Biostatistics, 2024, **25**, 933–946
https://doi.org/10.1093/biostatistics/kxae002
Advance Access publication February 9, 2024

OXFORD

Estimation of optimal treatment regimes with electronic medical record data using the residual life value estimator

Grace Rhodes[1,*], Marie Davidian[2], Wenbin Lu[2]

[1]Eli Lilly and Company, Indianapolis, IN 46204, USA
[2]Department of Statistics, North Carolina State University, SAS Hall, 2311 Stinson Dr, Raleigh, NC 27607, USA

*Corresponding author: Eli Lilly and Company, Indianapolis, IN 46204, USA. Email: grace.rhodes.stat@gmail.com

SUMMARY

Clinicians and patients must make treatment decisions at a series of key decision points throughout disease progression. A dynamic treatment regime is a set of sequential decision rules that return treatment decisions based on accumulating patient information, like that commonly found in electronic medical record (EMR) data. When applied to a patient population, an optimal treatment regime leads to the most favorable outcome on average. Identifying optimal treatment regimes that maximize residual life is especially desirable for patients with life-threatening diseases such as sepsis, a complex medical condition that involves severe infections with organ dysfunction. We introduce the residual life value estimator (ReLiVE), an estimator for the expected value of cumulative restricted residual life under a fixed treatment regime. Building on ReLiVE, we present a method for estimating an optimal treatment regime that maximizes expected cumulative restricted residual life. Our proposed method, ReLiVE-Q, conducts estimation via the backward induction algorithm Q-learning. We illustrate the utility of ReLiVE-Q in simulation studies, and we apply ReLiVE-Q to estimate an optimal treatment regime for septic patients in the intensive care unit using EMR data from the Multiparameter Intelligent Monitoring Intensive Care database. Ultimately, we demonstrate that ReLiVE-Q leverages accumulating patient information to estimate personalized treatment regimes that optimize a clinically meaningful function of residual life.

KEYWORDS: context vector; dynamic treatment regime; electronic medical record; MIMIC-III; precision medicine; Q-learning; random forest; residual life; sepsis.

1. INTRODUCTION

A dynamic treatment regime is a set of sequential decision rules that provides a formal process for making treatment decisions at a series of key decision points throughout disease progression. At each decision point, the decision rule accepts patient history as input and returns a recommended treatment from among the feasible options. An optimal treatment regime leads to the most favorable outcome on average when it is used to select treatments for a patient population. Identifying an optimal regime is an integral component of precision medicine, an approach to healthcare that focuses on tailoring treatment decisions to patient characteristics.

For patients with potentially life-threatening conditions, identification of an optimal treatment regime that maximizes remaining life is of particular interest. Formally, the remaining life of a

Received: January 19, 2023. **Revised:** November 10, 2023. **Accepted:** January 3, 2024

patient at time t, given the patient has survived up to time t, is referred to as residual life. Sepsis is a life-threatening medical condition that involves severe infections with organ dysfunction and is a leading cause of death worldwide (Singer et al., 2016). Although international guidelines for sepsis treatment have been established, treating septic patients remains highly challenging, as the heterogeneity of the septic patient population results in differing responses to medical intervention (László et al., 2015). Consequently, identifying optimal decision rules that maximize residual life and select treatment based on patients' individualized characteristics is highly desirable for septic patient care.

We are especially interested in identifying optimal treatment regimes for septic patients in the intensive care unit (ICU) using their electronic medical record (EMR) data. In particular, we study a data set of septic patients constructed from the Multiparameter Intelligent Monitoring Intensive Care database (MIMIC-III), a freely available database comprised of de-identified medical records for over 40,000 critical care patients (Johnson et al., 2016). Given a number of patient characteristics, including admission records, longitudinal vital signs, and longitudinal laboratory measurements, we seek to estimate an optimal treatment regime that maximizes residual life for the septic patients in MIMIC-III.

An overview of methods for estimating optimal treatment regimes based on censored time-to-event outcomes can be found in Chapter 8 of Tsiatis et al. (2020). One class of methods seeks to directly optimize an estimator for the value of a regime, where the value of a regime is defined in terms of the expectation of a given function of the event time, and optimization is conducted over all treatment regimes in a restricted class. Zhao et al. (2015), Bai et al. (2017), Cui et al. (2017), Zhou et al. (2021), and Wang et al. (2022) introduced direct optimization methods for regimes limited to only a single decision point, while Jiang et al. (2017a), Jiang et al. (2017b), Hager et al. (2018), Zhao et al. (2020), Xue et al. (2022), and Choi et al. (2023) presented methods for estimating optimal multistage regimes. In practice, implementing direct optimization methods with observational EMR data can be computationally prohibitive. First, EMR data often include a large number of covariates. Thus, the class of possible regimes derived from EMR data is likely to be large, rendering direct optimization computationally intensive. Second, estimators for the value of a regime often involve weighting by patients' propensities for being consistent with the given regime. In EMR data with a large number of possible regimes, the number of patients observed to be consistent with any given regime may be low, resulting in unstable value estimates. These complications become especially severe in settings with a large number of decision points.

Given the limitations of direct optimization, Q-learning has become a popular method for estimating optimal treatment regimes from observational data. Q-learning is a backward induction algorithm introduced in the reinforcement learning literature to solve multistage decision problems (Bellman, 1957; Watkins and Dayan, 1992). The Q-learning algorithm considers three key elements at each stage: a state, an action, and a reward. Q-learning estimates a sequence of decision rules, or a policy, that maximizes the expected sum of rewards. At each stage, a decision rule accepts the accumulated states and actions as input and returns a recommended action. Like direct optimization, Q-learning estimates become increasingly unstable as the number of stages increases. However, the Q-learning algorithm can more readily estimate optimal policies in settings with numerous covariates and low levels of observed consistency with studied policies.

In the clinical setting, a stage corresponds to a decision point, a state corresponds to patient history, an action corresponds to treatment, and a policy corresponds to a dynamic treatment regime. Moreover, a reward is defined to be some quantitative measure of patient health. The Q-learning algorithm estimates an optimal treatment regime that maximizes the expected sum of rewards using a backward recursive approach. First, the algorithm estimates an optimal decision rule at the final decision point, then the algorithm sequentially estimates an optimal decision rule at each previous decision point. At each decision point, a Q-model is posited for the expectation of the sum of current and future rewards, given patient history and treatment. To estimate optimal decision rules, each Q-model is optimized with respect to treatment, with all future treatment decisions taken to be optimal (Murphy, 2005).

The Q-learning framework has been extended to accommodate censored time-to-event outcomes. Zhao et al. (2011), Goldberg and Kosorok (2012), Huang et al. (2014), and Zhao et al. (2020) all introduced Q-learning methods to estimate optimal treatment regimes that maximize expected survival time beyond the first decision point, where survival time is restricted to a constant. For regimes comprised of a fixed number of decision points, Zhao et al. (2011) presented a Q-learning method that used support vector regression to fit the Q-models. Goldberg and Kosorok (2012) extended the Q-learning algorithm to accommodate a flexible number of decision points, positing Q-models of an unspecified functional form adjusted by inverse probability of censoring weights. Huang et al. (2014) implemented a related approach that relied on fitting accelerated failure time models. Zhao et al. (2020) extended the Q-learning algorithm presented by Goldberg and Kosorok (2012) to allow shared decision rule parameters across decision points. Recent innovations in the field include an imputation-based Q-learning approach presented by Lyu et al. (2023), which uses non- or semiparametric models to estimate optimal treatment rules, followed by multiple imputation to predict optimal potential survival times. Additionally, Illenberger et al. (2023) introduced a combined Q-learning and policy-search method to estimate optimal list-based treatment regimes with a constraint imposed on expected treatment costs. Notably, all of the presented Q-learning methods defined the reward at each decision point to be a measure of the incremental amount of time between the given decision point and the next decision point or failure, whichever comes first.

