2011 First IEEE International Conference on Healthcare Informatics, Imaging and Systems Biology

Applying an Instance-specific Model to Longitudinal Clinical Data for Prediction

Emily Watt, MLIS, James W. Sayre, PhD, Alex A.T. Bui, PhD
Medical Imaging Informatics, Dept. of Radiological Sciences
University of California, Los Angeles
Los Angeles, California
e-mail: ewatt@ucla.edu

Abstract – **Dynamic Bayesian Belief networks (DBNs) have been commonly used to represent temporal data in several domains; however, an ideal representation requires a near perfect mapping between the process being modeled and the DBN. Furthermore, DBNs assume a full set of observations collected at a fixed frequency. Bayesian model selection has arisen to address biased inference and underlying assumptions about the data (e.g., distribution, representativeness) to choose a model that best fits the given observations. Per patient case, a Bayesian model is generated to maximize specificity, and the collective set of models is averaged to fit all examples. This paper demonstrates the advantages of patient-specific modeling over a DBN-driven approach. Results evaluating this approach are presented based on models for two longitudinal clinical datasets (neuro-oncology, knee osteoarthritis). Largely, the patient-specific models show improved performance in prediction relative to the DBNs.**

Keywords-temporal modeling; Dynamic Bayesian Belief network; Bayesian model averaging; state-model; data mining; imputation; resampling

I. INTRODUCTION

Despite early efforts in the 1970s [1, 2] that explored the incorporation of uncertainty and conflicting information in decision making, quantitative models built from clinical datasets to support predictive tasks have yet to fully accommodate the unique presentation of a patient's characteristics and the assumptions made on clinical data (e.g., distribution, representativeness). Indeed, medical decision-making models are complicated by: the high-dimensional nature of clinical data; the varying presentation of patients and disease over time at the individual and global levels; and the trade-offs amongst multiple decision alternatives [3]. Other issues associated with the development of predictive models include the uncertainty with how data are collected and interpreted to construct the model; the effects of missing data (which is typical of routine clinical environments); and the evaluation of potential relationships between a patient's variables and presumed knowledge of patients with the same disease.

This paper contrasts the population-wide and instance-specific approaches to constructing a Bayesian belief network (BBN) from clinical, temporally-oriented datasets. Specifically, the model parameters (e.g., conditional probability tables) are learned in the context of missing data for: 1) a dynamic belief network (DBN), which is used as a population-wide temporal model; and 2) an instant-specific patient model developed using a Lazy Bayes Rule (LBR). The performances of the DBN- and LBR-based models are compared in predicting the evolution and longitudinal presentation of chronic diseases (brain cancer, knee osteoarthritis). The remainder of this paper is organized as follows. Section II reviews related work addressing challenges to the construction of context-specific models from clinical data. Section III describes the test datasets, and the methodology used to construct the DBN- and LBR-based models. Section IV presents the evaluation results and comparison of the two models' predictive power. Finally, Sections V and VI provide a discussion of the results and conclude with potential future work.

II. BACKGROUND AND PREVIOUS WORK

Classic statistical and machine learning methods typically induce a model from a set of observed training data, which is subsequently applied to prospective cases. Such approaches establish a *population-wide model*, as they use a group of previously seen cases as training data to derive an optimal set of relationships and statistical/probabilistic features to fit all prospective test cases. More recent research [4] has suggested the use of *instance-specific models*, which use features about a particular instance, in addition to information from previously seen cases, to train an optimal, local model for each observed training case. Learning models that are specific to the particular features of a given patient case has been shown to improve predictive performance [4]. Furthermore, instance-specific modeling has the potential to explain the disparities innate to disease progression and to an individual patient case for improved patient-centric treatment; ultimately, these methods may improve the predictive power of medical decision-making tools by more accurately portraying the context unique to a patient's record.

A. Instance-specific models

The traditional, *a priori* predictive model used in various domains of biomedical research is commonly generalized to the baseline set of patients in a disease population, but is not tailored to an individual patient. In [5] patient-specific models were tailored to the generic anatomic model with information about the specific individual. Knowledge about each nodule from a patient's baseline exam was used to update the generic features in the *a priori* model with patient-specific information (e.g., nodule volume, shape, location), reducing the uncertainty associated with information known about a specific patient. In the area of biological modeling,

978-0-7695-4407-6 2011

DOI 10.1109/HISB.2011.12

instance-specific models are used as a form of case-based reasoning, transferring a solution from a previously seen, most "similar" case, to the case at hand. The solution is then re-adapted per the outcome of its application to the new case and differences between the current case and "historical" precedent are accounted for [6].

