Federated Patient Hashing

Jie Xu,[1] **Zhenxing Xu,**[1] **Peter Walker,**[2] **Fei Wang**[1]
[1]Weill Cornell Medical College, Cornell University, USA
[2]U.S. Department of Defense Joint Artificial Intelligence Center, USA
{jix4002, zhx2005, few2001}@med.cornell.edu, peter.b.walker.mil@mail.mil

Abstract

Privacy concerns on sharing sensitive data across institutions are particularly paramount for the medical domain, which hinders the research and development of many applications, such as cohort construction for cross-institution observational studies and disease surveillance. Not only that, the large volume and heterogeneity of the patient data pose great challenges for retrieval and analysis. To address these challenges, in this paper, we propose a *Federated Patient Hashing* (FPH) framework, which collaboratively trains a retrieval model stored in a shared memory while keeping all the patient-level information in local institutions. Specifically, the objective function is constructed by minimization of a similarity preserving loss and a heterogeneity digging loss, which preserves both inter-data and intra-data relationships. Then, by leveraging the concept of Bregman divergence, we implement optimization in a federated manner in both centralized and decentralized learning settings, without accessing the raw training data across institutions. In addition to this, we also analyze the convergence rate of the FPH framework. Extensive experiments on real-world clinical data set from critical care are provided to demonstrate the effectiveness of the proposed method on similar patient matching across institutions.

Introduction

Nowadays more and more complex and heterogeneous healthcare data are becoming readily available. This provides an unprecedented opportunity for digging out insights from those data with the development of machine learning algorithms. However, due to privacy and other considerations, data access is very restricted, which limits the research and methodology development in many medical applications, such as cross-institution observational studies on cohort construction, clinical trial recruitment and disease surveillance (Lee et al. 2018; Vatsalan and Christen 2016). Therefore, it is crucial to develop an effective way to analyze fragmented and heterogeneous patient data from multiple institutions.

Hashing function, due to its noninvertible nature, is attractive in this scenario as they can map the sensitive data into compact binary codes. Rather than data-independent hashing methods (Athitsos et al. 2008; Chi, Li, and Zhu 2013; Indyk and Motwani 1998), there have been lots of research on learning hashing functions (*e.g.*, Laplacian Co-Hashing (LCH) (Zhang et al. 2010), Collaborative Hashing (CH) (Liu et al. 2014), Semantic Correlation Maximization (SCM) (Zhang and Li 2014), Spectral Hashing (SH) (Weiss, Fergus, and Torralba 2012), Anchor Graph Hashing (AGH) (Liu, Wang, and fu Chang 2011), and the latest deep hashing methods (Li, Wang, and Kang 2015; Zhu et al. 2016)) based on a given set of training data so that certain data characteristics (such as similarity) can be preserved. Analytics with the learned binary hash codes can lead to less memory consumption and short query time (Chi and Zhu 2017). However, the training process of these hashing functions is usually conducted on the entire data set. What's more, most methods focus on image or video retrieval, especially for the deep hashing approaches. Hence, they may not be suitable for healthcare settings where the patient data are heterogeneous and usually scattered across multiple different institutions (Denny et al. 2013; Newton et al. 2013).

Efficient learning of the retrieval model across institutions without exchanging raw patient data is also challenging. Although some data-independent hashing algorithms, *e.g.*, locality sensitive hashing (Indyk and Motwani 1998), can avoid accessing data, their computational complexity increases exponentially with the feature dimension and their retrieval accuracy is usually low comparing with learning to hash approaches. The most intuitive solution is to collaboratively learn a shared model stored in a central server while keeping all the raw training data on local clients, a.k.a. federated learning (Konečnỳ et al. 2016; Lee et al. 2018). Many researches have been conducted recently on different aspects of federated learning. For example, Konečnỳ *et al.* (Konečnỳ et al. 2016) and Caldas *et al.* (Caldas et al. 2018) worked on reducing the communication cost between local clients and server. Beyond that, aiming at non-independent identical distribution (non-IID) and unbalanced properties of the optimization, McMahan *et al.* (McMahan et al. 2016) proposed a practical method for the federated learning of deep networks based on iterative model averaging. Smith *et al.* (Smith et al.

2017) and Sahu *et al.* (Sahu et al. 2018) analyzed the convergence guarantee of federated optimization based on some strong assumptions.

In this paper, we propose a Federated Patient Hashing (FPH) model, which is trained in a distributed manner without data sharing across different institutions. The goal is to improve the local retrieval models by leveraging a more expansive view of patients. We also provide the convergence analysis as well. The main contributions are summarized as follows:

- Firstly, we formulate the general federated patient hashing problem, equipped with a similarity preserving loss and a heterogeneity digging loss. The former is used to preserve the similarity order, while the latter to capture the potential relationship within the heterogeneous data.
- Secondly, by leveraging the Bregman divergence, we develop both centralized and decentralized learning strategies to optimize our model in a federated manner without accessing the patient-level information across institutions and provide the convergence analysis as well based on some assumptions.
- Finally, we provide a specific loss function including a triplet ranking loss and a multi-modal consistency loss, and conduct empirical evaluations to validate the proposed framework. Our results on patients with Acute Kidney Injury (AKI) in critical care demonstrate the effectiveness of the proposed method.

Problem Definition

Notations

Let boldface lowercase letters like $\mathbf{z} \in \mathbb{R}^d$ be vectors with dimension d and boldface uppercase letters like $\mathbf{Z} \in \mathbb{R}^{d\times c}$ be matrices with size $d \times c$. The transpose of $\mathbf{Z}$ is denoted as $\mathbf{Z}^\top$ and the real numbers are denoted as uppercase letters like $Z \in \mathbb{R}$.

Problem Setting

Assume there are Q institutions (*e.g.*, hospitals), where the q-th institution contains the patient data $\mathbf{X}^{(q)} = [\mathbf{X}_1^{(q)\top}, ..., \mathbf{X}_M^{(q)\top}]^\top \in \mathbb{R}^{d\times N_q}$ with M modalities. The so-called modality is due to the heterogeneous of patient data, *e.g.*, the lab measures and clinical notes in electronic health records (EHRs) are usually regarded as different modalities. N_q is the number of patients in q-th institution and $d = \sum_{m=1}^{M} d_m$ is the feature dimension. $\mathbf{x}_{im}^{(q)}$ is the i-th column of $\mathbf{X}_m^{(q)}$, it represents the m-th modality of the i-th patient in q-th institution. $N = \sum_{q=1}^{Q} N_q$.

