the available data to check if the assumption holds is one of the criteria for selecting an optimal model. In addition, various approaches involve different ways to estimate parameters. Errors in the parameterization process arise due to limitations of the approach as well as the data source.

In this paper, we focus on developing probabilistic graphical models, which can encode a complex distribution over high dimensional space and capture uncertainty inherent in clinical data. Bayesian belief networks (BBNs), a type of probabilistic graph model, have been widely used for prediction tasks for diagnosis and prognosis. Naïve Bayes is a simplified version of BBNs that assumes the attributes are conditionally mutually independent given the class node, thus dramatically decreasing the complexity while still achieving high prediction accuracy in practice. We explore the development of predictive models using variants of the naïve Bayes approach.

F. Model Querying

The final component is an end-user application that allows target users (e.g., physicians, researchers) to interact with the model and pose questions. The user interface should be designed with the following requirements: 1) an intuitive query interface that enables users to easily input patient-specific evidence to generate a prediction that is personalized to individual cases; 2) a result interface that provides sufficient context and explanation for the clinician user to understand how the result was derived; and 3) a retrieval interface that helps users identify and explore other patient cases that are similar based on a measure derived from the probabilistic model itself.

Various interfaces that replace the graph-based representation of the model with more intuitive clinical interfaces have been proposed in the past, including the Visual Query Interface [12] and TraumaSCAN [13]. Other desired features include an intuitive representation of the temporality at patient-, document- and/or event-level.

TABLE I. A DESCRIPTION OF THE VARIABLES CONSIDERED IN THE MODEL BASED ON LITERATURE REVIEW

Variable Name	Type	States	Non-vasospasm (n=64) (min/median/max/mean)	Vasospasm (n=37) (min/median/max/mean)
Age	categorical	<30;31-50;51-70;71-90	16/57/88/58.2	34/51/87/54.5
Gender	binary	Male/Female	Female:73% Male: 27%	Female: 84% Male: 16%
Fisher CT Grade	categorical	1,2,3,4,5	1/2/5/2.5	1/3/4/2.9
Smoking History	binary	Yes/No	Yes:45%, No: 55%	Yes:27%, No: 73%
Aneurysm Location	categorical	Anterior choroidal, anterior cerebral, anterior communicating, basilar tip, basilar trunk, carotid artery…		
Hypertension	binary	Yes/No	Yes: 64%, No: 36%	Yes: 38%, No: 62%
Hyperglycemia	binary	Yes/No	Yes: 55%, No: 45%	Yes: 24%, No: 76%
Aneurysm Neck Size	categorical	Small/Wide/Fusiform	2/4/8/4.13	2/6/8/4.22
Dome Size	categorical	Small/Medium/Large/Giant	7/7/30/9.76	7/7/20/8.85

Medical events are usually reported with temporal information, which can be incorporated into prediction tasks. Such a feature allows users to capture the temporal patterns of events of interest across documents for the same patient and across patients as well, thus increasing the utility.

III. CASE STUDY: CEREBRAL VASOSPASMS

We demonstrate our approach on a specific task: to predict the vasospasm for aneurysm patients who are admitted with a ruptured intracranial aneurysm (ICA).

A. Clinical Problem

Our modeling efforts are geared towards assessing if patients having subarachnoid hemorrhage (SAH) are at risk of experiencing cerebral vasospasms. SAH is often caused by the rupture of an aneurysm in the blood vessel, resulting in an uncontrolled leakage of blood into the surrounding regions. One significant complication in patients with SAH is cerebral vasospasm, a condition in which a blood vessels spasm causes vasoconstriction. It potentially results in reduced cerebral perfusion, cerebral ischemia, potential infarction, and further deterioration of neurological function. The average mortality rate from vasospasm is approximately 33% [14]. Given the high mortality rate and the need to intervene quickly in situations when vasospasm occurs, our goal is to develop a model that predicts whether a patient with a ruptured aneurysm will develop vasospasms.

The process of decomposing this problem is as follows: we identify the target variable to be the occurrence of vasospasm (yes/no). Additionally, we identify evidence variables and their relationships; this information will be derived from various data sources discussed in the following section.