B. Dynamic Bayesian Networks

To formulate a comprehensive patient history, modeling temporal relationships throughout the record is necessary. Unfortunately the observations made during routine patient care may not necessarily conform across different individuals (i.e., observations may not be acquired at fixed points in time). Dynamic Bayesian networks (DBNs) [7-10] have been used to model the longitudinal aspects of (clinical) data over time. A DBN extends a classic BBN to model probability distributions over semi-finite collections of random variables, separated by input, hidden, and output variables of a state-space model [7]. In application, a DBN can be defined as a pair of Bayesian networks, where one BBN defines the prior probability of a variable $P(Z_1)$ in the system and the second BBN is a two-slice temporal belief network (TBN) defining the conditional probability of the variable using a directed acyclic graph (DAG) [8]. As a generic model, a DBN may not effectively describe previously unobserved data, due to the variation and uncertainty in clinical observations between patients within the same cohort.

C. Bayesian Model Averaging

Another problem with the DBN is that it utilizes a single model as an approximate best fit for the observed data. Notably, value-specific independencies are not modeled. Inferring from a single model may fail to account for information about effect sizes and predictions [11]; and different independence relationships may hold under only certain configurations of the network. In contrast, patient-specific models can capture context-specific independencies [12].

To this end, applying Bayesian model averaging (BMA) [13, 14] has been shown to provide overall better predictive ability than using a single model. The *patient-specific Markov blanket local structure* (PSMBl-MA) algorithm [15] is a specific type of patient-specific model-averaging (PSA) method that learns Markov blanket (MB) models using decision graph conditional probability distributions. PSMBl-MA derives the posterior distribution $P(Z^t|x^t, D)$ for the target variable Z^t, given the values of the other variables, $X^t = x^t$ for the case at hand and set of training data, D. The computation of the posterior distribution $P(Z^t|x^t, D)$ by Bayesian model averaging is given by:

$$P(Z^t|x^t, D) = \sum_{G \in M} P(Z^t|x^t, G, D)P(G|D)$$

where the sum is taken over all MB structures G, in the model space M. The first term in the summation, $P(Z^t|x^t, G, D)$, is the conditional probability $P(Z^t|x^t)$ compared with a MB that has structure G with parameters that are estimated from training data, D. The term $P(G|D)$ is the posterior probability, or weight, of the MB structure G given D. The conditional probability is thus derived from the weighted average of the posterior probabilities for all MB structures. This approach proffers a more complete model space, by averaging over a suitable set of models, in addition to value-specific relationships.

D. Lazy Bayesian Rule

The naïve Bayesian tree learner, *NBTree* [16], was developed to improve the performance of a naïve Bayesian classifier by incorporating decision tree learning methods. Although *NBTree* lessens the effect of the interdependence assumptions made by naïve Bayesian classifiers, it suffers from replication and fragmentation problems due to small disjuncts. The algorithm builds one tree that best fits all examples on average, but often the inadequate number of training examples at the leaves of the tree fails to describe the full breadth of examples [17].

The Lazy Bayes Rule (LBR) algorithm more recently emerged from the application of lazy learning to Bayesian tree induction [18], and was utilized to support patient-specific model selection. The LBR algorithm stores input training examples and only invokes an ideal Bayes' rule when classifying an unseen case [19]. The antecedent of the Bayesian rule is defined by a subspace of the instance space to which the test case belongs. The subspace is comprised of available training examples selected by the LBR algorithm, which is then used as a source of training data for the consequent of the Bayes rule, a local naïve Bayes classifier with which the test case is classified. Thus no explicit decision trees or rules are built at training time, which enables more instance-specific modeling. Hence, the ultimate objective of LBR is to grow the antecedent (attribute-value pairs, or conditions) of the Bayesian rule that best matches the test case, while decreasing the error of the local naïve Bayesian classifier in the consequent of the rule. The LBR algorithm has demonstrated a lower error rate than any alternative algorithm, as it relaxes conditional feature independence assumptions [19].

III. METHODS

The objective of this work is to evaluate DBN- and LBR-based models representing clinical information over time, comparing population-wide and instance-specific learning of the underlying conditional probabilities in the presence of missing data. The topology of the DBN was assumed to be known, specified by experts familiar with the data domain. Model variables, associated states, and relationships between variables were identified based on published literature and clinical relevance.

A. Data Collection and Preprocessing

Two different datasets were used: 1) a set of medical records for 50 neuro-oncology patients seen at the UCLA Medical Center, with confirmed diagnoses of glioblastoma multiforme (GBM); and 2) a combination of three datasets (MedHist00-05, Biomarkers00-05, JointSx00-05) from the Osteoarthritis Initiative (OAI), a database of over 4,700 subjects available for public access at http://www.oai.ucsf.edu. The former represents raw, clinical data with (unstructured)

free-text reports from radiology, pathology, laboratory, surgery, and oncology, in addition to more structured laboratory and demographic-related information. The latter is a collection of data tables covering a longitudinal observational study, which has information for subjects assessed at baseline and at specific follow-up visits (12-, 18-, 24-, 30-, and 36-months). OAI's datasets include demographics, imaging, and bio-specimen information related to the diagnosis and treatment of knee osteoarthritis (OA). While GBM is not standardized to any fixed time interval, the total number of reports provided a solid starting point for the amount of data needed to train our proposed models. The OAI data was chosen for its structured characteristic, and standardization to a fixed time frame and intervals. As such, patients could be characterized by a specific disease "stage," or set of features with a unique arrangement of values. The OAI data tables were chosen based on their longitudinal characteristics conducive to tracking variable changes over time.

