Predicting Adverse Drug Reactions on Distributed Health Data using Federated Learning

Olivia Choudhury[1], PhD, Yoonyoung Park[1], ScD, Theodoros Salonidis[2], PhD, Aris Gkoulalas-Divanis[3], PhD, Issa Sylla[1], BA, Amar K. Das[1], MD, PhD
[1]IBM Research Cambridge, Massachusetts, USA
[2]IBM T.J. Watson Research Center, Yorktown Heights, NY, USA
[3]IBM Watson Health, Cambridge, Massachusetts, USA

Abstract

Using electronic health data to predict adverse drug reaction (ADR) incurs practical challenges, such as lack of adequate data from any single site for rare ADR detection, resource constraints on integrating data from multiple sources, and privacy concerns with creating a centralized database from person-specific, sensitive data. We introduce a federated learning framework that can learn a global ADR prediction model from distributed health data held locally at different sites. We propose two novel methods of local model aggregation to improve the predictive capability of the global model. Through comprehensive experimental evaluation using real-world health data from 1 million patients, we demonstrate the effectiveness of our proposed approach in achieving comparable performance to centralized learning and outperforming localized learning models for two types of ADRs. We also demonstrate that, for varying data distributions, our aggregation methods outperform state-of-the-art techniques, in terms of precision, recall, and accuracy.

1 Introduction

Adverse Drug Reactions (ADRs) are a major concern for medical practictioners, healthcare system, and pharmaceutical industry. As patients can experience expected and sometimes unexpected negative outcomes from taking any drug, delayed detection of ADRs can pose life-threatening risks to patients; posing considerable legal, financial, and social repercussions to the manufacturing companies and regulatory agencies. The use of medical data, such as claims and electronic health records (EHR), has become common in providing rich insights on health services and supporting ADR investigation[1]. Advancements in machine learning and artificial intelligence have produced a number of analytic methods that can be applied to such high-dimensional data for the purpose of predicting adverse reactions[2]. However, making timely and accurate predictions remains a challenge. Due to the distributed nature of healthcare data, obtaining a sufficiently large dataset to detect rare events requires merging data from different data silos. Analyses generated from different data sources can be conflicting or imprecise[3], necessitating methods to appropriately aggregate results.

Prior work to resolve these issues often have limitations in their approach. The Food and Drug Administration (FDA)'s self-report Adverse Event Reporting System (FAERS)[4], collects ADR data into a traditional, centralized database. A single database approach is the most straightforward way to explore ADRs, but information owned by different entities is seldom shared due to significant privacy concerns. Moreover, creating and maintaining such a large data repository incurs resource and system-level constraints, including high latency and single points of vulnerability (failure, breach). To avoid such overhead and risks, the FDA created the Sentinel system to monitor the safety of its regulated products, using a distributed data network[5,6]. The network comprises multiple stakeholders, each maintaining a large claims database. Despite the distributed framework and large-scale data amassed from active participation of data partners, Sentinel has limited analytic capabilities. Limitations of other state-of-the-art systems include access to potentially small-scale, sparse, and low-quality hospital records[7]. In addition, current claims-based frameworks experience a time lag between ADR instance, claim submission, adjudication, and consolidation of the claim into a database. EHR data, collected in near real-time, is therefore a promising alternative, but comes with the aforementioned quality concerns. Hence, there is an unmet need for accurate, scalable, and efficient solutions for predicting ADRs using distributed health data, that also protects the privacy of patients.

To address this challenge, we present a federated learning-based framework that permits health data to be distributed across multiple sites. Federated learning[8] has brought a paradigm shift in the construction of machine learning models from distributed data sources maintained by various organizations. Under such a decentralized, collaborative learning setting, each site contributes to the computation of a global model while simultaneously shielding its own data from leakage to distrusted third parties. Our framework allows us to train a global model based on each site's local data, without ever moving the raw data from their respective sites. To the best of our knowledge, this is the *first implementation of federated machine learning algorithms that leverages distributed electronic health data for predicting ADRs*. ADR prediction itself brings significant challenges for federated learning due to the huge imbalance between the majority class of individuals who do not suffer from ADRs, and the minority class of individuals with severe ADRs. To address this issue, we propose two novel methods of aggregating model updates from the sites and compare their performance with that of the state-of-the-art alternative. To show the effectiveness of our proposed approach, we consider two use cases: (i) prediction of chronic opioid usage for patients taking opioid drugs, and (ii) prediction of extrapyramidal symptoms for patients taking antipsychotic drugs. We conduct a comprehensive experimental evaluation using real-world patient data.

The **key contributions** of our work include: (1) implementing federated models for ADR prediction based on three supervised learning algorithms; (2) proposing and implementing two novel methods of aggregating local model updates in a federated setup; (3) demonstrating the effectiveness of our approach in analyzing sensitive, distributed, and highly imbalanced real-world electronic health data; (4) conducting a comparative analysis to evaluate our approach against state-of-the-art alternatives; and (5) demonstrating scalability of our approach for varying number of sites, data size, and data distribution.