The task is to learn a hashing function that maps d-dimensional input $\mathbf{x} \in \mathbb{R}^d$ onto c-dimensional Hamming space $\mathbf{h} \in \mathcal{H} \equiv \{-1, +1\}^c$ through $\mathbf{h} = \text{sign}(f(\mathbf{x}; \mathbf{W}))$, where sign$(\cdot)$ denotes the element-wise sign function, and $f(\mathbf{x}; \mathbf{W}) = \mathbf{W}^\top\mathbf{x} : \mathbb{R}^d \rightarrow \mathbb{R}^c$ is a real-valued transformation[1].

During training process, exchanging patients' information across institutions is not allowed due to privacy policy. Meanwhile, the learnt mapping is usually stored in a shared memory which can be accessed by any institutions.

Overall Objective

As discussed earlier, the main objective is to learn compact hash codes to represent heterogeneous data across institutions without exchanging patient-level information. Besides the federated learning manner, there are two other things to consider: first, the similarity relationship between patients data should be preserved in the projected binary space; second, the heterogeneity of the patient data should be in view. In the following, we will detailedly formulate the federated patient hashing problem based on these two concepts.

Let $\mathbf{H}^{(q)}$ be the matrix of columns $\mathbf{h}_i^{(q)}, i = 1, ..., N_q$. After we access all the patients in Q institutions, the cumulative loss suffered on all the patients can be formulated as:

$$\min_{\mathbf{W}\in\mathcal{W}} \frac{1}{Q}\sum_{q=1}^{Q}\left(\mathcal{L}_s(\mathbf{X}^{(q)}; \mathbf{H}^{(q)}) + \sum_{m,m'=1}^{M} \mathcal{L}_h(\mathbf{H}_m^{(q)}; \mathbf{H}_{m'}^{(q)})\right) \quad (1)$$

where $\mathcal{L}_s$ is a general similarity preserving loss to preserve the similarity order, *i.e.,* minimizing the gap between the approximate nearest neighbor search result obtained from the input patient data space and the projected binary space. $\mathcal{L}_h$ can be seen as a heterogeneity digging loss, which is used to capture the potential relationship between different modalities of same patient.

Simple examples for these two kinds of losses are described in the next part.

Similarity Preserving Loss $\mathcal{L}_s$. Suppose we have the positive pairs by $(\mathbf{x}, \mathbf{x}^+) \in \mathcal{P}$ and negative pairs by $(\mathbf{x}, \mathbf{x}^-) \in \mathcal{N}$, where $(\mathbf{x}, \mathbf{x}^+)$ have same label and $\mathbf{x}^+$ belongs to the k-neighborhood of $\mathbf{x}$, $(\mathbf{x}, \mathbf{x}^-)$ have different labels. Simple examples for the similarity preserving loss can be:

- Pairwise similarity preserving loss:

$$\mathcal{L}_s = \mathsf{E}\{d_{\mathcal{H}}(\mathbf{h}, \mathbf{h}^+)|\mathcal{P}\} - \lambda\mathsf{E}\{d_{\mathcal{H}}(\mathbf{h}, \mathbf{h}^-)|\mathcal{N}\}, \quad (2)$$

where λ is a regularization parameter and $(\mathbf{h}, \mathbf{h}^+, \mathbf{h}^-)$ is the binary embedding of $(\mathbf{x}, \mathbf{x}^+, \mathbf{x}^-)$. The Hamming metric $d_{\mathcal{H}}(\mathbf{h}, \mathbf{h}^+) = c - \frac{1}{2}\sum_{i=1}^{c} |h_i + h_i^+|$ computes the embedding dissimilarity between the sample pair $(\mathbf{x}, \mathbf{x}^+)$. In practice, the conditional expectations $\mathsf{E}\{\cdot|\mathcal{P}\}$, $\mathsf{E}\{\cdot|\mathcal{N}\}$ are replaced by averages on a training set of positive pairs and negative pairs of embedding, respectively.

If we further denote the triplet constraints by $(\mathbf{x}, \mathbf{x}^+, \mathbf{x}^-) \in \mathcal{R}$, the mapping function is constructed such that the embedding from the same class are forced to be close to each other and those from different classes are pushed far away. Thus, another widely used triplet loss can be derived as following (Weinberger, Blitzer, and Saul 2006):

- Triplet ranking loss:

$$\mathcal{L}_s = \left[\mathsf{E}\{d_{\mathcal{H}}(\mathbf{h}, \mathbf{h}^+)|\mathcal{R}\} - \mathrm{E}\{d_{\mathcal{H}}(\mathbf{h}, \mathbf{h}^-)|\mathcal{R}\} + \varepsilon\right]_+ . \quad (3)$$

[1] Let assume the data be zero-centered, then we omit the bias term.

More complicated non-convex problems occur in neural networks, where the networks make prediction through a non-convex function of the feature vector instead of via the linear-in-the-feature mapping $\mathbf{W}^\top\mathbf{x}$, however, the resulting loss can still be written as $\mathcal{L}_s$ with the gradients computed efficiently using back-propagation (Konečný 2017).

Heterogeneity Digging Loss $\mathcal{L}_h$. For an individual patient, the data are usually of multiple forms, *e.g.*, structured data (including patient demographics, medications, lab results, *etc*), clinical notes (recorded by physicians and medical professionals), medical images, *etc*. A natural way is to view these multiple forms of data as multiple modalities.

Ideally, for each patient, we want the same retrieval results from any form of data. It could be satisfied by making different modalities of each sample share same hash codes. This could derive the following consistency loss.

- Multi-modal consistency loss:

$$\sum_{m,m'=1}^{M}\mathcal{L}_h = 2\sum_{m=1}^{M}\sum_{m'=m+1}^{M}\mathsf{E}\{d_{\mathcal{H}}(\mathbf{h}_m,\mathbf{h}_{m'})|\forall\mathbf{x}\}, \tag{4}$$

where $\mathbf{h}_m$ and $\mathbf{h}_{m'}$ are the hash codes of the m-th and m'-th modality of data $\mathbf{x}\in\mathbb{R}^d$, respectively.

Similarly, we can also construct the pairwise or the triplet ranking loss with regard to different modalities of positive and negative pairs. Let's take triplet loss as an example.

- Triplet modality ranking loss:

$$\mathcal{L}_h = \left[\mathsf{E}\{d_{\mathcal{H}}(\mathbf{h}_m,\mathbf{h}^{+}_{m'}) - d_{\mathcal{H}}(\mathbf{h}_m,\mathbf{h}^{-}_{m'})|\mathcal{R}\} + \varepsilon\right]_+ . \tag{5}$$

In this case, this triplet loss implies different modalities of the similar sample pairs should also be closed to each other, and accordingly, those from different classes are pushed far away.

