STAT447C Project Proposal

Rainie Fu

March 2025

1 Introduction

Targeted Maximum Likelihood Estimation (TMLE) is a widely used two-step procedure in causal inference, particularly in observational studies [7]. The first step of TMLE involves estimating an initial model, while the second (targeting) step adjusts this estimator using maximum likelihood estimation (MLE) to determine a fluctuation parameter (ϵ). This update of the initial estimator is done to achieve desirable properties such as double robustness and asymptotic efficiency [8, 11]. However, while MLE produces a point estimate for ϵ, it does not naturally quantify uncertainty around this parameter. A Bayesian alternative offers a promising solution: by replacing the point estimate with a posterior distribution over ϵ, it allows for richer uncertainty quantification [13], which could potentially lead to improvements in finite-sample performance and more robust conclusions.

This project, addressing the theme of "A careful and scientific comparison of a Bayesian estimator with another one, either Bayesian or non-Bayesian," investigates whether a Bayesian approach—either by summarizing the posterior distribution (e.g., via the mean or mode) or by integrating over the full posterior—can improve the targeting step in TMLE compared to the traditional MLE-based method. The central hypothesis is that Bayesian methods, by incorporating uncertainty about ϵ, can provide more accurate and reliable causal effect estimates, especially in small or complex samples where traditional methods may fall short [14, 15].

To achieve this, the project introduces priors on the fluctuation parameter ϵ and derives posterior distributions that propagate uncertainty throughout the estimation process, ultimately affecting the final causal effect estimates. The method will first be tested on a simulated dataset to assess its performance, with metrics including bias, standard error, mean squared error, and coverage. Following this, the approach will be applied to estimate the effect of right heart catheterization (RHC) on mortality using the SUPPORT study dataset [5]. Comparisons will be made between the Bayesian approach and traditional TMLE variants, such as vanilla TMLE and full cross-validated TMLE.

This is an individual project conducted by Rainie Fu. All code for this project is available in the public GitHub repository: https://github.com/RainieFu/STAT447CProject.

2 Method

2.1 Overview of Targeted Maximum Likelihood Estimation

In our context, we are interested in estimating the average treatment effect (ATE) of RHC on mortality within 180 days among critically ill patients.

Let Y denote the binary outcome (where $Y = 1$ indicates death within 180 days), A denote the treatment (where $A = 1$ indicates RHC was performed), and W represent the vector of covariates including demographic characteristics, disease categories, and physiological parameters that may confound the relationship between RHC and mortality.

The causal parameter of interest is the ATE, defined as:

$$\psi_0 = \mathbb{E}[Y(1) - Y(0)] \tag{1}$$

where $Y(a)$ represents the potential outcome that would have been observed had treatment A been set to value a.

TMLE estimates this parameter through a two-step procedure [11, 9]:

2.2 Initial Estimation

First, we estimate two nuisance parameters:

- The outcome regression: $\bar{Q}_0(A, W) = \mathbb{E}[Y|A, W]$, the predicted outcome.
- The propensity score: $g_0(W) = P(A = 1|W)$, the probability of patient receive the treatment.

For these initial estimates, we employ the Super Learner ensemble algorithm, which combines predictions from multiple machine learning algorithms to optimize predictive performance through cross-validation [10]. The initial estimates are denoted $\bar{Q}_n(A, W)$ and $g_n(W)$.

2.3 Targeting Step - Traditional Approach

The traditional targeting step aims to update the initial outcome regression estimate to reduce bias in the estimation of the causal parameter [8]. This is achieved by constructing a parametric submodel through the initial estimate:

$$\bar{Q}_n^{\epsilon}(A, W) = \text{logit}^{-1}(\text{logit}(\bar{Q}_n(A, W)) + \epsilon \cdot H(A, W)) \tag{2}$$

where $H(A, W)$ is the "clever covariate" defined as:

$$H(A, W) = \frac{A}{g_n(W)} - \frac{1 - A}{1 - g_n(W)} \tag{3}$$

The parameter ϵ is estimated using maximum likelihood estimation (MLE):

$$\hat{\epsilon}_{\text{MLE}} = \arg\max_{\epsilon} \sum_{i=1}^{n} \left[Y_i \log \bar{Q}_n^{\epsilon}(A_i, W_i) + (1 - Y_i) \log(1 - \bar{Q}_n^{\epsilon}(A_i, W_i)) \right] \tag{4}$$

