Private Cross-Silo Federated Learning for Extracting Vaccine Adverse Event Mentions

Pallika Kanani[1], Virendra J. Marathe[1], Daniel Peterson[1]
Rave Harpaz[2], and Steve Bright[2]

[1] Oracle Labs
[2] Oracle
{pallika.kanani,virendra.marathe,daniel.peterson
rave.harpaz,steve.bright}@oracle.com

Abstract. Federated Learning (FL) is quickly becoming a goto distributed training paradigm for users to jointly train a global model without physically sharing their data. Users can indirectly contribute to, and directly benefit from a much larger aggregate data corpus used to train the global model. However, literature on successful application of FL in real-world problem settings is somewhat sparse. In this paper, we describe our experience applying a FL based solution to the Named Entity Recognition (NER) task for an adverse event detection application in the context of mass scale vaccination programs. We present a comprehensive empirical analysis of various dimensions of benefits gained with FL based training. Furthermore, we investigate effects of tighter *Differential Privacy (DP)* constraints in highly sensitive settings where federation users must enforce DP to ensure strict privacy guarantees. We show that DP can severely cripple the global model's prediction accuracy, thus disincentivizing users from participating in the federation. In response, we demonstrate how recent innovation in *personalization* methods can help significantly recover the lost accuracy.

1 Introduction

Federated Learning (FL) is a distributed ML paradigm that enables multiple users to jointly train a shared model without sharing their data with any other users [4,30], offering advantages in both scale and privacy. In FL, multiple users wish to perform essentially the same task using ML, with a model architecture that is agreed upon in advance. Each user wants the best possible model for their individual use, but often has a limited budget for labeling their own data. Pooling the data of multiple users could improve model accuracy, because accuracy generally increases with increased training data. However user data cannot be shipped to a common model training facility due to bandwidth limitations or data privacy concerns. As a result, users locally train the shared (global) model on their local data, and thereafter send the updated model to the *federation server*. The federation server aggregates updates received from its users to improve the global model for all users.

Although the initial focus of FL has been on targeting millions of mobile devices [4], also called *cross-device FL*, the benefits of its architecture are evident even for institutional settings, also called *cross-silo FL* [28]. While cross-device FL is concerned with

both bandwidth consumption and data privacy, cross-silo federations and their users are considered well equipped with resources to handle bandwidth concerns, and data privacy is the primary objective. Our work focuses on the cross-silo FL setting.

Today our world grapples with safely rolling out massive scale vaccination programs to end a pandemic. Understanding adverse events related to these vaccines is critically important. These adverse events are often expressed in free text form, such as social media posts and reports provided to health care agencies and pharmaceutical companies. Currently, mentions of specific adverse events are extracted and coded manually, which is a time consuming, expensive and non-scalable process. Therefore, Machine Learning (ML) based methods to extract named entities (adverse events) automatically from such unstructured data are highly desirable.

Typically, more training data yield more accurate models. Unfortunately, collecting human annotations for building such Named Entity Recognition (NER) models is expensive, and particularly challenging given the need to maintain privacy of health records. One way to overcome this data scarcity issue would be for various agencies to share their data to build a joint model with combined data. However, privacy concerns, government regulation and data use agreements might not allow the data to leave individual organizational or geographical silos. Sharing user data with other users is absolutely not an option in these settings.

Cross-silo FL makes perfect sense to address such problems. Each vaccine provider's data remains in its private *silo*. At the same time, the provider can collaborate with other providers on a FL framework to collectively improve the NER model used for adverse event detection. Everyone benefits without violating data privacy. More specifically, for institutions participating in a federation as users, restricting data movement helps fulfill contractual obligations with their customers and comply with legal regulatory constraints on data movement [6,17].

However, restricting the provider's training data to its private silo does not guarantee complete privacy. Recent works have demonstrated that the data can indirectly leak out through model updates shipped by users to the federation server [3,42,44]. To combat this problem, researchers have proposed the addition of Differential Privacy (DP) [12,14,13] to FL [1,19,31,41].

Informally, DP aims to provide a bound on the variation in the model's output based on the inclusion or exclusion of a single data point used in its training set. This is done by introducing precisely calibrated noise in the training process. The method of noise calibration and injection varies between implementations [1,41], but is always structured to enforce the precise formal DP guarantee, which we define in section 2. We will refer to this process as "DP inducing noise injection" henceforth. This noise makes it difficult, even impossible, to determine whether any particular data point was used to train the model.

In settings where the federation server is trusted, DP enforcement is delegated to the federation server [41]. However, in settings where users do not trust even the federation server, DP may need to be enforced by the users locally [29]. While all this noise is structured to enforce formally provable privacy guarantees for each training data point [13], it can significantly degrade accuracy of model predictions. This degradation may happen to an extent that disincentivizes users from participating in the federation –

the global (noisy) model performs worse than a user-resident local model trained just on the user's dataset, which we call the *individual* model.

Another instance where the global model may perform worse than the individual model for a user is when the user's data distribution is different from most of the users, or the users collectively have non-IID training data [25,37]. There is a rapidly growing body of FL *Personalization* literature to address this problem [11,15,38,39,45,52], a handful of which addresses model degradation due to DP induced noise [45,52].

We are interested in applying this body of work to real-world problem settings. The health care sector is one such application domain that can leverage FL in significant ways. Indeed there is rapidly growing awareness and investment in FL at world-wide scale including consortiums [43] and public-private partnerships [26]. This is accompanied by the beginnings of applied research in this sector [36].

