Machine Teaching-based Efficient Labelling for Cross-unit Healthcare Data Modelling

Yang Wang[1,2], Xueping Peng[1]✉, Allison Clarke[2],
Clement Schlegel[2], and Jing Jiang[1]

[1] Australian AI Institute, University of Technology Sydney
yang.wang-17@student.uts.edu.au,
{xueping.peng, jing.jiang}@uts.edu.au
[2] Health Economics and Research Division, Australian Department of Health
{alvin.wang, allison.clarke, clement.schlegel}@health.gov.au

Abstract. A data custodian of a big organization (such as a Commonwealth Data Integrating Authority), namely teacher, can easily build an intelligent model which is well trained by comprehensive data collected from multiple sources. However, due to information security and privacy-related regulation requirements, full access to the well-trained intelligent model and the comprehensive training data is usually limited to the teacher only and not available to any unit (or branch) of that organization. Therefore, if a unit, namely student, needs an intelligent function similar to the trained intelligent model, the student has to train a similar model from scratch using the student's own dataset. Such a dataset is usually unlabelled, requiring a big workload on labelling. Inspired by the Iterative Machine Teaching, we propose a novel collaboration pipeline. It enables the teacher to iteratively guide the student to select samples that are most worth labelling from the student's own dataset, which significantly reduces the requirement for human labelling and, at the same time, prevents regulation and information security breaches. The effectiveness and efficiency of the proposed pipeline is empirically demonstrated on two publicly available healthcare datasets in comparison with baseline methods. This work has broad implications for the healthcare sector to facilitate data modelling in instances where the large labelled datasets are not accessible to each unit.

Keywords: Iterative Machine Teaching · Cross-units · Efficient labelling · Electronic Health Records.

1 Introduction

The use of large-scale complex health data, including Electronic Health Records (EHR), holds immense potential to better predict patient outcomes and understand disease cohorts [23, 25]. Although huge volumes of EHR are typically unlabelled and have privacy concerns, existing deep learning models [2, 7, 20, 16, 11, 19] have shown great success in healthcare applications by self-supervised learning [4, 15, 5, 18, 17]. However, deep learning models typically require a large amount of labelled data for training which is not always available in real-world settings. Although data linkage and sharing can sometimes reduce the requirement for human labelling, this is often hindered due to information security and privacy-related requirements and concerns [27, 14]. Consider the below scenario that often occurs in the real world:

A large data service provider in healthcare (e.g. a Commonwealth Data Integrating Authority) may have a large amount of valuable data on a wide range of health and welfare topics (e.g. linked comprehensive dataset). Such a data service provider may have responsibility to provide end-to-end data services to ensure strong evidence available to policymakers, service planners, researchers and the community. On the other hand, an approved data recipient (e.g. a small research team in a university) may only have limited access to a specific part of the linked comprehensive dataset for approved studies. It is much easier and more achievable for the large data service provider to train a high performing machine learning model using their comprehensive dataset. However, It is difficult for the small research team to train a similar model due to the limited data access.

Let's take a more specific example here: a small research group needs to train a machine learning model to classify patient cohorts (by disease) using their approved access to a ten percent sample of patient's pharmaceutical benefits claims data. However, this would require a large number of records to be labelled by humans (i.e. label disease type on thousands of pharmaceutical benefits claim history records) to construct a sufficient training data for achieving a good machine learning classifier for this specific task. Alternatively, it would be much easier for a large data integration organization to train the same machine learning classifier, because they may already have sufficient training data. For example, the disease type on each pharmaceutical benefits claim history can be easily found from a linked comprehensive dataset that is available to the integration organization (such as the diagnosis code from the linked hospital data). Therefore, the amount of time spent on constructing the required training dataset by humans for the large data integration organization is considerably smaller. However, in this example, given the information security and privacy-related regulation requirements, the small research group is not allowed to access any extra information so the large data integration organization won't be able to directly give the researchers the trained model and/or the required training dataset.

To overcome these limitations, this paper proposes a novel collaboration pipeline, namely **Ma**chine **Te**aching-based **Labelling** (MaTe-Labelling) framework. It enables the teacher to iteratively guide the student to select samples that are most worth labelling from the student's own dataset, which significantly reduces the requirement for human labelling and, at the same time, prevents regulation and information security breaches.