Alternative reinforcement learning algorithms have been proposed for estimating optimal treatment regimes with censored time-to-event outcomes. One approach, called A-learning, employs a similar backward recursive strategy to Q-learning, but estimates the optimal decision rule at each decision point by optimizing only the portion of the regression outcome that involves differences between treatments (Murphy, 2003; Robins, 2004). While this makes A-learning robust to misspecification of the outcome models, A-learning also requires modeling the probability of the observed treatments given patient history at each decision point. Simoneau et al. (2020) and Zhang et al. (2022) introduced A-learning methods to estimate optimal treatment regimes that maximize expected restricted survival time beyond the first decision point. Alternatively, Cho et al. (2022) presented a backward recursive algorithm to estimate an optimal treatment regime that maximizes either mean survival time or survival probability at a given timepoint. As opposed to Q-learning and A-learning, the method presented by Cho et al. (2022) does not express the value of a regime in terms of a sum of rewards. At each decision point, Cho et al. (2022) appends conditional survival probability information, rather than a reward, in a backward recursive fashion.

To estimate optimal treatment regimes with censored time-to-event outcomes, we present a new Q-learning method based on an intuitive reward definition. While the aforementioned Q-learning methods define the reward to be the incremental time between decision points or failure, we argue that in clinical application, a more interpretable reward is a patient's remaining survival time, or residual life. Using this reward specification, the Q-learning algorithm estimates an optimal treatment regime that maximizes the expected sum of residual life, where the sum is taken across all decision points reached by the patient. We refer to this quantity as "expected cumulative residual life." To formulate the proposed Q-learning algorithm, we first introduce the residual life value estimator (ReLiVE), which estimates the expected cumulative residual life of a patient under a fixed treatment regime. We then build on ReLiVE to present the residual life value estimator Q-learning method (ReLiVE-Q). By defining the reward at each decision point to be residual life, ReLiVE-Q estimates an optimal treatment regime that maximizes expected cumulative residual life. To guarantee a finite expectation, we study restricted residual life, which we compute based on event times restricted to a fixed constant.

In addition to reward interpretability, ReLiVE-Q offers two advantages that make it especially suitable for estimating optimal treatment regimes from EMR data with censored time-to-event outcomes. First, the relationship between patient history, treatment, and event times is inherently complex and unlikely to be well-represented by a parametric model. At each decision point, ReLiVE-Q fits flexible, nonparametric Q-models using the random forest algorithm (Breiman, 2001). Second, ReLiVE-Q incorporates longitudinal covariates into the Q-models using low-dimensional summaries of the trajectories, called context vectors, that are constructed using modern machine learning techniques (Rhodes et al., 2023). This is important because EMR data often contain a large number of longitudinal covariates measured over a potentially dense time grid, and the number and timing of measurements often differs among patients. Thus, including the entire patient history directly in the Q-models is often computationally prohibitive. Moreover, while summary statistics of the longitudinal covariates could be included in the Q-models, simple summaries are unlikely to provide an adequate synthesis of the complex information contained in the trajectories.

In Section 2, we present the statistical framework for ReLiVE and ReLiVE-Q, and in Section 3, we detail the ReLiVE and ReLiVE-Q methodology. In Section 4, we demonstrate the utility of ReLiVE-Q in a simulation study, and in Section 5, we apply ReLiVE-Q to estimate optimal treatment regimes for septic patients in MIMIC-III. In both the simulation study and MIMIC-III data application, we demonstrate that ReLiVE-Q estimates personalized treatment regimes that optimize a clinically meaningful function of residual life. We conclude with a discussion of implications and open problems in Section 6.

2. Statistical Framework

2.1. Potential outcomes and treatment regimes

We focus our study on settings where treatment decisions are made at fixed intervals, as is the scenario for the studied MIMIC-III data set. Consider a series of K decision points at which treatment decisions must be made, where the decision points occur at fixed times $\tau_1, \ldots, \tau_K$ with $\tau_1 = 0$. Let $\mathbf{x}_1$ denote the vector of covariates measured at τ_1. Then the patient history at decision point 1 is $\mathbf{h}_1 = (\tau_1, \mathbf{x}_1)$. For $k = 1, \ldots, K$, define $\mathcal{A}_k$ to be the finite set of all available treatment options at decision point k, and let $a_k \in \mathcal{A}_k$ be the treatment administered at decision point k. We assume all $a_k \in \mathcal{A}_k$ are feasible for all subjects at all decision points $k = 1, \ldots, K$, as this assumption holds for the treatments studied in MIMIC-III. For $k = 2, \ldots, K$, let $\mathbf{x}_k$ denote the vector of covariate information accumulated between decision points $k-1$ and k. Moreover, let $\bar{\tau}_k = (\tau_1, \ldots, \tau_k)$, $\bar{\mathbf{x}}_k = (\mathbf{x}_1, \ldots, \mathbf{x}_k)$, and $\bar{a}_k = (a_1, \ldots, a_k)$, where $k = 1, \ldots, K$, and let $\bar{a} = \bar{a}_K$. Then for $k = 2, \ldots, K$, a patient who does not experience the event of interest prior to decision point k will have patient history $\mathbf{h}_k = (\bar{\tau}_k, \bar{\mathbf{x}}_k, \bar{a}_{k-1})$ at decision point k. In contrast, a patient who experiences the event of interest at time t between decision points $k-1$ and k will have patient history $\mathbf{h}_j = (\bar{\tau}_{k-1}, \bar{\mathbf{x}}_{k-1}, \bar{a}_{k-1}, t)$ at each decision point $j = k, \ldots, K$.

For each decision point, $k = 1, \ldots, K$, let $d_k(\mathbf{h}_k)$ be a decision rule. If $\mathbf{h}_k$ indicates that the event of interest has not occurred prior to decision point k, then d_k maps $\mathbf{h}_k$ to $\mathcal{A}_k$. Else, if $\mathbf{h}_k$ indicates that the event of interest has occurred prior to decision point k, then $d_k(\mathbf{h}_k)$ returns null. For notational simplicity, suppress the dependence of the decision rule on patient history, and let $d_k = d_k(\mathbf{h}_k)$. Define treatment regime $d = (d_1, \ldots, d_K)$, and denote the set of all such possible treatment regimes as $\mathcal{D}$. For convenience, let $\bar{d}_k = (d_1, \ldots, d_k)$, where $k = 1, \ldots, K$.

To formalize the definition of the residual life that would occur if an individual were to follow a given regime $d \in \mathcal{D}$, we define the potential outcomes that would be achieved if a randomly selected patient were treated according to d. To this end, we first define the potential outcomes associated with a given sequence of treatments $\bar{a}$. Let $\kappa^*(\bar{a})$ be the potential number of decision points a randomly selected patient would reach if administered treatments $\bar{a}$, where $1 \leq \kappa^*(\bar{a}) \leq K$. For $k = 2, \ldots, \kappa^*(\bar{a})$, let $\mathbf{X}_k^*(\bar{a}_{k-1})$ be the vector of potential covariate information that would occur between decision points $k-1$ and k if a patient was administered treatments $\bar{a}_{k-1}$. Further, let $T^*(\bar{a}_{\kappa^*(\bar{a})})$ be the potential time-to-event that would be observed if a patient was administered treatments $\bar{a}_{\kappa^*(\bar{a})}$. Define the set of all possible potential outcomes to be $W^* = \{\text{For all possible } \bar{a} \in \mathcal{A}_1 \times \cdots \times \mathcal{A}_K : \kappa^*(\bar{a}), \mathbf{X}_2^*(a_1), \mathbf{X}_3^*(\bar{a}_2), \ldots, \mathbf{X}_{\kappa^*(\bar{a})}^*(\bar{a}_{\kappa^*(\bar{a})-1}), T^*(\bar{a}_{\kappa^*(\bar{a})})\}$.

As discussed in Sections 6.2.3 and 8.3.2 of Tsiatis et al. (2020), it can be shown that the potential outcomes associated with a given regime $d \in \mathcal{D}$ can be defined in terms of W^*. Let $T^*(d)$ be the potential time-to-event under regime d. To avoid infinite event times, also specify a restricted event time $L > 0$. Then $\kappa_L^*(d) = \max[k : \tau_k < \min\{T^*(d), L\}]$ is the potential number of decision points reached under regime d. For $k = 2, \ldots, \kappa_L^*(d)$, let $\mathbf{X}_k^*(\bar{d}_{k-1})$ be the vector of potential covariate information that would occur between decision points $k-1$ and k if a randomly selected patient was treated according to regime $\bar{d}_{k-1}$. Denote the set of potential outcomes associated with regime d as $W_d^* = \{\kappa_L^*(d), \mathbf{X}_2^*(d_1), \mathbf{X}_3^*(\bar{d}_2), \ldots, \mathbf{X}^*_{\kappa_L^*(d)}(\bar{d}_{\kappa_L^*(d)-1}), T^*(d)\}$.