Once $\hat{\epsilon}_{\text{MLE}}$ is obtained, the targeted estimate of the outcome regression becomes:

$$\bar{Q}_n^{*}(A, W) = \text{logit}^{-1}(\text{logit}(\bar{Q}_n(A, W)) + \hat{\epsilon}_{\text{MLE}} \cdot H(A, W)) \tag{5}$$

The ATE is then estimated as:

$$\hat{\psi}_{\text{TMLE}} = \frac{1}{n} \sum_{i=1}^{n} \left[\bar{Q}_n^{*}(1, W_i) - \bar{Q}_n^{*}(0, W_i) \right] \tag{6}$$

Inference is conducted using the influence function approach, providing asymptotically valid confidence intervals when either the outcome regression or propensity score model is correctly specified [11].

2.4 Targeting Step - Bayesian Approach

We propose a Bayesian alternative to the traditional targeting step that allows for more comprehensive uncertainty quantification [14, 16, 15]. Rather than obtaining a point estimate $\hat{\epsilon}_{\text{MLE}}$, we derive a posterior distribution for ϵ.

2.4.1 Bayesian Model Formulation

The likelihood function is:

$$p(Y|A, W, \epsilon) = \prod_{i=1}^{n} \left[\bar{Q}_n^{\epsilon}(A_i, W_i) \right]^{Y_i} \left[1 - \bar{Q}_n^{\epsilon}(A_i, W_i) \right]^{1 - Y_i} \tag{7}$$

The posterior distribution for ϵ is then:

$$p(\epsilon|Y, A, W) \propto p(\epsilon) \cdot p(Y|A, W, \epsilon) \tag{8}$$

We use Markov Chain Monte Carlo (MCMC) methods to obtain samples $\{\epsilon^{(1)}, \epsilon^{(2)}, \ldots, \epsilon^{(M)}\}$ from this posterior distribution.

2.4.2 Approach 1: Posterior Summarization

In our first Bayesian approach, we summarize the posterior distribution of ϵ using a single value such as the posterior mean:

$$\hat{\epsilon}_{\text{Bayes}} = \frac{1}{M} \sum_{m=1}^{M} \epsilon^{(m)} \tag{9}$$

For binary outcomes, the targeted estimate is then calculated as:

$$\bar{Q}_n^*(A, W) = \text{logit}^{-1}(\text{logit}(\bar{Q}_n(A, W)) + \hat{\epsilon}_{\text{Bayes}} \cdot H(A, W)) \tag{10}$$

For continuous outcomes, we apply a simpler update:

$$\bar{Q}_n^*(A, W) = \bar{Q}_n(A, W) + \hat{\epsilon}_{\text{Bayes}} \cdot H(A, W) \tag{11}$$

The ATE estimate is calculated in the same way for both outcome types:

$$\hat{\psi}_{\text{Bayes}} = \frac{1}{n} \sum_{i=1}^{n} \left[\bar{Q}_n^*(1, W_i) - \bar{Q}_n^*(0, W_i) \right] \tag{12}$$

This approach incorporates Bayesian inference in the estimation of ϵ while maintaining the standard TMLE structure for the final parameter estimate. It provides a more regularized estimator than MLE when the posterior distribution incorporates informative priors. If the prior is not informative, then the performance may likely be worse than the MLE.

2.4.3 Approach 2: Full Distribution Approach

In our second Bayesian approach, instead of summarizing the posterior distribution of ϵ with a single value, we integrate over its entire posterior distribution to obtain the final estimate. Let D denote the observed data.

For binary outcomes, the targeted estimate is computed as:

$$\bar{Q}_n^*(A, W) = \mathbb{E}_{\epsilon \sim p(\epsilon|D)} \left[\text{logit}^{-1}(\text{logit}(\bar{Q}_n(A, W)) + \epsilon \cdot H(A, W)) \right] = \int \text{logit}^{-1}\Big(\text{logit}(\bar{Q}_n(A, W)) + \epsilon \cdot H(A, W) \Big) \, p(\epsilon \mid D) \, d\epsilon. \tag{13}$$

For continuous outcomes, we apply:

$$\bar{Q}_n^*(A, W) = \mathbb{E}_{\epsilon \sim p(\epsilon|D)} \left[\bar{Q}_n(A, W) + \epsilon \cdot H(A, W) \right] = \int \Big[\bar{Q}_n(A, W) + \epsilon \cdot H(A, W) \Big] \, p(\epsilon \mid D) \, d\epsilon. \tag{14}$$

