Intervention-Driven Predictive Framework for Modeling Healthcare Data

Santu Rana*, Sunil Kumar Gupta*, Dinh Phung, Svetha Venkatesh

Center for Pattern Recognition and Data Analytics
Deakin University, Australia 3216
{santu.rana,sunil.gupta,dinh.phung,svetha.venkatesh}@deakin.edu.au

Abstract. Assessing prognostic risk is crucial to clinical care, and critically dependent on both diagnosis and medical interventions. Current methods use this augmented information to build a single prediction rule. But this may not be expressive enough to capture differential effects of interventions on prognosis. To this end, we propose a supervised, Bayesian nonparametric framework that simultaneously discovers the latent intervention groups and builds a separate prediction rule for each intervention group. The prediction rule is learnt using diagnosis data through a Bayesian logistic regression. For inference, we develop an efficient collapsed Gibbs sampler. We demonstrate that our method outperforms baselines in predicting 30-day hospital readmission using two patient cohorts - Acute Myocardial Infarction and Pneumonia. The significance of this model is that it can be applied widely across a broad range of medical prognosis tasks.

1 Introduction

Medical interventions cure us, and keep us alive. They form the cornerstone of modern medical practice. Doctors carefully study the clinical observations related to our illness, and perform interventions. To formulate the most effective post-discharge care plan, they assess the prognosis. For example, what is the risk of readmission? How long will this person live? Answering these questions requires risk prediction models.

A patient's condition captures usual risk factors that can then be used in prognostic models. But medical interventions performed on patients are confounding, changing the outcome and thus prediction rules. For example, different cancer treatments (such as radiotherapy, chemotherapy or their combinations) have different prognosis profiles for the same tumor type [1]. Similarly, prognosis of cardiac patients for different procedures are different [2]. Thus interventions should be taken into account when developing prediction models.

Traditionally, in the healthcare community both the patient conditions and interventions are augmented together and a single prediction rule is learnt [3]. A single rule, however, may not be expressive enough to capture differential rules

* Contributed equally

due to different interventions. Current predictive methods, such as logistic regression (LR), Support vector machine (SVM), Naïve Bayes (NB), and Random Forest (RF) require amalgamation of interventions with the patient condition variables, and suffer from the same limitation. At the other extreme, learning prediction rules for each intervention separately is not useful either - out of hundreds of unique interventions, all are not equally important and many of them are performed together (as groups of interventions) - for a variety of reasons including current treatment polices, hospital capacity and cost. This opens up the need to learn a set of intervention groups and group-specific risk prediction models.

Following this, we propose a nonparametric, supervised framework that uses a mixture distribution over interventions, learning a prediction model for each mixture component. A Dirichlet Process (DP) prior over interventions mixture is used allowing extraction of latent intervention groups, for which the number of groups is not known *a priori*. The outcome is then modeled as conditional on this latent grouping and patient condition data through a Bayesian logistic regression (B-LR). The use of DP also allows formation of new intervention groups when necessary, thus coping with changes in medical practice. In addition, the intervention based clustering inferred by the model is made predictive. This encourages formation of intervention groups that lead to a low prediction error. We refer to this model as DPM-LR. Efficient inference is derived for this model.

To evaluate our model, prediction of 30-day readmission on two retrospective cohorts of patients from an Australian hospital is considered: 2652 admissions related to Acute Myocardial Infarction (AMI) between 2007-2011 and 1497 admissions related to Pneumonia between 2009-2011. On both the cohorts, DPM-LR outperforms several baselines - dpMNL [4], Bayesian Logistic Regression, SVM, Naïve Bayes and Random Forest. We show that the intervention groups discovered using DPM-LR are clinically meaningful. We also illustrate that the highest risk factors identified by DPM-LR for different intervention groups are different, validating the necessity of intervention-driven predictive modeling.

In summary, our main contributions are:

- A nonparametric Bayesian, supervised prediction framework (DPM-LR) that explicitly models interventions and extracts latent groups by imposing a Dirichlet Process Mixture over interventions. The prognosis is modeled as conditional on this latent grouping and patient condition data through a Bayesian logistic regression (B-LR).
- Efficient inference for DPM-LR is derived and implemented.
- Validation on both synthetic and two real-world patient cohorts, demonstrating better performance by model over state-of-the-art baselines.

