Stabilizing High-Dimensional Prediction Models Using Feature Graphs

Shivapratap Gopakumar, Truyen Tran, Tu Dinh Nguyen, Dinh Phung, and Svetha Venkatesh

***Abstract*—We investigate feature stability in the context of clinical prognosis derived from high-dimensional Electronic Medical Records. To reduce variance in the selected features that are predictive, we introduce Laplacian-based regularization into a regression model. The Laplacian is derived on a feature graph that captures both the temporal and hierarchic relations between hospital events, diseases and interventions. Using a cohort of patients with heart failure, we demonstrate better feature stability and goodness-of-fit through feature graph stabilization.**

Index keywords: Stability, Predictive models, Biomedical computing, Electronic medical records.

I. Introduction

Stability promotes reliability - in performance, estimation or interpretability. Commonly, stability relates to robust performance against reasonable perturbations in data, achieved through diverse methods such as Jackknife, bootstrap or cross-validation [1]. The stability of selected features is often overlooked in prediction models – particularly if consistent performance alone is the goal.

But feature stability matters. Even when the prognosis performance is robust. When building models from high dimensional data, feature selection algorithms choose a small subset of features that maximizes model performance. These features, predictive of the prognosis, are important because they could be hypothesis generating thus meriting further investigation [2]. In clinical situations, explaining the prognosis is as important as the prognosis itself. Consequently, consistent predictors in spite of data resampling, are critical for clinical adoption. Feature stability is crucial not only in clinical prognosis – as example, stable biomarkers aid model reproducibility in bioinformatics [3], [4].

Building clinical prediction models from Electronic Medical Records (EMR) faces serious challenges for stable feature selection. EMR data is temporal, strongly correlated and high dimensional [5]. Each of these aspects makes this task challenging. High dimensional data calls for sparsity inducing feature selection [6]. However, automatic feature selection, particularly in clinical data, has been known to cause instability in features resulting in non-reproducible models [7]. This problem is further aggravated by strong correlations in EMR data. Sparse models often pick the strongest features from the chosen sample-set [8]. Under data resampling, an alternate feature from the correlated pair could be selected causing significant variations to the feature weights during each training run [9]. This problem is illustrated in Fig. 1 – the mean weights of the top 50 predictors from routine EMR data for 6 months readmission for heart failure is shown. The top predictors selected by lasso-regularized model [10] have large variance in feature weights under bootstraps (Fig. 1) - thus rendering them unusable in a clinical setting.

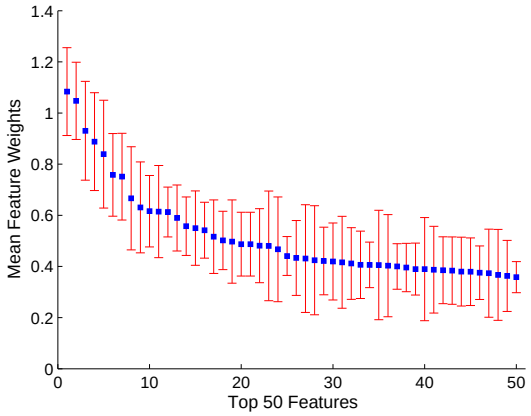


Figure 1: Feature instability due to data resampling. Mean weights vs standard deviation for the top 50 features selected by a lasso-regularized logistic regression model under bootstraps.

Addressing the open problem of stable feature selection in clinical settings we ask - Can we ensure the stability of predictors in a linear model for prognosis using EMR data? To measure the performance of this stability, we adopt variance in selected model parameters across data resamplings. For prognosis, we use a logistic regression model for 6 months readmission after heart failure - a deadly and costly disease with a majority of patients returning within a year after discharge. Automatic feature selection was achieved by the sparsity-promoting shrinkage method of lasso. To address our problem, we hypothesize that exploiting the inherent structures of EMR data to enforce statistical sharing may stabilize the prediction model. We consider temporal and hierarchical structures. Since features are accumulated over multiple time granularities (1 month, 3 months, etc.), features that lie in consecutive time periods are considered to be related. The hierarchies are exploited through the semantics in the ICD-10 tree[1] and the procedure cube (ACHI)[2] - codes that share similar prefix are considered to be related. We embed these relations in a feature graph. When the feature graph regularization term is added into the lasso model, weights assigned to related features will tend to be similar.

[1] http://apps.who.int/classifications/icd10

[2] https://www.aihw.gov.au/procedures-data-cubes/

The model was derived from feature rich EMR records of $1,405$ patients from Barwon health, a regional hospital in Australia. The model estimation was stabilized by utilizing a $3,338 \times 3,338$ feature graph constructed from hierarchical relations in ICD-10 disease tree and temporal abstraction of clinical events. We used the Jaccard Index [11] and Consistency Index [12] to measure feature stability of the top features selected, with and without graph stabilization. We further compared the robustness of the features from our graph stabilized model with state-of-the-art elastic net regularized model [8]. Both Jaccard Index and Consistency Index confirmed better feature stability for feature graph regularized model. The graph stabilized model also resulted in better goodness-of-fit as confirmed by Hosmer-Lemeshow test. The validation of our model on a held-out set resulted in an AUC of 0.66 (95%, CIs: $[0.60, 0.71]$) which is competitive against existing models that predict heart failure readmission [13], [14].

Our novelty is to identify the importance of the stable feature selection problem in clinical settings and to propose a solution based on additional regularization of a lasso model exploiting knowledge about hierarchical structures in disease and interventions and temporal relationships between events. Specifically, embedding these relations reduces the fragmentation of selection in the lasso model, delivering our goal of feature stability. The significance of our contribution is to reset the thinking of prognosis from "model performance only" to "model performance and feature stable models" - without these two components many of our advanced models will be rendered futile in a clinical setting.

A. Related Work

Despite advances in learning models for high dimensional data, stability in feature selection has received limited attention. Initial studies focused on comparing different feature selection algorithms based on stability of feature preferences [15], [16]. Kalousis *et. al*. [16] compared the stability of five popular feature selection algorithms on 11 datasets taken from three different application domains. Feature stability was investigated based on weight-scores, rank and selected feature subsets. No algorithm was found to be superior and it was concluded that feature stability depends significantly on the data set used. Another focus in stability studies is the development of various measures of stability. A recent survey [4] consolidated seven metrics for computing similarity measure of feature subsets.

Feature instability has been a serious concern to the bioinformatics domain, largely due to the nature of data. Early work on this topic proposed ensemble ranking, feature bias from prior knowledge, and grouping redundant data [2], [17]. Recent studies utilize prior biological knowledge and pathway information to enhance the stability of biomarkers. These information, compiled from many years of research, is made available through online databases like KEGG, HPRD, Pathway Commons, Reactome, BioCarta and BioCyc [18], [19]. Context specific data extracted from such databases can be used to create a graph network with nodes as genes or gene products and edges as interactions or relationships [18]. Such networks can be used to stabilize learning models by either a filter based approach or using embedded feature selection techniques [19].

The data in clinical prediction domain is similar to bioinformatics – features are correlated, high dimensional and size of cohorts under study are usually small ($p \gg n$). A natural solution to this problem is to select a subset of features from prior clinical knowledge. A recent study used only a subset of EMR features for predicting heart failure readmission [5]. Our work is inspired from the bioinformatics domain of using network information to stabilize high dimensional models. We differ from the traditional approach in feature stability by constructing the feature network graph from inherent structure and relations in the training data. The feature graph is used to stabilize a lasso regularized linear model for predicting heart failure readmission in 6 months. We wish to emphasize that the feature extraction process and construction of feature graphs depend solely on the hierarchical and temporal nature of EMR data and is not based on prior studies or predefined clinical knowledge.

II. Method

We present a stabilizing method for building prediction models from Electronic Medical Records (EMR). A typical EMR is very high dimensional. If consists of demographic information (e.g., age, gender and postcode) and time-stamped events (e.g., hospitalizations, ED visits, clinical tests, diagnoses, pathologies, medications and treatments). High dimensional data necessitate automatic feature selection. We choose the sparsity-promoting shrinkage method of lasso [20] as it is effective in handling very high-dimensional variables [21]. The methods are applicable to any member in the family of generalized linear models [22].

A. Feature Selection using Lasso

Let $\mathcal{D} = \{\boldsymbol{x}_\ell, y_\ell\}_{\ell=1}^n$ be the training dataset in which $\boldsymbol{x}_\ell \in \mathbb{R}^p$ denotes the high-dimensional feature vector of data instance ℓ and y_ℓ is the outcome (for example, the occurrence of future readmission). Our aim is to model the predictive distribution $P(y \mid \boldsymbol{x}; \boldsymbol{w})$ where $\boldsymbol{w} \in \mathbb{R}^p$ are sparse feature weights. The weights are estimated by maximizing the lasso penalized log-likelihood [10]:

$$\mathcal{L}_{\text{lasso}} = \frac{1}{n} \log \mathcal{L}(\boldsymbol{w}; \mathcal{D}) - \alpha \sum_i |w_i| \tag{1}$$

where $\log \mathcal{L}(\boldsymbol{w}; \mathcal{D}) = \sum_{\ell=1}^n \log P(y_\ell \mid \boldsymbol{x}_\ell, \boldsymbol{w})$ is the data log-likelihood, and $\alpha > 0$ is the penalty controlling the sparseness of the feature weights. Under lasso, weights of weak features are driven towards zeros, and thus the resulting model is sparse.

Unfortunately, sparsity-inducing models are susceptible to data variations resulting in loss of stability [7], [9]. For strong but highly correlated features, lasso often chooses one of the two [8], resulting in only a 0.5 chance for strongly predictive feature pairs. EMR-derived features amplify this selection instability because of (a) the high degree of redundancy in hospital-recorded events, and (b) a large portion of features

could be weakly predictive for some tasks, leading to low selection probabilities [23].

One popular solution to the correlated features is elastic net [8], which modifies the lasso regularization in Eq. (1) as follows:

$$\mathcal{L}_{\text{elastic.net}} = \frac{1}{n}\log \mathcal{L}\left(\boldsymbol{w}; \mathcal{D}\right) - \alpha\left(\lambda \sum_i |w_i| + (1-\lambda)\sum_i w_i^2\right) \tag{2}$$

where $\lambda \in [0,1]$ controls the relative contribution of the lasso term $\sum_i |w_i|$ and the ridge regression term $\sum_i w_i^2$. This method encourages features to have similar weights, and thus reducing the effect of selection by chance in pure lasso.

B. Stabilization using Feature Graph

We propose an alternative solution by encouraging shared statistical strength among correlated features. This is achieved by exploiting two relational structures in the EMR data. The first is the temporal relations that accumulates events (diagnoses and procedures) at different time granularities (Section II-C1). The second is the hierarchical structures captured through the disease classification semantics in the ICD-10 diagnosis tree and procedures codes. An undirected *feature graph* is then built with its edges representing the relations between two features.

Let $\boldsymbol{A} \in \mathbb{R}^{p\times p}$ be the incident matrix of the feature graph, i.e., $A_{ij} = 1$ if feature i and j are related and $A_{ij} = 0$ otherwise. Sharing statistical strength between any two related features is realized by enforcing the similarity in their weights, i.e., a graph-regularizing term is added to Eq. (1):

$$\mathcal{L}_{\text{Laplacian}} = \mathcal{L}_{\text{lasso}} - \frac{1}{2}\beta \sum_{ij} A_{ij}\left(w_i - w_j\right)^2 \tag{3}$$

where $\beta > 0$ is the correlation coefficient controlling the effect of the graph-based regularization. The graph-regularizer can be simplified as: $\frac{1}{2}\sum_{ij} A_{ij}\left(w_i - w_j\right)^2 =$

$$\begin{aligned} &= \sum_i \left(\sum_k A_{ik}\right) w_i^2 - \sum_i \sum_j A_{ij} w_i w_j \\ &= \boldsymbol{w}'\boldsymbol{L}\boldsymbol{w} \end{aligned}$$

where $\boldsymbol{L}$ is the Laplacian matrix of feature graph $\boldsymbol{A}$, i.e., $L_{ii} = \sum_j A_{ij}$ and $L_{ij} = -A_{ij}$ [24].

The Laplacian regularizer combats the instability in several ways. First, features of the same type tend to cluster, and thus their weights are more difficult to vary as a whole. Weaker features can thus borrow the statistical strength from the stronger ones. Second, two strongly correlated features must either be selected or suppressed jointly by the lasso.

C. Model Development

We present a framework for realizing the stabilization strategy described above. The framework consists of a training phase using data from the past and a validation phase using new admission data from the future (Fig. 2 for the workflow diagram). Our model development consists of three sub-phases: (i) multi-granular temporal feature extraction, (ii) feature graph construction based on the temporal relations and coding hierarchies, and (iii) model training with feature selection and feature graph regularization.

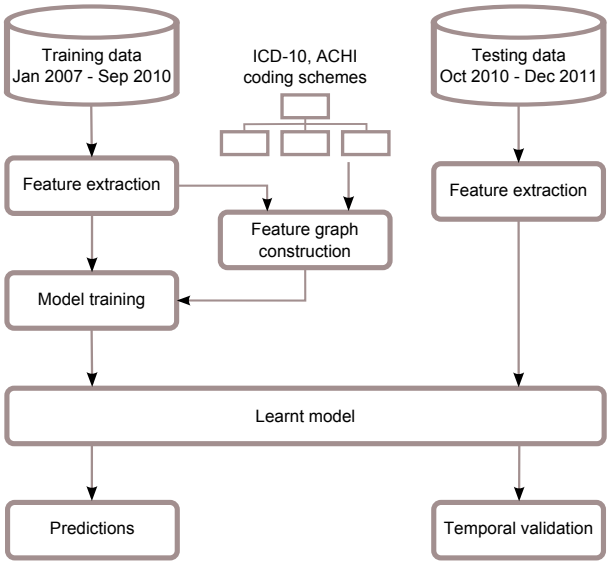


Figure 2: The workflow diagram of the framework for deriving graph-stabilized prediction models from Electronic Medical Records. Temporal feature relations and coding hierarchies were used to construct the feature graph (Fig. 4).

1) Multi-granular Temporal Feature Extraction : Feature extraction from EMR transforms inpatient time-stamped events (e.g., hospitalizations, clinical tests, diagnoses and treatments) into a high-dimensional feature vector at the index discharge. The challenges are that recorded events are sparse and irregular. As diseases progress in different paces, it is important to take multiple time scales into account. In addition, recent critical events carry more weight than mild conditions observed far back in the history. To this end, we employ the *one-sided convolutional filter bank* recently introduced in [25]. The filter bank summarizes event statistics over multiple time periods and granularities: (0-3), (3-6), (6-12), (12-24), (24-48), (48-72) months.

2) Feature Graph Construction: The feature graph is built by identifying connections between features that observe temporal and structural relations. Two features are connected if they satisfy one of the following two conditions. The first condition is the codes are identical and the periods are consecutive. This represents the disease progression over the time, for example, from the period of 3-6 months to the period of 0-3 months before the discharge. Alternatively, the periods are identical and the codes share the first two characters. This captures the diagnostic or therapeutic relations. For instance, two related features are the ICD-10 code *I25* (chronic ischaemic heart failure) and *I21* (acute myocardial infarction).

D. Validation Protocol

We validated the stabilization strategy on 6-month unplanned readmission prediction among patients suffering heart failure. As it is a binary outcome, logistic regression was used as the predictive model $P(y \mid \boldsymbol{x}; \boldsymbol{w})$.

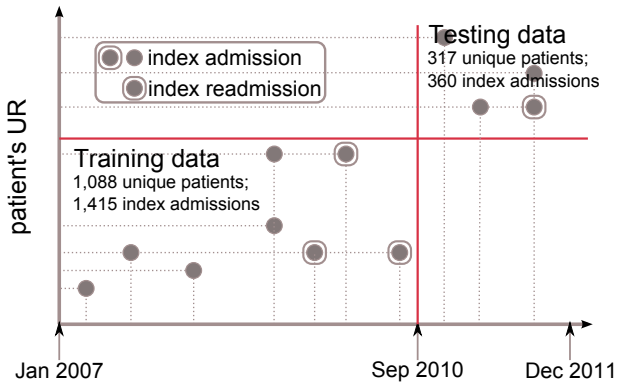


Figure 3: Training and test data: Time of hospitalization (x-axis) and unique patient id (y-axis), showing patient and temporal split. The temporal split of training and test data is made on 1st September 2010. The test and training set are disjoint in chosen patients.

1) Data: The data was collected from Barwon Health, a regional health service provider in Victoria, Australia. Ethics approval was obtained from the Hospital and Research Ethics Committee at Barwon Health (number 12/83) and Deakin University. Patient details are stored in EMR databases. The cohort of inpatients with heart failure contains $1,405$ unique patients with $1,885$ index admissions between January 2007 and December 2011. We identified patients as having heart failure if they had ICD-10 diagnosis code I50. Patients of all age groups were included. Inpatient deaths were excluded. We focused our study on emergency attendances or unplanned admissions of patients.

2) Temporal Validation: The model was externally validated in time [26]. That is, patients discharged prior to 1st September 2010 were used for training, and a separate set of those discharged afterward for testing (Fig. 3). This validation strategy was chosen because it better reflects the common practice of training the model in the past and using it in the future. Model performance was evaluated using measures of sensitivity (recall), specificity, precision, F-measure and AUC (area under the ROC curve) with confidence intervals based on Mann-Whitney statistic [27]. We used a predefined threshold to predict readmissions. The value of the threshold was chosen to maximize the F-measure computed from the training data.

3) Measuring Goodness-of-Fit: We used the Hosmer-Lemeshow test to measure the goodness-of-fit for our logistic regression models. The Hosmer-Lemeshow test [28] assesses the degree of fit by matching the observed probabilities with the estimated probabilities. The validation set is divided into G ordered groups based on estimated probability of outcome events. The Chi-squared test statistic is calculated by comparing the expected and observed number of outcome events in each group as:

$$\chi_{\text{HL}} = \sum_{g=1}^{G} \frac{(O_g - E_g)^2}{E_g(1 - E_g/n_g)} \tag{4}$$

where O_g= number of observed events in group g, E_g= number of expected events in group g, and n_g= number of observations in group g. For an ideal test, we have $G > 5$, $E_g > 5$ and $n_g = n_{g'}$, $(g, g') \in G$. When the significance of χ_{HL} is less than .05, we reject the null hypothesis which states there is no difference between estimated values and observed values. A large value for the test statistic with small significance (p-value < 0.05) indicates poor model fit while a small test statistic with large significance (p-value closer to 1) indicates a better fit [29].

4) Measuring Model Stability: Models were trained K times on K bootstraps. *Model estimation stability* is defined as variance in parameters. A measure is the signal-to-noise ratio (SNR):

$$\text{SNR}(i) = \frac{\bar{w}_i}{\sigma_i} \tag{5}$$

where $\bar{w}_i$ is the mean feature weight across bootstraps for feature i, and σ_i is its standard deviation. The higher absolute SNR, the more stable the feature is.

Feature selection stability is an alternative aspect of model stability. For each bootstrap, a subset of features is then formed by selecting top k features from the ranked list. Features were ranked by their importance. For each feature, importance was calculated as the product of its weight and the standard deviation in the training data [30]. The importance is thus scale-insensitive. We normalized the feature importance measures in the range of $[0, 100]$. Finally, we obtained a list of feature subsets $S = \{S_1, S_2, ..., S_K\}$ where $|S_i| = k$. To quantify selection stability, we used the Jaccard Index [11] and the Consistency Index [12].

- The *Jaccard Index* measures similarity as a fraction between cardinality of intersection and union feature subsets. Given two feature sets S_i and S_j, the pairwise Jaccard Index reads:

$$J_C(S_i, S_j) \quad = \quad \frac{|S_i \cap S_j|}{|S_i \cup S_j|} \tag{6}$$

  The Jaccard Index evaluating all K subsets was computed as follows:

$$J_S = \frac{2}{K(K-1)} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} J_C(S_i, S_j) \tag{7}$$

  Jaccard Index is bounded in $[0, 1]$.
- The *Consistency Index* corrects the overlapping due to chance. Considering a pair of subsets S_i and S_j, the pairwise Consistency Index I_C is defined as:

$$I_C(S_i, S_j) = \frac{rd - k^2}{k(d-k)} \tag{8}$$

  in which $|S_i \cap S_j| = r$ and d is the number of features (see Sec. II-A). Taking the average of all pairs, the overall

Consistency Index is:

$$I_S = \frac{2}{K(K-1)} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} I_C(S_i, S_j) \tag{9}$$

The Consistency Index is bounded in $[-1, +1]$.

III. Results

	Derivation	Validation
Number of admissions	1,415	369
Unique patients	1,088	317
Gender:		
Male	541 (49.7%)	155 (48.9%)
Female	547 (50.2%)	162 (51.1%)
Mean age (years)	78.3	79.4
Length of Stays:		
1-4 days	668 (61.4%)	209 (65.9%)
5 or more days	420 (38.6%)	108 (35.1%)

Table I: Training and validation cohorts characteristics.

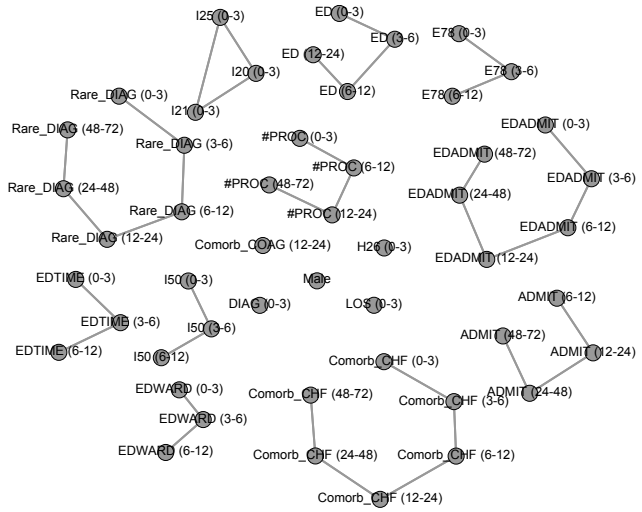


Figure 4: Feature subgraph of top risk factors. Numbers in brackets are time intervals, measured by months, before the index discharges. Factors selected are: *Male*; recent length of stay (*LOS*)); heart failure (*I50*, *Comord_CHF*); recent ischaemic heart diseases (angina pectoris (*I20*), acute myocardial infarction (*I21*), chronic ischaemic heart disease (*I25*)); any time rare diagnoses (*Rare_DIAG*); time stayed in emergency department (*EDTIME*); frequencies of emergency attendance (*ED*), unplanned admissions (*EDWARD, EDADMIT*), admissions (*ADMIT*), diagnoses (*DIAG*) and procedures (*#PROC*); and disorders of lipoprotein metabolism (*E78*) .

The characteristics of the training and validation cohort are summarized in Table I. The feature extraction process (Sec. II-C1) resulted $3,338$ features. The lasso-regularized regression model (Sec. 1) resulted in 142 risk factors which are positively predictive of unplanned rehospitalization following heart failure discharges.

Graph-based regularization (Sec. II-B) results in subgraphs being selected as a whole, as shown in Fig. 4. The question is how does it affect model performance and feature stability against data resampling?

A. Model Performance

The model performance was measured for different values of the lasso regularization term α and the Laplacian regularization term β. Table II reports other measures (sensitivity, specificity, precision, F-measure and AUC). Overall, the discriminative measures were not sensitive of the Laplacian factor β but depended critically on the lasso factor α. Fig. 5 displays the AUC in finer details for α. A good discrimination was achieved at $\alpha = .001$ and $\beta = .01$, where external validation resulted in an AUC of 0.66 (95%, CIs: $[0.6, 0.71]$). For the validation cohort, the Laplacian stabilized model was able to detect more true readmissions (sensitivity = 42.22 %) than lasso regularized model (sensitivity = 38.33 %). The overall classification accuracy for Laplacian stabilized model was 59.6% as opposed to 57.9 % for lasso regularized model.

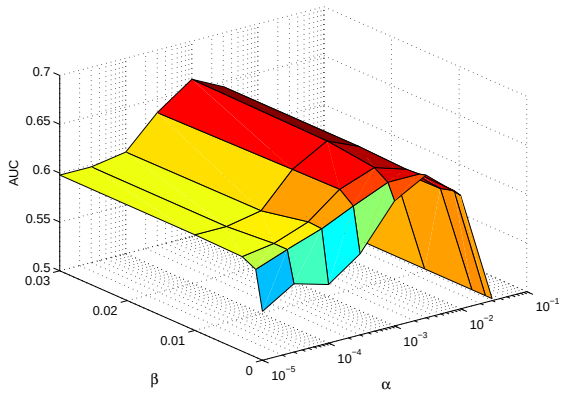


Figure 5: Effect of graph stabilized lasso regularization on area under the ROC curve (AUC). $\beta = 0$ reduces to the baseline lasso.

1) ROC Curve Analysis: The area under the ROC curve (AUC or c-statistic) can be used to compare different models fitted to the same data. As shown in Fig. 6, the application of Laplacian stabilization marginally improved the AUC over the lasso model. However a combination of elastic net and Laplacian was not able to improve the model discrimination.

2) Goodness-of-fit Statistics: We now compare the goodness-of-fit of models using Hosmer-Lemeshow (HL) test statistic. We divided our validation cohort into 10 groups defined by increasing order of estimated risk. Nine groups contained 37 observations, while one group contained 36. The expected frequencies in each group was more than five. Hence all conditions for reporting the HL test statistic was met [31]. Both Laplacian and combination of elastic net and Laplacian regularization resulted in small values of HL test statistic with $p > .05$ suggesting that these models fit the data quite well (see Table III).

Hyperparam.	Sens./Rec.	Spec.	Prec.	F-Meas.	AUC
$\alpha = \beta = 0$	0.49	0.59	0.54	0.51	0.54
$\alpha = .001$					
$\beta = .00$	0.41	0.79	0.62	0.51	0.62
$\beta = .01$	0.42	0.79	0.62	0.51	0.66
$\beta = .03$	0.44	0.76	0.66	0.53	0.66
$\alpha = .002$					
$\beta = 0.0$	0.49	0.73	0.66	0.55	0.65
$\beta = .01$	0.49	0.73	0.65	0.55	0.65
$\beta = .03$	0.48	0.72	0.62	0.54	0.64
$\alpha = .003$					
$\beta = 0.0$	0.46	0.76	0.64	0.54	0.62
$\beta = .01$	0.46	0.76	0.64	0.54	0.62
$\beta = .03$	0.45	0.75	0.63	0.53	0.62
$\alpha = .004$					
$\beta = 0.0$	0.44	0.77	0.66	0.53	0.63
$\beta = .01$	0.44	0.77	0.66	0.53	0.63
$\beta = .03$	0.43	0.78	0.65	0.52	0.63
$\alpha = .005$					
$\beta = 0$	0.46	0.81	0.69	0.55	0.63
$\beta = .01$	0.46	0.81	0.69	0.55	0.63
$\beta = .03$	0.45	0.82	0.69	0.55	0.63

Table II: The performance of model for various settings of lasso regularization term (α) and Laplacian regularization term (β) after model averaging from 50 bootstraps.

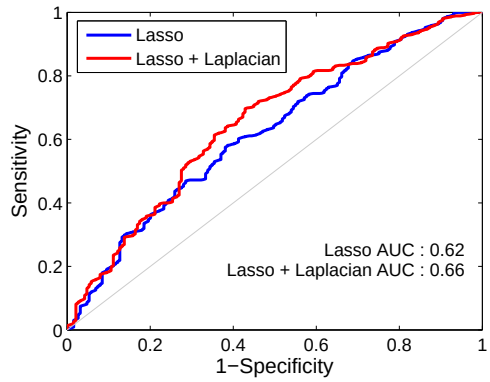


Figure 6: Comparing ROC plots.

B. Stability against Data Resampling

During this experiment, the lasso regularization term was fixed at $\alpha = .001$, corresponding to the value for maximum AUC of the model. Thus, feature stability through graph regularization is entirely controlled by the hyperparameter β in Eq. (3). The effect of β on feature stability is demonstrated in Fig. 7. Both Consistency Index and Jaccard Index confirmed improvements in feature stability with increasing graph penalty.

Next, we compared the stabilizing effect of regularization schemes. The feature graphs were applied not only for the lasso but also for the elastic net, thus creating four alternatives – lasso (baseline, no stabilizing), elastic net, Laplacian graph, and the combined elastic net + Laplacian graph. The hyperparameters were $\alpha = .001$, $\beta = .03$, and $\lambda = 0.1$ for elastic net.

Hosmer-Lemeshow test			
Model regularization	χ^2	**df**	**Significance**
Lasso	26.50	8	.0009
Lasso + Laplacian	7.23	8	.513
Elastic Net + Laplacian	6.25	8	.619

Table III: Measuring goodness-of-fit for logistic regression (df = degree of freedom). Small χ^2 values with large significance ($p > .05$) indicate better fit.

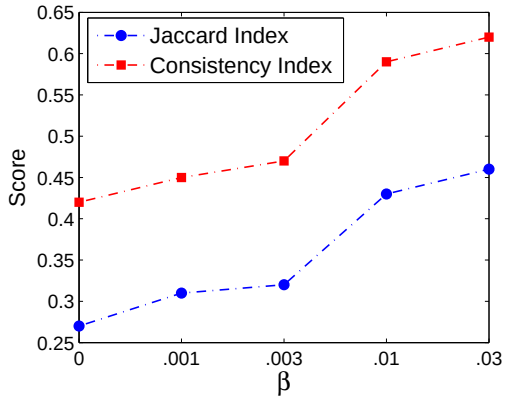
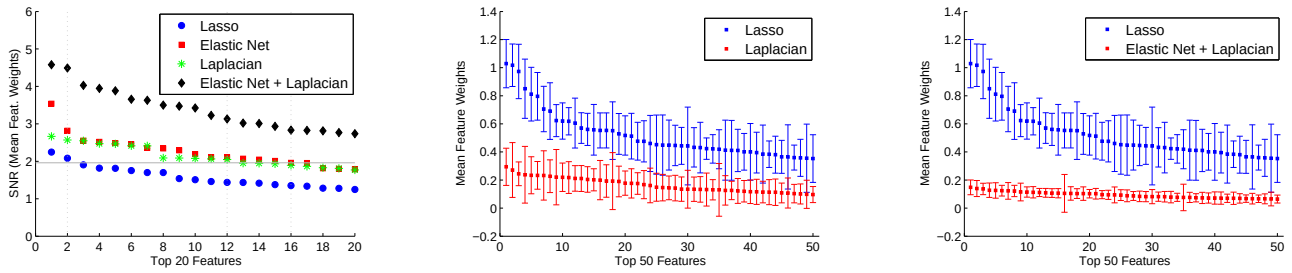


Figure 7: Effect of Laplacian regularization on feature stability for varying β , with $\alpha = .001$, and subset size of 100.

- For *model estimation stability*, the signal-to-noise ratios (SNR) of top individual feature weights are presented in Fig. 8a. Elastic net and Laplacian regularization both reduce weight variance significantly over the baseline lasso, and the Laplacian performs slightly better. With the combination of the elastic net and Laplacian, the effect is greatly amplified. At 95% CIs (approximately ± 1.96 std), lasso regularization identified 2 features, Laplacian identifies 12, elastic net 16 and the combination of Laplacian+elastic net regularization identified close to 50 features. Figs. 8b and 8c show a finer visual representation of the effect, clearly demonstrating the reduction in weight variance using the graph regularization.
- For *feature selection stability*, Consistency Index and Jaccard Index are reported in Fig. 9. Feature graph regularization consistently outperformed elastic net regularization for the top ranked features. Again, the combination of feature graph and elastic net resulted in the most stable set of features for all subset sizes.

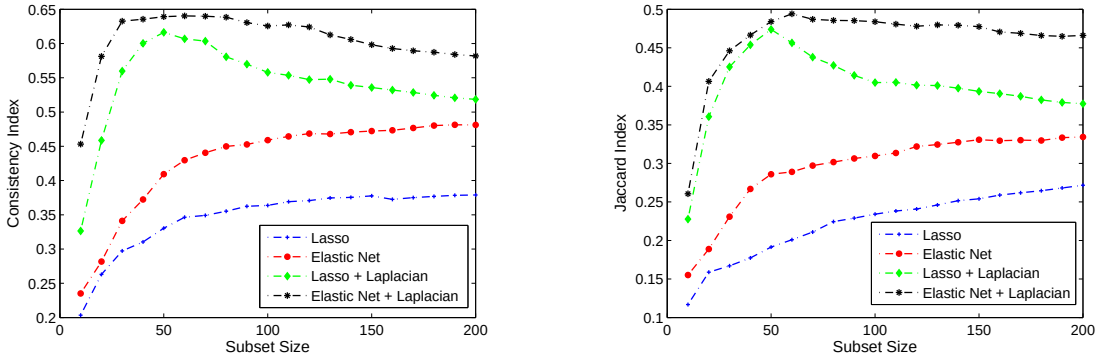
IV. Discussion and Conclusion

Although stability in feature selection is gaining importance [7], [16], [4], measuring the robustness of selected features in clinical prediction models has not been studied extensively. Feature stability facilitates reproducibility between model updates and generalization across medical studies. This is especially important in EMR-derived models due to its high-dimensional, dynamic and implementation-dependent nature.



(a) Signal-to-noise ratios. (b) Lasso versus Laplacian (c) Lasso versus elastic net + Laplacian.

Figure 8: Model estimation stability as measured by signal-to-noise ratios (SNR) of feature weights (Sub-fig. 8a). High value of SNR indicates more stability. Sub-figs. 8b and 8c elaborate the variances in more details.



(a) Consistency Index. (b) Jaccard Index.

Figure 9: Feature selection stability as measured by the Consistency Index (Fig. 9a) and Jaccard Index (Fig. 9b) for 6-month prediction. The plot compares the similarity in feature subsets generated by models with and without different stabilization under data variations. Larger indices imply more stability.

