Published in final edited form as:

Biom J. 2019 March ; 61(2): 442–453. doi:10.1002/bimj.201700181.

A cure-rate model for Q-learning: Estimating an adaptive immunosuppressant treatment strategy for allogeneic hematopoietic cell transplant patients

Erica E.M. Moodie^{1,*}, David A. Stephens², Shomoita Alam¹, Mei-Jie Zhang³, Brent Logan³, Mukta Arora⁴, Stephen Spellman⁵, and Elizabeth F. Krakow⁶

¹Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, QC H3A 1A2 Canada

²Department of Mathematics and Statistics, McGill University, Montreal, QC H3A 1A2 Canada

³Medical College of Wisconsin, Milwaukee, WI 53226 USA

⁴Department of Medicine, University of Minnesota, Minneapolis, MN 55455 USA

⁵Center for International Blood and Marrow Transplant Research, Minneapolis, MN 55401 USA

⁶Fred Hutchinson Cancer Research Center, Seattle, WA 98109 USA

Abstract

Cancers treated by transplantation are often curative, but immunosuppressive drugs are required to prevent and (if needed) to treat graft-versus-host disease. Estimation of an optimal adaptive treatment strategy when treatment at either one of two stages of treatment may lead to a cure has not yet been considered. Using a sample of 9,563 patients treated for blood and bone cancers by allogeneic hematopoietic cell transplantation drawn from the Center for Blood and Marrow Transplant Research database, we provide a case study of a novel approach to Q-learning for survival data in the presence of a potentially curative treatment, and demonstrate the results differ substantially from an implementation of Q-learning that fails to account for the cure-rate.

Keywords

Adaptive treatment strategy; Cure-rate; Dynamic treatment regime; Survival data; Q-learning

1 Introduction

Certain cancers of the blood and bone are often treated by allogeneic hematopoietic cell transplantation (AHCT). As part of the AHCT, immunosuppressive drugs are often administered sequentially to prevent and (if needed) to treat graft-versus-host disease (GVHD), a condition in which donor cells attack the recipient patient's normal tissues.

*Corresponding author: erica.moodie@mcgill.ca, Phone: +1 (514) 398-5520, Fax: +1 (514) 398-4503.

Conflict of Interest

The authors have declared no conflict of interest.

Immunosuppressants differ in which T-cell subsets they target and in the extent to which they lower the total T-cell count. For example, some drugs are known as *non-specific highly T-lymphodepleting therapies* (NHTL) such as antithymocyte globulin and alemtuzumab, may reduce the risk of GVHD [25, 31] but may also carry increased risk of opportunistic infections [4, 5, 3, 22] and, in some contexts, increased risk of cancer relapse compared to non-NHTL immunosuppressants like cyclosporine [19]. The delayed effects of different immunosuppressant regimens, their potential interactions, and their appropriateness given complex and evolving characteristics of individual patients remain poorly characterized. Moreover, because of limited financial resources, logistic challenges, and heterogeneity (yet relative scarcity) of patients who develop the most severe post-transplant complications there is a dearth of large randomized trials to guide practice that is tailored to individuals. The clinical question of how best to prevent and treat acute GVHD must therefore be answered by other means.

Adaptive treatment strategies (ATS, also called dynamic treatment regimes) are statistical approaches that operationalize the sequential decision-making inherent in adaptive clinical practice, sometimes also called “precision” or “personalized” medicine. An ATS offers a list of decision rules for how to modify treatment over time. These rules are based on a patient’s individual baseline characteristics and on how particular covariates (e.g. clinical characteristics such as lymphocyte count or grade of GVHD) vary after each specific treatment is administered. To date, much of the ATS literature has focused on the case of continuous outcomes [18, 21, 6, 32]. Though there has been some work in extending Q-learning to the survival setting [7, 11, 12, 13], most of the ATS literature for time-to-event data has focused on inverse weighted or classification based estimators [29, 10, 30, 24, 34], which may suffer from low statistical efficiency relative to approaches that directly model the survival time, and no method has specifically focused on the setting where treatment may be curative despite curative treatments being common in many fields, including oncology [8, 9]. We address this latter limitation, extending Q-learning to accommodate cure-rate models and apply this to the setting of prevention and treatment of GVHD following allogeneic hematopoietic cell transplantation.

In this article, we provide a case study of a novel approach to Q-learning for survival data in the presence of a potentially curative treatment. Additionally, we demonstrate that a simpler modelling choice that ignores the cure fraction is unable to uncover small but potentially important subgroups of individuals who might benefit from a particular class of immunosuppressants that are not beneficial for the population as a whole.

2 Methods: A cure-rate Q-learning model

Although ATS construction has been carried out for multiple treatment stages, we focus here on the two-stage case as we are motivated by the question of whether to (1) recommend NHTL or non-NHTL immunosuppressive therapies to prevent GVHD and (2) whether, for those who develop acute, steroid-refractory GVHD, to recommend NHTL or non-NHTL immunosuppressive therapies to treat the GVHD. Our notation is as follows, with $j=1,2$:

- Treatments, A_j , are taken over two stages, $A_j \in \mathcal{A}_j \equiv \{0, 1\}$ where 1 indicates NHTLs. Stage 1 ends with either (i) the occurrence of the event of interest (in our motivating example, it will be death or disease progression), (ii) transition to the second stage (triggered by a clinical event requiring second-stage treatment), or (iii) censoring.
- Pre-treatment covariates, X_j , are vector-valued.
- Let T_1 denote the time spent in the first stage, which is defined as the time from origin until either death, relapse, or initiation of the second-stage treatment.
- Let C_1 denote the censoring time of any individual lost to follow-up prior to observing any of death, disease progression, or initiation of the second stage treatment.
- Let $Y_1 = \min(T_1, C_1)$, and let δ_1 be an indicator of whether the event of interest occurred in the first interval.
- The indicator for a patient requiring second-stage treatment (GVHD salvage therapy) is η .
- For patients who do enter stage 2, let T_2 denote the time from salvage therapy until either death or disease progression, and C_2 denote the censoring time of any individual lost to follow-up after initiating salvage therapy. Then $Y_2 = \min(T_2, C_2)$, and $\delta_2 = 1[Y_2 = T_2]$ is an indicator of whether the event of interest occurred in the second interval.
- The total observation time for any individual is $Y = Y_1 + \eta Y_2$.

