Journal of Data Science 0 (0), 1–19
??? 2024
DOI: 10.6339/24-JDS1143
Data Science in Action

Physician Effects in Critical Care: A Causal Inference Approach Through Propensity Weighting with Parametric and Super Learning Methods

Yuan Bian[1], Yu Shi[1], Hui Guo[2], Grace Y. Yi[1,2], and Wenqing He[1,*]

[1]*Department of Statistical and Actuarial Sciences, University of Western Ontario, London, Ontario, Canada*

[2]*Department of Computer Science, University of Western Ontario, London, Ontario, Canada*

Abstract

Physician performance is critical to caring for patients admitted to the intensive care unit (ICU), who are in life-threatening situations and require high level medical care and interventions. Evaluating physicians is crucial for ensuring a high standard of medical care and fostering continuous performance improvement. The non-randomized nature of ICU data often results in imbalance in patient covariates across physician groups, making direct comparisons of the patients' survival probabilities for each physician misleading. In this article, we utilize the propensity weighting method to address confounding, achieve covariates balance, and assess physician effects. Due to possible model misspecification, we compare the performance of the propensity weighting methods using both parametric models and super learning methods. When the generalized propensity or the quality function is not correctly specified within the parametric propensity weighting framework, super learning-based propensity weighting methods yield more efficient estimators. We demonstrate that utilizing propensity weighting offers an effective way to assess physician performance, a topic of considerable interest to hospital administrators.

Keywords *generalized linear model; machine learning; maximum likelihood*

1 Introduction

The propensity score represents the conditional probability of receiving a particular treatment given a set of pre-treatment characteristics (Rosenbaum and Rubin, 1983). Propensity score weighting is widely employed to address pre-treatment imbalances in observed variables, allowing us to leverage observational or non-randomized data to estimate treatment effects on causal inference. Propensity scores are used to weight the samples from different treatment groups so that the distributions of observed covariates are similar across groups, thus minimizing the effects of observed confounding (McCaffrey et al., 2004; Austin, 2011). Extending propensity scores of binary treatment, Imbens (2000) proposed the generalized propensity score method to accommodate the multiple treatments.

Many methods focus on the context of binary treatment choices (Ding and Li, 2018). For multiple treatments, existing methods are primarily centered on parametric estimation of generalized propensity scores, utilizing multinomial logistic regression models (e.g., Imbens, 2000; Robins et al., 2000; Lechner, 2001; Spreeuwenberg et al., 2010; McCaffrey et al., 2013; Li and Li,

*Corresponding author. Email: whe@stats.uwo.ca.

Received April 13, 2024; Accepted June 8, 2024

2019). Spreeuwenberg et al. (2010) provided a practical step-by-step approach for causal modeling with multiple treatments. Li and Li (2019) proposed a unified propensity score weighting framework for causal inference with multiple treatments. By employing a general class of balancing weights, they weighted treatment groups to create a pseudo-population, termed the target population, where covariate distributions for the treatments were balanced. Zhou et al. (2022) developed the R package *PSweight*.

While parametric models are convenient, they are often impossible to be correctly specified in practice. Therefore, there is a preference for sufficiently flexible methods, such as machine learning methods, to capture the true form of the data generating process. In the case of binary treatment, recent studies on propensity score estimation have demonstrated that machine learning propensity models provide excellent performance in terms of both covariate balancing and effect estimation (e.g., McCaffrey et al., 2004; Setoguchi et al., 2008; Lee et al., 2010; Westreich et al., 2010; McCaffrey et al., 2013; Zivich and Breskin, 2021). For example, Setoguchi et al. (2008) examined propensity score estimation with neural networks, showing that neural networks generally produced the least biased estimates compared to many parametric methods. Lee et al. (2010) found that parametric propensity models with only main effects were generally adequate for covariate balancing, but their de-biasing capability was compromised if the models did not account for the interactions and non-linearities. Tree boosting and random forests substantially reduced bias and resulted in more consistent coverage. Westreich et al. (2010) presented an overview of machine learning-based propensity score estimation, recommending the use of boosting, neural networks, random forests, and support vector machines for estimating propensity scores. McCaffrey et al. (2004) described the use of the generalized boosted model (GBM) in implementing the inverse propensity score weighting to achieve superior balance properties with binary groups. McCaffrey et al. (2013) provided guidance on utilizing GBM for binary groups to estimate causal effects for multiple groups and presented diagnostic criteria for evaluating overall balance across these groups.

In practice, selecting the optimal machine learning method for a given application may be challenging, contingent on the unknown nature and characteristics of the data (Sarker, 2021). To simultaneously utilize different powerful models into consideration, van der Laan et al. (2007) introduced the super learning algorithm by creating a super learner that combines optimally weighted machine learning algorithms. Subsequently, Polley et al. (2021) and Coyle et al. (2022) respectively developed the R packages *SuperLearner* and *sl3*. Pirracchio et al. (2015) showed that using the super learning method may enhance covariate balance and reduce bias. Zivich and Breskin (2021) recommended using the super learning method in conjunction with a cross-fitting procedure.

This article employs analysis methods involving with parametric and super learning-based propensity weighting on critical care data. The goal is to assess the physician effects on patients in the intensive care unit (ICU), with the aim of enhancing the probability of patient survival by aligning the most suitable physicians with ICU admissions. The paper is organized as follows. Section 2 introduces the data source and gives the basic notations and assumptions. In Section 3, we exploit the propensity weighting method to analyze the critical care data and discuss the tilting function. Parametric and super learning methods are employed to determine the generalized weighted average physician effects in Section 4. In Section 5, we analyze the critical care data using the introduced methods. The article is concluded with discussion included in Section 6.

Table 1: Descriptive statistics of patients for physician groups.

	Number of patients	Percentage of Males	Percentage of Age $\geqslant 60$	ER is needed	Alive for discharge	Average SOFA
Physician 1	79	63.3%	41.8%	21.5%	22.8%	7.8
Physician 2	95	61.1%	43.2%	25.3%	30.5%	7.7
Physician 3	90	65.6%	43.3%	17.8%	30.0%	7.3
Physician 4	101	58.4%	57.4%	15.8%	43.6%	7.4
Physician 5	124	62.1%	45.2%	22.6%	44.4%	7.0
Total	489	62.0%	46.4%	20.7%	35.4%	7.4

2 Critical Care Data and Model Framework

2.1 Critical Care Data

The critical care data come as a subset from the 2022 case studies organized for the Statistics Society of Canada 2022 Annual Meeting (SSC, 2022), which include the information about 489 patients admitted to the intensive care unit (ICU). When the patients were admitted to the ICU, they were assigned to one of the five attending physicians on duty for the duration of their stay. For each patient in the dataset, the patient-specific information, including age, gender, sequential organ failure assessment (SOFA) score at admission, whether an emergent response (ER) is needed at admission, and whether the patient is alive or dead at ICU discharge, was recorded by the trained health record technicians. Table 1 shows the descriptive statistical information of the dataset.

Table 1 suggests that a higher proportion discharged patients if patients admitted to ICU are assigned to physician 4 or 5. However, it is not sensible to conclude that physician 4 or 5 has better performance on improving patients' survival probability than other physicians, because distinctive characteristics of patients may also be influential factors of their survival probability. While the proportion (15.8%) of patients assigned to physician 4 who need an emergent response at admission is smaller than that (22.6%) for physician 5, there is a higher proportion (57.4%) of patients aged 60 or older for physician 4 than that (45.2%) of the patients assigned to physician 5. Furthermore, the average SOFA score is 7.0 for physician 5, which is the smallest, indicating a favorable situation of patients for physician 5.

Directly comparing rough statistics for each variable does not enable us to assess the performance of different physicians. In this study, we are interested in assessing the physician effects on improving patients' survival probability. To this end, we formalize the problem by introducing the following framework and describe analysis methods accordingly.

2.2 Framework and Assumptions

Let Y denote the binary response variable, coded as 1 if a patient is alive for discharge from the ICU and 0 otherwise, and let $X = (X_1, \ldots, X_p)^\intercal$ denote the $p \times 1$ random vector of baseline covariates which takes values in $\mathcal{X}$ and has the joint density $f(x)$ with respect to the measure μ. Let A denote the random variable indicating which physician is assigned to a patient, with $A = j$ representing that physician j is assigned to a patient, and let $\mathcal{A} \triangleq \{1, \ldots, m\}$ denote the index set of m available physicians, where m is a finite integer equal or greater than 2. Let

$Y^*(j)$ denote the *potential patient outcome*, taking a value in $\{0, 1\}$, if physician j were to be assigned to the patient (Rubin, 1974). For $j \in \mathcal{A}$, let $\pi_j(X) = \Pr(A = j|X)$ denote the *generalized propensity score* (Imbens, 2000), representing the conditional probability of assigning physician j to a patient, given the patient's characteristics X. By construction, $\sum_{j=1}^{m} \pi_j(X) = 1$ holds for all $X \in \mathcal{X}$.