In this paper, we case study application of FL to the problem of vaccine adverse event detection, the first of its kind to the best of our knowledge. Importance of such a study cannot be understated in today's pandemic stricken world. Given the unprecedented speed at which new vaccines have been rolled out, it is crucial to automatically extract mentions of adverse events related to these vaccines from patient reports. We study implications of applying FL to train a Named Entity Recognition (NER) model on the Vaccine Adverse Event Reporting System (VAERS) dataset that we have annotated and partitioned by vaccine manufacturers. Each vaccine manufacturer acts as a federation user whose dataset is *siloed* in its private sandbox; all these sandboxes participate in our FL framework over multiple training rounds.

Our experiments reveal several interesting insights including general effectiveness of FL on model performance, effects of DP enforcement on model performance, and the value of personalization techniques to incentivize users to participate in FL. In particular, we show that FL improves average F1 value by 37.43% over the individual model, while enforcement of DP (DP-FL) degrades the FL model's average F1 by 25.17%. For one of the users, this degradation is so severe that the private FL model F1 is worse by 45.55% when compared with the individual model F1. This clearly makes DP-FL a non-starter for some users to join the federation. We study FL with *Fine-Tuning* (FT-FL) [52], a personalization approach that fine-tunes the global model at each user *after* the entire FL training process completes. Interestingly, contrary to prior work [52], simply augmenting fine-tuning to FL does not result in prediction accuracy improvement for the federation users. Instead, user accuracy degrades in most cases. However, somewhat surprisingly, fine-tuning in the presence of DP (FT-DP-FL) boosts user accuracy by 24.88%, compared to the individual model, to strongly incentivize users to join and stay with the federation. We also observe that vaccine reports related to different manufacturers have slightly different vocabulary (e.g. mentions of different vaccine names), and different distributions of adverse events, which aid FT-DP-FL in effectively recovering lost accuracy.

Even more interestingly, our findings indicate a unique *incentive structure* for users to join the federation. In particular, we find that users with small amount of training data, a.k.a. *small* users, have a strong incentive to join and stay with the federation even when DP is enforced without fine-tuning. This is because the user's private dataset is so small that any locally trained individual model performs poorly. Furthermore, even the global

model that is degraded because of DP inducing noise performs significantly better than the user's individual model. In short, small users have virtually no incentive to leave the federation, and may not require additional layers of personalization to improve the global model as long as there are enough participants in the federation.

For users with larger amount of data, the narrative is quite different. In particular, we observe that the global model's degradation due to DP inducing noise is significant enough to disincentivize those users from participating in the federation. As a result, if they opt for the additional layer of privacy through DP, the importance of personalization based enhancements, which salvage the accuracy lost due to DP inducing noise, cannot be understated.

In summary, this paper makes the following contributions:

- We present the first comprehensive study, to the best of our knowledge, on application of FL to the vaccine adverse event detection task in the field of pharmacovigilence on real-world data – the VAERS dataset.
- Our study examines benefits of FL based training, along with its robustness to user participation.
- We examine challenges posed by enforcement of differential privacy, to the extent that may disincentivize users from participating in a federation.
- We show that, unlike prior work [52], simply augmenting FL with personalization techniques, such as the aforementioned fine-tuning (FT-FL), does not necessarily improve prediction accuracy for FL users. In fact, it degrades prediction accuracy in our experiments. However, somewhat surprisingly, the same techniques (FT-DP-FL) turn out to be highly effective in recovering lost accuracy due to DP inducing noise injection. We furthermore show that personalization is robust to user participation uncertainties (e.g. users dropping out).
- We report an interesting new *incentive structure* amongst users participating in the federation, where users with small amount of training data are strongly incentivized to join and stay with the federation, whereas users with somewhat larger amounts of data require enhancements, such as FT-DP-FL, to overcome the pitfalls of DP inducing noise injection.
- Another surprising finding in our study is that fine-tuning based personalization is highly resilient to increasingly tighter margins for the differential privacy budget ($\epsilon < 1$).

The rest of the paper is structured as follows: We discuss background material and related work in section 2. The VAERS system used as the basis of this study is described in section 3. We describe our NER model used in an adverse event detection system, along with our FL framework and the personalization approach we use in section 4. Our comprehensive experiments and their analysis appears in section 5.

2 Background

Federated Learning (FL) In FL, a federation server initializes a global model and ships it to all participating users thereby initiating distributed training. Training happens over multiple rounds. In each round, each user, on receiving the global model re-trains the

model on its private data and sends back the resulting parameter updates to the federation server. The federation server aggregates updates from all users applying them to the global model, and then ships the revised model back to the users. The most widely used method of aggregation is FedAvg [30,40], where user parameter updates are averaged at the federation server and applied to the global model. Formally, FedAvg solves the following optimization problem:

$$\min_{w \in \mathcal{R}^d} f(w) \qquad where, f(w) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} f_i(w) \tag{1}$$

The function $f_i = \mathcal{L}(w; x_i, y_i)$ represents the local loss for each of the n federation users on the model w using the user's private data x_i, y_i.

Figure 1 shows the overall FL architecture. Users can dynamically join the federation or drop out. The framework is structured to be resilient to such changes. Noting privacy concerns, more recent work has proposed addition of differential privacy to FL [19,31,40].

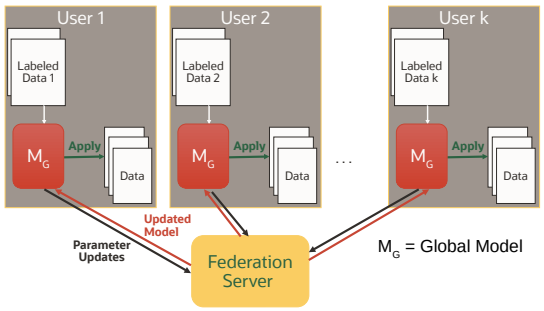

Fig. 1. The Federated Learning setting. M_G is the global model the federation server sends to users, each of which re-trains M_G on its private data and sends the updated model parameters back to the federation server.