Because sepsis is a life-threatening medical condition, we specify $T^*(d)$ to be the potential time of death under regime d. Then, restricted residual life under regime d at a given time τ is $\min\{T^*(d), L\} - \tau$. We specify the reward at each decision point $k = 1, \ldots, \kappa_L^*(d)$ to be the patient's restricted residual life under regime d at τ_k. Let the cumulative restricted residual life of a patient under regime d be the sum of the patient's restricted residual life under d, taken across all decision points reached by the patient. We define the value of regime $d \in \mathcal{D}$, $\mathcal{V}(d)$, to be the expected value of cumulative restricted residual life under d. That is, we define

$$\mathcal{V}(d) = E\left\{\sum_{j=1}^{K} I[\min\{T^*(d), L\} > \tau_j] \cdot \left[\min\{T^*(d), L\} - \tau_j\right]\right\}. \tag{2.1}$$

An optimal treatment regime $d^{\text{opt}} \in \mathcal{D}$ satisfies the condition that $\mathcal{V}(d^{\text{opt}}) \geq \mathcal{V}(d)$ for all $d \in \mathcal{D}$. Thus, an optimal regime d^{opt} maximizes expected cumulative restricted residual life across regimes $d \in \mathcal{D}$. Based on observed EMR data, we first aim to estimate $\mathcal{V}(d)$ for any fixed regime $d \in \mathcal{D}$. We then aim to estimate an optimal regime d^{opt} and its value $\mathcal{V}(d^{\text{opt}})$ from the observed data.

2.2. Data and assumptions

We now describe the observed data. As is conventional in the survival analysis literature, let $T > 0$ and $C > 0$ denote the potential times to death and censoring, respectively. We observe only $U = \min(T, C)$ and $\Delta = I(T < C)$, an indicator of whether death is observed $(\Delta = 1)$ or censored $(\Delta = 0)$. Accounting for the restricted lifetime $L > 0$, we observe the restricted outcome $U^L = \min(U, L)$ and associated indicator $\Delta^L = I\{\min(T, L) < C\} = \Delta + I(L < U)(1 - \Delta)$. A patient is observed to reach $\kappa_L = \max\{k : \tau_k < U^L\}$ decision points, where $1 \leq \kappa_L \leq K$. Then the observed history at decision points $k = 1, \ldots, \kappa_L$ is $\mathbf{H}_k = (\bar{\tau}_k, \bar{\mathbf{X}}_k, \bar{A}_{k-1})$, and the observed history at decision points $k = \kappa_L + 1, \ldots, K$ is $\mathbf{H}_k = (\bar{\tau}_{\kappa_L}, \bar{\mathbf{X}}_{\kappa_L}, \bar{A}_{\kappa_L}, U^L, \Delta^L)$. Given $i = 1, \ldots, m$ patients, the observed data are independent and identically distributed $(\kappa_{Li}, \mathbf{X}_{1i}, A_{1i}, \ldots, \mathbf{X}_{\kappa_{Li}i}, A_{\kappa_{Li}i}, U_i^L, \Delta_i^L)$.

We aim to estimate value (2.1) for a given regime $d \in \mathcal{D}$, and to estimate an optimal regime d^{opt} and its value, based on the observed data. Because $T^*(d)$ is defined in terms of W^*, we must be able to express relevant functionals of the distribution of the potential outcomes in terms of the observed data. Such expression is possible under three key, standard assumptions: the Stable Unit Treatment Value Assumption (SUTVA), the Sequential Randomization Assumption (SRA), and the positivity assumption. SUTVA, commonly referred to as the consistency assumption, states that the observed data are the same as those that would potentially be achieved under the treatment decisions observed to be administered. Formally, we assume $\kappa_L = \kappa_L^*(\bar{A}_{\kappa_L})$ and $T = T^*(\bar{A}_{\kappa_L})$. We also assume $\mathbf{X}_k = \mathbf{X}_k^*(\bar{A}_{k-1})$ for $k = 2, \ldots, \kappa_L$. SRA states that the treatment administered at each decision point is independent of the set of potential outcomes conditional on the history. For $k = 1, \ldots, K$, we assume $W^* \perp A_k | \mathbf{H}_k, \kappa_L \geq k$, where $\perp$ denotes independence. The positivity assumption states that all treatment options at each decision point are represented in the data. This is necessary to specify the distributions of the potential outcomes for all $d \in \mathcal{D}$ in terms of the observed data. Formally, we assume $P(A_k = a_k \mid \mathbf{H}_k = \mathbf{h}_k, \kappa_L \geq k) > 0$ for all $a_k \in \mathcal{A}_k$ and for all possible $\mathbf{h}_k$ such that $P(\mathbf{H}_k = \mathbf{h}_k, \kappa_L \geq k) > 0$, $k = 1, \ldots, K$. For simplicity, we also assume censoring is independent of treatment assignment, patient characteristics, restricted lifetime, and potential outcomes. See Section 6 for a discussion on relaxing this assumption.

3. METHODS

3.1. Methodological foundation of ReLiVE

We first introduce the methodological foundation for ReLiVE, our proposed estimator for value (2.1) of a fixed regime. For a given regime $d \in \mathcal{D}$, we construct an estimator for $\mathcal{V}(d)$ by defining K backward recursive Q-functions. We incorporate inverse-probability weights in the Q-functions to account for censoring. For $u > \tau_k$, let $\mathcal{K}_k(u|\mathbf{h}_k, a_k) = P(C > u \mid U^L > \tau_k, \mathbf{H}_k = \mathbf{h}_k, A_k = a_k)$, where $k = 1, \ldots, K$. Under SUTVA and the independent censoring assumption, $\mathcal{K}_k(u|\mathbf{h}_k, a_k) = P(C > u)/P(C > \tau_k)$. See Section 1 of the supplementary material for details.

For a fixed regime $d \in \mathcal{D}$, we define the Q-function at decision point K to be

$$Q_K^d(\mathbf{h}_K, a_K) = E\left\{ \frac{\Delta^L(U^L - \tau_K)}{\mathcal{K}_K(U^L|\mathbf{h}_K, a_K)} \middle| U^L > \tau_K, \mathbf{H}_K = \mathbf{h}_K, A_K = a_K \right\}.$$

Similarly, we define the Q-function at decision points $k = K - 1, \ldots, 1$, to be

$$Q_k^d(\mathbf{h}_k, a_k) = E\left\{ \frac{\Delta^L(U^L - \tau_k)}{\mathcal{K}_k(U^L|\mathbf{h}_k, a_k)} + \frac{I(U^L > \tau_{k+1})V_{k+1}^d(\mathbf{H}_{k+1})}{\mathcal{K}_k(\tau_{k+1}|\mathbf{h}_k, a_k)} \middle| U^L > \tau_k, \mathbf{H}_k = \mathbf{h}_k, A_k = a_k \right\},$$

where $V_k^d(\mathbf{h}_k) = Q_k^d\{\mathbf{h}_k, d_k(\mathbf{h}_k)\}$. Under SUTVA, SRA, and the positivity and independent censoring assumptions, $V_1^d(\mathbf{h}_1)$ is an unbiased estimator of $\mathcal{V}(d)$ if the true Q-functions are known. In Section 2 of the supplementary material, we provide a proof demonstrating $E\{V_1^d(\mathbf{h}_1)\} = \mathcal{V}(d)$. In practice, the Q-functions for regime d are unknown and must be estimated from the data. In Section 3.5, we present the value estimator ReLiVE, which uses a flexible approach to model and estimate the Q-functions.