Therefore, the data in a two-interval scenario are $(X_1, A_1, Y_1, \delta_1, \eta X_2, \eta A_2, \eta Y_2, \eta \delta_2, Y)$, and an adaptive treatment strategy is denoted $d_j = d_j(h_j) \in \mathcal{A}_j$. Where useful for concision and clarity, we will denote the covariate and treatment *history* at the start of interval as $H_1 = X_1$, $H_2 = (X_1, A_1, X_2)$, where H_2 will be used and defined only for those individuals for whom second-stage treatment is required (i.e. for whom $\eta = 1$). Note that our data setting can be viewed as a progressive illness-death model, where patients can transition from “health” (transplant receipt) to either “illness” (GVHD requiring salvage therapy) or death/relapse [2, p.28]. From this viewpoint, T_1 would be the sojourn time in the “health” state, and T_2 the occupation time of the “illness” state.

Throughout, we shall be working within the potential outcomes framework, where a potential outcome under treatment pattern a is the outcome that would be observed if treatments a were in fact given (whatever the actual observed treatment); potential outcomes are denoted $Y(a)$. We require the following assumptions: (1) Consistency: the potential outcome under any particular treatment sequence corresponds to the observed outcome if that treatment sequence were followed. (2) No unmeasured confounding: at each stage, treatment assignment is independent of a patient’s potential outcomes conditional on the observed treatment history and covariates. That is, $Y_1(a_1) \perp A_1/H_1$ and $Y_2(a_1, a_2) \perp A_2/H_2, \eta = 1$. (3) Subjects’ outcomes depend only on the treatments they received but not on treatments received by others, known as the no interference assumption. (4) Censoring is

independent and non-informative. These are the standard assumptions made in the causal inference setting.

In Q-learning, there are two key models that must be specified for each stage of treatment: (i) the contrast model (also called a blip, or structural nested mean model), which describes how the receipt of A_j affects outcome (both its main effect and any interactions with other covariates), and (ii) the main effects of all other covariates on outcome. The blip model is parameterized with ψ , the treatment-free model with β . Taken together, these make up what are called the Q-functions:

$$Q_2(h_2, a_2) = E[T | H_2 = h_2, A_2 = a_2],$$

$$Q_1(h_1) = E \left[\max_{a_2} Q_2(H_2, a_2) \middle| H_1 = h_1, A_1 = a_1 \right].$$

Q-learning may be viewed through the lens of potential outcomes. Suppose, for the sake of exposition, that there is no censoring so that $Y_j = T_j$ for $j = 1, 2$. Let $Y_1(a_1)$ denote the time to the end of the first stage if treatment $A_1 = a_1$ was given. Similarly, for those patients who develop GVHD that requires salvage therapy, let $T_2(a_1, a_2)$ denote the time from salvage therapy to disease recurrence or death that would be observed under prophylactic treatment a_1 and GVHD salvage treatment a_2 . Note, then, that $\max_{a_2} Q_2(h_2, a_2)$ corresponds to the expected counterfactual outcome under optimal second-stage treatment, i.e. the expected value of $T(a_1, a_2^{opt})$ where it should be noted that a_2^{opt} is a function of baseline and intermediate covariates as well as prior treatment so that $T(a_1, a_2^{opt})$ is shorthand for $T(a_1, a_2^{opt}(x_1, x_2, x_3))$. This is often referred to as a pseudo-outcome, and is denoted $\tilde{T}$. The optimal ATS is then given by $d_j(h_j) = \arg \max_{a_j \in \mathcal{A}_j} Q_j(h_j, a_j)$, $j = 1, 2$, which is a function only of the contrast model, provided the mean models take a (generalized) linear form with a monotone link function.

Q-learning proceeds according to the following general sequential algorithm for a two-stage example:

1. Propose a (typically parametric) Q-function for the second stage of treatment, say $Q_2(h_2, a_2; \beta_2, \psi_2)$, where h_2^β and h_2^ψ are two vector summaries of h_2 , such that h_2^β contains all covariates having a ‘main effect’ on the outcome and h_2^ψ contains all covariates that interact with treatment so as to be useful for tailoring treatment recommendations. The vectors h_2^ψ and h_2^β also contain a vector of 1s so as to ensure both an intercept and a main effect of treatment are included in the expected outcome model. For example, in a continuous outcome setting, a typical Q-function would take the form $Q_2(h_2, a_2; \beta_2, \psi_2) = h_2^\beta \beta_2 + a_2 h_2^\psi \psi_2$. Estimate β_2, ψ_2 (using some method appropriate to the outcome type, e.g. ordinary least squares for a continuous outcome), obtaining values $\hat{\beta}_2, \hat{\psi}_2$.

2. Set the stage-1 pseudo-outcome to $\tilde{y} = \max_{a_2} Q_2(h_2, a_2; \hat{\beta}, \hat{\psi})$
3. Propose a (typically parametric) Q-function for the first stage of treatment, say $Q_1(h_1, a_1; \beta_1, \psi_1)$, where h_1^β and h_1^ψ are two vector summaries of h_1 consisting, as in step 1, of covariates having a main effect and a treatment-interaction, respectively. Estimate β_1, ψ_1 .
4. Estimate the optimal ATS:

$$\hat{d}_j(h_j) = \arg \max_{a_j \in A_j} Q_j(h_j, a_j; \hat{\beta}_j, \hat{\psi}_j), j = 1, 2.$$

As noted above, the pseudo-outcome computed in step 2 may be interpreted as the outcome that would be seen for each individual had they received their actual observed stage 1 treatment followed by their optimal (dynamic) second-stage treatment. The backwards induction algorithm described above is generic. In its typical implementation, the outcome Y is continuous and ordinary least squares is used to estimate β_j, ψ_j in steps 1 and 3; however, the procedure has also been adapted to other outcome types [17, 14].

To adapt this general procedure to the specific setting of a time-to-event outcome where a fraction of the population is cured and individuals may be censored, steps 1–3 must be altered. One approach would be to model not just the expectation of the (pseudo)outcome, but an entire distribution so that both the distribution and the corresponding survival function can be modelled. Nevertheless, the same principle of recursive estimation applies. While it is possible to use semi-parametric models in steps 1 and 3, the prediction needed to generate the pseudo-outcome in step 2 is such that parametric models are often preferable as they provide more accurate predictions when the modelling assumptions are correct. Q-learning being a singly robust method, it is the case that a badly specified parametric model can lead to biased parameter estimates and poor ATSs; model diagnostics should always be employed.