In practice, a stable model will allow the clinician to have more confidence on the selected features and their predictive importance.

In this paper, we have introduced feature graphs and Laplacian regularization to regression models to enhance stability in feature selection. Laplacian feature graphs have been used in bioinformatics [18], [19] to improve feature stability, but with important differences. First, lasso was not used as an embedded feature selection method. The work [19], for example, employs a filter-based method where the feature selection does not occur during learning of model parameters. Second, feature graphs were often constructed based on prior knowledge of interaction between features (e.g., genes). In our method, the model estimation is stabilized using a feature graph constructed from latent clinical structures in the training data. Our work stands unique in the following aspects (i) generic construction of feature graphs from commonly available attributes in the medical database (ii) extensive numerical validation of model stability in both model estimation and feature selection.

Our experiments confirm that the stability of a high dimensional linear clinical prediction model can be improved by using temporal and structural relations in EMR database. The combination of Laplacian regularization with existing state-of-the-art binary elastic net resulted in most stable features without hurting the model discrimination. Thus with Laplacian regularization, more features can be confidently selected for prediction (Sub-fig. 8a). This is useful in the EMR setting because each patient typically has limited number of active features despite the huge number of features across the database. Having more confident features would make explanation for individual prediction easier.

With regards to performance, Laplacian regularization along with binary elastic net resulted in a model with a better fit against the validation cohort (as per Table III). The marginal increase in sensitivity and classification accuracy in Laplacian regularization can be attributed to grouping of correlated features.

With regards to feature stability, the improvement upon the elastic net demonstrates that feature graph is complementary to ridge regression. This could be explained by the fact that while ridge regression tends to encourage all weights to be similar and regressed toward zero, graph regularization only requires pairwise smoothness.

Our EMR-derived model achieved a discriminatory capacity (AUC = 0.66 for 6 months) comparable with or better than existing prediction models for rehospitalization following heart failure discharges [13], [14]. The model is derived from free

available administrative and medical data, making it readily implementable into existing EMR systems. Interestingly, the top predictors discovered by our model are consistent with the existing clinical studies. Our model ranked male gender highest on the importance scale [32], [33], [34]. Looking at the medical factors, the strong predictors include prior history of hospitalization (past emergencies, past emergency attend time), which are consistent with those in [32], [35], [33], [36], [34]. The comorbidities observed were occurrence of coagulopathy in the past year and occurrence of complicated diabetes in the past three months. Other major predictors for heart failure rehospitalization are heart failure [32], [33], [35], [36], lipoprotein metabolism disorders, angina pectoris, cataract, and chronic ischaemic heart diseases. Past number of procedures in a period of 3 months to 2 years was also ranked high.

The discrimination power, the automatic feature selection and stability control capacity suggest that the model can be used as a fast and inexpensive screening tool to select patients and risk factors for more in-depth clinical investigation. For example, through selected feature subgraphs, related risk factors can be collapsed to achieve more generality. It could serve as a first step in bridging the translational gap between bench and bedside [34] . We wish to emphasize that the entire prediction process is transparent as the model is capable of explaining what risk factors are involved in a risk estimate.

A. Study Limitations

We acknowledge the following limitations in our study. First, since our main focus was on stabilizing a high dimensional model, we did not concentrate on improving the accuracy. In our experiments, graph regularization contributed very little to improving model discrimination. Second, we did not investigate more complex relationship between variables in EMR data when building feature graphs. It is possible that exploiting structures like billing codes and lab tests may further enhance sharing of statistical strength between correlated features. Third, the model evaluation was not tested independently by other researchers. However, we have used temporal validation on unique patients, and it matches the common practice of learning models using past patients and predicting outcomes for future patient. Fourth, clinical measurements had a high degree of missingness, and hence were discarded. In review of the these limitations, we believe our derived model is conservative and may have underestimated the AUC of the validation cohort.

B. Conclusion

In this study, we tackle the seldom studied but notorious problem of feature instability in clinical prediction models. Stable model features translate to proper understanding of risk factors, and hence better confidence in prognosis. Our approach consists of a novel technique to mitigate the problem by utilizing feature graphs that link similar conditions/interventions and the same condition/intervention over multiple time periods. Our extensive experiments in predicting 6-month readmission in a heart failure cohort confirm that the application of feature graphs increases the stability of the selected feature subset and reduces the variation in feature weights. The performance of the readmission models derived from administrative hospital data is competitive against existing models developed on clinical data. Further, since our approach is based on commonly available administrative attributes, models can be readily implemented on top of existing EMR systems and portable across cohorts and institutions using similar EMR databases. We believe our stabilizing framework provides the first proof of concept in utilizing feature graphs in clinical setting and numerically validating stability for a clinical prediction model. Future work includes applying the same technique for a variety of cohorts and sites and prospective evaluation in practice.

REFERENCES

[1] B. Yu, "Stability," *Bernoulli*, vol. 19, no. 4, pp. 1484–1500, 2013.

[2] Y. Saeys, T. Abeel, and Y. Van de Peer, "Robust feature selection using ensemble feature selection techniques," in *Machine learning and knowledge discovery in databases*, pp. 313–325, Springer, 2008.

[3] W. Awada, T. M. Khoshgoftaar, D. Dittman, R. Wald, and A. Napolitano, "A review of the stability of feature selection techniques for bioinformatics data," in *Information Reuse and Integration (IRI), 2012 IEEE 13th International Conference on*, pp. 356–363, IEEE, 2012.

[4] T. M. Khoshgoftaar, A. Fazelpour, H. Wang, and R. Wald, "A survey of stability analysis of feature subset selection techniques," in *Information Reuse and Integration (IRI), 2013 IEEE 14th International Conference on*, pp. 424–431, IEEE, 2013.