3.2. Methodological foundation of ReLiVE-Q

Next, we introduce the methodological foundation for ReLiVE-Q, our proposed Q-learning method for estimating an optimal treatment regime and its value. Building on the Q-functions presented in Section 3.1, we characterize an optimal treatment regime $d^{\text{opt}} \in \mathcal{D}$ such that $\mathcal{V}(d^{\text{opt}}) \geq \mathcal{V}(d)$ for all $d \in \mathcal{D}$. At decision point K, define the Q-function for d^{opt} to be

$$Q_K^{d^{\text{opt}}}(\mathbf{h}_K, a_K) = E\left\{ \frac{\Delta^L(U^L - \tau_K)}{\mathcal{K}_K(U^L|\mathbf{h}_K, a_K)} \middle| U^L > \tau_K, \mathbf{H}_K = \mathbf{h}_K, A_K = a_K \right\},$$

and let $d_K^{\text{opt}}(\mathbf{h}_K) = \underset{a_K \in \mathcal{A}_K}{\operatorname{argmax}}\, Q_K^{d^{\text{opt}}}(\mathbf{h}_K, a_K)$. Similarly, at decision points $k = K - 1, \ldots, 1$, define

$$Q_k^{d^{\text{opt}}}(\mathbf{h}_k, a_k) = E\left\{ \frac{\Delta^L(U^L - \tau_k)}{\mathcal{K}_k(U^L|\mathbf{h}_k, a_k)} + \frac{I(U^L > \tau_{k+1})V_{k+1}^{d^{\text{opt}}}(\mathbf{H}_{k+1})}{\mathcal{K}_k(\tau_{k+1}|\mathbf{h}_k, a_k)} \middle| U^L > \tau_k, \mathbf{H}_k = \mathbf{h}_k, A_k = a_k \right\},$$

where $V_k^{d^{\text{opt}}}(\mathbf{h}_k) = Q_k^{d^{\text{opt}}}\{\mathbf{h}_k, d_k^{\text{opt}}(\mathbf{h}_k)\}$ and $d_k^{\text{opt}}(\mathbf{h}_k) = \underset{a_k \in \mathcal{A}_k}{\operatorname{argmax}}\, Q_k^{d^{\text{opt}}}(\mathbf{h}_k, a_k)$. Under SUTVA, SRA, and the positivity and independent censoring assumptions, it can be shown that $d^{\text{opt}} = (d_1^{\text{opt}}, \ldots, d_K^{\text{opt}})$ is an optimal treatment regime satisfying $\mathcal{V}(d^{\text{opt}}) \geq \mathcal{V}(d)$ for all $d \in \mathcal{D}$, and the value of an optimal regime is $\mathcal{V}(d^{\text{opt}}) = E\{V_1^{d^{\text{opt}}}(\mathbf{h}_1)\}$. See Sections 7.2.3–7.2.4 of Tsiatis et al. (2020) for details. In practice, the Q-functions for d^{opt} are unknown and must be estimated from the data. In Section 3.6, we present ReLiVE-Q, a Q-learning method that uses a flexible approach to estimate the Q-functions for d^{opt}.

3.3. Representation of longitudinal covariates

In practice, certain covariates, such as sex and birth date, are collected on patients only at baseline. Denote the vector of covariates measured only at baseline as $\mathbf{S}$. Other covariates, such as vital signs and laboratory values, may be measured repeatedly during a patient's follow up. Let $\mathcal{M}$ be the set of patient-specific measurement times, and denote the vector of longitudinal covariates measured at time $t \in \mathcal{M}$ as $\mathbf{Z}(t) = \{Z_1(t), \dots, Z_p(t)\}$. For longitudinal covariates measured at time $t = \tau_k$, $k = 1, \dots, \kappa_L$, we follow the convention that longitudinal covariates are measured immediately prior to making treatment decisions. Then $\mathbf{X}_1 = \{\mathbf{S}, \mathbf{Z}(0)\}$, and $\mathbf{X}_k = \{\mathbf{Z}(t) : \tau_{k-1} < t \leq \tau_k\}$ for $k = 2, \dots, \kappa_L$.

When estimating the Q-functions for ReLiVE and ReLiVE-Q, it is desirable to leverage all available patient information. Recall that EMR data often contain a large number of longitudinal covariates potentially measured over a dense time grid, and the number and timing of measurements often differs among patients. To incorporate the trajectories of the longitudinal covariates into the Q-models specified in Sections 3.5 and 3.6, it is expedient to construct lower-dimensional summary representations of the longitudinal trajectories. Let $f(\cdot)$ be a vector-valued function, and define vector $\boldsymbol{\zeta}(k) = f[\{\mathbf{Z}(t) : t \leq \tau_k\}]$ to be a function of the longitudinal covariate measurements accumulated by decision point k. Then, $\bar{\mathbf{X}}_1 = \{\mathbf{S}, \mathbf{Z}(0)\}$, and for $k = 2, \dots, \kappa_L$, we define $\bar{\mathbf{X}}_k = \{\mathbf{S}, \boldsymbol{\zeta}(k)\}$. The observed patient history is then given by $\mathbf{H}_1 = (\tau_1, \mathbf{X}_1) = \{\tau_1, \mathbf{S}, \mathbf{Z}(0)\}$ at decision point 1, and $\mathbf{H}_k = (\bar{\tau}_k, \bar{\mathbf{X}}_k, \bar{A}_{k-1}) = \{\bar{\tau}_k, \mathbf{S}, \boldsymbol{\zeta}(k), \bar{A}_{k-1}\}$ at decision points $k = 2, \dots, \kappa_L$. At decision points $k = 1, \dots, K$, we train the Q-models for ReLiVE and ReLiVE-Q over the predictor space $\bar{\mathbf{X}}_k$, as described in Section 3.4.

Intuitively, it is desirable to select a function $f(\cdot)$ that summarizes the longitudinal trajectories from baseline to decision point time. Several simple summary functions are commonly used in practice. The baseline vector $\boldsymbol{\zeta}^B(k)$ contains the longitudinal measurements recorded at baseline, the last-value carried forward vector $\boldsymbol{\zeta}^L(k)$ contains the most recently observed measurements, and the average vector $\boldsymbol{\zeta}^A(k)$ contains the average of the measurements observed through decision point k. Because these simple summary functions are unlikely to capture the complex nature of the longitudinal trajectories, we propose synthesizing the longitudinal covariates using window-specific context vectors constructed via long short-term memory (LSTM) autoencoders, as described in Rhodes et al. (2023). The window-specific context vector $\boldsymbol{\psi}_l(\tau)$ is an encoded representation of the trajectory of longitudinal covariate $Z_l(\cdot)$ from baseline to time $\tau > 0$, where $l = 1, \dots, p$. Specifically, we propose defining $\boldsymbol{\zeta}^C(k) = \{\boldsymbol{\psi}_1(\tau_k), \dots, \boldsymbol{\psi}_p(\tau_k)\}$. A detailed description of the studied forms of $\boldsymbol{\zeta}(\cdot)$ can be found in Section 3 of the supplementary material.

3.4. Q-function estimation strategy

We introduce methods to estimate the Q-functions for ReLiVE and ReLiVE-Q in Sections 3.5 and 3.6, respectively. Although extension to more general treatment settings is possible, we focus on settings with two treatment options such that $\mathcal{A}_k = \{0, 1\}$ for $k = 1, \dots, K$. We model the Q-functions using a non-parametric approach to account for the inherently complex relationship between patient history, treatment, and residual life. Specifically, we estimate $Q_k^d(\mathbf{h}_k, 0)$ and $Q_k^d(\mathbf{h}_k, 1)$ at each decision point $k = 1, \dots, K$ using distinct random forests, where the outcomes used to train the forests are presented in Sections 3.5 and 3.6. A random forest is an ensemble learning algorithm that conducts estimation by combining the output of numerous decision trees. For a detailed overview of the random forest algorithm, see Breiman (2001). At decision point 1, we train the random forests over the predictor space $\bar{\mathbf{X}}_1 = \{\mathbf{S}, \mathbf{Z}(0)\}$. Given a function $\boldsymbol{\zeta}(\cdot)$, at decision points $k = 2, \dots, K$, we train the random forests over the predictor space $\bar{\mathbf{X}}_k = \{\mathbf{S}, \boldsymbol{\zeta}(k)\}$. We train the random forest for $Q_k^d(\mathbf{h}_k, 0)$ on patients having $U^L > \tau_k$ and $A_k = 0$, and we train the random forest for $Q_k^d(\mathbf{h}_k, 1)$ on patients having $U^L > \tau_k$ and $A_k = 1$.

3.5. ReLiVE: value estimation for a fixed regime

We now introduce ReLiVE, which estimates value (2.1) for a fixed regime $d \in \mathcal{D}$. Let $\hat{\mathcal{K}}_k(u|\mathbf{h}_k, a_k) = \hat{G}(u)/\hat{G}(\tau_k)$ for $k = 1, \dots, K$, where $\hat{G}(t)$ is the Kaplan–Meier estimator for the survival function of censoring time C at time t (Kaplan and Meier, 1958). Then $\hat{\mathcal{K}}_k(u|\mathbf{h}_k, a_k) = \hat{\mathcal{K}}_k(u)$. Beginning with decision point K, given patient history $\mathbf{H}_K = \mathbf{h}_K$, we train the random forests for $Q_K^d(\mathbf{h}_K, 0)$ and $Q_K^d(\mathbf{h}_K, 1)$ to estimate the outcome $\{\Delta^L(U^L - \tau_K)\}/\hat{\mathcal{K}}_K(U^L)$.

Denote the random forest estimators of $Q_k^d(\mathbf{h}_k, 0)$ and $Q_k^d(\mathbf{h}_k, 1)$ as $\hat{Q}_k^d(\mathbf{h}_k, 0)$ and $\hat{Q}_k^d(\mathbf{h}_k, 1)$, respectively, for $k = 1, \ldots, K$. At decision points $k = K - 1, \ldots, 1$, given patient history $\mathbf{H}_k = \mathbf{h}_k$, we then train the random forests for $Q_k^d(\mathbf{h}_k, 0)$ and $Q_k^d(\mathbf{h}_k, 1)$ to estimate the outcome

$$\frac{\Delta^L(U^L - \tau_k)}{\hat{\mathcal{K}}_k(U^L)} + \frac{I(U^L > \tau_{k+1})\hat{V}_{k+1}^d(\mathbf{h}_{k+1})}{\hat{\mathcal{K}}_k(\tau_{k+1})},$$

where $\hat{V}_k^d(\mathbf{h}_k) = \hat{Q}_k^d\{\mathbf{h}_k, d_k(\mathbf{h}_k)\}$. Given m patients, we define ReLiVE as

$$\hat{\mathcal{V}}(d) = m^{-1}\sum_{i=1}^{m} \hat{V}_1^d(\mathbf{h}_{1i}),$$

where $\hat{V}_1^d(\mathbf{h}_{1i}) = \hat{Q}_1^d\{\mathbf{h}_{1i}, d_1(\mathbf{h}_{1i})\}$ for patient i with history $\mathbf{h}_{1i}$.

3.6. ReLiVE-Q: estimation of an optimal regime and its value

Next, we introduce ReLiVE-Q, which estimates an optimal treatment regime $d^{\text{opt}} \in \mathcal{D}$ and its value $\mathcal{V}(d^{\text{opt}})$. Beginning with decision point K, given patient history $\mathbf{H}_K = \mathbf{h}_K$, we train the random forests for $Q_K^{d^{\text{opt}}}(\mathbf{h}_K, 0)$ and $Q_K^{d^{\text{opt}}}(\mathbf{h}_K, 1)$ to estimate the outcome $\{\Delta^L(U^L - \tau_K)\}/\hat{\mathcal{K}}_K(U^L)$.

We let $\hat{d}_K^{\text{opt}}(\mathbf{h}_K) = \operatorname*{argmax}_{a_K \in \mathcal{A}_K} \hat{Q}_K^{d^{\text{opt}}}(\mathbf{h}_K, a_K) = I\{\hat{Q}_K^{d^{\text{opt}}}(\mathbf{h}_K, 1) > \hat{Q}_K^{d^{\text{opt}}}(\mathbf{h}_K, 0)\}$, where $a_K = 0$ is the standard treatment. At decision points $k = K - 1, \ldots, 1$, given patient history $\mathbf{H}_k = \mathbf{h}_k$, we then train the random forests for $Q_k^{d^{\text{opt}}}(\mathbf{h}_k, 0)$ and $Q_k^{d^{\text{opt}}}(\mathbf{h}_k, 1)$ to estimate the outcome

$$\frac{\Delta^L(U^L - \tau_k)}{\hat{\mathcal{K}}_k(U^L)} + \frac{I(U^L > \tau_{k+1})\hat{V}_{k+1}^{d^{\text{opt}}}(\mathbf{h}_{k+1})}{\hat{\mathcal{K}}_k(\tau_{k+1})},$$

where $\hat{V}_k^{d^{\text{opt}}}(\mathbf{h}_k) = \hat{Q}_k^{d^{\text{opt}}}\{\mathbf{h}_k, \hat{d}_k^{\text{opt}}(\mathbf{h}_k)\}$. We let $\hat{d}_k^{\text{opt}}(\mathbf{h}_k) = \operatorname*{argmax}_{a_k \in \mathcal{A}_k} \hat{Q}_k^{d^{\text{opt}}}(\mathbf{h}_k, a_k) = I\{\hat{Q}_k^{d^{\text{opt}}}(\mathbf{h}_k, 1) > \hat{Q}_k^{d^{\text{opt}}}(\mathbf{h}_k, 0)\}$, where $a_k = 0$ is the standard treatment. We estimate an optimal treatment regime to be $\hat{d}^{\text{opt}} = (\hat{d}_1^{\text{opt}}, \ldots, \hat{d}_K^{\text{opt}})$. Given m patients, we estimate the value of an optimal treatment regime to be $\hat{\mathcal{V}}(d^{\text{opt}}) = m^{-1}\sum_{i=1}^{m} \hat{V}_1^{d^{\text{opt}}}(\mathbf{h}_{1i})$, where $\hat{V}_1^{d^{\text{opt}}}(\mathbf{h}_{1i}) = \hat{Q}_1^{d^{\text{opt}}}\{\mathbf{h}_{1i}, \hat{d}_1^{\text{opt}}(\mathbf{h}_{1i})\}$ for patient i with history $\mathbf{h}_{1i}$.

3.7. Testing and validation procedures

To evaluate the estimation performance of ReLiVE-Q, we implement testing and validation procedures that compare value estimates for an optimal treatment regime to value estimates for two fixed treatment regimes, the observed treatment regime d^{obs} and the no treatment regime d^{no}. At each decision point $k = 1, \ldots, K$, d^{obs} specifies $d_k(\mathbf{h}_k) = a_k$, where a_k is the treatment observed in the data at decision point k, and d^{no} specifies $d_k(\mathbf{h}_k) = 0$. Thus, d^{obs} represents the treatment regime that was actually administered by clinicians in practice, and d^{no} represents the treatment regime that always administers the standard treatment. In the testing and validation procedures, we compare N value estimates for each regime $d \in \{d^{\text{opt}}, d^{\text{obs}}, d^{\text{no}}\}$. A detailed description of the testing and validation procedures is provided in Section 4 of the supplementary material.

In summary, the testing procedure obtains N value estimates $\hat{\mathcal{V}}^{\text{test}}(d)$ for each regime $d \in \{d^{\text{opt}}, d^{\text{obs}}, d^{\text{no}}\}$ using N unique training/testing data sets. The procedure computes $\hat{\mathcal{V}}^{\text{test}}(d)$ for the fixed regimes $d \in \{d^{\text{obs}}, d^{\text{no}}\}$ by conducting ReLiVE separately on the training and testing data sets. The testing procedure also estimates $\hat{d}^{\text{opt}}$ and its value $\hat{\mathcal{V}}^{\text{test}}(d^{\text{opt}})$ via a cross-validated adaptation of ReLiVE-Q. Moreover, the validation procedure obtains N value estimates $\hat{\mathcal{V}}^{\text{val}}(d)$ for each regime $d \in \{d^{\text{opt}}, d^{\text{obs}}, d^{\text{no}}\}$ by exploiting the known data generation process in a simulation study. Given a simulated data set of patient covariates and a regime d, the validation procedure randomly generates N potential survival times $T^*(d)$ for each patient. Using each of the N randomly generated sets of $T^*(d)$, the procedure computes $\hat{\mathcal{V}}^{\text{val}}(d)$ as the mean of $g\{T^*(d)\} = \sum_{j=1}^{K} I[\min\{T^*(d), L\} > \tau_j] \cdot [\min\{T^*(d), L\} - \tau_j]$ across patients.

We implement the testing and validation procedures using simulated data in Section 4, and we implement the testing procedure using the MIMIC-III data set in Section 5. In both sections, we repeat the testing and validation procedures four times, defining $\boldsymbol{\zeta}(\cdot)$ in terms of the baseline vector $\boldsymbol{\zeta}^B(\cdot)$, the last-value carried forward vector $\boldsymbol{\zeta}^L(\cdot)$, the average vector $\boldsymbol{\zeta}^A(\cdot)$, and the context vectors $\boldsymbol{\zeta}^C(\cdot)$, as described in Section 3.3. Note, $\hat{\mathcal{V}}^{\text{test}}(d^{\text{opt}})$, $\hat{\mathcal{V}}^{\text{test}}(d^{\text{obs}})$, and $\hat{\mathcal{V}}^{\text{test}}(d^{\text{no}})$ are dependent on the fitted Q-models, as is $\hat{\mathcal{V}}^{\text{val}}(d^{\text{opt}})$. Thus, these value estimates are dependent on the functional form of $\boldsymbol{\zeta}(\cdot)$. In contrast, $\hat{\mathcal{V}}^{\text{val}}(d^{\text{obs}})$ and $\hat{\mathcal{V}}^{\text{val}}(d^{\text{no}})$ are independent of the fitted Q-models, so these estimates do not depend on $\boldsymbol{\zeta}(\cdot)$. We evaluate the estimation performance of ReLiVE-Q by studying the results of the testing and validation procedures in Sections 4 and 5.

4. SIMULATION STUDY

4.1. Simulation strategy

We design a simulation study to evaluate the estimation performance of ReLiVE-Q. We provide a detailed description of the simulation strategy in Section 5 of the supplementary material. In summary, we conduct the testing and validation procedures described in Section 3.7 to obtain $N = 500$ value estimates for each studied regime using simulated data. We consider $K = 4$ decision points that occur at times $(\tau_1, \ldots, \tau_4) = (0, 3, 6, 9)$, and we generate a data set of $m = 10,000$ patients. For each patient, we generate a single treatment variable $A_k \in \mathcal{A}_k = \{0, 1\}$ at $k = 1, \ldots, 4$, and we generate a single covariate measured only at baseline, S. We consider ten measurement times $\mathcal{M} = (0, 1, \ldots, 9)$, and we generate three longitudinal covariates $Z_l(t) = B_l(t) + \epsilon_l(t)$ at each $t \in \mathcal{M}$, where $l = 1, 2, 3$. We generate censoring times C and impose a restricted lifetime of $L = 50$. We conduct two separate analyses, each using a distinct survival time generation process. First, we generate T according to an accelerated failure time (AFT) model. Second, we generate T according to a Cox proportional hazards model. We provide the code used to conduct the simulation study in the supplementary material.

4.2. Simulation results

The 500 testing value estimates $\hat{\mathcal{V}}^{\text{test}}(d)$ and validation value estimates $\hat{\mathcal{V}}^{\text{val}}(d)$ are plotted in Figures 1 and 2 for $d \in \{d^{\text{opt}}, d^{\text{obs}}, d^{\text{no}}\}$. In both the AFT and Cox analyses, the distributions of $\hat{\mathcal{V}}^{\text{test}}(d)$ and $\hat{\mathcal{V}}^{val}(d)$ are consistently higher for the optimal treatment regime than for the observed treatment regime or the no treatment regime. Thus, the simulation study supports that $\mathcal{V}(\hat{d}^{\text{opt}}) > \mathcal{V}(d^{\text{obs}})$ and $\mathcal{V}(\hat{d}^{\text{opt}}) > \mathcal{V}(d^{\text{no}})$, which allows us to conclude that ReLiVE-Q successfully produces reasonable estimates of an optimal treatment regime.

Moreover, in both the AFT and Cox analyses, Q-models fit with $\boldsymbol{\zeta}^B(\cdot)$ result in lower value estimates for the optimal treatment regime than those fit with $\boldsymbol{\zeta}^A(\cdot)$, $\boldsymbol{\zeta}^L(\cdot)$, or $\boldsymbol{\zeta}^C(\cdot)$. In the AFT analysis, Q-models fit with $\boldsymbol{\zeta}^L(\cdot)$ or $\boldsymbol{\zeta}^C(\cdot)$ result in the highest value estimates for the optimal treatment regime, while in the Cox analysis, Q-models fit with $\boldsymbol{\zeta}^C(\cdot)$ result in the highest. Thus, the simulation study supports that sophisticated functions are necessary to synthesize the trajectories of the longitudinal covariates, and that representing the longitudinal covariates with context vectors can lead to improved optimal treatment regime estimation via ReLiVE-Q.

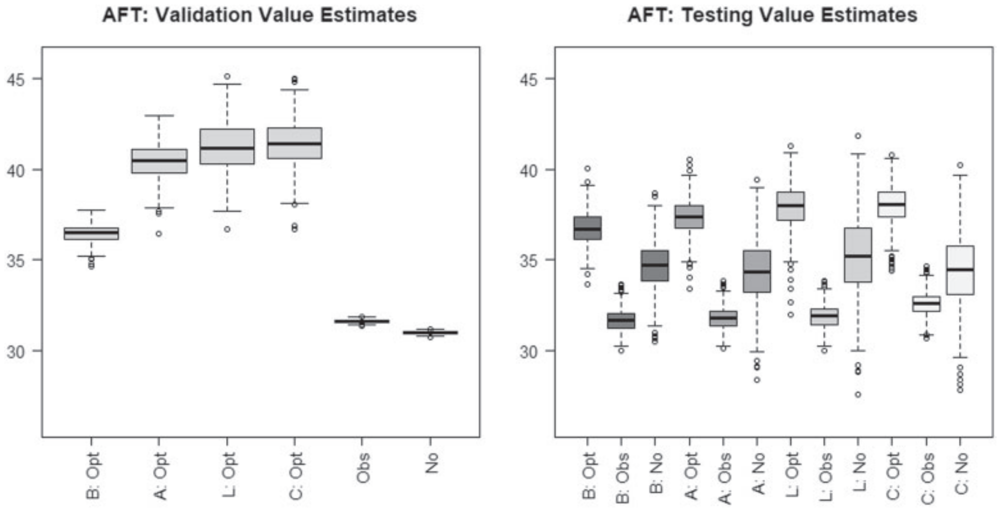

Figure 1. AFT simulation study: boxplots of the value estimates from the validation procedure (left) and the testing procedure (right) for an optimal treatment regime (Opt), the observed treatment regime (Obs), and the no treatment regime (No). For scenarios dependent on the Q-models, value estimates are presented using the baseline vector (B), the average vector (A), the last-value carried forward vector (L), and the context vector (C).

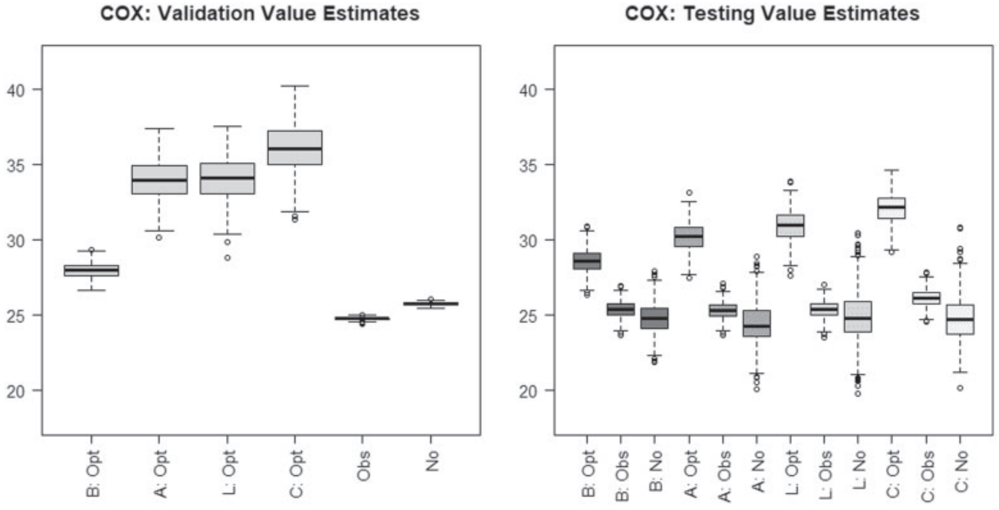

Figure 2. Cox simulation study: boxplots of the value estimates from the validation procedure (left) and the testing procedure (right) for an optimal treatment regime (Opt), the observed treatment regime (Obs), and the no treatment regime (No). For scenarios dependent on the Q-models, value estimates are presented using the baseline vector (B), the average vector (A), the last-value carried forward vector (L), and the context vector (C).

5. MIMIC-III DATA APPLICATION

5.1. MIMIC-III data description

Next, we evaluate the estimation performance of ReLiVE-Q by conducting the testing procedure described in Section 3.7 on the EMR data of septic patients in the MIMIC-III database. MIMIC-III is a freely available database comprised of deidentified medical records for over 40,000 patients who stayed in the critical care units at Beth Israel Deaconess Medical Center between 2001 and 2012 (Johnson et al., 2016). MIMIC-III contains data on patients' demographics, vital signs, laboratory measurements, medications, imaging reports, chart notes, procedure codes, diagnostic codes, hospital stay, and survival. For a complete description of the MIMIC-III database, refer to Johnson et al. (2016).

In 2016, the definitions and clinical criteria for sepsis and septic shock were updated in the *Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)* (Singer et al., 2016). Sepsis-3 defines sepsis as a "life-threatening organ dysfunction caused by a dysregulated host response to infection" and provides clinical criteria for diagnosing septic patients (Singer et al., 2016). Komorowski (2019) developed code to identify patients in MIMIC-III fulfilling the Sepsis-3 criteria. Komorowski's code pulls relevant physiological parameters for septic patients from up to 24 h preceding their sepsis diagnosis until 48 h after. The code aggregates the data into 4-h time windows, recording an appropriate summary statistic when several measurements were taken in the same time window. We use Komorowski's code to construct our studied data set of $m = 20,952$ patients, only 15% of whom have observed restricted survival times, i.e. $\Delta^L = 1$.

We estimate optimal treatment regimes for septic patients in the ICU using the MIMIC-III data set. We conduct the testing procedure described in Section 3.7 to obtain $N = 100$ value estimates for the studied regimes. Technical details are provided in Section 7 of the supplementary material. We consider $K = 10$ decision points that occur every 4 h at $(\tau_1, \tau_2, \dots, \tau_{10}) = (0, 4, \dots, 36)$. We study a single treatment $A_k \in \{0, 1\}$ at each decision point, $k = 1, \dots, 10$, where $A_k = 1$ if the patient is provided vasopressors in the 4-h time window $[\tau_k, \tau_k + 4)$ and $A_k = 0$ otherwise. We specify a restricted lifetime of $L = 40$, and we study a predictor space containing a single baseline covariate S and 13 longitudinal covariates $\mathbf{Z}(\cdot)$. The baseline covariate of interest is an indicator of whether the patient was previously admitted to the ICU during the given hospital stay. The longitudinal covariates of interest include a longitudinal indicator of mechanical ventilator dependence, as well as 12 longitudinal vital signs and laboratory values: albumin, calcium, magnesium, hemoglobin, arterial lactate, arterial pH, fraction of inspired oxygen (FiO_2), peripheral oxygen saturation (SpO_2), Sequential Organ Failure Assessment (SOFA) score, respiratory rate, heart rate, and systolic blood pressure.

5.2. MIMIC-III results

The 100 testing value estimates $\hat{\mathcal{V}}^{\text{test}}(d)$ are plotted in Fig. 3 for $d \in \{d^{\text{opt}}, d^{\text{obs}}, d^{\text{no}}\}$. The distribution of $\hat{\mathcal{V}}^{\text{test}}(d)$ is consistently higher for the optimal treatment regime than for the observed treatment regime or the no treatment regime. These results support that $\mathcal{V}(\hat{d}^{\text{opt}}) > \mathcal{V}(d^{\text{obs}})$ and $\mathcal{V}(\hat{d}^{\text{opt}}) > \mathcal{V}(d^{\text{no}})$. Thus, we can conclude that ReLiVE-Q successfully uses EMR data to produce reasonable optimal treatment regime estimates for septic patients in the ICU.

Again, Q-models fit with $\boldsymbol{\zeta}^B(\cdot)$ result in lower value estimates for the optimal treatment regime than those fit with $\boldsymbol{\zeta}^A(\cdot)$, $\boldsymbol{\zeta}^L(\cdot)$, or $\boldsymbol{\zeta}^C(\cdot)$. Moreover, Q-models fit with $\boldsymbol{\zeta}^C(\cdot)$ result in the highest value estimates for the optimal treatment regime. This application suggests that representing longitudinal covariates with context vectors can lead to improved optimal treatment regime estimation via ReLiVE-Q. In Section 8 of the supplementary material, we evaluate the importance of each studied covariate for accurately estimating the Q-functions.

6. DISCUSSION

In the simulation study and application to MIMIC-III, we demonstrate that the optimal treatment regime estimated via ReLiVE-Q results in higher estimates of value (2.1) than the no treatment and observed treatment regimes. Thus, we can expect cumulative restricted residual life to be higher on average for patients who follow the estimated optimal treatment regime, as compared to patients who always receive the standard treatment, and compared to patients who received the observed

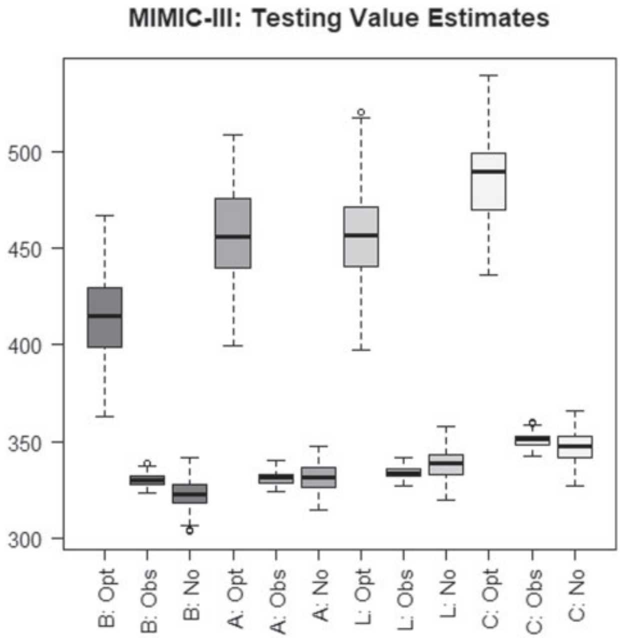

Figure 3. MIMIC-III application: boxplots of the value estimates from the testing procedure for an optimal treatment regime (Opt), the observed treatment regime (Obs), and the no treatment regime (No). Value estimates are presented from Q-models using the baseline vector (B), the average vector (A), the last-value carried forward vector (L), and the context vector (C).

treatments administered by clinicians. Thus, ReLiVE-Q leverages patient history to estimate personalized treatment regimes that maximize a clinically meaningful function of residual life. This finding is especially important for patients diagnosed with sepsis, as septic patients experience varying responses to treatment for the life-threatening condition. Moreover, we demonstrate that synthesizing longitudinal covariates with context vectors in ReLiVE-Q leads to improved optimal treatment regime estimation, as compared to using simpler summary statistics.

We limit our study to clinical settings with two treatment options that are feasible for all patients at all decision points, where decision points occur at fixed intervals. Extending our methods to accommodate a finite number of treatment options is straightforward. Our methods can be extended to settings where not all treatments are feasible for all patients under the feasible sets framework outlined in Tsiatis et al. (2020). By re-framing the decision point times as potential outcomes dependent on patient history, our methods can also be extended to accommodate subject-specific decision point times. See Sections 8.3.1–8.3.2 of Tsiatis et al. (2020) for an example of this approach. Further research is required to determine how ReLiVE-Q performs in these more complex clinical settings.

We assume censoring is independent of treatment assignment, patient characteristics, restricted lifetime, and potential outcomes, and we estimate the censoring weights via the Kaplan-Meier method. However, it may be possible to relax this assumption and take censoring to be noninformative in the sense that the cause-specific hazard of censoring as a function of patient history and potential outcomes does not depend on the potential outcomes. In this case, a time-dependent Cox model could be used to incorporate covariate information into the censoring weight estimates. See Section 8.3.2 of Tsiatis et al. (2020) for details.

In this article, we focus on accurately estimating optimal treatment regimes. Accordingly, we model the Q-functions using flexible, nonparametric random forests. A drawback to modeling the

Q-functions using random forests is that the estimated optimal treatment rules are not readily interpretable. ReLiVE-Q could be implemented with simpler Q-models to improve decision rule interpretability, though further research is required to determine how this would affect estimation.

ReLiVE-Q defines the reward at each decision point to be residual life. This quantity is more interpretable than the rewards used in the alternative Q-learning methods described in Section 1, defined as the incremental amount of time between the given decision point and the next decision point or failure. However, ReLiVE-Q's improvement in reward interpretability comes with the trade-off of a less interpretable value function. ReLiVE-Q defines the value of regime $d \in \mathcal{D}$ to be the expected value of cumulative restricted residual life under d summed across all decision points reached by the patient, while the Q-learning methods discussed in Section 1 more simply define the value of regime d to be the expected value of restricted residual life under d.

ACKNOWLEDGMENTS

The authors would like to thank Han Wang at North Carolina State University. Han Wang constructed the studied MIMIC-III data set.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Biostatistics Journal* online.

FUNDING

This project was supported by the National Institutes of Health grant T32 HL079896.

Conflict of Interest: None declared.

DATA AVAILABILITY

The simulation code is available at https://github.com/gmrhodes/ReLiVE-Q, and the associated output is available at https://figshare.com/collections/RL_Q-Learning/6340298.

REFERENCES

Bai, X., Tsiatis, A. A., Lu, W. and Song, R. (2017). Optimal treatment regimes for survival endpoints using a locally-efficient doubly-robust estimator from a classification perspective. Lifetime Data Anal 23(4), 585–604.

Bellman, R. (1957). Dynamic programming. Princeton, NJ: Princeton University Press.

Breiman, L. (2001). Random forests. Mach Learn 45, 5–32.

Cho, H., Holloway, S. T., Couper, D. J. and Kosorok, M. R. (2022). Multi-stage optimal dynamic treatment regimes for survival outcomes with dependent censoring. Biometrika 00(0), 1–16.

Choi, T., Lee, H. and Choi, S. (2023). Accountable survival contrast-learning for optimal dynamic treatment regimes. Sci Rep. 13, 2250.

Cui, Y., Zhu, R. and Kosorok, M. (2017). Tree based weighted learning for estimating individualized treatment rules with censored data. Electron J Stat 11(2), 3927–3953.

Goldberg, Y. and Kosorok, M. R. (2012). Q-learning with censored data. Ann Stat 40(1), 529–560.

Hager, R., Tsiatis, A. A. and Davidian, M. (2018). Optimal two-stage dynamic treatment regimes from a classification perspective with censored survival data. Biometrics 74(4), 1180–1192.

Huang, X., Ning, J. and Wahed, A. S. (2014). Optimization of individualized dynamic treatment regimes for recurrent diseases. Stat Med 33(14), 2363–2378.

Illenberger, N., Spieker, A. J. and Mitra, N. (2023). Identifying optimally cost-effective dynamic treatment regimes with a Q-learning approach. J R Stat Soc Ser C 72(2), 434–449.

Jiang, R., Lu, W., Song, R. and Davidian, M. (2017 a). On estimation of optimal treatment regimes for maximizing t-year survival probability. J R Stat Soc Ser B 79(4), 1165–1185.

Jiang, R., Lu, W., Song, R., Hudgens, M. G. and Naprvavnik, S. (2017b). Doubly robust estimation of optimal treatment regimes for survival data—with application to an HIV/AIDS study. Ann Appl Stat 11(3), 1763–1786.

Johnson, A. E. W, Pollard, T. J., Shen, L., Lehman, L. H, Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A. and Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. Sci Data 3. 160035.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. J Am Stat Assoc 53(282), 457–481.

Komorowski, M. (2019). AI Clinician. https://github.com/matthieukomorowski/AI_Clinician.

László, I., Trásy, D., Molnár, Z. and Fazakas J. (2015). Sepsis: from pathophysiology to individualized patient care. J Immunol Res. 2015:510436.

Lyu, L., Cheng, Y. and Wahed, A. S. (2023). Imputation-based Q-learning for optimizing dynamic treatment regimes with right-censored survival outcome. Biometrics **79**, 3676–3689.

Murphy, S. A. (2003). Optimal dynamic treatment regimes. J R Stat Soc Ser B. 65(2), 331–355.

Murphy, S. A. (2005). A generalization error for Q-learning. J Mach Learn Res. 6, 1073–1097.

Rhodes, G., Davidian, M. and Lu, W. (2023). Dynamic prediction of residual life with longitudinal covariates using long short-term memory networks. Ann Appl Stat. 17(3), 2039–2058.

Robins, J. M. Optimal structural nested models for optimal sequential decisions. In: Lin DY, Heagerty PJ, editors. Proceedings of the Second Seattle Symposium in Biostatistics: Analysis of Correlated Data, Volume 179, Lecture Notes in Statistics. New York, NY: Springer; 2004. p. 189–326.

Simoneau, G., Moodie, E. E. M., Nijjar, J. S. and Platt, R. W.; and the Scottish Early Rheumatoid Arthritis Inception Cohort Investigators. (2020). Estimating optimal dynamic treatment regimes with survival outcomes. J Am Stat Assoc. 115(531), 1531–1539.

Singer, M., Deutschman, C. S., Seymour, C. W., Shankar-Hari, M., Annane, D., Bauer, M., Bellomo, R., Bernard, G. R., Chiche, J.-D., Coopersmith, C. M., Hotchkiss, R. S., Levy, M. M., Marshall, J. C., Martin, G. S., Opal, S. M., Rubenfeld, G. D., van der Poll, T. and Vincent, J.-L. et al. (2016). The third international consensus definitions for sepsis and septic shock (sepsis-3). J Am Med Assoc 315(8), 801–810.

Tsiatis, A. A., Davidian, M., Holloway, S. T. and Laber, E. B. (2020). Dynamic treatment regimes: statistical methods for precision medicine. 1st ed. Boca Raton, FL: Chapman and Hall/CRC.

Wang, J., Zeng, D. and Lin, D. Y. (2022). Semiparametric single-index models for optimal treatment regimens with censored outcomes. Lifetime Data Anal. 28, 744–763.

Watkins, C. J. C. H. and Dayan, P. (1992). Q-learning. Mach Learn. 8, 279–292.

Xue, F., Zhang, Y., Zhou, W., Fu, H. and Qu, A. (2022). Multicategory angle-based learning for estimating optimal dynamic treatment regimes with censored data. J Am Stat Assoc 117(539), 1438–1451.

Zhang, Z., Yi, D. and Fan, Y. (2022). Doubly robust estimation of optimal dynamic treatment regimes with multicategory treatments and survival outcomes. Stat Med. 41(24), 4745–4960.

Zhao, Y., Zeng, D., Socinski, M. A. and Kosorok, M. R. (2011). Reinforcement learning strategies for clinical trials in non-small cell lung cancer. Biometrics 67(4), 1422–1433.

Zhao, Y. Q., Zeng, D., Laber, E. B., Song, R., Yuan, M. and Kosorok, M. R. (2015). Doubly robust learning for estimating individualized treatment with censored data. Biometrika 102(1), 151–168.

Zhao, Y.-Q., Zhu, R., Chen, G. and Zheng, Y. (2020). Constructing dynamic treatment regimes with shared parameters for censored data. Stat Med 39(9), 1237–1413.

Zhou, J., Zhang, J., Lu W. and Li, X. (2021). On restricted optimal treatment regime estimation for competing risks data. Biostatistics 22(2), 217–232.