2 Background

Hospital readmissions are common and costly. The 30-day readmission rate among the Medicare beneficiaries in the USA is estimated at 18%, costing $17 billion [5]. Some hospital readmissions are considered avoidable and thus 30-day readmission rates are used for benchmarking across hospitals, with financial

penalties for hospitals with high risk-adjusted rates [5]. Avoidable readmissions can be avoided by appropriately planning post-discharge care [6]. This requires accurate risk prediction.

Few models exist in the healthcare community to predict 30-day readmission risk in general medical patients [7,8,3]. All these methods employ Logistic Regression to derive a score based system for risk stratification using retrospective clinical and administrative data collected mainly from Electronic Health Records. Readmission prediction using other machine learning techniques such as SVM, Naïve Bayes and Random Forest have been studied respectively for heart-failure patients in [9] and for ten different diseases [10]. In all the methods, both the patients condition and interventions are augmented together to learn a single prediction rule.

A single rule, however, may not be sufficient to model the effect of different interventions. On the contrary, learning prediction rules for each intervention is not necessary - out of all the unique interventions, many of them are performed together and only a few latent groups exist. This gives rise to the need to learn the set of intervention groups and group-specific prediction models. The intervention grouping can be learnt using a mixture distribution with a Dirichlet Process prior to account for the unknown number of groups.

The use of Dirichlet process (DP) has been previously studied for modeling a set of classifiers under mixture model settings. In an attempt to develop a nonlinear classifier, Shahbaba and Neal [4] use DP as a prior for dividing data in clusters learning a separate linear classifier for each cluster. This model (dpMNL) learns nonlinear boundaries through a piecewise linear approximation. The idea from this model can be adapted for dividing patients for different intervention groups. Instead of using a single feature for both clustering and classification, we can use interventions to cluster the patients, and learn separate classifiers using patient condition features for each of the intervention groups.

3 Framework

We describe a prediction framework that learns a set of latent, predictive intervention groups and builds a prediction rule for each intervention group. In developing such a framework, our intention is to develop a predictive model that is flexible in modeling the effect of medical interventions on patient condition variables and outcome.

Typically, healthcare data has the following form: for each patient, we have a list of patient conditions (denoted by $\mathbf{x}$), a list of medical interventions (denoted by $\mathbf{i}$) and an outcome variable (denoted by y). We denote the data as $D = \{(\mathbf{x}_n, \mathbf{i}_n, y_n) \mid n = 1, \ldots, N\}$ where $\mathbf{x}_n \in \mathbb{R}^{M_x \times 1}$, $\mathbf{i}_n \in \mathbb{R}^{M_i \times 1}$.

To model the effect of interventions, we cluster the interventions into a set of predictive groups. A Dirichlet process mixture (DPM) over interventions is used to extract a set of latent intervention groups. The use of DPM allows us to form new intervention groups when necessary and thus copes with changes in hospital practices and policies. Further, the intervention-based clustering is

made predictive so that it encourages formation of intervention groups that lead to a low predictive error. Given such clustering, we learn a separate classifier for each intervention group. We refer to this model as DPM-LR.

The generative process of DPM-LR can be described as follows: A random probability measure G is drawn from a Dirichlet process $\mathrm{DP}(\alpha, H)$ where α is a positive concentration parameter and H is a fixed base measure. Since we are using a DP prior, the random measure G is discrete with probability one [11]. In stochastic process notation, we can write:

$$G \sim \mathrm{DP}(\alpha, H),\ \psi_n \sim G,\ \{\mathbf{x}_n, \mathbf{i}_n, y_n\} \sim \psi_n \tag{1}$$

Stick-breaking construction of Dirichlet process [12] often provides more intuitive and clearer understanding of DP-based models. Using stick-breaking notation, the above generative process can be written as:

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k} \tag{2}$$

where θ_k are independent random variables (also called "atoms") distributed according to H. Further, δ_{θ_k} denotes an atomic measure at θ_k and π_k are the "stick-breaking weights" such that $\sum_k \pi_k = 1$. For our model, the variable θ_k takes values in a product space of two independent variables ϕ_k and $\mathbf{w}_k$. Thus, we can explicitly write $\theta_k \equiv \{\phi_k, \mathbf{w}_k\}$. For DPM-LR model, the ϕ_k can be interpreted as k-th "intervention topic" while the $\mathbf{w}_k$ is the classifier weight vector for k-th intervention topic. We model $\phi_k \sim \mathrm{Dir}(\lambda)$, i.e. a Dirichlet distribution with parameter λ and $\mathbf{w}_k \sim \mathcal{N}\left(\mathbf{0}, \sigma_w^2 \mathbf{I}\right)$, i.e. a multivariate normal distribution with zero mean and single standard deviation parameter σ_w. The two representations (the stochastic and the stick-breaking) can be tied by introducing an indicator variable z_n such that $\psi_n \equiv \theta_{z_n}$. We summarize the generative process as:

