Published in final edited form as:

Proc Conf Assoc Comput Linguist Meet. 2019 August ; 2019: 283–291. doi:10.18653/v1/W19-5030.

Two-stage Federated Phenotyping and Patient Representation Learning

Dianbo Liu,
CHIP, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA, 02115

Dmitriy Dligach,
Loyola University Chicago, Chicago, IL, USA 60660

Timothy Miller
CHIP, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA, 02115

Abstract

A large percentage of medical information is in unstructured text format in electronic medical record systems. Manual extraction of information from clinical notes is extremely time consuming. Natural language processing has been widely used in recent years for automatic information extraction from medical texts. However, algorithms trained on data from a single healthcare provider are not generalizable and error-prone due to the heterogeneity and uniqueness of medical documents. We develop a two-stage federated natural language processing method that enables utilization of clinical notes from different hospitals or clinics without moving the data, and demonstrate its performance using obesity and comorbities phenotyping as medical task. This approach not only improves the quality of a specific clinical task but also facilitates knowledge progression in the whole healthcare system, which is an essential part of learning health system. To the best of our knowledge, this is the first application of federated machine learning in clinical NLP.

1 Introduction

Clinical notes and other unstructured data in plain text are valuable resources for medical informatics studies and machine learning applications in healthcare. In clinical settings, more than 70% of information are stored as unstructured text. Converting the unstructured data into useful structured representations will not only help data analysis but also improve efficiency in clinical practice (Jagannathan et al., 2009; Kreimeyer et al., 2017; Ford et al., 2016; Demner-Fushman et al., 2009; Murff et al., 2011; Friedman et al., 2004). Manual extraction of information from the vast volume of notes from electronic health record (EHR) systems is too time consuming.

To automatically retrieve information from unstructured notes, natural language processing (NLP) has been widely used. NLP is a subfield of computer science, that has been developing for more than 50 years, focusing on intelligent processing of human languages

dianbo.liu@childrens.harvard.edu.

(Manning et al., 1999). A combination of hard-coded rules and machine learning methods have been used in the field, with machine learning currently being the dominant paradigm.

Automatic phenotyping is a task in clinical NLP that aims to identify cohorts of patients that match a predefined set of criteria. Supervised machine learning is curently the main approach to phenotyping, but availability of annotated data hinders the progress for this task. In this work, we consider a scenario where multiple instituitions have access to relatively small amounts of annotated data for a particular phenotype and this amount is not sufficient for training an accurate classifier. On the other hand, combining data from these institutions can lead to a high accuracy classifier, but direct data sharing is not possible due to operational and privacy concerns.

Another problem we are considering is learning patient representations that can be used to train accurate phenotyping classifiers. The goal of patient representation learning is mapping the text of notes for a patient to a fixed-length dense vector (embedding). Patient representation learning has been done in a supervised (Dligach and Miller, 2018) and unsupervised (Miotto et al., 2016) setting. In both cases, patient representation learning requires massive amounts of data. As in the scenario we outlined in the previous paragraph, combining data from several institutions can lead to higher quality patient representations, which in turn will improve the accuracy of phenotyping classifiers. However, direct data sharing, again, is difficult or impossible.

To tackle the challenges we mentioned above, we developed a federated machine learning method to utilize clinical notes from multiple sources, both for learning patient representations and phenotype classifiers.

Federated machine learning is a concept that machine learning models are trained in a distributed and collaborative manner without centralised data (Liu et al., 2018a; McMahan et al., 2016; Bonawitz et al., 2019; Konečnỳ et al., 2016; Huang et al., 2018; Huang and Liu, 2019). The strategy of federated learning has been recently adopted in the medical field in structured data-based machine learning tasks (Liu et al., 2018a; Huang et al., 2018; Liu et al., 2018b). However, to the best of our knowledge, this work is the first time a federated learning strategy has been used in medical NLP.

We developed our two-stage federated natural language processing method based on previous work on patient representation (Dligach and Miller, 2018). The first stage of our proposed federated learning scheme is supervised patient representation learning. Machine learning models are trained using medical notes from a large number of hospitals or clinics without moving or aggregating the notes. The notes used in this stage need not be directly relevant to a specific medical task of interest. At the second stage, representations from the clinical notes directly related to the phenotyping task are extracted using the algorithm obtained from stage 1 and a machine learning model specific to the medical task is trained.