As with standard causal inference, we make the following three assumptions:

Assumption 1 (Consistency): $Y = Y^*(A)$;

Assumption 2 (Weak Unconfoundedness): $Y^*(j) \perp\!\!\!\perp I(A = j)|X \quad$ for $j \in \mathcal{A}$;

Assumption 3 (Overlap): $0 < \pi_j(X) < 1 \quad$ for $X \in \mathcal{X}$ and $j \in \mathcal{A}$,

where $I(\cdot)$ is the indicator function. Assumption 1 (Cole and Frangakis, 2009) ensures that the potential outcome $Y^*(A)$ is identical to the observed outcome if the patient is actually taken care of by physician A. In contrast to a stronger assumption, made by Rosenbaum and Rubin (1983), that $\{Y^*(j) : j \in \mathcal{A}\} \perp\!\!\!\perp A|X$, Assumption 2 (Imbens, 2000) suggests that conditional on patient characteristics X, the potential outcome $Y^*(j)$ is independent of the assignment indicator $I(A = j)$. Imbens (2000) showed that this weak version of unconfundedness assumption was sufficient for identification of the population-level estimand. Assumption 3 (Rosenbaum and Rubin, 1983; Imbens, 2004), also known as the positivity assumption, implies that each patient in the study population has non-zero probability to receive the care from any physician.

For $j \in \mathcal{A}$, let $Q_j(X) \triangleq E\{Y^*(j)|X\}$ denote the *quality function*, which represents the conditional mean of $Y^*(j)$, given X. It is the conditional probability that, possibly contrary to fact, the patient with characteristics X would be alive for discharge had the patient been treated by physician j. With Assumptions 1–3, the quality function can be expressed in terms of the observed outcome, with $Q_j(X) = E(Y|A = j, X)$ (Imbens, 2000), or equivalently, $P(Y = 1|A = j, X)$. As the conditional distribution of Y, given A and X, is usually unknown in applications, $Q_j(X)$ is often estimated based on observed measurements from a random sample, say $\mathcal{O} = \{\{X_i, A_i, Y_i\} : i = 1, \ldots, n\}$. To do so, we further make the *no-interference* assumption (Cox, 1958), also known as the *stable unit treatment value* (SUTV) assumption (Rubin, 1980, 1990), for the random sample:

Assumption 4 (No-Interference): $Y_i^*(j)$ is not affected by the treatment assignment to patient i' for any $i' \neq i$.

Assumption 4 says that the potential outcome for patient i does not depend on how treatment assignments are conducted for other patients. Assumptions 1–4 together allow us use measurements in $\mathcal{O}$ to estimate $Q_j(X)$ (Imbens and Rubin, 2015), which is typically carried out by modelling $Q_j(X)$ parametrically or using super learning with the introduction of an appropriate loss function.

3 Analysis Methods with Parametric Propensity Weighting

3.1 Propensity Weighting

It is often of interest to assess the marginal mean of $Y^*(j)$, which links with the quality function $Q_j(X)$ via $E\{Y^*(j)\} = E[E\{Y^*(j)|X\}] = E\{Q_j(X)\}$. However, accurately specifying $Q_j(X)$ is not feasible when the dimension of X is high (Imbens, 2000). In this paper, we consider a weighting framework, called propensity weighting (Li et al., 2018; Li and Li, 2019), to incorporate different distributions of covariates for patients with different physicians to assess physician effects.

To capitalize on the population which is more clinically relevant (Zhou et al., 2020) or easier to assess the physician effects on it, we take this population, called the *target* population, as

a reference level and let $g(x)$ denote the density function for the covariates of individuals in this population. Then we compare the density $f(x)$ of the covariates for the *study* population, defined in Section 2.2, to $g(x)$ by considering the ratio $h(x) \triangleq g(x)/f(x)$, which is defined for those x in the support of $f(x)$ and is nonnegative, and we call $h(x)$ the *tilting* function. While $h(x)$ is unknown just like $f(x)$, its introduction offers us an informative way to characterize the differences of the study population from the target population. With $h(x)$, one may be interested in considering a tilting-function-dependent expectation of the potential outcomes for physician j, defined as

$$Q_j^h \triangleq \frac{E\{Q_j(X)h(X)\}}{E\{h(X)\}},$$

which equals

$$\frac{\int_{\mathcal{X}} Q_j(x)h(x)f(x)d\mu(x)}{\int_{\mathcal{X}} h(x)f(x)d\mu(x)} = \frac{\int_{\mathcal{X}} Q_j(x)g(x)d\mu(x)}{\int_{\mathcal{X}} g(x)d\mu(x)}, \tag{1}$$

where $d\mu(x)$ represents the Lebesgue or the counting measure of X, depending on whether X is continuous or discrete (Li et al., 2018). When $Q_j(X)$ and $h(X)$ are independent, Q_j^h recovers $E\{Q_j(X)\}$.

To compare causal effects associated with different physicians, motivated by Li and Li (2019), we propose to consider the *generalized weighted average physician effect* (GWAPE) for physicians j' and j'':

$$\tau^h(j', j'') = \frac{Q_{j'}^h}{Q_{j''}^h}, \tag{2}$$

where j' and $j'' \in \mathcal{A}$, and Q_j^h is assumed to be greater than zero for all $j \in \mathcal{A}$. The difference of $\tau^h(j', j'')$ from 1 shows different effects associated with physicians j' and j''. To contextualize within the additive estimands framework proposed by Li and Li (2019), we introduce the log-GWAPE as

$$\lambda^h(j', j'') = \log\left(Q_{j'}^h\right) - \log\left(Q_{j''}^h\right),$$

then the exponential transformation of $\lambda^h(j', j'')$ yields $\tau^h(j', j'')$, for possibly further analysis. A similar approach can be applied to

$$\tau^h(j', j'') = \frac{Q_{j'}^h/\left(1 - Q_{j'}^h\right)}{Q_{j''}^h/\left(1 - Q_{j''}^h\right)},$$

when odds ratios are of interest.

To more closely describe the covariate distribution for the study population, we let $f_j(x) \triangleq f(x|A = j)$ denote the conditional density of X for an individual, given that physician j is assigned to the individual. Using Bayes rule, we have that $f_j(x) \propto f(x)\pi_j(x)$. In contrast to the construction of the tilting function, we introduce a weight, defined as:

$$w_j(x) \triangleq \frac{g(x)}{f_j(x)} \propto \frac{f(x)h(x)}{f(x)\pi_j(x)} = \frac{h(x)}{\pi_j(x)} \quad \text{for } j \in \mathcal{A}, \tag{3}$$

which yields that

$$f_j(x)w_j(x) = g(x) \quad \text{for all } j \in \mathcal{A}. \tag{4}$$

Expression (4) has an important implication that by attaching a weight to the conditional distribution $f_j(x)$ for each j, we reach a balance in mimicking the target population with density $g(x)$.

3.2 Specification of the Tilting Function

The derivation of (4), or the determination of the weights defined in (3), hinges on the knowledge of the tilting function $h(x)$, which is determined by the discrepancy of $f(x)$ from $g(x)$. While $g(x)$ may be specified to reflect a particular target population of interest, the unknownness of $f(x)$ in many applications makes $h(x)$ unknown. However, one may specify different forms for $h(x)$ in order to describe $f(x)$ on a scale relative to $g(x)$.

The simplest way is to specify the tilting function $h(x)$ as the constant function 1, which shows the identity of the covariate distributions for the study population and the target population, and thus, the target estimand $\tau^h(j', j'')$ is the pairwise average physician effect (PAPE). In this case, the resulting weights $w_j(x)$ in (3) become the standard inverse propensity weights (IPW) (Robins et al., 2000), $1/\pi_j(x)$, the reciprocal of the probability of assigning the jth physician to a patient.

Alternatively, one may set the tilting function $h(x)$ to be of a form that highlights certain features such as the treatment assignment. However, as the specification of $h(x)$ is constrained to make $h(x)f(x)$ be a density (i.e., $g(x)$), we often include a desired function form in $h(x)$ and then attach a normalizing constant, say c, such that $\int ch(x)f(x)dx = 1$ holds. This strategy was used by Li and Li (2019), for example, who took $h(x)$ as $\left\{\sum_{k=1}^{m} 1/\pi_j(x)\right\}^{-1}$. Clearly, $h(x)$ is large when the values of $\pi_j(x)$ are close to each other for all $j \in \mathcal{A}$, and $h(x)$ is small when some of $\pi_j(x)$ are close to 0. In other words, for the target population and the resulting weights, $h(x)$ gives the most relative weight to the covariate regions in which none of the $\pi_j(x)$ are close to 0, and $h(x)$ produces the least relative weight to the regions with a lack of overlap among at least one of dimension of $\pi_j(x)$. Thus, the individuals in the population having values of $cf(x)/\{\sum_{k=1}^{m} 1/\pi_k(x)\}$ for their covariate density form the *overlap* population (Li and Li, 2019), where c is a positive constant such that $cf(x)/\{\sum_{k=1}^{m} 1/\pi_k(x)\}$ is a legitimate density. The overlap population can be interpreted as a pseudo-population of patients, and the target estimand $\tau^h(j', j'')$ is the pairwise average physician effect among the overlap population (PAPO).

