Preserving Privacy when Learning Individualized Treatment Rules

Sensitive data can provide valuable insights while still remaining private.

Spencer Giddens

Abstract

Clinical trials and other medical studies routinely collect sensitive data from individual participants. In many cases, this information is then used to inform individualized treatment rules (ITRs). Outcome weighted learning (OWL) is a machine learning method for developing ITRs in a manner that ensures the best possible treatment benefit. Though beneficial in many ways, the use of OWL (or any other machine learning method for that matter) for fitting a treatment assignment model introduces privacy concerns, as it has been shown that models fitted on sensitive data are susceptible to attacks revealing private information. Differential privacy (DP) is a popular framework for providing mathematically provable privacy guarantees when operating on sensitive data. Though previous work has been devoted to applying DP to related methods, prior to this project it had not previously been applied to OWL, despite the clear reasons for doing so. To apply DP to OWL, we developed a novel algorithm approximating OWL and proved that it satisfies DP. We then assessed the performance of DP-OWL via simulation studies. The results demonstrate that it is possible to simultaneously fit an effective OWL model for individualized treatment assignment and provide privacy guarantees to medical study participants.

Why the Healthcare Industry Should Care

- Information collected via medical studies, such as clinical trials, typically has a quality unmatched by data obtained from other sources due to the carefully prescribed procedure for collection. As a result, such data should be used whenever possible. Considering methods for using these data with robust privacy guarantees could permit ethical data use in previously unexplored situations.
- In today's data-driven world, the public is more concerned about the privacy of their personal information than ever. Though well-executed clinical trials are extremely beneficial to society, they may also be a source of apprehension for participants regarding the use and protection of their personal data, which could discourage participation. Privacy-preserving analysis methods represent an ethical way to settle these concerns.

- By utilizing privacy-preserving counterparts to popular methods of analyzing clinical trial data, researchers can engender confidence in the public that sensitive data is handled appropriately.

Introduction

Clinical trials allow researchers to collect valuable information about the impact of a studied treatment on the trial participants, which can then be generalized to the population as a whole. Clinical trials can have many different goals, such as determining treatment safety, measuring treatment effectiveness, and uncovering treatment side effects. One such goal for some clinical trials is to learn which of two potential treatments will be most beneficial to an individual based on their personal characteristics. A model assigning treatments in such a way is known as an individualized treatment rule (ITR).

Outcome weighted learning

Outcome weighted learning (OWL) (Zhao et al., 2012) is a machine learning method for determining ITRs. In the OWL method, it is assumed that clinical trial participants are randomly assigned to one of two potential treatment groups. For each participant, a set of individual characteristics are measured before treatment, and the resulting post-treatment effect, otherwise known as the treatment benefit, is recorded. Using this training data, OWL fits a treatment assignment model that aims to use individual characteristics to choose the treatment most likely to provide the largest benefit. We do this by first defining a loss function to represent the discrepancy between a given treatment assignment function and the optimal treatment assignment. Lower values of the loss function correspond to better treatment assignments. Then, we use the measured treatment benefit to weight the loss function at each clinical trial participant data point, assigning more importance to individuals that saw greater benefits. A regularizer function with corresponding regularization constant is also used. The OWL method outputs the model that produces the smallest regularized loss.

Data collected through clinical trials are classified as health data, which is extremely sensitive and usually subject to legal protections. Thus, the utmost care must be taken to protect such data when presenting results. Even when the individual-level clinical trial data is not released directly, releasing a predictive model such as an OWL model fitted on such data can still be considered problematic, as such models are subject to attacks that can reveal private information from the dataset used to build them (Shokri et al., 2017, Zhao et al., 2021).

Differential privacy

Differential privacy (DP) (Dwork et al., 2006) is a rigorous mathematical framework for addressing privacy concerns when releasing aggregate statistics or building predictive models based on sensitive data. DP has become commonplace in recent years, being used in industry at

companies such as Apple (Apple, 2017), and by government organizations such as the US Census Bureau (US Census Bureau, 2021).

DP ensures privacy for a given mechanism (a general term for an algorithm, method, statistical computation, etc.) by injecting carefully calibrated random noise (such as from a Laplace or Gaussian distribution) into the results. The amount of noise used is controlled by a tunable parameter known as the privacy budget, denoted by ε (which is always positive). A privacy budget closer to zero means we are more stringent in our privacy requirements, so more noise is required. Likewise, a larger privacy budget permits a greater amount of privacy loss, so less noise is necessary. Thus, there is a tradeoff between the privacy and the utility of a DP-satisfying mechanism that must be balanced in applications.