After constructing an appropriate model for the patients retrieval task and heterogeneous data, another important issue is to apply this model to Q institutions in consideration of privacy concern. We will solve this problem in the next section.

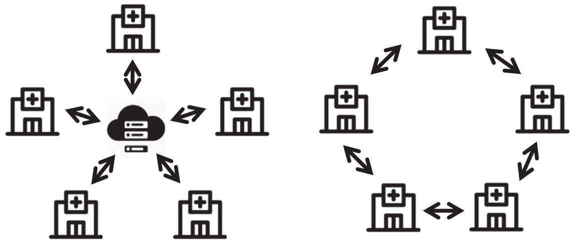
Optimization

To avoid exchanging patient-level information across institutions during training process, we propose to optimize (1) in a federated manner by two mechanisms, *i.e.*, centralized and decentralized learning strategy, as shown in Fig. 1.

Centralized Learning Strategy

Algorithm Description In centralized learning strategy, we assume there is one server (*i.e.*, a shared memory which can be accessed by any institutions) besides Q institutions. First, let's denote the local accumulated loss of q-th institution in Eq. (1) as follows:

$$\mathcal{L}(\mathbf{X}^{(q)};\mathbf{W}) = \mathcal{L}_s(\mathbf{X}^{(q)};\mathbf{H}^{(q)}) + \sum_{m,m'=1}^{M}\mathcal{L}_h(\mathbf{H}^{(q)}_{m};\mathbf{H}^{(q)}_{m'}). \tag{6}$$



(a) Centralized (b) Decentralized

Figure 1: Two learning strategies.

We assume that $\mathcal{L}(\mathbf{X};\mathbf{W})$ is differentiable in $\mathbf{W}\in\mathcal{W}$ after appropriate relaxation, and $\mathcal{W}$ is a closed convex subset. We denote the gradient of $\mathcal{L}$ with respect to $\mathbf{W}$ as $\nabla\mathcal{L}(\mathbf{X};\mathbf{W})$.

At iteration t, the server is connected with a set of institutions. After broadcasting model $\mathbf{W}^t$ to Q institutions, the server waits until receiving the updated model $\mathbf{W}^{t+1}_q$:

$$\mathbf{W}^{t+1}_q = \arg\min_{\mathbf{W}\in\mathcal{W}}\{\langle\nabla\mathcal{L}(\mathbf{X}^{(q)};\mathbf{W}^t),\mathbf{W}\rangle + \beta\mathcal{D}(\mathbf{W},\mathbf{W}^t)\}, \tag{7}$$

where $q=1,...,Q$ and $\mathcal{D}(\cdot)$ is the Bregman divergence generated by a differentiable 1-strongly convex function $h:\mathcal{W}\to\mathbb{R}$ with $\min_{\mathbf{W}\in\mathcal{W}} h(\mathbf{W})=0$, where $\mathcal{W}\subseteq\{\mathbf{W}:\|\mathbf{W}\|\leq\bar{B}\}$ (Lan, Lu, and Monteiro 2011). Also, suppose $D^2=\max_{\mathbf{U},\mathbf{V}\in\mathcal{W}}\mathcal{D}(\mathbf{U},\mathbf{V})$. In this paper, we choose:

$$\mathcal{D}(\mathbf{W},\mathbf{W}_q)\geq\frac{1}{2}\|\mathbf{W}-\mathbf{W}_q\|^2. \tag{8}$$

Finally, we update model $\mathbf{W}^t$ on the server via model averaging (Konečný et al. 2016):

$$\mathbf{W}^{t+1} = \frac{1}{Q}\sum_{q=1}^{Q}\mathbf{W}^{t+1}_q. \tag{9}$$

Our centralized learning strategy can be viewed as a stochastic Mirror Descent (SMD) method in federated environment, which we will call it Federated SMD (FSMD). The specific procedures are summarized in Alg. 1.

Algorithm 1 Centralized Learning Strategy to Minimize Eq. (1)

```
Input: $\mathcal{S}, \mathcal{R}, \mathbf{X}=[\mathbf{X}^{(1)},\mathbf{X}^{(2)},...,\mathbf{X}^{(Q)}]\in\mathbb{R}^{d\times N}$
Output: $\mathbf{W}\in\mathbb{R}^{d\times c}$
for <$t=1,...,T$> do
    for each institution $q$ in parallel do
        Pull $\mathbf{W}^t$ from the master;
        $\mathbf{W}^{t+1}_q = \arg\min_{\mathbf{W}\in\mathcal{W}}\{\langle\nabla\mathcal{L}(\mathbf{X}^{(q)};\mathbf{W}^t),\mathbf{W}\rangle + \beta\mathcal{D}(\mathbf{W},\mathbf{W}^t)\}$.
        Push $\mathbf{W}^{t+1}_q$ to the master;
    end for
    Update $\mathbf{W}^{t+1}=\frac{1}{Q}\sum_{q=1}^{Q}\mathbf{W}^{t+1}_q$.
end for
```

Convergence Analysis In order to analyze the convergence of the centralized learning strategy, we take the following commonly used assumptions.

Assumption 1. *(Bounded Second Moment)*

$$\|\nabla \mathcal{L}(\mathbf{X}^{(q)}; \mathbf{W})\|^2 \leq G^2, \quad \forall \mathbf{W}, q. \tag{10}$$

Assumption 2. *(Bounded Gradient Variance)*

$$\left\|\nabla \mathcal{L}(\mathbf{X}^{(q)}; \mathbf{W}) - \nabla \mathcal{L}(\mathbf{X}; \mathbf{W})\right\|^2 \leq \sigma^2, \forall \mathbf{W}, q. \tag{11}$$

Before introducing the main convergence result in Theorem 1, we first introduce a lemma that describes the behavior of Algorithm 1 at each iteration in q-th institution.

Lemma 1. *Under Assumptions 1 and 2, suppose that* $\mathbf{W}^* = \arg\min_{\mathbf{W} \in \mathcal{W}} \mathcal{L}(\mathbf{X}; \mathbf{W})$*, Algorithm 1 runs with convex* $\nabla \mathcal{L}(\mathbf{X}^{(q)}; \mathbf{W})$ *and step size* $\eta_s = \frac{\log S_q}{2\mu S_q}$ *on q-th institution satisfying the following inequality:*

$$E\|\mathbf{W}_q^{S_q} - \mathbf{W}^*\|^2 \leq \frac{\|\mathbf{W}_q^0 - \mathbf{W}^*\|^2}{S_q} + \frac{(G^2 + 2\beta^2 D^2)\log S_q}{2\mu^2 S_q} \tag{12}$$

where S_q *is the total number of iterations to optimize function (7) in q-th institution.*

It is easy to know that the right side of inequality (12) is zero when $S_q = \infty$. That is to say, the Algorithm 1 runs on q-th institution (*e.g.*, the inner **for** loop) is guaranteed to converge to local minima at a rate of $\mathcal{O}(\frac{\log S_q}{S_q})$. Next, we will derive the global convergence of Algorithm 1.

Theorem 1. *Under Assumptions 1 and 2, suppose that* $\mathbf{W}^* = \arg\min_{\mathbf{W} \in \mathcal{W}} \mathcal{L}(\mathbf{X}; \mathbf{W})$*, Algorithm 1 has the following convergence rate:*

$$E\|\mathbf{W}^T - \mathbf{W}^*\|^2 \leq \frac{\|\mathbf{W}^0 - \mathbf{W}^*\|^2}{S^T} + \frac{G^2 + 2\beta^2 D^2}{2\mu^2 S} \tag{13}$$

where $S = \min\{S_1, S_2, ..., S_Q\}$.

According to Theorem 1, the outer **for** loop of Algorithm 1 converges exponentially to the neighborhood of local minimum. It's worth noting that there is no communication assumptions for our centralized learning strategy, thus it is perfect for the proposed patient hashing scenario. However, the patient data is continuously accumulated, so it is natural to update the model using new patient data. To this end, in the next section, we introduce a decentralized learning strategy to update our model, which can make the most use of the new coming data.

Decentralized Learning Strategy

Algorithm Description In our decentralized learning strategy, we assume the algorithm sequentially goes over Q institutions, and before processing the q-th institution, produce an iterative $\mathbf{W}_q \in \mathcal{W}$.

We assume that $\mathcal{L}(\mathbf{X}; \mathbf{W})$ is L-smooth for any realization of $\mathbf{W}$. Namely, we assume that $\mathcal{L}(\cdot; \mathbf{W})$ is differential and that:

$$\|\nabla \mathcal{L}(\mathbf{W}_1) - \nabla \mathcal{L}(\mathbf{W}_2)\| \leq L\|\mathbf{W}_1 - \mathbf{W}_2\|. \tag{14}$$

Then, we solve the following problem to obtain the updated model $\mathbf{W}_{q+1}$:

$$\arg\min_{\mathbf{W} \in \mathcal{W}} \{\langle \nabla \mathcal{L}(\mathbf{X}^{(q)}; \mathbf{W}_q), \mathbf{W}\rangle + (L + \beta_{q+1})\mathcal{D}(\mathbf{W}, \mathbf{W}_q)\}. \tag{15}$$

The specific procedures are summarized in Alg. 2.

Algorithm 2 Decentralized Learning Strategy to Minimize Eq. (1)

```
Input: $\mathcal{S}, \mathcal{R}, \mathbf{X} = [\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, ..., \mathbf{X}^{(Q)}] \in \mathbb{R}^{d \times N}$
Output: $\mathbf{W} \in \mathbb{R}^{d \times c}$
for <$q = 1, ..., Q$> do
    $\mathbf{W}_{q+1} = \arg\min_{\mathbf{W} \in \mathcal{W}} \{\langle \nabla \mathcal{L}(\mathbf{X}^{(q)}; \mathbf{W}_q), \mathbf{W}\rangle + (L + \beta_{q+1})\mathcal{D}(\mathbf{W}, \mathbf{W}_q)\}$.
end for
```

Convergence Analysis Apparently, each iteration of Algorithm 2 has same convergence rate as the inner **for** loop of Algorithm 1 as long as $\mathcal{L}(\mathbf{X}; \mathbf{W})$ is based on same assumptions. To introduce the main convergence result in Theorem 2, we provide another lemma that could describe the behavior of Algorithm 2 at each iteration more detailedly.

Lemma 2. *Under Assumptions 1 and 2, suppose that* $\mathbf{W}^* = \arg\min_{\mathbf{W} \in \mathcal{W}} \mathcal{L}(\mathbf{X}; \mathbf{W})$*, and Algorithm 2 runs with a non-decreasing step size* $\beta_{q+1} \geq \beta_q$*, the iterations of Algorithm 2 satisfy the following inequality for all q:*

$$\begin{aligned}\mathcal{L}(\mathbf{W}_{q+1}) \leq\ & \mathcal{L}(\mathbf{W}^*) + (L + \beta_q)\mathcal{D}(\mathbf{W}^*, \mathbf{W}_q) \\ & - (L + \beta_{q+1})\mathcal{D}(\mathbf{W}^*, \mathbf{W}_{q+1}) + \frac{\|\mathbf{g}_q\|_*^2}{2\beta_q} \\ & + (\beta_{q+1} - \beta_q)D^2 + \langle \mathbf{g}_q, \mathbf{W}^* - \mathbf{W}_q\rangle\end{aligned} \tag{16}$$

where $\mathbf{g}_q = \nabla \mathcal{L}(\mathbf{X}^{(q)}; \mathbf{W}_q) - \nabla \mathcal{L}(\mathbf{X}; \mathbf{W}_q)$.

With Lemma 2, we could further derive the convergence rate of Algorithm 2.

Theorem 2. *Under Assumptions 1 and 2, suppose that each* $\mathbf{W}_q$ *is chosen from a fixed domain* $\mathcal{W} \subseteq \{\mathbf{W} : \|\mathbf{W}\| \leq \bar{B}\}$*, Algorithm 2 has the following convergence rate (Shamir 2016):*

$$\begin{aligned}& E\left[\frac{1}{Q}\sum_{q=1}^{Q} \mathcal{L}(\mathbf{X}; \mathbf{W}_q) - \mathcal{L}(\mathbf{X}; \mathbf{W}^*)\right] \\ & \qquad \leq \frac{2LD^2}{Q} + \frac{2\sigma D}{\sqrt{Q}} + \frac{2(12 + \sqrt{2}\bar{B}L)}{\sqrt{N}}.\end{aligned} \tag{17}$$

According to Theorem 2 and the definition of N and Q, we know that the right side of inequality (17) is zero when $Q = \infty$. While in practice, the number of institutions is usually limited. Then, we could sequentially goes over Q institutions for more than one pass to update the model (1).

Experiments

In this section, we evaluate the proposed model on a real-world clinical data set from critical care.

Table 1: MAP (%) of different hashing methods using 16 $\sim$ 64 bits on AKI case task.

# of Bits	LCH	CH	CCA	SCM	MLP$_{cen}$	FPH$_{de}$	FPH$_{cen}$
16	46.62 ± 0.17	47.74 ± 0.24	49.20 ± 0.21	53.52 ± 0.69	53.09 ± 3.96	53.30 ± 0.55	53.74 ± 1.57
32	47.28 ± 0.49	48.96 ± 0.05	49.60 ± 0.12	54.29 ± 0.36	54.46 ± 2.77	53.99 ± 0.90	54.40 ± 1.34
48	48.21 ± 0.39	49.39 ± 0.14	49.86 ± 0.11	53.94 ± 0.49	53.95 ± 2.29	53.91 ± 1.05	53.48 ± 0.52
64	48.71 ± 0.27	49.74 ± 0.05	50.07 ± 0.11	53.67 ± 0.50	54.58 ± 2.16	53.91 ± 0.97	53.26 ± 1.10

Table 2: MAP (%) of different hashing methods using 16 $\sim$ 64 bits on AKI stage task.

# of Bits	LCH	CH	CCA	SCM	MLP$_{cen}$	FPH$_{de}$	FPH$_{cen}$
16	42.43 ± 0.61	44.90 ± 0.16	43.71 ± 0.19	45.37 ± 0.22	47.49 ± 3.42	47.39 ± 0.61	48.30 ± 0.90
32	43.00 ± 0.67	46.27 ± 0.18	44.16 ± 0.16	45.31 ± 0.16	47.92 ± 1.66	47.93 ± 0.30	47.80 ± 1.04
48	43.44 ± 0.74	46.68 ± 0.14	44.62 ± 0.12	45.31 ± 0.13	48.25 ± 1.25	48.35 ± 0.65	48.66 ± 0.51
64	44.06 ± 0.59	47.10 ± 0.13	44.94 ± 0.07	45.27 ± 0.11	49.53 ± 0.60	48.55 ± 0.85	48.72 ± 0.63

Datasets

We evaluate the proposed model over EHR data acquired from Medical Information Mart for Intensive Care III (MIMIC-III) (Johnson et al. 2016). MIMIC-III is a freely and publicly-available database which encompasses a diverse and very large population of Intensive Care Unit (ICU) patients. Our experiments are designed to perform similar patient matching with regard to the onset of AKI and corresponding AKI stages in terms of severity during an ICU admission. As in (Li et al. 2018), a total of 16,558 ICU stays of 14,469 patients with structured data and electronic documentation are included in this study.

Evaluation Setup

We compare the proposed Federated Patient Hashing (FPH) model with the following methods: Canonical Correlation Analysis (CCA) (Hotelling 1992), Laplacian Co-Hashing (LCH) (Zhang et al. 2010), Collaborative Hashing (CH) (Liu et al. 2014), Semantic Correlation Maximization (SCM) (Zhang and Li 2014) and MLP (Zhang, Lai, and Feng 2018). CCA and SCM are two multi-modal hashing methods. CCA maps two views into a common latent space and SCM seamlessly integrates semantic labels into the hashing learning procedure. LCH simultaneously hashes both terms and documents according to their semantic similarities and CH fully explores the duality between two views. MLP is widely used in deep hashing methods to process text data.

For convenience, we represent our FPH method trained by centralized and decentralized strategy as "FPH$_{cen}$" and "FPH$_{de}$", respectively. Next, we will detailedly describe the specific loss function adopted in the experiments.

Similarity Preserving Loss $\mathcal{L}_s$. We choose the triplet ranking loss (3) with addition of a positive pair similarity preserving loss. Apparently, direct minimization of $\mathcal{L}_s$ is difficult since the term $\mathbf{h}$ involves a non-differentiable sign function, which is also inconsistent with our convergence analysis. Thus, we relax the problem by directly dropping the sigh function sign$(z) \approx z$, also let $\mathbf{D}_{\mathbf{x},\mathbf{y}} = (\mathbf{x}-\mathbf{y})(\mathbf{x}-\mathbf{y})^\top$, then we have $d_{\mathcal{H}}(\mathbf{h}, \mathbf{h}^+) = ||\mathbf{W}^\top\mathbf{x} - \mathbf{W}^\top\mathbf{x}^+||^2 = \text{Tr}(\mathbf{W}^\top\mathbf{D}_{\mathbf{x},\mathbf{x}^+}^\top\mathbf{W})$. Therefore, the similarity preserving loss $\mathcal{L}_s$ for the q-th institution can be formulated as:

$$\begin{aligned}\mathcal{L}_s(\mathbf{X}^{(q)};\mathbf{W}) &= \lambda_1 \mathsf{E}\left\{\text{Tr}(\mathbf{W}^\top\mathbf{D}_{\mathbf{x},\mathbf{x}^+}^\top\mathbf{W})\middle|\mathcal{S}^{(q)}\right\} \\ &+ \lambda_2\left[\mathsf{E}\left\{\text{Tr}(\mathbf{W}^\top(\mathbf{D}_{\mathbf{x},\mathbf{x}^-} - \mathbf{D}_{\mathbf{x},\mathbf{x}^+})^\top\mathbf{W})\middle|\mathcal{R}^{(q)}\right\} + \varepsilon\right]_+ .\end{aligned} \tag{18}$$

where $\mathcal{S}^{(q)}$ and $\mathcal{R}^{(q)}$ are the similar pairs and triplet constraints constructed from the patient data in q-th institution, respectively.

In the experiments, we choose one neighbor for each sample, *i.e.* $k = 1$. It's worth mentioning that patient data are usually with serious imbalance problem. For example, in the task of predicting the onset of a disease, the number of controls are far more than cases. In this case, we can reduce the imbalance problem to a certain extent by choosing more neighbors for the classes with less samples.

Heterogeneity Digging Loss $\mathcal{L}_h$. We choose the multi-modal consistency loss (4) in the experiment. Similarly, we relax the loss by dropping the sign function, then for any two modalities, we need to minimize $\|\mathbf{W}_i^\top\mathbf{X}_i - \mathbf{W}_j^\top\mathbf{X}_j\|_F^2, i, j = 1, ..., M$. Apparently, this leads to M dependent problems with M parameters $\mathbf{W}_i \in \mathbb{R}^{d_i \times b}, i = 1, ..., M$, which are tedious to solve. For convenience, we introduce the following two concatenate matrices:

$$\begin{aligned}&\mathbf{W} = [\mathbf{W}_1^\top, \mathbf{W}_2^\top, ..., \mathbf{W}_M^\top]^\top \in \mathbb{R}^{(d_1+...+d_M)\times b}, \\ &\tilde{\mathbf{X}}_m = [\mathbf{0}, ..., \mathbf{X}_m^\top, ..., \mathbf{0}]^\top \in \mathbb{R}^{(d_1+...+d_M)\times N}.\end{aligned} \tag{19}$$

It is clear that $\mathbf{W}_i^\top\mathbf{X}_i = \mathbf{W}^\top\tilde{\mathbf{X}}_i$. Let $\mathbf{Z}_{ij} = [\mathbf{0}, ..., \mathbf{0}, \mathbf{I}_i, \mathbf{0}, ..., \mathbf{0}, -\mathbf{I}_j, \mathbf{0}, ..., \mathbf{0}]$, $\tilde{\mathbf{X}} = [\tilde{\mathbf{X}}_1, ..., \tilde{\mathbf{X}}_M] = \text{diag}(\mathbf{X}_1, ..., \mathbf{X}_M)$ be a diagonal block matrix. Thus, we have the following heterogeneity digging loss for q-th institution:

$$\sum_{m,m'=1}^{M} \mathcal{L}_h = 2\sum_{m=1}^{M}\sum_{m'=m+1}^{M} \mathsf{E}\left\{\|\tilde{\mathbf{X}}^{(q)\top}\mathbf{W}\mathbf{Z}_{ij}\|_F^2\right\}. \tag{20}$$

Eq. (20) has only one parameter $\mathbf{W}$ to be solved. Such processing is mainly for the convenience of analysis. If the space complexity is a problem in reality, then this loss can still be viewed as multiple sub-problems.

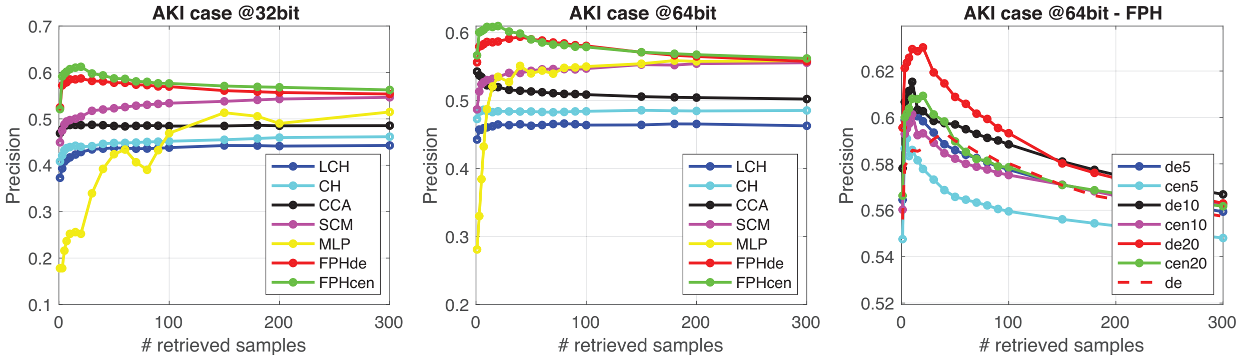


Figure 2: Precision curves on similar patient matching with regard to onset of AKI.

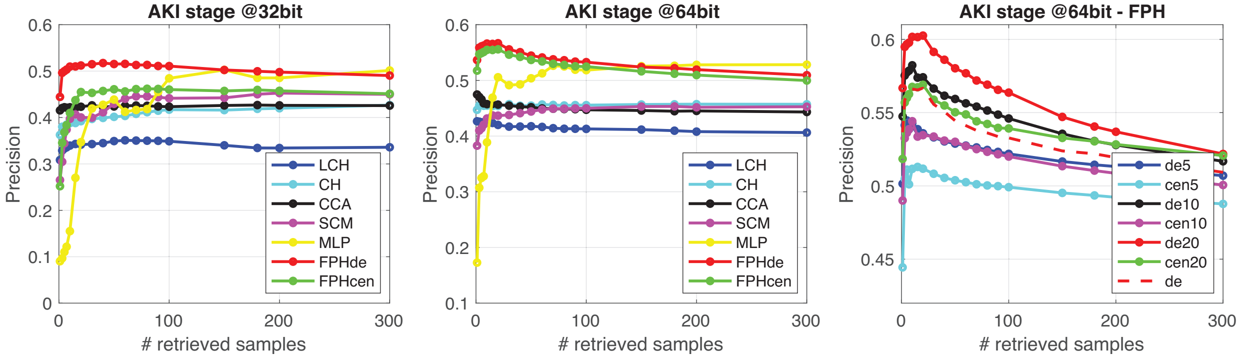


Figure 3: Precision curves on similar patient matching with regard to AKI stages in terms of severity.

The specific subgradient of $\mathcal{L}(\mathbf{X}^{(q)};\mathbf{W})$ with respect to $\mathbf{W}$ used in experiments for two learning strategies is:

$$\nabla\mathcal{L}(\mathbf{X}^{(q)};\mathbf{W}) = \frac{\alpha}{N_q}\sum_{i=1}^{M-1}\sum_{j=i+1}^{M} \mathsf{E}\left\{\tilde{\mathbf{X}}^{(q)}\tilde{\mathbf{X}}^{(q)\top}\mathbf{W}\mathbf{Z}_{ij}\mathbf{Z}_{ij}^{\top}\right\} + \lambda_1\mathsf{E}\left\{\mathbf{D}_{\mathbf{x},\mathbf{x}^+}^{\top}\mathbf{W}\middle|\mathcal{S}^{(q)}\right\} + \lambda_2\mathsf{E}\left\{(\mathbf{D}_{\mathbf{x},\mathbf{x}^-}-\mathbf{D}_{\mathbf{x},\mathbf{x}^+})^{\top}\mathbf{W}\middle|\mathcal{R}_+^{(q)}\right\}, \tag{21}$$

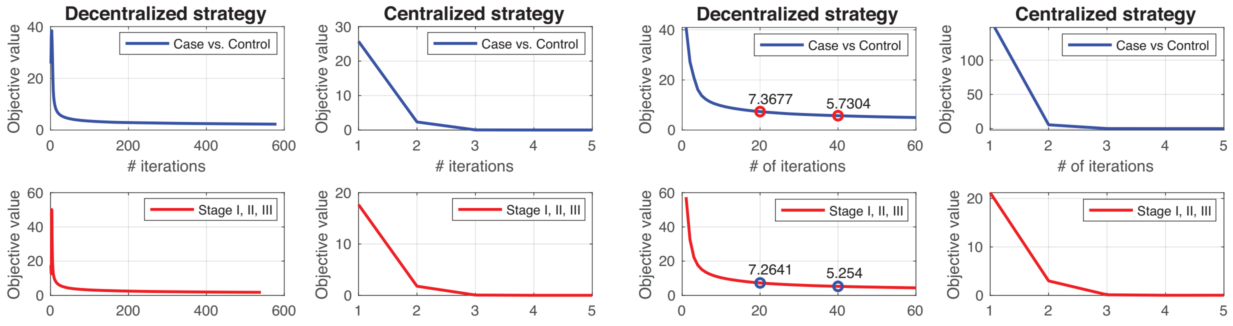
where $\mathcal{R}_+$ denotes the subset of constraints in $\mathcal{R}$ that is larger than 0.

Patient Matching: Onset of AKI & Corresponding Stage

Among 16558 ICU stays, 2785 were case patients who were diagnosed with AKI within a 7-day window after 24-hour observation. The remaining 13773 stays were controls (Li et al. 2018). In accordance with the AKI staging criteria from KDIGO (Kellum et al. 2012), the 2785 cases were further divided into three stages, *i.e.*, 1499 stays in stage I, 720 stays in stage II and 566 stays in stage III. In our setting, we treat structured data (*e.g.*, lab results and chart-events, *etc*) and unstructured data (*i.e.*, clinical notes) as two modalities. The preprocessing details are similar to (Xu et al. 2019). For unstructured data, clinical notes are first represented as "bag-of-words", and latent semantic analysis (LSA) is further applied to find the underlying meaning.

For all hashing function learners, the dataset is splited into 5 folds based on sample proportion, where 4 folds are used for training and 1 fold for testing. We repeat the process for five times, and gauge the mean of Mean Average Precision (MAP) (Buckley and Voorhees 2000) and standard deviation as final performance, as shown in Table 1 and Table 2. We abbreviate these two tasks, *i.e.*, similar patient matching on onset of AKI and corresponding stage, as "AKI case" and "AKI stage". In the experiments, we set $\lambda_1 = \lambda_2 = 0.5$, the other regularization parameters are tuned from range $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$. In addition, to simulate the cross-institution scenario for our algorithm, we randomly divide the training data into 20 subsets and avoid the communication across subsets in the training process. For decentralized learning strategy, we only access the full training data subsets by one full pass. From Tables 1 and 2, we can see that our model obtains good results even if we use less information (*i.e.*, avoiding exchanging patient-level information across institutions) than other compared methods.

We also compare precision performance using 32 and 64 bits on two tasks in left two subfigures in Figs. 2 and 3. The precision curves are plotted based on the first retrieved 300 patients. Apparently, our method achieves best results in both centralized and decentralized learning strategies. For the right subfigures in Figs.2 and 3, we choose a part of training subsets and optimize the model to converge (*i.e.*, we access the subsets for several passes). For example, "de5" means we choose 5 of 20 subsets to train the model to converge using decentralized learning strategy. The dash red



(a) # of samples are same (b) # of samples are different

Figure 4: Objective function value vs. number of iterations.

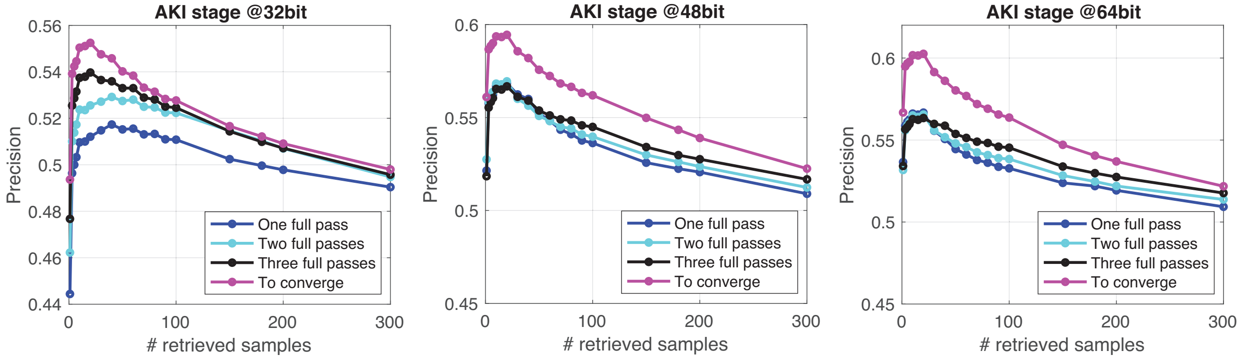


Figure 5: Precision curves on AKI stage task after accessing data for different passes by decentralized learning strategy.

line "de" represents we access the data for one full pass (*i.e.*, the "FPH$_{de}$" case reported in Tables 1 and 2). From the figure, we can see that our method obtains promising results even with a small part of data. Also, if we could access the data for multiple times and optimize the model to converge using decentralized strategy, our results would be better.

Further Analysis

Convergence of Learning Strategy. As we described before, we adopt two mechanisms, *i.e.*, centralized and decentralized learning strategy, to optimize our algorithm by considering the privacy of the sensitive data and theoretically prove the convergence as well.

To further illustrate the convergence of our learning strategies intuitively, we show the changes of objective values with the increase of iterations in Fig. 4. When we randomly divide the training data into 20 subsets to simulate the cross-institution scenario, we have two settings, *i.e.*, the number of samples in each subset is same or different. The objective function values versus number of iterations related to two settings are shown in Fig. 4. From the figure, we can see that whether the number of samples in each subset are same or not, the objective values of the proposed two learning strategies constantly decline. This property makes our model more applicable in practice.

Feasibility of Model Updating with New Samples. To test the feasibility of model updating with new samples by decentralized learning strategy, we plot the precision curves using 32, 48 and 64 bits on AKI stage task in Fig. 5. Different color lines in each subfigure represent the training data subsets can be accessed for different full passes when using decentralized learning strategy. Apparently, If we could optimize the model to converge, better results could be obtained. But the previous results have already shown the proposed algorithm obtains good results even if we access the data subsets for only one full pass. We recall the convergence analysis in the previous section, the convergence rate of the decentralized learning strategy we proved is related to the number of institutions. So it is natural to derive that with the increase of the number of institutions, the model would be better and close to the optimum.

In addition, to intuitively show the gap produced by insufficient iterations, let's take Fig. 4(a) as an example. Since we divide the training data for 20 subsets, it means the objective value at 20-th iteration is achieved when we access the data for one full pass. Apparently, the objectives decline most at first full pass. Therefore, if multiple accesses to data are not allowed in practice, we could sacrifice the performance obtained by sufficient iterations to some extent, and use the model updated by accessing the data for only one full pass.

