Title

Efficient and sparse feature selection for biomedical text classification via the elastic net: Application to ICU risk stratification from nursing notes

Authors

Marafino, Ben J
Boscardin, W John
Dudley, R Adams

Publication Date

2015-04-01

DOI

10.1016/j.jbi.2015.02.003

Peer reviewed

Journal of Biomedical Informatics 54 (2015) 114–120

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Efficient and sparse feature selection for biomedical text classification via the elastic net: Application to ICU risk stratification from nursing notes

Ben J. Marafino [a,b,*], W. John Boscardin [c,d], R. Adams Dudley [a,b,c,d]

[a] Philip R. Lee Institute for Health Policy Studies, School of Medicine, University of California, San Francisco, United States
[b] Center for Healthcare Value, University of California, San Francisco, United States
[c] Department of Epidemiology and Biostatistics, University of California, San Francisco, United States
[d] Department of Medicine, University of California, San Francisco, United States

ARTICLE INFO

Article history:
Received 9 October 2014
Accepted 8 February 2015
Available online 17 February 2015

Keywords:
Text mining
Feature selection
Elastic net
ICU
Risk stratification
Machine learning

ABSTRACT

Background and significance: Sparsity is often a desirable property of statistical models, and various feature selection methods exist so as to yield sparser and interpretable models. However, their application to biomedical text classification, particularly to mortality risk stratification among intensive care unit (ICU) patients, has not been thoroughly studied.
Objective: To develop and characterize sparse classifiers based on the free text of nursing notes in order to predict ICU mortality risk and to discover text features most strongly associated with mortality.
Methods: We selected nursing notes from the first 24 h of ICU admission for 25,826 adult ICU patients from the MIMIC-II database. We then developed a pair of stochastic gradient descent-based classifiers with elastic-net regularization. We also studied the performance-sparsity tradeoffs of both classifiers as their regularization parameters were varied.
Results: The best-performing classifier achieved a 10-fold cross-validated AUC of 0.897 under the log loss function and full L_2 regularization, while full L_1 regularization used just 0.00025% of candidate input features and resulted in an AUC of 0.889. Using the log loss (range of AUCs 0.889–0.897) yielded better performance compared to the hinge loss (0.850–0.876), but the latter yielded even sparser models.
Discussion: Most features selected by both classifiers appear clinically relevant and correspond to predictors already present in existing ICU mortality models. The sparser classifiers were also able to discover a number of informative – albeit nonclinical – features.
Conclusion: The elastic-net-regularized classifiers perform reasonably well and are capable of reducing the number of features required by over a thousandfold, with only a modest impact on performance.

1. Background and significance

Feature selection methods have recently been growing in importance within the fields of genomics, bioinformatics, and computational biology, where they have found wide utility in problems ranging from microarray DNA analysis to genome-wide association studies, among others [1–3]. During the course of microarray DNA analysis, for example, one common objective is to classify tumor samples from patients with cancer, based on the gene expression profiles of those samples. However, the number of genes under consideration is almost always much larger than the number of tumor samples, with only a small subset of these genes being putatively associated with the tumor classes. This problem hence exemplifies the so-called "$p \gg n$" setting [4,5], where one is faced with many more candidate features p than examples n. Thus, an ideal classifier for this problem setting should not only be accurate and exhibit other good performance characteristics – it should be able to select only those genes that make up the pathways present in the cancer of interest, easing interpretation of the resulting predictions, while also neglecting genes that are not discriminative of the outcome.

With this in mind, many parallels can be drawn between the example of microarray DNA analysis given above and the usual setting of biomedical text classification, where the goal may be to predict outcomes (i.e., mortality) from text or to perform information extraction [6], such as ongoing smoking status [7], or receiving a

* Corresponding author at: Philip R. Lee Institute for Health Policy Studies, School of Medicine, University of California, San Francisco, 3333 California Street, Suite 265, San Francisco, CA 94118, United States. Fax: +1 415 476 0705.
E-mail address: ben.marafino@ucsf.edu (B.J. Marafino).

http://dx.doi.org/10.1016/j.jbi.2015.02.003
1532-0464/

procedure such as mechanical ventilation in the ICU [8]. In these setting, nearly all input features derived from the underlying text are noisy in the sense that they carry little information about the outcome or clinical entity of interest, and therefore are not discriminative. The problem is also further complicated by the fact that the dimensionality of the input feature space often proves large – and can be increased even further by extracting bigram or higher-order *n*-gram features, or via other, more sophisticated methods of feature extraction.

Very generally, feature selection algorithms for linear models, including logistic regression and support vector machines (SVM) can be classified as follows: a method may carry out explicit feature selection by setting some feature weights or parameter estimates zero based on a set of criteria (which vary with the algorithm used). On the other hand, an algorithm may instead perform *shrinkage*, where the feature weights are smoothly shrunk toward zero while never being made exactly zero (ridge penalty), or *implicit* variable selection via a shrinkage process that allows for making at least some weights exactly zero (lasso), or while also performing ridge-like shrinkage (elastic net). In the cases of the ridge, lasso, and elastic net penalties, the shrinkage and selection effects are enforced by a constraint on the feature weights, or a *regularization penalty* that constrains how large the weights can be while they are being estimated. The lasso penalizes the sum of the absolute values of the weights (L_1 norm), while the ridge penalizes the square root of the sum of the squared weights (L_2 norm), and the elastic net combines these penalties into a linear combination of the L_1 and L_2 norms.

