Information Sciences 330 (2016) 245–259

Contents lists available at ScienceDirect

Information Sciences

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

A distributed ensemble approach for mining healthcare data under privacy constraints

Yan Li, Changxin Bai, Chandan K. Reddy*

Department of Computer Science, Wayne state University, Detroit, MI 48202, United States

ARTICLE INFO

Article history:
Received 16 June 2015
Revised 14 September 2015
Accepted 6 October 2015
Available online 20 October 2015

Keywords:
Privacy-preserving data mining
Ensemble learning
Electronic health record
Boosting
Machine learning
Healthcare

ABSTRACT

In recent years, electronic health records (EHRs) have been widely adapted at many healthcare facilities in an attempt to improve the quality of patient care and increase the productivity and efficiency of healthcare delivery. These EHRs can accurately diagnose diseases if utilized appropriately. While the EHRs can potentially resolve many of the existing problems associated with disease diagnosis, one of the main obstacles in effectively using them is the patient privacy and sensitivity of the medical information available in the EHR. Due to these concerns, even if the EHRs are available for storage and retrieval purposes, sharing of the patient records between different healthcare facilities has become a major concern and has hampered some of the effective advantages of using EHRs. Due to this lack of data sharing, most of the facilities aim at building clinical decision support systems using limited amount of patient data from their own EHR systems to provide important diagnosis related decisions. It becomes quite infeasible for a newly established healthcare facility to build a robust decision making system due to the lack of sufficient patient records. However, to make effective decisions from clinical data, it is indispensable to have large amounts of data to train the decision models. In this regard, there are conflicting objectives of preserving patient privacy and having sufficient data for modeling and decision making. To handle such disparate goals, we develop two adaptive distributed privacy-preserving algorithms based on a distributed ensemble strategy. The basic idea of our approach is to build an elegant model for each participating facility to accurately learn the data distribution, and then transfer the useful healthcare knowledge acquired on their data from these participators in the form of their own decision models without revealing and sharing the patient-level sensitive data, thus protecting patient privacy. We demonstrate that our approach can successfully build accurate and robust prediction models, under privacy constraints, using the healthcare data collected from different geographical locations. We demonstrate the performance of our method using the type-2 diabetes EHRs accumulated from multiple sources from all fifty states in the U.S. Our method was evaluated on diagnosing diabetes in the presence of insufficient number of patient records from certain regions without revealing the actual patient data from other regions. Using the proposed approach, we also discovered the important biomarkers, both universal and region-specific, and validated the selected biomarkers using the biomedical literature.

* Corresponding author. Tel.: +13135779005.
E-mail address: reddy@cs.wayne.edu (C.K. Reddy).

http://dx.doi.org/10.1016/j.ins.2015.10.011
0020-0255/© 2015 Elsevier Inc. All rights reserved.

1. Introduction

In recent years, electronic health records (EHRs) have been widely adapted at many healthcare facilities in an attempt to improve the quality of patient care and increase the productivity and efficiency of healthcare delivery. Typically, EHRs are generated and maintained within a particular healthcare facility, such as a hospital, clinic or physician's office. These EHRs can not only aid in various daily healthcare operations but also help in accurately diagnosing diseases, if utilized appropriately [32].

While the EHRs can resolve many of the existing problems associated with disease diagnosis, one of the main obstacles in effectively using them is the patient privacy and sensitivity of the medical information available in the EHR. EHRs generally contain patient information about demographics, diagnostics, medications, living habits and other health-related information. It is needless to say that most of these information is extremely sensitive [31]. EHR systems must abide the Health Insurance Portability and Accountability Act (HIPAA) [6] to preserve the privacy of patient information. Thus, the public EHR products must be certified by certain institutes such as Certification Commission for Healthcare Information Technology (CCHIT), and Office of the National Coordination for Health Information Technology (ONC) [38]. Due to these legal concerns and medical ethics [4,19], physicians are always keen about maintaining the highest possible standards while protecting patient privacy [36].

Due to these concerns, even if the EHRs are available for storage and retrieval purposes, sharing of patient records between different healthcare facilities has become a major concern and has hampered some of the effective advantages of using EHRs. Due to this lack of data sharing, most of the facilities do not have any other option but to build their own clinical decision support systems with limited amount of patient data available in their own EHRs for important diagnosis related decisions. In addition, it becomes quite infeasible for a newly established healthcare facility, small hospital, or rural hospital to build a robust decision making system due to lack of sufficient patient records. However, to make effective decisions from clinical data, it is indispensable to have large amounts of data to train the decision making models. As a result, these small and/or rural hospitals are less motivated, and in 2011, only 20.8% of them were using EHR systems [12]. In this regard, it is clear that there is a conflicting objective of maintaining patient privacy and having sufficient data for modeling and decision making.

The problem statement of the work that is being developed in this paper is as follows. Given healthcare data collected from multiple facilities, how do we obtain a decision model that leverages the knowledge from all the facilities without revealing any patient specific information from any of the individual facilities. In other words, the goal is to develop a knowledge based data integration mechanism in a privacy-preserving context.

To deal with these challenges, we propose a privacy-preserving data mining framework based on horizontally distributed healthcare data. The goal of our work is to build an effective decision making system that can utilize the knowledge from multiple facilities (or geographical locations) without revealing any patient-level information [24]. In our approach, an elegant model for each participating facility is accurately learnt for approximating the local data distribution, and the useful healthcare knowledge acquired on such local data is then transferred in the form of their own decision models without revealing and sharing the sensitive data, thus, protecting the patient privacy. Transferring knowledge between multiple locations is a common practice with the goal of improving the prediction power [22]. Our approach can successfully build accurate and robust prediction models, under privacy constraints, with healthcare data collected from different geographical locations. Each participator shares its own local model with others and builds a specific integrated model based on its own specifications.

Merely trying to acquire the entire knowledge available from all the participators without carefully accounting for the distribution differences can potentially degrade the performance of the classifier [33]. While the overall data distribution for a particular disease must be similar within different hospitals since we are dealing with the same disease, there will still be certain significant differences that arise due to the differences in the demographics of the patient population. Such distribution differences will play a vital role in building robust integrated decision making models that are specific to the population group. However, in the horizontal distribution privacy-preserving problem, there is no access to the original EHR data, one cannot directly measure the data distribution differences among participators but can only approximately say how large the difference might be by analyzing the difference between the models trained by each participator.

Most of the state-of-the-art privacy-preserving data mining methods for horizontally distributed data are built based on a star network, where an untrusted third-party is needed [11,40]. Each participator shares their statistical data distribution with a centralized agent, and this central agent is responsible for building the integrated decision model [24]. However, this framework suffers from two important issues: (1) participators need frequent information exchange with the central agent and hence the communication cost is high; (2) the decision model is built on the central agent whose aim is to provide a decision support for all participators, but it ignores the data distribution differences between the participators.

In addition, in the horizontal distributed privacy preserving data mining literature, there is no prior work on preventing "negative impact during integration". In our work, as each participator will build a specific integrated model based on their own specifications, they can selectively decide which of the models from other participators can be used to build the integrated model. We build local prediction models to represent data, detect the data distribution difference, and transfer knowledge. In our work, the AdaBoost algorithm is chosen to be the local learner for each participator because it is a simple yet effective classifier which is extensively studied in the literature.

The main contributions of our work are summarized as follows:

- Propose an adaptive distributed privacy-preserving data mining technique based on an ensemble strategy which can successfully acquire knowledge from multiple healthcare facilities without gaining access to the sensitive patient data.

- Propose a new integration scheme which does not require a third-party agent, and each participator can build its own integrated model based on its particular needs and can selectively prevent the "negative impact" during the integration process.
- Demonstrate the performance of the proposed distributed privacy-preserving ensemble method using type-2 diabetes patient records gathered from multiple sources from all the fifty states in the U.S. Evaluate our method on diagnosing diabetes in the presence of insufficient number of patient records from certain regions without revealing the actual patient data from other regions.
- Identify the important global and region-specific biomarkers for type-2 diabetes and validate the selected biomarkers using the biomedical literature.

The rest of this paper is organized as follows: In Section 2, we provide important relevant works on privacy-preserving data mining and also information about the type-2 diabetes disease. In Section 3, we propose our adaptive distributed privacy-preserving ensemble method. Our experimental results along with the important biomarkers discovered using the proposed work are presented in Section 4. Finally, Section 5 concludes our discussion with some future research directions.

2. Related work

In this section, we will provide the related literature in the field of privacy-preserving data mining and then discuss more about the type-2 diabetes disease.

2.1. Privacy-preserving data analysis

Recently, health data privacy has become an important area of research. As a result, privacy-preserving data mining has gained much more attention especially in the context of healthcare data analysis. The state-of-the-art methods for privacy-preserving data mining can be categorized under three types, namely, *randomization*, *k-anonymization*, and *distributed* privacy-preserving data mining [1]. In randomization approach, some techniques are used during the data collection process to induce perturbation in the selected attributes of the original data records for concealing certain sensitive information [3]. These techniques include swapping attribute values [14], principal component analysis (PCA) based techniques [18], adding random components which follow a known probability distribution [2], etc. However, in some applications such as healthcare, randomization methods are not sufficient. It has been shown that by incorporating certain publicly available data, the hidden sensitive information in the processed records can be recovered and hence the privacy will be compromised. The k-anonymity procedure [34] was proposed as a solution for this indirect de-identification issue. By using generalization and suppression techniques, the k-anonymity method guarantees that any attribute value can be indistinguishably backtracked to at least k records. In [7], k-Optimize algorithm has been developed to efficiently solve the problem of optimal k-anonymization in most cases.