B. Data Sources

Four sources of information were utilized in the construction of our model: an existing curated research database, unstructured reports from our institution's EHR, published literature on PubMed, and domain expert opinion. Collaborating with a clinician at our institution, we were granted access to a research database of patients that is manually maintained for the purpose of assessing the safety and efficacy of endovascular coiling of patients with ruptured aneurysms. The database contained 1,544 patients assessed over a period of a decade and characterized across 36 different variables including demographics, aneurysm location, coiling details, imaging follow-up, and clinical outcome. We also obtained Institutional Review Board approval to review the electronic health records of these patients along with a prospective cohort of patients who present to the clinic with cerebral vasospasm. In addition to available data, we performed a review of existing scientific literature touching upon the pathogenesis of cerebral vasospasms and have consulted clinical collaborators in interventional neuroradiology for feedback on the model and its clinical application.

C. Variable Identification

The exact mechanism by which vasospasm occurs is an active area of study. Using the top-down approach, we performed a PubMed search utilizing the keywords "cerebral vasospasm" and "predictor", which resulted in an initial set of 25 papers. By retrieving papers cited in the initial set, we identified a number of journal articles describing the clinical risk factors related to vasospasm. The amount of subarachnoid blood is a strong predictor of vasospasm [15, 16], which is clinically measured using the Fisher scale [17] based on the first non-contrast computed tomography scan after the patient is admitted. Other severity grading scales, such as the Hunt-Hess grade, World Federation of Neurological Surgeons (WFNS) grade, and modified Rankin scale have also been considered relevant for assessing a patient's risk of vasospasm [18]. Furthermore, correlations between age and vasospasm show that patients under the age of 50 have an increased likelihood of vasospasm occurrence [19] and females tend to have vasospasms more frequently

than men [20]. Medical history also yields important information: hyperglycemia, hypertension, and smoking history have been identified as having a positive correlation with vasospasms [15]. Computed tomography has been used to identify whether specific locations of SAH (e.g., Sylvian fissure) may cause vasospasms more frequently. Perfusion imaging identifies mean transit time and time to peak, which quantifies the severity of vasospasm occurrence [21]. Finally, recent genome wide association studies have identified potential proteins such as endothelin-1 and single-nucleotide polymorphisms that may be implicated in the pathogenesis of vasospasms [22].

We identified an initial set of nine variables considered significant in predicting the occurrence of vasospasm based on the literature review (Table 1). We then examined our curated database to determine whether it would support the creation of a model using these predictors. 6 out of 9 predictors were available, but some variables had a significant number of missing values. 49 of the 1,544 patients were documented as having an incident of cerebral vasospasm. All patients were treated with endovascular coiling. Of the 1,544 patients, 804 presented with SAH. Given the limitations of our existing database, we considered the following:

1) Inherent biases in the data

Given that the population was biased by treatment (all individuals underwent endovascular coiling), the original clinical question was revised to predict the risk of vasospasm given that the patient had undergone endovascular coiling. In addition, a literature review was conducted to assess the association between treatment and vasospasm, which revealed that the choice of treatment (surgery vs. coiling) is independent of vasospasms. The occurrence of vasospasm is often associated with SAH. To eliminate the confounding effect that vasospasm may have on our model, we decided to include only patients that present with SAH in our model.

2) Issues related to missing data

As mentioned earlier, not all entries in the curated database contained complete information. Our initial attempt at addressing this issue was to re-review the patient records to determine whether the information could be extracted from clinical reports. In addition, we also examined the patient records to extract values of variables that were not captured as part of the original data collection. Hypertension and smoking history were consistently reported in the patient's record. Hyperglycemia, while not consistently mentioned in the clinical reports, could be interpreted from the laboratory values or indirectly based on whether the patient has diabetes. We determined that 12 of 49 patients with vasospasm did not have identifiers that would allow us to re-review their records and thus, data for these cases would not be missing at random. Therefore, we removed those cases, resulting in a total of 37 vasospasm patients. Our review of literature revealed that the incidence rate of vasospasm among patients presenting with SAH is approximately 30% [23], and we randomly selected an additional 64 cases of non-vasospasm patients. Thus, our patient population consists of 101 cases characterized across 9 variables, as summarized in Table 1.