The average treatment effect (ATE) is then calculated by averaging the difference in the targeted predictions for treatment $A = 1$ and control $A = 0$ across all individuals. For both outcome types, this can be written as:

$$\hat{\psi}_{\text{Bayes}} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\epsilon \sim p(\epsilon|D)} \left[\bar{Q}_n^*(1, W_i) - \bar{Q}_n^*(0, W_i) \right]. \tag{15}$$

This approach fully propagates uncertainty in ϵ throughout the estimation process, which should leading to a more robust inference that accounts for the entire posterior distribution rather than relying on a point estimate [16].

3 Datasets

Two complementary datasets will be used to evaluate and compare the proposed Bayesian-targeted maximum likelihood estimation (TMLE) approach against traditional methods.

3.1 Right Heart Catheterization (RHC) Data

The RHC dataset originates from a large observational study—the SUPPORT study—which includes detailed clinical information on 5,735 critically ill patients [2]. Key outcome measures, the mortality up to 180 days, are available alongside a rich set of potential confounders (e.g., demographics, comorbidities, physiological and laboratory parameters).

3.2 Simulated Data

In parallel, we will use a simulated dataset designed to mimic a clinical scenario [12] —for example, evaluating the effect of statin treatment on atherosclerotic cardiovascular disease (ASCVD). This dataset is generated following a detailed data-generating mechanism that incorporates realistic distributions for confounders such as age, low-density lipoprotein (LDL) levels, diabetes status, and frailty. A risk score is computed and used to drive the treatment assignment through a logistic model, while the outcomes are derived under a potential outcomes framework [6] with a known true average treatment effect (ATE) of approximately –0.108. This simulated dataset allows us to benchmark estimator performance under controlled conditions and to systematically compare the proposed Bayesian TMLE variant against traditional point-estimation approaches.

Comparison with Prior Analyses

The closest analyses of the SUPPORT dataset have focused on traditional causal inference methods such as propensity score matching and regression adjustment [1]. These studies have shown that RHC is associated with increased mortality, but have not fully explored the potential benefits of machine learning-enhanced methods like TMLE [7, 5, 3] or Bayesian methods for uncertainty quantification. For example, the work by Akosile et al. (2018) [4] applied TMLE but did not consider the Bayesian alternative in the targeting step, which could offer more nuanced uncertainty estimates and potentially more robust causal conclusions.

Similarly, while the simulated dataset is used to assess the validity of different causal inference methods [12], prior research has typically evaluated TMLE in a more standard frequentist framework. Our inclusion of Bayesian methods, particularly the integration over the full posterior, represents a novel extension of these techniques that is not typically explored in the current literature.

By applying both the real-world RHC data and the simulated dataset, this project will offer a comprehensive evaluation of Bayesian methods in causal inference, comparing them directly to traditional TMLE methods and providing a new perspective on how uncertainty can be better incorporated into causal estimates.

4 Bayesian Workflow for TMLE Enhancement

The implementation of a Bayesian approach to TMLE involves careful consideration of the workflow to ensure robust and reliable results. Following the framework established by Gelman et al. (2020) [17], we outline a systematic workflow tailored specifically for this project.

4.1 Prior Specification and Assessment

For the fluctuation parameter ϵ, we will begin with a weakly informative prior, $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$, where σ_ϵ^2 is chosen to be sufficiently large to avoid strong influence on the posterior while providing some regularization. Although this parameter is data-dependent, which makes strong prior assumptions challenging, we can still benefit from a principled Bayesian approach to prior specification.

We will conduct prior predictive checks to ensure the chosen prior generates reasonable behavior in the targeted outcomes. This involves:

- Simulating from the prior distribution of ϵ
- Examining the implied distribution of targeted estimates before observing data
- Adjusting the prior variance σ_ϵ^2 if the prior predictive distributions exhibit implausible behavior

The goal is not necessarily to incorporate substantial domain knowledge, but rather to ensure computational stability while allowing the data to predominantly inform our posterior.

4.2 Computational Strategy and Validation

The computation for our Bayesian TMLE approach will rely on MCMC methods implemented in Stan. Our workflow will include:

4.2.1 MCMC Diagnostics

We will assess convergence using multiple chains with diverse initial values, examining both trace plots and rank histograms for the ϵ parameter. We will compute effective sample sizes (ESS) and $\hat{R}$ statistics to ensure reliable posterior approximation. For complex models, we may need to address potential divergences or other sampling pathologies by reparameterization or adjusting the MCMC algorithm parameters.