The third category, which is the distributed privacy preservation, is closely related to secure distributed computation which is a research topic in cryptography [30]. In this scenario, non-fully trusting collaborators can jointly compute useful aggregate statistics over the entire dataset without sharing the original data. Thus the privacy of the individual dataset can be protected from various other participators. Based on the partitioning mechanism, distributed privacy-preserving data analysis can be grouped into two sub-categories: vertically partitioned distributed privacy preservation and horizontally partitioned distributed privacy preservation [10]. In the vertically partitioned distributed privacy preservation problem, each participant has a different set of attributes for the same group of data objects and the primary goal is to build a robust model that can leverage the entire set of attributes without revealing the individual details of the features to other participators. More detailed work about this approach can be found in [13]. In the horizontally partitioned distributed privacy preservation, each participator has different subset of records with the same set of attributes. The work proposed in this paper falls into this horizontally partitioned category where the patient data is being collected at multiple facilities and each patient record will consist of the same set of attributes. One of the first works in this category presented a strategy that extends the popular ID3 algorithm to two-party privacy preservation [23]. Then, there were a number of approaches proposed for horizontally partitioned distributed privacy preservation using various machine learning algorithms such as naive Bayes classifier [21], support vector machine [40], and ensemble based classifiers [16]. Among these methods, the ensemble-based approaches have demonstrated good performance in terms of simultaneously preserving privacy and having good accuracy.

Gambs et al. [16] proposed MultBoost algorithm which is a boosting-based privacy-preserving data mining method. However, the main problem with these above-mentioned models is that they need a third-party (central agent) that participates in the protocol to revise and aggregate the integrated final models. This also leads to an extra cost in model implementation. In addition, as each participator must have frequent contact with the third-party, a significant portion of the time is spent on the communication between the participators rather than the actual computation itself. Most importantly, those methods are not adaptable, i.e., a new integrated model has to be learned from scratch by repeating the entire learning process every time when a new participator is added. In this paper, we propose a distributed ensemble based horizontally partitioned privacy preservation algorithm which overcomes these limitations by making each participator compute their own local models independently from each other. Therefore, each participator can take advantage of the aggregate statistics (or a basic prediction model) without compromising their own privacy. Also the communication cost is substantially reduced because a third-party is not needed in our model, and the participators will have to only recompute the models for the newly integrated components once a new participator is added.

2.2. Type-2 diabetes

Diabetes is the 7th leading cause of death in the United States [17] and the 8th leading cause of death in the world [39]. Around 29.1 million Americans (or 9.3% of the U.S. population) have diabetes [9]. In adults, type-2 diabetes (or insulin-dependent diabetes mellitus) accounts for approximately 90–95% of all diagnosed cases of diabetes. The risk for developing type-2 diabetes is mostly associated with older age, obesity, family history of diabetes, history of gestational diabetes, impaired glucose metabolism, physical inactivity, and race/ethnicity (African American, Alaska Native, American Indian, Asian American, Hispanic/Latino, or Pacific Islander American have a high prevalence of diabetes) [5,9,26].

There are several medical tests that can be performed to check if a person has diabetes or not. Currently, the American Diabetes Association (ADA) recommends four kinds of testing methods for diagnosing type-2 diabetes: (1) A1C test [27] which measures a person's average levels of blood glucose over a period of 3 months. (2) Fasting Plasma Glucose (FPG) test [28] which measures the blood glucose in a person who has fasted for more than 8 h. (3) Oral Glucose Tolerance Test (OGTT) [25] measures the blood glucose after a person fasts for at least 8 h and 2 h after the person drinks a liquid containing 75 g of glucose dissolved in water. (4) Random Plasma Glucose (RPG) test [28] which is performed during a regular health checkup.

Though these medical diagnoses tests are relatively accurate (patients really have diabetes when these tests are positive), they typically have poor sensitivity (will miss a lot of cases). In 2012, 8.1 million (27.8% of 29.1 million) people with diabetes were undiagnosed [9]. In addition, in the data that we used in our study, only 1/8th of the patients suffering from diabetes were diagnosed by those medical tests. In addition, such tests cause inconvenience to the patients. Hence, it is an important problem to diagnose these patients at early stages through their electronic health records (EHR) using data-driven methods.

To improve the diagnosis of type-2 diabetes, a machine learning community hosted a "Practice Fusion Diabetes Classification" challenge [20] in 2012 which aims at developing some advanced EHR based type-2 diabetes diagnostic support system. There were a total of 9948 patients, and among them 1904 patients suffer from diabetes. These patients come from all the 50 states and the District of Columbia in the U.S. and the Commonwealth of Puerto Rico. For each patient, the EHRs were comprehensively collected from the following primary sources of information [22]:

- **Demographic information** such as year of birth, gender, weight, and geographical location of each patient.
- **Diagnosis information** consists of the ICD9 Codes.
- **Allergy and immunization** consists of a list of allergies and vaccination records.
- **Laboratory information** consists of lab test observations for lab panels, and the lab test results received from lab facilities.
- **Medication and prescription** consists of the medication history, where each medicine is identified by National Drug Code (NDC), and prescription records.
- **Patient smoking status** is an ordinal status variable maintained on a yearly basis.
- **Transcripts** consist of visited document records including allergy, medication, and diagnosis information.

We integrated all the information from 17 different sources and constructed a consolidated repository of diabetes EHR which contained 536 variables in total after combining all the data from multiple sources. It should be noted that, among all these features, only the gender and location are categorical in nature; the lab test results and personal information are real-valued attributes; certain status attributes such as smoking status, and emergency status have ordinal values. A significant majority of the remaining features are count variables.

3. Proposed work

In this section, we will explain the proposed framework for privacy-preserving data mining. Before describing the details of our proposed distributed ensemble based approach, we will first review the working of the standard AdaBoost algorithm [15]. The AdaBoost algorithm will be used as our local learner for each participator along with the decision stump which will be used as the weak learner. In our approach, we choose the AdaBoost algorithm because it is a simple, flexible and effective classifier, and it is constructed using a weighted combination of weak classifiers. Before going to the details of the proposed methods, we will first introduce the notations used in this paper in Table 1.

AdaBoost is a well-known high-performance ensemble learning method which iteratively generates a strong classifier from a pool of weak hypotheses. In each iteration, a base classifier is learned, and based on the correctness of the prediction in each iteration, the weights of the training examples are updated. Thus, in the subsequent iteration, the base classifiers will focus on the misclassified samples in the previous iteration. The final ensemble classifier is a weighted combination of these base classifiers, where for each base classifier, the weight depends on the corresponding error rate. In other words, a classifier with lower error rate gets higher weight. Simple learners such as decision stumps (decision trees with only two leaf nodes) often perform well for AdaBoost [8]. Typically, a decision stump can be expressed using three parameters: (i) the name of the attribute to be tested (fn), (ii) the test threshold (θ), and (iii) the sign of the test ($\delta \in \{+1, -1\}$). Note that a decision stump can only handle one attribute (feature) at a time and hence the weights of weak classifiers can also be used to measure the ability to distinguish the corresponding features conditional on the learning task. The final boosted model will be a linear combination of these independent decision stumps, and this independence would not require each participator to have exactly the same set of features which is another primary advantage of our proposed model. Finally, this will allow us to select the common important features among all participators, and each participator can take advantage of the useful aggregated model statistics without sharing the original data.

Table 1
Notations used in this paper.

Notation	Description
$n^{(p)}$	Sampling size of pth participator
$D^{p}_{n^{(p)}}$	Dataset of pth participator
X	Data matrix
Y	Corresponding labels $\in \{+1, -1\}$
fn	Name of the attribute
θ	Test threshold
δ	Sign of the test $\in \{+1, -1\}$
$h^{p(t)}(\cdot)$	The weak classifier trained at tth iteration in pth participator
$\alpha^{p(t)}$	The weight of $h^{p(t)}(\cdot)$
$H^{p}(\cdot)$	The classifier trained in the pth participator
ϵ_{p}	Training error rate of pth participator
$\epsilon_{p}^{(q)}$	Testing error rate of qth model in the pth participator's training data
$S^{(p)}$	The index set of selected models for pth participator
λ_{p}	Instance proportion of pth participator
σ	Factor of the integrated weight

3.1. Adaptive distributed privacy preserving using boosting

In this section, we will discuss the proposed boosting based adaptive model for privacy-preserving data mining with horizontally partitioned data. In this case, each participant has different set of records with both common features and local unique attributes. Let $D^{1}_{n^{(1)}}, D^{2}_{n^{(2)}}, \ldots, D^{M}_{n^{(M)}}$ denote the datasets of M participators, the dataset of pth participator contains a total of $n^{(p)}$ samples, and it can be represented as

$$D^{p}_{n^{(p)}} = \left\{ \left(X^{p}_{1}, y^{p}_{1}\right), \left(X^{p}_{2}, y^{p}_{2}\right), \ldots, \left(X^{p}_{n^{(p)}}, y^{p}_{n^{(p)}}\right) \right\},$$

where each example $X^{p}_{i} = \left(x^{p}_{i1}, x^{p}_{i2}, \ldots, x^{p}_{iK^{(p)}}\right)$ is a covariate vector with $K^{(p)}$ components and each label $y^{p}_{i} \in \{+1, -1\}$ for the binary decision making that is being considered here. In our model, we employ AdaBoost algorithm to build an ensemble classifier for each participator, and these ensemble classifiers can be used to represent the data distribution of each participator. By sharing these local models with each other, all the participators can build their individual integrated model which takes advantage of other participators' knowledge without direct access to the datasets. Hence, the privacy of the individual datasets can be preserved.

