Privacy-Preserving Predictive Models for Lung Cancer Survival Analysis

Glenn Fung[1], Shipeng Yu[1], Cary Dehing-Oberije[2], Dirk De Ruysscher[2], Philippe Lambin[2], Sriram Krishnan[1], R. Rao Bharat[1]
[1] CAD and Knowledge Solutionis, Siemens Medical Solutions USA, Inc., Malvern, PA USA.
[2] MAASTRO clinic, the Netherlands.

Abstract

Privacy-preserving data mining (PPDM) is a recent emergent research area that deals with the incorporation of privacy preserving concerns to data mining techniques. We consider a real clinical setting where the data is horizontally distributed among different institutions. Each one of the medical institutions involved in this work provides a database containing a subset of patients. There is recent work that shows the potential of the PPDM approach in medical applications. However, there is few work in developing/implementing PPDM for predictive personalized medicine. In this paper we use real data from several institutions across Europe to build models for survival prediction for non-small-cell lung cancer patients while addressing the potential privacy preserving issues that may arise when sharing data across institutions located in different countries. Our experiments in a real clinical setting show that the privacy preserving approach may result in improved models while avoiding the burdens of traditional data sharing (legal and/or anonymization expenses).

1 Introduction

Privacy-preserving data mining (PPDM) is a recent emergent research area that deals with the incorporation of privacy preserving concerns to data mining techniques. We are particularly interested in a scenario when the data is horizontally distributed among different institutions. In the medical domain this means that each medical institution (hospitals, clinics, etc.) provides a database containing a complete (or almost complete) subset of item sets (patients). An efficient PPDM algorithm should be able to process the data from all the sources and learn data mining/machine learning models that take into account all the information available without sharing explicitly private information among the sources. The ultimate goal of a PPDM model is to perform similarly or identically to a model learned by having access to all the data at the same time.

P3DM'08, April 26, 2008, Atlanta, Georgia, USA.

There are have been a push for the incorporation of electronic health records (EHR) in medical institutions worldwide. There seems to be a consensus that the availability of EHR will have several significant benefits for health systems across the world, including: improvement of quality of care by tracking performance on clinical measures, better and more accurate insurance reimbursement, computer assisted diagnosis (CAD) tools, etc. Therefore, there is a constant increase on the number of hospitals saving huge amounts of data that can be used to build predictive models to assist doctors in the medical decision process for treatment, diagnosis, and prognosis among others. However, sharing the data across institutions becomes a difficult and tedious process that also involves considerable legal and economic burden on the institutions sharing the medical data.

In this paper we explore two privacy preserving techniques applied to learn survival predictive models for non-small-cell lung cancer patients treated with (chemo) radiotherapy. We use real data collected from patients treated on three European institutions in two different countries (the Netherlands and Belgium) to build our models. The framework we are describing in this paper allows to design/learn improved predictive models that perform better than the individual models obtained by using local data from only one institution, without addressing the local and international privacy preserving concerns that arise when sharing patient-related data. As far as we know, there is none previous work related to learning survival models for lung cancer radiation therapy addressing PP concerns.

The rest of the paper is organized as follows: in the next section, we introduced the notation used in the paper. In section 3 we present an overview of the related work. In sections 4.1 and 4.3 we present the overview of the two methods used for our predictive models: Newton-Lagrangian Support Vector Machines [5] and Cox Regression [3]. Later in sections 4.2 and 4.4, we present the technical details of the corresponding privacy preserving (PP) algorithms used. We conclude

the paper describing our application with experimental results performed in a real clinical setting and the conclusions.

2 Notation

We describe our notations now. All vectors will be column vectors unless transposed to a row vector by a prime $'$. For a vector $x \in R^n$ the notation x_j will signify either the j-th component or j-th block of components. The scalar (inner) product of two vectors x and y in the n-dimensional real space R^n will be denoted by $x'y$. The notation $A \in R^{m\times n}$ will signify a real $m \times n$ matrix. For such a matrix, A' will denote the transpose of A, A_i will denote the i-th row or i-th block of rows of A. A vector of ones in a real space of arbitrary dimension will be denoted by e. Thus for $e \in R^m$ and $y \in R^m$ the notation $e'y$ will denote the sum of the components of y. A vector of zeros in a real space of arbitrary dimension will be denoted by 0. For $A \in R^{m\times n}$ and $B \in R^{k\times n}$, a *kernel* $K(A, B')$ maps $R^{m\times n} \times R^{n\times k}$ into $R^{m\times k}$. In particular, if x and y are column vectors in R^n then, $K(x', y)$ is a real number, $K(x', B')$ is a row vector in R^k and $K(A, B')$ is an $m \times k$ matrix. The abbreviation "s.t." stands for "subject to".