2 Background

The sensitive and distributed nature of electronic health information in real-world scenarios motivate the need for a mechanism that can learn from data residing in silos, while accounting for data privacy. This compels us to explore the potential and value of federated learning for ADR prediction. Federated learning enables training a global model from distributed data, without having the sites exchanging any raw sensitive data. The global model is distributed to each site, where an instance is trained locally. The updates from locally trained instances are then aggregated to improve the global model, which is shared again with the sites for another round of training. This iterative process, illustrated in Figure 1, terminates when a performance criterion is met.

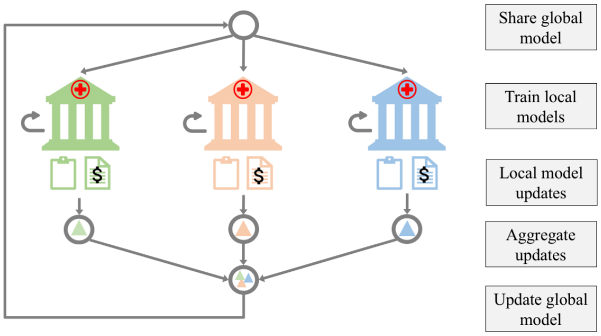


Figure 1: System design of federated learning for ADR prediction. Each site maintains electronic health records for a number of patients. Once a global model is shared with each site, it is trained on the site's local data. Updates to the local models' parameters are aggregated to improve the global model. This process is repeated until a convergence criterion for the global model is satisfied.

Initial implementations of federated learning were intended for image classification and language modeling on mobile devices[8,9]. Existing literature aims to improve the performance of deep networks in a federated setting[10–13]. There is currently very limited research focusing on the application of federated learning in healthcare. Recent work noted the effectiveness of federated models in predicting hospital admissions using EHR data[14]. However, the potential of federated learning in healthcare applications that make use of claims or EHR data for ADR prediction is yet to be explored. Moreover, the existing method of aggregating updates from local models[8] relies on the size, rather than the inherent characteristics, of the data. This approach may not work well in healthcare applications, which often deal with skewed, sparse, and imbalanced datasets. Hence, exploring the underlying characteristics of distributed data to improve the predictive capability of the global model is also an important research direction.

Unlike the methods that focus on surveillance of all potential ADRs for a given drug, specific prediction typically employs supervised learning algorithms[15–17]. Commonly used algorithms are logistic regression, random forest, decision trees, Support Vector Machine (SVM), and neural networks. Prior works on ADR prediction with machine learning methods are largely limited to centralized models, where all data are available to the researcher in a centralized data store. A majority of these works also lack evaluation on real-world datasets. For instance, distributed logistic regression based on multi-party computation, was studied using simulated data[18].

In this paper, we implement federated models based on three supervised classification algorithms: SVM, single-layer perceptron, and logistic regression, using stochastic gradient descent (SGD)-based optimization, which provides a generic approach for the algorithms to learn local models and aggregate their parameters to improve the global model and is the method currently supported by federated learning.

3 Methods

3.1 Data and cohort selection

To evaluate our approach, we used the Limited IBM MarketScan Explorys Claims-EMR Data Set (LCED). The data is procured from administrative claims and EHR data of over 1 million commercially insured patients over 5 years (2012-2017), each varying in lengths of medical activity. LCED contains patient-level features, such as demographics, diagnostic codes, outpatient prescription fills, laboratory results, and inpatient admission records. From this dataset, we defined our cohort based on the occurrences of two drug-ADR pairs during a 5 year period. Our selection was based on multiple factors: the number of patients taking the drugs of interest, the prevalence of the specific ADR of interest, and the feasibility of reliably detecting the ADR incidents using the LCED data.

Use case I: The first use case regards predicting chronic opioid usage (potential opioid use disorder) among patients who receive an opioid drug. Opioid abuse is currently one of the most pressing public health issues in the United States. The first exposure to opioids, for many who develop an abuse disorder, is a physician's prescription to alleviate symptoms of another condition. We closely follow Zhang et al[19] to construct the opioid user cohort by first identifying all patients who had one or more opioid prescriptions. We define opioid discontinuation as having 180 or more opioid-free days during one year (365 days) after the initial prescription, and chronic users of opioid as those who did not discontinue during the same period. Opioid-free day is calculated by subtracting from 365 days the sum of all subsequent opioid prescriptions' days-supply during the year following initiation. We excluded patients who had a diagnosis of opioid use disorder, cancer, or received hospice service in the year prior to the initiation, as their chronic opioid usage was expected. After identifying a number of risk factors for chronic opioid use from the literature, we hand-generate them from the data as predictive features. Examples include age, gender, smoking and alcohol habits, diagnoses of different types of pain (spinal injury, arthritis, etc.), surgical procedures, use of psychotropic drugs, and the morphine milligram equivalent (MME) of the initial opioid prescription normalized by its days-supply.

Use case II: The second use case regards predicting incidents of extra-pyramidal symptoms (EPS), a type of motor dysfunction, from the use of antipsychotic drugs. Broad off-label use of antipsychotics is well known despite the risk for developing ADRs like EPS. Similar to the opioid cohort, we start from identifying patients who had one or more prescription for an antipsychotic. We define the incidence of EPS following antipsychotic initiation using International Classification of Diseases (ICD) Ninth and Tenth Revision diagnosis codes. Examples of patient level predictive features identified from the literature include age, gender, smoking and alcohol habits, type of the first antipsychotic (first vs. second generation), other drug use (other psychotropic drugs, drugs related to Parkinson's Disease, etc.), comorbid diagnoses (dementia, psychiatric diagnoses, epilepsy, Parkinson's Disease, etc.), and number of prior hospital admissions. We generate these features using the data from the one-year period prior to the initiation of antipsychotic for each patient.