Differential Privacy (DP) *Differential Privacy* [13] is a mathematically quantifiable privacy guarantee for a data set used by a computation that analyzes it. While it originally emerged in the database and data mining communities, triggered by privacy concerns in Machine Learning (ML) [16,24,33,47,49], DP has garnered enormous traction in the ML community over the last decade [1,5,7,9,10].

In DP, the privacy guarantee applies to each individual item in the data set and is formally specified in terms of a pair of data sets that differ in at most one item. Specifically, consider an algorithm A such that $A : D \mapsto R$, where D and R are

respectively the domain and range of A. Now consider two data sets d and d' that differ from each other in exactly one data item. Such data sets are considered *adjacent* to each other in the DP literature. Algorithm A is said to be (ε, δ)-differentially private if the following condition holds true for all adjacent d and d' and any subset of outputs $O \subseteq R$:

$$P[A(d) \in O] \leq e^{\varepsilon} P[A(d') \in O] + \delta \quad (2)$$

Enforcement of DP typically translates into introduction of a "correction" in algorithm A to ensure that the differential privacy bound holds for any two adjacent inputs. This correction is commonly referred to as the *noise* introduced in the algorithm, its input, or output to ensure that the (ε, δ)-differential privacy bound holds. While a disciplined introduction of noise guarantees DP, the noise itself leads to accuracy degradation in the output produced by A. In the context of ML, the algorithm is a model being trained using sensitive private data sets, and accuracy degradation can significantly hamper the model's utility.

Personalization in FL The basic FL algorithm, FedAvg, assumes IID training data across all FL users. In fact, it is known to be quite effective in practice for such data distributions. However, FedAvg may perform poorly in the presence of non-IID user data [25,37]. A recent flurry of research addresses this problem using *personalization* techniques [11,15,38,39,45,52] that specialize training at each user, typically in the form of training an additional local model, or letting the local copy of the global model "drift" from the global model in a constrained fashion. This enables the local model to fit better to the user's local data distribution thereby delivering a better performing model.

Adverse Event Mention Extraction By some estimates, adverse drug reactions are among the leading causes of death in the developed world. Reports of adverse events are a critical source of information for tracking and studying adverse events associated with medicinal products. However, portions of the sought information is only available in unstructured format. The use of and necessity of automated methods for extracting mentions of drug adverse events from unstructured text is widely recognized in pharmacovigilance [23]. Several different genres of text are tackled in this line of research, including social media [21,32], biomedical literature [34,50], clinical narratives [22,35] and drug labels [46]. More recently, use of state of the art deep learning technology for NER have been proposed [20].

3 Vaccine Adverse Event Reporting System

Drug and vaccine safety surveillance relies predominantly on spontaneous reporting systems. These systems are comprised of reports of suspected drug/vaccine adverse events (potential side effects) collected from healthcare professionals, consumers, and pharmaceutical companies, and maintained largely by regulatory and health agencies. Among other, these systems are used to detect possible safety problems – called "signals" – that may be related to a vaccination or the consumption of a drug. In the US, the prominent surveillance system for vaccines is the U.S. Centers for Disease Control and

Prevention (CDC) and the Food and Drug Administration (FDA) Vaccine Adverse Event Reporting System (VAERS), created in 1990.

The VAERS data (de-identified) is publicly available in structured format. Each VAERS report includes the name of (and additional information about) the administrated vaccine, a list of adverse events related to the vaccine, dates, and limited demographic information about the patient receiving the vaccine (e.g., age, gender). Importantly, the report also includes a textual narrative describing the adverse event. For example,

"*Shortly after patient was vaccinated, she started to feel an itching, tingling feeling in her throat. Fearing that it was an allergic reaction, I called 911. The patient remained alert, talking and breathing normally until paramedics arrived, though she stated that she started to feel additional tingling in her arms and chest.*"

In this example, the following token spans would be annotated as adverse events: "itching", "tingling feeling in her throat", "allergic reaction", "tingling in her arms and chest".

Most of the data collected in VAERS is currently processed by humans for downstream applications. Adverse event reports, whether they're forms, emails, articles, or other source documents, do not arrive in structured format, which means they have to be entered manually into safety systems. This manual data entry can take hours and represents a significant cost to the organization. Free-text narratives take the most time, requiring a manual sift through every sentence to find relevant information and then enter it into the correct field. With the rapidly increasing volume of such data this human effort is becoming prohibitive and calls for the increased use of automated methods such as NER. In addition, pharmacovigilance data such as that available in and similar to VAERS originates from private siloed sources, motivating the need for privacy preserving distributed approaches such as FL.

4 Model and Framework

4.1 NER based on Recurrent Neural Networks

The recurrent neural network (RNN) architecture we used to perform NER is based on a commonly applied BiLSTM architecture. The architecture consists of three major components: (1) a word representation layer made of word embeddings, (2) two stacked layers of bidirectional long short-term memory (LSTM) cells, and (3) a feedforward layer that performs the final BIO sequence labeling.

Pre-trained word embeddings were used to seed the network's word embedding layer. These were generated using Word2Vec applied to the sentences comprising the VAERS NER dataset described in section 5. Dropout regularization was implemented between each of the three major network components. The dropout rate was 0.4.