More specifically, the above-mentioned large data service provider is considered to be the teacher, and the approved data recipient (i.e. the small research team) is considered to be the student. In each iteration, the teacher leverages

MaTe-Labelling to construct an optimal sample set that is selected only from the data that the student has access to. Similar to the optimization task of the Iterative Machine Teaching, the optimal sample set is carefully selected by solving an optimization task that minimizes the difficulty of the selected samples and, at the same time, maximizes their usefulness [13]. Such an optimal sample set would then be returned to the student. After being labelled by domain experts, it becomes the most efficient training set for the student model in that iteration, outperforming any training set created by labelling without teacher guidance. Given the optimal sample sets are only selected from the data that the student has access to, there is no extra information released to the student.

Our main contributions are summarized as follows:

- We propose a novel Machine Teaching-based Labelling (MaTe-Labelling) framework. It enables iterative guidance on the student to select samples that are most worth labelling, which reduces the large human efforts for labelling.
- MaTe-Labelling enables teacher to provide efficient data services and strong guidance to student without releasing any extra information, which effectively prevents regulation and information security breaches.
- Extensive experiments are conducted on two public health datasets to demonstrate effectiveness and efficiency of the proposed pipeline.

The remainder of this paper is organised as follows: Section 2 briefly reviews the related work on iterative machine teaching, interactive machine learning and active leaning. Section 3 describes the proposed model. Section 4 presents the experiments and results for EHR data from three patient cohorts and Section 5 concludes the paper by summarising the research.

2 Related Work

2.1 Iterative Machine Teaching

Traditional machine teaching is to solve the problem of finding an optimal (usually minimal) training set given a machine learning algorithm (the student model) and a target [28, 29]. Iterative Machine Teaching was proposed afterwards and extends the traditional machine teaching from batch setting to iterative setting, enabling iterative student model to achieve faster convergence [13]. Specifically, the optimization task of the Iterative Machine Teaching is to minimize the difficulty of the selected samples and, at the same time, maximize their usefulness.

2.2 Interactive Machine Learning

Interactive machine learning has been proposed as a promising field in visual analytics [6, 12, 26], which couples human input with machines in the learning process. Recently, machine teaching has been combined with interactive machine learning to improve human teacher by giving teaching guidance via performing a classification task by showing examples [3]. To address the crowdsourcing problem, a model called STRICT [24] has been introduced as an efficient algorithm for selecting examples to teach crowd workers to better classify the query. These studies consider a very different setting where the learner is not iterative and does not have a particular optimization algorithm [13].

2.3 Active Learning

Active learning (also called query learning) enables the learner to choose the data from which it learns and ask an oracle for its label, which performs better with less training [22, 21]. Active learning is different from machine teaching in the sense that active learners explore the optimal parameters by itself rather than being guided by the teacher. They therefore have different sample complexities [1, 28, 13].

3 Methodology

This section starts with notations of several important concepts and settings in the paper. The remainder mainly focuses on details of the proposed pipeline consisting of machine teaching and example selection.

3.1 Notations and Settings

Notations. We denote an example for the teacher as (x, y) while the same example for the student as $(\widetilde{x}, \widetilde{y})$. We assume the representation spaces of $x \in \mathcal{X}$ and $\widetilde{x} \in \widetilde{\mathcal{X}}$ are the same, and $y = \widetilde{y}$. v^* and w^* are teacher's optimal model and student's optimal model, respectively. In this paper, we assume v^* is the same as w^*. The initial parameter is denoted as w_0, loss function as $\ell(f(x), y)$, learning rate as η_t over time (and initial η_0) and the trackability of the parameter as w^t, where t denotes the t-th iteration.

Settings. The paper introduces the following settings to describe the proposed model.

- ***Student's Components***: The initial parameter w_0, loss function, optimization algorithm, representation, model, learning rate η_t and the trackability of the parameter w^t.
- ***Model***: The teacher uses a model with parameter v^* (w^* for student's space) that is taught to the student. w and v do not necessarily lie in the same space, but in this paper, they are equivalent and interchangeably used.
- ***Communication***: The teacher can only communicate with the student via examples. In this paper, the teacher provides one example x^t in the iteration t.