Cohort	# Patients	Mean Age (SD)	Female (%)	# Patients with ADR	Class ratio
Opioid	1,161,048	48.6 (18.2)	57.8	69,863	1:15.6
Antipsychotic	86,674	50.1 (21.7)	60.5	1,314	1:65

Table 1: Summary of the two drug-ADR pair cohorts. SD: Standard deviation

Table 1 provides summary statistics of the two cohorts. The imbalance between the majority (non-ADR) and minority (ADR) classes was consistent with our expectation, since severe ADRs are rare with most of the drugs available on the market. Our primary analyses focus on the opioid cohort, where a better balance is achieved with a larger sample size. We examine the antipsychotic cohort as an example of small, highly imbalanced data that one can encounter frequently in the healthcare domain. The implication of severe class imbalance is discussed in the following sections.

3.2 ADR prediction model

Consider a general binary classification problem, where features, denoted by x_k (for the k^{th} feature), are drawn from a feature space $\mathbb{X}$. The corresponding labels y_k are drawn from the label space $\mathbb{Y} := \{-1, 1\}$. Let the features corresponding to positive labels be denoted by $\mathbb{X}_+$ and those corresponding to negative labels by $\mathbb{X}_-$, that is

$$\mathbb{X}_+ = \{x_k \in \mathbb{X} : y_k = +1\} \quad \text{and} \quad \mathbb{X}_- = \{x_k \in \mathbb{X} : y_k = -1\}.$$

For any $x_k^+ \in \mathbb{X}_+$ and $x_k^- \in \mathbb{X}_-$, the objective of binary classification is to construct a function $f : \mathbb{X} \to \mathbb{Y}$ such that

$$f(x_k^+) = +1 \quad \text{and} \quad f(x_k^-) = -1.$$

In this paper, we denote cases of ADR as labels $y_k = +1$, and cases of non-ADR as $y_k = -1$.

3.2.1 Cost-sensitive learning

Class imbalance is intrinsic to ADR prediction. Since most classification algorithms assume balanced class distributions or equal misclassification costs, they fail to represent the characteristics of imbalanced data and are more likely to classify new observations to the majority class[20]. For ADR prediction, the cost of a false negative classification should be much higher than that of a false positive classification. Recent work on imbalanced learning can be categorized into sampling methods[21], cost-sensitive methods[22], and active learning methods[23]. As discussed in[24], sampling methods, such as undersampling the majority class or oversampling the minority class, either discard potentially useful data or can lead to overfitting. Since our dataset does not comprise unlabeled samples, active learning is not applicable. Hence, to mitigate the challenge of skewed data distribution, we incorporate cost-sensitive learning, wherein we increase the cost associated with misclassifying a minority class sample. Specifically, if C_{FN} and C_{FP} denote the cost of false negative and false positive in a cost matrix, respectively, then we set $C_{\text{FN}} > C_{\text{FP}}$. The magnitude of cost depends on the problem at hand and we determine their values using grid search.

3.2.2 Centralized model

For the purpose of binary classification of samples into ADR and non-ADR cases, we consider three supervised classification methods: SVM, single-layer perceptron, and logistic regression . We implemented these algorithms using scikit-learn version 0.20.2. To establish benchmark results, we first evaluate the performance of the classifiers in a centralized learning approach. This represents the scenario of gathering data from multiple sites for training a machine learning model. For each cohort, we split its entire dataset into two parts: 70% for training, and 30% for testing, where $\mathbb{X}_{\text{train}}$ and $\mathbb{Y}_{\text{train}}$ denote the feature and label sets for training, and $\mathbb{X}_{\text{test}}$ and $\mathbb{Y}_{\text{test}}$ denote the feature and label sets for testing. As the splits are stratified, the proportion of positive and negative cases in each split is the same as the entire dataset. After standardizing the features, we use 5-fold cross-validation to train the models on $\mathbb{X}_{\text{train}}$ and $\mathbb{Y}_{\text{train}}$, and test them on $\mathbb{X}_{\text{test}}$. To incorporate cost-sensitive learning, we update the $class\_weight$ parameter in scikit-learn based on class frequencies.

3.2.3 Localized model

Since healthcare and biomedical data is rife with sensitive information, sharing such data across sites or transferring it to a centralized database is often restricted. In such cases, a site has to rely on its own data for predictive analytics. We consider this scenario while designing localized models for ADR prediction. We train each classifier on a site's data, without leveraging data from other sites. Let us suppose there are N sites, representing hospitals or data owners. We use horizontal partitioning to split the training data into N disjoint subsets. We partition $\mathbb{X}_{\text{train}}$ into $\{\mathbb{X}^i_{\text{train}}\}_{i=1}^N$, where $\cup_{i=1}^N \mathbb{X}^i_{\text{train}} = \mathbb{X}_{\text{train}}$ and $\mathbb{X}^i_{\text{train}} \cap \mathbb{X}^j_{\text{train}} = \emptyset, \forall i, j \in \{1, ..., N\}$, for $i \neq j$. We follow the same logic to partition the corresponding label set $\mathbb{Y}_{\text{train}}$ into $\{\mathbb{Y}^i_{\text{train}}\}_{i=1}^N$. In the case of localized learning, the classifiers are trained on a single site's data $\{\mathbb{X}^i_{\text{train}}\}_{i=1}^N$ and $\{\mathbb{Y}^i_{\text{train}}\}_{i=1}^N$, and tested on $\mathbb{X}_{\text{test}}$. The limited availability of data may fail to account for detection of rare events[25]. We consider the results obtained from the localized models for benchmark analysis with federated and centralized learning models.