[5] D. He, S. C. Mathews, A. N. Kalloo, and S. Hutfless, "Mining high-dimensional administrative claims data to predict early hospital readmissions," *Journal of the American Medical Informatics Association*, pp. amiajnl–2013, 2013.

[6] J. Ye and J. Liu, "Sparse methods for biomedical data," *ACM SIGKDD Explorations Newsletter*, vol. 14, no. 1, pp. 4–15, 2012.

[7] P. C. Austin and J. V. Tu, "Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality," *Journal of clinical epidemiology*, vol. 57, no. 11, pp. 1138–1146, 2004.

[8] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[9] H. Xu, C. Caramanis, and S. Mannor, "Sparse algorithms are not stable: A no-free-lunch theorem," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 1, pp. 187–193, 2012.

[10] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[11] R. Real and J. M. Vargas, "The probabilistic basis of Jaccard's index of similarity," *Systematic biology*, vol. 45, no. 3, pp. 380–385, 1996.

[12] L. I. Kuncheva, "A stability index for feature selection.," in *Artificial Intelligence and Applications*, pp. 421–427, 2007.

[13] S. B. and et al Ross JS, Mulvey GK, "Statistical models and patient predictors of readmission for heart failure: A systematic review," *Archives of Internal Medicine*, vol. 168, no. 13, pp. 1371–1386, 2008.

[14] V. Betihavas, P. M. Davidson, P. J. Newton, S. a. Frost, P. S. Macdonald, and S. Stewart, "What are the factors in risk prediction models for rehospitalisation for adults with chronic heart failure?," *Australian critical care : official journal of the Confederation of Australian Critical Care Nurses*, vol. 25, pp. 31–40, Feb. 2012.

[15] P. Křížek, J. Kittler, and V. Hlaváč, "Improving stability of feature selection methods," in *Computer Analysis of Images and Patterns*, pp. 929–936, Springer, 2007.

[16] A. Kalousis, J. Prados, and M. Hilario, "Stability of feature selection algorithms: a study on high-dimensional spaces," *Knowledge and information systems*, vol. 12, no. 1, pp. 95–116, 2007.

[17] Z. He and W. Yu, "Stable feature selection for biomarker discovery," *Computational biology and chemistry*, vol. 34, no. 4, pp. 215–225, 2010.

[18] C. Li and H. Li, "Network-constrained regularization and variable selection for analysis of genomic data," *Bioinformatics*, vol. 24, no. 9, pp. 1175–1182, 2008.

[19] Y. Cun and H. Fröhlich, "Network and data integration for biomarker signature discovery via network smoothed t-statistics," *PloS one*, vol. 8, no. 9, p. e73074, 2013.

[20] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[21] A. Y. Ng, "Feature selection, l1 vs. l2 regularization, and rotational invariance," in *Proceedings of the twenty-first international conference on Machine learning*, p. 78, ACM, 2004.

[22] P. McCullagh and J. Nelder, *Generalized linear models*. Chapman & HallCRC, 1989.

[23] N. Meinshausen and P. Bühlmann, "Stability selection," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 72, no. 4, pp. 417–473, 2010.

[24] F. R. Chung, *Spectral graph theory*, vol. 92. AMS Bookstore, 1997.

[25] T. Tran, D. Phung, W. Luo, R. Harvey, M. Berk, and S. Venkatesh, "An integrated framework for suicide risk prediction," in *Proc. of ACM Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD)*, 2013.

[26] D. G. Altman, Y. Vergouwe, P. Royston, and K. G. Moons, "Prognosis and prognostic research: validating a prognostic model," *BMJ: British Medical Journal*, vol. 338, no. 7708, pp. 1432–1435, 2009.

[27] Z. Birnbaum, "On a use of the Mann-Whitney statistic," tech. rep., DTIC Document, 1955.

[28] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*. Wiley. com, 2013.

[29] F. C. Pampel, *Logistic regression: A primer*, vol. 132. Sage, 2000.

[30] J. H. Friedman and B. E. Popescu, "Predictive learning via rule ensembles," *The Annals of Applied Statistics*, pp. 916–954, 2008.

[31] C.-Y. J. Peng, K. L. Lee, and G. M. Ingersoll, "An introduction to logistic regression analysis and reporting," *The Journal of Educational Research*, vol. 96, no. 1, pp. 3–14, 2002.

[32] M. Chin, H. Marshall, M. Goldman, *et al.*, "Correlates of early hospital readmission or death in patients with congestive heart failure," *The American journal of cardiology*, vol. 79, no. 12, pp. 1640–1644, 1997.

[33] H. M. Krumholz, E. M. Parent, N. Tu, V. Vaccarino, Y. Wang, M. J. Radford, and J. Hennen, "Readmission after hospitalization for congestive heart failure among medicare beneficiaries," *Archives of Internal Medicine*, vol. 157, no. 1, p. 99, 1997.

[34] R. Amarasingham, B. J. Moore, Y. P. Tabak, M. H. Drazner, C. A. Clark, S. Zhang, W. G. Reed, T. S. Swanson, Y. Ma, and E. A. Halm, "An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data," *Medical care*, vol. 48, no. 11, pp. 981–988, 2010.

[35] H. M. Krumholz, Y.-T. Chen, Y. Wang, V. Vaccarino, M. J. Radford, and R. I. Horwitz, "Predictors of readmission among elderly survivors of admission with heart failure," *American heart journal*, vol. 139, no. 1, pp. 72–77, 2000.

[36] G. M. Felker, J. D. Leimberger, R. M. Califf, M. S. Cuffe, B. M. Massie, K. F. Adams Jr, M. Gheorghiade, and C. M. O'Connor, "Risk stratification after hospitalization for decompensated heart failure," *Journal of cardiac failure*, vol. 10, no. 6, pp. 460–466, 2004.