Clinicians spend a significant amount of time reviewing clinical notes. This time can be saved or reduced with reasonably designed NLP technologies. One such task is phenotying from medical notes. In this study, we demonstrated, using phenotyping from clinical note as a clinical task (Conway et al., 2011; Dligach and Miller, 2018), that the method we

developed will make it possible to utilize notes from a wide range of hospitals without moving the data.

The ability to utilize clinical notes distributed at different healthcare providers not only benefits a specific clinical practice task but also facilitates building a learning healthcare system, in which meaningful use of knowledge in distributed clinical notes will speed up progression of medical knowledge to translational research, tool development, and healthcare quality assessment (Friedman et al., 2010; Blumenthal and Tavenner, 2010). Without the needs of data movement, the speed of information flow can approach real time and make a rapid learning healthcare system possible (Slutsky, 2007; Friedman et al., 2014; Abernethy et al., 2010).

2 Methods

2.1 Study Cohorts

Two datasets were used in this study. The MIMIC-III corpus (Johnson et al., 2016) was used for representation learning. This corpus contains information for more than 58,000 admissions for more than 45,000 patients admitted to Beth Israel Deaconess Medical Center in Boston between 2001 and 2012. Relevant to this study, MIMIC-III includes clinical notes, ICD9 diagnostic codes, ICD9 procedure codes, and CPT codes. The notes were processed with cTAKES[1] to extract UMLS[2] unique concept identifiers (CUIs). Following the cohort selection protocol from (Dligach and Miller, 2018), patients with over 10,000 CUIs were excluded from this study. We obtained a cohort of 44,211 patients in total.

The Informatics for Integrating Biology to the Bedside (i2b2) Obesity challenge dataset was used to train phenotyping models (Uzuner, 2009). The dataset consists of 1237 discharge summaries from Partners HealthCare in Boston. Patients in this cohort were annotated with respect to obesity and its comorbidities. In this study we consider the more challenging *intuitive* version of the task. The discharge summaries were annotated with obesity and its 15 most common comorbidities, the presence, absence or uncertainty (questionable) of which were used as ground truth label in the phenotyping task in this study. Table 1 shows the number of examples of each class for each phenotype. Thus, we build phenotyping models for 16 different diseases.

2.2 Data Extraction and feature choice

At the representation learning stage (stage 1), all notes for a patient were aggregated into a single document. CUIs extracted from the text were used as input features. ICD-9 and CPT codes for the patient were used as labels for supervised representation learning.

At the phenotyping stage (stage 2), CUIs extracted from the discharge summaries were used as input features. Annotations of being present, absent, or questionable for each of the 16 diagnoses for each patient were used as multi-class classification labels.

[1] https://ctakes.apache.org
[2] https://www.nlm.nih.gov/research/umls/

2.3 Two-stage federated natural language processing of clinical notes

We envision that clinical textual data can be useful in at least two ways: (1) for pre-training patient representation models, and (2) for training phenotyping models.

In this study, a patient representation refers to a fixed-length vector derived from clinical notes that encodes all essential information about the patient. A patient representation model trained on massive amounts of text data can be useful for a wide range of clinical applications. A phenotyping model, on the other hand, captures the way a specific medical condition works, by learning the function that can predict a disease (e.g., asthma) from the text of the notes.

Until recently, phenotyping models have been trained from scratch, omitting stage (1), but recent work (Dligach and Miller, 2018) included a pretraining step, which derived dense patient representations from data linking large amounts of patient notes to ICD codes. Their work showed that including the pre-training step led to learning patient representations that were more accurate for a number of phenotyping tasks.

Our goal here is to develop methods for federated learning for both (1) pre-training patient representations, and (2) phenotyping tasks. These methods will allow researchers and clinicans to utilize data from multiple health care providers, without the need to share the data directly, obviating issues related to data transfer and privacy.

To achieve this goal, we design a two-stage federated NLP approach (Figure 1). In the first stage, following (Dligach and Miller, 2018), we pre-train a patient representation model by training an artificial neural network (ANN) to predict ICD and CPT codes from the text of the notes. We extend the methods from (Dligach and Miller, 2018) to facilitate federated training.