Let $\tilde{f}(t|X)$ be the population distribution of event times. We propose the use of a mixture model to capture the event process at each stage of analysis, with Q-functions following from the assumption that the event time distribution takes the form

$$\tilde{f}(t|X = x) = \pi(x)1_\infty(t) + (1 - \pi(x))(1 - 1_\infty(t))f(t|x) \quad (1)$$

where $\pi(x)$ is the probability of being cured, $1_\infty(t)$ is a point mass at infinity representing the disease-free survival time in the cured fraction of the population, $f(t|x)$ is the density function of disease-free survival times for the uncured fraction of the population, and X are covariates used to model the probability of cure and of the distribution of event times amongst those who are not cured. In fact, the covariates used to model the cure probability and (disease-free) survival distribution need not be the same. We shall model $f(t|X)$ parametrically with $X = H_j, A_j$ at each stage, and further posit a logistic regression for the

cure probability such that $\pi(X) = \{1 + \exp(-\theta^T X)\}^{-1}$. The density $\tilde{f}$ can incorporate censored event times for those who aren't cured in the standard way, replacing $f(t/X)$ with $f(t/X)^\delta S(t/X)^{(1-\delta)}$ where $S(\cdot/X)$ is the survival function corresponding to density $f(\cdot/X)$.

With the outcome model characterized by Equation (1), we can adapt the typical Q-learning algorithm as follows:

1. Propose a parametric cure-rate model, $\tilde{f}$, for the second-stage outcome Y_2 of the form described in Equation (1), such that we can derive a Q-function as the its expectation, say $Q_2(h_2, a_2; \theta_2, \zeta_2)$, where ζ_2 parameterizes the uncured event-time distribution, $f(t_2/h_2, a_2)$, and contains the outcome model parameters (β_2, ψ_2) described above as well as any other necessary distributional parameters (e.g. a shape parameter). Estimate θ_2, ζ_2 by maximum likelihood.
2. In a time-to-event context, the pseudo-outcome is the sum of the observed time in the first stage plus the longest possible expected survival time spent in the second:

- (a) For each individual and each treatment option, estimate the optimal predicted stage 2 outcome, $\tilde{t}_2$, as the maximum expected survival time, marginalizing over each individual's probability of being cured:

$$\tilde{t}_2 = \max_{a_2 \in \mathcal{A}_2} \left\{ \pi(h_2, a_2; \hat{\theta}_2) r + (1 - \pi(h_2, a_2; \hat{\theta}_2)) E_f(T | H_2 = h_2, A_2 = a_2; \hat{\zeta}_2) \right\},$$

where r is the individual's residual life expectancy and $E_f(T | H_2 = h_2, A_2 = a_2; \hat{\zeta}_2)$ is the expected survival time calculated using the distribution $f(t_2, | h_2, a_2; \hat{\zeta}_2)$.

- (b) Calculate the stage-1 pseudo-outcome as $\tilde{y}_1 = y_1 + \eta \tilde{t}_2$ i.e. the sum of the observed stage 1 outcome and the optimal predicted stage 2 outcome. Note that $\tilde{y}_1 = y_1$ whenever $\eta = 0$, and $\tilde{y}_1 = y_1 + \tilde{t}_2$ otherwise.
3. Propose a parametric cure-rate model for the first-stage pseudo-outcome $\tilde{Y}_1$ of the form described in Equation (1), such that we can derive a Q-function as the its expectation, say $Q_1(h_1, a_1; \theta_1, \zeta_1)$. Estimate θ_1, ζ_1 by maximum likelihood.
4. Estimate the optimal DTR:

$$\hat{d}_j(h_j) = \arg \max_{a_j} Q_j(h_j, a_j; \hat{\theta}_j, \hat{\zeta}_j); j = 1, 2.$$

The models for the Q-functions in steps 1 and 4 are not restricted to follow the same parametric form; standard model checking and selection methods may be employed to assess model fit.

In some cases, it may be possible to estimate the residual life expectancy in step 2(a) from the available data. In many cases, including our case study, this is not the case as follow-up time is insufficient to accurately model the expected outcome in the cured portion of the population. In such cases, external information from, say, the World Health Organization may be leveraged to provide an estimate of the expected age- and sex-adjusted lifetime remaining. Further, if the residual life expectancy is much greater than the typical event times observed in the data, one may wish to redefine the outcome to be a reward or utility that is not the event time itself, but rather some function of that event time by truncation since the residual life expectancy based on a healthy population may be overly optimistic or because of a desire to ensure follow-up time remains consistent in the sense of having a maximum observation time (“end of study”) across the study sample both for observed and pseudo-outcomes.

3 Analysis of the CIBMTR data

3.1 The CIBMTR Data

The Center for Blood and Marrow Transplant Research (CIBMTR) consists of 450 health care institutions worldwide that report longitudinal data on hematopoietic stem cell transplants. The CIBMTR database has records on more than 425,000 transplant recipients. The CIBMTR collects information on variables such as age and sex of the transplant recipient, disease type, date of diagnosis, pre-transplantation disease stage, graft source, treatment regimen, blood lab work, development of GVHD, disease progression and survival, secondary malignancies, and cause of death. Data are collected at specified time points: before transplantation, then 100 days, 6 months and annually after transplantation or until death.

3.1.1 Identification and selection of eligible patients—Inclusion criteria for our analysis were: allogeneic stem cell transplant; all ages; both sexes; year of transplant between 1995–2007 inclusive; related or unrelated donor; any degree of human leukocyte antigen (HLA) match; any graft source; any conditioning regimen; and a diagnosis of acute myeloid leukemia or myelodys-plastic syndrome as the indication for transplant. The only exclusion criteria were having an ex-vivo T-cell depleted or CD34-selected graft or having a syngeneic transplant. Using these criteria, the CIBMTR identified 11,141 patients who had been retrieved for a study on chronic GVHD [1], of whom 9,563 met our eligibility criteria.