Firstly, the standard AdaBoost algorithm is applied to each participator for T boosting iterations, and set of T weak classifiers will be learned. In the decision stump case, each weak classifier can be represented by the variable set (fn, θ, δ), where fn denotes the name of the selected attribute, θ is the decision threshold of the attribute fn, and δ is the sign of the decision. For any instance X_i, the hypothesis $h(X_i)$, which is made by decision stump, is either $+1$ or -1. For the pth participator, the ensemble classifier H^p $(\cdot)$ is a set of T weak classifiers:

$$\{h^{p(1)}(\cdot)\alpha^{p(1)}, h^{p(2)}(\cdot)\alpha^{p(2)}, \ldots, h^{p(T)}(\cdot)\alpha^{p(T)}\},$$

where $h^{p(t)}(\cdot)$ is the weak classifier at tth iteration and $\alpha^{p(t)}$ is the corresponding weight of that weak classifier. For a particular test instance X_i, the binary prediction made by the pth participator's local model can be defined as

$$\mathcal{H}^{p}(X) = \text{sign}\left\{H^{p}(X)\right\} = \text{sign}\left\{\sum_{t=0}^{T} \alpha^{p(t)} h^{p(t)}(X_i)\right\}, \tag{1}$$

where sign{·} is the signum function. Once the local models are built, all the participators share their own models and sample sizes with each other, and hence each participator will receive $M-1$ models from other participators. Then, each participator will build an integrated model independently based on their specific requirements. For example, some facilities might want to maximize the sensitivity of their models instead of the area under curve (AUC) value for their clinical decision support system. Using our ensemble based approach, they can use their own cost matrix to select the decision threshold for the integrated model that is being built in their own facility. The P2P network based distributed framework of the proposed ensemble system is shown in Fig. 1, and we can see that no third-party (central agent) is needed to build an integrated model in our proposed framework.

As each participator will build its local model independently, when a new participator is added, it will learn a local model by itself and share this trained model with other participators. Then, each existing participator will decide whether it needs to update the integrated model based on the new local model's performance on their own data. We can see that this P2P network based framework is more adaptable and can provide more autonomy for participators than the star network based distributed privacy-preserving framework. Each participator integrates the final model independently, and hence the integrated model will take care of the specific needs of their own local data. In summary, having an individualized integrated model separately at each participator will tailor the model to their own local data. Only the integration component will have to be selectively computed again when a new participator is added, rather than re-training the entire model from scratch.

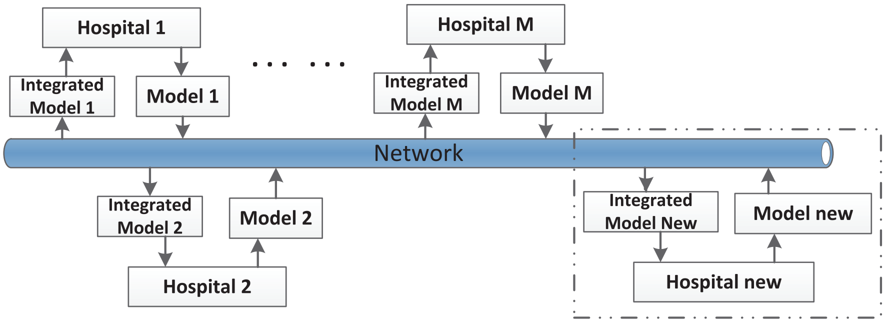


Fig. 1. Block diagram representing various participators, the local and the integrated models built at each participator.

Each participator receives $M-1$ models and another important issue that needs to be resolved for a specific participator is how to integrate those local models and its own model in an appropriate manner so that the integrated model performs better than the local model. Also, there is a slight risk that by integrating various unwanted (or poor) models, the performance of the integrated model might deteriorate. To prevent such "negative impact" during integration, when the model integration is performed for the *p*th participator, first one has to decide which of the other local models can be used to build the final integrated model.

The basic idea here is to exclude the models from the other participators whose data distribution is very different from the data distribution of this *p*th participator. Since looking into the original data distribution of the other participators is infeasible due to patient privacy, we resort to applying all the other local models on the training set of *p*th participator, $D^p_{n^{(p)}}$, and compare the error rate of each local model with the training error rate of *p*th participator's trained model. Note that the models received from the other participators might include some decision stumps for their own unique local features. Hence, before using those models, the *p*th participator should select a suitable subset of common decision stumps based on *fn*. For the *p*th participator, the error rate of *q*th participator's model is given by

$$\epsilon_p^{(q)} = \frac{1}{n^{(p)}}\left[\sum_{i=1}^{n^{(p)}} I\Big(\text{sign}\Big\{\hat{H}^q(X_i^p)\Big\} \neq y_i^p\Big)\right], \tag{2}$$

where $\hat{H}^q(\cdot)$ is the selected subset of decision stumps from $H^q(\cdot)$, the model trained by *q*th participator, $I(\cdot)$ is the indicator function. The training error rate of the *p*th participator's trained model is given by

$$\epsilon_p = \frac{1}{n^{(p)}}\left[\sum_{i=1}^{n^{(p)}} I\Big(\text{sign}\Big\{H^p(X_i^p)\Big\} \neq y_i^p\Big)\right]. \tag{3}$$

For every local model, we compute the difference between $\epsilon_p^{(q)}$ and ϵ_p. If $(\epsilon_p^{(q)} - \epsilon_p)$ is less than a certain threshold τ, then we can assume that the data distributions in *p*th participator and *q*th participator are similar, and we can use *q*th participator's trained model to build the integrated model for *p*th participator. The integrated model only uses other local models (not the other participators' data) and selects the ones with similar data distribution to that of the local data; thus, it indirectly strengthens the data distribution. By doing this, it avoids the overfitting problem in the presence of only a few samples in the local data. For our experiments, we set $\tau = 0.2$ which is relatively a large threshold. This is because ϵ_p (the training error rate of AdaBoost) is very small; thus $(\epsilon_p + 0.2)$ is a reasonable threshold for $\epsilon_p^{(q)}$ (the testing error rate of *q*th model in the *p*th participator's training data) as the *q*th model is not trained from the training data of *p*th participator.

Now that the model selection is complete, for *p*th participator we get an index set of selected models ($S^{(p)}$) which will be used in the model integration. It should be noted that *p* also belongs to $S^{(p)}$. We will now explain the actual integration of the selected models. The model integration should not only take advantage of the aggregate statistics from other selected local models but also consider the specificity of the participator. To achieve this goal, we establish the following three criteria for the proposed ensemble based model integration in the *p*th participator:

1. More weight should be assigned to the participator's own local model compared to any other local model with same sample size. This is because of the fact that the data from other participators might follow a slightly different distribution and hence one cannot give more importance to the other local models. The assumption here is that by adding other local models, one can strengthen the local data distribution. This is important especially since the participator weights suffer from insufficient data.
2. The weight of the *p*th participator's own local model is positively related to $n^{(p)}$. This is because the more the number of samples, the fewer the mistakes that are caused by insufficient sampling.
3. The weight of the *p*th participator's own local model should have an upper bound. Otherwise, the weight of selected participator's model will become insignificant, and the final integrated model will be the same as *p*th participator's own model and integration becomes unnecessary.

Considering these guidelines, the weight of each participator will be updated when building the integrated model for each participator. To illustrate this concept, let $\lambda_q \sigma_q^{(p)}$ denote the weight assigned to the qth participator's model when we compute the integrated model for pth participator, where $p \in \{1, 2, \ldots, M\}$, $q \in S^{(p)}$, and $\lambda_q = \frac{n^{(q)}}{\sum_{s \in S^{(p)}} n^{(s)}}$ is the instance proportion of qth participator. $\sigma_q^{(p)}$ is updated as follows:

$$\sigma_q^{(p)} = \begin{cases} \frac{\lambda_q}{\lambda_{\max}^2} \cdot \lceil \frac{\lambda_{\max}^2}{\lambda_{\min}} \rceil & \text{if } q = p \\ 1 & \text{if } q \neq p \end{cases}, \tag{4}$$

where $\lambda_{\min} = \min(\lambda_q)$ is the minimum proportion among all the selected participators, $\lambda_{\max} = \max(\lambda_q)$ is the maximum proportion, and $\lceil \cdot \rceil$ denotes the ceiling function. We will now prove that the proposed weight update mechanism $\lambda_q \sigma_q^{(p)}$ will satisfy all the three criteria mentioned earlier.