For the UCLA GBM model, 2,970 reports were downloaded and processed: for each patient, the body of each free-text report was extracted using file processing and parsing algorithms. The reports for each patient were chronologically ordered, and then assembled into a single file for all patients. Words/phrases occurring with high frequency across the records were highlighted and included in a keyword list. These keywords were used to identify the high-level features or variables that a domain expert defined for the model. Each report was then manually reviewed to extract changes in state, or the associated text, for a variable over time. States for the level features *contrast* and *symptoms*, for example, were represented as "improving," "worsening," "stable/no change," or "recurrent"; change was characterized by the presence or replacement of a different keyword. To signify the clinical progression of disease, a particular configuration of feature values collected per patient represented data for only one time slice (e.g., a patient newly diagnosed will not have recurrence). Thus, approximately 80-90% of data was missing across all variables. To address this issue, last observation carried forward (LOCF) was used to impute missing values in each patient case per time slice, with the assumption that data was missing at random and unchanged since the last observation. LOCF has been used for analysis [20] because of its simplicity, ease of implementation, and the assertion that the bias resulting from carrying observations forward yields a "conservative" analysis. The data was then re-sampled using non-parametric, iterative bootstrapping (i.e., case resampling) to bolster statistical power [21]. A new sample was drawn by applying LOCF to each case that met a predefined fixed time interval. For example, after every fourth day a patient observation was made, the last original observation following a four-day span was carried forward to derive a bootstrap sample. This method generated an additional 2,980 bootstrap examples using a week-long time interval for resampling.

The OAI dataset was already standardized. With the exception of a quality of life (QoL) variable being nominal and multidimensional, all other chosen features were binary, and no variables required discretization. As some subjects had no data recorded in certain follow-up studies, data for all patients were combined into a single file using file processing scripts. Approximately 8% of the data were missing; in contrast to the GBM dataset, null value placeholders were substituted in these cases, rather than impute new values, to better preserve the underlying distribution of the observed data.

TABLE I. FEATURES EXTRACTED FROM THE GBM DATASET

Feature	Attributes
Contrast	New, stable, worsening, improving, recurrent
Symptoms	New, stable, worsening, improving
Chemotherapy	Temodar, Avastin, Tarceva, carboplatin, BCNU, etoposide, irinotecan, other, combination
Radiation	Yes, no
Medication	Decadron, Dilantin, Keppra, Kytril, synthroid, Ativan, combination
Performance	S1, S2, S3, S4
Complications	Hemorrhoids, pneumonia, breast nodule, seizure, other, paralysis, aphasia, lung complication, sepsis, deep vein thrombosis (DVT), UTI
Labs	Yes, no
Tumor location	Frontal lobe, temporal lobe, parietal, occipital, tempo-roparietal, fronto-temporal, fronto-parietal, temporo-occipital, occipito-parietal
Surgical pathology	Yes, no
Resection	First, second, third, fourth
GBM	Yes, no

TABLE II. FEATURES EXTRACTED FROM THE OAI DATASET

Feature	Attributes
Knee pain	Yes, no
Knee replacement	Yes, no
Quality of life	None, mild moderate, severe, extreme
Pain medication	Yes, no
History of arthritis	Yes, no
Knee arthritis	Yes, no

B. Model Construction, Evaluation Metrics

A different model was generated for each dataset. Observations for the GBM model (Fig. 1a) were not standardized to fixed time intervals, so a three-slice temporal model was defined for this initial work: initial diagnosis and treatment (T = 1); first tumor recurrence (T = 2); and an ensuing change in status, either improvement or progression (T = 3). These stages characterize the disease's progression, which was marked by a transition in the patient's presentation of variables. The GBM model contains 12 multivariate and binary variables (Table 1) measured at each stage, for a total of 36 variables. The model can thus be seen as a static BBN replicated for the number of total time slices, with arcs between slices that only point forward in time. For the OA model (Fig. 1b), we used a total of 38 variables across a six-slice model. Not all variables (Table 2) were collected at each time stage: some variables measured at baseline were explicitly not measured in every follow-up study. Other variables not measured at baseline but collected in follow-up studies were excluded in this model.