Conclusion

In this paper, we focus on designing a federated patient hashing framework, which queries the potentially distributed, heterogeneous patient data scattered in multiple institutions, without exchanging patient-level information. Our framework includes a similarity preserving loss and a heterogeneity digging loss, in order to preserve both intra-data and inter-data relationships. Meanwhile, two learning mechanisms including decentralized and centralized strategies are adopted to optimize the proposed model in a federated manner with the proof of convergence. Finally, extensive experiments on a real-world clinical data set with simulated federated environment demonstrate the effectiveness and convergence of the proposed method.

Acknowledgments

This work is supported by NSF IIS 1716432, 1750326, ONR N00014-18-1-2585 and Amazon Web Service (AWS) Machine Learning for Research Award.

References

Athitsos, V.; Alon, J.; Sclaroff, S.; and Kollios, G. 2008. Boostmap: An embedding method for efficient nearest neighbor retrieval. *IEEE TPAMI* 30(1):89–104.

Buckley, C., and Voorhees, E. M. 2000. Evaluating evaluation measure stability. In *SIGIR*, 33–40. ACM.

Caldas, S.; Konečny, J.; McMahan, H. B.; and Talwalkar, A. 2018. Expanding the reach of federated learning by reducing client resource requirements. *arXiv preprint arXiv:1812.07210*.

Chi, L., and Zhu, X. 2017. Hashing techniques: A survey and taxonomy. *ACM Computing Surveys (CSUR)* 50(1):11.

Chi, L.; Li, B.; and Zhu, X. 2013. Fast graph stream classification using discriminative clique hashing. In *PAKDD*, 225–236. Springer.

Denny, J. C.; Bastarache, L.; Ritchie, M. D.; Carroll, R. J.; Zink, R.; Mosley, J. D.; Field, J. R.; Pulley, J. M.; Ramirez, A. H.; Bowton, E.; et al. 2013. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nature biotechnology* 31(12):1102.

Hotelling, H. 1992. Relations between two sets of variates. In *Breakthroughs in statistics*. Springer. 162–190.

Indyk, P., and Motwani, R. 1998. Approximate nearest neighbors: towards removing the curse of dimensionality. In *STOC*, 604–613. ACM.

Johnson, A. E.; Pollard, T. J.; Shen, L.; Li-wei, H. L.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Celi, L. A.; and Mark, R. G. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data* 3:160035.

Kellum, J. A.; Lameire, N.; Aspelin, P.; Barsoum, R. S.; Burdmann, E. A.; Goldstein, S. L.; Herzog, C. A.; Joannidis, M.; Kribben, A.; Levey, A. S.; et al. 2012. Kidney disease: improving global outcomes (kdigo) acute kidney injury work group. kdigo clinical practice guideline for acute kidney injury. *Kidney international supplements* 2(1):1–138.

Konečnỳ, J.; McMahan, H. B.; Yu, F. X.; Richtárik, P.; Suresh, A. T.; and Bacon, D. 2016. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*.

Konečnỳ, J. 2017. Stochastic, distributed and federated optimization for machine learning. *arXiv preprint arXiv:1707.01155*.

Lan, G.; Lu, Z.; and Monteiro, R. D. 2011. Primal-dual first-order methods with $\mathcal{O}(1/\epsilon)$ iteration-complexity for cone programming. *Mathematical Programming* 126(1):1–29.

Lee, J.; Sun, J.; Wang, F.; Wang, S.; Jun, C.-H.; and Jiang, X. 2018. Privacy-preserving patient similarity learning in a federated environment: Development and analysis. *JMIR medical informatics* 6(2):e20.

Li, Y.; Yao, L.; Mao, C.; Srivastava, A.; Jiang, X.; and Luo, Y. 2018. Early prediction of acute kidney injury in critical care setting using clinical notes. In *BIBM*, 683–686. IEEE.

Li, W.-J.; Wang, S.; and Kang, W.-C. 2015. Feature learning based deep supervised hashing with pairwise labels. *arXiv preprint arXiv:1511.03855*.

Liu, X.; He, J.; Deng, C.; and Lang, B. 2014. Collaborative hashing. In *CVPR*, 2139–2146.

Liu, W.; Wang, J.; and fu Chang, S. 2011. Hashing with graphs. In *ICML*.

McMahan, H. B.; Moore, E.; Ramage, D.; Hampson, S.; et al. 2016. Communication-efficient learning of deep networks from decentralized data. *arXiv preprint arXiv:1602.05629*.

Newton, K. M.; Peissig, P. L.; Kho, A. N.; Bielinski, S. J.; Berg, R. L.; Choudhary, V.; Basford, M.; Chute, C. G.; Kullo, I. J.; Li, R.; et al. 2013. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the emerge network. *JAMIA* 20(e1):e147–e154.

Sahu, A. K.; Li, T.; Sanjabi, M.; Zaheer, M.; Talwalkar, A.; and Smith, V. 2018. On the convergence of federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127*.

Shamir, O. 2016. Without-replacement sampling for stochastic gradient methods: Convergence results and application to distributed optimization. *arXiv preprint arXiv:1603.00570*.

Smith, V.; Chiang, C.-K.; Sanjabi, M.; and Talwalkar, A. S. 2017. Federated multi-task learning. In *NeurIPS*, 4424–4434.

Vatsalan, D., and Christen, P. 2016. Privacy-preserving matching of similar patients. *JBI* 59:285–298.

Weinberger, K. Q.; Blitzer, J.; and Saul, L. K. 2006. Distance metric learning for large margin nearest neighbor classification. In *NeurIPS*, 1473–1480.

Weiss, Y.; Fergus, R.; and Torralba, A. 2012. Multidimensional spectral hashing. In *ECCV*, 340–353. Springer.

Xu, Z.; Chou, J.; Zhang, X. S.; Luo, Y.; Isakova, T.; Adekkanattu, P.; Ancker, J. S.; Jiang, G.; Kiefer, R. C.; Pacheco, J. A.; et al. 2019. Identification of predictive sub-phenotypes of acute kidney injury using structured and unstructured electronic health record data with memory networks. *arXiv preprint arXiv:1904.04990*.

Zhang, D., and Li, W.-J. 2014. Large-scale supervised multimodal hashing with semantic correlation maximization. In *AAAI*.

Zhang, D.; Wang, J.; Cai, D.; and Lu, J. 2010. Laplacian co-hashing of terms and documents. In *ECIR*, 577–580. Springer.

Zhang, X.; Lai, H.; and Feng, J. 2018. Attention-aware deep adversarial hashing for cross-modal retrieval. In *ECCV*, 591–606.

Zhu, H.; Long, M.; Wang, J.; and Cao, Y. 2016. Deep hashing network for efficient similarity retrieval. In *AAAI*.