Though previous research has been devoted to the development of DP counterparts to some popular statistical and machine learning methods (Chaudhuri et al., 2011, Kifer et al., 2012), there have been no previous efforts to extend such DP guarantees to OWL. To accomplish this, we first develop an extension to previous DP algorithms that encompasses OWL and prove that the extension does in fact satisfy the requirements of DP. We then figure out ways to extract the best possible utility from the algorithm (i.e. ensure the algorithm assigns as many individuals as possible to their optimal treatment) for a given fixed privacy budget. Finally, we demonstrate the performance of the algorithm on a simulation dataset designed to approximate clinical trial data.

From the perspective of a clinical trial research practitioner who analyzes clinical trial data, the results of this research are important. Practitioners do not currently have access to a method of fitting an OWL model with privacy guarantees. DP-OWL would be valuable both to ensure the highest standards of privacy are kept and to instill confidence in clinical trial participants.

Methods

Our research work contributes to solving the problem of privacy concerns in clinical trial data by producing a novel algorithm that can be used to fit a DP-OWL model. The algorithm first uses the standard, non-DP procedure for OWL (Zhao et al., 2012) to find a set of coefficients for an individualized treatment assignment function. The algorithm then adds random noise to these coefficients and releases the resulting function.

The scale of the random noise is a function of the tunable privacy budget ε and a value known as the global sensitivity of the coefficients. Statistics with larger global sensitivities require more noise to achieve DP. In general, deriving the global sensitivity of these coefficients is non-trivial. Previous work derived the global sensitivity for a problem related to OWL (Chaudhuri et al., 2011). However, this work did not generalize to the case of OWL, where the loss function values at each data point are weighted. The algorithm we developed extends the previous work to include this case.

As part of our work, we provided a rigorous mathematical proof that, under certain conditions, our algorithm gives DP guarantees. These conditions include conditions from the previous methods (Chaudhuri et al., 2011), but also extend to include the condition that each of the weights used by OWL are bounded above by a fixed constant. Additionally, we state modifications that are necessary to the OWL loss function for the DP proof to be valid, specifically, that a smooth approximation to the loss function should be used.

To strengthen our methods, we provide guidelines for the crucial feature selection and hyperparameter tuning steps. Usually, these steps rely on the dataset itself. However, in the DP framework, using the dataset for these steps incurs privacy loss. Thus, we suggest either allocating a portion of the privacy budget to perform these steps and using the remaining budget to fit the OWL model, or alternatively, using a similar, public dataset to perform the steps and reserving the entire privacy budget for the model fitting. For our simulations, we take the latter approach, using a simulated public dataset of size 1000 to select features and tune the regularization constant.

Though our simulation study results (discussed further in the next section) indicate that it is possible to obtain reasonably high accuracy and treatment value while providing DP guarantees, our approach does have some limitations. First, our approach uses only the most conservative form of DP. There also exist practical relaxations of DP that could be explored that are known to improve utility in some cases. Additionally, the feature selection and hyperparameter tuning for our simulations rely on the assumption that there exists a similar, public dataset, and thus represent a "best-case" scenario.

Findings

Figure 1. The mean (with 95% confidence intervals) optimal treatment assignment accuracy rate computed on the simulated testing set via DP-OWL as a function of the privacy budget ε and sample size n. Simulations without DP (denoted by $\varepsilon = \infty$) are also shown for reference.