3 Related Work

As a consequence of the recent advances of network computing, there has been recently great interest in privacy-preserving data mining techniques. An extensive review of PPDM techniques can be found in [14]. Most of the available data mining techniques require and assume that there is complete access to all data at all times. This may not be true for example, in an uncentralized distributed medical setting where for each data source or institution, there are local procedures in place to enforce privacy and security of the data. If this is the case, there is a need to use efficient data mining and machine learning techniques that can use data across institutions while complying with the non-disclosure nature of the available data. There are two main kinds of data partitioning when dealing with distributed setting where PPDM is needed: a) the data is partitioned vertically, this means that all institutions have some subset of features (predictors, variables) for all the available patients. When this is the case, several techniques have been proposed to address the issue including: adding random perturbations to the data [2, 4]. The other popular PPDM setting occurs when the data is partitioned horizontally among institutions, that means that different entities hold the same input features for different groups of individuals. This case have been addressed in [16, 15] by privacy-preserving SVMs and induction tree classifiers. There are several other recently proposed privacy preserving classifying techniques including cryptographically private SVMs [7], wavelet-based distortion [10]. There is recent work that shows the potential of the approach [6, 12] in medical settings. However , there is few work in developing/implementing PPDM for predictive personalized medicine.

4 Privacy-Preserving Predictive Models (PPPM)

In this section we introduce two PP predictive models, namely PP Support Vector Machines and PP Cox Regression. We first give an overview of the two techniques in sections 4.1 and 4.3, and then present the PP versions in sections 4.2 and 4.4.

4.1 Overview of Support Vector Machines. We describe in this section the fundamental classification problems that lead to the standard quadratic Support vector machine (SVM) formulation that minimizes a quadratic convex function. We consider the problem of classifying m points in the n-dimensional real space R^n, represented by the $m \times n$ matrix A, according to membership of each point A_i in the classes +1 or -1 as specified by a given $m \times m$ diagonal matrix D with ones or minus ones along its diagonal. For this problem, the standard support vector machine with a linear kernel AA' [13] is given by the following quadratic program for some $\nu > 0$:

$$\begin{array}{rl} \min\limits_{(w,\gamma,y)\in R^{n+1+m}} & \nu e'y + \frac{1}{2}w'w \\ \text{s.t.} & D(Aw - e\gamma) + y \geq e \\ & y \geq 0. \end{array} \tag{4.1}$$

As depicted in Figure 1, w is the normal to the bounding planes:

$$\begin{array}{rcrcl} x'w & - & \gamma & = & +1 \\ x'w & - & \gamma & = & -1, \end{array} \tag{4.2}$$

and γ determines their location relative to the origin. The first plane above bounds the class +1 points and the second plane bounds the class -1 points when the two classes are strictly linearly separable, that is when the slack variable $y = 0$. The linear separating surface is the plane

$$x'w = \gamma, \tag{4.3}$$

midway between the bounding planes (4.2). If the classes are linearly inseparable then the two planes bound the two classes with a "soft margin" determined by a nonnegative slack variable y, that is:

$$\begin{array}{rcrcll} x'w - \gamma & + & y_i & \geq & +1, & \text{for } x' = A_i \text{ and } D_{ii} = +1, \\ x'w - \gamma & - & y_i & \leq & -1, & \text{for } x' = A_i \text{ and } D_{ii} = -1. \end{array} \tag{4.4}$$

The 1-norm of the slack variable y is minimized with weight ν in (4.1). The quadratic term in (4.1), which is twice the reciprocal of the square of the 2-norm distance $\frac{2}{\|w\|}$ between the two bounding planes of (4.2) in the n-dimensional space of $w \in R^n$ for a *fixed* γ, maximizes that distance, often called the "margin". Figure 1 depicts the points represented by A, the bounding planes (4.2) with margin $\frac{2}{\|w\|}$, and the separating plane (4.3) which separates $A+$, the points represented by rows of A with $D_{ii} = +1$, from $A-$, the points represented by rows of A with $D_{ii} = -1$. For this paper we used Newton-

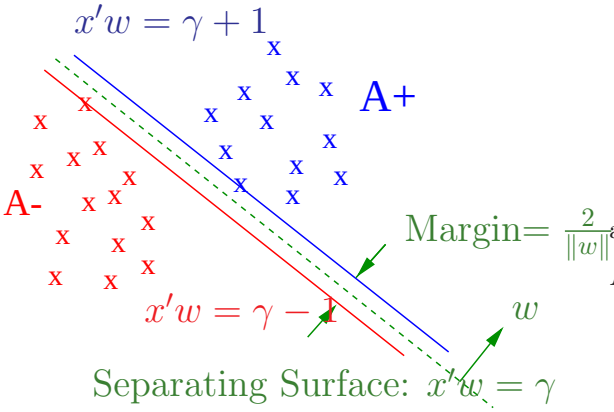

Figure 1: **The bounding planes (4.2) with margin $\frac{2}{\|w\|}$, and the plane (4.3) separating $A+$, the points represented by rows of A with $D_{ii} = +1$, from $A-$, the points represented by rows of A with $D_{ii} = -1$.**

