Learning in Clinical Trial Settings

Zoe Fowler ZFOWLER3@GATECH.EDU
Georgia Institute of Technology
Atlanta, GA, USA

Kiran Kokilepersaud KPK6@GATECH.EDU
Georgia Institute of Technology
Atlanta, GA, USA

Mohit Prabhushankar MOHIT.P@GATECH.EDU
Georgia Institute of Technology
Atlanta, GA, USA

Ghassan AlRegib ALREGIB@GATECH.EDU
Georgia Institute of Technology
Atlanta, GA, USA

Abstract

This paper presents an approach to active learning that considers the non-independent and identically distributed (non-i.i.d.) structure of a clinical trial setting. There exists two types of clinical trials: retrospective and prospective. Retrospective clinical trials analyze data after treatment has been performed; prospective clinical trials collect data as treatment is ongoing. Traditional active learning approaches are often unrealistic in practice and assume the dataset is i.i.d. when selecting training samples; however, in the case of clinical trials, treatment results in a dependency between the data collected at the current and past visits. Thus, we propose prospective active learning to overcome the limitations present in traditional active learning methods, where we condition on the time data was collected. We compare our proposed method to the traditional active learning paradigm, which we refer to as retrospective in nature, on one clinical trial dataset and one non-clinical trial dataset. We show that in clinical trial settings, our proposed method outperforms retrospective active learning.

Keywords: Clinical Trials, Active Learning, Non-iid Datasets

1. Introduction

Clinical trials exist to determine the effect of certain treatments on a population through a strict set of procedures and regimens (Chien et al. (2022)). This differs from other medical settings, where there is no constraint on the treatment provided to patients. As indicated in Figure 1, there are two main types of clinical trials used by researchers: retrospective and prospective (Song and Chung (2010)). These clinical trials differ in their direction in time. For example, retrospective clinical trials examine previously collected data from past patient visits, as shown in Figure 1. Therefore, retrospective clinical trials have immediate access to all collected data (Song and Chung (2010)). On the other hand, prospective clinical trials are executed over time as patients receive treatment. Thus, prospective clinical trials are a laborious, time-consuming process, where a patient is diagnosed and treated at each visit.

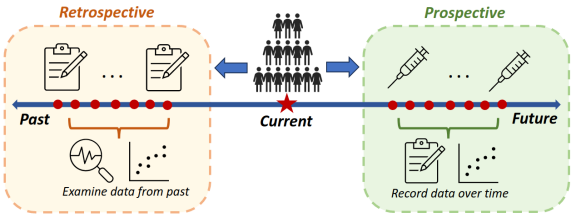

Figure 1: Retrospective vs Prospective Clinical Trial Description

Machine learning is a solution to alleviate the immense diagnosing burden placed on medical professionals in Figure 1 (McDonald et al. (2015), Erickson et al. (2017)), though these approaches are rarely adopted in clinical practice due to a large labeled data requirement (Roh et al. (2019)). To this end, active learning is frequently used when machine learning applications impose a budget for the amount of data to label. Active learning is an iterative process that uses query strategies to acquire a smaller subset of training samples across multiple training rounds (Settles (2009)). It mirrors the prospective clinical trial routine depicted in Figure 1, where a round of training an active learning algorithm parallels a patient visit in the clinical trial process.

Even though active learning resembles the flow of prospective clinical trials, it is rarely adapted to clinical trial data, and even when it is, active learning is not performed in a way that matches the data collection process (Fowler et al. (2023)). Specifically, traditional active learning attempts view the active learning process for medical data from a retrospective lens, where the entirety of the collected data is an option to query from (Logan et al. (2022), Smailagic et al. (2018)). Thus, these approaches contribute to unrealistic assumptions of active learning performance in real clinical trial scenarios, as the model queries and relies on data samples it otherwise would not have access to.