With $h(x) = \left\{\sum_{k=1}^{m} 1/\pi_j(x)\right\}^{-1}$, the resulting weights $w_j(x) \propto \{1/\pi_j(x)\}/\{\sum_{k=1}^{m} 1/\pi_k(x)\}$, is called *generalized overlap* weights (GOW) (Li and Li, 2019). Li and Li (2019) showed that among all the weights defined through (3), GOW minimizes the total asymptotic variances of all pairwise comparisons, and has the best finite sample efficiency in estimating GWAPE. Interestingly, if the physician groups are almost balanced in the size of individuals and the covariate distribution so that $\pi_j(x) \approx 1/m$ for all $j \in \mathcal{A}$, we have that $h(x) \approx 1/m^2$, suggesting that $g(x) \approx f(x)$. In other words, the overlap population well approximates the study population, and PAPO is close to PAPE (Li and Li, 2019). A discussion on the specification of $h(x)$ can be found in Li and Li (2019).

4 Generalized Propensity Scores and Quality Functions

4.1 Estimation with Parametric Modeling

In practice, the generalized propensity scores $\pi_j(X_i)$ are not known and need to be estimated from the observed data $\mathcal{O}$. To this end, the generalized propensity scores are commonly modeled parametrically. Consider the multinomial logistic regression model:

$$\pi_j(X_i) = \frac{\exp\left(\alpha_j + \beta_j^{\mathsf{T}} X_i\right)}{1 + \sum_{k=1}^{m-1} \exp\left(\alpha_k + \beta_k^{\mathsf{T}} X_i\right)} \quad \text{for } j = 1, \ldots, m-1$$

and

$$\pi_m(X_i) = \frac{1}{1 + \sum_{k=1}^{m-1} \exp\left(\alpha_k + \beta_k^{\mathsf{T}} X_i\right)}, \tag{5}$$

where $\theta_{\mathrm{P}} \triangleq (\alpha_1, \ldots, \alpha_{m-1}, \beta_1^{\mathsf{T}}, \ldots, \beta_{m-1}^{\mathsf{T}})^{\mathsf{T}}$ is the vector of the model parameters.

The maximum likelihood method can be employed to the data $\mathcal{O}$ to estimate θ_{P}, and let $\hat{\theta}_{\mathrm{P}}$ denote the resulting estimator (or estimate). Consequently, for $j = 1, \ldots, m$ and given X, the estimated generalized propensity scores, denoted $\hat{\pi}_j(X)$, can be obtained from (5) with θ_{P} replaced by $\hat{\theta}_{\mathrm{P}}$. For $j = 1, \ldots, m$ and $i = 1, \ldots, n$, let $\hat{w}_j(X_i)$ denote the resulting estimate of the weight $w_j(X_i)$ defined in (3).

By the argument of Li and Li (2019), Q_j^h in (1) can be consistently estimated by

$$\hat{Q}_j^h = \frac{\sum_{i=1}^n \hat{w}_j(X_i) I(A_i = j) Y_i}{\sum_{i=1}^n \hat{w}_j(X_i) I(A_i = j)}, \tag{6}$$

and therefore, the GWAPE can be consistently estimated by

$$\hat{\tau}^h(j', j'') = \frac{\hat{Q}_{j'}^h}{\hat{Q}_{j''}^h}. \tag{7}$$

Alternatively, as Li and Li (2019) suggested, the estimation of Q_j^h in (1) can be augmented by using

$$\hat{Q}_j^{h,aug} \triangleq \hat{Q}_j^h - \frac{\sum_{i=1}^n \left\{ I(A_i = j) - \hat{\pi}_j(X_i) \right\} \hat{w}_j(X_i) \hat{Q}_j(X_i)}{\sum_{i=1}^n \hat{h}(X_i)}, \tag{8}$$

where $\hat{Q}_j(X_i)$ is an estimate of $E\{Y^*(j)|X_i\}$, which can be obtained by modeling the relationship between the potential outcomes and covariates X_i for each physician group.

To this end, for $j = 1, \ldots, m$, we employ the logistic model:

$$Q_j(X_i) = \frac{1}{1 + \exp\left(\zeta_j + \xi_j^{\mathsf{T}} X_i\right)} \text{ for } i \text{ with } I(A_i = j), \tag{9}$$

where $\theta_j \triangleq (\zeta_j, \xi_j^{\mathsf{T}})^{\mathsf{T}}$ represents the vector of model parameters for the jth group, and we write $\theta \triangleq (\theta_1^{\mathsf{T}}, \ldots, \theta_m^{\mathsf{T}})^{\mathsf{T}}$ to denote the vector of all involved parameters for the quality function. Estimation of θ can be obtained by the maximum likelihood method, and estimates $\hat{Q}_j(X_i)$ can thereby be obtained from (9) with θ_j replaced by its estimate.

When both the model (5) for $\pi_j(X_i)$ and the outcome regression model (9) for $Q_j(X_i)$ are correctly specified, the augmented estimator in (8) is semiparametrically efficient (Li and Li, 2019). Moreover, when the tilting function $h(x) = 1$, the augmented estimator (8) is doubly robust in the sense that if either (5) or (9) is correctly specified, the augmented estimator (8) is a consistent estimator for $E\{Y^*(j)\}$. However, when the tilting function $h(x) \neq 1$, the augmented estimator (8) is not necessarily doubly robust, but its variability may be reduced in comparison with (6). With the augmented estimators in (8), GWAPE is estimated as

$$\hat{\tau}^{h,aug}(j', j'') = \frac{\hat{Q}_{j'}^{h,aug}}{\hat{Q}_{j''}^{h,aug}}. \tag{10}$$

4.2 Estimation with Super Learning

While parametric propensity modeling offers us a convenient way to compare the pairwise average physician effects, it hinges on the assumption that the generalized propensity scores and the quality function are correctly specified. Results from parametric models, however, are vulnerable to model misspecification. To alleviate this issue, we alternatively consider a more flexible method, called *super learning* (van der Laan et al., 2007), an ensemble method that creates a weighted combination of various candidate learners (e.g., machine learning methods) to build a super learner. To achieve asymptotically optimal performance, the super learning method minimizes the cross-validated risk for a user-specified loss function (van der Laan et al., 2007; Polley and van der Laan, 2010). Empirical experience suggests that the super learning method performs satisfactorily in many applications (Luedtke and van der Laan, 2016). We now employ the super learning method to estimate the generalized propensity scores $\pi_j(X_i)$ and quality functions $Q_j(X_i)$.

First, we describe the idea of the super learning method. Let Z denote the outcome of interest and let X represent a vector of covariates to be used to predict Z. The goal of the super learning is to find a function or a vector function of X, denoted $\psi(X) : \mathcal{X} \to \mathbb{R}^d$ to predict Z well, where $\mathcal{X}$ represents the input space as defined in Section 2.2, and d is a positive integer. Let $L : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$ denote the loss function of using $\psi(X)$ to predict Z. Now consider two cases where in case 1, we take Z to be A for modeling generalized propensity scores and in case 2 we take Z to be Y for modeling quality functions.