3.1.2 Definition of outcome and treatments at the two decision stages—The event time was measured from the time of graft infusion (day 0), and the event of interest is the composite outcome of cancer recurrence (or persistence) or death. We define the stage 1 outcome T_1 to be the time from graft infusion to the event of interest or GVHD requiring salvage therapy (progression to stage 2), and may be subject to censoring. T_2 is defined as the time from GVHD salvage therapy to death or cancer persistence or recurrence, and is also subject to censoring. The final outcome was thus taken to be the disease-free survival time, defined as survival time in months after day 0 without confirmation of disease persistence, relapse, or death.

The CIBMTR data contain more than 14 classes of agents used for GVHD prophylaxis and/or systemic treatment of acute GVHD; therefore, we simplified the treatments to (a) non-specific, highly T-lymphodepleting (NHTL) therapeutics vs. (b) the ensemble of non-NHTL therapies, or what we shall refer to as “standard” treatment. For this analysis, first-line therapy for acute GVHD was considered to be any 2 systemic therapies with any number of topical treatments; a continuation of a prophylaxis regimen with any number of systemic drugs without increasing their doses; or any number of topical therapies. Salvage treatment was defined as at least 3 systemic therapies, at least one of which was a systemic corticosteroid, and at least one of which was newly prescribed and not simply continued from prophylaxis. Because there was little variability in first-line GVHD treatment (95% received corticosteroids and 85% received calcineurin inhibitors) and because of the clinical uncertainty in how to treat steroid-refractory GVHD, we focused on salvage as the relevant second stage decision. That is, our aim is to construct an ATS which recommends NHTL or standard therapy prior to transplant, and NHTL or standard therapy as salvage therapy for steroid-refractory GVHD, where decisions at each stage depend on current patient information.

3.1.3 Potential tailoring variables—We considered three classes of variables as potentially valuable for tailoring GVHD prophylaxis and prevention: patient- and disease-specific characteristics, donor characteristics, and treatment-related factors. At the prophylaxis stage, the specific variables are age; Karnofsky/Lansky performance status; disease status (early, intermediate, advanced); cytomegalovirus status, donor relatedness, HLA match (well, partial, mis-matched), graft source; conditioning intensity (myeloblastic vs. not), total-body irradiation (yes/no), and other components of prophylaxis (methotrexate, mycophenolate, or corticosteroids). At the post-GVHD salvage treatment stage, the tailoring variables that we considered were all those considered for the prophylaxis stage, as well as grade of GVHD, time from transplant to GVHD, and the use of 4 or more systemic GVHD treatments. Note that the Karnofsky/Lansky performance status was recorded only at the time of transplant and not at the time of diagnosis of steroid-refractory GVHD. Nevertheless, we use this variable as a potential tailoring variable at both stages of the analysis.

Tailoring variables, denoted h_j^w above, are also likely to act as confounding variables, and so adjustment for confounding effects of all of the above-mentioned variables was performed in the analytic models so that $H_j^w \subset H_j^\beta$; the full list of covariates can be found in Table 1 with summary statistics given in Table 1 of the Online Supplementary Materials.

All variables included in both the cure-fraction and the survival models were chosen based on substantive considerations and subject matter knowledge. No variable selection was undertaken in this analysis.

3.2 Analysis: Q-functions, the pseudo-outcome, and missing data

Exploratory analyses suggested the Pareto model fit the CIBMTR data well (see Figure 1; a Weibull model was also explored but found inferior both graphically and via Akaike’s information criterion). Thus, Pareto models were fit at both the first and second stage of the

analysis. The cure probability and the Pareto mean (scale parameters) were modelled as a generalized linear function in which all tailoring and confounding variables were included as main effects and interactions between the treatment and the tailoring variables were also included. Censoring was assumed to be non-informative. All parameters were estimated via maximum likelihood in R version 3.3.2 (see Online Supplementary Materials).

Specifically, we estimate the parameters in the following cure-rate Pareto likelihood:

$$\left\{ \pi(h_j^\theta) \mathbb{1}_\infty(y) + (1 - \pi(h_j^\theta))(1 - \mathbb{1}_\infty(y)) f(y_j | h_j^\beta, h_j^\psi, \beta_j, \psi_j, \alpha_j) \right\}^{\delta_j} \times \left\{ \pi(h_j^\theta) \mathbb{1}_\infty(y) + (1 - \pi(h_j^\theta))(1 - \mathbb{1}_\infty(y)) [1 - F(y_j | h_j^\beta, h_j^\psi, \beta_j, \psi_j, \alpha_j)] \right\}^{1 - \delta_j} \quad (2)$$

for intervals $j = 1, 2$, thus adapting Equation (1) to accommodate censoring. In the above likelihood, we assume the following model specifications:

- $\pi(h_j^\theta) = 1/(1 + \exp(-h_j^\theta))$ is the probability of being cured in the j th stage, modelled as a function of covariates h_j^θ ;
- $f(y_j | h_j^\beta, h_j^\psi; \beta_j, \psi_j, \alpha_j)$ is a Pareto probability density function with shape α_j and scale λ_j , where $\log \lambda_j = (h_j^\beta \beta_j + a_j h_j^\psi \psi_j)$ that is

$$f(y_j | h_j^\beta, h_j^\psi; \beta_j, \psi_j, \alpha_j) = \frac{\alpha_j \lambda_j^{\alpha_j}}{(\lambda_j + y)^{\alpha_j + 1}}$$

with corresponding cumulative distribution function $F(y_j | h_j^\beta, h_j^\psi; \beta_j, \psi_j, \alpha_j)$ - this model also allows the Pareto shape parameter to differ by treatment group A_j ; and

- covariate vectors $h_j^\theta, h_j^\beta, h_j^\psi$ as given in Table 1.

To compute the pseudo-outcome in step 2 of the Q-learning algorithm that has been modified to accommodate a cured fraction as outlined in the previous section, residual life expectancy was required for each participant who entered the second stage of treatment. This was calculated based on age- and sex-adjust life-tables, with life-expectancy reduced by 30% [15] as AHCT patients who survive treatment known to have life expectancies that are reduced relative to healthy individuals. This residual life expectancy was then truncated at 10 years, to avoid ‘imputing’ remaining lifetime for individuals that persisted well beyond the range of follow-up in the observed data.