In the second stage, a phenotyping machine learning model is trained in a federated manner using clinical notes that are distributed across multiple sites for the target phenotype. In this stage, the notes mapped to fixed-length representations from stage (1) are used as input features and whether the patient has a certain disease is used as a label with one of the three classes: presence, absence or questionable.

In the following sections, we first describe a simple notes pre-processing step. We then discuss the method for pre-training patient representations and the method for training phenotyping models. Finally, we describe our framework for performing the latter two steps in a federated manner.

2.4 Pre-processing

All of our models rely on standardized medical vocabulary automatically extracted from the text of the notes rather than on raw text.

To obtain medically relevant information from clinical notes, Unified Medical Language System (UMLS) concept unique identifiers (CUIs) were extracted from each note using Apache cTAKES (https://ctakes.apache.org). UMLS is a resource that brings together many

health and biomedical vocabularies and standardizes them to enable interoperability between computer systems.

The Metathesaurus is a large, multi-purpose, and multi-lingual vocabulary that contains information about biomedical and health related concepts, their various names, and the relationships among them. The Metathesaurus structure has four layers, Concept Unique Identifies(CUIs), Lexical (term) Unique Identifiers (LUI), String Unique Identifiers (SUI) and Atom Unique Identifiers (AUI). In this study, we focus on CUIs, in which a concept is a medical meaning. Our models use UMLS CUIs as input.

2.5 Representation learning

We adapted the architecture from (Dligach and Miller, 2018) for pre-training patient representations. A deep averaging network (DAN) that consists of an embedding layer, an average pooling layer, a dense layer, and multiple sigmoid outputs, where each output corresponds to an ICD or CPT code being predicted.

This architecture takes CUIs as input and is trained using binary cross-entropy loss function to predict ICD and CPT codes. After the model is trained, the dense layer can be used to represent a patient as follows: the model weights are frozen and the notes of a new patient are fed into the network; the patient representation is collected from the values of the units of the dense layer. Thus, the text of the notes is mapped to a fixed-length vector using a pre-trained deep averaging network.

Algorithm 1:

Two-stage federated natural language processing

Stage 1

Input: MIMIC3 data clinical notes distributed at 10 simulated sites, Representation learning model

Output: 174 ICD or CPT codes

Extract CUIs from each patient's clinical notes using cTAKE.

for $t \in 1$ **to** T **do**

  for $k \in 1$ **to** K *in parallel* **do**

    Train patient representation learning model f_k

  end

  aggregate models from all sites by $W_{ag}^{t} = \sum_{k=1}^{K} \frac{n_k}{N} w_k^t$

end

;

Stage 2

Input: i2b2 clinical notes for obesity comorbidities distributed at 3 sites, phenotyping machine learning model

Output: 1 single binary output (one of the comorbidities)

Extract CUIs from each clinical notes using cTAKES.

for $t \in 1$ **to** $T^{'}$ **do**

  for $k \in 1$ **to** $K^{'}$ *in parallel* **do**

    Train phenotyping model $f_k^{'}$

  end

  aggregate models from all sites by $W_{ag}^{'t} = \sum_{k=1}^{K^{'}} \frac{n_k^{'}}{N^{'}} w_k^{'t}$

end

2.6 Phenotyping

A linear kernel Support Vector Machine (SVM) taking input from representations generated using the pre-trained model from stage 1 was used as the classifer for each phenotype of interest. No regularization was used for the SVM and stochastic gradient descent was used as the optimization algorithm.

2.7 Federated machine learning learning on clinical notes

To train the ANN model in either stage 1 or stage 2, we simulated sending out models with identical initial parameters to all sites such as hospitals or clinics. At each site, a model was trained using only data form that site. Only model parameters of the models were then sent back to the analyzer for aggregation but not the original training data. An updated model is generated by averaging the parameters of models distributively trained, weighted by sample size (Konečnỳ et al., 2016; McMahan et al., 2016). In this study, sample size is defined as the number of patients.