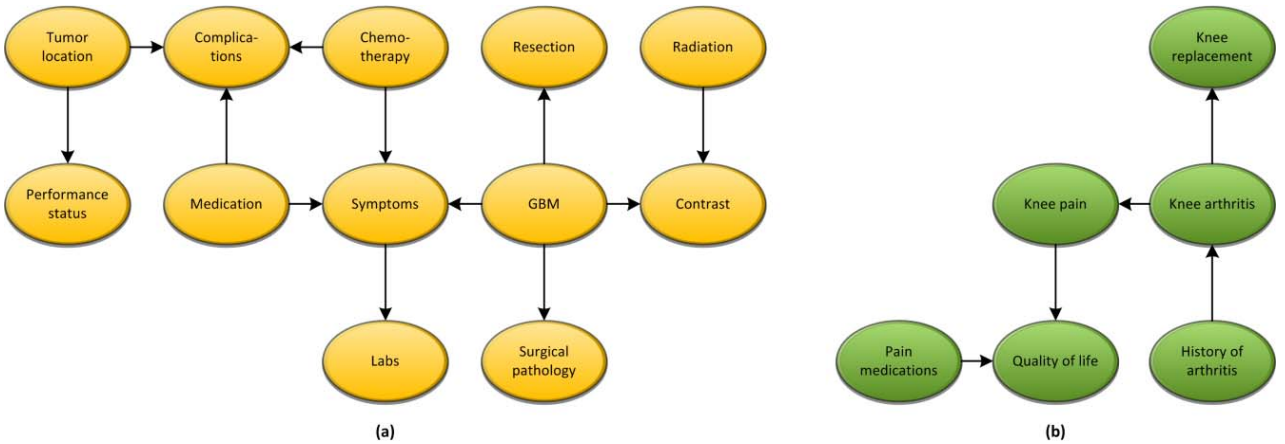

Figure 1. Single time slices from each of the models generated for the clinical datasets. Each model is replicated several times to represent different time points in the course of the patient's diagnosis and treatment. (a) The dynamic Bayesian network (DBN) for glioblastoma multiforme. (b) The DBN for osteoarthritis.

MB-modified versions of only the GBM model were built to determine whether the high-dimensionality of the network (i.e., number of states per variable) affected predictive power. A biostatistician initially provided the domain expert knowledge to perform a non-parametric analysis with classification and regression trees (CART) on the full GBM network. The preliminary analysis showed that increasing the number of regression trees was necessary to improve classifier performance. Consequently, the input space was reduced by pruning the GBM model, ultimately to obtain more accurate predictive results for our evaluation. Our expert selected a subset of variables with value of importance of over 60% per the results of the CART analysis, pruning the network to a total of 18 nodes. This new subset formed the MB-modified model (contrast, symptoms, chemotherapy, medication, tumor location, performance status). Further receiver operator characteristic (ROC) analysis on the MB model clearly supported the need for imputation to handle the missing data in the GBM dataset: the area under the curve (AUC) for ROC analysis improved from 70-77% with the original 36-variable network and unimputed data to 80-95% with the pruned MB model and imputed data. The binary nature of the OA model's variables and overall fewer input variables were not amenable to the same MB pruning and evaluation.

With the network topology and datasets, both the GBM and OAI DBNs were processed using BayesiaLab software [22]. Variables, associations, and variable states were manually formulated using BayesiaLab tools to specify the network. BayesiaLab was responsible for automatically learning the associated CPTs from the processed data used as training data.

IV. Results

Evaluation metrics for both models were based on a general measure of predictive power, or classification accuracy. Precision, or positive predictive value (PPV), was also included to measure the proportion of each model's correctly identified true positive cases. Both metrics were assessed through a stratified tenfold cross-validation.

A. DBN Evaluation

1) UCLA GBM model

Two rounds of evaluation were conducted for the GBM model. Evaluation was performed on an arbitrarily selected set of target variables in the model, predicting the likelihood of a given state. Prior evidence used to predict the target variable state included observation of all variables in the network per time slice, as shown in the associated tables.

The first evaluation involved using priors for the complete 36-variable network. The performance varied from 60-77% in predicting the presentation of "Contrast (enhancement)," "Symptoms," "Chemotherapy," and "Medication" in the second (T = 2; initial progression) (Table 3) and third (T = 3; change in health status, improving or worsening) (Table 4) time slices. In general, the DBN demonstrated improvement in prediction at T = 3 (given observance of variables in prior time slices as priors), a potential reflection of the data gradually stabilizing over time: the progression of GBM is more predictable in later stages of the disease. The results of the MB DBN evaluation are comparable for T = 2, and slightly better for T = 3 (with the exception of predicting medication, where accuracy dropped marginally from 77% to 76%). Overall, the MB demonstrated only a slight improvement in prediction for the DBN, which could indicate that the MB failed to model necessary relationships contained in the full network model needed for predictive accuracy. Precision ranged from about 60-80% across the GBM DBM.