In case 1, we employ $\psi(X) = (\psi_1(X), \ldots, \psi_m(X))^\mathsf{T}$ to estimate $(\pi_1(X), \ldots, \pi_m(X))^\mathsf{T}$, and take the loss function to be the multinomial logistic loss function

$$L\{A, \psi(X)\} = -\sum_{j \in \mathcal{A}} I(A = j) \log\{\psi_j(X)\}.$$

In case 2, we use $\psi(X)$ to estimate $Q_j(X)$ for $j = 1, \ldots, m$ by splitting the data set into m disjoint subsets, and then estimating $Q_j(X)$ on the jth subset for $j = 1, \ldots, m$. In this case, the loss function is set as the logistic loss function

$$L\{Y, \psi(X)\} = -Y \log\{\psi(X)\} - (1 - Y) \log\{1 - \psi(X)\}.$$

Consider a set of basic machine learning methods, denoted $\mathcal{K}$, and let K represent the cardinality of $\mathcal{K}$. The super learner algorithm is formulated using the following steps (Polley and van der Laan, 2010):

1. Randomly partition the data set $\mathcal{O} = \{(X_i, Z_i),\ i = 1, \ldots, n\}$ into V disjoint equal sized subsamples, where V is a user-specified positive integer greater than 1. For $v = 1, \ldots, V$, take the vth subsample as the validation set, indexed by $V(v)$, and the remaining $V - 1$ subsamples to be the training data, indexed by $T(v)$. That is, $\bigcup_{v=1}^{V} V(v) = \{1, \ldots, n\}$ and $V(v_1) \cap V(v_2) = \emptyset$ for $v_1 \neq v_2$.
2. For $v = 1, \ldots, V$ and $k = 1, \ldots, K$, fit the kth machine learning method in $\mathcal{K}$ to the training set $T(v)$, and let $\hat{\psi}_{k,T(v)}$ represent the resulting estimator of ψ. For $i \in V(v)$, calculate the predicted value for X_i using the kth algorithm and let $\hat{\psi}_{k,T(v)}(X_i)$ denote it.
3. Determine the weight vector $\hat{\gamma} = (\hat{\gamma}_1, \ldots, \hat{\gamma}_K)^\mathsf{T}$ satisfying $\sum_{k=1}^{K} \hat{\gamma}_k = 1$ and $\hat{\gamma}_k \geqslant 0$ for $k = 1, \ldots, K$ by minimizing the cross-validated risk:

$$\hat{\gamma} = \underset{\substack{\gamma_k \geqslant 0, k=1,\ldots,K, \\ \sum_{k=1}^{K} \gamma_k = 1}}{\operatorname{argmin}} \sum_{v=1}^{V} \sum_{i \in V(v)} \left\{ Z_i - \sum_{k=1}^{K} \gamma_k \hat{\psi}_{k,T(v)}(X_i) \right\}^2.$$

4. Finally, for $k = 1, \ldots, K$, fit the kth algorithm to the complete data set $\mathcal{O}$ to obtain estimator $\hat{\psi}_k$ for ψ. The super predictor function is then created by combining these fits with the weight vector $\hat{\gamma}$ obtained in the previous step:

$$\hat{\psi}(X) = \sum_{k=1}^{K} \hat{\gamma}_k \hat{\psi}_k(X).$$

With $\pi_j(X_i)$ and $Q_j(X_i)$ estimated by super learning, one may be tempted to directly substitute them into (6) and (8) as in Section 4.1, but this procedure has limitations. As the super learning method aims to minimize the risk for estimating generalized propensity scores $\pi_j(X_i)$ and quality functions $Q_j(X_i)$, instead of "targeting" the GWAPE we hope to estimate (van der Laan and Rose, 2011), it is useful to investigate the bias-and-variance trade-off carefully. Moreover, the super learning procedure offers us conveniently implemented estimators whose performance is good, as demonstrated numerically. The statistical properties, such as consistent variance estimates, remain unclear (van der Laan and Rose, 2011). As a remedy, we employ the bootstrap method to obtain variance estimates in our following data analysis.

4.3 Assessing Balance and Overlap

Parametric modeling approaches are usually more effective than nonparametric methods in handling large-dimensional data. The challenge, however, is to decide the appropriate interactions and polynomial terms among the covariates to capture possible non-linearity relationships (McCaffrey et al., 2013). On the other hand, machine learning methods automatically incorporate non-linearity and interactions of covariates, outperforming parametric models in many applications for binary treatment groups (e.g., McCaffrey et al., 2004; Setoguchi et al., 2008; Lee et al., 2010; Westreich et al., 2010; McCaffrey et al., 2013; Pirracchio et al., 2015; Zivich and Breskin, 2021). The performance of machine learning methods depends on the the choice of hyperparameters, usually carried out by cross validation in the non-causal framework.

To access the adequacy of the model used to estimate generalized propensity scores, it is useful to check the balance by examining the estimated version of (4) (e.g., McCaffrey et al., 2013). We now describe two popular metrics to check for balance.

By (4), one way to check balance is to inspect if, for each physician level, the weighted covariate mean deviates from that of the target population. For $k = 1, \ldots, p$ and $j = 1, \ldots, m$, let $\bar{X}_{k,j} = \{\sum_{i=1}^{n} I(A_i = j) X_{i,k} \hat{w}_j(X_i)\} / \{\sum_{i=1}^{n} I(A_i = j) \hat{w}_j(X_i)\}$ denote the weighted mean of covariate X_k for the jth physician group, where $\hat{w}_j(X_i) = \hat{h}(X_i)/\hat{\pi}_j(X_i)$ with $\hat{h}(X_i)$ denoting the estimator of the pre-specified tilting function. Further, let $\bar{X}_{k,P} = \left\{\sum_{i=1}^{n} X_{i,k} \hat{h}(X_i)\right\} / \left\{\sum_{i=1}^{n} \hat{h}(X_i)\right\}$ denote the mean of covariate X_k for the target population, and let $S_{X_k} = \sqrt{\left(\sum_{j=1}^{m} S^2_{X_{k,j}}\right)/m}$ denote the averaged weighted (or unweighted) sample standard deviation of covariate X_k, with $S^2_{X_{k,j}}$ denoting the unbiased weighted (or unweighted) sample variance of covariate X_k for the jth physician group. McCaffrey et al. (2013) and Li and Li (2019) defined the population standardized differences (PSD) for covariate X_k as

$$\text{PSD}_{k,j} = \frac{|\bar{X}_{k,j} - \bar{X}_{k,P}|}{S_{X_k}},$$

and proposed to use the maximum PSD, $\max_j |\text{PSD}_{k,j}|$, as the balance metric for each covariate X_k.

Alternatively, by that the balancing property (4) implies the pairwise balancing:

$$f_{j'}(x)w_{j'}(x) = f_{j''}(x)w_{j''}(x) \text{ for } j' \neq j'', \tag{11}$$

we check the maximum pairwise absolute standardized differences (ASD) (McCaffrey et al., 2013; Li and Li, 2019) for each covariate X_k, defined as

$$\max_{j'<j''} |\text{ASD}_{k,j',j''}|, \text{ where } \text{ASD}_{k,j',j''} = \frac{|\bar{X}_{k,j'} - \bar{X}_{k,j''}|}{S_{X_k}}.$$

ASD allows us to assess the similarity of a physician group to the others in terms of covariate means.

Austin and Stuart (2015) suggested a rule of thumb to determine adequate balance, requiring the maximum PSD (or maximum pairwise ASD) of all covariates to be less than 0.1. In the case of parametric models, if some covariates, say $X_{\tilde{k}}$, exhibit inadequate balance, higher-order terms of $X_{\tilde{k}}$ and/or the interaction terms involving $X_{\tilde{k}}$ with $X_{\tilde{k}'}$ ($\tilde{k}' \neq \tilde{k}$) may be added to model (5). The new model is then re-fitted and re-evaluated using the maximum PSD (or maximum pairwise ASD) until satisfactory balance is achieved. For machine learning methods, if the desired balance is not attained, alternative choices of hyperparameters should be explored, and the new model is re-evaluated iteratively until an acceptable balance is reached.

On the contrary, a scenario may arise where the estimated generalized propensity scores for some patients are extreme, nearing zero or one. This issue, known as *lack of overlap*, implies that some physicians may only be suitable for certain patients, while some patients rarely or never receive treatments from particular physicians. Consequently, PAPE may correspond to an infeasible intervention, rendering causal effects non-identifiable (Petersen et al., 2012; Li and Li, 2019). Extreme estimated generalized propensity scores $\hat{\pi}_j(X)$ lead to extreme estimated IPW $1/\hat{\pi}_j(X)$, resulting in poor balance, biased estimates, and excessive variance of the IPW estimators (Austin and Stuart, 2015; Li et al., 2018, 2019; Li and Li, 2019).

Alternatively, by design of its tilting function, GOW automatically bypasses the issue of extreme estimated generalized propensity scores. GOW achieves this by down-weighting subjects with generalized propensity scores close to 0 or 1 and prioritizing subjects in the middle range of the distribution of generalized propensity scores (Li and Li, 2019). In applications, visually checking the overlap assumption is usually difficult when the number of covariates is greater than two (McCaffrey et al., 2013). McCaffrey et al. (2013) proposed an alternative method to assess whether the patients in each physician group had substantial probability of receiving treatment from each physician. This involves comparing the distributions of the estimated generalized propensity scores $\hat{\pi}_j(X)$ for different groups of patient assigned to each physician. This method is to be used in our following analysis.