After model aggregation, the updated model was sent out to all sites again to repeat the global training cycle (Algorithm 1). Formally, the weight update is specified by:

$$W_{ag}^{t} = \sum_{k=1}^{K} \frac{n_k}{N} W_k^t \quad (1)$$

where W_{ag} is the parameter of aggregated model at the analyzer site, K is the number of data sites, in this study the number of simulated healthcare providers or clinics. n_i is the number of samples at the i^{th} site, N is the total number of samples across all sites, and W_i is the parameters learned from the i^{th} data site alone. t is the global cycle number in the range of [1,T]. The algorithm tries to minimize the following objective function:

$$\underset{f}{argmin}\left(-\sum_{j=1}^{N}\sum_{p=1}^{M}\left[y_{jp}logf\left(x_{jp}\right)+\left(1-y_{jp}\right)log\left(1-f\left(x_{jp}\right)\right)\right]\right)$$

Where x_j is the feature vector of CUIs. and y is the class label. p is the output number and M is the total number of outputs. f is the machine learning model such as artificial neural network or SVM. Codes that accompany this article can be found at our github repository [3].

3 Experiments

To imitate real world medical setting where data are distributed with different healthcare providers, we randomly split patients in MIMIC-III data into 10 sites for stage 1 (federated representation learning). The training data of i2b2 was split into 3 sites for stage 2 (phenotype learning) to mimic obesity related notes distributed with three different healthcare providers. i2b2 notes were not included in the representation learning as in clinic settings information exchange routes for disease-specific records are often not the same as general medical information and ICD/CPT codes were not available for i2b2 dataset.

Experiments were designed to answer three questions:

1. Whether clinical notes distributed in different silos can be utilized for patient representation learning without data sharing
2. Whether utilizing data from a wide range of sources will help improve performance of phenotyping from clinical notes
3. Whether models trained in a two-stage federated manner will have inferior performance to models trained with centralized data.

To answer these questions, two-stage NLP algorithms were trained. Performance of models trained using only i2b2 notes from one of the three sites were compared with two-stage federated NLP results. Furthermore, performance of machine learning models using distributed or centralized data at patient representation learning stage or phenotyping stage were compared.

[3] https://github.com/kaiyuanmifen/FederatedNLP

4 Results

4.1 Two-stage federated natural language processing improves performance of automatic phenotyping

We looked at the scenarios where no representation learning was performed. In those cases, the standard TF-IDF weighted sparse bag-of-CUIs vectors were used to represent i2b2 notes. The sparse vectors were used as input into the phenotyping SVM model. We also looked at the scenarios where representation learning was performed by predicting ICD codes. For each of these conditions, we trained our phenotyping models using centralized vs. federated learning. Finally, we considered a scenario where the phenotyping model was trained using the notes from a single site (the metrics we report were averaged across three sites).

To summarize, seven experiments were conducted:

1. No representation learning + centralized phenotyping learning
2. No representation learning + federated phenotyping learning where i2b2 training data were randomly split into 3 silos
3. No representation learning + single source phenotyping learning, where i2b2 data were randomly split into 3 silos, but phenotyping algorithm was only trained using data from one of the silos
4. Centralized representation learning + centralized phenotyping learning
5. Centralized representation learning + federated phenotyping learning
6. Federated representation learning + centralized phenotyping learning,where MIMIC-III data were randomly split into 10 silos
7. Federated representation learning + federated phenotyping learning, where MIMIC-III data were randomly split into 10 silos and i2b2 data into 3 silos (Table 2).

The results of our experiments are shown in Table 3. First of all, we looked at whether phenotyping model training can be conducted in a federated manner without compromising performance. When only i2b2 data from one of three silos was used for phenotyping training (experiment 3), the F1 score of 0.542 was achieved. When data from all three i2b2 sites were used for phenotyping model training (experiment 1) the F1 score improved to 0.634, which suggests that more data did improve the model. If we assume data from the three i2b2 silos can not be moved and aggregated together, the model trained in a federated manner (experiment 2) achieved a comparable F1 score of 0.632. This suggested federated learning worked for phenotyping model training.

Previous work showed that using learned representations from clinical notes from a different source using a transfer learning strategy helps to improve the performance of phenotyping NLP models (Dligach and Miller, 2018). When patient representations learned from centralized MIMIC-III notes were used as features and centralized phenotyping training was conducted (experiment 4), the phenotyping performance increased significantly with F1 score of 0.714, which was consistent with previous findings (Dligach and Miller, 2018).