Two examples of explicit feature selection are stepwise regression, and exhaustive best-subset methods, which have been widely applied within the biostatistical and epidemiological literature [9], but have enjoyed somewhat less application elsewhere, particularly to high-dimensional learning and other, related contexts [10]. One reason for this is that exhaustive best-subset methods suffer from the limitation that with p candidate input features, the algorithm must train $2^p - 1$ classifiers (less the null classifier). While forward and backward stagewise methods have worst-case $O(p^2)$ complexity, they are not guaranteed to select the best possible permutation of input features [11,12]. In either case, when p is in the hundreds of thousands to millions of features, as is usually the case when dealing with text-based problems, these approaches quickly become computationally infeasible. In addition, stepwise regression performs poorly when features are correlated; in practice, it does not exhibit a grouping effect, a desirable property of a feature selection method, where correlated features tend to be included together in a final model [13].

Examples of shrinkage-based methods for linear models include ridge regression [14] and the lasso [15]. Prior to the development of the lasso, ridge regression enjoyed preeminence among shrinkage methods, in part because the ridge penalty, represented as the L_2 norm of the vector parameter estimates, resembles the usual ordinary least squares objective and thus is easier to optimize [16]. In contrast, optimizing those non-smooth objective functions that include the lasso penalty requires specialized algorithms, e.g. least-angle regression (LARS) [17]. Moreover, ridge regression has generally been found to outperform lasso and exhibit a grouping effect when $p < n$; it is only when p grows larger than n that the performance of ridge regression begins to degrade and it no longer handles correlated features well [4]. However, the lasso penalty enforces automatic feature selection by forcing at least some features to be zero, as opposed to ridge regression, where only shrinkage is performed. Nevertheless, the use of the lasso proves problematic when at least some features are highly correlated. In this case, the lasso will select from among these features at random. Moreover, given n training examples, the lasso is capable of selecting only at most n features [13].

The elastic net, in contrast, represents a compromise between the ridge and lasso penalties [13]. Indeed, the elastic net penalty is simply written as a linear combination of these two penalties: the lasso penalty term acts to encourage sparsity in the parameter estimates of the resulting model, while the ridge term acts to "average out" the parameter estimates of correlated features, which imposes a grouping effect [4,18]. Hence, the elastic net performs both shrinkage (although milder than that obtained via ridge regression) and automatic feature selection. Depending on the preferences of the user and the properties of the underlying problem, the elastic net penalty can be smoothly adjusted so as to give more weight to either the lasso or ridge penalties. Compared to the lasso, the elastic net is able to yield a model including more features p than training examples n, but with possibly far fewer than would be selected via the ridge penalty alone (depending on the parameter settings chosen), which, taken with the fact that the elastic net exhibits a grouping effect [13], represents a clear advantage over either method.

Despite holding promise for feature selection and model development within the usual problem setting of biomedical text classification, the elastic net has yet to be applied toward these problems in clinical NLP. In particular, the elastic net allows those relevant word or *n*-gram features associated with the outcome to be discovered far more easily, and hence has the potential to supersede existing "black-box" approaches [8,19,20], e.g., unregularized SVMs, which are widely used in clinical NLP Furthermore, it is also possible that regularization techniques applied to classifiers could be used to validate and improve on what we term the "expert input" approach [7,21], where potentially relevant word or *n*-gram features are manually selected and extracted before a classifier is trained. In particular, Walsh and Hripcsak's recent work [21] constitutes an example of the use of "expert input": while they developed a series of classifiers with the aid of the lasso, using a combination of free text and a series of clinical features, these text features were manually chosen in advance based on their potential association with readmission, before combining them with other features with which to train the classifier.

Here, we describe the application of elastic net regularization to a pair of classifiers developed to predict mortality risk among adult ICU patients based on the free text of their first 24 h of nursing notes, and we report a sample of the relevant features discovered by these classifiers. (A full list of the features of one such classifier, along with their coefficients, is included in the Supplementary Information.) Nursing notes constitute a good candidate source of information for mortality risk prediction, as they contain a detailed and regularly-updated record of the interventions performed, medications administered, vital signs, and physical examination findings, all of which carry highly specific information about the patient's dynamic physiological state and eventual outcome. We then characterize what we term the *sparsity-performance tradeoff* of both classifiers as the elastic net regularization parameter is varied. We also report examples of informative features found by the classifier, and compare them to what is currently known of predictors of ICU mortality. Finally, we also compare the performance of our models to an existing method of ICU risk stratification based on the Simplified Acute Physiology Score (SAPS) and validated on the same dataset.

2. Methods

The nursing notes were derived from the Multiparameter Intelligent Monitoring in Intensive Care-II (MIMIC-II) database, version 2.6. The MIMIC-II database contains complete sets of clinical free text notes from roughly 40,000 ICU stays for nearly 33,000 patients at Beth Israel Deaconess Medical Center (BIDMC) in Boston,

Massachusetts, dating from between 2001 and 2008 [22,23]. For each adult ICU patient, we selected and combined all nursing notes dated within 24 h of the first recorded ICU admission time. On this basis, 25,826 adult patients and their notes were selected, of which 2099 died in the hospital prior to discharge at any point following their ICU admission; mortality was determined via the ICUSTAY_EXPIRE_FLG variable.