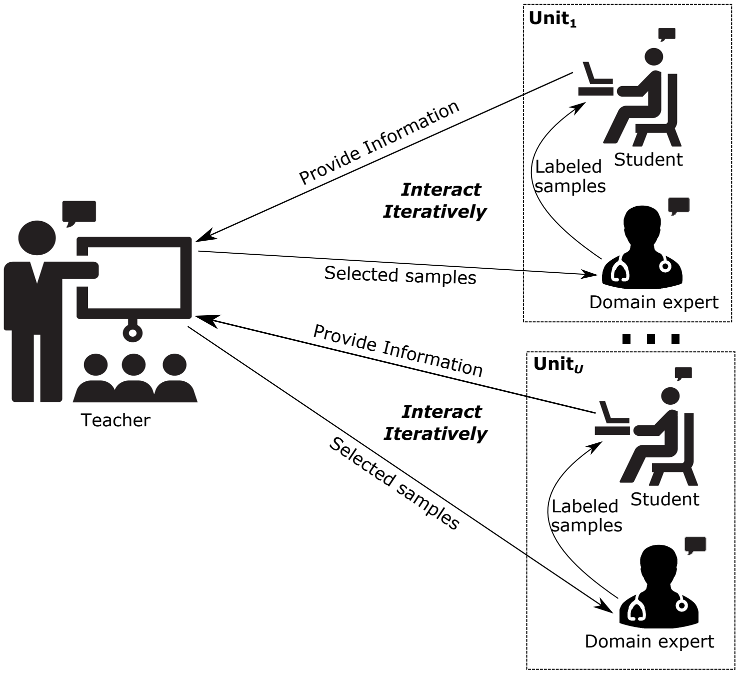

Fig. 1. The structure of the proposed model MaTe-Labelling.

- ***Loss Function***: The teacher and student share the same loss function. We assume this is a convex loss function $\ell(f(x), y)$, and the best model is usually found by minimizing the expected loss below: $w^* = \arg\min_{w} \ \mathbb{E}_{(x,y)} \left[\ell(\langle w, x \rangle, y) \right]$, where the sampling distribution $(x, y) \sim \mathbb{P}(x, y)$.
- ***Algorithm***: The student uses the stochastic gradient descent to optimize the model. The iterative update is $w^{t+1} = w^t - \eta_t \dfrac{\partial \ell(\langle w, x \rangle, y)}{\partial w}$.

3.2 Model Structure

As shown in Figure 1, the whole model architecture of MaTe-Labelling consists of one Teacher and a set of Units and each Unit includes a Student and a Domain expert. In each iteration, a Student from an Unit first provides information about the Student's data access and current parameters to the Teacher. The Teacher then solves an optimization task to be able to select an optimal sample set from the data (usually unlabeled) that the Student has access to. Then the selected optimal sample set is returned to the Unit for Domain expert labelling. The labelled sample set would then be used to update the Student model in the Unit.

3.3 Machine Teaching

An teacher has access to the student's feature space, model, loss function and optimization algorithm [13]. In specific, teacher's (x, y) and student's $(\widetilde{x}, \widetilde{y})$ share the same representation space.

Teaching algorithm. In order to make the student model converge faster with a smaller set of samples, the algorithm will start with looking into the difference between the current student parameter and the teacher parameter w^* during each iteration:

$$\begin{aligned} \|w^{t+1} - w^*\|_2^2 &= \left\| w^t - \eta_t \frac{\partial \ell(\langle w, x \rangle, y)}{\partial w} - w^* \right\|_2^2 \\ = \|w^t - w^*\|_2^2 &+ \eta_t^2 \left\| \frac{\partial \ell(\langle w^t, x \rangle, y)}{\partial w^t} \right\|_2^2 \\ - 2\eta_t &\left\langle w^t - w^*, \frac{\partial \ell(\langle w^t, x \rangle, y)}{\partial w^t} \right\rangle \\ = \|w^t - w^*\|_2^2 &+ \eta_t^2 T_1(x, y|w^t) - 2\eta_t T_2(x, y|w^t) \end{aligned} \tag{1}$$

where $T_1(x, y|w^t) = \left\| \frac{\partial \ell(\langle w^t, x \rangle, y)}{\partial w^t} \right\|_2^2$ and $T_2(x, y|w^t) = \left\langle w^t - w^*, \frac{\partial \ell(\langle w^t, x \rangle, y)}{\partial w^t} \right\rangle$. Based on the decomposition of the parameter error, the teacher aims to choose a particular example (x, y) such that $\|w^{t+1} - w^*\|_2^2$ is most reduced compared to $\|w^t - w^*\|_2^2$ from the last iteration. Thus the general strategy for the teacher is to choose an example (x, y), such that $\eta_t^2 T_1 - 2\eta_t T_2$ is minimized in the t-th iteration:

$$\underset{x \in \mathcal{X}, y \in \mathcal{Y}}{\arg\min} \; \eta_t^2 T_1(x, y|w^t) - 2\eta_t T_2(x, y|w^t). \tag{2}$$