3.2.4 Federated model

In this paper, we focus on classification models that can be trained using gradient descent optimization, as currently supported by federated learning. Similarly to the scenario of localized model, for N sites, we randomly partition the training data into N disjoint subsets of feature set $\{\mathbb{X}^i_{\text{train}}\}_{i=1}^N$ and corresponding label set $\{\mathbb{Y}^i_{\text{train}}\}_{i=1}^N$. Let T denote the rounds of aggregating local model updates. For stochastic gradient descent, let η, E, and $Batch$ denote the learning rate, number of epochs, and batch based on a given batch size B, respectively. Let $F_i(w)$ be the local loss function of the i^{th} site with respect to its model parameter w. As described in Section 2, a global model is shared with each site, which trains the model on its local data. During local model training, based on given η, E, and $Batch$, at each site, we compute average gradient ($\triangledown F_i(w)$) with respect to its current model parameter w. We then compute weighted average to aggregate the parameter updates from the local models. The process is repeated until a convergence criterion, such as minimization of loss function, is satisfied. The process of training the global model only relies on updates from the local models, rather than raw data residing at the sites. Algorithm 1 presents the core algorithm of federated learning, where the weight w^i_D, used to compute weighted average, depends on the model aggregation method used. We first implement and evaluate the performance of the state-of-the-art model averaging approach, known as federated averaging[9]. It computes a weighted average based on the fraction of data residing at each site. For this case, the weight w^i_D is equal to $\frac{|D_i|}{|D|}$, where $|D_i|$ and $|D|$ denote the size of data at the i^{th} site and the entire dataset, respectively. Such an approach may fail to consider the inherent characteristics of data distribution at the sites. For the use case of ADR prediction, federated averaging would not account for imbalanced data and the varying distribution of ADR cases across sites. Since such scenarios are common when dealing with real-world health data, particularly in predicting rare events, it is important to explore other aggregation approaches.

3.2.5 Aggregation of local model updates

In this paper, we propose two novel methods of aggregating local model updates. The first method is particularly designed for training data with imbalanced classes. For each site, we estimate the class ratio of its training data to assign a corresponding weight, as denoted by w^i_D. This would imply that sites with cases of rare events, would have higher impact when improving the global model. For the second approach, we consider loss per sample, the change in the loss function during local model training. Since a gradient descent-based method attempts to minimize the loss function, we determine its rate of convergence. This is measured by the metric *epoch*, which is the maximum number of passes over the training data until convergence. Based on each site's epoch and training data size, we assign a weight, corresponding to w^i_D, for future aggregation. Using this approach, sites which require less training samples to reach convergence faster, will be assigned a higher weight during aggregation.

To evaluate these methods, we create a separate partition of the training data, based on the opioid cohort, to represent unequal distribution of class labels, as shown in Table 2. We do not conduct the same experiment with the antipsychotic cohort due to the limited number of minority class labels (ADR).

Site #	1	2	3	4	5	6	7	8	9	10
# ADR	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000
# Non-ADR	5,000	5,000	5,000	100,000	150,000	200,000	250,000	16,258	16,258	16,258
Class Ratio	1:1	1:1	1:1	1:20	1:30	1:40	1:50	1:3.2	1:3.2	1:4.1

Table 2: Partitioning of the opioid cohort training data with varying class ratio.

4 Experimental Evaluation

In this section, we present experimental results to evaluate our proposed approach. We discuss the evaluation metrics we used in this study, followed by a comparative analysis to demonstrate the effectiveness of the proposed system.

4.1 Evaluation metrics

To measure the predictive capability of the centralized, localized, and federated learning models, we compute *precision*, *recall*, and *accuracy* scores. As noted in prior work[21,26], precision and recall are better indicators for models dealing with imbalanced data. We also report the runtime incurred in training the models for each setup. All experiments were run on an Intel(R) Xeon(R) E5-2683 v4 2.10 GHz CPU equipped with 16 cores and 64 GB of RAM.

Algorithm 1 Federated Learning Model for ADR Prediction

```
function UPDATEGLOBALMODEL
    initialize $w_0$
    for $t = 1$ to $T$ do
        for $i = 1$ to $N$ do
            $w^i_{t+1}$ = UPDATELOCALMODEL($i, w_t$)
        $w_{t+1} = \sum_{i=1}^{N} w^i_D * w^i_{t+1}$

function UPDATELOCALMODEL($i, w$)
    for $e = 1$ to $E$ do
        for $b \in Batch$ do
            $w = w - \eta \nabla F_i(w)$
    return $w$
```

For the set of experiments comparing centralized, localized, and federated learning models, we examine the observed differences in performance metrics in two ways: (a) by calculating the % relative error of federated learning and localized learning with respect to centralized learning when using federated averaging[9], and (b) by testing the statistical significance of the difference using the Wilcoxon signed-rank test at 0.05 significance level.