D. Information Extraction

To facilitate the extraction of relevant variables, we have created an automated tool that utilized a combination of regular expressions and dictionary lookup to identify mentions of variables and extract pertinent values from free-text reports. For our vasospasm model, we have adapted the tool to identify and extract values related to smoking history, hypertension, and hyperglycemia. The tool, which is built upon the UIMA framework, contains a collection of annotators that tokenize documents into sentences followed by pattern matching and dictionary lookup to identify relevant variables and extract associated values. A filtering annotator removes documents that do not contain information about smoking history, hypertension, and hyperglycemia and removes any markups encoded in the document (i.e., hypertext markup language tags). A sentence detector adopted from OpenNLP's maximum entropy-based approach is used. Regular expressions are used to identify variables that can be characterized using a finite set of patterns (e.g., whether a patient is a smoker, size of aneurysm). A subset of records has been manually reviewed to enumerate possible ways a variable can be reported. For more complex variables (e.g., aneurysm location), we leverage a part-of-speech annotator to identify noun phrases that can then be classified using a dictionary lookup annotator. For example, the anatomical location "inferior cerebellar artery" can also be expressed as the abbreviation ICA or more specifically, PICA (posterior inferior cerebellar artery). Finally, we utilize an approach based on ConText [24] to determine whether a variable is negated or historical.

To briefly describe the performance of our annotator, the total precision, recall, and accuracy of the annotator in extracting values of variables listed in Table 1 are 83.3% (115/138), 97.5% (115/118), and 91.4% (277/303), respectively. Several reasons for false positives have been identified: 1) the annotator was not able to identify negations perfectly. Sentences such as, "she is apparently a nonsmoker and nondrinker," or "glucose will be tightly regulated to prevent hyperglycemia," would be incorrectly identified as positive because the "non" prefix and "prevent" are not accounted for; and 2) the annotator was not able to assign the experiencer consistently (e.g., conditions reported about other family members are incorrectly assigned to the patient). The false negative rate is small (3/303).

E. Contextual Model

An important source of domain knowledge that is translated into the variables, relationships, and parameters of the model comes from results of controlled trials and experimental studies. We are developing methods to incorporate results from published research studies [25]. Each study is manually reviewed, extracting contextual information such as: 1) study hypothesis and observed (outcome) variables; 2) study population characteristics; 3) experimental method, including specifics of assay techniques, specific platforms, normalization methods; 4) statistical tests; and 5) study conclusion, including hypothesized pathways and explanations relating the study variables and outcomes. The information is used to