The network was implemented on PyTorch6 and trained using stochastic mini-batch gradient descent with the Adam optimizer for a pre-defined number of iterations. Each iteration processed a batch of 256 randomly selected sentences. The network was trained for a total of 20 epochs, each epoch consisting of number of sentences in the training set / batch size iterations.

4.2 Federated Learning Framework

We have implemented our own FL simulation framework, on PyTorch6, that hosts the federation server and users on the same computer. The framework supports several federated aggregation protocols, including FedAvg and FedSGD [30], of which we use FedAvg in our evaluation. The framework is extendable to support other custom aggregation protocols [11,15,38,45,52].

Trust Model Considerations and Differential Privacy The decision to train a ML model using the FL framework requires careful analysis of privacy considerations for users' data. More specifically, the *meaning* of the term "data privacy" in a given setting needs to be precisely understood since it has profound implications on techniques required to enforce the desired data privacy. For instance, in some settings, simply restricting user data to its private silo is sufficient for the use case. On the other hand, in settings involving highly sensitive private data (e.g. health records of individuals), it may be desirable to ensure that even the parameter updates shipped from the user silo to the federation server cannot be reverse engineered by any means, external to the user, to determine the user's training data records. Ultimately, the level of privacy protection must be agreed upon by all parties involved. While an exhaustive treatment of a taxonomy of such *trust models* in FL is beyond the scope of this paper, we assume that personal health records describing an adverse reaction to a vaccine are highly sensitive private material. Consequently, they must be protected using techniques guaranteeing the strictest data privacy.

In the FL setting, these data records would be hosted in a participating pharmaceutical company's silo. The pharmaceutical company's silo performs the role of a user in the federation. We view Differential Privacy (DP) as an appropriate tool to enforce privacy guarantees to individuals' health records. However, more careful analysis of how DP is enforced in FL settings is required. Other technologies such as secure multi-party computation [51] and homomorphic encryption [18] may be worth considering, but are beyond the scope of this work. Additional security technologies such as end-to-end encryption may be necessary to augument to the DP solution, but is also outside the scope of this work.

We assume a trust model where users do not trust the federation server, and enforce DP *locally* on the parameter updates shipped back to the server. To enforce DP locally, we use the algorithm proposed by Abadi et al. [1] that injects gaussian noise (calculated using their moments accountant algorithm) in parameter gradients during local training at each user. Noisy gradients lead to noisy parameter updates, which are eventually shipped from the user to the federation server.

Interestingly, since users can possess datasets with different sizes, the computed noise, which is a function of the dataset size, varies considerably from user to user. For instance, the noise introduced for a user with a handful of data points is much higher than the noise introduced by a user with a much larger private dataset. However, FedAvg smoothes out the noisy updates through the parameter aggregation process (averaging, in our case). The resulting model that each user receives is much more robust. Note that our implementation of DP covers the privacy of each narrative, but we assume that there

is not enough information in the data to link multiple narratives relating to the same person.

Personalization through Fine Tuning The main allure of FL for a user is the promise of significant prediction accuracy improvements over a locally trained *individual* model. While parameter aggregation through FL can significantly improve accuracy of the global model, introduction of noise to enforce DP can severely compromise that improvement. The degradation can be severe enough to make users reconsider their decision to join the federation, and deter new users from joining the federation. Furthermore, data distributions across users may have significant side effects on the global model's prediction accuracy: If a user's dataset has a significantly different distribution than most of the federation users, the global model may perform worse than a locally trained individual model. If users of a federation have non-IID data, the resulting global model may be ineffective [37].

Many researchers have recently proposed different forms of *personalization* approaches to remedy the disparate data distribution problem [2,8,27,38,39,45,48,52]. Just two of these works [45,52], to the best of our knowledge, propose personalization approaches as solutions to model degradation due to DP inducing noise. Among the proposed personalization approaches, we focus on FL with *Fine Tuning* [52]: FT-FL for fine tuning on top of plain FL, and FT-DP-FL for fine tuning on top of FL with DP enforcement at the user. In this approach each user continues training, without noise, the local copy of the global differentially private model *after* the FL training process has completed.

The fine tuning based parameter updates are private to each user and are not shared with the federation. As a result, the fine tuned local models may diverge from the global model at varying degrees in order to better fit the users' private data. While endlessly fine tuning the global model can lead to the model converging to a locally trained individual model, care must be taken to ensure that the fine-tuned model does not deteriorate. This can be achieved through standard hyperparameter tuning techniques.

5 Experiments

5.1 Dataset

We used a total of 17,841 narratives submitted to VAERS through the years 2015-2017 to form the NER data set used for this study. The narratives were automatically annotated for adverse event named entities using the list of adverse events supplied with each report. In total the NER data set used for this study comprised of 87,730 sentences and 39,139 annotated adverse event named entities. In our experiments, we split the data randomly into train, validation, tune and test sets in the proportion 60%, 10%, 10%, and 20% respectively. We used the validation set to decide early stopping in the fine tuning algorithm and tuned the rest of parameters on the tune set. We refer to "large manufacturers" as those with more than 1000 VAERS reports in this data and "small manufacturers" as those with fewer reports to reflect the availability of training data in each user's silo. In the rest of this paper, we use the terms 'manufacturer' and 'user' interchangeably.