5 Analysis of Critical Care Data

5.1 Determination of GWAPE Using Parametric Models

Here we use the development in Section 3 to analyze the critical care data in Section 2.1 using the parametric models in Section 4.1, where the generalized propensity score $\pi_j(X_i)$ is modeled by (5), and the quality function $Q_j(X_i)$ is modeled by (9). Applying the notation in Section 2.2 to the critical care data to assess the performance of 5 physicians on patient care, we have that $m = 5$, $\mathcal{A} = \{1, \ldots, 5\}$, and four baseline covariates, *Age*, *Gender*, *Admission Type*, and *SOFA* are considered, where X_1 indicates whether a patient's age is larger than 60; X_2 represents the

gender of a patient; X_3 indicates whether an emergent response is needed at the admission; and X_4 stands for a patient's SOFA score at admission.

The goal here is to assess the average physician care effects on patients using various parametric propensity weighting methods described in Section 3. The first two methods, called the IPW and IPW-aug methods, respectively, specify $h(x)$ to be 1 and respectively employ (6) and (8) for the estimation. The third and fourth methods, called the GOW and GOW-aug methods, respectively, set $h(x)$ to be $\left\{\sum_{k=1}^{m} 1/\pi_k(x)\right\}^{-1}$ and respectively utilize (6) and (8) for the estimation. Employing the maximum likelihood method, we obtain the estimated generalized propensity scores $\hat{\pi}_j(x)$ for $j = 1, \ldots, m$.

As noted in Section 3.2, the target population for IPW-based methods (i.e., IPW and IPW-aug methods) is the entire study population, whereas the target population for GOW-based methods (i.e., GOW and GOW-aug methods) is the overlap population. To assess the overlap assumption that $0 < \Pr(A = j|X) < 1$ for all $X \in \mathcal{X}$ for this dataset with the dimension of X equal to 4, we employ the approach suggested by McCaffrey et al. (2013), as detailed in Section 4.3 to assess if all m data subsets are sufficiently similar by comparing the distributions of the estimated propensity scores across the m data subsets. If those m data subsets are sufficiently close, then the overlap assumption is likely to be satisfied. Here, m data subsets are formed by the measurements for the patients who are assigned to the same physician, i.e., all the patients form m groups, with group k including the patients assigned to physician k for $k = 1, \ldots, m$. In particular, using the kth data subset, for $j = 1, \ldots, m$, we determine $\hat{\pi}_j(X)$, defined after (5), which is a random variable due to its being a function of random vector X, and let W_{jk}^{P} denote it, where $k = 1, \ldots, m$. For $j = 1, \ldots, m$, we utilize R package *PSweight* (Zhou et al., 2022), available at https://cran.r-project.org/web/packages/PSweight/index.html to draw smoothed density estimates for W_{jk}^{P} by letting $k = 1, \ldots, m$. We report the results in Figure 1, where the results for physicians $1, \ldots, 5$ are respectively displayed in subfigures (a)-(e), in which five curves in each subfigure are obtained using the five data subsets respectively.

Although the curves in each of five subfigures in Figure 1 do not coincide, as expected, they all show similar supports, roughly ranging from 0.1 to 0.3, and exhibit fairly similar shapes, suggesting that all the m data subsets are fairly similar. By the arguments in Section 3.2, the overlap population reasonably well-approximates the entire population, meaning that the estimation results from IPW-based methods and GOW-based methods may be close.

To assess the performance of the propensity weighting methods on balancing the covariates, we calculate the maximum pairwise ASD and the maximum PSD for each of 4 covariates using the unweighted, IPW and GOW methods, and report the results in Figure 2, where the vertical dashed line at 0.1 is taken as a conventional threshold to show the imbalance level; values below 0.1 indicate a negligible imbalance. Clearly, for either IPW or GOW, the ASDs and PSDs for four covariates are all below 0.1, suggesting balance is reached by the adjustment derived from IPW or GOW methods. On the contrary, the covariates are not balanced if no adjustment is made, as shown by the ASD and PSD values for all the four covariates except the PSD value for gender.

We now analyze the data using the parametric propensity weighting methods, described in Section 4.1, and display the results in the left panel of Table 2, where we conduct all pairwise comparisons for 5 physicians, indicated by $\tau^h(j, k)$ with $j < k$ and $j, k \in \{1, \ldots, 5\}$, and we report the estimated GWAPE, their associated 95% confidence intervals, constructed from using the square root of twenty-five bootstrap sample variance, and the length of each confidence interval. The inclusion of 1 by a confidence interval shows an insignificant difference between the average effect of the two compared physicians at the significance level 0.05. The two naive parametric

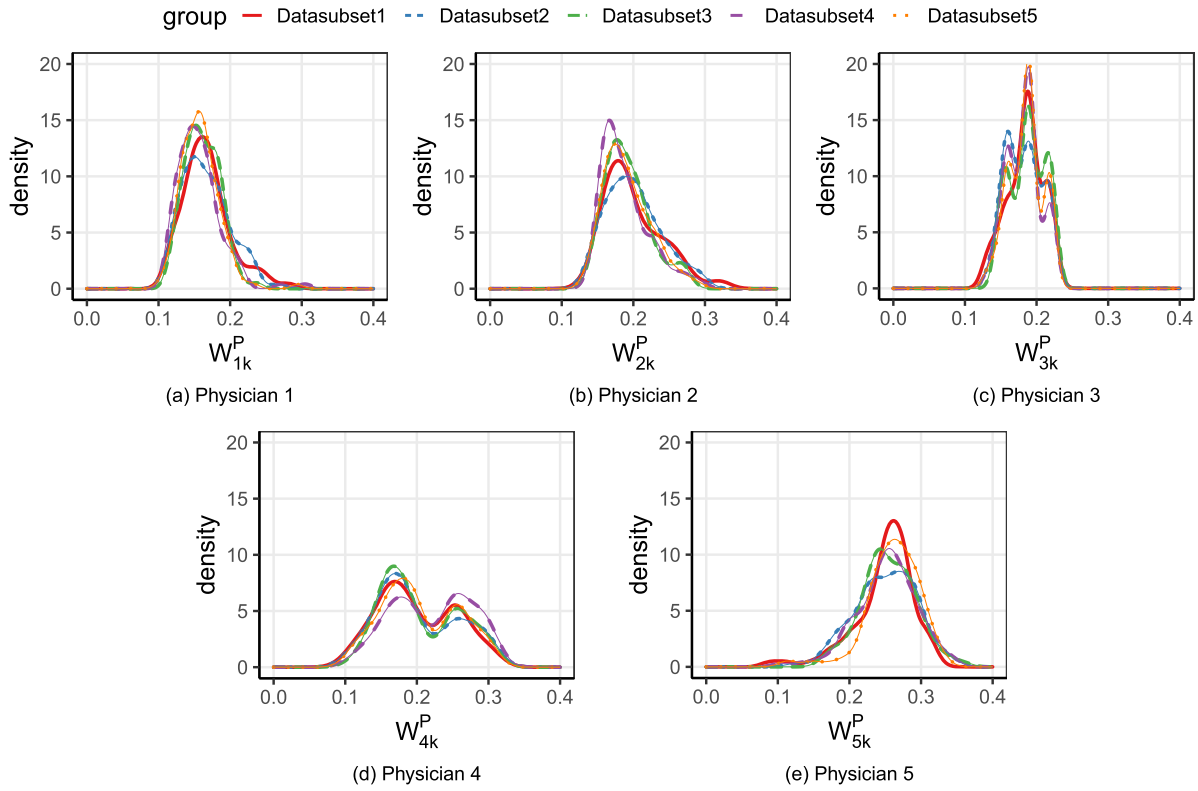
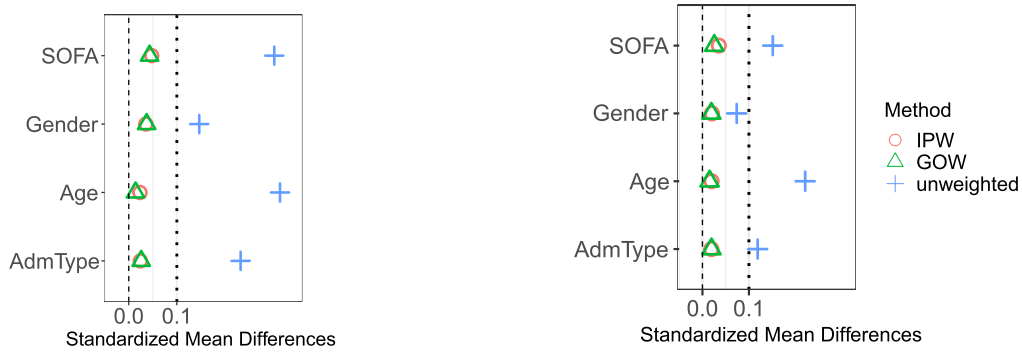


Figure 1: Smoothed density estimates for W^{P}_{jk} for $j = 1, \ldots, m$. The five curves in each subfigure are obtained using data subset k with $k = 1, \ldots, m$.