$$\pi \sim \mathrm{GEM}(\alpha),\ (\phi_k, \mathbf{w}_k) \overset{\text{iid}}{\sim} H(\lambda, \sigma_w),\ H(\lambda, \sigma_w) = \mathrm{Dir}(\lambda) \times \mathcal{N}\left(\mathbf{0}, \sigma_w^2 \mathbf{I}\right) \tag{3}$$

For $n = 1, \ldots, N$

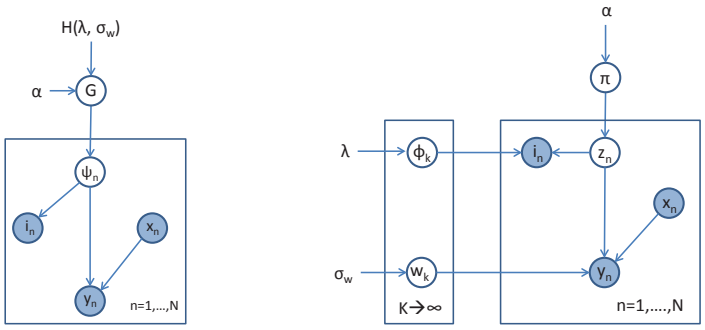
$$z_n \sim \mathrm{Discrete}(\pi),\ \mathbf{i}_n \mid z_n, \phi \sim \Pi_{m=1}^{M_i} \mathrm{Discrete}(\phi_{z_n}) \tag{4}$$

$$y_n \mid \mathbf{x}_n, z_n, \mathbf{w} \sim \mathrm{Ber}\left(f\left(\mathbf{w}_{z_n}^{\mathsf{T}} \mathbf{x}_n\right)\right) \tag{5}$$

where GEM distribution is named after the first letters of Griffiths, Engen and McCloskey [13]. $\mathrm{Ber}(.)$ and $\mathrm{Dir}(.)$ denote the Bernoulli and Dirichlet distributions, respectively and $f(.)$ denotes the logistic function. Graphical representations of DPM-LR is shown in Figure 1.

4 Inference

The inference of parameters in a fully Bayesian model is performed by sampling them from their joint posterior distribution, conditioned on the observations. For DPM-LR model, this distribution does not take a closed form. A popular way to



(a) (b)

Fig. 1: Graphical representation of the DPM-LR (a) the stochastic process view (b) the stick-breaking view.

circumvent this problem is to approximate this distribution using Markov chain Monte Carlo (MCMC) sampling. Asymptotically, the samples obtained using MCMC are guaranteed to come from the true posterior distribution [14]. We use Gibbs sampling (a MCMC variant) - an algorithm that iteratively samples a set of variables conditioned upon the remaining set of variables and the observations. The MCMC parameter state space consists of the variables $\{\pi, z, \phi, \mathbf{w}\}$ and the hyperparameters λ, α and σ_w. To improve the sampler mixing, we integrate out π, ϕ and only sample variables $\{\mathbf{z}, \mathbf{w}\}$ and the hyperparameter α. The hyperparameters λ and σ_w are fixed to one. After the sampler convergence, we finally estimate ϕ as it provides useful insights into different intervention groups.

Since our model uses a Dirichlet process (DP) prior, Gibbs sampling of variable ϕ conditioned on other variables remains identical to the standard DP mixture model. However, due to the changes in the generative process caused by altering the model into a supervised setting, the Gibbs sampling updates for the variables z and $\mathbf{w}$ need to be derived.