The smallest value of $\eta_t^2 T_1 - 2\eta_t T_2$ is $-\|w^t - w^*\|_2^2$. If the teacher achieves this, it means that we have reached the teaching goal after this iteration. However, it usually cannot be done in just one iteration, because of the limitation of teacher's capability to provide examples. T_1 and T_2 have some nice intuitive interpretations:

Difficulty of an example. T_1 quantifies the difficulty level of an example. The difficulty level is not related to the teacher w^*, but is based on the current parameters of the learner w^t. From another perspective, the difficulty level can also be interpreted as the information that an example carries. Essentially, a difficult example is usually more informative. In such sense, our difficulty level has similar interpretation to curriculum learning, but with different expression.

Usefulness of an example. T_2 quantifies the usefulness of an example. Concretely, T_2 is the correlation between discrepancy $w^t - w^*$ and the information (difficulty) of an example. If the information of the example has large correlation with the discrepancy, it means that this example is very useful in this teaching iteration.

3.4 Example Selection

Teacher can access a comprehensive training set with labels generated by both domain experts and linked data that only teacher has access to. Due to information security and privacy-related regulation requirements, student from any individual unit does not have access to the same comprehensive training set that the teacher has access to. In addition, each unit is only allowed to see it's own data at all times. To this end, the optimal samples that are returned by teacher to a unit are only selected from the data that the very unit has access to. There is no extra information released to any unit. We take unit u as an example to formalize example select as below,

$$(x_u^t, y_h^t) = \underset{x_u \in \mathcal{X}_u, y_h \in \mathcal{Y}_h}{\arg\min} \; \eta_t^2 \left\| \frac{\partial \ell(\langle w_u^t, x_u \rangle, y_h)}{\partial w_u^t} \right\|_2^2 - 2\eta_t \left\langle w_u^t - w^*, \frac{\partial \ell(\langle w_u^t, x_u \rangle, y_h)}{\partial w_u^t} \right\rangle \tag{3}$$

where $\mathcal{X}_u$ denotes the set of samples collected from or available to unit u, and $\mathcal{Y}_h$ denotes corresponding labels for $\mathcal{X}_u$ that are available to the student, but not to the student, w_u^t represents the parameters of the student on unit u in the iteration t.

The optimal sample(s) selected by the teacher would be sent to the domain expert in unit u to be labelled. We set y_u^t as the label of x_u^t, thus, the student model update on unit u can be formalized as following,

$$w_u^t = w_u^{t-1} - \eta_t \frac{\partial \ell\left(\langle w_u^{t-1}, x_u^t \rangle, y_u^t\right)}{\partial w_u^{t-1}}. \tag{4}$$

The proposed MaTe-Labelling algorithm is summarized in Alg.1.

4 Experiments

In this section, we conduct experiments on two real world medical claim datasets to evaluate the performance of the proposed MaTe-Labelling. Compared with the baseline models, MaTe-Labelling yields better performance on different evaluation strategies.

4.1 Data Description

Dataset. We conducted comparative studies on two real-world datasets in the experiments, which are the MIMIC-III [9] and MIMIC-IV [8] databases.

- The MIMIC-III dataset [9] is an open-source, de-identified dataset of ICU patients and their EHRs between 2001 and 2012. The diagnosis codes in the dataset follow the ICD9 standard.
- The MIMIC-IV dataset [8] is an update to M-III, which incorporates contemporary data and improves on numerous aspects of M-III. The dataset consists of the medical records of 73,452 patients between 2008 and 2019.