(a) ASD values (b) PSD values

Figure 2: ASD values and PSD values, respectively displayed by (a) and (b), are obtained for the four covariates using the parametric IPW, GOW and unweighted methods.

propensity weighting methods, i.e., IPW and GOW, perform similarly and produce results in close agreement with those yielded from the two augmented parametric propensity weighting methods, IPW-aug and GOW-aug. The results indicate that physicians 4 and 5 have better performance on improving the probability of survival for patients admitted to ICU, though the improvement of physician 5 over physician 2 is not statistically significant.

Table 2: Estimates of GWAPE and their associated 95% confidence interval, via parametric propensity weighting and super learning methods. The highlights indicate the difference between parametric and super-learning methods.

		Parametric method				Super learning method			
		IPW	IPW-aug	GOW	GOW-aug	IPW-SL	IPW-aug-SL	GOW-SL	GOW-aug-SL
$\tau^h(1,2)$	Point estimate	0.753	0.737	0.757	0.740	0.712	0.722	0.723	0.733
	95% C.I.	(0.389, 1.458)	**(0.378, 1.437)**	(0.393, 1.459)	**(0.380, 1.442)**	(0.439, 1.154)	**(0.608, 0.858)**	(0.452, 1.158)	**(0.623, 0.862)**
	Length of C.I.	1.070	1.059	1.065	1.063	0.715	0.250	0.706	0.239
$\tau^h(1,3)$	Point estimate	0.796	0.786	0.801	0.789	0.751	0.726	0.770	0.746
	95% C.I.	(0.440, 1.439)	**(0.447, 1.385)**	(0.446, 1.438)	**(0.447, 1.391)**	(0.488, 1.155)	**(0.617, 0.854)**	(0.504, 1.176)	**(0.642, 0.868)**
	Length of C.I.	0.999	0.938	0.992	0.944	0.667	0.237	0.672	0.226
$\tau^h(1,4)$	Point estimate	0.494	0.492	0.496	0.493	0.480	0.480	0.491	0.491
	95% C.I.	(0.288, 0.847)	(0.289, 0.836)	(0.290, 0.849)	(0.289, 0.842)	(0.290, 0.793)	(0.413, 0.559)	(0.298, 0.808)	(0.429, 0.562)
	Length of C.I.	0.559	0.547	0.559	0.552	0.503	0.146	0.510	0.133
$\tau^h(1,5)$	Point estimate	0.536	0.532	0.533	0.530	0.506	0.499	0.516	0.509
	95% C.I.	(0.299, 0.958)	(0.297, 0.952)	(0.300, 0.949)	(0.297, 0.947)	(0.322, 0.794)	(0.439, 0.568)	(0.332, 0.803)	(0.450, 0.577)
	Length of C.I.	0.659	0.655	0.650	0.650	0.472	0.129	0.471	0.127
$\tau^h(2,3)$	Point estimate	1.056	1.068	1.058	1.066	1.055	1.004	1.064	1.018
	95% C.I.	(0.679, 1.644)	(0.684, 1.668)	(0.680, 1.646)	(0.685, 1.659)	(0.696, 1.598)	(0.898, 1.124)	(0.706, 1.604)	(0.918, 1.129)
	Length of C.I.	0.966	0.984	0.966	0.974	0.902	0.226	0.898	0.211
$\tau^h(2,4)$	Point estimate	0.656	0.668	0.655	0.667	0.674	0.665	0.678	0.669
	95% C.I.	**(0.440, 0.979)**	(0.448, 0.996)	**(0.440, 0.977)**	(0.449, 0.991)	**(0.422, 1.078)**	(0.606, 0.730)	**(0.426, 1.078)**	(0.607, 0.738)
	Length of C.I.	**0.539**	0.548	**0.537**	0.452	**0.656**	0.124	**0.652**	0.130
$\tau^h(2,5)$	Point estimate	0.711	0.722	0.704	0.717	0.710	0.691	0.714	0.695
	95% C.I.	(0.472, 1.070)	**(0.476, 1.094)**	(0.467, 1.062)	**(0.471, 1.090)**	(0.488, 1.034)	**(0.629, 0.759)**	(0.497, 1.025)	**(0.628, 0.769)**
	Length of C.I.	0.598	0.617	0.595	0.619	0.546	0.130	0.528	0.141
$\tau^h(3,4)$	Point estimate	0.621	0.626	0.620	0.626	0.639	0.662	0.637	0.658
	95% C.I.	(0.433, 0.892)	(0.444, 0.881)	(0.432, 0.889)	(0.442, 0.886)	(0.420, 0.973)	(0.603, 0.726)	(0.418, 0.973)	(0.606, 0.714)
	Length of C.I.	**0.459**	0.437	**0.457**	0.444	**0.553**	0.123	**0.555**	0.108
$\tau^h(3,5)$	Point estimate	0.673	0.676	0.666	0.672	0.674	0.688	0.671	0.682
	95% C.I.	(0.471, 0.962)	(0.480, 0.952)	(0.461, 0.962)	(0.471, 0.959)	(0.481, 0.943)	(0.629, 0.753)	(0.479, 0.939)	(0.626, 0.744)
	Length of C.I.	0.491	0.471	0.501	0.488	0.462	0.124	0.460	0.119
$\tau^h(4,5)$	Point estimate	1.084	1.081	1.075	1.075	1.053	1.039	1.052	1.038
	95% C.I.	(0.796, 1.475)	(0.791, 1.477)	(0.800, 1.444)	(0.793, 1.457)	(0.759, 1.461)	(0.986, 1.095)	(0.759, 1.459)	(0.986, 1.093)
	Length of C.I.	**0.679**	0.685	**0.644**	0.664	**0.702**	0.109	**0.700**	0.107

Comparing the results from the two IPW-based methods for the study population, the IPW method shows that the proportions of being alive for ICU discharged patients assigned to physicians 1, 2 and 3 are respectively 0.494, 0.656 and 0.621 times of that for patients assigned to physician 4, and the IPW-aug method estimates those proportions to be 0.492, 0.668 and 0.626, respectively; all those differences are statistically significant at the significance level 0.05. Regarding the results of GOW-based methods for the overlap population, the GOW method indicates that the proportions of being alive for ICU discharged patients assigned to physicians 1 and 3 are 0.533 and 0.666 times of that for patients assigned to physician 5, respectively, and the GOW-aug method suggests those proportions to be respectively 0.530 and 0.672; all those differences are statistically significant at the significance level 0.05.

5.2 Determination of GWAPE Using Super Learning

In contrast to Section 5.1, we begin by estimating the generalized propensity scores $\pi_j(X)$ with $j = 1\ldots, m$ by employing the super learning method to four candidate learners that are derived from the tree ensemble methods. To be specific, we employ XGBoost and Random Forests, each with two sets of hyperparameter values, as the four candidate learners for the super learning method. For the XGBoost method, we set the hyperparameter controlling the learning rate to 0.001; the maximum number of boosting iterations to 20000; and the maximum depth of a tree to 2 and 1, respectively. For the random forest method, we set the hyperparameter controlling the number of features to possibly split at in each node of a tree to 3; and the maximal tree depth to 200 and 100, respectively.

Both the Random Forest and the XGBoost methods utilize multiple samplings through iteratively re-weighted or bootstrap procedures to enhance the performance of the single classification and regression tree (CART) algorithms and reduce overfitting. CART algorithms aim to partition the ICU data into suitable regions, ensuring patients are as homogeneous as possible within a region.

Analogous to the procedure in Section 5.1, we divide the original dataset into m subsets. For X in the kth data subset, we denote the super learned propensity score $\hat{\pi}_j(X)$ as W_{jk}^{SL} for $j, k = 1, \ldots, m$. We then generate m approximated density estimates for W_{jk}^{SL} with $k = 1, \ldots, m$. The results are reported in Figure 3, where the smoothed density estimates for physicians $1, \ldots, 5$ are shown in subfigures (a)-(e), respectively. Similar to the discussion in Section 5.1, the five approximated density functions in each of subfigure (a)-(e) exhibit a considerable overlap. Consequently, we may infer that the overlapping population well approximates the entire population. Therefore, the estimation results obtained from the super learning method, coupled with IPW-based methods and GOW-based methods, are likely to be closely aligned.

We compare the propensity estimates produced by the super learning method with those derived from the parametric method in Figure 4. The generalized propensity scores obtained from the super learning method, typically ranging from 0.05 to 0.4, generally exhibit similarities but a slightly greater dispersion compared to those produced by the parametric method, which typically fall within the range of 0.1 to 0.3 for most cases.

In Figure 5, the ASDs and PSDs for the four covariates under various weighting methods are presented in the two respective panels. Similar to the parametric case discussed in Section 5.1, we observe that the imbalance is negligible when employing IPW or GOW methods, while the covariates remain unbalanced in the absence of any adjustment.