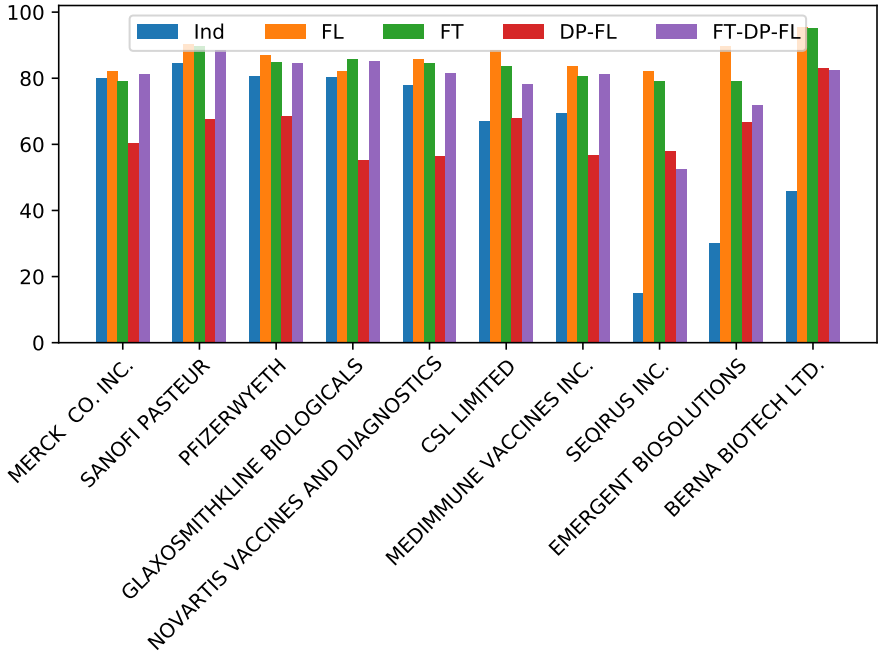

Fig. 2. F1 per manufacturer for different methods for $\epsilon = 2.0$

5.2 Experimental Setup

As the first baseline for our experiments, we train Individual models (*Ind*), i.e. assume that each manufacturer only uses their own training set, and test on their respective test set. This baseline represents the case in which the manufacturer chooses not to participate in the federation at all. *FL* is the federated learning model trained in a collaborative fashion across users using the FedAvg algorithm. This model is then fine tuned for each user using the protocol described in section 4, which yield a set of models, one per manufacturer, that we call *FT*. Next, we introduce DP to the *FL* model, as described in section 4. We use $\epsilon = 2.0$ for this first set of experiments as it is considered a fairly conservative privacy setting in the literature [1] and calculate the sigma values suitable per user. We call this private federated learning variant *DP-FL*. Finally, we fine tune this private FL model and call it *FT-DP-FL*.

The training parameters for all of these algorithms were tuned using a separate tuning dataset. We use a learning rate of 0.01 and train all the federated models for 20 rounds of FedAvg, with additional 20 epochs for the fine tuning variants at each manufacturer. For evaluation, we compute the precision, recall, and F1 of each token label on a 1-vs-all basis. The values reported are the mean F1 score (henceforth called F1) for the labels at the beginning or inside of an adverse event mention.

We ask the following questions as part of this study. Does *FL* perform better than *Ind* models across users? What happens when differential privacy is introduced? Does

personalization help improve accuracy over *FL* and mitigate *DP-FL*'s accuracy loss enough to re-incentivize users to participate in the federation? If fine-tuning based personalization helps mitigate accuracy loss due to DP, how robust is it to varying parameters of DP? Finally, we ask if the federation is stable enough for the uncertainties of real world, such as users dropping out? We also analyze the incentive structure that emerges for users with varying amounts of training data.

5.3 Private Federated Learning with Personalization

Figure 2 shows the F1 values for each of the described models on the individual users' test sets. Note that the manufacturers on the x-axis are sorted based on the size of their training sets. As we can see, the FL model consistently outperforms *Ind* models for each of the users, including large manufacturers with a lot of training data. As table 1 shows, the amount of error reduction over the *Ind* model for each user is substantial. Contrary to findings by Yu et. al. [52], in our case, personalization based on fine tuning *FT-FL* performs worse than *FL* in most cases. As we add noise related to differential privacy to the federated learning model, F1 values drop significantly across the board. This makes participation for larger manufacturers in the federation unattractive, since the *DP-FL* model ends up performing worse than their *Ind* models. However, applying fine tuning in this case helps bring it back up to the point, where it is again advantageous for each party to participate in the federation. This shows that personalization based approach can help mitigate the loss of accuracy from introducing differential privacy.

Vaccine Manufactuer	**Num Reports**	**Individual F1**	**FL**		**FT-DP-FL**	
			F1	**Error Red.**	**F1**	**Error Red.**
Merck Co. Inc.	7638	80.10	82.00	9.55%	81.20	5.53%
Sanofi Pasteur	3352	84.60	90.40	37.66%	88.40	24.68%
Pfizer-Wyeth	2428	80.50	87.00	33.33%	84.60	21.03%
Glaxo-Smithkline Biologicals	2289	80.20	82.20	10.10%	85.30	25.76%
Novartis Vaccines And Diagnostics	1183	77.80	85.80	36.04%	81.50	16.67%
CSL Limited	465	67.10	88.50	65.05%	78.30	34.04%
Medimmune Vaccines Inc.	265	69.30	83.50	46.25%	81.10	38.44%
Seqirus Inc.	111	15.00	82.10	78.94%	52.60	44.24%
Emergent Biosolutions	58	30.10	89.70	85.26%	71.90	59.80%
Berna Biotech Ltd.	52	45.80	95.40	91.51%	82.50	67.71%

Table 1. F1 and Error Reduction with Federated Learning and Private Federated Learning with Fine Tuning. 'Vaccine Manufacturer' is a field in the public VAERS database that identifies the manufacturer of the vaccine reported in the VAERS form. There is no relationship between this field and the reporter. 'Num VAERS Reports' does not represent the rate of adverse events associated with the manufacturer or its products and cannot be used to estimate such rates. The statistics are based on a sample of reports submitted to VAERS between 2015-2017 whose MedDra coded adverse events appeared in the narrative. Because the statistics are based on a carefully selected sample, the distribution of reports shown may not represent the true distribution of reports associated with different vaccine manufacturers.