By treating active learning in a retrospective manner, relevant correlations that exist across time amongst data samples are also disregarded. Traditional query strategies sample data based on solely a criteria unique to each strategy, thus treating all data samples as i.i.d. (Settles (2009), Wang and Ye (2015)). However, in a clinical trial setting, the data collection and treatment process is non-i.i.d since at any given visit, the current patient information depends on medical data from past visits. Hence, traditional active learning approaches lose relevant information (such as deteriorating health), as they sample data regardless of which visit it was collected from.

In this work, we argue instead that the *active learning process for clinical trial data must be viewed from the prospective clinical trial lens in Figure 1* due to the nature of the data collection and treatment administration process. We refer to traditional active learning methods as retrospective active learning, which assumes access to the entirety of the collected data. In contrast, we refer to prospective active learning as active learning methods that follow the structure of prospective clinical trials by querying data sequentially based on patient visit number. Hence, we consider the data collection process of clinical trials that creates time-related dependencies across data in clinical trial datasets, and we aim to understand reasons as to why machine learning with clinical trial data must be different from other non-i.i.d. medical datasets through insights about the treatment process. The

contributions are as follows: (1) We convert both retrospective and prospective clinical trials into active learning frameworks by following Figure 1, (2) We compare the performance for this setup against retrospective active learning on the classification of Optical Coherence Tomography (OCT) from an ophthalmology prospective clinical trial, and (3) We compare the performance of prospective active learning on a non-clinical trial chest x-ray dataset.

2. Relating the Concept of non-i.i.d. to Clinical Trials

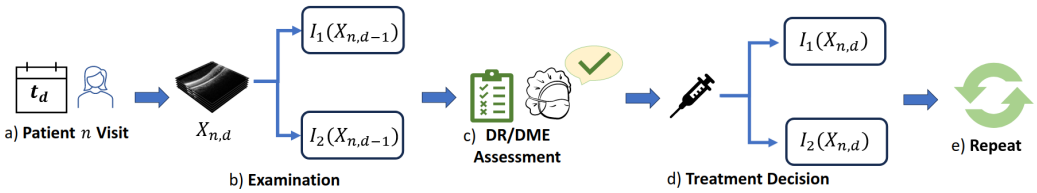

Figure 2: Prospective clinical trial data collection process

A clinical trial dataset is i.i.d. if a patient's data deposited at a particular visit has no dependency on previous visit data. However, in prospective clinical trial settings, the effect of a certain treatment is analyzed across several visits, and the resulting clinical trial dataset is non-i.i.d. To understand the non-i.i.d. nature of clinical trial datasets, consider an ophthalmology prospective clinical trial for Diabetic Retinipathy (DR) and Diabetic Macular Edema (DME) assessment, illustrated in Figure 2. In this clinical trial, assume we have a set of patients $P = \{p_1, \cdots, p_N\}$, where each $p_n \in P$ represents a unique patient. Additionally, each p_n is associated with a set of visits $T = \{t_1, \cdots, t_D\}$. At a visit t_d, a patient p_n is assessed through examination of the patient's OCT data $X_{n,d}$, indicating the patient number n and visit number d the data originated from. Clinical trials typically then administer a uniform treatment to the population. In Figure 2, patients receive an eye injection I_1 or I_2 corresponding to which disease the patient is diagnosed with. This differs from what is done in hospitals, where patients with the same disease do not necessarily receive the same treatment. In clinical trials, however, a patient's OCT data at a particular visit can be represented as $X_{n,d} = I_1(X_{n,d-1})$ or $X_{n,d} = I_2(X_{n,d-1})$, as $X_{n,d}$ is a result of the treatment decision at patient p_n's previous visit t_{d-1} and is thus non-i.i.d.

We argue that by disregarding the non-i.i.d. structure of a clinical trial dataset, performance of active learning approaches suffer due to the overall order in which the data is analyzed. As depicted in Figure 2, clinical trials have a distinct treatment being administered, which differs based on disease diagnosis. We term this an intervention, represented by I_1 for DR or I_2 for DME. In the context of classification, estimating the intervention function by utilizing machine learning algorithms is a viable way to classify a data sample as a particular class. However, if the data is observed non-sequentially, the intervention estimated is likely not equivalent to the actual intervention, resulting in an incorrect classification result since the data at a particular visit is dependent on all previous intervention functions applied during past visits. By creating a framework that queries images in a sequential manner based on patient visit number, we utilize the non-i.i.d. property of clinical trials resulting from the treatment administration process depicted in Figure 2.