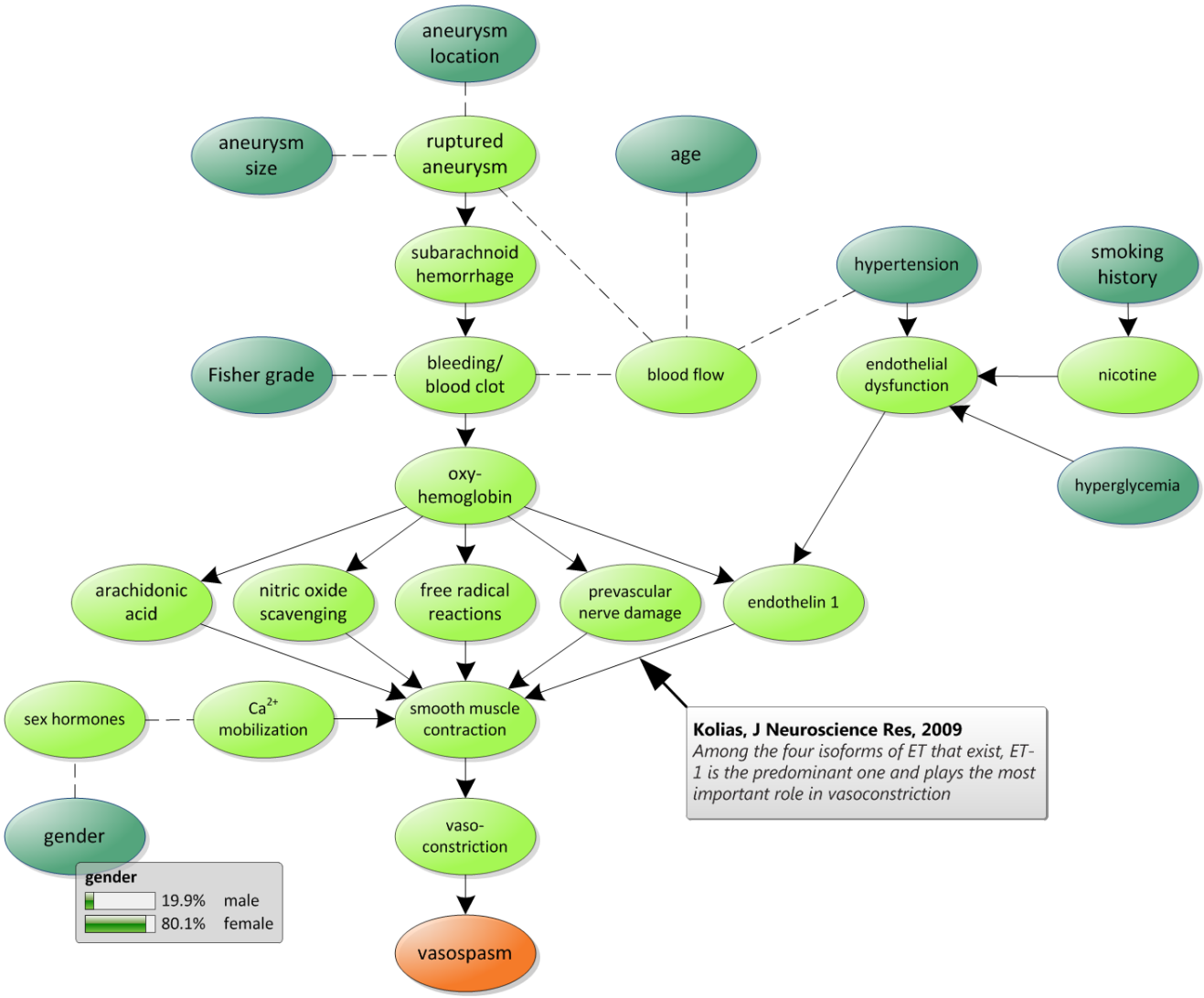

Figure 2. A schematic of the structured representation for explaining the factors related to cerebral aneurysms. The representation formalizes the contextual knowledge drawn from literature and domain expert opinion. Arrows indicate relationships that are drawn from experimental literature. Dotted lines represent postulated relationships. Dark nodes represent variables with available data.

supplement variables derived from available data with domain knowledge that can potentially be used to generate explanations.

Figure 2 depicts our current understanding of cerebral vasospasms, placing the variables derived from available data in the context of relationships and other factors identified from literature review. Using PubMed, we conducted a literature search using keywords "pathogenesis of vasospasm", "subarachnoid hemorrhage", and "vasospasm". Papers that elucidate the mechanism between ruptured aneurysms to cerebral vasospasm at the tissue and molecular levels were reviewed manually. The goal was to determine the underlying biological factors that are associated with the identified nine predictors. For example, [26, 27] were used to clarify the relationship between gender and vasoconstriction, which were found to be influenced by sex hormones, Ca^{2+} mobilization, and smooth muscle contraction. Dark green nodes represent variables that are associated with conditional probabilities computed from available data. Solid edges represent direct relationships documented in literature, while dotted edges represent relationships postulated by domain experts that do not have supporting evidence from literature.

While the contextual model incorporates variables representing clinical observables, image findings, pathology, treatments, and genomic analysis, our focus is currently to build a model from clinical observables. The model is represented as a concept graph explicitly capturing variables, relationships, and attributes. A subset of variables may be associated with a conditional probability table, if data to compute the probability distribution is available. We are exploring methods to estimate probabilities from structured patient data (using algorithms such as expectation maximization) or from values reported in literature using statistical meta-analysis.

F. Model Creation

Four different approaches were used to generate predictive models from the structured data: logistic regression, naïve Bayes, augmented naïve Bayes, and tree-augmented naïve Bayes. Two types of experiments were conducted: 1) varying the number of predictors used to train each of the models; 2) using one set of predictors, and comparing the performance of different models. BayesiaLab was used to train the naïve Bayes model and its variants, and IBM SPSS Statistics for the logistic regression model. We selected subsets of predictors based on the strength of correlation between the variable and outcome by computing the Pearson correlation coefficient. Three subsets of predictors were generated: 1) all 9 predictors listed in Table 1; 2) 6 predictors that exclude aneurysm location, dome size, and neck size; and 3) 3 predictors-hypertension, hyperglycemia, and Fisher CT grade.