It is interesting to note that for small manufacturers, with an exception of one with very small amount of evaluation data, it is always beneficial to participate in the federation, even for *DP-FL*, with or without personalization. For large manufacturers however, the DP is only attractive in the presence of the mitigation offered by fine-tuning based personalization (*FT-DP-FL*).

5.4 Stability of Federation against Users Leaving

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
M1	0.0	0.9	1.8	0.4	1.0	2.1	1.8	0.4	1.0	0.0
M2	-0.3	0.0	0.4	0.5	1.4	1.6	1.6	-0.4	3.2	-1.5
M3	-0.1	0.5	0.0	0.1	0.1	0.9	1.4	1.9	1.0	-1.5
M4	-0.6	0.8	0.2	0.0	2.6	-0.2	3.5	1.3	1.0	0.0
M5	-0.5	-0.1	-0.1	2.9	0.0	0.6	0.6	-1.9	1.0	0.0
M6	-0.8	0.0	0.2	-0.5	-0.4	0.0	1.6	-1.1	2.1	0.0
M7	-0.5	0.5	-0.3	-0.5	0.1	0.7	0.0	0.4	1.0	-1.5
M8	-0.7	0.3	0.3	-0.1	-0.5	0.0	-0.5	0.0	0.8	0.0
M9	-0.4	0.1	0.2	0.0	0.4	0.1	0.9	0.9	0.0	4.5
M10	-1.0	0.0	-0.2	-0.2	-0.2	0.3	-1.3	-1.1	0.0	0.0

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
M1	0.0	0.1	0.4	1.9	-2.4	1.4	2.9	-8.3	0.3	15.8
M2	-0.1	0.0	0.6	1.6	-1.5	-1.6	0.5	-2.5	1.4	22.5
M3	0.5	0.5	0.0	2.1	-1.7	0.2	-1.3	-1.2	-1.2	3.7
M4	-0.3	-0.3	0.2	0.0	-0.1	-4.3	0.7	-1.3	-0.4	18.7
M5	-0.1	0.0	-0.3	1.0	0.0	-0.3	-0.3	-1.9	-0.8	0.5
M6	-0.2	-0.5	0.3	1.6	-1.9	0.0	-1.5	-0.3	-0.5	4.2
M7	-0.5	0.1	0.3	2.2	-1.2	-2.8	0.0	-0.5	0.9	28.9
M8	0.5	-0.5	0.8	0.6	0.0	-4.0	-0.9	0.0	5.2	15.8
M9	-0.5	-0.5	0.3	1.0	-2.5	-3.3	-3.5	-2.4	0.0	4.1
M10	-0.1	-0.2	1.0	0.9	-1.8	-3.2	-0.1	-1.4	2.2	0.0

Table 2. Stability of Private FL with Fine Tuning performance when a single user leaves. M1-M10 are manufacturers sorted in descending order by size. Each row represents a manufacturer that is leaving the federation. Each Column represents the difference between F1 values under full federation and this reduced federation for that manufacturer. The table on the left represents FL and the table on the right represents FT-DP-FL

Building a federation across organizations can be challenging in the real world due to a variety of factors. For instance, users may discontinue their participation in the federation. We simulate this scenario and study the effect of one of the manufacturers leaving the federation. As we can see from Table 2, both federated learning and private federated learning with fine tuning are fairly stable against such a change, with the exception of a few manufacturers with very small amount of training and test data. In other words, no single manufacturer has disproportionally large impact on the overall accuracy gains from participating in the federation.

5.5 Federation of Small Manufacturers

Another scenario that we simulate is the one where only participants with small amount of training data agree to collaborate. In this case, we do not have the advantage of the large amount of training data from any of the larger manufacturers. To better understand if such a federation is still advantageous, we compare the F1 values for small manufacturers in two different scenarios: one, in which they are a part of a large federation with all manufacturers, and second, in which they are a part of a federation with only the small manufacturers.

Figure 3 shows these comparisons for FL and FT-DP-FL respectively. As is clear from the bar chart, even in the case of a federation with just the small manufacturers, most

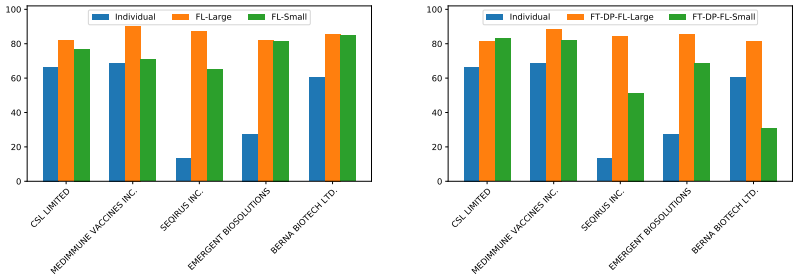

Fig. 3. F1 for small manufacturers when they are a part of a larger federation vs. a federation of only small manufacturers. The graph on left is for FL and the one on right is for FT-DP-FL

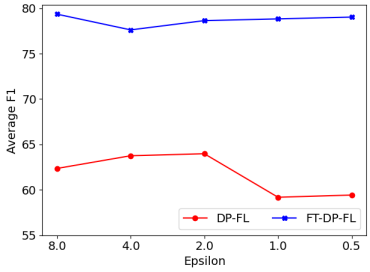

Fig. 4. Average F1 across users for the two differentially private FL variants.

of the manufacturers benefit significantly from participating. In fact, the performance of all manufacturers in the small federation closely tracks their performance in the large federation, with one exception.