2) OAI model

For the OAI dataset, one evaluation was performed on the full DBN and the remaining evaluations used selected previous months' observations to predict the target variables at the 36-month follow-up stage (Table 5). The binary nature of the variables and lower dimensionality of the state-space improved prediction rates overall, relative to the GBM model. Furthermore, in some situations, the results of the OAI evaluation suggest that including additional prior information may weaken overall DBN prediction rates. For instance, when only using the 30-month observations to predict the state of the patient at 36 months, the prediction rate was

higher or comparable in all categories than when using additional past information (e.g., using both 24- and 30-month observations; or the full network). This implies that while more prior evidence may be helpful to learn each case, the added evidence may not be needed when the model is used, or more data is needed to improve BBN training. Precision results for the OAI DBN are shown in Table 6.

B. LBR Evaluation

1) UCLA GBM model

Next, a series of LBR evaluations were conducted using *Weka* [23]. The precision for target variables were determined at T = 2 and T = 3 using combinations of information to learn the associated probabilities: from the MB of the prior time slice; from all nodes from the immediately prior step; and from all nodes from the current and immediately prior step, minus the targeted predicted value. The results (Tables 3, 4) show that with each added time step, the prediction rate improves. The results also demonstrate improvements across all variables in prediction with the LBR algorithm relative to the DBN. For example, for predicting the state of chemotherapy for T = 2, the LBR-based approach ranged from 72-85% given different levels of evidence, whereas the DBN approach achieved only 59-60%. The comparatively lower predictive power of the DBN suggests that expert knowledge captured in the population-wide model is inadequate compared to a more specific, patient-centered model to describe the specific presentation and treatment or diagnostic paths for GBM. The LBR evaluation demonstrates clear advantages in using the patient-specific model selection for fitting the GBM data. Furthermore, in evaluating the model at T = 3, evidence earlier than the time step immediately prior (i.e., T = 1) has relatively little predictive power compared to more recent information (i.e., T = 2). The evaluation using evidence at T = 2, 3 further supports this finding: evidence at T = 1, 2 is outperformed by using only the evidence in T = 2: the added evidence in T = 1 does not help improve predictive accuracy. This may imply that the immediately previous state is the only prior evidence needed to predict future states. The MB-version of the baseline model for evaluating T = 2 demonstrates only comparable performance to the full baseline topology; the MB-version of the T = 2 model for evaluating T = 3 out performs the full baseline network, which suggests the volatility of baseline or earlier data. Precision per variable in the LBR evaluation ranged from 62-70% at T = 2, and 61-76% at T = 3 with the MB-modified model for T = 1.

2) OAI model

The results of the LBR evaluation on the OAI data are shown in Table 5. Again, evaluation was conducted on the complete network to predict the 36-month target variables; as well as using data from prior observation time points to predict the 36-month presentation. There is relatively clear consistency in the predictive accuracy of each variable, and results are improved over the DBN. Overall, the structured characteristic of this dataset may be more robust to the effects of overtraining and data sparseness. However, the high degree of accuracy is tempered by the fact that most of the variables are binary, and have fewer bins to complicate the prediction task. Still, based on these evaluations, LBR exhibits better predictive rates over the DBN's global learning model.

V. Discussion

The results of this study demonstrate the possible advantages of the Lazy Bayesian Rule algorithm as a technique for selecting an instance-specific predictive model for a given set of data. Patient-specific models that take into account a particular configuration of variable values for every patient case can more effectively model the unique specificities of disease. Indeed, a model must either implicitly or explicitly perform variable selection, choosing the most appropriate subset of a domain's variables for use. For every variable that is chosen, one must also choose the variable's representation within the model (e.g., categorical, continuous, nominal, number/type of attributes) and its relationship to other chosen model variables. A population-wide model tends to include only the predictors that on average provide the best predictive performance. An instance-specific model, on the other hand, may include variable values that are highly predictive for the case at hand but are not applicable to the general population-wide model. Thus, for rare cases the typical population-wide model may predict poorly, whereas an instance-specific model can do well.

The LBR approach, which incorporates Bayesian rule induction, can be readily applied to different domains, as shown in this work with both GBM and knee OA. Moreover, the approaches demonstrated in this study are unique in their application to multivariate clinical variables, particularly in modeling GBM. Most network models in the clinical domain are limited to a priori knowledge and are applied to outcomes or diagnostic models, rather than state models. While predicting outcomes holds significant clinical value, state models provide a full context to understanding the patient's changing health over time, in relation to specific treatments, interventions, other variables, etc.

Two results of this study are of potential significance. First, the instance-specific LBR algorithm outperforms a traditional DBN in terms of predictive accuracy. The patient-specific technique that LBR employs includes information from unique training cases that have varying variable-state configurations, resulting in the construction of a more robust state-space. Thus, the LBR model construction may be a more ideal approach when handling clinical data. Second, the results of evaluating the LBR suggest that the full patient history and set of associated data may not be necessary for accurate prediction (at least, in some scenarios). Using state values from the time step or stage immediately prior to the current state demonstrates improved accuracy over added information from multiple time steps. This contradicts customary assumptions that the full context of a patient record (i.e., history) will always improve predictive accuracy.