We randomly selected 81(80%) patients to form the training dataset with the remaining 20 patients used as test cases. The average precision and area under the receiver operating characteristic (AUROC) curve was obtained using a 5-fold cross-validation. Our initial results demonstrated that the tree-augmented naïve Bayes approach achieved the best performance of all of the models, obtaining a total precision of 90% and an AUROC value of 0.925 with predictors: hypertension, hyperglycemia, and Fisher CT grade.

IV. Discussion

In this paper, we have discussed an approach for utilizing information from multiple data sources to generate a predictive model. We have described our efforts towards implementing a probabilistic model for predicting risk of vasospasm for patients with subarachnoid hemorrhage.

Throughout this paper, we have emphasized the important role that context plays in the entire process. Context provides: 1) evidence for explicitly modeling the relationships between variables; 2) a means for ensuring that the available data is being interpreted properly (i.e., the motivation and method by which a variable is derived); and 3) a method for integrating evidence across multiple biological scales and generating explanations based on this information.

We have encountered several challenges related to utilizing unstructured clinical records to generate the model through our initial efforts:

- Uncertainty. An approach to capture and characterize the inherent uncertainty in clinical notes is needed. For example, sentences such as "she most likely has pulmonary venous hypertension" and "the patient is a 72-year-old female with a history of a possible systemic hypertension" are ambiguous as to whether the patient has a specific condition. One approach to resolving this issue is developing methods to co-reference information extracted from other time points to utilize other mentions as a way of confirming whether the patient has a condition.
- Conflicting information. Clinical notes may contain conflicting information across different documents. For example, one document reported "she has no history of alcohol abuse, but she does smoke approximately half to one pack a day for several years." For the same patient, another document said "social history: she denies any use of tobacco, alcohol, or recreational drugs." Based on these statements, it is unclear whether the patient is a smoker. Extending the contextual model to support multiple values of a variable across time could provide sufficient context to determine the correct value.
- Temporality. While clinical notes contain temporal information, our current representation does not incorporate time-varying variables. The change of certain variable values (e.g. aneurysm growth rate) is potentially significant predictors of the target variable. We are presently extending the contextual model to capture temporal information, providing support for the creation of models that are time-varying (e.g., dynamic Bayesian belief networks).
- Sample size. Some of the trends reported in literature conflicted with the trends identified in the patient population for this case study. For example, literature reveals that smoking history is a positive predictor of vasospasm, but in our dataset, we find the opposite correlation. We believe this is the result of our population being small and consisting overwhelmingly of non-smokers. However, these discrepancies also emphasize the need to capture sufficient context to determine whether a study's conclusions (e.g., based on its experimental design) are generally applicable to other populations.

Several future research directions will be addressed, including:

- Error propagation. Each component introduces some error, which may be a result of human error, inherent uncertainty of the approach, or the nature of the problem. For example, when considering data from a limited population such as a single institution, the difference between the sample population and the national population leads to errors. When extracting information from the EHR, the annotator may not completely extract relevant information leading to poor recall. During model development, using algorithms that are not well-suited for the type of data being examined may lead to low accuracy and incorrect network topologies. These errors accumulate as each component of our approach is performed. We plan to examine how error can be characterized at each step and identify ways of addressing them so that their propagation is limited.
- Multi-scale disease models. Besides clinical observations documented in the textual portions of the EHR, other data types such as medical imaging, pathology, and genomics provide opportunities to further elucidate the biological mechanisms that may improve the prediction accuracy of developed models. To combine clinical observations with variables derived from other biological scales (i.e., tissue/organ, cellular, molecular), an approach such as hierarchical Bayesian

networks may be explored [28]. Furthermore, the numerous uncertainties that remain about how different biological levels are related emphasize the need of a contextual model to capture relevant information.

