DISSERTATION PROPOSAL

Jingyi (Kyra) Gan

"Modern Methods in Healthcare"

Thursday, November 19, 2020
10:00am EST
Zoom: https://cmu.zoom.us/j/95957798799?pwd=SzVYTndJWkVWbjhHNzhZdlZTVVdCZz09

In this dissertation, we use tools from stochastic learning, causal inference, and machine learning to analyze problems in healthcare with the goal of reducing healthcare expenditure and improve the quality of care.

In the first chapter, we study the problem of personalized treatment for Opioid Use Disorder (OUD) using wearable devices when the budget is limited. Wearable devices have the potential to revolutionize treatments for OUD by measuring patient responses to different treatment regimens in real-time, enabling the development of personalized treatments. A variety of wearable devices with different features, sensitivities, and costs are available. Whether such devices are practical and cost-effective to incorporate in treatments for OUD, and if so how they should be used, are critical questions. To investigate these questions, we build a finite-horizon, non-stationary Constrained Partially Observable Markov Decision Process (CPOMDP). To facilitate the solution of our model, we provide a novel budget reformulation that finds all optimal solutions lying on the original formulation's solution's convex hull. Next, we show our reformulation can be solved using a binary search in conjunction with an exact POMDP algorithm. We apply those elements, using extracted transition matrices and rewards from past literature, to perform a numerical study to investigate the value of incorporating different wearables in treatments for OUD under scenarios described by different levels of budget, wearable precision, and patient Treatment Adherence (TA). We find that wearables can be valuable at moderate budgets for patients with low or moderate TA; this benefit increases as the wearable accuracy increases. Outside of these settings, either the marginal benefit of wearables is negligible relative to their cost, or their use increases the patients' risk of overdose to an unacceptable degree.

In the second chapter, we study the novel question in causal inference of how to deconfound the observational data efficiently to estimate the causal effects accurately. The theory of causal inference is a foundational topic with direct impact in healthcare. Given only data generated by a standard confounding graph with unobserved confounder, the Average Treatment Effect (ATE) is not identifiable. To estimate the ATE, a practitioner must then either (a) collect deconfounded data; (b) run a clinical trial; or (c) elucidate further properties of the causal graph that might render the ATE identifiable. In this paper, we consider the benefit of incorporating a large confounded observational dataset (confounder unobserved) alongside a small deconfounded observational dataset (confounder revealed) when estimating the ATE. Our theoretical results suggest that the inclusion of confounded data can significantly reduce the quantity of deconfounded data required to estimate the ATE to within a desired accuracy level. Moreover, in some cases---say, genetics---we could imagine retrospectively selecting samples to deconfound. We demonstrate that by actively selecting these samples based upon the (already observed) treatment and outcome, we can reduce sample complexity further. Our theoretical and empirical results establish that the worst-case relative performance of our approach (vs. a natural benchmark) is bounded while our best-case gains are unbounded. Next, we demonstrate the benefits of selective deconfounding using a large real-world dataset related to

genetic mutation in cancer. Finally, we introduce an adaptive version of the problem and propose three adaptive heuristics.

In the third chapter, we focus on solving real-world problems where we collaborate with physicians. This chapter is motivated by improving the gap between machine learning research in healthcare and what has been implemented in practice. This project will consist of multiple sub-projects. In the first project, we collaborate closely with Dr. Patel and Dr. Novelli from University of Pittsburgh Medical Center in predicting the 30-day readmission risk for patients with Sickle Cell Disease (SCD). Reducing preventable hospital readmissions in SCD could potentially improve outcomes and decrease healthcare costs. In a retrospective study of electronic health records, we hypothesized Machine Learning (ML) algorithms may outperform standard readmission scoring systems (LACE and HOSPITAL indices). Participants (n=446) included patients with SCD with at least one unplanned inpatient encounter between January 1, 2013, and November 1, 2018. Patients were randomly partitioned into training and testing groups. Unplanned hospital admissions (n=3299) were stratified to training and testing samples. Potential predictors (n=486), measured from the last unplanned inpatient discharge to the current unplanned inpatient visit, were obtained via both data-driven methods and clinical knowledge. Three standard ML algorithms, Logistic Regression (LR), Support Vector Machine (SVM), and Random Forest (RF) were applied. Prediction performance was assessed using the C-statistic, sensitivity, and specificity. In addition, we reported the most important predictors in our best models. In this dataset, ML algorithms outperformed LACE (C-statistic 0.6, 95%CI 0.57-0.64) and HOSPITAL (C-statistic 0.69, 95%CI 0.66-0.72), with the RF (C-statistic 0.77, 95%CI 0.73-0.79) and LR (C-statistic 0.77, 95%CI 0.73-0.8) performing the best. ML algorithms can be powerful tools in predicting readmission in high risk patient groups.

Our next project will focus on immunotherapies. We are currently collaborating with two research teams at Hillman Cancer Center and one research group at Genentech. We expect at least one of these three projects to lead to a publishable paper that will complete the third chapter (and so the Thesis).