4.2 Comparative analysis

We compare precision, recall, and accuracy of the federated learning (FL) model with two benchmark models: centralized learning (CL) and localized learning (LL). We execute the experiments 10 times for each setup. Figures 2, 3, and 4 report these metrics for the two datasets. As seen in Figure 2, SVM and perceptron yield similar scores and perform better than logistic regression. For all classifiers and datasets, federated learning exhibits comparable performance with respect to centralized learning. At the same time, due to the lack of sufficient training data, localized models do not perform well. It must be noted that the precision score for the antipsychotic data is higher than that for the opioid data. This is due to having a lower number of false positives, possibly because the former dataset is severely imbalanced with a class ratio of $1 : 65$.

Figure 3 presents the recall scores of the models for the two datasets. Perceptron generated the highest recall score, followed by SVM and logistic regression. Federated learning performed as good as centralized learning, and outperformed the localized learning models. Since the implementation of cost-sensitive learning reduced the cases of false negative, even with such imbalanced data, centralized and federated learning models for SVM and perceptron achieved high recall.

Figure 4 compares the accuracy score of the three models for the given dataset. Similarly to previous observations, Perceptron and SVM perform better than logistic regression. Our federated learning approach achieves comparable and better accuracy than centralized and localized learning models, respectively.

We observe that the % relative error values from federated learning are smaller than those from localized learning (Table 3). To put the numbers into a context, a difference of 5% in recall can translate to missing 5 out of 100 ADR cases compared to using centralized learning. Higher recall is desirable given the potential cost of missing severe ADR cases, and therefore federated learning with low % relative error is preferred. Based on statistical testing, in both opioid and antipsychotic data, the performance of centralized learning and federated learning is comparable for all three metrics for all classifiers. On the other hand, the performance of localized learning is inferior compared to either centralized or federated learning for the three evaluated metrics (all p values < 0.05).

In Table 4, we report the running time (in seconds) incurred in training the models for different setups. As expected, centralized learning requires a lot of time as it involves training the models on the entire training dataset. Federated learning requires significantly less time to train the models. Localized learning models train on a subset of the data on a single round, due to which they incur the lowest running time. For both datasets, perceptron required higher running time, compared to SVM and logistic regression. Due to the considerably large scale of opioid data, it consistently required more time to train the models.

To demonstrate the scalability of federated learning models, we further measure their predictive capability, in terms of precision and recall, for a varying number of sites and data sizes. As the number of sites increases, the size of training data residing at each site proportionally decreases. Due to the imbalanced nature of the data, this has a pronounced impact on the recall score, as evident in Figure 5. This scenario also accounts for the ability of the system to handle varying sizes of training data.

As previously discussed, we partition the opioid cohort such that the sites have a varying distribution of ADR and non-ADR cases (see class ratios in Table 2). We compare the effectiveness of our two proposed aggregation methods, in terms of precision, recall, and accuracy, with respect to default averaging (without weights) and federated averaging (based on data size). As seen in Table 5, for all evaluation metrics, our methods, particularly aggregation based on loss per sample, outperforms the state-of-the-art method of aggregation. This result implies that for skewed datasets, it is very important to consider the underlying characteristics of the data when aggregating local models.

		Precision		**Recall**		**Accuracy**	
Dataset	**Classifier**	**FL vs CL**	**LL vs CL**	**FL vs CL**	**LL vs CL**	**FL vs CL**	**LL vs CL**
Opioid	SVM	2.84(1.64)	6.44(1.59)	3.10(2.58)	8.49(3.88)	3.96(2.33)	13.34(2.50)
	Perceptron	1.11(.73)	7.16(2.49)	7.32(5.45)	9.06(6.06)	5.31(3.99)	12.88(4.58)
	LogReg	1.86(1.11)	11.28(2.56)	3.21(2.81)	11.55(3.37)	2.66(2.54)	12.55(4.49)
Antipsychotic	SVM	1.61(1.04)	11.82(2.28)	2.24(1.49)	15.71(2.55)	4.47(3.63)	16.45(2.50)
	Perceptron	2.81(1.68)	12.73(3.53)	2.39(1.95)	14.08(5.00)	4.87(2.50)	11.22(4.19)
	LogReg	1.42(1.26)	11.13(2.05)	2.21(1.99)	12.70(4.60)	4.19(3.90)	7.90(6.13)

Table 3: Comparison of relative error (%) for federated learning (FL) and localized learning (LL) with respect to centralized learning (CL). The values denote average (standard deviation) over 10 iterations.

	Opioid			**Antipsychotic**		
	CL	**FL**	**LL**	**CL**	**FL**	**LL**
SVM	612.8 (8.5)	122.2 (3.4)	4.1 (.3)	170.1 (1.8)	17.3 (0.9)	3.2 (.1)
Perceptron	842.8 (9.0)	117.6 (2.9)	6.3 (.7)	169.1 (1.3)	19.1 (1.4)	4.6 (.7)
Logistic Regression	513.7 (6.4)	102.7 (3.4)	4.8 (.6)	147.8 (2.3)	14.2 (.3)	3.9 (.7)

Table 4: Time (in seconds) incurred in training the centralized learning (CL), federated learning (FL), and localized learning (LL) models using SVM, perceptron, and logistic regression. The times denote average (standard deviation) over 10 iterations.