We then proceed to estimate the quality function $Q_j(X_i)$ using the super learning method. Similar to $\pi_j(X)$, we employ XGBoost and Random Forests, each with two sets of different

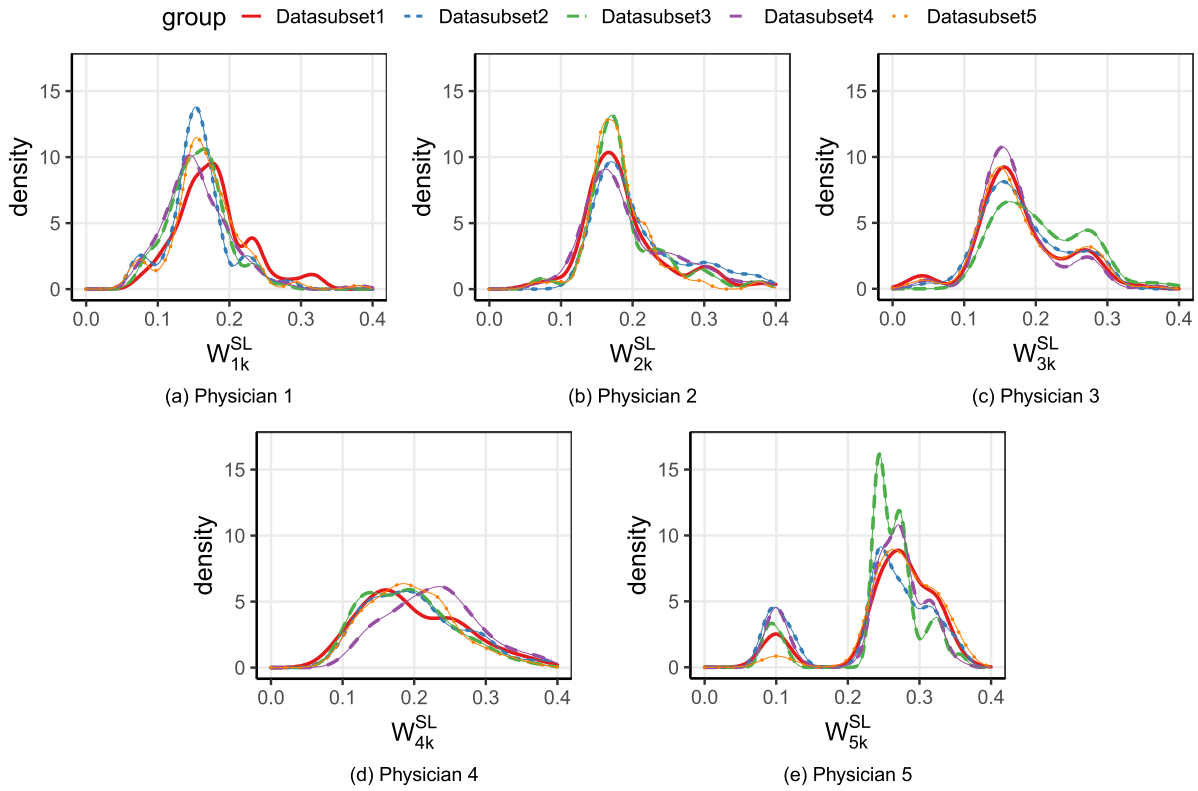


Figure 3: Smoothed density estimates for W^{SL}_{jk} for $j = 1, \ldots, m$. The five curves in each subfigure are obtained using the data subsets $k = 1, \ldots, m$, respectively.

hyperparameters, as the four candidate learners for super learning. With $\hat{\pi}^{\text{SL}}_j(X_i)$ and $\hat{Q}^{\text{SL}}_j(X_i)$ estimated, we incorporate them into (6) and (8) and present the summarized results in Table 2. These four methods, utilizing super learning with IPW, IPW augmented, GOW, and GOW augmented, are denoted as IPW-SL, IPW-aug-SL, GOW-SL, and GOW-aug-SL, respectively.

We now analyze the data using the super learning propensity weighting methods, described in Section 4.2, and display the results in Table 2. The super learning methods produce estimates in close agreement with those yielded from the parametric methods. They, however, tend to have tighter confidence intervals, constructed using bootstrap standard errors, derived from using twenty-five bootstrap samples, especially for the super learning augmented methods. Some 95% confidence intervals for the parametric methods and super learning methods lead to different results for statistical significance. For example, for $\tau^h(1,2)$, $\tau^h(1,3)$ and $\tau^h(2,5)$, the 95% confidence intervals for parametric augmented methods include 1, while those for super learning methods exclude 1; and for $\tau^h(2,4)$, the 95% confidence intervals for parametric methods exclude 1, while those for super learning methods include 1. Overall, though some differences exist, we reach the same conclusion as in Section 5.1 that the treatment of physicians 4 and 5 result in higher probability of survival for patients admitted to ICU.

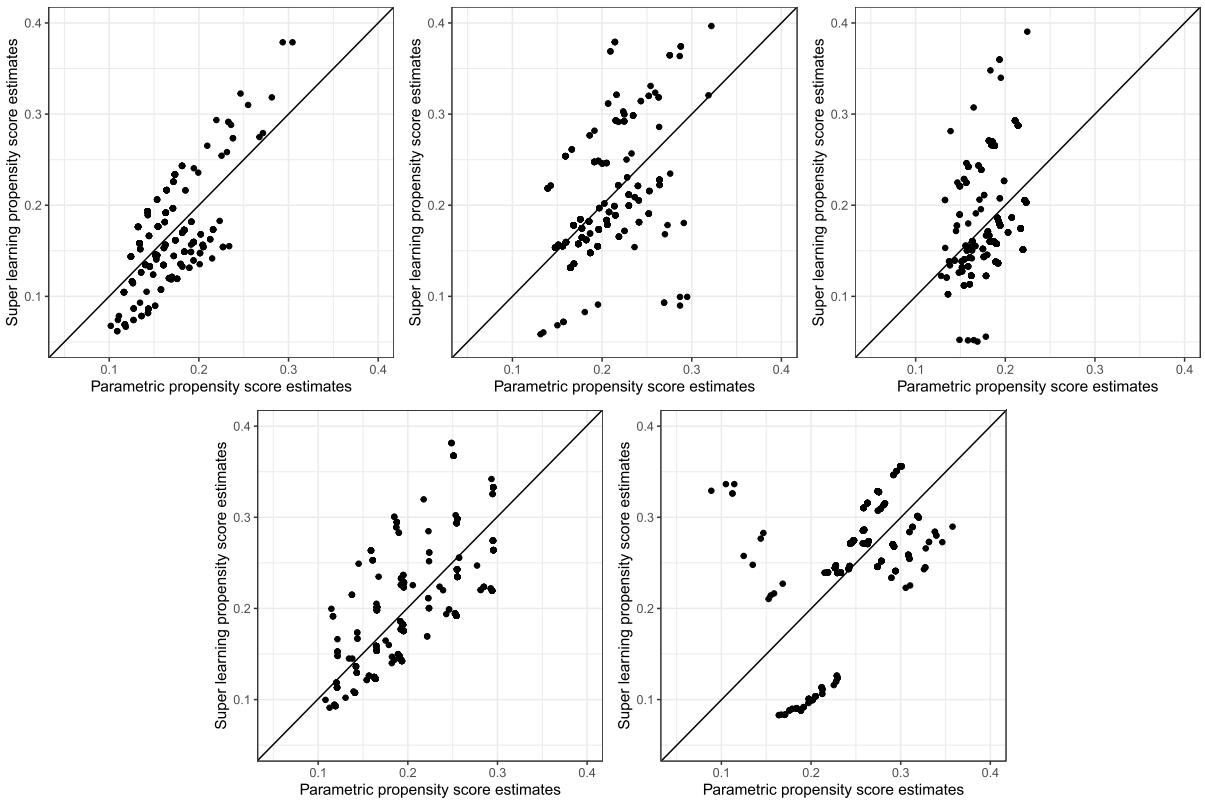
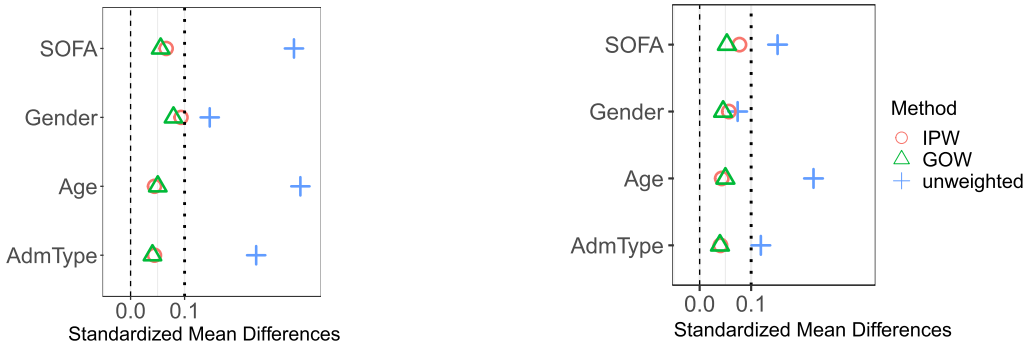


Figure 4: Scatterplot of propensity scores estimated by the super learning versus parametric method for all observations with respect to the 5 physician groups.