Next, each set of notes for a single patient (i.e., each document) was processed in order to extract unique unigram and bigram features. We have previously reported [8] that extracting 3-grams or higher-order n-grams (i.e. setting n greater than 2) does not improve classifier performance. Moreover, extracting these higher-order n-grams drastically increases the time needed for feature extraction and classifier training. We removed numbers, punctuation, and neutral stop-words (e.g. 'and', 'the') and performed stemming on each word. The feature counts – i.e., the counts of each unigram and bigram – for each document were then extracted and mapped to their term frequency-inverse document frequency (tf-idf) values, i.e., the count of a feature in that document, divided by the number of total notesets in which it appears. Essentially, each document was transformed into a vector of tf-idf values of features, which were then used for classification.

We then trained, by way of stochastic gradient descent (SGD) [24], two different classifiers – essentially logistic regression and a linear support vector machine – which are distinguished only by their loss functions. The loss function defining logistic regression, with respect to a single training example y_i is given by $L(\beta, y_i) = \log(1 + \exp(-y_i \beta^T x))$, or the *log loss*, while the loss function giving the linear SVM is $L(\beta, y_i) = \max(0, 1 - y_i \beta^T x)$, which is also commonly known as the *hinge loss*. In both cases, the objective function J to be minimized took the form

$$J(\beta) = \alpha(\lambda_1 |\beta|_1 + \lambda_2 |\beta|_2^2) + \sum_{\text{all } y_i} L(\beta, y_i)$$

where α gives the overall regularization strength, the regularization terms $|\beta|_1$ and $|\beta|_2^2$ correspond to the L_1 and L_2 norms of the vector of weights or parameter estimates β, respectively, and the rightmost cost term gives the average loss by summing over all training examples y_i. For brevity, we express both elastic net penalty hyperparameters λ_1 and λ_2 in terms of one hyperparameter, $\lambda = \lambda_1/(\lambda_1 + \lambda_2)$, which represents the ratio of L_1 to L_2 regularization strength imposed on the classifier, and ranges from 0 to 1. Setting $\lambda = 1$ results in only L_1 regularization, while $\lambda = 0$ puts full weight on the L_2 penalty term. However, values of λ between 0 and 1 correspond to a linear combination of the two penalty terms. We interpret relative feature influence as proportional to the absolute values of the parameter estimates $\hat{\beta}_i$. We also considered a feature *selected* if its parameter estimate or weight learnt by a classifier was nonzero (no matter how small).

We trained all classifiers via SGD with constant learning rate $\eta = 1.4$. The performance metric used was the area under the ROC curve (AUC), and we performed 10-fold nested cross-validation to estimate AUC for different values of λ [25]. The optimal values of λ and η for each classifier were also determined by 10-fold cross-validation over a reasonable parameter grid of values for λ and η. We used a constant value of $\alpha = 10^{-5}$ in all of our experiments.

In order to better characterize each classifier's ability to induce sparsity, we also made use of an additional sparsity measure, defined as follows. Denote the sets of all unigram and bigram features having nonzero coefficients as U and B, respectively. Then the quantity $M(\lambda)$ defined as

$$M(\lambda) = \frac{|\text{elements of } B \text{ containing an unigram from } U|}{|B|}$$

where the operation $|A|$ denotes the cardinality – i.e., the number of elements – of a set A. As the input feature space contains both unigrams and bigrams, it is possible – and often the case – that bigram features selected by a sparse classifier contain unigrams already included in the model, which would, in a sense, render those features less informative. For example, a classifier could preferentially select the bigram "metabolic acidosis" over just "acidosis", which could also refer to respiratory acidosis, an unrelated condition, and "metabolic", a unigram that would likely not be selected, as the term is not discriminative of the outcome by itself.

On the other hand, it is possible that a differently trained classifier would select both "metabolic acidosis" and "acidosis", which could complicate interpretation of the resulting predictions. Ideally, as metabolic and respiratory acidosis denote the only types of acidosis, the classifier should select only "metabolic acidosis" and "respiratory acidosis", and not just "acidosis" or "metabolic" or "respiratory". Therefore, the ideal sparse classifier would select only those bigrams that carry maximal marginal information content, i.e., those that carry information not contained in the unigrams already selected; thus, we hypothesize that $M(\lambda)$ denotes, though somewhat crudely, how sparse the feature space of selected *bigrams* is. If $M(\lambda)$ is zero, then all the bigrams selected by the classifier very likely do not contain information already present in the unigrams already selected; conversely, if $M(\lambda) = 1$, then all bigrams are redundant, as they repeat one or at most two unigram feature(s).