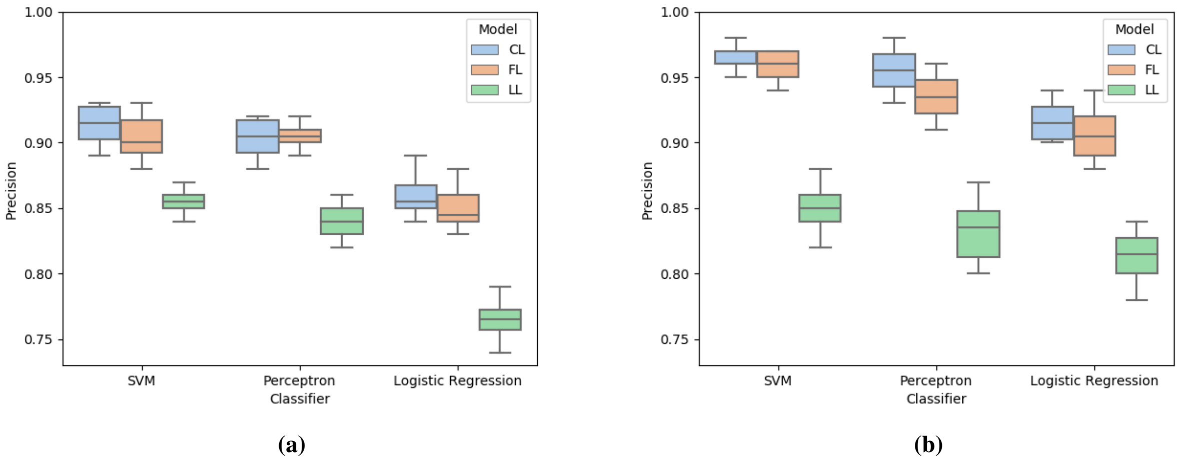
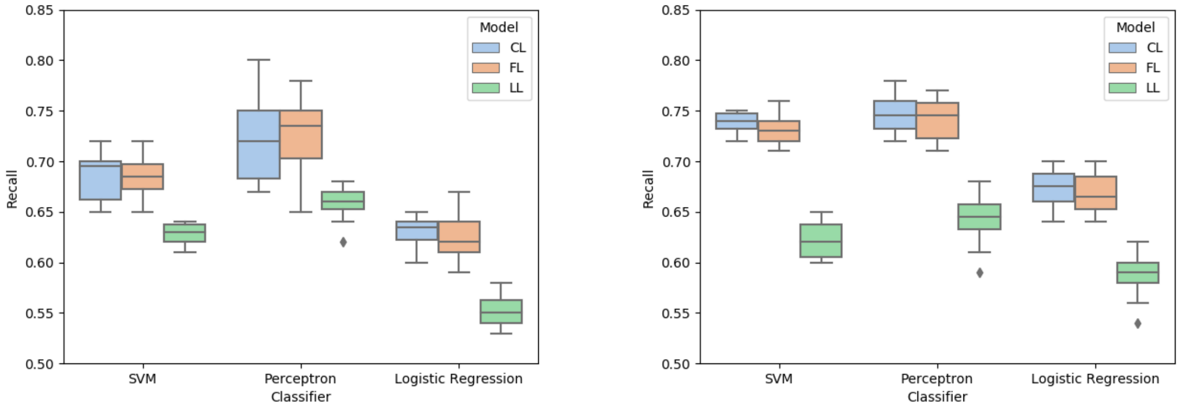


Figure 2: Comparison of precision score for centralized learning (CL), federated learning (FL), and localized learning (LL) models using SVM, perceptron, and logistic regression with (a) opioid data and (b) antipsychotic data.

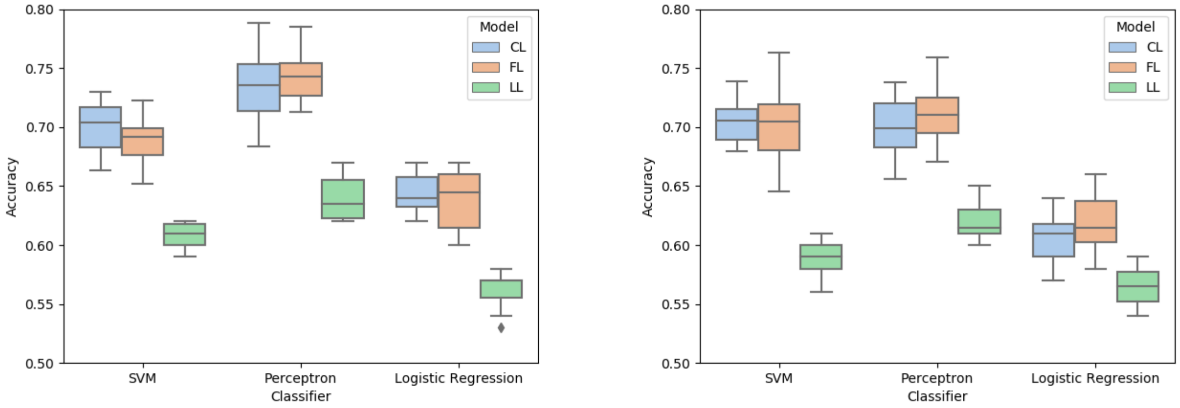
5 Discussion

The availability of electronic health data brings countless opportunities to investigate and predict ADR, provided that the hurdles in gathering and using such data are overcome. In this work, we proposed and evaluated the use of federated



(a) (b)

Figure 3: Comparison of recall score for centralized learning (CL), federated learning (FL), and localized learning (LL) models using SVM, perceptron, and logistic regression with (a) opioid data and (b) antipsychotic data.