3. Methodology

For our experiments, we utilize the `OLIVES` (Prabhushankar et al. (2022)) OCT clinical trial dataset for DR and DME detection. We also use the `NIH ChestX-ray8` (Wang et al. (2017)) non-clinical trial dataset for Consolidation detection in chest x-ray scans. Further details on the dataset and training can be found in Appendix A.

For our results, we adapt the retrospective and prospective clinical trial framework presented in Figure 2 in terms of active learning. We present the layout of this framework in Figure 3. Specifically, in retrospective active learning, we implement the acquiring of a batch of b new data samples X^* as $X^* = \underset{x_{t_1},\ldots,x_{t_b} \in X_{\text{pool}}}{\arg\max} \; a_r(x_{t_1},\ldots,x_{t_b})$, where a_r is the acquisition function for the retrospective setting as shown in Figure 3. In the case of retrospective active learning, this is simply a pre-existing query strategy.

For prospective active learning, the non-i.i.d. structure of the data is taken into consideration. Our selection of samples in this case is $X^* = \underset{x_{t_1},\ldots,x_{t_b} \in X_{\text{pool}}}{\arg\max} \; a_p(x_{t_1},\ldots,x_{t_b}|d)$, where we augment the previous equation such that we condition the selection based on the time or visit number d of when the data was collected. This results in a_p, which constrains the selection space on which pre-existing strategies can operate. To understand the difference

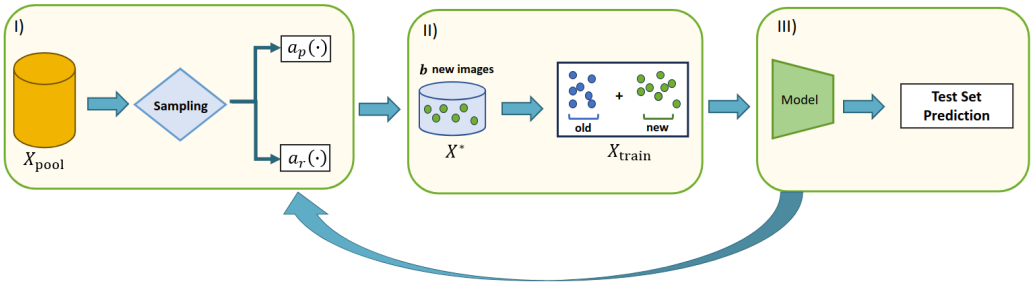

Figure 3: Active learning framework: I) Choose query strategy and sample from X_{pool} II) Add sampled data X^* to the training set X_{train} III) Evaluate model on test set

between a_r and a_p, consider the `OLIVES` dataset. At the first visit, OCT scans will be collected from N patients, resulting in $N \times 49$ OCT scans. When performing prospective active learning via a_p, we only have $N \times 49$ OCT scans to sample from in the first round. For the next round, there are the remaining OCT scans from the first visit not selected during the first round, as well as $N \times 49$ OCT scans corresponding to the patients' second visit, that can be sampled during the second round of training the model. In other words, each round d of training an active learning model corresponds to patients' visit d and past (unused) visit data. Retrospective active learning instead has access to the $N \times 49$ OCT scans from the first visit, the $N \times 49$ OCT scans from the second visit, and so forth. Hence at the first round there are $D \times N \times 49$ OCT scans to choose from, where D represents the total number of visits recorded in the dataset. Afterwards, retrospective active learning has access to the unused OCT scans from the previous rounds.