Lagrangian SVM (NSVM), an algorithm based on an essentially equivalent formulations of this classification problem [5]. In this formulation, the square of 2-norm of the slack variable y is minimized with weight $\frac{\nu}{2}$ instead of the 1-norm of y as in (4.1). In addition the distance between the planes (4.2) is measured in the $(n+1)$-dimensional space of $(w, \gamma) \in R^{n+1}$, that is $\frac{2}{\|(w,\gamma)\|}$. Measuring the margin in this $(n+1)$-dimensional space instead of R^n induces strong convexity and has little or no effect in general on the problem.

4.2 Privacy Preserving SVMs. For our privacy preserving application we chose to use a technique on random kernel mappings recently proposed by Mangasarian and Wild on [11]. The algorithm is based on two simple basic ideas:

1. **The use of reduced kernel mappings** [9, 8], where the kernel centers are randomly chosen. Instead of using the complete kernel function $K(A, A') : R^{m\times n} \rightarrow R^{m\times m}$ as it is usually done in kernel methods they propose the use of a reduced kernel $K(A, B') : R^{m\times n} \rightarrow R^{m\times \tilde{m}}$, where $B \in R^{\tilde{m}\times n}$ is a completely random matrix with fewer rows than the number of available features, $(\tilde{m} < n)$.
2. **Each entity makes public only a common randomly generated linear transformation of the data** given by the matrix product of its privately held matrix of data rows multiplied by the transpose of a common random matrix B for linear kernels, and a similar kernel function for nonlinear kernels. In our experimental setting, we assumed that all the available patient data is normalized between 0 and 1 and therefore the elements of B were generated according to a normal distribution with mean zero, variance one and standard deviation one.

Next, we formally introduce the PPSVM algorithm as presented in [11]

ALGORITHM 4.1. **Nonlinear PPSVM Algorithm**

(I) All q entities agree on the same random matrix $B \in R^{\tilde{m}\times n}$ with $\tilde{m} < n$ for security reasons as justified in the explanation immediately following this algorithm. All entities make public the class matrix D (labels) where $D_{ll} = \pm, l = 1, \ldots, m$ for the each of the data matrices $A_i, i = 1, \ldots, q$ that they all hold.

(II) Each entity generates its own privately held random matrix $B_{.j} \in R^{\tilde{m}\times n_j}$, $j = 1, \ldots\ldots, p$, where n_j is the number of input features held by entity j.

(III) Each entity j makes public its nonlinear kernel $K(A_j, B')$. This does not reveal A_j but allows the public computation of the full nonlinear kernel:
(4.5)

$$K(A, B') = K\left(\begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_q \end{bmatrix}, B'\right) = \begin{bmatrix} K(A_1, B') \\ K(A_2, B') \\ \vdots \\ K(A_q, B') \end{bmatrix}$$

(IV) A publicly calculated linear classifier $K(x', B')u - \gamma = 0$ is computed by any linear hyperplane based classification or regression method method such as the ones presented in sections 4.1 and 4.3.

(V) For each new *$x \in R^n$, obtained by an entity, that entity privately computes $K(x', B')$ and classifies the given x according to the sign of $K(x', B')u - \gamma$.*

Note that algorithm 4.1 works for any kernel with the following associative property:

$$K\left(\begin{bmatrix} C \\ D \end{bmatrix}, F\right) = \begin{bmatrix} K(C, F) \\ K(D, F) \end{bmatrix}$$

Which is, in particular, the case of the linear kernel $K(A, B') = AB'$ and that we will use for the rest of the paper.