We used McNemar's test for paired nominal data to compare classifiers. Furthermore, in order to compare our model to existing methodology, we also made use of a logistic regression model utilizing the Simplified Acute Physiology Score, version 1 (SAPS-I) [26]. The SAPS model includes as its components laboratory values, ventilator settings, age, Glasgow Coma Score (GCS), among others. The SAPS model used the patient's highest recorded value of SAPS within the first 24 h of their ICU stay. Modified Hosmer–Lemeshow goodness-of-fit tests [27] were used to assess calibration for all classifiers, i.e., how well a classifier's predicted probabilities of mortality agreed with the actual probabilities by decile of risk.

The Committee on Human Research of the University of California, San Francisco deemed this study exempt from review.

3. Results

A total of 101,806 nursing notes were selected for 25,826 patients within the first 24 h following ICU admission. Following processing, the mean length of the combined 24 h of nursing notes for each patient was 468 words (median 334, interquartile range 249–466). Extracted were 1,842,522 candidate input features, of which 91,317 were unigrams and 1,751,205 were bigrams. On this dataset, the SAPS-based logistic regression classifier achieved an AUC of 0.791.

The sparsity-performance tradeoff curves of both text-based classifiers are presented in Fig. 1. As expected, the SGD classifier with log loss selected all features at $\lambda = 0$, resulting in an AUC of 0.897. At $\lambda = 1$, just 465 features were selected, which represents just 0.00025% of the number of candidate input features, or almost a ten-thousandfold reduction, and the AUC for this classifier was 0.889. Under the log loss, intermediate values of lambda resulted in a smooth and robust tradeoff between sparsity and performance.

Interestingly, the same classifier equipped with the hinge loss selected just over half (970,319 out of 1,842,522, or 53%) of the candidate feature space at $\lambda = 0$; the resultant AUC with these features was 0.850. At $\lambda = 1.0$, 345 features were selected, which proved fewer than the log-loss classifier at the same value of λ, and resulted in an AUC of 0.876. Compared to the log loss, the

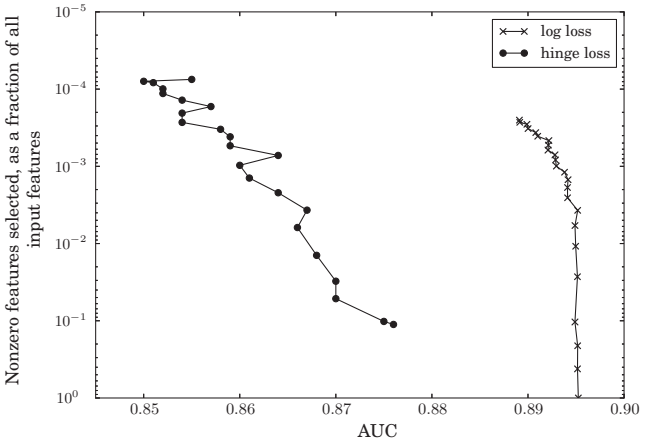

Fig. 1. The sparsity-performance tradeoffs curves of both classifiers tested as λ is varied from 0 to 1. Note that the y-axis, which denotes the fraction of nonzero features selected by either classifier for a given value of λ, has been scaled logarithmically. The optimal frontier lies in the upper right, where sparsity and performance are both maximized. Figure adapted from [28].

hinge loss also appeared less robust to changes in the size of the selected feature space. The log loss also displayed a more optimal (i.e., shallower) sparsity-performance tradeoff compared to the hinge loss (Fig. 1.) The differences in AUC between the two classifiers were significant at each value of λ by McNemar's test. A non-significant Hosmer–Lemeshow statistic was obtained for all classifiers for each value of λ studied at the 0.05 confidence level, indicating adequate calibration.

The empirical cumulative distribution functions of nonzero features of the classifier under log loss are presented in Fig. 2 for some choices of λ, as well as for the unregularized classifier. The shrinkage effect of the ridge regularizer clearly dominates at smaller values of λ; at $\lambda = 0$, compared to the unregularized classifier, the effect is fairly strong, with roughly 95% of nonzero features having parameter estimates lying in the interval $[-0.1, 0.1]$. In contrast, the lasso penalty obtained at $\lambda = 1$ tends to result in a wider range of parameter estimates for the features selected, which eases their interpretation. The effects of the lasso regularizer also clearly

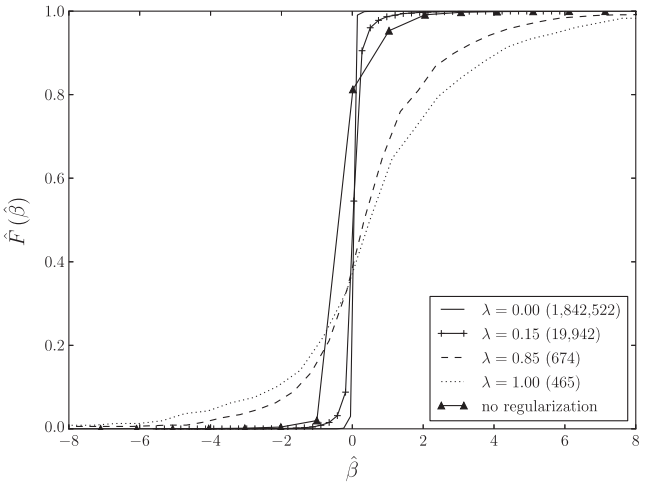

Fig. 2. Plot of empirical cumulative distribution functions (ECDFs) $\hat{F}(\hat{\beta})$ of *nonzero* coefficients $\hat{\beta}$ for classifiers equipped with the log loss, for different values of λ. For $\lambda = 0$ and $\lambda = 0.15$, the shrinkage effect of L_2 regularization can clearly be seen compared to the case of no regularization (line with triangles). At $\lambda = 1$, the lasso-equivalent solution is obtained, which is also the sparsest, as expected. The number in parentheses denotes the number of features selected, or the number of nonzero features of the vector of parameter estimates $\hat{\beta}$.