4. Results

A comparison between retrospective and prospective active learning on the OLIVES dataset using the Resnet-18 architecture (He et al. (2016)) and a query size of 256 is shown in Figure 4 across 5 seeds for the following query strategies: Random, Margin, and Least Confidence. In these plots, the x-axis corresponds to the number of rounds, while the y-axis corresponds to the performance accuracy on the test set. Overall, for majority of the query strategies the prospective active learning setting outperforms the traditional retrospective setting. Because prospective active learning queries data in a sequential manner, we hypothesize that prospective active learning distinguishes between the two types of treatment administered to the two disease classes. To further validate this claim, we assess the impact of the order in which visits are integrated into the sampling process in prospective active learning.

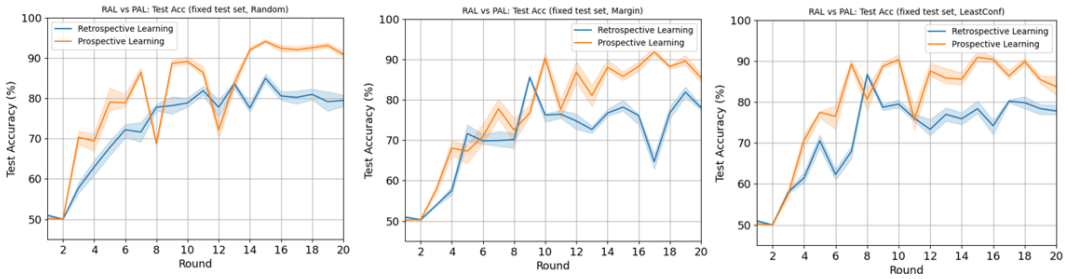

Figure 4: Retrospective active learning (RAL) vs Prospective active learning (PAL) test accuracy results averaged over 5 seeds.

We define a new experimental setup, where the prospective active learning framework randomly queries new visits non-sequentially at each round. Results for this setup are presented in Figure 5. In Figure 5, prospective active learning fails to consistently outperform traditional retrospective active learning, implying that the order in which these images are added to the model's training pool is significant. In particular, it implies that the model must receive the images consecutively in order to provide accuracy benefits over retrospective active learning, indicating that each patient visit can not be treated indepedently.

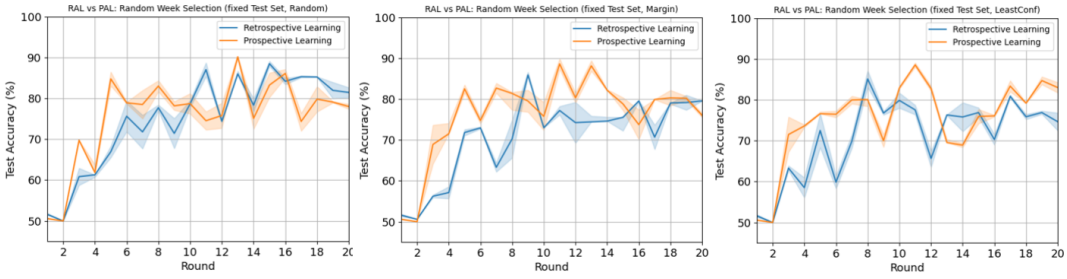

Figure 5: Retrospective active learning (RAL) vs Prospective active learning (PAL) test accuracy results averaged over 5 seeds, where PAL queries visits non-sequentially.

Lastly, we compare the performance of retrospective and prospective active learning in a non-clinical trial medical dataset, the NIH ChestX-ray8 dataset, in Figure 6 using the

Densenet-121 architecture (Huang et al. (2017)) and a query size of 384. Figure 6 shows consolidation detection results in chest x-ray scans for the Random, Margin, and Least Confidence query strategies averaged over 5 seeds. We report results using the AUROC score, as done in the original dataset paper (Wang et al. (2017)). As shown in Figure 6, prospective active learning appears to offer no performance gain on this dataset. The causes for this are most likely the difference in the data collection process. The `NIH ChestX-ray8` dataset is constructed from numerous radiology reports. In other words, no set treatment is administered to patients in the `NIH ChestX-ray8` dataset. When a dataset follows the clinical trial structure in Figure 2, the results indicate that querying data in a sequential manner boosts performance results due to the model better able to distinguish between interventions administered to the different disease classes, like what is shown in Figure 4 with the `OLIVES` dataset. The results in Figure 6, however, show that in a non-clinical trial dataset, prospective active learning does not offer the same benefits as observed in Figure 4, due to an inconsistency in treatment administered across disease classes.