Missing data were addressed by multiple imputation via chained equations [26]. Default settings were used in the mice package in R, so that binary variables were imputed using logistic regression and all other variable types were imputed using predictive mean matching, with all covariates and the outcome being used to predict that which is missing in

turn. Due to the computationally intensive nature of the analysis and the large sample size, five imputations were used. Estimates (parameter coefficients) from both the cure and survival components of the Q-function across the five imputed datasets were combined by a simple average as per standard multiple imputation procedures (“Rubin’s rules”). Measures of uncertainty including standard errors and confidence intervals were computed using a non-parametric bootstrap with 300 resamples. The bootstrap procedure fully accounted for all steps in the estimation procedure, including the imputation and the plug-in of second-stage estimates into the pseudo-outcome used for the first-stage estimation.

3.3 Results

This case study was motivated by the evident asymptote indicating a cured fraction of individuals in Kaplan-Meier plots for disease-free survival time in the CIBMTR data (Figure 1). In the Kaplan-Meier plots, there appeared to be little difference in the disease-free survival times under the two prophylactic treatments, while at the second stage, salvage treatment with NHTL therapy was associated with a much lower fraction of cured individuals. However, confounding was not taken into account in these plots.

We jointly estimated the probability of being cured and the disease-free survival distribution assuming Pareto models using Equation (2). As it is only those variables that interact with treatment that inform treatment tailoring, we present the relevant cure log odds ratios and Pareto scale parameter estimates in 2, along with 95% confidence intervals calculated by bootstrap percentiles for both intervals’ Q-function parameters. We observe a large coefficient on the interaction between NHTL and disease stage (which is statistically different from zero at the second stage; Table 2), indicating that the odds of a cure are 1.35 times higher for patient with intermediate or advanced disease if treated with NHTL prophylaxis relative to standard treatment, and 1.09 times higher for patient with advanced disease if treated with NHTL salvage therapy relative to standard therapeutic drugs.

The interpretation of the accelerated failure time model parameters follows along standard lines. For example, for a patient at the reference level of all covariates, we see that NHTL prophylaxis alters expected life by a factor of $\exp(-0.097) = 0.908$ [95% CI: 0.952–1.315] if not cured, indicating a non-significant decrease in expected disease-free survival of approximately 10%. Similarly, from these results, it can be seen that transplant recipients with an HLA partial match or mismatch, or with intermediate or advanced disease status, may benefit from NHTL therapies at either stage as the sum of the main effect of treatment along with the associated covariate are positive, indicating an increase in the expected disease-free survival time if not cured if given NHTL treatments. For example, a patient with advanced disease status would expect to see an increase in expected disease-free survival of approximately 15–16% (as $\exp(-0.097+0.242) = 1.156$) if treated with NHTL.

In Figure 2, we present results indicating the difference in the probability of being cured if given NHTL versus standard treatment (top row), the difference in expected disease-free survival if given NHTL versus standard treatment conditional on a patient not being cured (middle row), and the difference in expected disease-free survival if given NHTL versus standard treatment, marginalized over the probability of a patient being cured (bottom row). In these figures, patients whose point estimates favour NHTLs are indicated by black

crosses; note, however, that the majority of the 95% confidence intervals (constructed by pointwise bootstrap percentile intervals) include zero. The exception is at the first stage, where the models suggest that 13.5% of individuals would be significantly more likely to experience a cure if they received prophylactic NHTL treatment (and a further 49.1% show benefit that is not significantly different from 0). Because of this increase in the probability of being cured, and the relatively small loss in disease-free survival time if not cured, we can see from the bottom left-hand panel that 38.5% of individuals are predicted to have improved outcomes if receiving NHTLs at the first stage, whereas a much more modest fraction – only 5.3% – might benefit from NHTL salvage therapy. However improvements in both the probability of being cured and the expected disease-free survival at the second stage if given NHTLs do not differ from significantly from zero, suggesting that the cautious strategy would be to avoid NHTLs for salvage treatment.

These findings differ substantially from those obtained from fitting Pareto models in the Q-learning analysis that do not include a cured fraction: an analysis that does not allow for a fraction of the population to be cured suggests that 22.2% of individuals would have improved outcomes if receiving prophylactic NHTLs and 11.7% might benefit from NHTL salvage therapy (see Table 2 of the Online Supplementary Materials). Incorrect or inappropriate modelling of the distribution can lead to considerably different treatment recommendations. This is, of course, not surprising to analysts familiar with Q-learning: the method is not robust to mis-specification of the Q-functions (i.e. it is a ‘singly robust’ approach).

4 Discussion

In this paper, we have presented a case study of estimation of an adaptive treatment strategy that required extension to the current implementations of Q-learning. In particular, we developed Q-functions for a censored outcome that (i) allowed for a fraction of the population to be cured; (ii) fit a parametric model for those who were not deemed to be cured under optimal treatment to facilitate the computation of a pseudo-outcome; and (iii) incorporated covariate effects into both components of the model, while seeking external data to compute a pseudo-outcome for those who were deemed to be cured under optimal treatment.

In our analysis of the CIBMTR data, we found that, in spite of an overall detrimental effect of NHTL prophylaxis and NHTL salvage on disease-free survival, more than a third of patients may benefit from NHTL prophylaxis, and a smaller subset may benefit from NHTL salvage (5%), when considering the expected disease-free survival (averaged over the probability of being cured). Despite the large sample size, our analysis may have had insufficient power to detect the treatment-covariate interactions needed for developing an ATS. Nevertheless, harnessing the full information in the data by modelling the survival curve rather than a dichotomized outcome [13] is an important methodological development that allows better use of the available data.

As clinical trials are typically too small to be used to learn about ATSs with more than a very small number of tailoring variables, researchers are increasingly turning to large data

sources such as the CIBMTR registry. Unfortunately, whenever using data collected for purposes other than research, there is the possibility of unmeasured confounding. Validation of a developed rule can be challenging. For example, examining outcomes within subgroups of individuals who were consistent with different treatment strategies (including the estimated optimal ATS) does not provide evidence of the superiority of one strategy over another without additional adjustments since observed treatments are subject to confounding, whether using the original data or an external, validation dataset. When good models exist for the underlying biological process such as in the context of warfarin treatment, where pharmacokinetics and dynamics are well understood, the impact of an estimated ATS may be estimated by simulating new data under a proposed strategy [20]. However, it is generally accepted that adaptive treatment strategies that are estimated from non-experimental sources such as the CIBMTR or any similar registry should be considered exploratory rather than definitive. Further studies or a randomized trial comparing a proposed ATS to some other treatment regime such as standard care would generally be warranted prior to implementing an ATS in clinical practice as there remains the possibility of confounding. Future work will include the development of sensitivity analyses to evaluate the impact of missing confounder information on an estimated ATS.