5.6 Robustness to Differential Privacy Noise

Next, we study the effectiveness of personalization in recovering from the accuracy loss resulting from differential privacy noise. We vary the parameter ϵ and measure F1 averaged across users for two of the algorithm variants: differentially private federated learning (DP-FL) and the fine tuned differentially private federated learning (FT-DP-FL). As we can see from Figure 4, average F1 for DP-FL deteriorates significantly for values of ϵ less than 2. However, even in these cases, the personalized version, FT-DP-FL manages to retain its performance. We believe this is an important finding that provides significant latitude to differentially private FL frameworks to further tighten the privacy budget of ϵ without compromising utility.

6 Conclusion

Extracting mentions of vaccine adverse events using machine learning methods is an extremely urgent task right now. Federated Learning is a promising approach for

breaking down organizational and geographical barriers to collaboration on building very effective models to solve this problem. Our work demonstrates that the loss of accuracy incurred through adding additional layers of privacy can be mitigated by introducing personalization. We show that manufacturers with dataset of all different sizes can benefit from participating in such a federation and that it is stable to potential real world changes. In the future, we would like to investigate other approaches to personalization applied to this problem domain.

References

1. M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318, 2016.
2. M. G. Arivazhagan, V. Aggarwal, A. K. Singh, and S. Choudhary. Federated learning with personalization layers. *CoRR*, abs/1912.00818, 2019.
3. E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov. How to backdoor federated learning. In *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pages 2938–2948. PMLR, 2020.
4. K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. Kiddon, J. Konecný, S. Mazzocchi, H. B. McMahan, T. V. Overveldt, D. Petrou, D. Ramage, and J. Roselander. Towards federated learning at scale: System design. *CoRR*, 2019.
5. N. Carlini, C. Liu, Ú. Erlingsson, J. Kos, and D. Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium*, pages 267–284, 2019.
6. California consumer privacy act (ccpa). *https://oag.ca.gov/privacy/ccpa*.
7. K. Chaudhuri, C. Monteleoni, and A. D. Sarwate. Differentially private empirical risk minimization. *The Journal of Machine Learning Research*, 12:1069–1109, July 2011.
8. Y. Deng, M. M. Kamani, and M. Mahdavi. Adaptive personalized federated learning. *CoRR*, abs/2003.13461, 2020.
9. Differential Privacy Team. Learning with Privacy at Scale, *https://machinelearning.apple.com/2017/12/06/learning-with-privacy-at-scale.html*, 2017.
10. C. Dimitrakakis, B. Nelson, Z. Zhang, A. Mitrokotsa, and B. I. P. Rubinstein. Differential privacy for bayesian inference through posterior sampling. *The Journal of Machine Learning Research*, 18(1):343–381, Jan. 2017.
11. C. T. Dinh, N. H. Tran, and T. D. Nguyen. Personalized federated learning with moreau envelopes. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, virtual*, 2020.
12. C. Dwork. Differential privacy. In *Automata, Languages and Programming, 33rd International Colloquium, ICALP*, pages 1–12, 2006.
13. C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Conference on Theory of Cryptography*, TCC'06, pages 265–284, 2006.
14. C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, Aug. 2014.
15. A. Fallah, A. Mokhtari, and A. Ozdaglar. Personalized federated learning: A meta-learning approach, 2020.

16. M. Fredrikson, S. Jha, and T. Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22Nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1322–1333, 2015.
17. General data protection regulation (gdpr). *https://gdpr-info.eu/*.
18. C. Gentry. Fully homomorphic encryption using ideal lattices. In *Proceedings of the Forty-first Annual ACM Symposium on Theory of Computing*, pages 169–178, 2009.
19. R. C. Geyer, T. Klein, and M. Nabi. Differentially Private Federated Learning: A Client Level Perspective. *CoRR*, abs/1712.07557, 2017.
20. J. M. Giorgi and G. D. Bader. Transfer learning for biomedical named entity recognition with neural networks. *Bioinformatics*, 34(23):4087–4094, 2018.
21. H. Gurulingappa, A. M. Rajput, A. Roberts, J. Fluck, M. Hofmann-Apitius, and L. Toldo. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of Biomedical Informatics*, 45(5):885–892, 2012.
22. K. Haerian, D. Varn, S. Vaidya, L. Ena, H. Chase, and C. Friedman. Detection of pharmacovigilance-related adverse events using electronic health records and automated methods. *Clinical Pharmacology and Therapeutics*, 92(2):228–234, 2012.
23. R. Harpaz, A. Callahan, S. Tamang, Y. Low, D. Odgers, S. Finlayson, K. Jung, P. LePendu, and N. Shah. Text mining for adverse drug events: the promise, challenges, and state of the art. *Drug safety : an international journal of medical toxicology and drug experience*, 2014.
24. B. Hitaj, G. Ateniese, and F. Perez-Cruz. Deep models under the gan: Information leakage from collaborative deep learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 603–618, 2017.
25. K. Hsieh, A. Phanishayee, O. Mutlu, and P. B. Gibbons. The non-iid data quagmire of decentralized machine learning. *CoRR*, abs/1910.00189, 2019.
26. Innovatice medices initiative: Europe's partnership for health. *https://www.imi.europa.eu*.
27. Y. Jiang, J. Konecný, K. Rush, and S. Kannan. Improving federated learning personalization via model agnostic meta learning. *CoRR*, abs/1909.12488, 2019.
28. P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, R. G. L. D'Oliveira, S. E. Rouayheb, D. Evans, J. Gardner, Z. Garrett, A. Gascón, B. Ghazi, P. B. Gibbons, M. Gruteser, Z. Harchaoui, C. He, L. He, Z. Huo, B. Hutchinson, J. Hsu, M. Jaggi, T. Javidi, G. Joshi, M. Khodak, J. Konecný, A. Korolova, F. Koushanfar, S. Koyejo, T. Lepoint, Y. Liu, P. Mittal, M. Mohri, R. Nock, A. Özgür, R. Pagh, M. Raykova, H. Qi, D. Ramage, R. Raskar, D. Song, W. Song, S. U. Stich, Z. Sun, A. T. Suresh, F. Tramèr, P. Vepakomma, J. Wang, L. Xiong, Z. Xu, Q. Yang, F. X. Yu, H. Yu, and S. Zhao. Advances and open problems in federated learning. *CoRR*, abs/1912.04977, 2019.
29. S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. D. Smith. What can we learn privately? *CoRR*, abs/0803.0924, 2008.
30. J. Konecný, B. McMahan, and D. Ramage. Federated optimization: Distributed optimization beyond the datacenter. *CoRR*, abs/1511.03575, 2015.
31. J. Konecný, H. B. McMahan, D. Ramage, and P. Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. *CoRR*, abs/1610.02527, 2016.
32. I. Korkontzelos, A. Nikfarjam, M. Shardlow, A. Sarker, S. Ananiadou, and G. H. Gonzalez. Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts. *Journal of Biomedical Informatics*, 62:148–158, 2016.
33. A. Korolova. Privacy violations using microtargeted ads: A case study. In *2010 IEEE International Conference on Data Mining Workshops*, pages 474–482, 2010.
34. R. Leaman, L. Wojtulewicz, R. Sullivan, A. Skariah, J. Yang, and G. Gonzalez. Towards internet-age pharmacovigilance: Extracting adverse drug reactions from user posts in health-related social networks. In *Proceedings of the 2010 Workshop on Biomedical Natural*