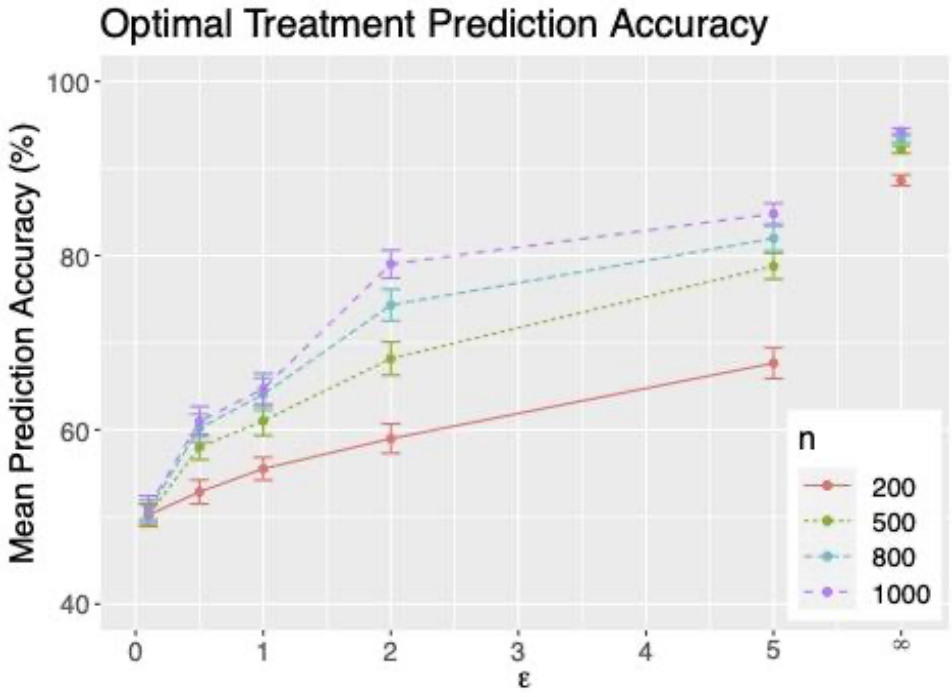
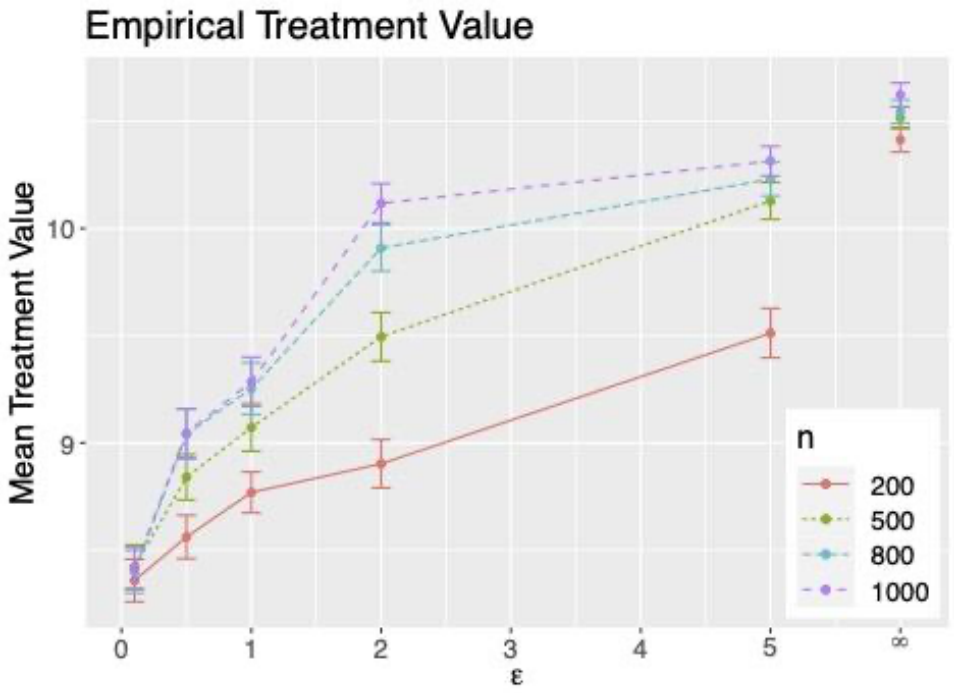


Figure II. The mean (with 95% confidence intervals) empirical treatment value computed on the simulated testing set via DP-OWL as a function of the privacy budget ε and sample size *n*. Simulations without DP (denoted by $\varepsilon = \infty$) are also shown for reference.



To evaluate the performance of our DP-OWL algorithm, we simulate randomized clinical trial data, including randomized individual characteristics, treatment assignments, and treatment benefits using an underlying ground truth treatment assignment function. We also simulate a public dataset based on the same settings, which we use to perform feature selection and hyperparameter tuning. We then measure the performance of our DP-OWL algorithm for various privacy budgets and sample sizes. The performance of the model output from the DP-OWL algorithm is measured by computing the accuracy of the model relative to the ground truth function, as well as the average empirical treatment value of the model.

The accuracy results are presented in Figure I and the average treatment value results are reported in Figure II. They are reported for various combinations of the privacy budget and the sample size. The non-DP method ($\varepsilon = \infty$) is also presented as a baseline for reference.

The accuracy results in Figure I show that the optimal treatment prediction accuracy increases both as the privacy budget increases and as the sample size increases. These results are to be expected as a larger privacy budget corresponds to less stringent privacy requirements and therefore lower required noise. Likewise, a larger sample size corresponds to a smaller global sensitivity, and therefore lower required noise. The results also match our expectation that the privacy-preserving OWL approach is less accurate than the non-private approach, as there is

always a tradeoff between privacy and accuracy. The empirical treatment value results in Figure II match the accuracy results in visual appearance as well as interpretation, meaning that the two metrics can be used interchangeably.