Algorithm 1 MaTe-Labelling Algorithm

```
Randomly initialize the student and teacher parameter $w^0$;
Train teacher with comprehensive training set to get optimal teacher parameter $w^*$;
Set $t = 1$ and the maximal iteration number $T$, $u = 0$ and the total unit number $U$;
while $u < U$ do
  while $w_u^t$ has not converged or $t < T$ do
    Solve the optimization (e.g., pool-based teaching):
      $$(x_u^t, y_h^t) = \underset{x_u \in \mathcal{X}_u, y_h \in \mathcal{Y}_h}{\arg\min} \quad \eta_t^2 \left\| \frac{\partial \ell(\left\langle w_u^t, x_u \right\rangle, y_h)}{\partial w_u^t} \right\|_2^2 - 2\eta_t \left\langle w_u^t - w^*, \frac{\partial \ell(\left\langle w_u^t, x_u \right\rangle, y_h)}{\partial w_u^t} \right\rangle$$
    Domain expert labels $x_u^t$ as $y_u^t$ to perform the update:
      $$w_u^t = w_u^{t-1} - \eta_t \frac{\partial \ell\left(\left\langle w_u^{t-1}, x_u^t \right\rangle, y_u^t\right)}{\partial w_u^{t-1}}.$$
    $t \leftarrow t + 1$
  end while
  $u \leftarrow u + 1$
end while
```

Cohort Identification. Patients were included in analysis if they had at least one cohort-specific International Classification of Diseases (ICD)-9 diagnosis code of 140-239 for cancer, 428 for heart failure and 249-250 for diabetes. Table 1 shows the statistical details of the three cohorts in the datasets, where the selected patients made at least two visits and the labels are identified by ICD codes in an indexed visit. If a visit includes a cohort-specific ICD-9 code (considered to be the index visit), the previous visits which represent a patient's sequential history are used as input. To evaluate the performance of the algorithm, we perform classification on the selected data and the classification is based on the original labels.

Table 1. Statistics of the datasets.

Cohort	**MIMIC-III**	**MIMIC-IV**
# of diabetes patients	943	10,640
# of heart failure patients	1,021	13,551
# of cancer patients	1,333	6,167

4.2 Experimental Setup

Performance Metric and Baseline. We evaluate the convergence performance with following metric: the average classification accuracy on testing set over U units. We compare our proposed framework to a random labelling strategy, which is teacher-free.

Implementation Details. We implement all the approaches with Pytorch 1.7.0. For the training models, we use Adam [10] with 1 patient per iteration on both MIMIC-III and MIMIC-IV. We randomly split the data into a training set and test set with ratio of 80% to 20%. The drop-out strategies (the drop-out rate is 0.1) are used for all the approaches. We set dimension $d = 200$ for all the baselines and the proposed model.

4.3 Experimental Results

Prediction Performance of Trained Student. Table 2. shows the average testing accuracy and corresponding standard deviations of the MaTe-Labelling pipeline compared with the baseline for the predictive tasks over three cohorts in the two MIMIC datasets. The results show that the proposed MaTe-Labelling pipeline outperforms baseline Random strategy on both MIMIC-III and MIMIC-IV datasets. It is obvious that the benefits of machine teaching to import global knowledge to students on individual units by selecting useful and informative samples. Specifically, the average testing accuracy of MaTe-Labelling increases by 9.2% on the task of Cancer vs. Diabetes compared to Random strategy.

Table 2. Performance comparison of classification tasks.

Dataset	**Model**	**Testing Accuracy (%)**		
		Canc. vs Diab.	**Canc. vs Heart.**	**Diab. vs Heart.**
MIMIC-III	Random	65.81 ± 5.68	73.19 ± 3.81	80.63± 1.00
	MaTe-Labelling	**75.01** ± 1.73	**74.47** ± 3.74	**82.36**± 2.62
MIMIC-IV	Random	75.58 ± 6.41	73.10 ± 7.00	78.06 ± 4.63
	MaTe-Labelling	**78.99** ± 2.49	**74.54** ± 5.15	**79.47** ± 5.93

Testing Accuracy over Iterations. Fig. 2 and 3 depict the testing accuracy for all models over three prediction (Cancer vs. Diabetes, Cancer vs. Heart Failure and Diabetes vs. Heart Failure) tasks on both MIMIC datasets with iteration number varying from 1 to 200. The two figures show that MaTe-Labelling outperforms the baseline model with increasing iteration number.