4.2.2 Posterior Predictive Checks

To validate the computational correctness of our implementation, we will conduct posterior predictive checks comparing simulated data from our fitted model to the observed data. This will help identify any potential issues in the model specification or implementation.

4.3 Model Evaluation and Refinement

The Bayesian workflow is inherently iterative. For our Bayesian TMLE implementation, we will:

1. Fit the initial model with our proposed prior
2. Validate the computation using the diagnostics described above
3. Assess model adequacy through posterior predictive checks
4. Refine the model if necessary by modifying priors or model structure
5. Compare different modeling choices using Bayesian cross-validation

We will pay particular attention to the sensitivity of our results to prior choices by conducting a prior sensitivity analysis, varying the prior variance σ_ϵ^2 to determine how strongly it influences our conclusions.

4.4 Comparison Framework

A critical aspect of our Bayesian workflow is the systematic comparison between our Bayesian TMLE approach and the traditional MLE-based TMLE. This comparison will span several dimensions:

- Calibration and width of uncertainty intervals
- Computational efficiency and reliability
- Behavior in small sample settings

For simulated data, where the true causal effect is known, we will evaluate both approaches on metrics such as mean squared error, coverage probability, and interval width. For the SUPPORT study dataset, where the ground truth is unknown, we will focus on the stability of estimates across different model specifications and their sensitivity to modeling choices.

4.5 Integration with Project Phases

The Bayesian workflow will be integrated across both phases of our project:

Simulation Phase: We will first implement and refine our workflow on simulated data, where we can validate against known ground truth.

Application Phase: When applying our method to the SUPPORT study data, we will maintain the same workflow principles while accounting for the additional complexities of real-world data, including possible outliers, and additional confounders.

Through this systematic Bayesian workflow, we aim to develop an enhancement to TMLE that robustly accounts for uncertainty in the targeting step, with the ultimate goal of preserving the key properties of double robustness and asymptotic efficiency, though their preservation in the Bayesian context remains an open challenge to explore.

References

[1] Connors, A. F., Speroff, T., Dawson, N. V., Thomas, C., Harrell, F. E., Wagner, D., et al (1996). The effectiveness of right heart catheterization in the initial care of critically ill patients. SUPPORT Investigators. JAMA, 276(11), 889-897.

[2] SUPPORT Principal Investigators. A controlled trial to improve care for seriously ill hospitalized patients: The Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT). JAMA. 1995;274(20):1591-1598.

[3] Keele, L., & Small, D. S. (2021). Comparing covariate prioritization via matching to machine learning methods for causal inference using five empirical applications. The American Statistician, 75(4), 355-363.

[4] Akosile M, Zhu H, Zhang S, Johnson NP, Lai D, Zhu H. Reassessing the effectiveness of right heart catheterization (RHC) in the initial care of critically ill patients using targeted maximum likelihood estimation. Int J Clin Biostat Biom. 2018;4(1):018.

[5] Mondol MH, Karim ME. Towards robust causal inference in epidemiological research: Employing double cross-fit TMLE in right heart catheterization data. American Journal of Epidemiology. 2024.

[6] Zivich PN, Breskin A. Machine learning for causal inference: on the use of cross-fit estimators. Epidemiology. 2021;32(3):393-401.

[7] Schuler MS, Rose S. Targeted maximum likelihood estimation for causal inference in observational studies. Am J Epidemiol. 2017;185(1):65-73.

[8] van der Laan, M. J., & Rose, S. (2011). Targeted Learning: Causal Inference for Observational and Experimental Data. Springer Series in Statistics.

[9] van der Laan, M. J., & Rubin, D. (2006). Targeted maximum likelihood learning. The International Journal of Biostatistics, 2(1).

[10] van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. Statistical Applications in Genetics and Molecular Biology, 6(1).

[11] Zheng, W., & van der Laan, M. J. (2010). Asymptotic Theory for Cross-validated Targeted Maximum Likelihood Estimation. U.C. Berkeley Division of Biostatistics Working Paper Series, 273.

[12] Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using Simulation Studies to Evaluate Statistical Methods. Statistical Methods in Medical Research, 28(11), 3250-3271.