There are several other statistical issues that must be taken into consideration in any ATS analysis. For example, researchers will want to balance feasibility and parsimony in model construction. We opted to perform no model selection at all, using only variables for tailoring and confounding adjustment that were suggested by author Krakow, a blood and bone marrow transplant physician. Alternative approaches could include the use of shrinkage (see, for example, [23] for an example of penalized Q-learning with a continuous outcome), or the use of some form of information criterion such as the one that has been developed a method closely related to Q-learning [33]. While a shrinkage approach should retain the oracle property of correct variable selection, other approaches such as variable-by-variable selection based on confidence intervals or Wald tests in particular will be subject to inflation of the type I error rate.

As noted above, Q-learning depends on correct model specification. When the blip model is mis-specified, it may be the case that the ATS that is estimated is optimal within the class of treatment rules implied by this blip [27], however there are no guarantees on consistency for Q-learning under model mis-specification [16], even when all confounding variables have been measured and included in the Q-function model. Many alternatives to Q-learning have been proposed. For example, inverse weighting has been used in a survival context [28, 12] to accommodate covariate-dependent censoring, however these approaches have not been implemented in a cure-rate setting. We are currently pursuing doubly-robust estimation in a cure-rate setting.

Equally important to the statistical developments is the dissemination of the principles of ATSs into the medical community, and working with physicians and health-care providers and those who maintain registries to ensure adequate infrastructure to collate health records across institutes with differing platforms and formats for electronic medical records. Further, as such methods become more accepted in clinical practice, there may be increased interest in dynamic learning of ATSs, which could potentially raise a number of challenges that are

both statistical (e.g. incorporating uncertainty into updated ATSS) and practical (e.g. maintaining records of each ATS used at each point in time in case of complaints or legal action).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

This work was funded by a grant from the Canadian Institutes of Health Research. We are grateful to the two anonymous referees and the associate editor whose thoughtful feedback helped to improve the clarity of our work.

The CIBMTR is supported primarily by Public Health Service Grant/Cooperative Agreement 5U24CA076518 from the National Cancer Institute (NCI), the National Heart, Lung and Blood Institute (NHLBI) and the National Institute of Allergy and Infectious Diseases (NIAID); a Grant/Cooperative Agreement 4U10HL069294 from NHLBI and NCI; a contract HHS250201200016C with Health Resources and Services Administration (HRSA/DHHS); two Grants N00014-17-1-2388 and N0014-17-1-2850 from the Office of Naval Research; and grants from *Actinium Pharmaceuticals, Inc.; *Amgen, Inc.; *Amneal Biosciences; *Angiocrine Bioscience, Inc.; Anonymous donation to the Medical College of Wisconsin; Astellas Pharma US; Atara Biotherapeutics, Inc.; Be the Match Foundation; *bluebird bio, Inc.; *Bristol Myers Squibb Oncology; *Cel-gene Corporation; Cerus Corporation; *Chimerix, Inc.; Fred Hutchinson Cancer Research Center; Gamida Cell Ltd.; Gilead Sciences, Inc.; HistoGenetics, Inc.; Immucor; *Incyte Corporation; Janssen Scientific Affairs, LLC; *Jazz Pharmaceuticals, Inc.; Juno Therapeutics; Karyopharm Therapeutics, Inc.; Kite Pharma, Inc.; Medac, GmbH; MedImmune; The Medical College of Wisconsin; *Mediware; *Merck & Co, Inc.; *Mesoblast; MesoScale Diagnostics, Inc.; Millennium, the Takeda Oncology Co.; *Miltenyi Biotec, Inc.; National Marrow Donor Program; *Neovii Biotech NA, Inc.; Novartis Pharmaceuticals Corporation; Otsuka Pharmaceutical Co, Ltd. - Japan; PCORI; *Pfizer, Inc.; *Pharmacyclics, LLC; PIRCHE AG; *Sanofi Genzyme; *Seattle Genetics; Shire; Spectrum Pharmaceuticals, Inc.; St. Baldrick's Foundation; *Sunesis Pharmaceuticals, Inc.; Swedish Orphan Biovitrum, Inc.; Takeda Oncology; Telomere Diagnostics, Inc.; and University of Minnesota. The views expressed in this article do not reflect the official policy or position of the National Institute of Health, the Department of the Navy, the Department of Defense, Health Resources and Services Administration (HRSA) or any other agency of the U.S. Government.

Corporate Members

References

- Arai S, Arora M, Wang T, and et al. Increasing incidence of chronic graft-versus-host disease in allogeneic transplantation: A report from the Center for International Blood and Marrow Transplant Research. *Biology of Blood and Marrow Transplantation*, 21:266–274, 2015. [PubMed: 25445023]
- Beyersmann J, Allignol A, and Schumacher M. *Competing Risks and Multistate Models with R* Springer, New York, 2012.
- Brunstein CG, Weisdorf DJ, DeFor T, and et al. Marked increased risk of Epstein-Barr virus-related complications with the addition of antithymocyte globulin to a nonmyeloablative conditioning prior to unrelated umbilical cord blood transplantation. *Blood*, 108(8):2874–2880, 2006. [PubMed: 16804113]
- Chakrabarti S, Mackinnon S, Chopra R, and et al. High incidence of cytomegalovirus infection after nonmyeloablative stem cell transplantation: potential role of Campath-1H in delaying immune reconstitution. *Blood*, 99(12):4357–4363, 2002. [PubMed: 12036862]
- Chakrabarti S, Mautner V, Osman H, and et al. Adenovirus infections following allogeneic stem cell transplantation: incidence and outcome in relation to graft manipulation, immunosuppression, and immune recovery. *Blood*, 100(5):1619–1627, 2002. [PubMed: 12176880]
- Chakraborty B and M Moodie EE. *Statistical Methods for Dynamic Treatment Regimes: Reinforcement Learning, Causal Inference, and Personalized Medicine*. Springer-Verlag, New York, NY, 2013.
- Goldberg Y and Kosorok MR. Q-learning with censored data. *The Annals of Statistics*, 40:529–560, 2012. [PubMed: 22754029]