The results demonstrate that accuracy above 80% can be achieved with a privacy budget as low as $\varepsilon = 2$ for a sample size of 1000. If the privacy budget increases to $\varepsilon = 5$, 80% accuracy can be achieved for a sample size as low as 500. These results are encouraging as they conclusively show that our DP-OWL algorithm can be used to provide privacy guarantees to medical study participants while still using the data to fit effective treatment assignment models.

Practical Applications

Our DP-OWL algorithm can be directly used by those who administer and analyze medical study data to utilize the data more ethically. Given a scenario in which an ITR is desired (e.g., a clinical trial aiming to determine which of two drugs will be best for certain groups of people), a practitioner employing our DP-OWL algorithm instead of other non-private methods is able to obtain a reasonably accurate ITR while also providing provable guarantees of privacy to the individuals choosing to participate in the study.

Practitioners choosing to implement the DP-OWL algorithm would also be able to advertise the additional privacy protections to those participating in the study. This would likely increase confidence for study participants in the security of their data, which may even lead to more willing participation.

OWL methods are also applicable outside of the medical study setting. One alternative application is in developing ITRs for online advertising, where the ad shown represents the "treatment." The optimal ad to show each individual may differ based on characteristics of the ad viewer. ITRs for this scenario can also be found using OWL methods, and the data used is also sensitive. As privacy concerns for personal online information are prevalent, using our DP-OWL algorithm in these cases would provide a more ethical alternative that could improve consumer trust.

Conclusion and Next Steps

We have presented our work on developing and proving valid a novel privacy-preserving algorithm for finding optimal ITRs known as DP-OWL. Simulation studies analyzing the performance of the DP-OWL algorithm in a clinical trial setting showed favorable results. Specifically, despite the necessary tradeoff between privacy and utility incurred with DP, our work demonstrates that the DP-OWL algorithm produces quality ITR models for reasonable privacy budgets and sample sizes while simultaneously ensuring privacy protections to study

participants. These privacy protections are increasingly important in today's data-driven world to ensure ethical data use.

There are still a few questions related to this work that remain unanswered and could warrant future research. Since the DP-OWL algorithm struggles to achieve high accuracy and treatment value for low privacy budgets and sample sizes, it would be worth examining ways to increase the utility of the algorithm for those cases. Practical relaxations of DP and alternative approaches to adding noise in the DP-OWL algorithm could be explored to solve this issue. Additionally, it would be valuable to explore and analyze the performance of DP-OWL in other scenarios that seek optimal ITRs, such as online advertising. These scenarios also use sensitive data, and may have larger sample sizes, which would make them good candidates for DP applications. Empirically examining the performance of our DP-OWL algorithm on data from those regimes could lead to beneficial insights and applications.

References

Apple. (2017, December). *Learning with Privacy at Scale*. Apple Machine Learning Research. Retrieved March 29, 2023, from https://machinelearning.apple.com/research/learning-with-privacy-at-scale

Chaudhuri, K., Monteleoni, C., & Sarwate, A. D. (2011). Differentially Private Empirical Risk Minimization. *Journal of Machine Learning Research*, *12*(29), 1069-1109.

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating Noise to Sensitivity in Private Data Analysis (S. Halevi & T. Rabin, Eds.). *Theory of Cryptography*, 265-284. 10.1007/11681878_14

Kifer, D., Smith, A., & Thakurta, A. (2012, June). Private Convex Empirical Risk Minimization and High-dimensional Regression (S. Mannor, N. Srebro, & R. C. Williamson, Eds.). *Proceedings of the 25th Annual Conference on Learning Theory*, *23*, 25.1-25.40.

Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership Inference Attacks Against Machine Learning Models. *2017 IEEE Symposium on Security and Privacy*, 3-18. 10.1109/SP.2017.41

US Census Bureau. (2021, June 9). *Census Bureau Sets Key Parameters to Protect Privacy in 2020 Census Results*. Census Bureau. Retrieved March 29, 2023, from https://www.census.gov/newsroom/press-releases/2021/2020-census-key-parameters.html

Zhao, B. Z. H., Agrawal, A., Coburn, C., Asghar, H. J., Bhaskar, R., Kaafar, M. A., Webb, D., & Dickinson, P. (2021). On the (In)Feasibility of Attribute Inference Attacks on Machine Learning Models. *2021 IEEE European Symposium on Security and Privacy*, 232-251. 10.1109/EuroSP51992.2021.00025

Zhao, Y., Zeng, D., Rush, A. J., & Kosorok, M. R. (2012). Estimating Individualized Treatment Rules Using Outcome Weighted Learning. *Journal of the American Statistical Association*, *107*(499), 1106-1118. JSTOR. http://www.jstor.org/stable/23427417