The completeness of the data may have also impacted the predictive accuracy at earlier stages of GBM; missing data may have a more significant role in poorer predictive accura-

TABLE III. PREDICTIVE POWER RESULTS FOR GBM MODELS (T =2)

Feature	**Regular *DBN (predicting T = 2)***		***LBR*-based *DBN (predicting T = 2)***		
	Full network	*MB with imputed data*	*MB only, from T = 1*	*All nodes, from T = 1*	*All nodes, from T = 1,2*
Contrast	62%	61%	66%	66%	77%
Symptoms	73%	72%	81%	77%	89%
Chemotherapy	60%	59%	72%	80%	85%
Medication	67%	67%	72%	74%	87%

TABLE IV. PREDICTIVE POWER RESULTS FOR GBM MODELS (T=3)

Feature	**Regular *DBN (predicting T = 3)***		***LBR*-based *DBN (predicting T = 3)***			
	Full network	*MB with imputed data*	*All nodes, from T = 1*	*MB only, from T = 2*	*All nodes, from T = 1, 2*	*All nodes, from T = 2, 3*
Contrast	68%	73%	66%	80%	79%	79%
Symptoms	79%	82%	77%	89%	89%	91%
Chemotherapy	73%	74%	75%	86%	86%	88%
Medication	77%	76%	74%	86%	85%	89%

TABLE V. PREDICTION RESULTS FOR OAI MODELS

Feature	**Regular *DBN***				***LBR*-based *DBN***			
	Full network	*36|30 months*	*36|30|24 months*	*36|30|24|18 months*	*Full network*	*36|30 months*	*36|30|24 months*	*36|30|24|18 months*
Quality of life (QoL)	71%	71%	72%	72%	74%	74%	75%	76%
Knee pain	85%	98%	87%	86%	99%	100%	99%	100%
Right knee symptom	72%	79%	74%	73%	79%	81%	80%	80%
Left knee symptoms	70%	77%	71%	71%	78%	79%	78%	79%
Knee replacement	88%	98%	94%	94%	98%	99%	98%	98%

cy at T = 2 versus T = 3. For example, data was often missing earlier in the patient's history due to the fact that the patient was transferred from another hospital (and thus, the history was not well-documented locally); or the patient's tumor had already progressed significantly. Also, our evaluation did not normalize the patients by a time scale defining clear boundaries between the three stages defined in this work; the number of data points per stage was dependent on the length of each stage, and was unique to each patient. Imposing a specific temporal framework on the cases (i.e., three time slices) in the GBM model ignored potential variation and other useful trend information inherent to given individuals. Moreover, in models involving a large number of variables, several values are likely to be missing, so a potentially large amount of otherwise informative data is discarded. To handle missing data we could use complete case analysis and other forms of imputation. Imputation-based methods other than LOCF attempt to fill in missing values using the observed mean value of the variable, or less naïve strategies based on a predicted value from regression analysis on known variables. Multiple imputation methods (i.e., filling with more than one value) have also been developed [24] to avoid biasing the variances of imputed variables. Still, imputation schemes are far from perfect: filling in values with averages per variable preserves the sample means, but distorts the covariance structure, shifting estimated variances and co-variances to zero. Imputing predicted values from regression analysis also tends to favor observed correlations, biasing them away from zero. Likelihood-based models are another method that could be used to handle missing data [25], [26], by inferring a model's parameters from the existing data that most likely explain the observed data. These models facilitate longitudinal analyses, and are more robust against potential bias from missing data compared to the LOCF approach.

To further validate the accuracy of our results, additional evaluation methods should be investigated to assess both DBN and patient-specific models. Namely, collecting outcomes-related variables to evaluate the models would be the ultimate measure of utility for these models. In addition, incorporating outcomes information may allow us to refine our models for improved prognostic accuracy. For example, partial predictive scoring (PPS) was introduced [27] as an empirical measure to evaluate model performance for BMA. A higher PPS score indicates that a set of events would occur with higher probability in that set. PPS-based metrics may be adapted here to provide further insight into the models' capacity to correctly infer values over time.

The use of domain-expert knowledge for the feature selection process is acknowledged as being not only time intensive but irreproducible and inapplicable to datasets in other domains. The dependence on expert-driven approaches to feature selection may also not necessarily help exhibit the efficacy of the instance-specific model for prediction, which is the focus of this work. In short, the performance of the models may be only the result of careful data specification and feature selection, rather than the characteristics of the instance-specific models themselves. While the supervised approach of feature extraction ensured a degree of consistency and completeness across extracted concepts/features,

evaluation of other instance-specific models and population-wide models may lessen the effects of feature selection bias.