dominate those of the ridge at $\lambda = 0.85$; while some shrinkage is observed, the classifier selects roughly the same number of features (in terms of orders of magnitude) as it does at $\lambda = 1$.

A plot of $M(\lambda)$ versus λ is depicted in Fig. 3. The classifier under hinge loss displays markedly lower values of $M(\lambda)$ for all values of λ, compared to the log loss. This effect was most marked for $\lambda = 1.0$. However, the increased diversity of bigram features selected by the hinge classifier did not lead to any improvements in performance. We also present a selection of the features selected by a log-loss classifier with $\lambda = 0.85$ in Table 1.

4. Discussion

All classifiers under both loss functions yielded reasonable results, with AUCs ranging from 0.85 to 0.90, and compared well to the performance of a logistic model based on SAPS, which achieved an AUC of 0.791. The best-performing classifier yielded an AUC of 0.897 when furnished with the log loss. These results compare well with some past studies; in [29], physicians were able to achieve roughly similar prediction performance only via manual chart review, and in [30], where Lehman et al. utilized topic models of text [31] (a dimensionality reduction technique) to stratify risk among patients from the same dataset, the resulting AUC was 0.78. Equipping the classifier with a hinge loss did result in somewhat more sparse models (Fig. 1) compared to the log loss. For both classifiers, the tradeoffs inherent in adjusting λ appear favorable and suggest that a substantially more sparse and interpretable model can be achieved with only a modest performance cost.

The classifier equipped with the hinge loss function exhibited two interesting behaviors related to the sparsity of the resulting models. First, with full L_2 regularization, only 970,319 features are selected out of a possible 1,842,522 (53%), so at $\lambda = 0$, the feature space is already more sparse compared to the log-loss classifier, which selects all features. Second, as measured by $M(\lambda)$, a larger proportion of the bigram features it selects appear to be "informative", compared to the classifier with log loss, and this effect holds for all values of λ (Fig. 3), although this did not translate into improved predictive performance (Fig. 1.) These results are consistent with other studies into the hinge loss, e.g. in [32]. The precise reasons for these behaviors are unclear, but seem to be related to the behavior of the hinge loss function and its derivative, which both act to encourage sparsity.

Among the more influential features uncovered by the sparser classifiers (with larger values of λ, Table) include those physical examination signs associated with neurological status, specifically

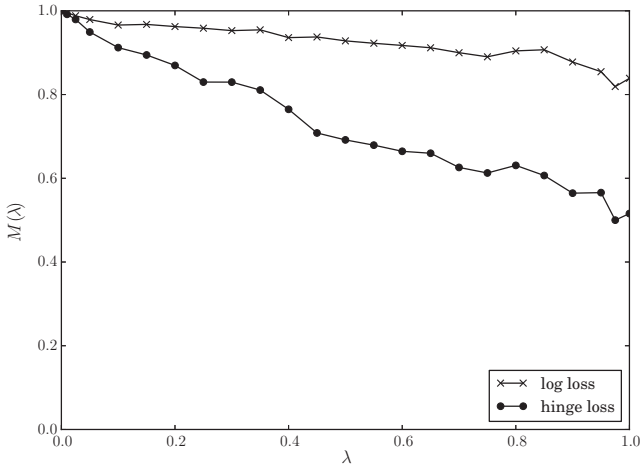

Fig. 3. Plot of $M(\lambda)$ for both classifiers as λ varies from 0 to 1.

Table 1
A selection of destemmed features selected by the sparse elastic net-regularized classifier with $\lambda = 0.85$. We reviewed and assigned a subset of these features with their associated weights to the ad-hoc categories above. The full list of features is given in the Supplementary Information. Abbreviations and terms: GCS, Glasgow Coma Score; MAE, moves all extremities; CMV, continuous mandatory ventilation; SIMV, synchronized intermittent mandatory ventilation; CPAP, continuous positive airway pressure; a-fib, atrial fibrillation; OOB, out of bed; FFP, fresh frozen plasma, used to correct coagulopathies; mannitol, an osmotic diuretic used to lower intracranial pressure; duoderm, a type of wound dressing often applied to decubitus or pressure ulcers (bed sores); lactulose, a laxative; zosyn, trade name of a common broad-spectrum antibiotic used in the treatment of, for example, pneumonia.