(a) ASD values

(b) PSD values

Figure 5: ASD values and PSD values, respectively displayed by (a) and (b), are obtained for the four covariates using the super learning IPW, GOW and unweighted methods.

6 Discussion

In this paper, we employ the propensity weighting methods to analyze critical care data. By specifying the tilting function, we obtain the estimates of GWAPE for the target population. The data analysis suggests that both parametric and super learning-based propensity weighting

methods offer effective tools for determining physician effects. When the generalized propensity or/and the quality function is not correctly specified under the parametric propensity weighting framework, super learning based propensity weighting methods lead to more efficient estimators. However, there might be hesitancy in using super learning methods in practice due to the black-box nature of machine learning-based candidate learners, which obscures the interpretability of the methods. Causal inference for discovering physician effects, an approach of interest to hospital administrators and patients, provides a way to assess the performance of physicians.

It is worth emphasizing that although the focus here is on pairwise mean comparison and discovering physician effects on the target population, the proposed procedures can be easily extended to the *Q-learning* framework (Watkins, 1989) if personalized physician recommendation is the target. Specifically, for a patient with covariates X, the recommended physician can be obtained as the physician maximizing a consistent estimator of the quality function $Q_j(X)$, denoted $\hat{Q}_j(X)$:

$$\hat{A}^{\text{opt}} = \underset{j \in \mathcal{A}}{\text{argmax}}\, \hat{Q}_{A=j}(X).$$

Moreover, Schulz and Moodie (2021) demonstrated that the weighted ordinary least squares regression of Y using either IPW or GOW yields consistent estimators for $Q_j(A)$, allowing for the recommendation of a physician with the most beneficial outcome. However, relying solely on point estimates may be insufficient to inform decisions. Therefore, it is also interesting to further incorporate the conformal inference techniques to construct valid prediction intervals for the recommended physician. Discussions from Lei and Candès (2021) may be adapted to accommodate the scenario involving multiple physicians here.

The parametric and machine learning based propensity weighting framework has promising applications. The framework can be used to establish a principle for assessing the performance of employees, corporations, or organizations in a specific aspect and to make further recommendations based on that principle.

Supplementary Material

The R code for this paper can be found at the Journal of Data Science website.

Acknowledgement

The authors thank the reviewers for their constructive comments.

Funding

This research is partially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC). Yi is Canada Research Chair in Data Science (Tier 1). Her research was undertaken, in part, thanks to funding from the Canada Research Chairs Program.

References

Austin PC (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46: 399–424. https://doi.org/10.1080/00273171.2011.568786

Austin PC, Stuart EA (2015). Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*, 34: 3661–3679. https://doi.org/10.1002/sim.6607

Cole SR, Frangakis CE (2009). The consistency statement in causal inference: A definition or an assumption? *Epidemiology*, 20: 3–5. https://doi.org/10.1097/EDE.0b013e31818ef366

Cox DR (1958). *Planning of Experiments*. Wiley, New Jersey.

Coyle JR, Hejazi NS, Malenica I, Phillips RV, Sofrygin O (2022). sl3: Modern pipelines for machine learning and Super Learning. https://github.com/tlverse/sl3. R package version 1.4.4.

Ding P, Li F (2018). Causal inference: A missing data perspective. *Statistical Science*, 33: 214–237. https://doi.org/10.1214/18-STS645

Imbens GW (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, 87: 706–710. https://doi.org/10.1093/biomet/87.3.706

Imbens GW (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86: 4–29. https://doi.org/10.1162/003465304323023651

Imbens GW, Rubin DB (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, New York.

Lechner M (2001). Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In: *Econometric Evaluation of Labour Market Policies* (Lechner, M, Pfeiffer, F, eds.), 43–58. Springer, New York.

Lee BK, Lessler J, Stuart EA (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine*, 29: 337–346. https://doi.org/10.1002/sim.3782

Lei L, Candès EJ (2021). Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society, Series B, Statistical Methodology*, 83: 911–938. https://doi.org/10.1111/rssb.12445

Li F, Li F (2019). Propensity score weighting for causal inference with multiple treatments. *Annals of Applied Statistics*, 13: 2389–2415. https://doi.org/10.1214/19-AOAS1282

Li F, Morgan KL, Zaslavsky AM (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113: 390–400. https://doi.org/10.1080/01621459.2016.1260466

Li F, Thomas LE, Li F (2019). Addressing extreme propensity scores via the overlap weights. *American Journal of Epidemiology*, 188: 250–257. https://doi.org/10.1093/aje/kwy201

Luedtke AR, van der Laan MJ (2016). Super-learning of an optimal dynamic treatment rule. *The International Journal of Biostatistics*, 12: 305–332. https://doi.org/10.1515/ijb-2015-0052

McCaffrey DF, Griffin BA, Almirall D, Slaughter ME, Ramchand R, Burgette LF (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine*, 32: 3388–3414. https://doi.org/10.1002/sim.5753

McCaffrey DF, Ridgeway G, Morral AR (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, 9: 403–425. https://doi.org/10.1037/1082-989X.9.4.403

Petersen ML, Porter KE, Gruber S, Wang Y, Van Der Laan MJ (2012). Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research*, 21: 31–54. https://doi.org/10.1177/0962280210386207

Pirracchio R, Petersen ML, van der Laan MJ (2015). Improving propensity score estimators' robustness to model misspecification using super learner. *American Journal of Epidemiology*, 181: 108–119. https://doi.org/10.1093/aje/kwu253

Polley EC, LeDell E, Kennedy C, Lendle S, van der Laan MJ (2021). Superlearner: Super learner prediction. https://CRAN.R-project.org/package=SuperLearner. R package version 2.0-28.

Polley EC, van der Laan MJ (2010). Super learner in prediction. U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 266.

Robins JM, Hernán MA, Brumback B (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11: 550–560. https://doi.org/10.1097/00001648-200009000-00011

Rosenbaum PR, Rubin DB (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70: 41–55. https://doi.org/10.1093/biomet/70.1.41

Rubin DB (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66: 688–701. https://doi.org/10.1037/h0037350

Rubin DB (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association*, 75: 591–593.

Rubin DB (1990). Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, 5: 472–480.

Sarker IH (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, 2: Article 160.

Schulz J, Moodie EEM (2021). Doubly robust estimation of optimal dosing strategies. *Journal of the American Statistical Association*, 116: 256–268. https://doi.org/10.1080/01621459.2020.1753521

Setoguchi S, Schneeweiss S, Brookhart MA, Glynn RJ, Cook EF (2008). Evaluating uses of data mining techniques in propensity score estimation: A simulation study. *Pharmacoepidemiology and Drug Safety*, 17: 546–555. https://doi.org/10.1002/pds.1555

Spreeuwenberg MD, Bartak A, Croon MA, Hagenaars JA, Busschbach JJV, Andrea H, et al. (2010). The multiple propensity score as control for bias in the comparison of more than two treatment arms: An introduction from a case study in mental health. *Medical Care*, 48: 166–174. https://doi.org/10.1097/MLR.0b013e3181c1328f

SSC (2022). Developing a physician performance model in critical care: Assessing quality and value. https://ssc.ca/en/case-study/developing-a-physician-performance-model-critical-care-assessing-quality-and-value. Accessed: 2022-09-10.

van der Laan MJ, Polley EC, Hubbard AE (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6: Article 25.

van der Laan MJ, Rose S (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, New York.

Watkins CJ (1989). Learning from delayed rewards, Ph.D. thesis, Cambridge.

Westreich D, Lessler J, Funk MJ (2010). Propensity score estimation: Neural networks, support vector machines, decision trees (cart), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology*, 63: 826–833. https://doi.org/10.1016/j.jclinepi.2009.11.020

Zhou T, Tong G, Li F, Thomas LE, Li F (2022). PSweight: An R package for propensity score weighting analysis. *The R Journal*, 14: 282–300. https://doi.org/10.32614/RJ-2022-011

Zhou Y, Matsouaka RA, Thomas LE (2020). Propensity score weighting under limited overlap and model misspecification. *Statistical Methods in Medical Research*, 29: 3721–3756. https://doi.org/10.1177/0962280220940334

Zivich PN, Breskin A (2021). Machine learning for causal inference: On the use of cross-fit estimators. *Epidemiology*, 32: 393–401. https://doi.org/10.1097/EDE.0000000000001332