When a federated approach was applied in both representation learning and phenotyping stages, the algorithm achieved F1 score of 0.724. It is worth pointing out that F1 scores from experiment 7, where both representation and phenotyping training were conducted in a federated manner, were not statistically different from F1 scores of experiment 4 over multiple rounds of experiment using different data shuffling and initialization. In comparison, when only data from a single simulated silo was used, the average F1 score 0.634. When the centralized approach was taken at both stages, the precision, recall and F1 score were 0.718, 0.711 and 0.714 respectively. These results suggested utilizing clinical notes from different silos in a federated manner did improve accuracy of the phenotyping NLP algorithm, and the performance is comparable to NLP trained on centralized data. The performance of federated NLP on each single obesity commodity were shown in Table 3. It is necessary to point out that it was impractical to conduct federated phenotyping training when the number of “questionable” cases for many diseases are small (Table 1). This is true for many diseases in the i2b2 dataset. In such situation, “questionable” cases were excluded from the training and testing process. Instead of 3-class classification, a 2-class binary classification of “presence” or “absence” were conducted. Therefore, the performance metrics can not be directly compared with results in the original i2b2 challenge, though the scores were similar.

5 Conclusion

In this article, we presented a two-stage method that conducts patient representation learning and obesity comorbidity phenotyping, both in a federated manner. The experimental results suggest that federated training of machine learning models on distributed datasets does improve performance of NLP on clinical notes compared with algorithms trained on data from a single site. In this study, we used CUIs as input features into machine learning models, but the same federated learning strategies can also be applied to raw text.

Acknowledgement

Research reported in this publication was supported by the National Library Of Medicine of the National Institutes of Health under Award Number R01LM012973. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Abernethy Amy P, Etheredge Lynn M, Ganz Patricia A, Wallace Paul, German Robert R, Neti Chalapathy, Bach Peter B, and Murphy Sharon B. 2010. Rapid-learning system for cancer care. Journal of Clinical Oncology, 28(27):4268. [PubMed: 20585094]

Blumenthal David and Tavenner Marilyn. 2010. The meaningful use regulation for electronic health records. New England Journal of Medicine, 363(6):501–504.

Bonawitz Keith, Eichner Hubert, Grieskamp Wolfgang, Huba Dzmitry, Ingerman Alex, Ivanov Vladimir, Kiddon Chloe, Konecny Jakub, Mazzocchi Stefano, McMahan H Brendan, et al. 2019. Towards federated learning at scale: System design. arXiv preprint arXiv:1902.01046.

Conway Mike, Berg Richard L, Carrell David, Denny Joshua C, Kho Abel N, Kullo Iftikhar J, Linneman James G, Pacheco Jennifer A, Peissig Peggy, Rasmussen Luke, et al. 2011. Analyzing the heterogeneity and complexity of electronic health record oriented phenotyping algorithms. In AMIA annual symposium proceedings, volume 2011, page 274. American Medical Informatics Association.

Demner-Fushman Dina, Chapman Wendy W, and McDonald Clement J. 2009. What can natural language processing do for clinical decision support? Journal of biomedical informatics, 42(5):760–772. [PubMed: 19683066]

Dligach Dmitriy and Miller Timothy. 2018. Learning patient representations from text. arXiv preprint arXiv:1805.02096.

Ford Elizabeth, Carroll John A, Smith Helen E, Scott Donia, and Cassell Jackie A. 2016. Extracting information from the text of electronic medical records to improve case detection: a systematic review. Journal of the American Medical Informatics Association, 23(5):1007–1015. [PubMed: 26911811]

Friedman Carol, Shagina Lyudmila, Lussier Yves, and Hripcsak George. 2004. Automated encoding of clinical documents based on natural language processing. Journal of the American Medical Informatics Association, 11(5):392–402. [PubMed: 15187068]

Friedman Charles, Rubin Joshua, Brown Jeffrey, Buntin Melinda, Corn Milton, Etheredge Lynn, Gunter Carl, Musen Mark, Platt Richard, Stead William, et al. 2014. Toward a science of learning systems: a research agenda for the high-functioning learning health system. Journal of the American Medical Informatics Association, 22(1):43–50. [PubMed: 25342177]

Friedman Charles P, Wong Adam K, and Blumenthal David. 2010. Achieving a nationwide learning health system. Science translational medicine, 2(57):57cm29–57cm29.

Huang Li and Liu Dianbo. 2019. Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records. arXiv preprint arXiv:1903.09296.