The presented work represents our initial attempt towards developing a methodical approach to promote individually tailored medicine. One potential application of this work is to utilize the developed models to identify similar patient cases, utilizing the underlying probability distributions as a measure of similarity. For example, given a particular instantiation of the model, how can we predict a patient's survival time based on similarities to the probability distribution of past patient cases generated by the model? Our group has pursued applying this approach to build predictive models for other complex diseases such as non-small cell lung cancer and glioblastoma multiforme, an aggressive form of brain cancer.

ACKNOWLEDGMENTS

The authors would like to thank Drs. Ricky Taira, James Sayre, and Fernando Vinuela for insightful discussions and domain expertise. We would also like to acknowledge the individuals who reviewed and provided feedback on this manuscript. This work is funded in part by the National Institute of Biomedical Imaging and Bioengineering 5R01EB000362.

REFERENCES

[1] B. E. Himes, Y. Dai, I. S. Kohane, S. T. Weiss, and M. F. Ramoni, "Prediction of chronic obstructive pulmonary disease (COPD) in asthma patients using electronic medical records," *J Am Med Inform Assoc,* vol. 16, no. 3, pp. 371-379, 2009.

[2] S. Pakhomov, N. Shah, P. Hanson, S. Balasubramaniam, and S. A. Smith, "Automatic quality of life prediction using electronic medical records," in AMIA Annu Symp Proc, 2008, pp. 545-9.

[3] G. K. Savova, J. E. Olson, S. P. Murphy, V. L. Cafourek, F. J. Couch, M. P. Goetz, J. N. Ingle, V. J. Suman, C. G. Chute, and R. M. Weinshilboum, "Automated discovery of drug treatment patterns for endocrine therapy of breast cancer within an electronic medical record," *J Am Med Inform Assoc,* vol. 19, no. e1, pp. e83-9, 2012.

[4] A. Coden, G. Savova, I. Sominsky, M. Tanenblatt, J. Masanz, K. Schuler, J. Cooper, W. Guan, and P. C. De Groen, "Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model," *J Biomed Inform,* vol. 42, no. 5, pp. 937-949, 2009.

[5] M. Fiszman, D. Demner-Fushman, H. Kilicoglu, and T. C. Rindflesch, "Automatic summarization of MEDLINE citations for evidence-based medical treatment: A topic-oriented evaluation," *J Biomed Inform,* vol. 42, no. 5, pp. 801-813, 2009.

[6] D. Zhao, and C. Weng, "Combining PubMed knowledge and EHR data to develop a weighted bayesian network for pancreatic cancer prediction," *J Biomed Inform,* vol. 44, no. 5, pp. 859-68, 2011.

[7] N. Hoot, and D. Aronsky, "Using Bayesian networks to predict survival of liver transplant patients," in Proc AMIA Symp, 2005, pp. 345-9.

[8] D. Aronsky, M. Fiszman, W. W. Chapman, and P. J. Haug, "Combining decision support methodologies to diagnose pneumonia," in Proc AMIA Symp, 2001, pp. 12-6.

[9] M. Boeker, H. Stenzhorn, K. Kumpf, P. Bijlenga, S. Schulz, and S. Hanser, "The@ neurIST ontology of intracranial aneurysms: providing terminological services for an integrated IT infrastructure," in Proc AMIA Symp, 2007, pp. 56-60.

[10] J. L. Schafer, and J. W. Graham, "Missing data: our view of the state of the art," *Psychol Methods,* vol. 7, no. 2, pp. 147-77, 2002.

[11] D. Ferrucci, and A. Lally, "UIMA: an architectural approach to unstructured information processing in the corporate research environment," *Natural Language Engineering,* vol. 10, no. 3-4, pp. 327-348, 2004.

[12] W. Hsu, A. A. T. Bui, R. K. Taira, and H. Kangarloo, "Integrating Imaging and Clinical Data for Decision Support," *Handbook of Research on Advanced Techniques in Diagnostic Imaging and Biomedical Applications*, 2009.

[13] O. I. Ogunyemi, J. R. Clarke, N. Ash, and B. L. Webber, "Combining geometric and probabilistic reasoning for computer-based penetrating-trauma assessment," *J Am Med Inform Assoc,* vol. 9, no. 3, pp. 273-82, 2002.