- Goldman AI. Survivorship analysis when cure is a possibility: a Monte Carlo study. *Statistics in Medicine*, 3:153–163, 1984. [PubMed: 6463452]
- Goldman AI. The cure model and time confounded risk in the analysis of survival and other timed events. *Journal of Clinical Epidemiology*, 44:1327–1340, 1991.
- Guo Xiang and Tsiatis Anastasios A.. A weighted risk set estimator for survival distributions in two-stage randomization designs with censored survival data. *The International Journal of Biostatistics*, 1:1–15, 2005.
- Huang X and Ning J. Analysis of multi-stage treatments for recurrent diseases. *Statistics in Medicine*, 31:2805–2821, 2012. [PubMed: 22826139]
- Huang Xueling, Ning Jing, and Wahed Abdus S.. Optimization of individualized dynamic treatment regimes for recurrent diseases. *Statistics in Medicine*, 33 no. 14:2363–2378, 2014. [PubMed: 24510534]
- Krakov EF, Hemmer M, Wang T, Logan B, Aurora M, Spellman S, Couriel D, Alousi A, Pidala J, Last M, Lachance S, and Moodie EEM. Tools for the precision medicine era: How to develop highly personalized treatment recommendations from cohort and registry data using Q-learning. *American Journal of Epidemiology*, 186:160172, 2017.
- Laber EB, Linn KA, and Stefanski LA. Interactive model building for Q-learning. *Biometrika*, 101(4): 831–847, 2014. [PubMed: 25541562]
- Martin PJ, Counts GW, Jr, Appelbaum FR, and et al. Life expectancy in patients surviving more than 5 years after hematopoietic cell transplantation. *Journal of Clinical Oncology*, 28:1011–1016, 2010. [PubMed: 20065176]
- Moodie EEM, Chakraborty B, and Kramer MS. Q-learning for estimating optimal dynamic treatment rules from observational data. *Canadian Journal of Statistics*, 40:629–645, 2012. [PubMed: 23355757]
- Moodie EEM, Dean N, and Sun YR. Q-learning: Flexible learning about useful utilities. *Statistics in Biosciences*, 2014. doi: 10.1007/s12561-013-9103-z.
- Murphy SA. Optimal dynamic treatment regimes (with discussion). *Journal of the Royal Statistical Society, Series B*, 65:331–366, 2003.
- Pavletic SZ, Kumar S, Mohty M, de Lima M, Foran JM, Pasquini M, Zhang M-J, Giralt S, Bishop MR, and Weisdorf D. NCI first international workshop on the biology, prevention, and treatment of relapse after allogeneic hematopoietic stem cell transplantation: Report from the committee on the epidemiology and natural history of relapse following allogeneic cell transplantation. *Biology of Blood and Marrow Transplantation*, 16(7):871890, 2010.
- Rich B, Moodie EEM, Stephens DA, and Platt RW. Simulating sequential multiple assignment randomized trials to generate optimal personalized warfarin dosing strategies. *Clinical Trials*, 11:435–444, 2014. [PubMed: 24464036]
- Robins JM. Optimal structural nested models for optimal sequential decisions In Lin DY and Heagerty P, editors, *Proceedings of the Second Seattle Symposium on Biostatistics*, pages 189–326, New York, 2004 Springer.
- Sabo Roy T, Scalora Allison F, Portier David, Fletcher Devon, Tessier Jeffrey M, Clark William B, Chung Harold M, McCarty John M, Roberts Catherine H, and Toor Amir A. Immune reconstitution in anti-thymocyte globulin conditioned unrelated donor stem cell transplantation. *Blood*, 122(21):2071–2071, 2013.
- Song R, Wang W, Zeng D, and Kosorok MR. Penalized Q-learning for dynamic treatment regimes. *Statistica Sinica*, 25:901–920, 2015. [PubMed: 26257504]
- Tang Xinyu and Wahed Abdus S.. Cumulative hazard ratio estimation for treatment regimes in sequentially randomized clinical trials. *Statistics in Biosciences*, doi:10.1007/s12561-013-9089-6:1–18, 2013.
- Theurich S, Fischmann H, Shimabukuro-Vornhagen A, Chemnitz JM, Holtick U, Scheid C, Skoetz N, and von Bergwelt-Baildon M. Polyclonal anti-thymocyte globulins for the prophylaxis of graft-versus-host disease after allogeneic stem cell or bone marrow transplantation in adults. *Cochrane Database Syst Rev*, 9:CD009159, 2012.
- van Buuren Stef. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, 16:219–242, 2007. [PubMed: 17621469]

- Van der Laan MJ and Petersen ML. Causal effect models for realistic individualized treatment and intention to treat rules. *The International Journal of Biostatistics*, 3, 2007.
- Wahed AS and Thall PF. Evaluating joint effects of induction-salvage treatment regimes on overall survival in acute leukemia. *Journal of the Royal Statistical Society, Series C*, 62:67–83, 2013.
- Wahed AS and Tsiatis AA. Optimal estimator for the survival distribution and related quantities for treatment policies in two-stage randomized designs in clinical trials. *Biometrics*, 60:124–133, 2004. [PubMed: 15032782]
- Wahed AS and Tsiatis AA. Semiparametric efficient estimation of survival distributions in two-stage randomisation designs in clinical trials with censored data. *Biometrika*, 93:163–177, 2006.
- Walker Irwin, Schultz Kirk R., Toze Cynthia L., Margaret Kerr Holly, Moore John, Szwajcer David, Ronan Foley Stephen, Kuruvilla John, Panzarella Tony, Couture Flix, Roy Jean, Couban Stephen, Devins Gerald, Popradi Gizelle, Lee Stephanie J., Tay Jason, Nevill Thomas J., Elemery Mohamed, and Gallagher Genevieve. Thymoglobulin decreases the need for immunosuppression at 12 months after myeloablative and nonmyeloablative unrelated donor transplantation: Cbmtg 0801, a randomized, controlled trial. *Blood*, 124(21):38–38, 2014.
- Wallace MP and Moodie EEM. Doubly-robust dynamic treatment regimen estimation via weighted least squares. *Biometrics*, 71:636–664, 2015. [PubMed: 25854539]
- Wallace MP, Moodie EEM, and Stephens DA. Model validation and selection for personalized medicine using dynamic weighted ordinary least squares. *Statistical Methods in Medical Research*, 26:1641–1653, 2017. [PubMed: 28486872]
- Zhao YQ, Zeng D, Laber EB, Song R, Yuan M, and Kosorok MR. Doubly robust learning for estimating individualized treatment with censored data. *Biometrika*, 102:151–168, 2015. [PubMed: 25937641]

Figure 1:
Kaplan-Meier curves for the first (left panel) and second (right panel) stages of treatment.
No adjustment is made for covariates or, in the first stage, for second-stage treatment.
Unadjusted Pareto fits of the curves are overlaid.