4.1 Sampling z

We sample the variable z_n from Gibbs conditional posterior integrating out π and ϕ from the model. For the assignment of z_n, there are *two* possibilities: (1) the intervention $\mathbf{i}_n$ is assigned to an existing intervention cluster, i.e. given K clusters, z_n takes a value between 1 and K (2) the intervention $\mathbf{i}_n$ is assigned to a *new* intervention cluster, i.e. z_n is set to $K+1$. For the *former* case, the Gibbs sampling updates can be obtained from the following posterior distribution:

$$\begin{aligned} &\text{For } k = 1, \ldots, K \\ &\quad p\left(z_n = k \mid \ldots\right) = p\left(z_n = k \mid \mathbf{z}^{-n}, \mathbf{i}^{-n}, \mathbf{i}_n, y_n, \mathbf{x}_n, \mathbf{w}\right) \\ &\quad \propto \underbrace{p\left(\mathbf{i}_n \mid z_n = k, \mathbf{z}^{-n}, \mathbf{i}^{-n}\right)}_{\text{intervention likelihood}} \underbrace{p\left(y_n \mid z_n = k, \mathbf{x}_n, \mathbf{w}\right)}_{\text{class likelihood}} \underbrace{p\left(z_n = k \mid \mathbf{z}^{-n}\right)}_{\text{predictive prior}} \end{aligned} \tag{6}$$

In the above posterior, three terms interact: *intervention likelihood* (how likely is the cluster k for intervention $\mathbf{i}_n$ given other interventions), *class likelihood* (if $\mathbf{i}_n$ is assigned to cluster k, how small would be the classification error for the k-th cluster) and the predictive prior (the prior probability of an intervention being assigned to the cluster k given other assignments). For the case when z_n is assigned to a *new* cluster, the Gibbs sampling updates can be obtained from the following posterior distribution:

$$\begin{aligned} &p\left(z_n = K+1 \mid \ldots\right) \\ &\propto \underbrace{p\left(\mathbf{i}_n \mid z_n = K+1\right)}_{\text{intervention likelihood}} \underbrace{p\left(y_n \mid z_n = K+1, \mathbf{x}_n\right)}_{\text{class likelihood}} \underbrace{p\left(z_n = K+1 \mid \alpha\right)}_{\text{predictive prior}} \end{aligned} \tag{7}$$

The class likelihood term in the above expression requires integrating out a Bernoulli likelihood with respect to $\mathbf{w}_{K+1}$. We approximate this integral numerically using Monte Carlo samples of $\mathbf{w}_{K+1}$.

4.2 Sampling $\mathbf{w}_k$

Using the generative process of (3-5), the Gibbs conditional posterior of $\mathbf{w}_k$ can be written as:

$$\begin{aligned} p\left(\mathbf{w}_k \mid ...\right) &= p\left(\mathbf{y}^k \mid \mathbf{w}_k, \mathbf{X}^k\right) p\left(\mathbf{w}_k \mid \sigma_w\right) \\ &\propto \left[\Pi_{i=1}^{n_k} \left(s_i^k\right)^{y_i^k} \left(1 - s_i^k\right)^{1-y_i^k}\right] e^{-\mathbf{w}_k^\mathsf{T}\mathbf{w}_k/2\sigma_w^2} \end{aligned} \tag{8}$$

where we define $\mathbf{X}^k \triangleq \{\mathbf{x}_n \mid z_n = k\}$, which contains the patient condition data from the k-th intervention group and $\mathbf{x}_i^k$ is the i-th data column of $\mathbf{X}^k$. Further, we have $N^k \triangleq \#\{n \mid z_n = k\}$ and $s_i^k \triangleq f\left(\mathbf{w}_k^\mathsf{T}\mathbf{x}_i^k\right)$. The direct sampling from the above posterior is not possible as this does not reduce to any standard distribution. However, we can approximate the density using Laplace approximation [15,16]. The idea is to find the mode of the posterior distribution through an optimization procedure and then fitting a Gaussian with its mean at the computed mode. Instead of optimizing the posterior directly, we optimize the *logarithm* of the posterior (results are unaltered due to monotonicity of logarithm), for which it is possible to compute the first and the second derivatives in closed form. The first and the second derivatives of the log posterior are given as:

$$\nabla_{w_k} \ln p\left(\mathbf{w}_k \mid ...\right) = \sum_{i=1}^{n_k} \left(y_i^k - s_i^k\right) \mathbf{x}_i^k - \frac{1}{\sigma_w^2} \mathbf{w}_k \tag{9}$$

$$\nabla_{w_k}^2 \ln p\left(\mathbf{w}_k \mid ...\right) = -\mathbf{X}^k \mathbf{D}_s\left(\mathbf{w}_k\right) \left(\mathbf{X}^k\right)^\mathsf{T} - \frac{\mathbf{I}}{\sigma_w^2} \tag{10}$$

where $\mathbf{D}_s\left(\mathbf{w}_k\right) \triangleq \operatorname{diag}\left(\left[s_1^k\left(1 - s_1^k\right), \ldots, s_{N^k}^k\left(1 - s_{N^k}^k\right)\right]\right)$ is a diagonal matrix with entries between 0 and 1. For the above optimization, we use quasi-Newton