Category	Feature	Weight
GCS & neurological status	Posturing	7.39
	Unresponsive	6.39
	Corneal (reflexes)	5.65
	Non-reactive	2.27
	Follows commands	−1.72
	Pleasant	−3.39
	Denies (pain)	−4.23
	MAE	−5.81
Ventilation status & modes of ventilation	Intubated	9.41
	CMV	1.02
	SIMV	−1.42
	CPAP	−4.63
	Wean	−8.94
	Extubated	−13.7
Hemodynamic instability	Levophed	5.14
	Pressors	4.78
	Vasopressin	4.05
	Hypotensive	3.27
	Dopamine	2.67
	Hemodynamically stable	−1.97
Other overall indicators of prognosis	a-fib	5.22
	(cardiac) arrest	2.31
	Septic	0.90
	Metabolic acidosis	0.87
	Incision	−0.41
	OOB	−3.14
Physical exam findings	Jaundice (d)	4.84
	Ascites	4.61
	Mottled (a sign of poor tissue perfusion, seen in, e.g., septic shock)	4.40
	Tachypnea	3.39
	Anasarca	1.75
	Flatus	−0.43
Indicators of comorbidities not explicitly documented in notes	FFP	5.14
	Mannitol	3.66
	Duoderm	2.34
	Lactulose	2.27
	Zosyn	1.25
Presence of family members at bedside (and possible proxy for patient age)	Daughter	2.47
	Son	2.42
	Wife	0.78
	Parents	−0.09
	Father	−1.09
	Mother	−1.45

level of consciousness and motor function, and which overlap with certain components of the Glasgow Coma Score (Table). Furthermore, the classifiers also selected features that serve as indicators of poor prognosis, such as those related to hemodynamic instability and the use of vasopressors, which also form a major component of the SOFA risk scoring system [33] commonly used in ICUs.

The management of mechanical ventilation is crucial to ICU outcomes, and this was also reflected in our models (Table). Even the sparsest models retained those features serving as indicators of ventilation status, as well as those marking the progression of weaning a patient from ventilation, to the point where they are extubated. While the precise protocols differ among ICUs [34], the process of ventilator weaning usually progresses in stages. When patients are first ventilated, they are often so extremely ill that they cannot cooperate and cannot be slowly put to sleep because they could not survive the slow respiratory rate created. Therefore, they need to be paralyzed briefly and put on a form of ventilation called controlled mechanical ventilation (CMV), in which all the work of breathing is done by the ventilator. As their respiratory function improves, they transition to synchronized intermittent mandatory ventilation (SIMV) and often then on to continuous positive airway pressure (CPAP) ventilation, following which the patient may then finally be capable of breathing on their own and is extubated [35]. Furthermore, since CPAP can also be applied without intubation, its presence in a note could also denote a trial of externally applied CPAP to prevent ever having to go to CMV, SIMV, or other forms of intubated ventilation, and this indicates that the patient is not as ill as one on CMV or SIMV. Accordingly, we observed the selection of features corresponding to these modes, and their weights—CMV higher than SIMV higher than CPAP—were commensurate with the mortality risk implied by each mode (Table).

Our classifiers also were able to glean a set of interventions that appear to serve as proxies for comorbidities not explicitly documented in the nursing note that indicate poor prognoses. For example, in order to treat a patient exhibiting increased intracranial pressure (ICP), a physician would likely order the administration of mannitol, an osmotic diuretic that is widely used to lower ICP, and a nurse would carry out this order. However, the nurse's note will usually not document the increased ICP – at least not explicitly, in that an ICP value might be given in their note although no judgment is made as to whether it is abnormal or not – whereas their note will almost always document the administration of mannitol. Other examples of selected features themselves also serve as outright indicators of poor prognosis, such as metabolic acidosis, midline shift, and cardiac arrest.

An interesting set of features that we did not expect to discover, but nevertheless were selected by the sparser classifiers, were related to a patient's family visitors and relationship status (Table). It is common for a patient's family to visit them in the hospital, and a nurse usually documents these visits in their notes. We observed that terms such as "wife", "son", and "daughter" were associated with mortality, while "father", "mother", and "parent" carried negative associations. One explanation for this observation is that those patients having parents present at their bedside are more likely to be younger, and consequently have better prognoses compared to patients whose spouse or children are visiting, the presence of whom implies increased age, on average. This is consistent with the findings of other studies that have developed clinical models to predict mortality risk among ICU patients [36–39], which have found a positive association between age and mortality.

We also remark that these classifiers are performing a form of information extraction; given a set of notes, not only can the extracted features that overlap with those learnt for the outcome be used to generate predictions – those features associated with the predicted outcome and in the text of a given note can be extracted and presented to users This capability could prove useful in other contexts with more complex outcomes having longer time horizons for intervention, such as that of readmission. In that case, the information – in the form of *n*-gram features and potentially also phrases – extracted would be more likely to be actionable by, and thus be utilized by, providers and care transition teams. For example, a regularized classifier developed to predict readmission risk based on text would be able to quickly find and "tag" charts that predicted positive also with those features contributing to the increased risk. In such a problem setting, we envision such a classifier learning groups of features corresponding to the presence of complex histories involving substance abuse, medication

non-compliance, mental illness, or similar, enabling these patients to receive more targeted transitional care.