VI. Conclusions

This paper presents a comparison of a DBN-based approach to modeling clinical data over time, against an LBR-based methodology. The former can be seen as a population-wide model, which creates a globally-optimized set of probabilities for a given set of training data; while the latter uses an instance-based framework that averages locally optimized models given specific examples of data. The two approaches are compared given two disparate datasets: a clinically-derived neuro-oncology dataset; and an observational study involving knee osteoarthritis. Using a series of comparative tenfold cross-validation studies, the LBR-based approach outperformed the DBN in correctly predicting future states of high-level features. Moreover, this investigation found that in some cases, using the full patient history as prior evidence may not be useful in predicting future states – in fact, having only the immediately prior stage of data may be sufficient. The measured PPVs for each model, demonstrate the models' strong correlation with which it can approximate to the population, while utilizing case-specific information.

Future work will concentrate on optimizing instance-specific model performance. Specifically, structuring the data in a temporal framework that respects the clinical evolution of the patient and disease, and extracting features that best represent the population are pivotal for optimal prediction. The data imputation (i.e., LOCF) approach used to create a dataset for training the model may not hold in the clinical setting [28], as the patient state may have changed before the observation is carried forward. Therefore, LOCF may bias estimates of treatment effects and the associated standard errors, based on the amount of missing data per feature. A few research studies [29], [30] have evaluated how direct likelihood and multiple imputation schemes and LOCF may impact final results, but other resampling techniques have yet to be completely explored for nominal variables (i.e., clinical text data).

Further evaluations could be conducted on varying configurations of prior evidence to determine their effect on model performance. The window of time that separates the prior evidence and prospective event or stage must be considered; namely, the more time that has elapsed between events with existing data and any future time point will be more difficult to predict. A more thorough evaluation would omit more time points that separate prior and a future state to be predicted. To assist in this assessment we will include added statistical metric mores to fully describe the performance of each model; negative predictive value and ROC analysis would first be considered to determine whether the models can correctly distinguish cases that do not belong in a particular class. Also, as researchers have found [26], LBR is computationally inefficient if large numbers of objects are to be classified from a single training set. As our evaluation was based on classifying one feature at a time, real-world prediction that would be worth replicating would involve classifying a feature, given a more specific configuration of variables. Clinically, this predictive accuracy would hold more significance and practicality if one were to consider a treatment decision (e.g., assessing whether a specific treatment would be more effective in optimizing a given variable's state, as per parametric sensitivity analysis). While the preliminary results from evaluating recall and precision of instance-specific models are promising, the strength of the study could improve with statistically significant results based on evaluating clinical outcomes prediction with instance-specific models.

We may also combine an unsupervised feature extraction technique, based on statistical likelihoods (e.g., topic models), with an application of a standardized vocabulary to construct a more robust set of features for modeling. For example, semi-supervised learning methods can be used: the approach would utilize pre-labeled concepts from a similar classification of documents to the extraction concepts for our corpus of clinical documents [31]. Applying a standardized biomedical concept mapping program to extract concepts such as Metamap could help obtain a more standardized set of concepts between supervised and unsupervised extracted data. Because this work relies on one domain expert the approach is acknowledged for high potential in biased feature extraction .

Acknowledgments

This research was supported in part by the NLM Training Grant LM007356 and NIH R01 EB000362. We would also like to thank Juan Eugenio Iglesias and Dr. Corey Arnold for their input on the intellectual direction of the work.

The OAI is a public-private partnership comprised of five contracts (N01-AR-2-2258; N01-AR-2-2259; N01-AR-2-2260; N01-AR-2-2261; N01-AR-2-2262) funded by the National Institutes of Health, a branch of the Department of Health and Human Services, and conducted by the OAI Study Investigators. Private funding partners include Merck Research Laboratories; Novartis Pharmaceutical Corporation, GlaxoSmithKline; and Pfizer, Inc. Private sector funding for the OAI is managed by the Foundation for the National Institutes of Health. This manuscript was prepared using an OAI public use data set and does not necessarily reflect the opinions or view of the OAI investigators, the NIH, or the private funding partners.

References

[1] G.A. Miller, E. Galanter and K.H. Pribram, Plans and the structure of behavior. New York: Holt, 1960.

[2] A. Newell and H.A. Simon, Human problem solving. Englewood Cliffs, NJ: Prentice-Hall, 1972.

[3] D.I. Coggon and C.N. Martyn, "Time and chance: The stochastic nature of disease causation," Lancet, vol. 365, 2005, pp. 1434-1437.

[4] S. Visweswaran and G.F. Cooper, "Learning instance-specific predictive models," Journal of Machine Learning Research, vol. 9999, 2010, pp. 3333-3369.

[5] M.S. Brown, M.F. Gray, J.G. Goldin, R.D. Suh, J.W. Sayre and D.R. Aberle, "Patient-specific models for lung nodule detection and surveillance in CT images", IEEE Transactions on Medical Imaging, vol. 20, 2001, pp. 1242-1250.