The results in Fig. 2 show that the student model can converge much faster using the example provided by the teacher and labelled by domain expert, showing the effectiveness of our MaTe-Labelling pipeline. Particularly, we find that the MaTe-Labelling consistently achieves faster convergence than the random labelling over task of Cancer vs. Diabetes on MIMIC-III dataset. The results over task of Diabetes vs. Heart Failure also show that the MaTe-Labelling is much stable than random labelling.

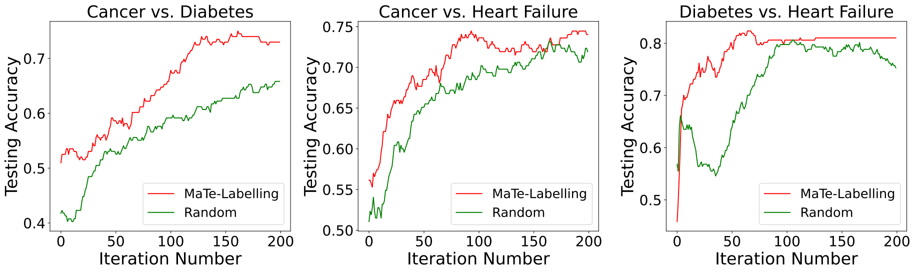

Fig. 2. Average Testing Accuracy on MIMIC-III with 3 units.

In contrast to the results on MIMIC-III, Fig. 3 show that the random strategy achieves better performance when the iteration number is less than 50. One possible reason is that the size of MIMIC-IV is much larger than MIMIC-III and learning ability of the student model is weak, thus, the MaTe-Laballing obtains the lower performance. However, we find that MaTe-Laballing can converge much faster with increasing iteration number when the number is larger than about 50. We also observe the similar stability issue about random strategy over task of Cancer vs. Diabetes.

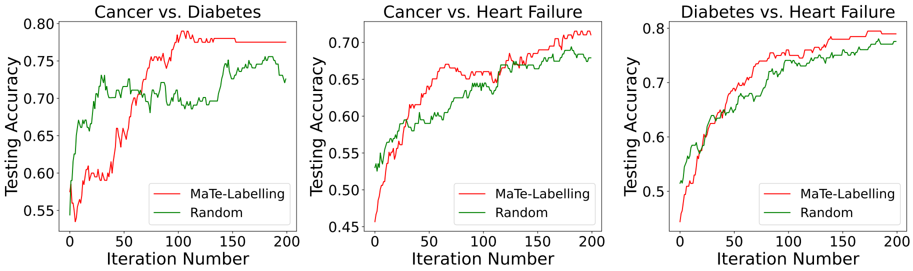

Fig. 3. Average Testing Accuracy on MIMIC-IV with 3 units.

5 Conclusion

In this paper, we have proposed a novel Machine Teaching-based Labelling (MaTe-Labelling) framework. It enables iterative guidance for the student to select samples that are most worth labelling, which largely reduces human efforts for labelling. On the other hand, MaTe-Labelling has also enabled teacher to provide efficient data services and strong guidance to student without releasing any extra information, which effectively prevents regulation and information security breaches. The effectiveness and efficiency of the proposed pipeline has been empirically demonstrated on two publicly available healthcare datasets in comparison with baseline methods. This work has broad implications for the healthcare sector to facilitate data modelling in instances where the large labelled datasets are not accessible to each unit.

Acknowledgements. This research is supported by an Australian Government Research Training Program Scholarship. We also thank the Australian Government Department of Health for supporting this work.