As stated in [11], it is important to note than in the the above algorithm no entity j reveals its data nor its components of a new testing data point. When $\bar{m} < n$, there is an infinite number of matrices $A_i \in R^{mi \times n}$ in the solution set of the equation $A_i B' = P_i$, when B and P_i are given. This claim can be justified by the well-known properties of under-determined systems of linear equations. Furthermore, the following proposition which is originally stated and proved in [11] is aimed to formally support the claim presented above:

PROPOSITION 4.2. **(infinite solutions of $A_i B' = P_i$ if $\bar{m} < n$)** *Given the matrix product $P_i' = AiB' \in R^{m_i \times \bar{m}}$, where $A_i \in R^{m_i \times n}$ is unknown and B is a known matrix in $R^{\bar{m} \times n}$ with $\bar{m} < n$, there are an infinite number of solutions, including:*

$$\begin{pmatrix} n \\ \bar{m} \end{pmatrix}^{m_i} = \left(\frac{n!}{(n-\bar{m})!\bar{m}!} \right)^{m_i}$$

possible solutions $A_i \in R^{m_i \times n}$ to the equation $A_i B' = P_i$. Furthermore, the infinite number of matrices in the affine hull of these $\begin{pmatrix} n \\ \bar{m} \end{pmatrix}^{m_i}$ matrices also satisfy $A_i B' = P_i$.

4.3 Overview of Cox Regression. Cox regression, or the Cox propositional-hazards model, is one of the most popular algorithms for survival analysis [3]. Apart from baing a classification algorithm which directly deal with binary or multi-class outcomes, Cox regression defines a semi-parametric model to directly relate the predictive variables with the real outcome which is in general the survival time (e.g., in years).

Let T represent survival time. The so-called *hazard function* is a representation of the distribution of survival times, which assesses the instantaneous risk of demise at time t, conditional on survival to that time:

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr[(t \leq T < t + \Delta t) | T \geq t]}{\Delta t}.$$

The Cox regression model assumes a linear model for the log-hazard, or as a multiplicative model for the hazard:

$$\log h(t) = \alpha(t) + w'x, \tag{4.6}$$

where x denote the covariates for each observation, and the *baseline hazard* $\alpha(t)$ is unspecified. This model is semi-parametric because while the baseline hazard can take any form, the covariates enter the model linearly. Now given any two observations x_i and x_j, from the definition of hazard function we can get

$$\frac{h(t_i)}{h(t_j)} = \exp[w'(x_i - x_j)],$$

which is independent of time t. The baseline hazard $\alpha(t)$ also does not affect the hazard ratio. This is why the Cox model is a *proportional-hazards model*.

And Cox has showed in [3] that even though the baseline hazard is unspecified, the Cox model can still be estimated by the method of *partial likelihood*. It is also possible to extract an estimate of the baseline hazard after having fit the model.

4.4 Privacy Preserving Cox Regression. The main idea of the privacy preserving SVM is to perform a random mapping of the original predictive variables into a new space, and then perform standard SVM on the new space. Since in the Cox regression the interaction between the parameter of the models and the data is linear, we can also apply the same idea presented in section 4.2 for the *privacy preserving Cox regression*. Given the random matrix B and assuming that we are using a linear kernel, equation 4.6 is slightly changed to:

$$\log h(t) = \alpha(t) + w'xB', \tag{4.7}$$

Again it is important to note, that to our knowledge, this is the first time that privacy preserving techniques are applied for survival analysis methods.

5 Application: 2-Year Survival Prediction for Non-Small Cell Lung Cancer Patients

Radiotherapy, combined with chemotherapy, is treatment of choice for a large group of non-small cell lung cancer (NSCLC) patients. The treatment is not restricted to patients with mediastinal lymph node metastasis, but is also indicated for patients who are inoperable because of their physical condition. In addition, the marginal role of radiotherapy and chemotherapy for the survival of NSCLC patients has been changed into one of significant importance. Improved radiotherapy treatment techniques allow an increase of the radiation dose, while at the same time more effective chemoradiation schemes are being applied. These developments have lead to an improved outcome in terms of survival. Although the introduction of FDG-PET scans has enabled more accurate detection of positive lymph nodes and distant metastases, leading to stage migration, the TNM staging system is still highly inaccurate for the prediction of survival outcome for this group of patients [1]. In summary, an increasing number of patients is being treated successfully with (chemo) radiation, but

an accurate estimation of the survival probability for an individual patient, taking into account patient, tumor as well as treatment characteristics and offering the possibility for treatment decision-making, is currently not available.

At present, generally accepted prognostic factors for inoperable patients are performance status, weight loss, presence of comorbidity, use of chemotherapy in addition to radiotherapy, radiation dose and tumor size. For other factors such as gender and age the literature shows inconsistent results, making it impossible to draw definitive conclusions. In these studies CT-scans were used as the major staging tool. However, the increasing use of FDG-PET scans offers the possibility to identify and use new prognostic factors. In a recent study it was shown that number of involved nodal areas quantified by PET-CT was an important prognostic factor [1]. We performed this retrospective study to develop and validate several prediction models for 2-year survival of NSCLC patients, treated with (chemo) radiotherapy, taking into account all known prognostic factors. To the best of our knowledge, this is the first study of prediction models for NSCLC patients treated with (chemo)radiotherapy

5.1 Patient Population. Between May 2002 and January 2007, a total number of 455 inoperable NSCLC patients, stage I-IIIB, were referred to MAASTRO clinic to be treated with curative intent. Clinical data of all these patients were collected retrospectively by reviewing the clinical charts. If PET was not used as a staging tool, patients were excluded from the study. This resulted in the inclusion of 399 patients. The primary gross tumor volume ($\mathrm{GTV_{primary}}$) and nodal gross tumor volume ($\mathrm{GTV_{nodal}}$) were calculated, as delineated by the treating radiation oncologist, using a commercial radiotherapy treatment planning system (Computerized Medical Systems, Inc, CMS). The sum of $\mathrm{GTV_{primary}}$ and $\mathrm{GTV_{nodal}}$ resulted in the GTV. For patients treated with sequential chemotherapy these volumes were calculated using the post-chemotherapy imaging information. The creation of the volumes was based on PET and CT information only; bronchoscopic findings were not taken into account. The number of positive lymph node stations was assessed by the nuclear medicine specialist using either an integrated FDG-PET-CT scan or a CT-scan combined with FDG-PET-scan. T-stage and N-stage were assessed using pre-treatment CT, PET and mediastinoscopy when applicable. For patients treated with sequential chemotherapy stage as well as number of positive lymph node stations was assessed using pre-chemotherapy imaging information.

Additionally, a smaller number of patients treated at the other two centers, the Gent hospital and the Leuven hospital, were also collected for this study. There are respectively 112 and 40 patients from the Gent and Leuven hospitals, and the same set of clinical variables as the MAASTRO patients were measured.

5.2 Radiotherapy Treatment Variables. No elective nodal irradiation was performed and irradiation was delivered 5 days per week. Radiotherapy planning was performed with a Focus (CMS) system, taking into account lung density and according to ICRU 50 guidelines. There were four different radiotherapy treatment regimes applied for these patients in this retrospective study, therefore to account for the different treatment time and number of fractions per day, the equivalent dose in 2 Gy fractions, corrected for overall treatment time (EQD2,T), was used as a measure for the intensity of chest radiotherapy 5.8. Adjustment for dose per fraction and time factors were made as follows:

$$\mathrm{EQD2,T} = D\left(\frac{d+\beta}{2+\beta}\right) - \gamma \max(0, T - T_k), \tag{5.8}$$

where D is the total radiation dose, d is dose per fraction, $\beta = 10$ Gy, T is overall treatment time, T_k is the accelerated repopulation kick-off time which is 28 days, and γ is the loss in dose per day due to repopulation which is 0.66 Gy/day.

5.3 Experimental Setup. In this paper we focus on 2-year survival prediction for these NSCLC patients, which is the most interesting prediction from clinical perspective. The survival status was evaluated in December 2007. The following 6 clinical predictors are used to build the prediction models: gender (two groups: male/female), WHO performance status (three groups: $0/1/\geq 2$), lung function prior to treatment (forced expiratory volume, in the range of $17 \sim 139$), number of positive lymph node stations (five groups: $0/1/2/3/\geq 4$), natural logarithm of GTV (in the range of $-0.17 \sim 6.94$), and the equivalent dose corrected by time (EQD2,T) from (5.8). The mean values across patients are used to impute the missing entries if some of these predictors are missing for certain patients. To account for the very different number of patients from the three sites, a subset of MAASTRO patients were selected for the following study. In the following we use the names "MAASTRO", "Gent" and "Leuven" to denote the data from the three different centers.

For the SVM methods, since they can only deal with binary outcome, we only use the patients with 2-year follow-up and create an outcome for them with $+1$ meaning they survived 2 years, and -1 meaning they didn't survive 2 years. This setting leads to 70, 37 and

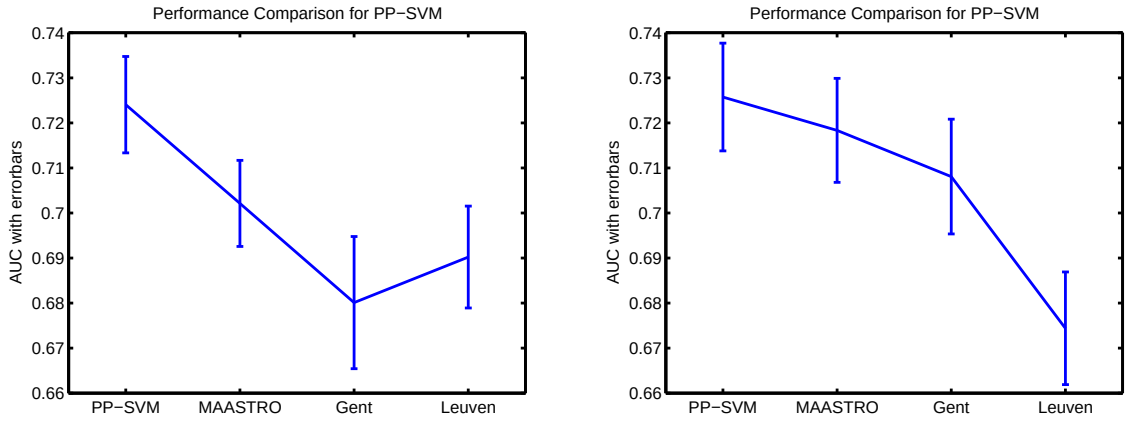

Figure 2: AUC comparison for privacy preserving SVMs with 40% (left) and 60% (right) training patients. The error bars are calculated based on 100 times of random splits of the data.

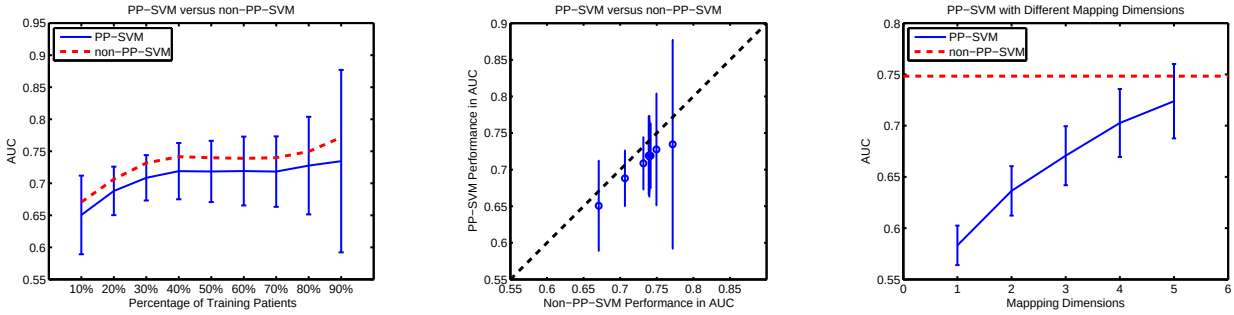

Figure 3: AUC comparison between PP-SVMs and non PP-SVMs (which explicitly use all the training data from different centers, and thus upper-bound the predictive performance of PP-SVMs). We compare the two with different percentages of training patients (left), in a scatter plot (middle), and with different dimensions $\bar{m}$ for PP-SVMs (right) for a 40% split.

23 patients for the MAASTRO, Gent and Leuven sets, respectively. For the Cox regression methods, we can potentially use all the patients with the exact number of survived years, and do right censoring for those patients who are still alive. Under this setting we end up with 80, 85 and 40 patients for MAASTRO, Gent and Leuven, respectively.

Under the privacy preserving setting, we are interested in assessing the predictive performance of a model combining the patient data from the three centers together, compared to the models trained based on each of these centers. The data combination needs to be done in a way that sensitive information is not uncovered. Therefore for our experiments we trained the following 4 models under each configuration:

- **PP model**: Apply the privacy preserving techniques we have introduced and train a model using combined data from the three centers.
- **MAASTRO, Gent and Leuven models**: Train models using only the MAASTRO, Gent and Leuven training patients repectively.

For each of the configurations, we vary the percentage of training patients in each of the centers, and report the Area Under the ROC Curve (AUC) for the test patients. Note that the testing was performed using all the test patients from all centers.

6 Results

In Figure 2 we show the results for privacy preserving SVM models, with 2 example training percentages (40% and 60%). The other percentages yield similar results. The error bars are over 100 runs with random split of training/test patients for each center, and each time a random B matrix of dimensionality 5×6 is used for the PP-SVM models. As can be seen, the PP-SVM models achieve the best performance compared to other single-center based models. This is mainly because PP-SVM models are able to use more data in model training, at

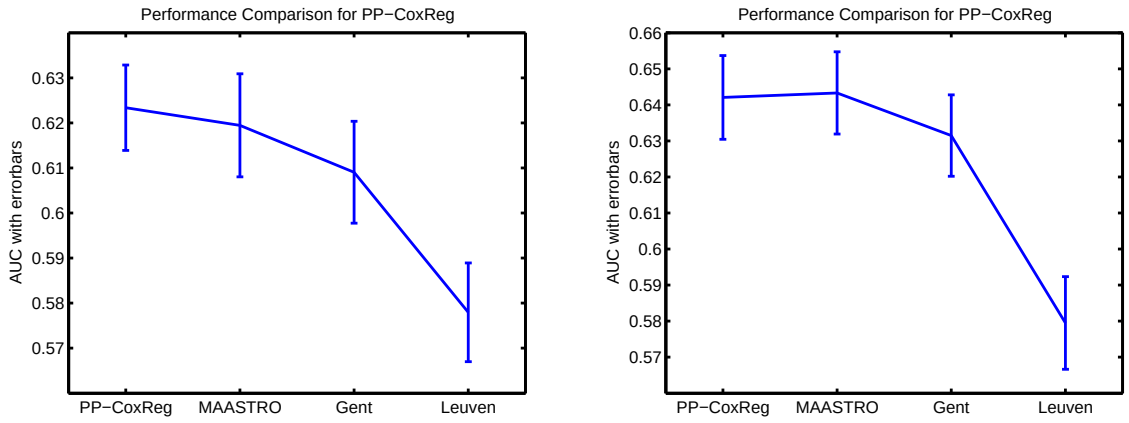

Figure 4: AUC comparison for privacy preserving Cox regression models with 40% (left) and 60% (right) training patients. The error bars are calculated based on 100 times of random splits of the data.

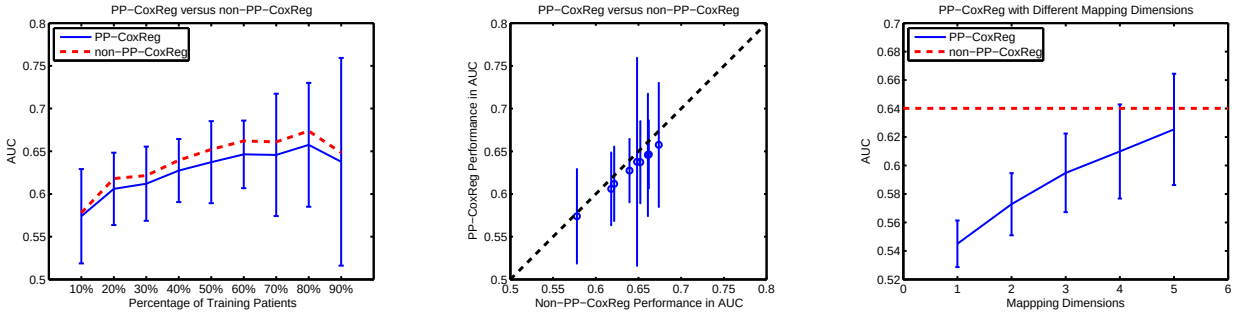

Figure 5: AUC comparison between PP-CoxReg and non PP-CoxReg (which explicitly use all the training data from different centers, and thus upper-bound the predictive performance of PP-CoxReg). We compare the two with different percentages of training patients (left), in a scatter plot (middle), and with different dimensions $\bar{m}$ for PP-CoxReg (right) in a 40% split.

the same time without violating the privacy regulations. When we increase the training percentages, all models will improve (compare Figure 2 right to left), and the single-center based models have a higher improvement. However the PP-SVM models still perform the best.

It is easy to realize that PP-SVM will end up with a performance loss compared to a non PP-SVM model, which explicitly combines all the training patients from different centers and does not preserve privacy. This is because in PP-SVMs a random matrix B projects each patient into a lower dimensional space (for privacy preserving purpose), and thus leads to information loss. To empirically evaluate how much performance loss the PP-SVMs have, we show a more extensive comparison in Figure 3. On the left we show the comparison with different percentages of the training/test splits, and as can be seen the gaps between PP-SVMs and non PP-SVMs are not very big. This indicates PP-SVMs can achieve similar predictive performance while satisfying the privacy preserving requirement. The scatter plot in the middle is another way to visualize these results. On the right we vary the mapping dimensions $\bar{m}$ for the B matrix we used in PP models, and as expected, bigger $\bar{m}$ yield better predictive performance. Therefore, in practice we normally choose $\bar{m} = n - 1$ to maximize the performance of the PP models (which still perfectly satisfies the privacy preserving requirements). From this comparison we see that there is a big error bar for different B matrices, and one interesting future work is to identify the best B matrix for PP models.

In Figure 4 we also empirically evaluate the results for privacy preserving Cox regression models, also with the 2 example training percentages (40% and 60%). They have the same trend as we have seen in Figure 2, but it is interesting that with a higher percentage of training data (e.g., 60% on the right), PP-CoxReg performs the same as the model trained using only MAASTRO training patients. This indicates PP-

CoxReg model is more sensitive to the different characteristics of the data from different centers. In practice, we need to carefully investigate the different data distributions to estimate the benefits of combining them.

We also empirically compare the PP Cox regression models with non PP-CoxReg models in Figure 5. As can be seen, the gaps between PP-CoxReg and non PP-CoxReg models are even smaller than those between PP-SVM and non PP-SVM models, meaning PP-CoxReg models are more accurate toward the non privacy preserving solutions. In practice we still need to choose $\bar{m} = n - 1$ to maximize the PP-CoxReg performance, and to choose the best B matrix if possible.

7 Discussion and Conclusions

We have applied a simple recently proposed PP technique in a real clinical setting where data is shared across three European institutions in order to build more accurate predictive models than the ones obtained using only data from one institute. We have extended the previously proposed PP algorithm (originally suggested for SVM) to cox regression. As far as we know this is the first work that addresses privacy preserving concerns for survival models. The work presented here is based on preliminary results and we are already working on designing improved algorithms to address several concerns that arise when performing our experiments. One of the concerns that arise (as shown in section 6) is how to address the impact of the variability of the matrix B on the performance of the predictive models. For that, we are currently experimenting with formulations in which the B matrix is intended not only to "de-identify" the data but also to optimally improve model performance. Another relevant concern that we are looking into is, how to weight the importance of data from different institutions, assuming that the reliability of the data or the labels varies among institutions.

References

[1] Dehing-Oberije C, De Ruysscher D, van der Weide H, and et al. Tumor volume combined with number of positive lymph node stations is a more important prognostic factor than tnm stage for survival of non-small-cell lung cancer patients treated with (chemo)radiotherapy. *Int J Radiat Oncol Biol Phys*, (in press).

[2] K. Chen and L. Liu. Privacy preserving data classification with rotation perturbation. In *Proceedings of the Fifth International Conference of Data Mining (ICDM'05)*, pages 589–592. IEEE, 2005.

[3] D. R. Cox. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*, Series B 34:187–220, 1972.

[4] Wenliang Du, Yunghsiang Han, and Shigang Chen. Privacy-preserving multivariate statistical analysis: Linear regression and classification. In *Proceedings of the Fourth SIAM International Conference on Data Mining*, pages 222–233, 2004. http://citeseer.ist.psu.edu/du04privacypreserving.html.

[5] G. Fung and O. L. Mangasarian. Finite Newton method for Lagrangian support vector machine classification. *Special Issue on Support Vector Machines. Neurocomputing*, 55:39–55, 2003.

[6] Gang Kou, Yi Peng, Yong Shi, and Zhengxin Chen. Privacy-preserving data mining of medical data using data separation-based techniques. *Data Science Journal*, 6:429–434, 2007.

[7] S. Laur, H. Lipmaa, and T. Mielikäinen. Cryptographically private support vector machines. In L. Ungar, M. Craven, D. Gunopulos, and T. Eliassi-Rad, editors, *Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 618–624, 2006.

[8] Y.-J. Lee and S.Y. Huang. Reduced support vector machines: A statistical theory. *IEEE Transactions on Neural Networks*, 18:1–13, 2007.

[9] Y.-J. Lee and O. L. Mangasarian. RSVM: Reduced support vector machines. In *Proceedings of the First SIAM International Conference on Data Mining*, 2001.

[10] L. Liu, J. Wang, Z. Lin, and J. Zhang. Wavelet-based data distortion for privacy-preserving collaborative analysis. Technical Report 482-07, Department of Computer Science, University of Kentucky, Lexington, KY 40506, 2007. http://www.cs.uky.edu/ jzhang/pub/MINING/lianliu1.pdf.

[11] O. L. Mangasarian and E. Wild. Privacy-preserving classification of horizontally partitioned data via random kernels. Technical Report 07-03, Computer sciences department, university of Wisconsin - Madison, Madison, WI, 2007.

[12] G. Schadow, S. J. Grannis, and C. J. McDonald. Privacy-preserving distributed queries for a clinical case research network. pages 55–65, 2002.

[13] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, second edition, 2000.

[14] V. Verykios, E. Bertino, I. Fovino, L. Provenza, Y. Saygin, and Y. Theodoridis. State-of-the-art in privacy preserving data mining. *SIGMOD*, 33:50–57, 2004.

[15] Ming-Jun Xiao, Liu-Sheng Huang, Yong-Long Luo, and Hong Shen. Privacy preserving id3 algorithm over horizontally partitioned data. In *PDCAT '05: Proceedings of the Sixth International Conference on Parallel and Distributed Computing Applications and Technologies*, pages 239–243, Washington, DC, USA, 2005. IEEE Computer Society.

[16] Hwanjo Yu, Xiaoqian Jiang, and Jaideep Vaidya. Privacy-preserving svm using nonlinear kernels on horizontally partitioned data. In *SAC '06: Proceedings of the 2006 ACM symposium on Applied computing*, pages 603–610, New York, NY, USA, 2006. ACM.