A limitation of our study is that our classifiers used only the text of nursing notes as input features, and did not include other types of notes or features that could be built from structured data elements present in the EMR, such as physiological vital signs or laboratory values. It is possible that including these additional data elements could improve classifier performance, although it is not yet clear how to best combine these structured data elements with text under the influence of regularization, nor how the resulting models would be interpreted. In addition, beyond the grouping effect, the classifier did not explicitly account for correlations between features or groups of features; such correlations, if found and determined to be clinically relevant, could uncover novel interactions between risk factors not yet recognized by clinicians. Furthermore, the interpretability of our models was somewhat complicated by the presence of nonclinical predictors (such as those relating to the presence of family at the bedside) selected by both classifiers; a filter list of terms and an associated methodology to generate such a list (potentially via comparison to nonclinical corpora serving as references) could be developed to restrict terms to only those that are clinically relevant, and to minimize the impact of variation in documentation styles between different care units and caregivers.

Finally, while feature selection methods largely do not figure into existing clinical NLP approaches making use of free text, we believe they have the potential to substantially improve models based on NLP. First, feature selection approaches such as ours could be used to validate the features utilized by classifiers previously considered as "black boxes" and to evaluate their clinical relevance, ultimately improving their usability and generalizability (or portability). As clinical decision support and other systems come to rely on classifiers derived from clinical text, it is imperative that the output of such systems be clinically relevant and interpretable, and thus actionable by providers. Second, while classifiers based on "expert input" may perform adequately when the features associated with the outcome are well characterized, and in conjunction with other non-text clinical features, such methods run the risk of missing out on potentially interesting predictors latent in the text and hence yielding suboptimal classifier performance. These approaches also hamper portability of the resulting classifier, as a classifier relying on features resulting from "expert input" must necessarily be re-validated on new corpora with different sets of manually derived text features in order to ensure optimal performance.

Lastly, several recent studies and reviews [7,40–42] have investigated and emphasized the portability of classifiers based on clinical free text and other NLP systems. Indeed, the need for clinical NLP systems to be portable as they mature out of the lab and begin to integrate and make use of data from multiple institutions will only grow. Even though we were not able to obtain data from different institutions in order to validate our hypothesis, regularization methods have the potential to restrict the feature space so as to avoid overfitting and thus improve classifier generalizability, and in turn, its portability. However, regularization by itself likely will not serve as a panacea to the question of classifier portability; other, synergistic approaches, such as model blending or ensembling [43–45], could be used to improve the portability of a clinical free-text-based classifier that must necessarily generalize well between multiple institutions.

5. Conclusion

Applying elastic-net regularization to classifiers based on clinical free text reduced the number of features selected by more than a thousandfold, thereby making those classifiers more easily interpretable, while sparing performance. The features selected were also clinically relevant, and correlated well with what is currently known about ICU outcomes. In addition, by avoiding overfitting, regularized text classifiers have the potential to improve on the usability and portability of existing methods within the field of clinical NLP.

Contributors

All authors meet the ICMJE criteria for authorship. Their contributions were: BJM: conception and design, acquisition of data, analysis and interpretation of data, drafting the article, and final approval of the version to be published. WJB: analysis and interpretation of data, critical revision of the manuscript for important intellectual content, and final approval of the version to be published.

RAD: conception and design, analysis and interpretation of data, critical revision of the manuscript for important intellectual content, and final approval of the version to be published.

Funding

This work was supported by Innovations Fund of the Philip R. Lee Institute for Health Policy Studies, the Center for Healthcare Value, both at the University of California, San Francisco, and the Andrew Grove Family Foundation. The funders were not involved in the design, conduct, or evaluation of the research.

Ethics approval

The study was deemed to be exempt from review by the Committee on Human Research of the University of California, San Francisco.

Acknowledgments

The authors are especially grateful to Bethany Percha of Stanford University and the anonymous reviewers for their helpful comments on earlier drafts of this paper.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbi.2015.02.003.

References

[1] Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics 2007;23(19):2507–17.
[2] Guyon I, Elisseeff A. An introduction to variable and feature selection. J Machine Learn Res 2005;3:1157–82.
[3] West M, Blanchette C, Dressman H, et al. Predicting the clinical status of human breast cancer by using gene expression profiles. Proc Nat Acad Sci USA 2001;98(20):11462–7.
[4] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. 2nd ed. Springer; 2008 [Springer Series in Statistics].
[5] Candes E, Tao T. The Dantzig selector: statistical estimation when *p* is much larger than *n*. Ann Statist 2007;35:2313–51.
[6] Savova GK, Masanz JJ, Ogren PV, et al. Mayo Clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010;17:507–13.
[7] Khor R, Yip WK, Bressel M, et al. Practical implementation of an existing smoking detection pipeline and reduced support vector machine training corpus requirements. J Am Med Inform Assoc 2013 [e-pub].
[8] Marafino BJ, Davies JM, Bardach NS, et al. *N*-gram support vector machines for scalable procedure and diagnosis classification, with applications to clinical free text data from the intensive care unit. J Am Med Info Assoc 2014. http://dx.doi.org/10.1136/amiajnl-2014-002694 [Apr 30].