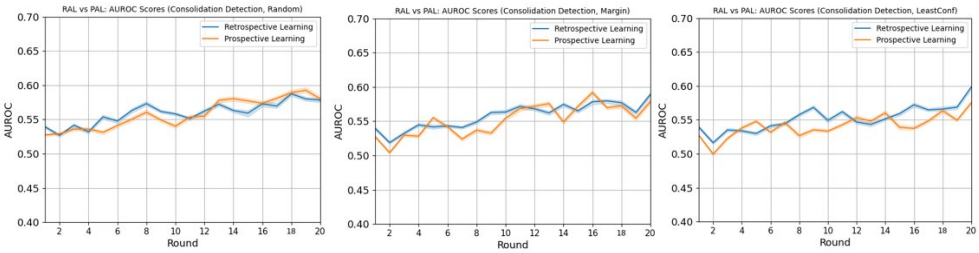

Figure 6: Retrospective active learning (RAL) vs Prospective active learning (PAL) AUROC scores for consolidation detection averaged over 5 seeds.

5. Conclusion

Our work shows that a prospective active learning framework attains higher accuracy than traditional active learning approaches when applied to a clinical trial dataset that has known interventional and time dependencies. Our work suggests that machine learning in clinical trial settings should be performed to align with the data collection process, as sampling sequentially allows the model to learn the interventions across the data. The value of this approach lies in exploiting information readily known during the clinical trial treatment process, such as the time of a patient's visit and the uniform, distinct treatment given to patients depending on their disease status. This differs from other clinical settings, such as in hospitals, where the uniformity in treatment ceases to exist. Future work could investigate the effect of incorporating other time-relevant clinical information when patient visit numbers may not be available.

Acknowledgments

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-2039655. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors(s) and do not necessarily reflect the views of the National Science Foundation.

References

Isabel Chien, Nina Deliu, Richard Turner, Adrian Weller, Sofia Villar, and Niki Kilbertus. Multi-disciplinary fairness considerations in machine learning for clinical trials. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 906–924, 2022.

Bradley J Erickson, Panagiotis Korfiatis, Zeynettin Akkus, and Timothy L Kline. Machine learning for medical imaging. *Radiographics*, 37(2):505–515, 2017.

Zoe Fowler, Kiran Kokilepersaud, Mohit Prabhushankar, and Ghassan AlRegib. Clinical trial active learning. In *Proceedings of the 14th ACM Conference on Bioinformatics, Computational Biology and Health Informatics (ACM-BCB)*, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

Yash-yee Logan, Ryan Benkert, Ahmad Mustafa, and Ghassan AlRegib. Patient aware active learning for fine-grained oct classification. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3908–3912. IEEE, 2022.

Robert J McDonald, Kara M Schwartz, Laurence J Eckel, Felix E Diehn, Christopher H Hunt, Brian J Bartholmai, Bradley J Erickson, and David F Kallmes. The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. *Academic radiology*, 22(9):1191–1198, 2015.

Mohit Prabhushankar, Kiran Kokilepersaud, Yash-yee Logan, Stephanie Trejo Corona, Ghassan AlRegib, and Charles Wykoff. Olives dataset: Ophthalmic labels for investigating visual eye semantics. *Advances in Neural Information Processing Systems*, 35: 9201–9216, 2022.

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

Yuji Roh, Geon Heo, and Steven Euijong Whang. A survey on data collection for machine learning: a big data-ai integration perspective. *IEEE Transactions on Knowledge and Data Engineering*, 33(4):1328–1347, 2019.

Burr Settles. Active learning literature survey. 2009.

Asim Smailagic, Hae Young Noh, Pedro Costa, Devesh Walawalkar, Kartik Khandelwal, Mostafa Mirshekari, Jonathon Fagert, Adrián Galdrán, and Susu Xu. Medal: Deep active

learning sampling method for medical image analysis. *arXiv preprint arXiv:1809.09287*, 2018.

Jae W Song and Kevin C Chung. Observational studies: cohort and case-control studies. *Plastic and reconstructive surgery*, 126(6):2234, 2010.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

Zheng Wang and Jieping Ye. Querying discriminative and representative samples for batch mode active learning. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 9(3):1–23, 2015.

Appendix A. Dataset and Additional Training Details

A.1 Dataset Details

For our experiments, we utilize the OLIVES (Prabhushankar et al. (2022)) dataset. The OLIVES dataset consists of 78,189 OCT images, collected from two different clinical trials. The two clinical trials are denoted as PRIME, which recruited patients with DR, and TREX, which recruited patients with DME. The patients in PRIME and TREX received two distinct treatments, allowing the assumptions from Figure 2 to reflect the structure of the actual dataset. As illustrated in Figure 2, the DR patients in PRIME receive treatment I_1, whereas the DME patients in TREX receive a different treatment I_2. The train set for this dataset includes 76 unique eyes, utilizing the remaining 20 eyes for our test set. The test set is composed of 2,000 images from various visits and is balanced between disease classes.

We also perform experiments on the NIH ChestX-ray8 dataset (Wang et al. (2017)), consisting of 14 disease classes. Due to the large imbalance of disease classes, the dataset was pre-processed to include only 8 of the more prevalent diseases. These include: Effusion, Infiltration, Mass, Nodule, Atelectasis, Pneumothorax, Pleural Thickening, and Consolidation. Furthermore, the dataset presents a 'No Finding' label when an image has none of these 8 diseases; due to the extreme prevalence of images associated with a 'No Finding' label, 5,000 of these images were randomly sampled to be included in our pre-processed dataset. The resulting dataset has 19,796 unique patients, resulting in 64,535 chest x-ray scans. We then include 10,862 unique patients in our train set, 5,959 patients in our test set, and the remaining 2,975 patients in our validation set. The test set is composed of 20,380 images from various visits.

A.2 Training Details

For our OCT experiments, we utilize the Pytorch implementation of the Resnet18 architecture (He et al. (2016)), with the linear layer being altered to accommodate binary classification. Thus, the input into the Resnet18 model is a batch of OCT images, and the output is a binary classification decision corresponding to the presence of either DR or DME. We train this model each round for DR/DME detection until a minimum accuracy of 97% is met. The Adam optimizer was used during training with a learning rate of 0.00015, following the training routine for OCT images from (Prabhushankar et al. (2022), Logan et al. (2022)). All OCT images were resized to 128×128 and were normalized with $\mu = 0.1706$ and $\sigma = 0.2112$. For Figures 4, 5, a query size of 256 is used, initializing the model with 128 samples.

For the chest x-ray experiments, we use a Densenet-121 architecture pretrained on ImageNet in PyTorch, with the linear layer being altered to include eight outputs and a sigmoid nonlinearity applied afterwards (Huang et al. (2017), Rajpurkar et al. (2017)). The output of this approach is a sequence of binary labels corresponding to the presence or absence of the following eight diseases: Effusion, Infiltration, Mass, Nodule, Atelectasis, Pneumothorax, Pleural Thickening, and Consolidation. The training regimen was adopted from (Rajpurkar et al. (2017)), where the model is trained each round until there has been no improvement

on the validation set in 3 epochs. We use stochastic gradient descent as our optimizer, with a learning rate of 0.001. All chest x-ray images were resized to 224×224 and normalized with the ImageNet mean and standard deviation. For Figure 6, a query size of 384 is used, initializing the model with 200 samples.