[6] F. Verdenius and J. Broeze, "Generalised and instance-specific modelling for biological systems", Environmental Modelling and Software, vol. 14, 1999, pp. 339-348.

[7] Z. Ghahramani, “Learning Bayesian networks”, in Adaptive Processing of Sequences and Data Structures, vol. 1387, 1998, pp. 168-197.

[8] K.P. Murphy, Dynamic bayesian networks: representation, inference, and learning. Doctoral dissertation, 2002.

[9] R. Dahlhaus and M. Eichler, “Causality and graphical models for time series”, in Highly structured stochastic systems, P. Green, N. Hjort and S. Richardson, eds. Oxford University Press, 2000.

[10] N. Friedman, M. Linial I. Nachman and D. Pe’er, “Using Bayesian networks to analyze expression data”, Journal of Computational Biology, vol. 7, 2000, pp. 601-620.

[11] D.W. Aha, D. Kibler and M.K. Albert, Instance-based learning algorithms, Machine Learning, vol. 6, 1991, pp. 37-66.

[12] C. Boutilier, N. Friedman, M. Goldszmidt and D. Koller, “Context-specific independence in Bayesian networks”, Proc. Annual Conference on Uncertainty in Artificial Intelligence, Reed College/Morgan Kaufmann, 1996, pp. 1-4.

[13] J.A. Hoeting, D. Madigan, A.E. Raftery and C.T. Volinsky, “Bayesian model averaging: a tutorial,” Statistical Science, vol. 14, 1999, pp. 384-417.

[14] C.T. Volinsky, D. Madigan, A.E. Raftery and R.A. Kronmal, “Bayesian model averaging in proportional hazard models: assessing the risk of a stroke,” Journal of the Royal Statistical Society, vol. 46, 1997, pp. 433-448.

[15] S. Visweswaran and G.F. Cooper, “Patient-specific models for predicting the outcomes of patients with community aquired pneumonia”, Proc. Annual American Medical Informatics Symposium (AMIA ’05), 2005, pp. 759-763.

[16] R. Kohavi, "Scaling up the accuracy of Naive-Bayes classifiers: a decision-tree hybrid", in Proceedings of the Second International Confrence on Knowledge Discovery and Data Mining, 1996.

[17] R.C. Holte, L.E. Acker and B.W. Porter, "Concept learning and the problem of small disjuncts," In Proc. of the 11th Intl Joint Conferrence on Artificial Intelligence,1989, pp. 8183-818.

[18] Z. Zheng and G.I. Webb, “Lazy learning of Bayesian rules”, Machine Learning, vol. 41, 2000, pp. 53-84.

[19] D.W. Aha, Lazy learning. Dordrecth: Kluwer Academic, 1997.

[20] R. Little and D. Rubin, Statistical Analysis with Missing Data. New York: Wiley & Sons, 1987.

[21] J. Carpenter and J. Bithell, "Bootstrap confidence intervals: when, which, what? a practical guide for medical statisticians," Statistics in Medicine, vol. 19, 2000, pp. 1141-1164.

[22] P. Munteanu, L. Jouffe and P. Wuillemin, Bayesia Lab, 2001.

[23] I.H. Witten. Weka: practical machine learning tools and techniques with Java implementations, 1999.

[24] P.W. Lavori, R. Dawson and D. Shera, “A multiple imputation strategy for clinical trials with truncation of patient data”, Statistics in Medicine, vol. 14, 1995, pp. 1913-1925.

[25] A.P. Dempster, N.M. Laird and D.B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm”, Journal of Research Statistical Society, vol.39, 1977, pp. 1-38.

[26] J.L. Schafer, Analysis of incomplete missing data. London: Chapman & Hall, 1997.

[27] C.H. Mallinckrodt, W.S. Clairk, R.J. Carroll and G. Molenberghs. “Assessing response profiles from incomplete longitudinal clinical trial data under regulatory conditions”, Journal of Biopharmaceutical Statistics, vol. 13, 2003, pp. 179-190.

[28] O. Siddiqui and M.W. Ali, “A comparison of the random effects pattern mixture model with last observation carried forward (LOCF) analysis in longitudinal clinical trials with dropouts,” Journal of Biopharmaceutical Statistics, vol. 8, 1998 pp. 545-563.

[29] C. Beunckens, G. Molenberghs and M.G. Kenward, "Direct likelihood analysis versus simple forms of imputation for missing data in randomized clinical trials," Clinical Trials, vol. 2, 2005, pp. 379-386.

[30] G. Verbeke and G. Molenberghs, Linear Mixed Models for Longitudinal Data. New York: Springer, 2000.

[31] R. Raina, A. Battle, H. Lee, B. Packer and A.Y. Ng, "Self-taught learning: transfer learning from unlabeled data," Proceedings of the 24th International Conference on Machine Laerning, 2007, pp. 759-766.