[9] Variable selection: current practice in epidemiological studies. Eur J Epidemiol 2009;24(12):733–6.
[10] Whittingham MJ, Stephens PA, Bradbury RB, et al. Why do we still use stepwise modelling in ecology and behaviour? J Anim Ecol 2006;75:1182–9.
[11] Miller AJ. Subset selection in regression. Chapman & Hall; 1990.
[12] Hocking RR. The analysis and selection of variables in linear regression. Biometrics 1976:32.
[13] Zhou H, Hastie T. Regularization and variable selection via the elastic net. J R Statist Soc B 2005;67(2):301–20.
[14] Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. Technometrics 1970;12:55–67.
[15] Tibshirani R. Regression shrinkage and selection via the lasso. J R Statist Soc B 1994;58(1):267–88.
[16] Boyd B, Vandenberghe L. Convex optimization. New York, NY, USA: Cambridge University Press; 2004.
[17] Efron B, Hastie T, Johnstone I, et al. Least angle regression. Ann Stat 2004;32:407–99.
[18] Mol CD, Vito ED, Rosasco L. Elastic net regularization in learning theory. J Complex 2009;25:201–30.
[19] Wright A, McCoy AB, Henkin S, et al. Use of a support vector machine for categorizing free-text notes: assessment of accuracy across two institutions. J Am Med Info Assoc 2013;20:887–90.
[20] Garla V, Taylor C, Brandt C. Semi-supervised clinical text classification with Laplacian SVMs: an application to cancer case management. J Biomed Inform 2013;46:869–75.
[21] Walsh C, Hripcsak G. The effects of data sources, cohort selection, and outcome definition on a predictive model of risk of thirty-day hospital readmissions. J Biomed Inform 2014;52:418–26.
[22] Saeed M, Villarroel M, Reisner AT, et al. Multiparameter intelligent monitoring in intensive care II (MIMIC-II): a public-access ICU database. Crit Care Med 2011;39:952–60.
[23] Goldberger AL, Amaral LAN, Glass L, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 2000;101:e215–20.
[24] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. J Mach Learn Res 2011;12:2825–30.
[25] Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinform 2006;7:91.
[26] Le Gall J-R, Loirat P, Alperovitch A, et al. A simplified acute physiology score for ICU patients. Crit Care Med 1984;12:975–7.
[27] Hosmer DW, Lemeshow S, Sturdivant RX. Applied logistic regression. 3rd ed. New York, NY, USA: Wiley; 2013.
[28] McMahan HB. Follow-the-regularized-leader and mirror descent: equivalence theorems and L1 regularization. In: Proceedings of the 14th international conference on artificial intelligence and statistics (AISTATS); 2011.
[29] McClish DK, Powell SH. How well can physicians estimate mortality in a medical intensive care unit? Med Decis Making 1989;9:125–32.
[30] Lehman LW, Saeed M, Long W, et al. Risk stratification of ICU patients using topic models inferred from unstructured progress notes. AMIA Annu Symp Proc 2012;2012:505–11.
[31] Teh YW, Jordan MI, Beal MJ, et al. Hierarchical Dirichlet processes. J Acoust Soc Am 2005;101(476):1566–82.
[32] Moore RC, Denero J. L1 and L2 regularization for multiclass hinge loss models. In: Symposium on machine learning in speech and natural language processing; 2011.
[33] Vincent JL, Moreno R, Takala J, et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the working group on sepsis-related problems of the European Society of Intensive Care Medicine. Intensive Care Med 1996;22:707–10.
[34] Esteban A, Frutos F, Tobin MJ, et al. A comparison of four methods of weaning patients from mechanical ventilation. N Engl J Med 1995;332:345–50.
[35] McConville JF, Kress JP. Weaning patients from the ventilator. N Engl J Med 2012;367:2233–9.
[36] Knaus WA, Wagner DP, Draper EA, et al. The APACHE III prognostic system: risk prediction of hospital mortality for critically ill hospitalized adults. Chest 1991;100:1619–36.
[37] Higgins TL, Teres D, Copes WS, et al. Assessing contemporary intensive care unit outcome: an updated mortality probability admission model (MPM_0-III). Crit Care Med 2007;35:827–35.
[38] Kuzniewicz MW, Vasilevskis EE, Lane R, et al. Variation in ICU risk-adjusted mortality: impact of methods of assessment and potential confounders. Chest 2008;133:1319–27.
[39] Erickson S, Vasilevskis EE, Kuzniewicz MW, et al. The effect of race and ethnicity on outcomes among patients in the intensive care unit: a comprehensive study involving socioeconomic status and resuscitation preferences. Crit Care Med 2011;39:429–35.
[40] Sarker A, Gonzalez G. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. J Biomed Inform 2015;53:196–207.
[41] Chapman WW, Cohen KB. Current issues in biomedical text mining and natural language processing. J Biomed Inform 2009;42:757–9.
[42] Cohen KB, Yu H, Bourne PE, et al. Translating biology: text mining tools that work. Pac Symp Biocomput 2008;13:551–5.
[43] Polikar R. Ensemble-based systems in decision making. IEEE Circ Syst 2006;6:21–45.
[44] Jacobs RA, Jordan MI, Nowlan SJ, et al. Adaptive mixtures of local experts. Neural Comput 1991;3:79–87.
[45] Kuncheva IJ. Combining pattern classifiers: methods and algorithms. New York (NY, USA): Wiley Interscience; 2005.