References

1. Balcan, M.F., Hanneke, S., Vaughan, J.W.: The true sample complexity of active learning. Machine learning **80**(2), 111–139 (2010)
2. Baytas, I.M., Xiao, C., Zhang, X., Wang, F., Jain, A.K., Zhou, J.: Patient subtyping via time-aware lstm networks. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 65–74 (2017)
3. Cakmak, M., Thomaz, A.L.: Eliciting good teaching from humans for machine learners. Artificial Intelligence **217**, 198–215 (2014)
4. Choi, E., Bahadori, M.T., Song, L., Stewart, W.F., Sun, J.: Gram: graph-based attention model for healthcare representation learning. In: SIGKDD. pp. 787–795. ACM (2017)
5. Choi, E., Bahadori, M.T., Sun, J., Kulas, J., Schuetz, A., Stewart, W.: Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In: NeurIPS. pp. 3504–3512 (2016)
6. Fails, J.A., Olsen Jr, D.R.: Interactive machine learning. In: Proceedings of the 8th international conference on Intelligent user interfaces. pp. 39–45 (2003)
7. Gao, J., Xiao, C., Wang, Y., Tang, W., Glass, L.M., Sun, J.: Stagenet: Stage-aware neural networks for health risk prediction. In: Proceedings of The Web Conference 2020. pp. 530–540 (2020)
8. Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L.A., Mark, R.: Mimic-iv (version 0.4). PhysioNet (2020)
9. Johnson, A.E., Pollard, T.J., Shen, L., Li-wei, H.L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G.: Mimic-iii, a freely accessible critical care database. Scientific data **3**, 160035 (2016)
10. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
11. Lee, D., Yu, H., Jiang, X., Rogith, D., Gudala, M., Tejani, M., Zhang, Q., Xiong, L.: Generating sequential electronic health records using dual adversarial autoencoder. Journal of the American Medical Informatics Association **27**(9), 1411–1419 (2020)
12. Liu, M., Jiang, L., Liu, J., Wang, X., Zhu, J., Liu, S.: Improving learning-from-crowds through expert validation. In: IJCAI. pp. 2329–2336 (2017)

13. Liu, W., Dai, B., Humayun, A., Tay, C., Yu, C., Smith, L.B., Rehg, J.M., Song, L.: Iterative machine teaching. In: International Conference on Machine Learning. pp. 2149–2158. PMLR (2017)
14. Long, G., Shen, T., Tan, Y., Gerrard, L., Clarke, A., Jiang, J.: Federated learning for privacy-preserving open innovation future on digital health. arXiv preprint arXiv:2108.10761 (2021)
15. Ma, F., You, Q., Xiao, H., Chitta, R., Zhou, J., Gao, J.: KAME: Knowledge-based attention model for diagnosis prediction in healthcare. In: CIKM. pp. 743–752. ACM (Oct 2018)
16. Nguyen, P., Tran, T., Wickramasinghe, N., Venkatesh, S.: Deepr: A convolutional net for medical records (2016)
17. Peng, X., Long, G., Pan, S., Jiang, J., Niu, Z.: Attentive dual embedding for understanding medical concepts in electronic health records. In: IJCNN. pp. 1–8. IEEE (2019)
18. Peng, X., Long, G., Shen, T., Wang, S., Jiang, J., Zhang, C.: Bitenet: Bidirectional temporal encoder network to predict medical outcomes. In: 2020 IEEE International Conference on Data Mining (ICDM). pp. 412–421. IEEE (2020)
19. Peng, X., Shen, T., Wang, S., Niu, Z., Zhang, C., et al.: Mimo: Mutual integration of patient journey and medical ontology for healthcare representation learning. arXiv preprint arXiv:2107.09288 (2021)
20. Pham, T., Tran, T., Phung, D., Venkatesh, S.: Deepcare: A deep dynamic memory model for predictive medicine. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. pp. 30–41. Springer (2016)
21. Settles, B.: Active learning. Synthesis Lectures on Artificial Intelligence and Machine Learning (2012)
22. Settles, B.: Active learning literature survey (2009)
23. Shickel, B., Tighe, P.J., Bihorac, A., Rashidi, P.: Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. IEEE J Biomed Health Inform **22**(5), 1589–1604 (2018)
24. Singla, A., Bogunovic, I., Bartók, G., Karbasi, A., Krause, A.: Near-optimally teaching the crowd to classify. In: International Conference on Machine Learning. pp. 154–162. PMLR (2014)
25. Song, L., Cheong, C.W., Yin, K., Cheung, W.K., Cm, B.: Medical concept embedding with multiple ontological representations. In: IJCAI. pp. 4613–4619 (2019)
26. Wang, Y., Long, G., Peng, X., Clarke, A., Stevenson, R., Gerrard, L.: Interactive deep metric learning for healthcare cohort discovery. In: Australasian Conference on Data Mining. pp. 208–221. Springer (2019)
27. Zhao, J., Chen, Y., Zhang, W.: Differential privacy preservation in deep learning: Challenges, opportunities and solutions. IEEE Access **7**, 48901–48911 (2019)
28. Zhu, X.: Machine teaching for bayesian learners in the exponential family. In: NIPS. pp. 1905–1913 (2013)
29. Zhu, X.: Machine teaching: An inverse problem to machine learning and an approach toward optimal education. In: AAAI. vol. 29 (2015)