Theorem 1. *The weight update mechanism will assign more weight to pth participator's own model compared to any other local models if they have the same sample size (criterion 1).*

Proof. If $q \neq p$ then $\sigma_q^{(p)} = 1$; that is to say the weight of other participator's local model is λ_q. Now let us consider the value of $\lambda_q \sigma_q^{(p)}$, the weight of pth participator's own model, i.e., when $q = p$. We note that the λ_q is bounded as follows. $\lambda_{\min} \leq \lambda_q \leq \lambda_{\max}$. We can see that the minimum possible value for $\sigma_p^{(p)}$ is obtained when $\lambda_p = \lambda_{\min}$. Now, this value $\sigma_p^{(p)} = \frac{\lambda_{\min}}{\lambda_{\max}^2} \cdot \lceil \frac{\lambda_{\max}^2}{\lambda_{\min}} \rceil \geq 1 \Rightarrow \lambda_p \sigma_p^{(p)} \geq \lambda_p$. Thus, if $\lambda_p = \lambda_q$, we have $\lambda_p \sigma_p^{(p)} \geq \lambda_q \sigma_q^{(p)}$. This means that for any participator, in the model integration procedure, its own model will have more weight compared to any of the other selected participators once they have the same sample size. □

Theorem 2. *The weight update mechanism will ensure that the weight of pth participator's own model is positively related to $n^{(p)}$ (criterion 2).*

Proof. For pth participator, since $S^{(p)}$ will be a constant set after model selection, $\lambda_{\min}$ and $\lambda_{\max}$ are also constant values. When $q = p$, $\lambda_p \sigma_p^{(p)} = \frac{\lambda_p^2}{\lambda_{\max}^2} \cdot \lceil \frac{\lambda_{\max}^2}{\lambda_{\min}} \rceil$ is the weight of pth participator's own model and contains only one variable $\lambda_p = \frac{n^{(p)}}{\sum_{s \in S^{(p)}} n^{(s)}}$, which is proportional to $n^{(p)}$ since the denominator is just a normalization factor. So, $\lambda_p \sigma_p^{(p)}$ is positively related to $n^{(p)}$. □

Theorem 3. *The weight update mechanism will ensure that the weight of other participator's local model will not become insignificant thus making the model integration unnecessary (criterion 3).*

Proof. When $q = p$, $\lambda_p \sigma_p^{(p)}$, is the weight of pth participator's own local model. We note that the λ_p is bounded as follows: $\lambda_{\min} \leq \lambda_p \leq \lambda_{\max}$. We observe that the maximum possible value for $\lambda_p \sigma_p^{(p)} = \lceil \frac{\lambda_{\max}^2}{\lambda_{\min}} \rceil$ is obtained when $\lambda_p = \lambda_{\max}$. This limitation ensures that the weight of other participators' high-performance weak learner will not become insignificant. □

Algorithm 1 outlines our proposed BOPPID (BOosting-based Privacy-Preserving Integration of Distributed data) algorithm. First, each participator employs standard AdaBoost algorithm to train their own local model and obtain the training error (lines 2–3). Then, they share their trained models with other participators (line 4). After the model sharing process, each participator will independently build its own integration model. In lines 7–13, each participator applies all the other local models to its own training data and selects a subset of "good" models from other participators for the subsequent integration process. The sample size proportion of the selected local models is calculated in line 14. Finally, the integration model is built using Eq. (4) (lines 16–23).

3.2. Complexity analysis

The computational complexity of proposed BOPPID depends on the AdaBoost algorithm in line 2. For pth participator, the overall cost of AdaBoost in all T iterations is $\Theta(K^{(p)}(T + \log n^{(p)}))$ if the weak learner is a decision stump, and the training error rate can be calculated in $\Theta(n^{(p)})$. In the integration procedure for each participator, the cost of selecting a subset of the beneficial local models is $\Theta(Mn^{(p)})$. In the worst case scenario, where all M participators are selected, the computational cost of updating the weights for all participators is $\Theta(M)$ and generating the final classifier $H^{q^*}(\cdot)$ takes $\Theta(MT)$ time. Thus, the total computational complexity of the pth participator is $\Theta(K^{(p)}(T + \log n^{(p)}) + MT + Mn^{(p)})$. As each participator trains its local model and does the final integration independently from other participators, it is possible to perform these computations in parallel. In addition, note that the model integration can only start after all the participators complete the computation of their own local models, so the computational complexity of the proposed model depends on the participator with the largest number of samples. Hence, the total computational cost of BOPPID is $\Theta(K^{\max}(T + \log n^{\max}) + MT + Mn^{\max})$, here $n^{\max}$ and $K^{\max}$ stand for the largest sample size and largest feature dimensions among all the M participators, respectively.

For the communication cost, let the network latency for transferring a model is f and the communication cost for sending and receiving a model is g and h respectively. For M participators, each of them will send their own model to $M - 1$ other participators, and receive $M - 1$ models from other participators. Hence, the total communication cost for BOPPID is $(M - 1)(f + g + h)$.

Algorithm 1: BOosting-based Privacy-Preserving Integration of Distributed data (BOPPID).

```
Input: The training sets of $M$ participators $\{D^1_{n^{(1)}}, \ldots, D^M_{n^{(M)}}\}$, Number of boosting iterations ($T$), and
Threshold for error rate increase ($\tau$).
Output: The set of $M$ final classifiers $\{H^{1^*}(\cdot), \ldots, H^{M^*}(\cdot)\}$.
foreach p in M participators parallel do
    $H^p(\cdot) \leftarrow$ AdaBoost($D^p_{n^{(p)}}, T$);
    $\epsilon_p \leftarrow \frac{1}{n^{(p)}}\left[\sum_{i=1}^{n^{(p)}} I\left(\mathrm{sign}\left\{H^p(X_i^p)\right\} \neq y_i^p\right)\right]$;
    Share $H^p(\cdot)$ with all other participators
end
foreach p in M participators parallel do
    Initialize the Index set of selected participators $S \leftarrow \{p\}$;
    for $q \leftarrow 1$ to $M$ do
        $\epsilon_p^{(q)} \leftarrow \frac{1}{n^{(p)}}\left[\sum_{i=1}^{n^{(p)}} I\left(\mathrm{sign}\left\{\hat{H}^q(X_i^p)\right\} \neq y_i^p\right)\right]$;
        if $(\epsilon_p^{(q)} - \epsilon_p) \leq \tau$ then
            Add $q$ in $S^{(p)}$
        end
    end
    $\Lambda = \left\{\lambda_q = \frac{n^{(q)}}{\sum_{s\in S} n^{(s)}} | q \in S\right\}$;
    Find $\lambda_{\max}, \lambda_{\min}$;
    foreach q in S do
        if $p = q$ then
            $\sigma_q^{(p)} = \frac{\lambda_q}{\lambda_{\max}^2} \cdot \lceil \frac{\lambda_{\max}^2}{\lambda_{\min}} \rceil$;
        else
            $\sigma_q^{(p)} = 1$;
        end
    end
    return $H^{p^*}(\cdot) = \sum_{q\in S}\sum_{t=1}^{T} h^{q(t)}(\cdot) \cdot \alpha^{q(t)}\sigma_q^{(p)}\lambda_q$
end
```

Now let us consider the scenario where a new participator is added, and the number of participators grows from M to $M+1$. In our model, only the integration part has to be recomputed, and hence the computational cost for the existing M participators is $\Theta((M+1)T + (M+1)n^{max})$, and for the new participator the computation cost is $\Theta(K^{\max}(T + \log n^{\text{new}}) + MT + Mn^{\max})$. Here n^{new} and $K^{\max}$ correspond to the sample size and feature dimension of the new participator, respectively. Because all of the existing M participators will send their own models to the new participator and receive a new model from the new participator, the communication cost for the existing M participators is $(f+g+h)$ and for the new participator, the communication cost is $M(f+g+h)$. However, if a new participator is added in the existing methods such as MultBoost, a new integrated model has to be retrained from the beginning. Therefore, the computation cost for all $M+1$ participators is $\Theta(K^{(p)}(T + \log(\sum_{p=1}^{M+1} n^{(p)})) + MT)$ [16] and communication cost becomes $2T(f+g+h)$. Hence, we can clearly observe that our proposed model is more efficient when a new participator is added (which is a more practical scenario).

3.3. Adaptive variant of parallel boosting

To strengthen our work, we also propose a new variant of an existing parallel boosting algorithm and make it applicable to the context of mining EHR data. In our previous work on parallel boosting [29], we developed a parallel variant of the boosting algorithm called AdaBoost.PL. Though originally designed for improving the computational efficiency, the overall layout of AdaBoost.PL can fit into the distributed privacy-preserving setting as it does not directly communicate the data from one participator to the others. The AdaBoost.PL algorithm follows the *MapReduce* workflows, where the original dataset is equally partitioned into M parts and mapped into M workers to learn M models which will then be reduced (integrated) into a final model. For each worker, its weak classifiers will be sorted with increasing order of $\alpha^{p(t)}$ values (line 3 in Algorithm 2). The basic intuition of sorting the workers' weak classifiers with respect to their weights is to place the classifiers with similar correctness at the same sorted level. This is a critical component of the AdaBoost.PL since this will ensure that like-minded classifiers will be merged during each boosting iteration. After sorting, $h^{p^*(t)}(\cdot)$ is the weak learner of pth participator at the tth iteration, and $\alpha^{p^*(t)}$ denotes the corresponding weight. In this paper, we generalize the idea of AdaBoost.PL and propose a new variant, AdaBoost.PL.V2, which can handle the diversity of sample size of each participator in the privacy-preserving context. It should be

Algorithm 2: AdaBoost.PL.V2.

```
Input: Training sets of M participators $\{D^1_{n(1)}, D^2_{n(2)}, \ldots, D^M_{n(M)}\}$, and
Number of boosting iterations ($T$).
Output: The final classifiers $H(\cdot)$

1  foreach p in M participators parallel do
2      $H^p(\cdot) \leftarrow \mathrm{AdaBoost}(D^p_{n(p)}, T)$;
3      $H^{p^*}(\cdot) \leftarrow$ the weak classifiers in $H^p(\cdot)$ sorted w.r.t. $\alpha^{p(t)}$;
4      Send $H^{p^*}(\cdot)$ to central agent;
5  end
6  Initialize $\Lambda = \left\{\lambda_p = \frac{n^{(p)}}{\sum_{m=1}^{M} n^{(m)}} \,|\, p = 1, 2, \ldots, M\right\}$;
7  for $t \leftarrow 1$ to $T$ do
8      $h^{(t)}(\cdot) \leftarrow \mathrm{Merge}\big(h^{1^*(t)}(\cdot), \ldots, h^{M^*(t)}(\cdot)\big)$;
9      $\alpha^t \leftarrow \sum_{p=1}^{M} \alpha^{p^*(t)} \lambda_p$;
10 end
11 return $H(\cdot) = \sum_{t=1}^{T} \alpha^t h^{(t)}(\cdot)$
```

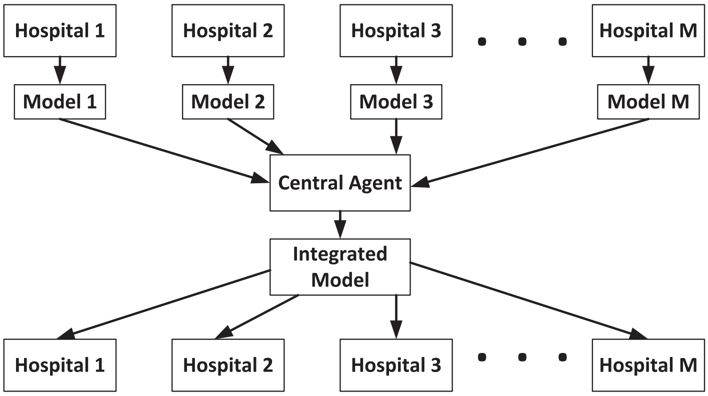


Fig. 2. The framework of star network based horizontal distributed privacy-preserving method.

noted that similar to AdaBoost.PL, the proposed AdaBoost.PL.V2 is also built on a star network, which is shown in Fig. 2. Most of state-of-the-art horizontally distributed privacy-preserving methods are also built on this star network, so that we can see that there is only one integration model that is being built for all hospitals.

Thus, compared with BOPPID, AdaBoost.PL.V2 is less adaptive and cannot preserve the data characteristics of each participator. AdaBoost.PL.V2 discards the “map” part of the AdaBoost.PL because the datasets are distributed and stored by each participator, and α^t is the weighted average (instead of the standard average used in AdaBoost.PL) of $\alpha^{p^*(t)}$ for all possible p. The weights are based on the sample size proportion (line 9 in Algorithm 2).

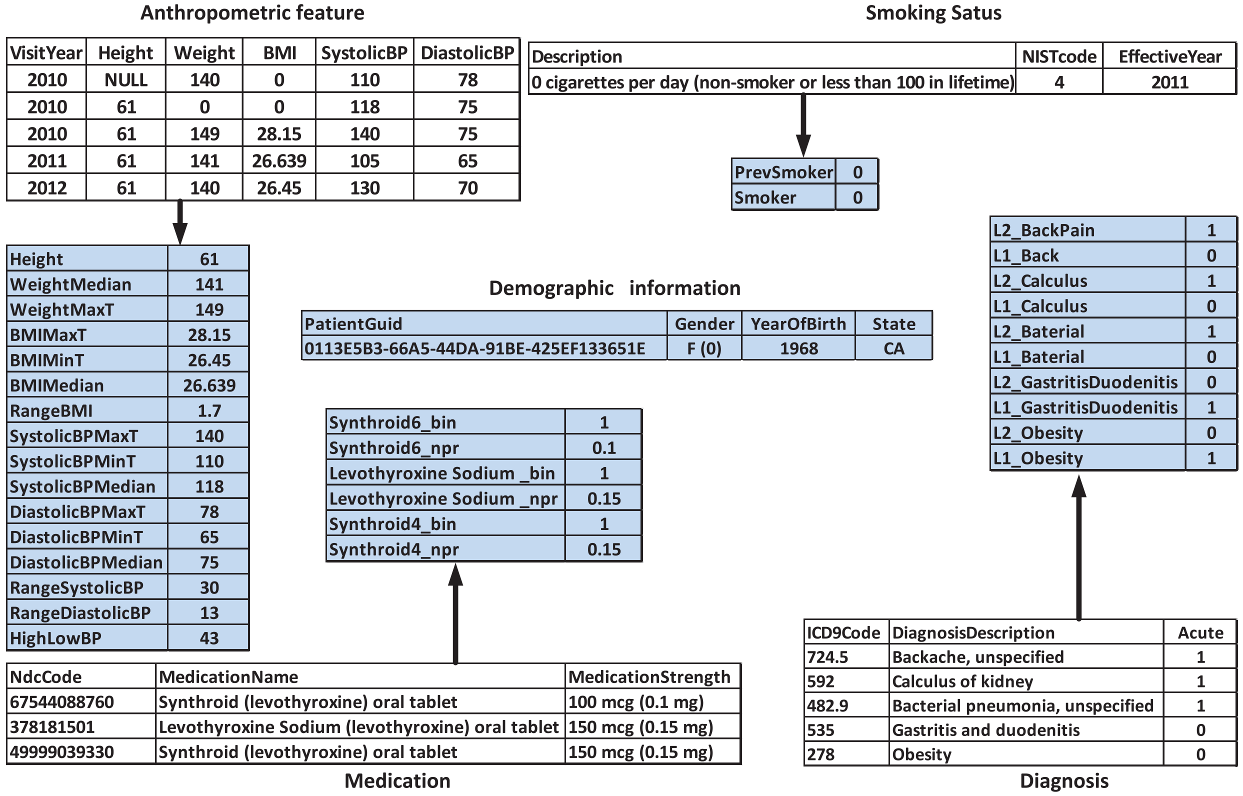
The merged classifier, $h^{(t)}(\cdot)$ is a *ternary classifier*, a variant of weak classifier, which can return ‘+1’, ‘−1’, and ‘0’ as a way of abstaining from answering [35]. It is built by taking a simple majority vote among weak classifiers of the worker as follows:

$$h^{(t)}(X_i) = \begin{cases} \mathrm{sign}\big(\sum_{p=1}^{M} h^{p^*(t)}(X_i)\big) & \text{if } \sum_{p=1}^{M} h^{p^*(t)}(X_i) \neq 0 \\ 0 & \text{otherwise} \end{cases}. \tag{5}$$

The ternary classifier will answer ‘0’ if equal number of positive and negative predictions are made by the workers’ weak classifiers. Otherwise, it will answer the majority prediction. In line 8, the weight of the ternary classifier is obtained by the weighted average of the corresponding classifier weights. Once all the ternary classifiers for T rounds are generated, the algorithm returns their weighted combination as the final classifier.

4. Experimental results

In this section, we demonstrate the performance of the proposed approach using real-world EHRs of diabetes patients. We first present the clinical feature transformation performed on the EHR data. We will then compare the performance of our proposed model against the state-of-the-art distributed privacy preserving ensemble prediction models. In addition, we also perform a detailed study on the biomarkers selected by the proposed algorithm on this data. It shows that our proposed model



Anthropometric feature

VisitYear	Height	Weight	BMI	SystolicBP	DiastolicBP
2010	NULL	140	0	110	78
2010	61	0	0	118	75
2010	61	149	28.15	140	75
2011	61	141	26.639	105	65
2012	61	140	26.45	130	70

Height	61
WeightMedian	141
WeightMaxT	149
BMIMaxT	28.15
BMIMinT	26.45
BMIMedian	26.639
RangeBMI	1.7
SystolicBPMaxT	140
SystolicBPMinT	110
SystolicBPMedian	118
DiastolicBPMaxT	78
DiastolicBPMinT	65
DiastolicBPMedian	75
RangeSystolicBP	30
RangeDiastolicBP	13
HighLowBP	43

Smoking Satus

Description	NISTcode	EffectiveYear
0 cigarettes per day (non-smoker or less than 100 in lifetime)	4	2011

PrevSmoker	0
Smoker	0

Demographic information

PatientGuid	Gender	YearOfBirth	State
0113E5B3-66A5-44DA-91BE-425EF133651E	F (0)	1968	CA

Synthroid6_bin	1
Synthroid6_npr	0.1
Levothyroxine Sodium _bin	1
Levothyroxine Sodium _npr	0.15
Synthroid4_bin	1
Synthroid4_npr	0.15

NdcCode	MedicationName	MedicationStrength
67544088760	Synthroid (levothyroxine) oral tablet	100 mcg (0.1 mg)
378181501	Levothyroxine Sodium (levothyroxine) oral tablet	150 mcg (0.15 mg)
49999039330	Synthroid (levothyroxine) oral tablet	150 mcg (0.15 mg)

Medication

L2_BackPain	1
L1_Back	0
L2_Calculus	1
L1_Calculus	0
L2_Baterial	1
L1_Baterial	0
L2_GastritisDuodenitis	0
L1_GastritisDuodenitis	1
L2_Obesity	0
L1_Obesity	1

ICD9Code	DiagnosisDescription	Acute
724.5	Backache, unspecified	1
592	Calculus of kidney	1
482.9	Bacterial pneumonia, unspecified	1
535	Gastritis and duodenitis	0
278	Obesity	0

Diagnosis

Fig. 3. Clinical feature transformation for a single patient.

can successfully select some important ‘universal’ and ‘region-specific’ biomarkers which can be used in the medical domain to improve the diagnosis of type-2 diabetes.

4.1. *Clinical feature pre-processing*

We now explain the clinical feature transformation procedure which transforms the raw EHR data to a format that is more suitable for data analysis. The original data contains patient records for a total of 9948 patients, and each patient can be identified using a specific serial number. In Fig. 3, we show the feature creation procedure from EHR data by considering a sample patient. In this example, we use several records from different categories for a particular patient (with PatientGuid “0113E5B3-66A5-44DA-91BE-425EF133651E”).

We extracted all the features for distinct anthropometric variables present in the data. In this example, we consider height, weight, body mass index (BMI), systolic blood pressure (SystolicBP), and diastolic blood pressure (DiastolicBP). To tackle the problem of multiple anthropometric values for the same patient, we represent each anthropometric feature using a set of summary statistics (maximum, minimum, and median). Note that, in this example, some anthropometric features are missing. It should be noted that certain missing values (like height in this example) are easy to fill. For example, we can confidently replace ‘NULL’ by 61 and calculate the corresponding BMI by using $\frac{\text{Weight (lb)}}{(\text{Height (in.)})^2} \times 703$. However, other missing values like weight cannot be filled-in, and hence we leave them empty. “HighLowBP” is the difference between the median of systolic blood pressure and the median of diastolic blood pressure. Lab variables are also generated using a similar approach. In addition, we also created a variable which counts the number of times the lab was conducted for the patient. ICD-9 code is a list code for International Statistical Classification of Diseases, and each code presents a disease description. With the acute indicator, two binary variables can be created to reflect whether the patient has a special disease or not, and if the disease is present, whether it is acute or not (“L2_” represents acute and “L1_” represents not acute). In this example, the codes 278 and 724.5 represent obesity and backache, respectively. Because this patient suffered from serious backache only once, we give the value of 1 to “L2_ BackPain” feature. For the medication information, we created two variables for each distinct medication (according to NDC) given to the patient; one feature represents the number of times the individual medication was given to the patient, and the other feature represents the dose. There are several states corresponding to smoking status which can be represented using ordinal numbers, and we build two new features, one for the current smoking state and another for the smoking history. When the patient has never smoked in the past, the status of “0 cigarettes per day” is denoted by 0.

Table 2
Demographic statistics of the top 14 states based on patient population.

State	Total no. of patients	No. of diabetes patients	No. of non-diabetes patients	Prevalence rate (%)
Arizona (AZ)	295	56	239	18.98
California (CA)	1917	258	1659	13.46
Florida (FL)	804	158	646	19.65
Illinois (IL)	412	75	337	18.20
Michigan (MI)	344	61	283	17.73
Missouri (MO)	702	104	598	14.81
New Jersey (NJ)	575	119	456	20.70
Nevada (NV)	495	135	360	27.27
New York (NY)	557	167	390	29.98
Ohio (OH)	515	111	404	21.55
Pennsylvania (PA)	250	67	183	26.80
South Dakota (SD)	325	40	285	12.31
Texas (TX)	897	243	654	27.09
Virginia (VA)	484	79	405	16.32

Table 3
Performance comparison of different algorithms based on the AUC values (along with their standard deviation).

State	LOCAL_Ada	MultBoost	AdaBoost.PL.V2	BOPPID
AZ	0.799 ± 0.103	0.831 ± 0.085	0.845 ± 0.096	**0.865 ± 0.088**
CA	0.895 ± 0.046	0.844 ± 0.060	0.870 ± 0.064	**0.902 ± 0.049**
FL	0.882 ± 0.065	0.861 ± 0.059	0.878 ± 0.056	**0.899 ± 0.053**
IL	0.849 ± 0.071	0.876 ± 0.076	0.885 ± 0.053	**0.896 ± 0.046**
MI	0.881 ± 0.065	0.926 ± 0.038	0.923 ± 0.037	**0.934 ± 0.038**
MO	0.914 ± 0.034	0.886 ± 0.052	0.902 ± 0.048	**0.921 ± 0.040**
NJ	0.802 ± 0.040	0.835 ± 0.051	0.847 ± 0.042	**0.857 ± 0.048**
NV	0.863 ± 0.045	0.877 ± 0.033	0.891 ± 0.029	**0.901 ± 0.019**
NY	0.858 ± 0.060	0.857 ± 0.038	0.873 ± 0.061	**0.891 ± 0.056**
OH	0.885 ± 0.041	0.879 ± 0.066	0.892 ± 0.053	**0.900 ± 0.050**
PA	0.816 ± 0.082	0.809 ± 0.117	0.873 ± 0.071	**0.885 ± 0.071**
SD	0.892 ± 0.072	0.920 ± 0.067	0.929 ± 0.056	**0.937 ± 0.047**
TX	0.854 ± 0.040	0.838 ± 0.038	0.845 ± 0.027	**0.871 ± 0.027**
VA	0.789 ± 0.056	0.827 ± 0.045	0.838 ± 0.072	**0.863 ± 0.059**

In summary, based on the generated features which are listed in tables with pale blue color, one can observe that following our transformation procedure not only helps in immensely reducing the dimensionality and complexity of the raw EHR data, but also summarizes the complex EHRs into a succinct representation which is then used for building prediction models.

4.2. Experimental setup

We will demonstrate the performance of the proposed algorithms on improving the diagnosis of the diabetes condition. Hence, for our experiments, we selected the top 14 states based on the total patient population in each state (which will be considered as a participator). Table 2 shows the demographic statistics of these 14 selected states.

We compare our proposed adaptive distributed privacy-preserving ensemble system with the MultBoost [16] and local AdaBoost (LOCAL_Ada) algorithms. LOCAL_Ada algorithm is a local model obtained by applying the standard AdaBoost on each participator independently. For all the algorithms, the number of boosting iterations is set to 100. In Table 3, we provide the performance results of AUC values from different algorithms using 10-fold cross validation. The best results are being highlighted in bold.

In Table 4, we present the comparison of *F*-measure values which is calculated as follows:

$$F\_\text{measure} = \frac{2 \times \text{Sensitivity} \times \text{Precision}}{\text{Sensitivity} + \text{Precision}}, \tag{6}$$

where Sensitivity $= \frac{\text{TP}}{\text{TP+FN}}$, and Precision $= \frac{\text{TP}}{\text{TP+FP}}$; therefore, a high value of *F*-measure indicates that both precision and sensitivity are reasonably high. It should be noted that the positive individuals are diabetics and the negative individuals are non-diabetics. Thus, here TP corresponds to the number of diabetics correctly predicted by the classifier, FN corresponds to the number of diabetics wrongly predicted as non-diabetics, and FP is the number of nondiabetics wrongly predicted as diabetics.

We observe that for all the 14 states, our proposed algorithm can diagnose type-2 diabetes more accurately compared with the other algorithms. This demonstrates that we have successfully transferred useful medical knowledge and clinical information within all the participators under the privacy constraints without revealing the data itself from each participator. This also means that those participators can take advantage of the aggregated global models without compromising the privacy of individual patients.

Table 4
Performance comparison of different algorithms based on the *F*-measure values (along with their standard deviation).

State	LOCAL_Ada	MultBoost	AdaBoost.PL.V2	BOPPID
AZ	0.541 ± 0.141	0.592 ± 0.084	0.612 ± 0.148	**0.631 ± 0.151**
CA	0.592 ± 0.092	0.498 ± 0.049	0.562 ± 0.079	**0.608 ± 0.090**
FL	0.664 ± 0.098	0.615 ± 0.085	0.643 ± 0.083	**0.679 ± 0.092**
IL	0.572 ± 0.102	0.643 ± 0.136	0.630 ± 0.110	**0.659 ± 0.088**
MI	0.619 ± 0.110	0.708 ± 0.087	0.707 ± 0.087	**0.720 ± 0.091**
MO	0.654 ± 0.069	0.608 ± 0.097	0.646 ± 0.094	**0.667 ± 0.103**
NJ	0.534 ± 0.059	0.583 ± 0.058	0.585 ± 0.056	**0.589 ± 0.063**
NV	0.668 ± 0.049	0.691 ± 0.058	0.705 ± 0.058	**0.715 ± 0.043**
NY	0.699 ± 0.086	0.682 ± 0.054	0.709 ± 0.099	**0.731 ± 0.093**
OH	0.657 ± 0.074	0.639 ± 0.104	0.688 ± 0.080	**0.701 ± 0.084**
PA	0.637 ± 0.090	0.623 ± 0.128	0.693 ± 0.092	**0.721 ± 0.112**
SD	0.666 ± 0.190	0.696 ± 0.159	0.691 ± 0.140	**0.732 ± 0.107**
TX	0.661 ± 0.065	0.628 ± 0.050	0.639 ± 0.035	**0.675 ± 0.042**
VA	0.493 ± 0.088	0.512 ± 0.075	0.560 ± 0.096	**0.566 ± 0.079**

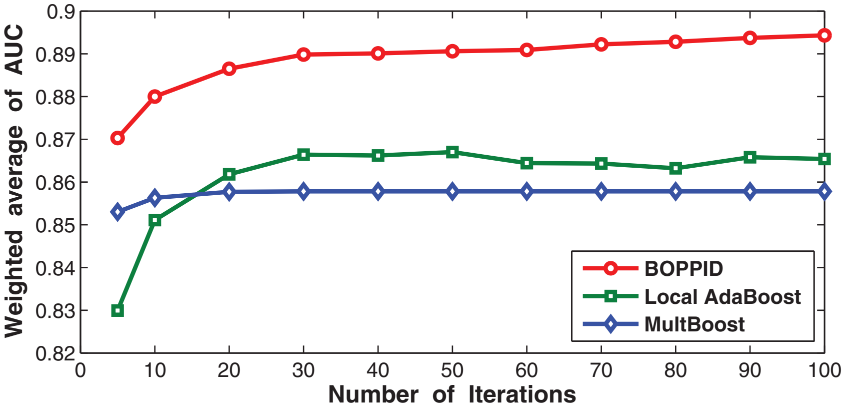


Fig. 4. Comparison of weighted average AUC values of Local AdaBoost, MultBoost, and BOPPID algorithms on the diabetes dataset for 100 iterations.

Fig. 4 shows the 10-fold cross validation results of weighted average AUC values at different iterations of the algorithm. The weighted average AUC is defined as

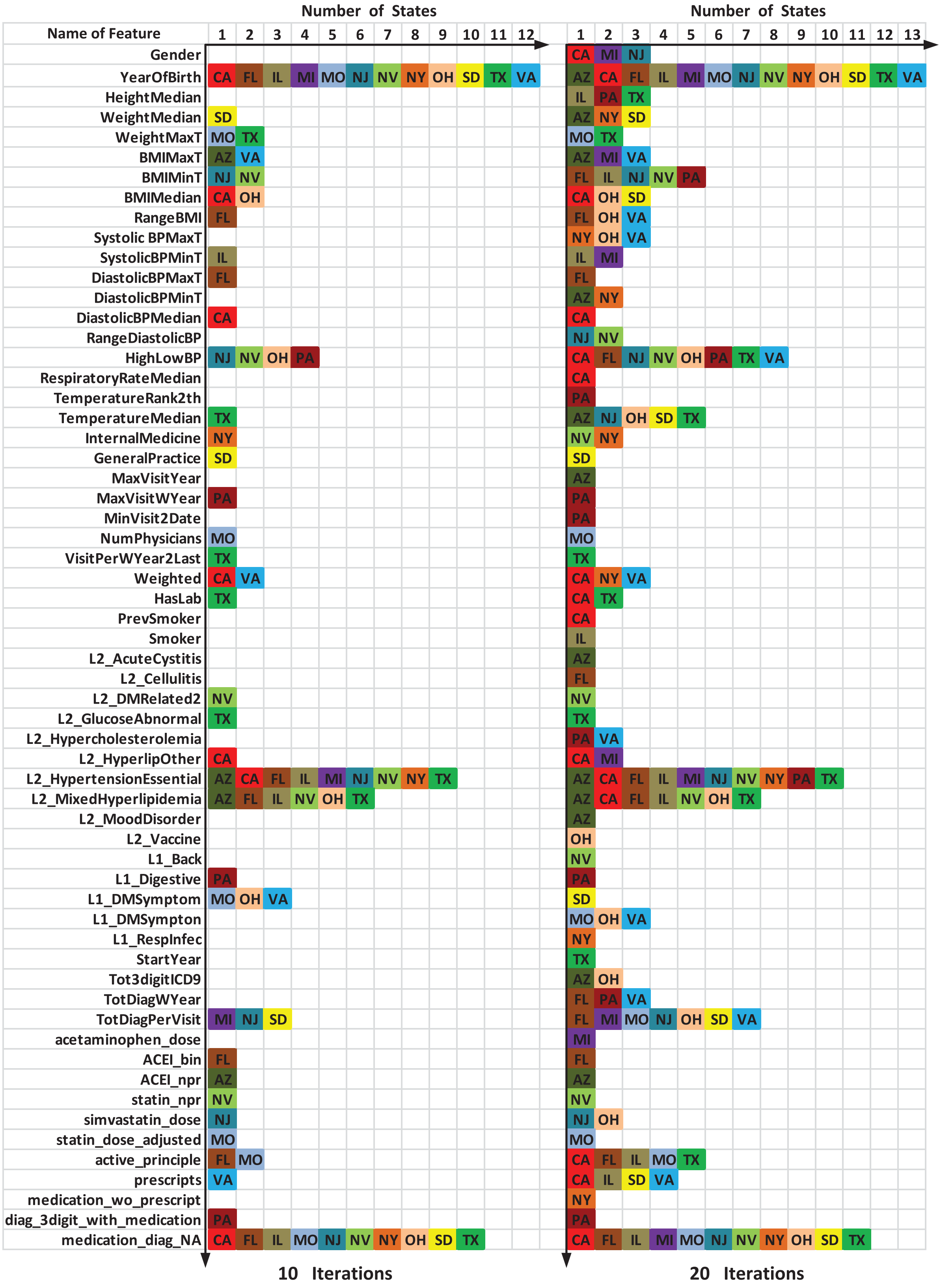
$$\mathrm{AUC}_{\mathrm{avg}} = \sum_{p=1}^{M} \mathrm{AUC}^{(p)} \frac{n^{(p)}}{\sum_{m=1}^{M} n^{(m)}}, \tag{7}$$

where $\mathrm{AUC}^{(p)}$ is the AUC value of pth participator, $n^{(p)}$ is the number of patients in the pth participator. The figure shows that the proposed algorithm always works better than MultBoost and Local AdaBoost, and only a few selected important features (or iterations) can provide a good prediction model.

4.3. Biomarker discovery

Biomarkers are important indicators used to diagnose a particular disease in a clinical setting. From Fig. 4, we can see that the weighted average AUC of our proposed model achieves 0.88 after 10 iterations and approximately 0.89 after 20 iterations. This indicates that the top selected features in each state can well represent the corresponding data distribution. It should be mentioned that since the weak classifier used in our algorithm is a decision stump, each iteration will correspond to the usage of only one feature with a simple split. In Fig. 5, we show the selected biomarkers for each state after 10 iterations (in the left) and 20 iterations (in the right). In the 10 iterations scenario, we find 39 important biomarkers, and in the 20 iterations case, this number becomes 60. For comparison, we present these two settings under a same coordinate system, where the vertical axis is the index of selected features and the horizontal axis corresponds to the list of states that contain this attribute in their top feature list. Fig. 5 clearly shows the selected important features and the corresponding state.

From Fig. 5, we can conclude that there are a total of 6 universal biomarkers: “YearOfBirth” which represents the age of patient, “HighLowBP” measures the difference between systolic blood pressure and diastolic blood pressure, “L2_HypertensionEssential” is a biomarker that reflects whether the patient has been diagnosed with hypertension or not, “L2_MixedHyperlipidemia” is an indicator that tells whether the patient has suffered with mixed hyperlipidemia or not, “TotDiagPerVisit” indicates the number



	Number of States												Number of States												
Name of Feature	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	13
Gender													CA	MI	NJ										
YearOfBirth	CA	FL	IL	MI	MO	NJ	NV	NY	OH	SD	TX	VA	AZ	CA	FL	IL	MI	MO	NJ	NV	NY	OH	SD	TX	VA
HeightMedian													IL	PA	TX										
WeightMedian	SD												AZ	NY	SD										
WeightMaxT	MO	TX											MO	TX											
BMIMaxT	AZ	VA											AZ	MI	VA										
BMIMinT	NJ	NV											FL	IL	NJ	NV	PA								
BMIMedian	CA	OH											CA	OH	SD										
RangeBMI	FL												FL	OH	VA										
Systolic BPMaxT													NY	OH	VA										
SystolicBPMinT	IL												IL	MI											
DiastolicBPMaxT	FL												FL												
DiastolicBPMinT													AZ	NY											
DiastolicBPMedian	CA												CA												
RangeDiastolicBP													NJ	NV											
HighLowBP	NJ	NV	OH	PA									CA	FL	NJ	NV	OH	PA	TX	VA					
RespiratoryRateMedian													CA												
TemperatureRank2th													PA												
TemperatureMedian	TX												AZ	NJ	OH	SD	TX								
InternalMedicine	NY												NV	NY											
GeneralPractice	SD												SD												
MaxVisitYear													AZ												
MaxVisitWYear	PA												PA												
MinVisit2Date													PA												
NumPhysicians	MO												MO												
VisitPerWYear2Last	TX												TX												
Weighted	CA	VA											CA	NY	VA										
HasLab	TX												CA	TX											
PrevSmoker													CA												
Smoker													IL												
L2_AcuteCystitis													AZ												
L2_Cellulitis													FL												
L2_DMRelated2	NV												NV												
L2_GlucoseAbnormal	TX												TX												
L2_Hypercholesterolemia													PA	VA											
L2_HyperlipOther	CA												CA	MI											
L2_HypertensionEssential	AZ	CA	FL	IL	MI	NJ	NV	NY	TX				AZ	CA	FL	IL	MI	NJ	NV	NY	PA	TX			
L2_MixedHyperlipidemia	AZ	FL	IL	NV	OH	TX							AZ	CA	FL	IL	NV	OH	TX						
L2_MoodDisorder													AZ												
L2_Vaccine													OH												
L1_Back													NV												
L1_Digestive	PA												PA												
L1_DMSymptom	MO	OH	VA										SD												
L1_DMSympton													MO	OH	VA										
L1_RespInfec													NY												
StartYear													TX												
Tot3digitICD9													AZ	OH											
TotDiagWYear													FL	PA	VA										
TotDiagPerVisit	MI	NJ	SD										FL	MI	MO	NJ	OH	SD	VA						
acetaminophen_dose													MI												
ACEI_bin	FL												FL												
ACEI_npr	AZ												AZ												
statin_npr	NV												NV												
simvastatin_dose	NJ												NJ	OH											
statin_dose_adjusted	MO												MO												
active_principle	FL	MO											CA	FL	IL	MO	TX								
prescripts	VA												CA	IL	SD	VA									
medication_wo_prescript													NY												
diag_3digit_with_medication	PA												PA												
medication_diag_NA	CA	FL	IL	MO	NJ	NV	NY	OH	SD	TX			CA	FL	IL	MI	MO	NJ	NV	NY	OH	SD	TX		

Fig. 5. Selected biomarkers for each state.

of complications the patient had, and "medication_diag_NA" indicates whether the patient was previously diagnosed with type-2 diabetes through medical tests. These 6 selected features meet the well known common knowledge in the clinical domain. The clinical study suggests that hypertensive persons are more predisposed to the development of diabetes than normotensive persons. Moreover, up to 75% of cardiovascular disease (CVD) in diabetes is attributed to hypertension [37] and older people are more likely to suffer from type-2 diabetes [26]. Medical diagnoses tests are relatively accurate, i.e., most of the patients who were diagnosed with diabetes by the medical tests actually have diabetes. However, the sensitivity of these tests is low, i.e., a lot of diabetic patients remain undetected through these tests.

Our results can also reflect that overweight or obese is closely related to type-2 diabetes [26], but for different states it is represented in different ways. For most of the states the body mass index (BMI) (related to features 6–9) is more important than the direct body weight, while for few other states direct body weight plays an important role as well. It can be seen from our results, that our approach can provide a more accurate set of "region-specific" biomarkers in addition to the global markers. Such analysis on region-specific biomarkers in comparison to the overall biomarkers (which are typically known in the medical literature) can provide new medical insights into specific regional patient data which can thus lead to new discoveries in the clinical domain.

5. Conclusions and future work

Healthcare in the United States is currently undergoing a revolutionary transformation and EHRs are playing a vital role in improving the quality of patient care. However, due to the patient privacy and the sensitivity of the patient information that is being stored in the EHRs, there are considerable barriers to EHR proliferation. In this paper, we developed a novel distributed privacy-preserving integration method based on an ensemble based strategy for building robust prediction models from EHR data. Using our approach, each participator will need to reveal only limited model information built using the local data without revealing any patient-specific information. The local models built from each participator will be effectively combined to build an integrated model that is specific to each participator. This integrated model is built by taking advantage of aggregate knowledge from all the participators instead of the local participator's data alone. Compared to other existing works, our proposed framework can prevent the "negative impact" during integration from multiple sources, which typically happens when the data distribution differences among the participators are very large, by building a specific integration model for each participator with only a few selected local models instead of taking all the local models. We demonstrated the performance of our proposed method using the type-2 diabetes EHR data and have shown that constructing a global model which utilizes other local models performs better compared to the original local models built on each participator's data separately. In addition, we also successfully identified some important universal/global and region-specific biomarkers.

In the future, we will try to find more accurate ways to measure the difference of data distributions among multiple participators under privacy constraints. We also plan to extend our work with the idea of *k*-anonymity, so that we can measure the data distributions more accurately without compromising on patient privacy.

Acknowledgments

This work was supported in part by the National Science Foundation, United States grants IIS-1242304, IIS-1231742, IIS-1527827 and the National Institutes of Health, United States grant R21CA175974.

References

[1] C.C. Aggarwal, S.Y. Philip, A General Survey of Privacy-Preserving Data Mining Models and Algorithms, Springer, 2008.
[2] D. Agrawal, C.C. Aggarwal, On the design and quantification of privacy preserving data mining algorithms, in: Proceedings of the Twentieth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, ACM, 2001, pp. 247–255.
[3] R. Agrawal, R. Srikant, Privacy-preserving data mining, ACM SIGMOD Record. 29 (2) (2000) 439–450.
[4] M. Alther, C.K. Reddy, Clinical decision support systems, in: C.K. Reddy, C.C. Aggarwal (Eds.), Healthcare Data Analytics, Chapman and Hall/CRC Press, 2015.
[5] American Diabetes Association, et al., Diagnosis and classification of diabetes mellitus, Diab. Care 31 (Suppl. 1) (2008) S55–S60.
[6] An Act, Health insurance portability and accountability act of 1996, Public Law 104 (1996) 191.
[7] R.J. Bayardo, R. Agrawal, Data privacy through optimal k-anonymization, in: Proceedings. 21st International Conference on Data Engineering, 2005 (ICDE 2005), IEEE, 2005, pp. 217–228.
[8] R. Caruana, A. Niculescu-Mizil, Data mining in metric space: an empirical analysis of supervised learning performance criteria, in: Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2004, pp. 69–78.
[9] Centers for Disease Control and Prevention, et al. National diabetes statistics report: estimates of diabetes and its burden in the United States, 2014, US Department of Health and Human Services, Atlanta, GA (2014).
[10] C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin, M.Y. Zhu, Tools for privacy preserving distributed data mining, ACM SIGKDD Explorations Newslett. 4 (2) (2002) 28–34.
[11] C. Clifton, M. KantarcioÇ§lu, A. Doan, G. Schadow, J. Vaidya, A. Elmagarmid, D. Suciu, Privacy-preserving data integration and sharing, in: Proceedings of the 9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, ACM, 2004, pp. 19–26.
[12] C.M. DesRoches, C. Worzala, M.S. Joshi, P.D. Kralovec, A.K. Jha, Small, nonteaching, and rural hospitals continue to be slow in adopting electronic health record systems, Health Affairs 31 (5) (2012) 1092–1099.
[13] C. Dwork, K. Nissim, Privacy-preserving datamining on vertically partitioned databases, in: Advances in Cryptology—CRYPTO 2004, Springer, 2004, pp. 528–544.
[14] S.E. Fienberg, J. McIntyre, Data swapping: variations on a theme by Dalenius and Reiss, in: Privacy in Statistical Databases, Springer, 2004, pp. 14–29.
[15] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, J. Comput. Syst. Sci. 55 (1) (1997) 119–139.
[16] S. Gambs, B. Kégl, E. Aïmeur, Privacy-preserving boosting, Data Mining Knowl. Discov. 14 (1) (2007) 131–170.
[17] D.L. Hoyert, J. Xu, et al., Deaths: preliminary data for 2011, Natl. Vital Stat. Rep. 61 (6) (2012) 1–51.

[18] Z. Huang, W. Du, B. Chen, Deriving private information from randomized data, in: Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, ACM, 2005, pp. 37–48.

[19] A.K. Jha, C.M. DesRoches, E.G. Campbell, K. Donelan, S.R. Rao, T.G. Ferris, A. Shields, S. Rosenbaum, D. Blumenthal, Use of electronic health records in us hospitals, New Engl. J. Med. 360 (16) (2009) 1628–1638.

[20] kaggle.com, Practice fusion diabetes classification: identify patients diagnosed with type 2 diabetes, 2012. https://www.kaggle.com/c/pf2012-diabetes.

[21] M. Kantarcıoglu, J. Vaidya, C. Clifton, Privacy preserving naive bayes classifier for horizontally partitioned data, in: IEEE ICDM Workshop on Privacy Preserving Data Mining, 2003, pp. 3–9.

[22] Y. Li, B. Vinzamuri, C.K. Reddy, Constrained elastic net based knowledge transfer for healthcare information exchange, Data Mining Knowl. Discov. 29 (4) (2015) 1094–1112.

[23] Y. Lindell, B. Pinkas, Privacy preserving data mining, in: Advances in Cryptology—CRYPTO 2000, Springer, 2000, pp. 36–54.

[24] G. Mathew, Z. Obradovic, Distributed privacy-preserving decision support system for highly imbalanced clinical data, ACM Trans. Manage. Inform. Syst. (TMIS) 4 (3) (2013) 12.

[25] M. Matsuda, R.A. DeFronzo, Insulin sensitivity indices obtained from oral glucose tolerance testing: comparison with the euglycemic insulin clamp, Diab. Care 22 (9) (1999) 1462–1470.

[26] National Diabetes Information Clearinghouse, Am I at risk for type 2 diabetes? Taking steps to lower your risk of getting diabetes, NIH Publication No. 12-4805, 2012.

[27] National Diabeties Information Clearinghouse, The A1C test and diabetes, NIH Publication No. 147816, 2014.

[28] National Diabeties Information Clearinghouse, Diagnosis of diabetes and prediabetes, NIH Publication No. 144642, 2014.

[29] I. Palit, C.K. Reddy, Scalable and parallel boosting with MapReduce, IEEE Trans. Knowl. Data Eng. 24 (10) (2012) 1904–1916.

[30] B. Pinkas, Cryptographic techniques for privacy-preserving data mining, ACM SIGKDD Explorations Newslett. 4 (2) (2002) 12–19.

[31] R. Rahman, C.K. Reddy, Electronic health records: a survey, in: C.K. Reddy, C.C. Aggarwal (Eds.), Healthcare Data Analytics, Chapman and Hall/CRC Press, 2015.

[32] C.K. Reddy, C.C. Aggarwal, Healthcare Data Analytics, CRC Press, 2015.

[33] C.K. Reddy, Y. Li, A review of clinical prediction models, in: C.K. Reddy, C.C. Aggarwal (Eds.), Healthcare Data Analytics, Chapman and Hall/CRC Press, 2015.

[34] P. Samarati, Protecting respondents identities in microdata release, IEEE Trans. Knowl. Data Eng. 13 (6) (2001) 1010–1027.

[35] R.E. Schapire, Y. Singer, Improved boosting algorithms using confidence-rated predictions, Mach. Learn. 37 (3) (1999) 297–336.

[36] S.R. Simon, R. Kaushal, P.D. Cleary, C.A. Jenter, L.A. Volk, E.J. Orav, E. Burdick, E.G. Poon, D.W. Bates, Physicians and electronic health records: a statewide survey, Arch. Internal Med. 167 (5) (2007) 507–512.

[37] J.R. Sowers, M. Epstein, E.D. Frohlich, Diabetes, hypertension, and cardiovascular disease an update, Hypertension 37 (4) (2001) 1053–1059.

[38] US Department of HealthHuman Services and others, Health information technology: Initial set of standards, implementation specifications, and certification criteria for electronic health record technology, Fed. Regist. 75 (8) (2010) 13.

[39] World Health Organization, The top 10 causes of death, 2012. http://www.who.int/mediacentre/factsheets/fs310/en/.

[40] H. Yu, X. Jiang, J. Vaidya, Privacy-preserving SVM using nonlinear kernels on horizontally partitioned data, in: Proceedings of the 2006 ACM Symposium on Applied Computing, ACM, 2006, pp. 603–610.