(L-BFGS) method as it converges faster compared to steepest-descent given good initializations. The optimization solution (denoted as $\mathbf{w}_k^*$) is used as mean of the approximating Gaussian. The covariance matrix of the Gaussian is computed (in closed form) by taking the negative of the inverse of the Hessian of the log posterior, i.e. $\Sigma_{\mathbf{w}_k}^* = -\left[\nabla_{w_k}^2 \ln p\left(\mathbf{w}_k \mid ...\right)\right]^{-1}$. Given $\mathbf{w}_k^*$ and $\Sigma_{\mathbf{w}_k}^*$, the posterior samples of $\mathbf{w}_k$ are drawn from $\mathcal{N}\left(\mathbf{w}_k^*, \Sigma_{\mathbf{w}_k}^*\right)$.

4.3 Sampling $\boldsymbol{\phi_k}, \boldsymbol{\alpha}$

Sampling ϕ_k is not necessary for the prediction. However, since it provides useful insights into different intervention groups, we finally estimate (after the sampler convergence) it as $\hat{\phi}_{m,k} = \frac{n_{m,k}+\lambda}{\sum_{m=1}^{M_i}(n_{m,k}+\lambda)}$ where $n_{m,k}$ is the number of occurrences of the m-th intervention in the k-th group. Sampling of the hyperparameter α remains same as in standard DPM model. Further details can be found in [17].

4.4 Prediction for new observations

After training the model with data $D = \{(\mathbf{x}_n, \mathbf{i}_n, y_n) \mid n = 1, \ldots, N\}$, we have samples $\left\{\mathbf{w}^{(l)}, \mathbf{z}^{(l)}\right\}_{l=1}^{L}$. Given a new observation $\left\{\tilde{\mathbf{x}}, \tilde{\mathbf{i}}\right\}$, its outcome $\tilde{y}$ can be sampled from the following distribution:

$$p\left(\tilde{y} \mid \tilde{\mathbf{x}}, \tilde{\mathbf{i}}\right) \approx \frac{1}{L} \sum_{l=1}^{L} \sum_{\tilde{z}=1}^{K} p\left(\tilde{y} \mid \tilde{z}, \tilde{\mathbf{x}}, \mathbf{w}^{(l)}\right) p\left(\tilde{z} \mid \tilde{\mathbf{i}}, \mathbf{z}^{(l)}, \alpha\right) \tag{11}$$

The posterior $p\left(\tilde{z} \mid \tilde{\mathbf{i}}, \mathbf{z}^{(l)}, \alpha\right)$ can be computed similar to the corresponding terms in (6) as the model is not updated during the test phase.

5 Experiments

We perform experiments with a synthetic dataset and two hospital datasets. Baseline methods used for comparison are first presented followed by results on synthetic data. Finally, evaluation is performed on two patient cohorts.

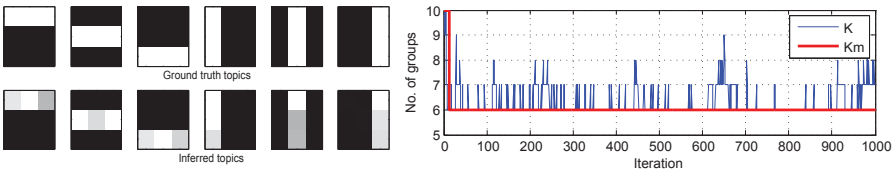
5.1 Baselines

We compare the predictive performance of DPM-LR with the following methods: (a) Standard DP-Multinomial Logit model, with Gaussian observation model (dpMNL)[4]. The method learns a nonlinear classifier with data constructed by augmenting patients condition and intervention features (b) An adaptation of dpMNL with Multinomial observation model (referred to as dpMNL(MM)) (c) Bayesian Logistic Regression (B-LR) (d) SVM with linear kernel (Linear-SVM) (e) SVM with 3rd order polynomial kernel (Poly3-SVM) (f) Naïve Bayes (g) Random Forest. Weka implementation [18] is used for the SVM, Naive Bayes and the Random Forest. For all the baselines, the feature vector is created by merging the patient condition and intervention features.

5.2 Experiments with Synthetic Data

The synthetic dataset spans 5 years with 100 unique patients per year. Nine different interventions are considered. Six intervention topics are created from horizontal and vertical bar patterns of a 3x3 matrix (Fig 2a). Per patient "intervention" feature is synthesized by sampling an intervention topic from a uniform mixture distribution and then sampling 4 interventions from the selected intervention topic. Each intervention topic is considered as an intervention group. The classification weight vector of each group is sampled from a 50-variate Normal distribution. The "patient condition" feature is randomly sampled from a set of 10 distinct random binary vectors. The label (or outcome) is computed by combining the group-specific classifier with patient data following (5). The prediction task is to predict labels for the patients in the 5th year. Default settings from Weka is used for SVM, Naïve Bayes and the Random Forest.

DPM-LR outperforms all the baselines (Table 1) in terms of AUC (Area under the ROC curve). DPM-LR outperforms (AUC 0.942) the closest contender, Random Forest (AUC 0.873). The performance of standard dpMNL (AUC 0.630) with Gaussian observation model was poor, however, the adapted version with multinomial observation model did reasonably well (AUC 0.836). All the other methods performed poorly (AUC$<$0.750). Figure 2b shows the number of intervention topics sampled over 1000 Gibbs iterations (including 500 burnins). It can be seen that the convergence to the true number of topics is achieved quickly (*i.e.* the mode of the number of groups (K_m) remains unchanged after about 50 iterations), implying stable estimate of the posterior. Intervention topics inferred by the DPM-LR closely match true intervention topics (Figures 2a).



(a) (b)

Fig. 2: Experiments on synthetic data (a) 6 intervention topics - True (top) and inferred (below) (b) Number of intervention topics (K) over Gibbs iterations and its running mode (K_m).

5.3 Experiments with Hospital Data

The data is collected from a large public hospital[1] in Australia. The hospital patient database provides a single point of access for information on patient hospitalizations, emergency department visits, in-hospital medications and treatments. Detailed records of these patient interactions with the hospital system are available through the EMR. This includes International Classification of Disease

[1] Ethics approval obtained through University and the hospital – Number 12/83

Methods	DPM-LR	dpMNL	dpMNL (MM)	B-LR	Linear-SVM	Poly3-SVM	Naive Bayes	Random Forest
AUC	**0.942**	0.630	0.836	0.719	0.568	0.735	0.686	0.873

Table 1: AUC for prediction on the synthetic dataset. Training is performed with 400 patients and testing with the remainder 100 patients.

10 (ICD-10) codes[2], Diagnosis-related Group (DRG) codes of each admission, ICD-10 codes for each emergency visit, details of procedures, and departments that have been involved in the patient's care. Other information includes demographic data (age, gender, and occupation) and details of the patient's access to primary care facilities.

Cohort 1: Acute Myocardial Infarction (AMI) The patient cohort consists of 2652 consecutive admissions with confirmed diagnosis of Acute Myocardial Infarction (AMI) admitted between 1st January 2007 and 31st December 2011. For each patient, we have a sequence of interactions with the hospital system. Of these, the discharge corresponding to an admission with primary reason for admission as AMI is treated as assessment points (APs) from which prediction is made. Patient records prior to an AP are used to construct features. The "patient condition" feature contains demographic (age, gender and occupation) and disease information (ICD-10 codes) for each admission, accumulated at four different time scales - past 1 month, past 3 months, past 6 months and past 1 year. The "intervention" feature consists of procedure codes associated with only the current admission. The label is set to one if there are any readmissions in 30-day period following an AP with a cardiac related diagnosis. Readmission rate in this cohort varied from 11.7% (2007) to 4.8% (2011).

Experimental Results Patient data from 2007-2010 are used for training and patient data from 2011 for testing. The comparative results with the baselines (Table 2) shows that DPM-LR outperforms all other methods.

DPM-LR is better (AUC 0.677) than the the closest contender dpMNL(MM) (AUC 0.641) by a significant margin. This is followed by dpMNL (AUC 0.635) and B-LR (AUC 0.607). All other methods have AUC less than 0.6. Surprisingly, more complex models such as SVM with polynomial kernel and the Random Forest perform the worst.

Table 3 lists the 5 strongest risk factors for the three intervention groups. These risk factors are the patient condition features that correspond to the largest positive weights in the linear regression model. We can see from the table that the strongest risk factors for different intervention groups are different. This vindicates the need of modeling intervention-specific prediction rules.

[2] `http://www.who.int/classifications/icd10/`

Cohort 2 - Pneumonia This cohort consists of 1497 admissions with confirmed diagnosis of Pneumonia, admitted between 1st January 2009 and 31st December 2011. Similar to AMI, the discharges corresponding to an admission with primary reason for admission as Pneumonia is treated as the assessment points (APs) from which prediction is made. Patient records prior to an AP are used to construct the features, in a similar fashion as in the AMI cohort described in the previous section. The label is set to one if there are any readmissions in 30-day period following an AP with respiratory related diagnosis. Readmission rate in this cohort varied between 5-6% over the study years (2009-2011).

Experimental Results The model is trained using patient data from 2009-2010 and then tested on patient data from 2011. The comparative results with the baselines are presented in Table 4. Once again, DPM-LR outperforms (AUC 0.667) the closest contender dpMNL(MM) (AUC 0.664).

DPM-LR learns two intervention groups. The risk factors corresponding to these two intervention groups are different (Table 5) - a point that was also observed for AMI cohort.

Methods	**DPM-LR**	**dpMNL**	**dpMNL (MM)**	**B-LR**	**Linear-SVM**	**Poly3-SVM**	**Naive Bayes**	**Random Forest**
AUC	**0.677**	0.635	0.641	0.607	0.576	0.516	0.577	0.566

Table 2: AUC for 30-day readmission prediction for the AMI cohort. Patient data from 2007-2010 is used for training. Test year is 2011.

6 Conclusion

We present a novel predictive framework for modeling healthcare data in the presence of medical interventions. This framework automatically discovers the latent intervention groups and builds group-specific prediction rules. A Dirichlet process mixture used over the intervention groups ensures that new groups are created when a new intervention is introduced. The prediction rule is learnt using patients condition data through a Bayesian logistic regression. Efficient inference is derived for this model. Experiments demonstrate that this method outperforms state-of-the-art baselines in predicting 30-day hospital readmission on two cohorts - Acute Myocardial Infarction and Pneumonia. As a future work, it would be interesting to explore the performance improvement through sharing across various intervention groups using Bayesian shared subspace learning [19].

References

1. M. Al-Sarraf, M. LeBlanc, P. Giri, K. K. Fu, J. Cooper, T. Vuong, A. A. Forastiere, G. Adams, W. A. Sakr, D. E. Schuller, and J. F. Ensley, "Chemoradiotherapy versus radiotherapy in patients with advanced nasopharyngeal cancer: phase iii randomized intergroup study," *Journal of Clinical Oncology*, vol. 16, no. 4, pp. 1310–1317, 1998.

Intervention group	Top 5 strongest risk factors for readmission
Coronary angioplasty with stenting, Coronary angiography, Examination procedures on ventricle, Generalised allied health interventions	Hypertension in the past 1 month Retired and pensioner Congestive heart failure in the past 1 month Obesity in the past 1 month Fluid and electrolyte disorders in the past 1 month
Coronary artery bypass, Coronary angiography, Examination procedures on ventricle, Generalised allied health interventions	Metastatic cancer in the past 1 month Depression in the past 1 month Diabetes, complicated in the past 1 month Congestive heart failure in the past 3 month Peripheral vascular disease in the past 1 month
No intervention	Obesity in the past 1 month Metastatic cancer in the past 1 year Solid tumor without metastasis in the past 3 month Fluid and electrolyte disorders in the past 3 month Age above 90 years

Table 3: Strongest risk factors associated with a 30-day readmission risk in the AMI cohort for three main intervention groups - *Coronary angioplasty with stenting, Coronary artery bypass, and No intervention.*

Methods	DPM-LR	dpMNL	dpMNL (MM)	B-LR	Linear-SVM	Poly3-SVM	Naive Bayes	Random Forest
AUC	**0.667**	0.590	0.664	0.640	0.523	0.511	0.635	0.561

Table 4: AUC for 30-day readmission prediction (pneumonia). Training is with patient data from 2007-2010. Test is with patient data from 2011.

Intervention group	Top 5 strongest risk factors for readmission
Generalized allied health intervention, Administration of blood and blood products, Administration of pharmacotherapy.	Iron deficiency anaemia in the past 1 month Lower respiratory infection in the past 1 month Angina pectoris in the past 1 month Acute kidney failure in the past 1 month Fluid and electrolyte disorders in the past 1 month
No intervention	Intestinal disorders in the past 1 month Congestive heart failure in the past 1 month Age between 70-80 years Acute myocardial infarction in the past 1 month Acute kidney failure in the past 1 month

Table 5: Strongest risk factors associated with a 30-day readmission risk in the pneumonia cohort for two main intervention groups.

2. E. L. Hannan, M. J. Racz, G. Walford, R. H. Jones, T. J. Ryan, E. Bennett, A. T. Culliford, O. W. Isom, J. P. Gold, and E. A. Rose, "Long-term outcomes of coronary-artery bypass grafting versus stent implantation," *New England Journal of Medicine*, vol. 352, no. 21, pp. 2174–2183, 2005.
3. J. Donzé, D. Aujesky, D. Williams, and J. L. Schnipper, "Potentially avoidable 30-day hospital readmissions in medical patientsderivation and validation of a prediction modelpotentially avoidable 30-day hospital readmissions," *JAMA internal medicine*, vol. 173, no. 8, pp. 632–638, 2013.
4. B. Shahbaba and R. Neal, "Nonlinear models using dirichlet process mixtures," *The Journal of Machine Learning Research*, vol. 10, pp. 1829–1850, 2009.
5. S. F. Jencks, M. V. Williams, and E. A. Coleman, "Rehospitalizations among patients in the medicare fee-for-service program," *New England Journal of Medicine*, vol. 360, no. 14, pp. 1418–1428, 2009.
6. E. H. Bradley, L. Curry, L. I. Horwitz, H. Sipsma, Y. Wang, M. N. Walsh, D. Goldmann, N. White, I. L. Piña, and H. M. Krumholz, "Hospital strategies associated with 30-day readmission rates for patients with heart failure," *Circulation: Cardiovascular Quality and Outcomes*, vol. 6, no. 4, pp. 444–450, 2013.
7. M. Omar Hasan MBBS, D. O. Meltzer, S. A. Shaykevich, C. M. Bell, P. J. Kaboli, A. D. Auerbach, T. B. Wetterneck, V. M. Arora, and J. L. Schnipper, "Hospital readmission in general medicine patients: a prediction model," *Journal of general internal medicine*, vol. 25, no. 3, pp. 211–219, 2010.
8. C. van Walraven, I. A. Dhalla, C. Bell, E. Etchells, I. G. Stiell, K. Zarnke, P. C. Austin, and A. J. Forster, "Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community," *Canadian Medical Association Journal*, vol. 182, no. 6, pp. 551–557, 2010.
9. N. Meadem, N. Verbiest, K. Zolfaghar, J. Agarwal, S.-C. Chin, and S. B. Roy, "Exploring preprocessing techniques for prediction of risk of readmission for congestive heart failure patients," 2013.
10. S. Cholleti, A. Post, J. Gao, X. Lin, W. Bornstein, D. Cantrell, and J. Saltz, "Leveraging derived data elements in data analytic models for understanding and predicting hospital readmissions," vol. 2012, p. 103, 2012.
11. T. S. Ferguson, "A bayesian analysis of some nonparametric problems," *The annals of statistics*, pp. 209–230, 1973.
12. J. Sethuraman, "A constructive definition of dirichlet priors," DTIC Document, Tech. Rep., 1991.
13. J. Pitman, *Combinatorial stochastic processes*. Springer-Verlag, 2006, vol. 1875.
14. W. R. Gilks, "Full conditional distributions," *Markov chain Monte Carlo in practice*, pp. 75–88, 1996.
15. N. E. Breslow and D. G. Clayton, "Approximate inference in generalized linear mixed models," *Journal of the American Statistical Association*, vol. 88, no. 421, pp. 9–25, 1993.
16. C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*. springer New York, 2006, vol. 1.
17. M. Escobar and M. West, "Bayesian density estimation and inference using mixtures," *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 577–588, 1995.
18. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explorations*, vol. 11, 2009.
19. S. Gupta, D. Phung, and S. Venkatesh, "A Bayesian nonparametric joint factor model for learning shared and individual subspaces from multiple data sources," in *Proc. of SIAM Int. Conference on Data Mining (SDM)*, 2012, pp. 200–211.