(a) (b)

Figure 4: Comparison of accuracy score for centralized learning (CL), federated learning (FL), and localized learning (LL) models using SVM, perceptron, and logistic regression with (a) opioid data and (b) antipsychotic data.

learning to address the limitations of ADR prediction frameworks based on centralized learning. We demonstrated that SVM and perceptron perform better than logistic regression with respect to precision, recall, and accuracy. Perceptron has higher recall values, making it the preferable classifier for ADR detection, where false negatives generally have more significant consequences than false positives. We also demonstrated that the performance of federated learning models is comparable to that of centralized learning, implying that a federated learning framework can be used to predict ADR without compromising the model performance, while bypassing the challenges associated with centralized learning. An important finding of our evaluation regards the quality of our proposed aggregation approach with loss to sample ratio weighting, which achieves superior performance compared to state-of-the-art federated averaging. This approach is advantageous in federated learning applications with real-world health data, where severe class imbalance is a norm, rather than an exception.

In this paper, we focused on classification algorithms that are amenable to distributed solution using gradient descent, as currently supported by the federated learning paradigm. In the future, we plan to extend our federated learning framework to other types of algorithms, such as decision trees and gradient boosting, as well as applications where large-scale distributed datasets are common and deep learning models are applicable. We will leverage other charac-

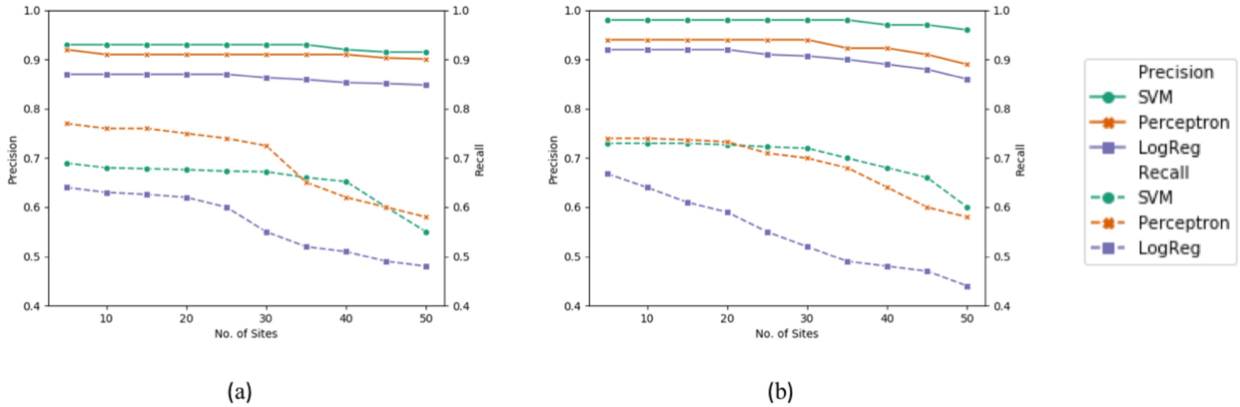


Figure 5: Effect of varying number of sites on precision and recall scores of federated learning models (SVM, perceptron, logistic regression) with (a) opioid data and (b) antipsychotic data.

Aggregate	Precision			Recall			Accuracy		
	SVM	Perc	LR	SVM	Perc	LR	SVM	Perc	LR
Average	.93 (.01)	.91 (.02)	.91 (.01)	.63 (.02)	.68 (.02)	.59 (.01)	.64 (.02)	.68 (.03)	.59 (.01)
Fed Avg	.93 (.02)	.91 (.01)	.90 (.02)	.58 (.03)	.64 (.01)	.54 (.02)	.58 (.01)	.64 (.02)	.54 (.02)
Class ratio	.94 (.01)	.92 (.01)	.90 (.01)	.72 (.02)	.68 (.01)	.61 (.02)	.71 (.01)	.67 (.01)	.62 (.01)
Loss/sample	.94 (.01)	.92 (.01)	.91 (.01)	.75 (.02)	.69 (.01)	.63 (.01)	.74 (.01)	.69 (.01)	.63 (.02)

Table 5: Comparison of our proposed aggregation methods (based on class ratio and loss/sample) with respect to averaging and federated averaging methods. For SVM, perceptron (Perc), and logistic regression (LR), we report the average (standard deviation) values of precision, recall, and accuracy scores.

teristics of data, such as quality, relevance, and rate of generation, to determine the impact of sites when aggregating their local model updates. We will also explore potential approaches for tuning hyperparameters of the global model in a federated setup. We intend to work on approaches for privacy-preserving federated learning, which protect patients' privacy against adversarial attacks, in addition to not exchanging raw data while training the models.