Figure 2:

CIBMTR results: optimal treatment recommendations for each patient in the sample ordered by the difference in cure probability (top row) or in survival probabilities (middle row). In each row, the left-hand panel indicates the first stage (prophylaxis) while the right-hand panel indicates stage 2 (salvage). In all figures, positive values, indicated with black crosses, favour NHTL treatments. Top row: difference in probability of a cure if given NHTL vs if given standard treatment. Middle row: difference in expected disease-free survival, if not cured (y-axes truncated to 100 months). Bottom row: scatter plot of the difference in

probability of a cure if given NHTL vs if given standard treatment, against the difference in expected disease-free survival if not cured.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1:

Covariates used to model the cure probability and the scale of the Pareto density for the CIBMTR analysis.

Stage, j	Cure model variables (H_j^θ)	Pareto scale model variables (H_j^ψ)	
			(H_j^β)
1	A_1 , age, Karnofsky/Lansky status, HLA match, disease risk score, $A_1 \times$ HLA match, $A_1 \times$ disease risk	Sex match, graft type, related donor, HLA match, CMV status ¹ , disease risk score, conditioning regimen, TBI ²	h_1^ψ , age, Karnofsky/Lansky status, PGVHcor ³
2	A_2 , age, Karnofsky/Lansky status, HLA match, disease risk score, $A_2 \times$ HLA match, $A_2 \times$ disease risk	Sex match, graft type, related donor, HLA match, CMV status ¹ , disease risk score, conditioning regimen, PGVHcor ³ , 4ISP ⁴ , time to acute GVHD, grade of GVHD, A_1	h_2^ψ , age, Karnofsky/Lansky status, TBI ²

¹ CMV status: Both donor and recipient negative for cytomegalovirus² TBI: Total-body irradiation³ PGVHcor: Corticosteroids given with GVHD prophylactic treatment⁴ 4ISP: At least 4 systemic immunosuppressive therapies given for acute GVHD

Table 2:

CIBMTR Cure-Rate Pareto Q-functions: Maximum likelihood estimators and 95% bootstrap confidence intervals of NHTL and tailoring variable parameters

	Stage 1	Stage 2
<i>Cure-probability parameters (θ)</i>		
NHTL	-0.065 (-0.226, 0.115)	-1.652 (-2.846, -1.056)
NHTL×Partially matched	0.310 (0.045, 0.608)	-0.788 (-2.191, 0.348)
NHTL×Mismatched	-0.120 (-0.463, 0.228)	-0.622 (-2.062, 0.576)
NHTL×Intermediate disease status	0.234 (-0.086, 0.461)	1.431 (0.263, 2.588)
NHTL×Advanced disease status	0.233 (-0.058, 0.516)	1.742(0.406, 3.132)
<i>Survival scale parameters (λ)</i>		
NHTL	-0.097 (-0.487, 0.274)	0.652 (-0.355, 3.411)
NHTL×Sex:F-M ¹	-0.070 (-0.342, 0.196)	-0.023 (-0.560, -0.523)
NHTL×Sex:M-F ¹	-0.002 (-0.266, 0.236)	-0.279 (-0.862, -0.261)
NHTL×Sex:M-M ¹	0.073 (-0.146, 0.309)	-0.086 (-0.554, -0.492)
NHTL×Graft: Cord blood	-0.148 (-0.600, 0.247)	-0.413 (-1.502, 0.523)
NHTL×Graft: Periph. Blood	-0.021 (-0.224, 0.180)	0.146 (-0.210, 0.565)
NHTL×Unrelated donor	0.161 (-0.047, 0.364)	0.116 (-0.264, 0.614)
NHTL×Partially Matched	-0.009 (-0.207, 0.249)	-0.133 (-0.578, 0.280)
NHTL×Mismatched	0.032 (-0.265, 0.328)	0.706(0.167, 1.384)
NHTL×CMV status ²	-0.054 (-0.243, 0.192)	-0.173 (-0.583, 0.174)
NHTL×Intermediate disease status	0.146 (-0.092, 0.408)	-0.107 (-0.571, 0.371)
NHTL×Advanced disease status	0.242 (0.057, 0.438)	0.002 (-0.384, 0.350)
NHTL×RIC/NMA conditioning ³	0.213 (0.003, 0.429)	-0.187 (-0.738, 0.255)
NHTL×TBI ⁴	0.080 (-0.123, 0.276)	-
NHTL×PGVHcor ⁵	-	-0.173 (-0.654, 0.203)
NHTL×(4ISP ⁶)	-	0.005 (-0.437, 0.436)
NHTL×Time to acute GVHD	-	-0.009 (-0.424, 0.527)
NHTL×Grade III-IV	-	-0.021 (-1.086, 0.612)
NHTL×NHTL Proph.	-	-0.262 (-0.690, 0.181)
<i>Shape parameter (α)</i>		
Standard	1.396 (1.230, 1.606)	1.771 (1.404, 2.479)
NHTL	1.944 (1.634, 2.499)	3.442 (2.596, 39.025)

¹Sex: Donor sex - recipient sex, where F indicates female and F, male

²CMV status: Both donor and recipient negative for cytomegalovirus

³RIC/NMA: Reduced Intensity Conditioning/Non-Myeloablative

⁴TBI: Total-body irradiation

⁵PGVHcor: Corticosteroids given with GVHD prophylactic treatment

⁶ 4ISP: At least 4 systemic immunosuppressive therapies given for acute GVHD

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript