Modeling Heart Procedures from EHRs: An Application of Exponential Families

Shuo Yang*, Fabian Hadiji†, Kristian Kersting‡, Shaun Grannis§ and Sriraam Natarajan¶
*School of Informatics, Computing and Engineering
Indiana University, Bloomington, USA
Email: shuoyang@indiana.edu
†TU Dortmund University, Dortmund, Germany
Email: fabian@hadiji.com
‡TU Darmstadt University, Darmstadt, Germany
Email: kersting@cs.tu-darmstadt.de
§Regenstrief Institute, Bloomington, USA
Email: sgrannis@regenstrief.org
¶The University of Texas at Dallas, Dallas, USA
Email: Sriraam.Natarajan@utdallas.edu

***Abstract*—In order to facilitate better estimations on coronary artery disease conditions of a patient, we aim to predict the number of Angioplasty (a coronary artery procedure) by taking into account all the information from his/her Electronic Health Record (EHR) data. For this purpose, two exponential family members—multinomial distribution and Poisson distribution models—are considered, which treat the target variable as categorical-valued and count-valued respectively. From the perspective of exponential family, we derive the functional gradient boosting approach for these two distributions and analyze their assumptions with real EHR data. Our empirical results show that Poisson models appear to be more faithful for modeling the number of this procedure.**

***Keywords*-clinical events prediction; exponential family; probabilistic models; EHR;**

I. Introduction

Coronary Artery Disease (CAD) is the leading cause of death for people in the United States and killing more than 370,000 people annually [1]. A serious condition a lot of CAD patients suffer from is the narrowed or blocked artery by plaque. One of the common procedures to restore blood flow through narrowed arteries is *Angioplasty*. However, a considerable amount of patients treated with angioplasty often need repeat revascularization procedures [2]. While there is research investigating the predictive factors of restenosis [3]–[6], there is no significant research that focuses on the factors over a substantial period of time to discover their effect on the number of procedures. We do not only aim to investigate the association between the restenosis and predictive factors, but also to predict the count of restenosis. This allows us to better estimate a new patient's condition by analyzing his/her historical health record data. Since the number of angioplasties a patient needs usually associates with the severity and complexity of his/her CAD condition, modeling and predicting the count of angioplasties can provide better understanding on the future CAD conditions and hence facilitate better treatment plans in advance.

Faithfully modeling such count data could potentially provide a better way to exploit the large-scale health records data in order to reveal the development patterns of certain diseases or facilitate the discoveries of the significant influential factors on a specific medical condition. An important property of such count data is that these variables typically only take the non-negative integer values. Most of the previous work that employed machine learning algorithms modeled such count variables as multinomial distributed [7]. The key issue with this assumption is that it places an upper bound on the counts (number of readmissions or procedures for instance) of the target. While in many domains, it is possible to obtain a reasonable upper bound, in other domains, this requires guessing a good one.

Another common assumption that most prior work assumed is that of the Gaussian assumption over these count variables. However, low counts can lead to the left tail of the Gaussian distribution predicting negative values for these counts. Subsequently, there are research directions [8]–[10] that model such data from a different perspective by assuming that these counts are distributed according to a Poisson. Such models have been extended to learning dependencies among multiple Poisson variables as Poisson graphical models. Poisson models have been increasingly employed in the context of the count data [11]–[13]. However, little research has been performed on comparing these two different probability distribution assumptions in real-world medical domains.

In this paper, we target the number of *Angioplasty* as it is a significant indicator for the severity of patients' coronary disease conditions and aim to examine various probabilistic graphical models on predicting the count of *Angioplasty* by learning from the historical medical conditions of the CAD patients. To model such count variables, we employed probabilistic graphical models with two different assumptions on the conditional probability distributions of the target variable. One is multinomial distribution assumption which assumes

that each sample is independently extracted from an identical categorical distribution, and the numbers of samples falling into each category follow a *multinomial distribution*. Another assumption is the *Poisson distribution* which states that each sample is a instance of a count variable following Poisson distribution with one parameter λ.

It must be mentioned that such assumptions on the data probability distributions cannot be examined by simply employing certain statistical analysis methods, such as data transformations, on the interested variable alone. This is because most of the time we also aim to know how other variables in the domain influence the target variable, and which assumption more faithfully fits the conditional probability distributions of the target variable given the variables on which it depends.

Consequently, we propose a set of experiments to model the dependencies among the variables and at the same time examine these probability distribution assumptions by employing probabilistic graphical models with the two different probability distributions from exponential family. Our key aim in this work is to investigate the impact of different assumptions in modeling this challenging task. As a general form of most probabilistic graphical models, exponential family [14] provides a better perspective from which the fundamental connection between their theories as well as learning algorithms can be revealed. We will take the perspective of exponential families and illuminate the theories and boosting learning approaches for these two probability distribution assumptions.

We make the following key contributions: first, we consider multiple base learners in the context of gradient boosted multinomial and Poisson models for the task of predicting the number of Angioplasty procedures; second, we consider different types of gradient updates for learning Poisson models; third, we consider different types of modeling for the predictive features when learning these models. Finally, we perform comprehensive analyses on the real EHR data and demonstrate the usefulness of exponential family distributions for this interesting task.

The rest of the paper is organized as follows: we first present the background on gradient-boosting approaches and exponential family concepts. Next, from the perspective of exponential families, we derive the gradient updates for both multinomial and Poisson distributions based on additive updates. We then explain the EHR data that is employed for the task. Finally, we present our extensive real-world empirical evaluation before concluding the paper.

II. Background

A. Functional Gradient Boosting

The standard gradient ascent approach learns graphical models by optimizing a loss function w.r.t. the natural parameters in the probability distribution assumptions on the target variable, e.g. the parameters $\{p_1, p_2, ..., p_k\}$ in the multinomial distribution or the parameter λ in the Poisson distribution. Typically, it starts with initial parameters and iteratively adds the gradient of an objective function w.r.t. the parameters, till it finds the optimal parameters that most faithfully fit the data.

Friedman proposed the functional gradient boosting [15] which defines a regression function ψ w.r.t. each example x_i. The gradients of the objective function are derived with respect to the function parameters $\psi(x_i)$. The key idea behind this approach is that, instead of computing the overall gradient, the gradients are approximated by computing them for each example, i.e.

$$\Delta(x_i) = \frac{\partial LL(\mathbf{x})}{\partial \psi(x_i)}. \tag{1}$$

These gradients can be easily understood as the difference between the true label and the current predicted probability of an example. These gradients are computed for each example and form the regression dataset. To generalize from these regression examples, in each iteration t, a regression function $\hat{\psi}_t$ (generally regression tree) is learned to fit to the gradients (which effectively become the weight of each example). The final model,

$$\psi_t = \psi_0 + \hat{\psi}_1 + \cdots + \hat{\psi}_t \tag{2}$$

is a sum over these regression trees. Unlike AdaBoost, this FGB approach learns a *probabilistic* classifier. Also, most off-the-shelf regression learners can be used for fitting the examples in each gradient step.

B. Graphical Models as Exponential Family

Exponential Family [14] provides a general way to represent the probability distribution as a density p continuous w.r.t some base measure $h(x)$ [16]. This base measure $h(x)$ could be the counting measure or the ordinary Lebesgue measure on $\mathbb{R}$. Many common distributions belong to the exponential family, such as the normal, exponential, gamma, Dirichlet, Bernoulli, Poisson, multinomial (with fixed number of trials) and etc. Generically, there could be numerous different distributions that are consistent with the observed data, so this problem can be converted to find the distribution that faithfully fits the data while has the maximal Shannon entropy [16]. Hence, the exponential family is defined as a family of probability distributions whose probability mass function (only consider the discrete variables in this paper) can be represented with a parameterized collection of density functions:

$$p_X(x|\theta) = h(x)\exp\{\langle\theta, \phi(x)\rangle - A(\theta)\},$$

where $\phi(x)$ is a collection of *sufficient statistics*, θ is an associated vector of *exponential parameters* and $A(\theta)$ is the *log partition function* which ensures $p_X(x|\theta)$ being properly normalized.

III. Learning Exponential Family Models with Functional Gradient Boosting

In this paper, we aim to answer the following question: which assumption of a probability distribution more faithfully fits the number of a patient's cumulative $Angioplasty$ surgeries given his/her historical medical conditions? To this effect, we employ graphical models of exponential family with different assumptions on the conditional probability distribution of the target variable. We denote the target discrete variable as y

(i.e. number of $Angioplasty$), the other attributes related to the patient's medical conditions as variable set X. As in graphical models, each node corresponds to one random variable and we will use nodes and variables interchangeably. We refer to those variables which have dependence correlations with the target variable as parent nodes if they are captured by the graphical models. The subscript i indexes the samples indicating the i^{th} patient. The subscript $_{y;X}$ indicates the corresponding sufficient statistics (i.e. $\phi(y;X)$) or exponential parameters (i.e. $\theta_{y;X}$) in a certain graphical model which could be a linear function such as in Logistic Regression and Naive Bayes or a non-linear models such as Decision Trees, Neural Networks and Poisson Dependency Networks. So, we can re-write the exponential family in terms of one example as:

$$p_{y_i;X_i}(y_i;X_i|\theta) = \\ h(y_i;X_i)\exp\{\langle\theta_{y_i;X_i},\phi(y_i;X_i)\rangle - A(\theta_{y_i;X_i})\}. \tag{3}$$

Since the target variable y has discrete values, we choose to evaluate two exponential family members– multinomial distribution and Poisson distribution. We present the gradient updates for the two distributions in the context of our learning algorithm - gradient boosting in the following subsections.

A. Functional Gradient Boosting for Graphical Models with Multinomial Distribution Assumption

First, we will present the objective function for optimizing the log-likelihood under the multinomial distribution assumption and the corresponding gradient. If we assume that given the observations in the learned discriminative model, the $Angioplasty$ count follows a categorical distribution with parameters $\{p_{y^1},...,p_{y^r}\}$ where $\sum_k p_{y^k}=1$ and $\{y^1,...,y^r\}$ indicate possible values y can take, the probability distribution over the number of all possible values is a multinomial distribution:

$$p(N_{y^1},...,N_{y^r};p_{y^1},...,p_{y^r}) = \frac{\Gamma(\sum_k N_{y^k}+1)}{\prod_k \Gamma(N_{y^k}+1)}\prod_{k=1}^{r} p_{y^k}^{N_{y^k}},$$

where $y \in \{0, \quad 1, \quad ..., \quad r-1\}$ and $r-1$ is the maximum number of $Angioplasty$ a patient can have in the EHR. According to the definition in Equation (3), the parameters in the general form of exponential family can be calculated through the natural parameters in this multinomial distribution by:

$$\begin{aligned}
&\phi(y;X) = [N_{y^1}, \quad ..., \quad N_{y^r}]^T \\
&\theta_{y;X} = [\ln p_{y^1}, \quad ..., \quad \ln p_{y^r}]^T \\
&h(y;X) = \frac{(\sum_{k=1}^r N_{y^k})!}{\prod_{k=1}^r N_{y^k}!} \\
&A(\theta_{y;X}) = 0
\end{aligned}$$

The gradients of the logarithm of the exponential family probability function w.r.t. the natural parameter p_{y^k} equals to:

$$\begin{aligned}
&\frac{\partial LL}{\partial p_{y^k}} \\
&= \frac{\partial}{\partial p_{y^k}}\ln h(y;X) + \langle\theta_{y;X},\phi(y;X)\rangle - A(\theta_{y;X}) \\
&= \langle\frac{\partial\theta_{y;X}}{\partial p_{y^k}},\phi(y;X)\rangle - \frac{\partial A(\theta_{y;X})}{\partial p_{y^k}} \\
&= \frac{1}{p_{y^k}}N_{y^k}.
\end{aligned} \tag{4}$$

Because of the simplex constraint $\sum_k p_{y^k}=1$ and the value range of a probability $0 \le p_{y^k} \le 1$, in order to optimize the parameters separately while keeping these constraints satisfied, the multinomial regression models are usually converted into the form of a log-linear model with a new set of parameters $\{\beta_{y^1}, \quad ..., \quad \beta_{y^r}\}$ which satisfies $p_{y^k} = \frac{e^{\beta_{y^k}}}{\sum_{k'} e^{\beta_{y^{k'}}}}$.

To learn such log-linear models with multinomial variables with functional gradient boosting, we need to view this probability distribution assumption from the perspective of a single patient $(y_i;X_i)$ each of which attached with a function parameter $\psi_{y_i;X_i}$. Instead of optimizing the log-likelihood function w.r.t. the natural parameters $\{\beta_{y^1}, \quad ..., \quad \beta_{y^r}\}$, functional gradient boosting optimizes the objective function w.r.t. the function parameter $\psi_{y_i;X_i}$ for every example. Under the multinomial distribution assumption, the probability of a single patient is defined as a function of parameter $\psi_{y_i;X_i}$ as: $p_{y_i;X_i} = \frac{e^{\psi_{y_i;X_i}}}{\sum_k e^{\psi_{y_i=y^k;X_i}}}$. The probability for this patient to have k times of $Angioplasty$ surgery can be treated as a categorical distribution, in which case the sufficient statistics in the general exponential family form (3) can be derived as $\phi(y_i;X_i) = [I_{y_i=y^1}, \quad ..., \quad I_{y_i=y^r}]^T$ where I is the indicator function. So the optimization problem w.r.t. p_{y^k} can be converted to an optimization problem w.r.t. ψ_{y^k}. The point-wise gradients of the logarithm of function (3) w.r.t. $\psi_{y_i;X_i}$ can be derived by substituting Equation (4) as:

$$\begin{aligned}
\frac{\partial LL_i}{\partial\psi_{y_i=y^k;X_i}} &= \frac{1}{p_{y_i;X_i}}\cdot\frac{\partial p_{y_i;X_i}}{\partial\psi_{y_i=y^k;X_i}} \\
&= I(y_i=y^k;X_i) - p(y_i=y^k;X_i).
\end{aligned} \tag{5}$$

This is the gradient function employed by regression models with multinomial distribution assumption when learning with functional gradient boosting. At each step, a regression function is fit to these gradients. We consider two types of regression functions - regression trees and neural networks.

B. Functional Gradient Boosting for Graphical Models with Poisson Distributions

The other case we consider is the Poisson distribution assumption, in which we employ discriminative Poisson model whose structure is similar as a reversed Naive Bayes but the conditional probability distribution of the target variable follows Poisson distribution. In other word, we assumed that given the parent nodes in the learned discriminative Poisson

model, the probability distribution of the $Angioplasty$ count follows Poisson distribution:

$$p(y) = \frac{\lambda^y e^{-\lambda}}{y!},$$

where $y \in \{0, \ 1, \ 2 \ ...\}$. The components in the general exponential family form (Equation (3)) can be converted from the point-wise perspective as :

$$\begin{aligned}
&\phi(y_i; X_i) = y_i \\
&\theta_{y_i;X_i} = \ln \lambda_{y_i;X_i} \\
&h(y_i; X_i) = \frac{1}{y_i!} \\
&A(\theta_{y_i;X_i}) = e^{\theta_{y_i;X_i}} = \lambda_{y_i;X_i}
\end{aligned}$$

Then, the point-wise gradients of the log-likelihood function w.r.t. the natural parameter $\lambda_{y_i;X_i}$ can be derived as:

$$\begin{aligned}
&\frac{\partial LL_i}{\partial \lambda_{y_i;X_i}} \\
&= \frac{\partial}{\partial \lambda_{y_i;X_i}} \ln h(y_i; X_i) + \langle \theta_{y_i;X_i}, \phi(y_i; X_i) \rangle - A(\theta_{y_i;X_i}) \\
&= \langle \frac{\partial \theta_{y_i;X_i}}{\partial \lambda_{y_i;X_i}}, \phi(y_i; X_i) \rangle - \frac{\partial A(\theta_{y_i;X_i})}{\partial \lambda_{y_i;X_i}} \\
&= \frac{1}{\lambda_{y_i;X_i}} y_i - 1.
\end{aligned} \tag{6}$$

Here, $\lambda_{y_i;X_i}$ is the Poisson parameter for the probability distribution of the target variable of the i^{th} patient given the learned discriminative Poisson model and the values of y_i's parent nodes X_i, which are embodied by the subscript $_{y_i;X_i}$.

For functional gradient boosting, we define the function parameters as $\psi_{y_i;X_i}$ so that $\lambda_{y_i;X_i} = e^{\psi_{y_i;X_i}}$, in order to satisfy the constraints on λ which is $\lambda > 0$. Since exponential is a monotonic increasing function, the optimal Poisson parameters λ can be found by boosting the log-likelihood function with respect to the function parameters ψ. Based on Equation (6), the point-wise gradients of the log-likelihood function of (3) w.r.t. $\psi_{y_i;X_i}$ is derived as:

$$\begin{aligned}
\frac{\partial LL_i}{\partial \psi_{y_i;X_i}} &= \frac{\partial LL_i}{\partial \lambda_{y_i;X_i}} \cdot \frac{\partial \lambda_{y_i;X_i}}{\partial \psi_{y_i;X_i}} \\
&= (\frac{1}{\lambda_{y_i;X_i}} y_i - 1) \cdot e^{\psi_{y_i;X_i}} = y_i - \lambda_{y_i;X_i}.
\end{aligned} \tag{7}$$

Equation (7) is the gradient function we used to optimize a discriminative Poisson model with additive functional gradient boosting which we denote as *DPM-AGB*.

Hadiji et al. [10] defined the functional parameter as $\psi_{y_i;X_i} = \lambda_{y_i;X_i}$ and adapted a self-adaptive step size η into the additive functional gradient update function $\psi_t = \psi_{t-1} + \eta \cdot \Delta_t$, where η is equivalent to the functional parameter from previous iteration (i.e. $\eta = \psi_{t-1}$), and further proved that this new update formula is equivalent to a multiplicative update step with formula:

$$\psi^{t+1}(y_i; X_i) = \psi^t(y_i; X_i) \cdot E_{X_i, y_i} \left[\frac{y_i}{\lambda_i(y_i; X_i)} \right]. \tag{8}$$

We also evaluated the performance of the discriminative Poisson model with their multiplicative gradient boosting approach, which is denoted as *DPM-MGB* in the following sections. The algorithm for *DPM-MGB* is as shown in Figure 1 and we omit the algorithm of *DPM-AGB* since it is only different from *DPM-MGB* in the Function **GenDPMEgs** where it generates the examples with Equation (7) instead of Equation (8) and updates the model with additive step $F_m := F_{m-1} + \Delta_m$ in Function **DPM-AGB**.

Fig. 1. Multiplicative Gradient Boosting for Discriminative Poisson Models

function DPM-MGB($Data$)
  for $1 \leq m \leq M$ **do**
    ▷ Iterate through M gradient steps
    $S :=$GenDPMEgs($Data; F_{m-1}$)
    ▷ Generate examples
    $\Delta_m :=$FitRegressTree(S)
    ▷ Regression Tree learner
    $F_m := F_{m-1} \cdot \Delta_m$ ▷ Update model
  end for
end function

function GenDPMEgs($Data, F$)
  $S := \emptyset$
  for $1 \leq i \leq N$ **do** ▷ Iterate over all examples
    $\Delta(y_i; \mathbf{x}_i) := \frac{y_i}{\lambda_i(y_i; X_i)}$
    $S := S \cup [(y_i), \Delta(y_i; ; \mathbf{x}_i))]$
  end for
return S ▷ Return regression examples
end function

function FitRegressionTree(S)
  Tree := createTree($P(X)$)
  Beam := {root(Tree)}
  while $numLeaves(Tree) \leq L$ **do**
    Node := popBack(Beam)
    ▷ Node w/ worst score
    C := createChildren(Node)
    ▷ Create children
    BN := popFront(Sort(C, S))
    ▷ Node w/ best score
    addNode(Tree, Node, BN)
    ▷ Replace Node with BN
    insert(Beam, BN.left, BN.left.score)
    insert(Beam, BN.right, BN.right.score)
  end while
return Tree
end function

IV. Predicting Angioplasty from Real EHR

As mentioned earlier, we evaluated the distribution assumption on a specific task - that of predicting treatments of coronary artery disease. We used the electronic health

record extracted from *Regenstrief* Institute. We extracted data from 5991 patients for a total of 4928799 health records spanning from 1973 to 2015. We defined the target variable as $Angioplasty$ since it is a strong indicator for coronary artery disease and assigned its value with the number of Angioplasty through the trajectory of the patient's medical record.

We considered medical measurements, procedures, medications as well as certain health behaviors before the first Angioplasty as the predicting features (parents) of the target variable *Angioplasty*. Table I shows the features and their value types. As can be seen, there are two types of variables - Boolean and numeric variables. Given that, for a certain medical test, there could be multiple measurements before the first Angioplasty procedure, the obvious question is which measurement should we choose?

TABLE I
DOMAIN DESCRIPTION

Attribute Name	Value Type
Procedures for EKG	Boolean
Beta-Blocker Medications	Boolean
Lipid Lowering Medication	Boolean
Tricyclic Anti-Depressant	Boolean
Calcium Channel Blocker Medication	Boolean
DM Medication	Boolean
HTN Medication	Boolean
Post Acute Coronary Syndrome	Boolean
Presence of Atrial Fib	Boolean
History of Alcoholism	Boolean
Smoke Status	Boolean
Diastolic Blood Pressure	Numeric
Systolic Blood Pressure	Numeric
High-Density Lipoproteins (HDL)	Numeric
Low-Density Lipoprotein (LDL)	Numeric
A1C test	Numeric
Triglyceride	Numeric
Body Mass Index	Numeric

To answer this, consider Figure 2 which shows a sample EHR trajectory of one patient with all records that are related to him/her from the EHR. Each node along the time-line represents a medical event happened at that time point, e.g. having a *Procedures for EKG* at 3/16/1995, taking HTN Medication at 6/19/2001 or being tested for A1C at 10/31/2009. The red nodes indicate the occurrences of $Angioplasty$. We use different colors (orange and purple) to indicate that there are two kinds of value types of the features, boolean and numeric.

As mentioned earlier, the time point when the first $Angioplasty$ procedure was performed is used to determine the end time point of the segment over which the other sequential events are extracted and aggregated. For the boolean valued variables (usually the usage of certain medications or procedures), we employed two different aggregation functions: i). *Indicator* which is 1 if a certain event happens at least once within the segment (i.e. before the first $Angioplasty$) and 0, otherwise; ii). *Count* which is the number of occurrences for that event along the chosen segment of the patient's EHR trajectory.

For the case of numeric variables, we considered three different ways for aggregating the multiple observations before the first Angioplasty - i). *Min* which is the min value of all the recorded values through the segment for a certain medical measurement; ii). *Max* which is the max value of all the measured values within the segment; iii). *Mean* which is the mean value of those values. We also considered the final measurement before the first Angioplasty as the representative measure of the observations till that time point and denoted it as *Latest* in our results. Finally, if a patient never had $Angioplasty$, the target variable has the value of 0 and the values of other variables are calculated by aggregating their corresponding values over the entire trajectory instead of a segment.

As mentioned earlier, we considered both Poisson and Multinomial distributions with gradient boosting as the learning algorithm. Inside the gradient boosting, we considered two types of base learners for Multinomials - Trees and Neural networks [17] denoted as REPTree-B and NN-B respectively in our results. For the Poisson assumption, we considered two different update manners - multiplicative boosting [10] (denoted as DPM-MGB) and additive boosting(denoted as DPM-AGB).

We aim to answer the following questions explicitly:

- Q1: Which of the two exponential family distributions faithfully models the number of Angioplasty procedures?
- Q2: Does the choice of base learner or the update type impact the modeling ability of the learned distributions?
- Q3: Does the choice of aggregation function matter when predicting the number of procedures?

To evaluate these questions, we calculated the Mean Square Error (MSE)

$$MSE = \frac{\sum_{i=1}^{N}[1 - p(y_i = \hat{y}_i; X_i)]^2}{N},$$

where $\hat{y}_i$ is the true value of the target variable of the i^{th} patient and N is the number of patients. We also calculated the mean Log-likelihood (LL) for DPMs, since the other classifiers often predict the probability for $Angioplasty$ being high values as 0, hence the LL score cannot be reported for them. We performed 5-fold cross-validation using the data from 5991 patients and aggregated their results.

The first observation on Table II is that the Poisson model with multiplicative updates tends to have lower MSE than their multinomial counterparts. Its superior performance is statistically significant at the 3% significance level except for the case of Indicator Mean when the *p*-value equals 0.1. DPM-AGB has slightly higher MSE than multinomial boosting with regression trees (RepTree-B). However, the differences between their performance are NOT statistically significant when applying the aggregators: Count Min ($p\text{-}value = 0.588$), Indicator Min ($p\text{-}value = 0.088$) and Latest ($p\text{-}value = 0.077$). We hypothesize that this is due to the sensitiveness of the step-size parameter of the gradient steps in additive Poisson models. The multiplicative models appear more robust to the step size (an observation made by Hadiji et al. [10]). So we answer Q1

Fig. 2. Sample Trajectory of One EHR. The red circles indicate the events of $Angioplasty$; purple circles indicate the events of medical procedures with binary values; orange circles indicate the events of medical measurements with continuous values. Events before the first $Angioplasty$ would be aggregated with certain aggregators according to their value types.

TABLE II
MSE OF PREDICTIONS FROM BOOSTING DIFFERENT CLASSIFIERS

Numeric	Boolean	REPTree-B	NN-B	DPM-AGB	DPM-MGB
Min	Count	0.071	0.705	0.072	**0.066**
	Indicator	0.064	0.819	0.067	**0.058**
Max	Count	0.066	0.614	0.069	**0.063**
	Indicator	0.063	0.994	0.066	**0.059**
Mean	Count	0.064	0.255	0.068	**0.061**
	Indicator	0.063	0.513	0.068	**0.061**
Latest	Latest	0.065	0.438	0.068	**0.061**

cautiously in that Poisson models appear more suited for this task if they are robust.

With respect to the base learners employed, it appears that the neural networks do not perform as well as trees when boosted - possibly due to overfitting. The trees serve as better weak learners for the boosted model. For the gradient updates, the multiplicative models appear to exhibit better performance than additive models which are sensitive to their step-sizes. Thus Q2 can be answered in favor of trees and multiplicative updates.

Finally, for the comparison of different aggregators, we considered the model with the best performance (i.e. DPM-MGB). The differences between the performances of Indicator and Count are NOT statistically significant at 3% significance level, which indicates that the Indicator function is as useful as Count for boolean variables (in that knowing whether an event happened appears to be as informative as the number of times the event happened). Same results are observed with numeric variables as well. Thus answering Q3, there does not seem to be too much difference which is statistically significant in the choice of aggregation functions .

It is intriguing to check if different aggregators can result in vastly different models. To this effect, we analyze the first tree learned by each Poisson model. The first tree is a good indicator of the most important set of features that are employed in the prediction task. As can be seen clearly in Figure 3, all the settings use the same feature in the root, which is to check if the patient had post acute coronary syndrome. Alcohol consumption appears on the left subtree in all the aggregators and the Triglyceride appears on the right. It is also instructive to note that the regression values at the leaves are similar as well. This demonstrates that the aggregators do not necessarily influence the final learned model.

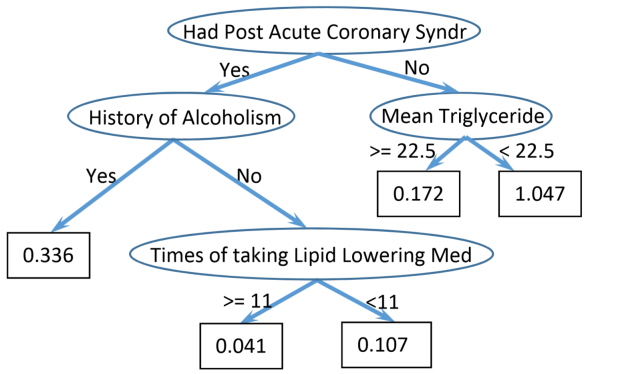
In summary, our results are encouraging in that extensive feature engineering is not necessary when employing a reasonable probability distribution for modeling Angioplasties. Our results show that modeling the count distributions using Poisson regression model allows us to faithfully capture the number of procedures (as the high log-likelihood numbers show in Table III).

TABLE III
LOG-LIKELIHOOD OF RESULTS FROM DPM-MGB

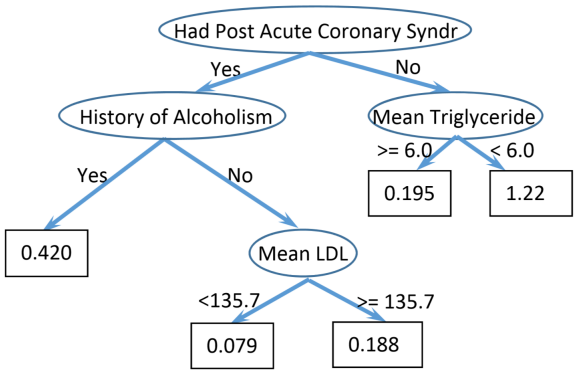
LL	Min	Max	Mean	Latest
Count	-0.486	-0.37	-0.384	\
Indicator	-0.412	-0.399	-0.414	\
Latest	\	\	\	-0.409

V. CONCLUSION

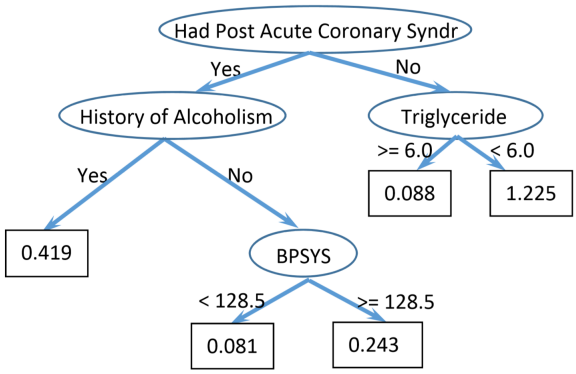
We considered a real application of graphical models - that of predicting the number of Angioplasties on a patient given historical data about the patient. We employed two different types of assumptions—Poisson and multinomials—for modeling this task. It is encouraging that going beyond multinomials and respectively Gaussian (as an approximation) for discrete random variables as observed recently in the literature yields interesting insights to the underlying data. We demonstrated this with a challenging real-world high-impact task. Our results showed that Poisson models are reasonable in modeling count data and appear to be better than multinomials when used with a multiplicative boosting approach. Our results also showed that good modeling assumptions can reduce the need for feature engineering. We intend to evaluate this algorithm on a larger EHR with over 50,000 patients. In addition, we aim to learn a full dependency network that will model several procedures simultaneously (i.e., joint learning). Finally, extending this work to move beyond EHRs and include other data sources such as social network, behavioral data and possibly genetic and imaging data remains interesting directions for future research.



(a)



(b)



(c)

Fig. 3. Sample Learned Poisson Regression Tree for each of the aggregator. (a) is the Count-Mean, (b) is the Indicator-Mean and (c) is the Latest. It is worth noting that in all the three cases, the top levels appear to be very similar demonstrating that the learned mode is indeed robust for all settings.

ACKNOWLEDGMENTS

SY and SN gratefully acknowledge National Science foundation grant no. IIS-1343940. SN also gratefully acknowledges the support of Indiana University, Precision Health Initiative. FH and KK were with the TU Dortmund and supported by the German Science Foundation (DFG) within the Collaborative Research Center SFB 876 "Providing Information by Resource-Constrained Data Analysis", project B4 "Analysis and Communication for dynamic traffic prognosis", when the research was done.

REFERENCES

[1] "Deaths: Final data for 2013," *NVSR*, vol. 64, no. 2, p. 119, 2016.

[2] F. Alfonso, R. A. Byrne, F. Rivero, and A. Kastrati, "Current treatment of in-stent restenosis," *Journal of the American College of Cardiology*, vol. 63, no. 24, pp. 2659–2673, 2014.

[3] M. Gaudry, J.-M. Bartoli, L. Bal, R. Giorgi, M. De Masi, P.-E. Magnan, and P. Piquet, "Anatomical and technical factors influence the rate of in-stent restenosis following carotid artery stenting for the treatment of post-carotid endarterectomy stenosis," *PLOS ONE*, vol. 11, no. 9, pp. 1–15, 09 2016.

[4] A. Kastrati, MD, A. Schmig, MD, S. Elezi, MD, H. Schhlen, MD, J. Dirschinger, MD, M. Hadamitzky, MD, A. Wehinger, MD, J. Hausleiter, MD, H. Walter, MD, and F.-J. Neumann, MD, "Predictive factors of restenosis after coronary stent placement," *Journal of the American College of Cardiology*, vol. 30, no. 6, pp. 1428–1436, 1997.

[5] R. E. Kuntz, C. Gibson, M. Nobuyoshi, and D. S. Baim, "Generalized model of restenosis after conventional balloon angioplasty, stenting and directional atherectomy," *Journal of the American College of Cardiology*, vol. 21, no. 1, pp. 15 – 25, 1993.

[6] O. Iida, M. Takahara, Y. Soga, K. Suzuki, K. Hirano, D. Kawasaki, Y. Shintani, N. Suematsu, T. Yamaoka, S. Nanto, and M. Uematsu, "Shared and differential factors influencing restenosis following endovascular therapy between {TASC} (trans-atlantic inter-society consensus) {II} class a to c and d lesions in the femoropopliteal artery," *JACC: Cardiovascular Interventions*, vol. 7, no. 7, pp. 792 – 798, 2014.

[7] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[8] E. Yang, G. Allen, Z. Liu, and P. K. Ravikumar, "Graphical models via generalized linear models," in *NIPS*, 2012, pp. 1358–1366.

[9] E. Yang, P. Ravikumar, G. I. Allen, and Z. Liu, "On Poisson graphical models," in *NIPS*, 2013, pp. 1718–1726.

[10] F. Hadiji, A. Molina, S. Natarajan, and K. Kersting, "Poisson dependency networks: Gradient boosted models for multivariate count data," *Machine Learning*, vol. 100, no. 2-3, pp. 477–507, 2015.

[11] J. Ma, K. Kockelman, and P. Damien, "A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods," *Accident Analysis and Prevention*, vol. 40, no. 3, pp. 964–975, 5 2008.

[12] B. M. Bolker, M. E. Brooks, C. J. Clark, S. W. Geange, J. R. Poulsen, M. H. H. Stevens, and J.-S. S. White, "Generalized linear mixed models: a practical guide for ecology and evolution." *Trends in ecology & evolution*, vol. 24, no. 3, pp. 127–35, Mar. 2009.

[13] G. I. Allen and Z. Liu, "A Local Poisson Graphical Model for Inferring Networks From Sequencing Data," *IEEE Transactions on NanoBioscience*, vol. 12, no. 3, pp. 189–198, Sep. 2013.

[14] E. B. Andersen, "Sufficiency and exponential families for discrete sample spaces," *Journal of the American Statistical Association*, vol. 65, no. 331, pp. 1248–1255, 1970.

[15] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, 2001.

[16] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," *Found. Trends Mach. Learn.*, vol. 1, no. 1-2, pp. 1–305, Jan. 2008.

[17] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009.