References

1. Nicholas P Tatonetti, P Ye Patrick, Roxana Daneshjou, and Russ B Altman. Data-driven prediction of drug effects and interactions. *Science translational medicine*, 4(125):125ra31–125ra31, 2012.
2. Peter B Jensen, Lars J Jensen, and SoÃÿren Brunak. Mining electronic health records: Towards better research applications and clinical care. *Nature Reviews Genetics*, 13:395–405, 2012.
3. David Madigan, Patrick B Ryan, Martijn Schuemie, Paul E Stang, J Marc Overhage, Abraham G Hartzema, Marc A Suchard, William DuMouchel, and Jesse A Berlin. Evaluating the impact of database heterogeneity on observational study results. *American journal of epidemiology*, 178(4):645–651, 2013.
4. FDA Adverse Event Reporting System (FAERS). `https://www.fda.gov/drugs/informationondrugs/ucm135151.htm`, Accessed: February 2019.
5. The Sentinel Initiative. `https://www.sentinelinitiative.org/`, Accessed: February 2019.
6. Qoua L. Her, Jessica M. Malenfant, Sarah Malek, Yury Vilk, Jessica Young, Lingling Li, Jeffery Brown, and Sengwee Toh. A Query Workflow Design to Perform Automatable Distributed Regression Analysis in Large Distributed Data Networks. *eGEMs*, 2018.
7. Bruce K Bayley, Tom Belnap, Lucy Savitz, Andrew L Masica, Nilay Shah, and Neil S Fleming. Challenges in using electronic health record data for cer: Experience of 4 learning organizations and solutions applied. *Medical Care*, 51:S80–S86, 08 2013.

8. Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.
9. H Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, et al. Communication-efficient learning of deep networks from decentralized data. *arXiv preprint arXiv:1602.05629*, 2016.
10. Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1175–1191. ACM, 2017.
11. H Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. *arXiv preprint arXiv:1710.06963*, 2017.
12. Shiqiang Wang, Tiffany Tuor, Theodoros Salonidis, Kin K Leung, Christian Makaya, Ting He, and Kevin Chan. When edge meets learning: Adaptive control for resource-constrained distributed machine learning. In *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*, pages 63–71. IEEE, 2018.
13. Stacey Truex, Nathalie Baracaldo, Ali Anwar, Thomas Steinke, Heiko Ludwig, and Rui Zhang. A hybrid approach to privacy-preserving federated learning. *arXiv preprint arXiv:1812.03224*, 2018.
14. Theodora S Brisimi, Ruidi Chen, Theofanie Mela, Alex Olshevsky, Ioannis Ch Paschalidis, and Wei Shi. Federated learning of predictive models from federated Electronic Health Records. *International journal of medical informatics*, 112:59–67, 2018.
15. Tao Hoang, Jixue Liu, Elizabeth Roughead, Nicole Pratt, and Jiuyong Li. Supervised signal detection for adverse drug reactions in medication dispensing data. *Computer Methods and Programs in Biomedicine*, 161:25–38, 2018.
16. Jing Zhao, Aron Henriksson, Lars Asker, and Henrik" Boström. Predictive modeling of structured electronic health records for adverse drug event detection. *BMC Medical Informatics and Decision Making*, 15(4):S1, 2015.
17. Felix Hammann, Heike Gutmann, Nadine Vogt, Christoph Helma, and Juergen Drewe. Prediction of Adverse Drug Reactions Using Decision Tree Modeling. *Clinical Pharmacology and Therapeutics*, 88(1):52–59, 2010.
18. Luk Arbuckle, Khaled El Emam, Saeed Samet, Robyn Tamblyn, Craig Earle, and Murat Kantarcioglu. A secure distributed logistic regression protocol for the detection of rare adverse drug events. *Journal of the American Medical Informatics Association*, 20(3):453–461, 08 2012.
19. Jinghe Zhang, Vijay S. Iyengar, Dennis Wei, Bhanukiran Vinzamuri, Hamsa Bastani, Alexander R. Macalalad, Anne E. Fischer, Gigi Yuen-Reed, Aleksandra MojsiloviÄĞ, and Kush R. Varshney. Exploring the Causal Relationships between Initial Opioid Prescriptions and Outcomes. *AMIA Workshop on Data Mining for Medical Informatics*, 2017.
20. Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge & Data Engineering*, (9):1263–1284, 2008.
21. Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.
22. Salman H Khan, Munawar Hayat, Mohammed Bennamoun, Ferdous A Sohel, and Roberto Togneri. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE transactions on neural networks and learning systems*, 29(8):3573–3587, 2018.
23. Maciej Zięba and Jakub M Tomczak. Boosted SVM with active learning strategy for imbalanced data. *Soft Computing*, 19(12):3357–3368, 2015.
24. Gary M Weiss, Kate McCarthy, and Bibi Zabar. Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs? *DMIN*, 7:35–41, 2007.
25. Jenna Wiens, John Guttag, and Eric Horvitz. A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions. *Journal of the American Medical Informatics Association*, 21(4):699–706, 2014.
26. Nitesh V Chawla, Nathalie Japkowicz, and Aleksander Kotcz. Special issue on learning from imbalanced data sets. *ACM Sigkdd Explorations Newsletter*, 6(1):1–6, 2004.