Huang Li, Yin Yifeng, Fu Zeng, Zhang Shifa, Deng Hao, and Liu Dianbo. 2018. Loadaboost: Loss-based adaboost federated machine learning on medical data. arXiv preprint arXiv:1811.12629.

Jagannathan Vasudevan, Mullett Charles J, Arbogast James G, Halbritter Kevin A, Yellapragada Deepthi, Regulapati Sushmitha, and Bandaru Pavani. 2009. Assessment of commercial nlp engines for medication information extraction from dictated clinical notes. International journal of medical informatics, 78(4):284–291. [PubMed: 18838293]

Johnson Alistair EW, Pollard Tom J, Shen Lu, Li-wei H Lehman, Feng Mengling, Ghassemi Mohammad, Moody Benjamin, Szolovits Peter, Celi Leo Anthony, and Mark Roger G. 2016. Mimic-iii, a freely accessible critical care database. Scientific data, 3:160035. [PubMed: 27219127]

Konečnỳ Jakub, McMahan H Brendan, Yu Felix X, Richtárik Peter, Suresh Ananda Theertha, and Bacon Dave. 2016. Federated learning: Strategies for improving communication efficiency. arXiv preprint arXiv:1610.05492.

Kreimeyer Kory, Foster Matthew, Pandey Abhishek, Arya Nina, Halford Gwendolyn, Jones Sandra F, Forshee Richard, Walderhaug Mark, and Botsis Taxiarchis. 2017. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. Journal of biomedical informatics, 73:14–29. [PubMed: 28729030]

Liu Dianbo, Miller Timothy, Sayeed Raheel, and Mandl Kenneth. 2018a. Fadl: Federated-autonomous deep learning for distributed electronic health record. arXiv preprint arXiv:1811.11400.

Liu Dianbo, Sepulveda Nestor, and Zheng Ming. 2018b. Artificial neural networks condensation: A strategy to facilitate adaption of machine learning in medical settings by reducing computational burden. arXiv preprint arXiv:1812.09659.

Manning Christopher D, Manning Christopher D, and Schütze Hinrich. 1999. Foundations of statistical natural language processing. MIT press.

McMahan H Brendan, Moore Eider, Ramage Daniel, Hampson Seth, et al. 2016. Communication-efficient learning of deep networks from decentralized data. arXiv preprint arXiv:1602.05629.

Miotto Riccardo, Li Li, Kidd Brian A, and Dudley Joel T. 2016. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Scientific reports, 6:26094. [PubMed: 27185194]

Murff Harvey J, FitzHenry Fern, Matheny Michael E, Gentry Nancy, Kotter Kristen L, Crimin Kimberly, Dittus Robert S, Rosen Amy K, Elkin Peter L, Brown Steven H, et al. 2011. Automated identification of postoperative complications within an electronic medical record using natural language processing. Jama, 306(8):848–855. [PubMed: 21862746]

Slutsky Jean R. 2007. Moving closer to a rapid-learning health care system. Health affairs, 26(2):w122–w124. [PubMed: 17259193]

Uzuner Özlem. 2009. Recognizing obesity and comorbidities in sparse data. Journal of the American Medical Informatics Association, 16(4):561–570. [PubMed: 19390096]