Language Processing, BioNLP@ACL 2010, Uppsala, Sweden, July 15, 2010, pages 117–125. Association for Computational Linguistics, 2010.
35. P. LePendu, S. Iyer, A. Bauer-Mehren, R. Harpaz, J. Mortensen, T. Podchiyska, T. Ferris, and N. Shah. Pharmacovigilance using clinical notes. *Clinical Pharmacology and Therapeutics*, 93:547–555, 2013.
36. X. Li, Y. Gu, N. Dvornek, L. H. Staib, P. Ventola, and J. S. Duncan. Multi-site fmri analysis using privacy-preserving federated learning and domain adaptation: Abide results. *Medical Image Analysis*, 65:101765, 2020.
37. X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang. On the convergence of fedavg on non-iid data. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.
38. P. P. Liang, T. Liu, Z. Liu, R. Salakhutdinov, and L. Morency. Think locally, act globally: Federated learning with local and global representations. *CoRR*, abs/2001.01523, 2020.
39. Y. Mansour, M. Mohri, J. Ro, and A. T. Suresh. Three approaches for personalization with applications to federated learning. *CoRR*, abs/2002.10619, 2020.
40. H. B. McMahan, E. Moore, D. Ramage, and B. A. y Arcas. Federated learning of deep networks using model averaging. *CoRR*, abs/1602.05629, 2016.
41. H. B. McMahan, D. Ramage, K. Talwar, and L. Zhang. Learning differentially private language models without losing accuracy. *CoRR*, abs/1710.06963, 2017.
42. L. Melis, C. Song, E. D. Cristofaro, and V. Shmatikov. Inference attacks against collaborative learning. *CoRR*, abs/1805.04049, 2018.
43. New research consortium seeks to accelerate drug discovery using machine learning to unlock maximum potential of pharma industry data *https://www.janssen.com/emea/new-research-consortium- seeks-accelerate-drug-discovery-using-machine-learning-unlock-maximum*.
44. M. Nasr, R. Shokri, and A. Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE Symposium on Security and Privacy, SP 2019, San Francisco, CA, USA, May 19-23, 2019*, pages 739–753. IEEE, 2019.
45. D. W. Peterson, P. Kanani, and V. J. Marathe. Private federated learning with domain adaptation. *CoRR*, abs/1912.06733, 2019.
46. K. Roberts, D. Demner-Fushman, and J. M. Tonning. Overview of the TAC 2017 adverse reaction extraction from drug labels track. In *Proceedings of the 2017 Text Analysis Conference, TAC 2017, Gaithersburg, Maryland, USA, November 13-14, 2017*. NIST, 2017.
47. R. Shokri, M. Stronati, C. Song, and V. Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18, 2017.
48. V. Smith, C.-K. Chiang, M. Sanjabi, and A. Talwalkar. Federated multi-task learning, 2017.
49. F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart. Stealing machine learning models via prediction apis. In *Proceedings of the 25th USENIX Conference on Security Symposium*, pages 601–618, 2016.
50. R. Winnenburg, A. Sorbello, A. Ripple, R. Harpaz, J. Tonning, A. Szarfman, H. Francis, and O. Bodenreider. Leveraging medline indexing for pharmacovigilance - inherent limitations and mitigation strategies. *Journal of Biomedial Informatics*, 2015.
51. A. C. Yao. How to generate and exchange secrets. In *27th Annual Symposium on Foundations of Computer Science*, pages 162–167, 1986.
52. T. Yu, E. Bagdasaryan, and V. Shmatikov. Salvaging federated learning by local adaptation. *CoRR*, abs/2002.04758, 2020.