[13] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). Bayesian Data Analysis (3rd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/b16018

[14] Oganisian, A., & Roy, J. A. (2021). A practical introduction to Bayesian estimation of causal effects: Parametric and nonparametric approaches. Statistics in Medicine, 40(2), 518-551. https://doi.org/10.1002/sim.8761

[15] Li, F., Ding, P., & Mealli, F. (2022). Bayesian Causal Inference: A Critical Review. arXiv:2206.15460v3 [stat.ME].

[16] Lattimore, F., & Rohde, D. (2019). Causal inference with Bayes rule. arXiv:1910.01510v2 [stat.ML].

[17] Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., & Modrák, M. (2020). Bayesian Workflow. arXiv:2011.01808 [stat.ME].

A Datasets Used in This Study

This appendix provides an overview of the datasets used in this study.

A.1 Right Heart Catheterization (RHC) Dataset

[2]

The RHC dataset originates from the SUPPORT study, which includes detailed clinical information on 5,735 critically ill patients. Table 1 shows a selection of key variables from the first 6 rows of this dataset.

Table 1: First 6 rows of the RHC dataset (selected variables)

	Death	RHC.use	Disease.category	age	sex	APACHE.score	Length.of.Stay
1	0	0	Other	[70,80)	Male	46	9
2	1	1	MOSF	[70,80)	Female	50	45
3	0	1	MOSF	[-Inf,50)	Female	82	60
4	1	0	ARF	[70,80)	Female	48	37
5	1	1	MOSF	[60,70)	Male	72	2
6	0	0	Other	[80, Inf)	Female	38	7

Table 2: Clinical parameters for the first 6 rows of the RHC dataset

	Albumin	Creatinine	PaO2vs.FIO2	Heart.rate	blood.pressure
1	3.50	1.20	68.00	124	41
2	2.60	0.60	218.31	137	63
3	3.50	2.60	275.50	130	57
4	3.50	1.70	156.66	58	55
5	3.50	3.60	478.00	125	65
6	3.10	1.40	184.19	134	115

The RHC dataset includes numerous variables related to demographics, clinical measurements, and outcomes. 47 variables in total:

- **Outcome Variable**: Death (within 180 days)
- **Treatment Variable**: RHC.use (whether right heart catheterization was performed)
- **Demographic Variables**: age, sex, race, income, education level
- **Clinical Variables**: Disease category, APACHE score, Glasgow Coma Score, vital signs (heart rate, blood pressure, respiratory rate, temperature)
- **Laboratory Values**: Albumin, hematocrit, bilirubin, creatinine, sodium, potassium, PaO2/FIO2 ratio
- **Comorbidities**: Cancer, cardiovascular disease, congestive heart failure, pulmonary disease, etc.

A.2 Simulated Dataset

[6]

The simulated dataset is designed to mimic a clinical scenario evaluating the effect of statin treatment on atherosclerotic cardiovascular disease (ASCVD). Table 3 shows the complete data for the first 6 rows of this dataset.

A.3 Variable Descriptions

A.3.1 RHC Dataset Key Variables

Death Binary outcome indicating whether the patient died (1) or not (0) within 180 days

RHC.use Treatment variable indicating whether right heart catheterization was performed (1) or not (0)

Table 3: First 6 rows of the simulated dataset

sim_id	Y	statin	age	ldl_log	diabetes	risk_score	risk_score_cat
1	0	0	46	4.89	0	0.087	2
2	1	0	49	4.70	0	0.058	1
3	1	0	47	5.30	0	0.079	2
4	1	0	45	4.76	0	0.098	2
5	1	0	47	4.75	0	0.076	2
6	1	0	49	4.83	0	0.059	1

Disease.category Patient's primary disease category:

- ARF: Acute Respiratory Failure
- MOSF: Multiple Organ System Failure
- Other: Other conditions

APACHE.score Acute Physiology and Chronic Health Evaluation score, a severity of illness scoring system

Length.of.Stay Duration of hospital stay in days

A.3.2 Simulated Dataset

Y Binary outcome variable (e.g., occurrence of cardiovascular event)

statin Treatment variable indicating whether statin was administered (1) or not (0)

age Age of the patient in years

ldl_log Log-transformed low-density lipoprotein cholesterol level

diabetes Binary indicator for diabetes status (1 = yes, 0 = no)

risk_score Calculated cardiovascular risk score

risk_score_cat Categorized risk score (1 = low risk, 2 = higher risk)