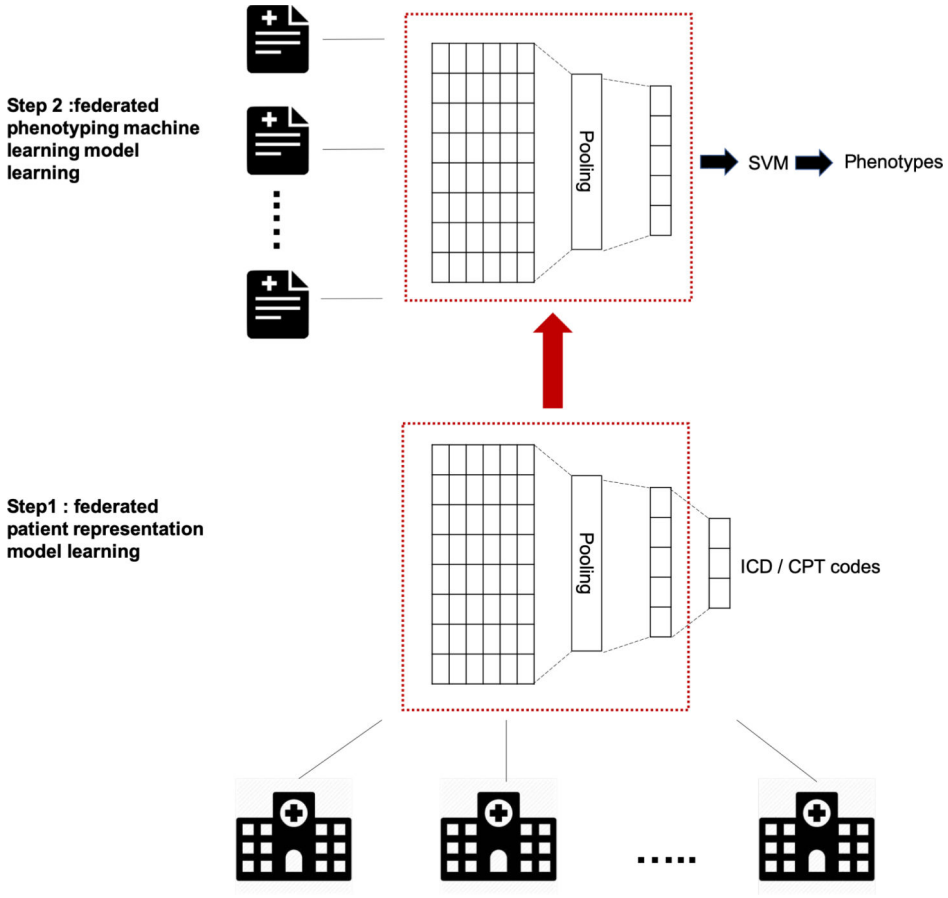

Figure 1:

Two stage federated natural language processing for clinical notes phenotyping. In the first stage, a patient representation model was trained using an artificial neural network (ANN) to predict ICD and CPT codes from the text of the notes from a wide range of healthcare providers. The model without output layer was then used as ”representation extractor” in the next stage. In the second stage, a phenotyping support vector machine model was trained in a federated manner using clinical notes for the target phenotype distributed across multiple silos.

Table 1:

i2b2 cohort of obesity comorbidities

Disease	#Absence	#Presence	#Questionable
Asthma	86	596	0
CAD	391	265	5
CHF	308	318	1
Depression	142	555	0
Diabetes	473	205	5
GERD	144	447	1
Gallstones	101	609	0
Gout	94	616	2
Hypercholesterolemia	315	287	1
Hypertension	511	127	0
Hypertriglyceridemia	37	665	0
OA	117	554	1
OSA	99	606	8
Obesity	285	379	1
PVD	110	556	1
Venous Insufficiency	54	577	0

Table 2:

Performance of different experiments

Experiment	Patient representations	Phenotyping	Precision	Recall	F1
1	Bag-of-CUIs	Centralized	0.649	0.627	0.634
2	Bag-of-CUIs	Federated	0.650	0.623	0.632
3	Bag-of-CUIs	Single source	0.552	0.540	0.542
4	Centralized learned	Centralized	0.749	0.714	0.726
5	Centralized learned	Federated	0.743	0.713	0.723
6	Federated learned	Centralized	0.729	0.716	0.715
7	Federated learned	Federated	**0.753**	**0.715**	**0.724**

Table 3:

Performance of two-stage federated NLP in obesity comobidity phenotyping by disease

Disease	Prec	Rec	F1
Asthma	0.941	0.919	0.930
CAD	0.605	0.606	0.605
CHF	0.583	0.588	0.585
Depression	0.844	0.774	0.801
Diabetes	0.879	0.873	0.876
GERD	0.578	0.543	0.558
Gallstones	0.775	0.619	0.650
Gout	0.948	0.929	0.938
Hypercholesterolemia	0.891	0.894	0.892
Hypertension	0.877	0.854	0.865
Hypertriglyceridemia	0.725	0.519	0.524
OA	0.531	0.520	0.525
OSA	0.627	0.594	0.609
Obesity	0.900	0.894	0.897
PVD	0.590	0.604	0.596
Venous Insufficiency	0.763	0.712	0.734
Average	0.753	0.715	0.724