[14] R. L. Macdonald, *Cerebral vasospasm: advances in research and treatment*: Thieme Medical Pub, 2005.

[15] C. G. Harrod, B. R. Bendok, and H. H. Batjer, "Prediction of cerebral vasospasm in patients presenting with aneurysmal subarachnoid hemorrhage: a review," *Neurosurgery,* vol. 56, no. 4, pp. 633-54, 2005.

[16] J. G. de Oliveira, J. Beck, C. Ulrich, J. Rathert, A. Raabe, and V. Seifert, "Comparison between clipping and coiling on the incidence of cerebral vasospasm after aneurysmal subarachnoid hemorrhage: a systematic review and meta-analysis," *Neurosurgical review,* vol. 30, no. 1, pp. 22-31, 2007.

[17] J. A. Frontera, J. Claassen, J. M. Schmidt, K. E. Wartenberg, R. Temes, E. S. Connolly, R. L. Macdonald, and S. A. Mayer, "Prediction of Symptomatic Vasospasmafter Subarachnoid Hemorrhage: the Modified Fisher Scale," *Neurosurgery,* vol. 59, no. 1, pp. 21-7, 2006.

[18] E. Bor-Seng-Shu, M. de-Lima-Oliveira, M. J. Teixeira, and R. B. Panerai, "Predicting symptomatic cerebral vasospasm after aneurysmal subarachnoid hemorrhage," *Neurosurgery,* vol. 69, no. 2, pp. E501-2, 2011.

[19] C. Charpentier, G. Audibert, F. Guillemin, T. Civit, X. Ducrocq, S. Bracard, H. Hepner, L. Picard, and M. C. Laxenaire, "Multivariate analysis of predictors of cerebral vasospasm occurrence after aneurysmal subarachnoid hemorrhage," *Stroke,* vol. 30, no. 7, pp. 1402-1408, 1999.

[20] M. Hohlrieder, M. Spiegel, J. Hinterhoelzl, K. Engelhardt, B. Pfausler, A. Kampfl, H. Ulmer, P. Waldenberger, I. Mohsenipour, and E. Schmutzhard, "Cerebral vasospasm and ischaemic infarction in clipped and coiled intracranial aneurysm patients," *Eur J Neurol,* vol. 9, no. 4, pp. 389-399, 2002.

[21] V. Lefournier, A. Krainik, B. Gory, F. Derderian, P. Bessou, B. Fauvage, J. F. Le Bas, and J. F. Payen, "Perfusion CT to quantify the cerebral vasospasm following subarachnoid hemorrhage," *J Neuroradiol,* vol. 37, no. 5, pp. 284-291, 2010.

[22] H. Kim, E. Crago, M. Kim, P. Sherwood, Y. Conley, S. Poloyac, and M. Kerr, "Cerebral vasospasm after sub‐arachnoid hemorrhage as a clinical predictor and phenotype for genetic association study," *Int J Stroke*, 2012.

[23] K. Oyama, and L. Criddle, "Vasospasm after aneurysmal subarachnoid hemorrhage," *Nurs Crit Care,* vol. 24, no. 5, pp. 58-67, 2004.

[24] H. Harkema, J. N. Dowling, T. Thornblade, and W. W. Chapman, "ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports," *J Biomed Inform,* vol. 42, no. 5, pp. 839-851, 2009.

[25] M. Tong, J. Wu, S. Chae, A. Chern, W. Speier, W. Hsu, A. Bui, and R. Taira, "Computer-assisted systematic review and interactive visualization of published literature," in Radiological Society of North America (RSNA) Annual Meeting, Chicago, IL, 2010.

[26] J. K. Crews, and R. A. Khalil, "Gender‐Specific Inhibition of Ca2+ Entry Mechanisms of Arterial Vasoconstriction by Sex Hormones," *Clin Exp Pharmacol Physiol,* vol. 26, no. 9, pp. 707-715, 1999.

[27] G. M. Rubanyi, A. Johns, and K. Kauser, “Effect of estrogen on endothelial function and angiogenesis,” *Vascul Pharmacol,* vol. 38, no. 2, pp. 89-98, 2002.

[28] E. Gyftodimos, and P. A. Flach, "Hierarchical bayesian networks: A probabilistic reasoning model for structured domains." pp. 23-30.