ADVANCED ANALYTICS FOR PREDICTING SURVIVAL AND FACILITATING PRECISION MEDICINE IN CHECKPOINT IMMUNOTHERAPY

Hadi Beyhaghi

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Health Policy and Management in the Gillings School of Global Public Health.

Chapel Hill
2022

Approved by:

Kristen Hassmiller Lich

Antonia Bennett

Stacie Dusetzina

Anna Kahkoska

Michael R. Kosorok

ABSTRACT

Hadi Beyhaghi: Advanced Analytics for Predicting Survival and Facilitating Precision Medicine in Checkpoint Immunotherapy
(Under the direction of Kristen Hassmiller Lich)

Checkpoint immunotherapy drugs, either individually or in combination with other drugs, have become standard of care for many cancers. The long-term survival impacts of these drugs on a subset of treated population and their unique survival dynamics have challenged traditional statistical methods to model long-term survival impacts and estimate individualized treatment rules. In this dissertation, I proposed novel machine learning techniques that can be used to tackle some of these challenges.

This research addresses the following three aims using patient-level data from a checkpoint immunotherapy clinical trial for advanced melanoma: (1) Develop a novel individual-level survival extrapolation method for right-censored observations, and compare the predictive accuracy of the proposed method with population-level standard parametric models. (2) Compare the accuracy of survival extrapolation models that directly model heterogeneity of treatment response to the accuracy of the proposed survival extrapolation model from Aim 1 that incorporates cure fraction models at the individual level. (3) Estimate individualized treatment rules (ITRs) and calculate survival and cost impacts associated with implementing them in the trial cohort compared to the survival and cost impacts associated with universal use of the trial-recommended treatment.

The Aim1 paper provides a tutorial that introduces the kernel-weighted survival forest (KWSF) model, a novel survival extrapolation method that uses patient-level characteristics to estimate individualized survival function. The findings showed that compared to standard

parametric models, KWSF more accurately predicted survival beyond the available trial follow up. The results of the Aim2 paper showed that compared to models that use standard parametric extrapolation, cure fraction models and KWSF with cure fraction extrapolation function were more accurate in predicting survival in the immunotherapy arm of the trial. The KWSF model with a cure fraction survival extrapolation function demonstrated comparable accuracy with cure fraction models, while uniquely allowing for estimating individual-level survival functions. The findings of Aim3 paper showed that compared to allocating treatment based on the average treatment effect from a clinical trial, treatment allocation based on the estimated ITRs resulted in higher survival gains and lower direct treatment costs, which is likely to persist even when considering the cost of implementing individualized treatment.

To my parents Fatemeh and Hosseinali for their love and inspiration

ACKNOWLEDGEMENTS

I would like to express my gratitude to my committee members Kristen Hassmiller Lich, Michael Kosorok, Antonia Bennett, Stacie Dusetzina, and Anna Kahkoska. I would specially like to thank my committee chair and adviser (Kristen Hassmiller Lich) for her support and guidance throughout my dissertation process. My sincere gratitude extends to Michael Kosorok who has taught me so much during our many regular meetings. The other members of my dissertation committee, Antonia Bennett, Stacie Dusetzina, and Anna Kahkoska, all taught me so much about research and effective scientific communication. I also owe a huge thank you to Yifan Cui for his continuous support and tremendous help with the coding in this dissertation and for his technical expertise and insightful feedback.

I am very grateful to the Bristol Myers Squibb WWHEOR group whose doctoral fellowship program benefitted me in many ways. I would like to acknowledge Bristol Myers Squibb Clinical Trial Data Transparency Program for proving patient-level data from CA184-024 Study, which were instrumental for the analyses in this dissertation. I would like to specifically thank Srividya Kotapati for her support and mentorship, as well as James Shaw and Kyna Gooden who helped me during my fellowship. I'd like to thank former and current faculty from Health Policy and Management especially Sally Stearns, George Pink, Paula Song, Marisa Domino and Andrea Biddle for giving me so many opportunities to learn and improve myself through course work, and research.

My gratitude also extends to my employer, Novavax who gave me both flexibility and financial support to continue my dissertation progress while growing in my professional

career. In particular, I’d like to thank Seth Toback for his support and mentorship. Lastly, I would like to thank my family and friends for their love, support, and continuous encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AIC	Akaike Information Criterion
CART	Classification and Regression Tree
CTLA-4	Cytotoxic T-Lymphocyte Associated Protein 4
DTIC	Dacarbazine
ECOG	Eastern Cooperative Oncology Group
EHR	Electronic Health Records
ERT	Extremely Randomized Trees
FDA	Food and Drug Administration
ICER	Incremental Cost-Effectiveness Ratio
IO	Immuno-Oncology
ITR	Individualized Treatment Rule
KM	Kaplan-Meier
KWSF	Kernel-Weighted Survival Forest
LDH	Lactate Dehydrogenase
LY	Life Year
MCM	Mixture Cure Model
MSE	Mean Squared Error
NMCM	Non-Mixture Cure Model
OS	Overall Survival
OWL	Outcome-Weighted Learning
RCT	Randomized Controlled Trial
RIST	Recursively-Imputed Survival Trees
SPM	Standard Parametric Model
WTP	Willingness to Pay

CHAPTER 1: INTRODUCTION

Clinical trials of treatments that influence survival can face data limitations as a result of time and budget constraints. Specifically, results are commonly reported before key events, for example death, are observed for every participant in a clinical trial (i.e., some individuals are right-censored). In the presence of significant censoring, extrapolation beyond trial follow-up duration is necessary to estimate the complete survival impact of a new intervention.[1]

Extrapolation beyond trial follow-up becomes particularly challenging when interventions' mechanisms of action result in more complex hazard functions, due, for example, to heterogenous survival effects for a subset of patients. Treatment with checkpoint immunotherapy drugs is a prime example of such interventions. These drugs target the immune system checkpoints (molecules on certain immune cells that need to be activated or inactivated to start an immune response). Numerous trials have shown the positive impact of these drugs on overall survival; hence these drugs are considered the standard of care in many cancers.[2,3] Checkpoint immunotherapy drugs typically trigger a durable response in a subset of patients, which translates to long-term survival for those patients.[4] There is ongoing investigation into understanding the reasons behind the heterogeneity in response and what can be done to increase the response rate.[5]

When used as a clinical input, for example in cost-effectiveness models, data from checkpoint immunotherapy trials are usually less mature than they need to be (i.e., the overall survival (OS) curves have yet to reach the median survival point).[6,7] Therefore, substantial extrapolation is required, making the plausibility of the extrapolated portion of

alternative models far more important than the fit to the observed data.[1,8,9] Numerous methods have been suggested and used to extrapolate survival beyond clinical trial follow-up especially in the context of cost-effectiveness analysis of checkpoint immunotherapy drugs.[10] Because cost-effectiveness models for these interventions typically require a lifetime horizon, understanding the long-term survival impact of different interventions is key to determine the relative economic and clinical value of checkpoint immunotherapy drugs. The majority of the survival extrapolation methods focus on modeling survival at the population level (e.g., standard parametric models, piecewise models) or subsets of population (e.g., cure models and landmark/response-based models).[11-13]

Given the level of heterogeneity associated with checkpoint immunotherapy treatment response, it is important to develop models that are capable of estimating individual-level survival functions that can accurately predict survival beyond available follow up of a clinical trial (i.e., individual-level extrapolated survival functions). Such models leverage patient characteristics to estimate survival functions and can serve as the foundation for individual-level simulations to model the relative health and economic value of different treatment strategies, taking into account individual variation in treatment responses and ultimate survival outcomes.

Additionally, individual-level extrapolated survival functions can be used to develop individualized treatment rules (ITRs) to inform precision and personalized medicine strategies. An ITR is a data-driven decision algorithm that recommends treatment according to patient characteristics in a way that, if implemented in practice, can maximize the health outcome(s) of interest at both individual and population level.[14] Because of the high cost and a potential for serious adverse effects associated with checkpoint immunotherapy treatments, identifying which patients will benefit from these drugs has become increasingly critical. Therefore, ITRs may play a critical role in selecting patients for immunotherapy.

Clinical trial data are commonly used to construct ITRs; however, estimating ITRs particularly for checkpoint immunotherapy treatments can be challenging due to heterogeneous response to treatment, numerous potential outcome predictors, and a limited follow-up of clinical trials.[15] Individual-level extrapolated survival functions can provide direct input for ITR estimation models while allowing these models to better capture the long-term survival impacts of different interventions. Additionally, to convince patients, healthcare providers, and healthcare systems to adopt these individualized treatment models in real-world practice, evidence needs to be generated to show the potential economic and clinical benefits of implementing such strategies compared to treating patients based on the average treatment effect of a clinical trial.[16] Such evidence is lacking in the literature as the majority of published papers focus on the methodological aspect of developing ITRs.[17]

My proposed research seeks to fill the research gaps described above by developing a new method for estimating individual-level extrapolated survival functions, estimate the predictive accuracy of the new method compared to population-level methods, and proposing a novel approach to incorporate these individual-level extrapolated survival curves in an ITR estimation model. Furthermore, I estimate the potential economic and survival impact associated with implementing the estimated ITRs, using clinical trial data. My long-term goal is to help patients, physicians, and healthcare systems make better-informed decisions about novel cancer treatments and ultimately improve patients' lives, while potentially reducing (and not substantially increasing) healthcare resource use. My central hypothesis is that using the estimated ITRs in treatment decisions has the potential to improve patient's outcomes and reduce healthcare costs compared to treatment allocation based on average treatment effects from clinical trials. My research addresses the following three aims:

Aim 1: To develop a novel individual-level survival extrapolation method for right-censored observations, and compare the predictive accuracy of the proposed method with population-level standard parametric models.

To achieve Aim 1, I develop, describe and implement a novel survival extrapolation method that combines a non-parametric survival model based on extremely randomized trees with kernel-weighted parametric extrapolation. I then compare the accuracy of the resulting survival predictions with the results of population-level standard parametric extrapolation by estimating the mean squared error (MSE) associated with each model's estimates. While the proposed method is more computationally complex, I hypothesize that compared to standard parametric models, it confers greater accuracy in estimating individual-level long-term survival effects. This aim is written as a tutorial with the objective of making this methodological innovation accessible to decision modelers.

Aim 2: To compare the accuracy of survival extrapolation models that are designed to directly model heterogeneity of treatment response (i.e., cure fraction models and response-based/landmark models) to the accuracy of the proposed survival extrapolation model from Aim 1 that incorporates cure fraction models at the individual level.

A class of approaches have been introduced recently, which are designed to accommodate some heterogeneity in population-level survival extrapolation – including cure fraction models and response-based/landmark models. In general, these approaches segment the population into more homogenous groups (e.g., based on their objective response to treatment in the response-based models) and use different models to extrapolate the survival for each group. The cure fraction models have been shown to have high accuracy in extrapolating population-level overall survival for checkpoint immunotherapy treatment.[12,18] Although these methods offer more flexibility, they typically require longer follow-up to detect populations with markedly different survival.[18] The proposed individual-level survival extrapolation method in Aim 1 has the capability to utilize

similar models to improve the prediction accuracy; however, it is not clear whether using these functions in individual-level extrapolation offers more accurate survival predictions. Therefore, it is possible that these population-level approaches may have similar accuracy to the proposed individual-level extrapolation methods.

In Aim 2, I implement several variations of the individual-level survival extrapolation model described in Aim 1 with mixture and non-mixture cure fraction models as extrapolation functions. Parallel to the approach in Aim 1, I estimate the population-level extrapolated survival functions using cure models (mixture and non-mixture) and landmark/response-based model and then compare the accuracy of the resulting survival predictions with the corresponding results from individual-level extrapolation by estimating the MSE associated with each model's estimates across multiple time points. While the proposed method in Aim 1 is uniquely capable of providing individual-level extrapolated survival curves, I hypothesize that compared to population-level survival extrapolation methods that directly model heterogeneity, an individual-level extrapolation that uses similarly flexible survival functions confers similar or greater accuracy in estimating individual-level long-term survival effects.

Aim 3: To estimate ITRs using most accurate survival projections from Aims 1 and 2 and calculate survival and cost impacts associated with implementing these ITRs in the trial cohort compared to the survival and cost impacts associated with universal use of the trial-recommended treatment, with the goal of maximizing overall survival among patients with advanced melanoma.

To achieve this aim, I used the most accurate individual-level extrapolated survival estimates from methods proposed in aims 1 and 2 as inputs in an outcome-weighted learning algorithm, an innovative classification approach that uses support vector machine techniques, to develop ITRs that maximize patient survival.[19] I describe the characteristics of the subgroup who is assigned to each treatment, and estimate the direct treatment cost

(payer perspective) and survival impact of two distinct scenarios in the cohort of patients studied in the advanced melanoma trial: (1) where treatment allocation is based on the average treatment effect of the clinical trial, and (2) where treatment allocation is based on the estimated ITRs. Considering these results, I discuss the net monetary benefit of individualizing treatment in practice more broadly, taking into account costs and feasibility of implementing individualized treatment in real-world practice. The main hypothesis of this aim is that compared to allocating treatment based on the average treatment effect from a clinical trial, treatment allocation based on the estimated ITRs results in higher survival gained and lower direct treatment cost, which is likely to persist even when considering the cost of implementing individualized treatment and willingness to pay for life-years gained.

The first two aims are methodological, addressing gaps in decision science methods through the use of predictive analytics (machine learning algorithms) and setting the stage for informed decision-making. The third aim builds on this foundation and other novel machine learning methods to inform cancer treatment decision making by assessing the cost and survival impacts of individualized treatment versus treatment assignment based on clinical trial results.

For all three aims of this study, I used patient-level data from CA184-024 Study “A Multi-Center, Randomized, Double-Blind, Two-Arm, Phase III Study in Patients with Untreated Stage III (Unresectable) or IV Melanoma Receiving Dacarbazine plus 10 mg/kg of Ipilimumab vs. Dacarbazine with Placebo”. In this trial, a total of 502 subjects were randomized (250 to ipilimumab plus DTIC and 252 to DTIC monotherapy).[20] The CA184-024 is one of the longest running phase III trials of checkpoint immunotherapy for advanced melanoma. Available trial data includes minimum follow-up of 5 years.[21] This phase III trial provides rich data to validate predictive algorithms for extrapolating survival data. In this trial, similar to other checkpoint immunotherapy drugs, ipilimumab was found to produce long-term survival benefits in a subgroup of patients treated.[22] While the treatments evaluated in

this clinical trial are no longer considered as standard of care for advanced melanoma, the proposed methods provide prototype approaches that can be used to inform future treatment decisions especially in the context of checkpoint immunotherapy.

Results from these three aims have implications for decision analysis methods, clinical care and more specifically precision medicine, and policy development. This work serves as a case example of novel methodologic approaches to predict long-term survival impacts of checkpoint immunotherapy treatments beyond trial follow up that account for individual-level heterogeneity in treatment response. Lastly, although treatment decisions involve a number of complex and inter-related factors, application of the proposed predictive models may provide valuable individualized information that can improve decision making in the clinical setting.

The remainder of this dissertation is organized as follows: Chapters 2-4 are individual manuscripts that correspond to Aims 1-3. These chapters are concise and intended to be submitted to peer-reviewed journals and therefore, formatted as such. Chapter 5 presents a summary of key insights and implications from this research: I discuss the strengths and weaknesses of this work and its relevance to practice, policy, and future research.

REFERENCES

1. Latimer NR. Survival analysis for economic evaluations alongside clinical trials--extrapolation with patient-level data: inconsistencies, limitations, and a practical guide. Med Decis Mak Int J Soc Med Decis Mak. 2013 Aug;33(6):743–54.

2. Azoury SC, Straughan DM, Shukla V. Immune Checkpoint Inhibitors for Cancer Therapy: Clinical Efficacy and Safety. Curr Cancer Drug Targets. 2015;15(6):452–62.

3. He X, Xu C. Immune checkpoint signaling and cancer immunotherapy. Cell Res. 2020 Aug;30(8):660–9.

4. Gemmen E, Parmenter L. Special Considerations for the Analysis of Patient-Level Immuno-Oncology Data. 2018;4(1):23–4.

5. Saenger YM, Wolchok JD. The heterogeneity of the kinetics of response to ipilimumab in metastatic melanoma: patient cases. Cancer Immun. 2008 Jan 17;8:1.

6. Guyot P, Ades AE, Beasley M, Lueza B, Pignon JP, Welton NJ. Extrapolation of Survival Curves from Cancer Trials Using External Information. Med Decis Mak Int J Soc Med Decis Mak. 2017 May;37(4):353–66.

7. Kim H, Goodall S, Liew D. Health Technology Assessment Challenges in Oncology: 20 Years of Value in Health. Value Health J Int Soc Pharmacoeconomics Outcomes Res. 2019 May;22(5):593–600.

8. Huang M, Latimer N, Zhang Y, Mukhopadhyay P, Ouwens M, Briggs A. Estimating the Long-term outcomes associated with Immuno-oncology therapies: challenges and approaches for overall survival extrapolations. Value Outcomes Spotlight. 2018;4(1):28–30.

9. Connock M, Hyde C, Moore D. Cautions regarding the fitting and interpretation of survival curves: examples from NICE single technology appraisals of drugs for cancer. PharmacoEconomics. 2011 Oct;29(10):827–37.

10. Hawkins N, Grieve R. Extrapolation of Survival Data in Cost-effectiveness Analyses: The Need for Causal Clarity. Med Decis Mak Int J Soc Med Decis Mak. 2017 May;37(4):337–9.

11. Gibson E, Koblbauer I, Begum N, Dranitsaris G, Liew D, McEwan P, et al. Modelling the Survival Outcomes of Immuno-Oncology Drugs in Economic Evaluations: A Systematic Approach to Data Analysis and Extrapolation. PharmacoEconomics. 2017 Dec;35(12):1257–70.

12. Bullement A, Latimer NR, Bell Gorrod H. Survival Extrapolation in Cancer Immunotherapy: A Validation-Based Case Study. Value Health J Int Soc Pharmacoeconomics Outcomes Res. 2019 Mar;22(3):276–83.

13. Latimer N. NICE DSU technical support document 14: survival analysis for economic evaluations alongside clinical trials-extrapolation with patient-level data. 2011;

14. Xu Y, Greene TH, Bress AP, Sauer BC, Bellows BK, Zhang Y, et al. Estimating the optimal individualized treatment rule from a cost-effectiveness perspective. Biometrics. 2022 Mar;78(1):337–51.

15. Cui Y, Zhu R, Kosorok M. Tree based weighted learning for estimating individualized treatment rules with censored data. Electron J Stat. 2017;11(2):3927–53.

16. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J. 2019 Jun;6(2):94–8.

17. Kosorok MR, Laber EB. Precision medicine. Annual review of statistics and its application. 2019 Mar;6:263

18. Othus M, Bansal A, Koepl L, Wagner S, Ramsey S. Accounting for Cured Patients in Cost-Effectiveness Analysis. Value Health J Int Soc Pharmacoeconomics Outcomes Res. 2017 Apr;20(4):705–9.

19. Zhao Y, Zeng D, Rush AJ, Kosorok MR. Estimating Individualized Treatment Rules Using Outcome Weighted Learning. J Am Stat Assoc. 2012 Sep 1;107(449):1106–18.

20. Robert C, Thomas L, Bondarenko I, O'Day S, Weber J, Garbe C, et al. Ipilimumab plus dacarbazine for previously untreated metastatic melanoma. N Engl J Med. 2011 Jun 30;364(26):2517–26.

21. Maio M, Grob JJ, Aamdal S, Bondarenko I, Robert C, Thomas L, et al. Five-year survival rates for treatment-naive patients with advanced melanoma who received ipilimumab plus dacarbazine in a phase III trial. J Clin Oncol Off J Am Soc Clin Oncol. 2015 Apr 1;33(10):1191–6.

22. Péron J, Lambert A, Munier S, Ozenne B, Giai J, Roy P, et al. Assessing Long-Term Survival Benefits of Immune Checkpoint Inhibitors Using the Net Survival Benefit. J Natl Cancer Inst. 2019 Nov 1;111(11):1186–91.

CHAPTER 2: KERNEL-WEIGHTED SURVIVAL FOREST FOR PROJECTING INDIVIDUAL-LEVEL MORTALITY: A TUTORIAL AND DEMONSTRATION USING DATA FROM AN IMMUNOTHERAPY TRIAL FOR ADVANCED MELANOMA

Introduction

Clinical trials of treatments that influence survival can face data limitations as a result of time and/or budget constraints. Specifically, trial results are commonly reported before key events, for example death, are observed for every participant in a clinical trial (i.e., some individuals are right-censored). In the presence of censoring, extrapolation beyond trial follow-up is necessary to estimate the complete survival impact of a new intervention.[1]

Extrapolation beyond trial follow-up becomes particularly challenging when an intervention's mechanism of action results in more complex hazard functions, due for example, to heterogenous survival effects for a subset of patients. Treatment with checkpoint immunotherapy drugs is a prime example of such interventions. These drugs target the immune system checkpoints -- molecules on certain immune cells that need to be activated or inactivated to start an immune response -- and are becoming the standard of care in many cancers.[2,3] Checkpoint immunotherapy drugs trigger a durable response in a subset of patients, which translates to long-term survival for some – but not all – patients.[4] When used as a clinical input, for example in cost-effectiveness models, data from checkpoint immunotherapy trials are typically less mature than they need to be. Specifically, the overall survival (OS) curves have yet to reach the median survival point.[5,6] Therefore, substantial extrapolation is required, making the plausibility of the extrapolated portion of alternative models far more important than the fit to the observed data.[1,7,8]

Numerous methods have been suggested and used to extrapolate survival beyond clinical trial follow-up, especially in the context of cost-effectiveness analysis of checkpoint immunotherapy drugs.[9] However, the majority of these methods focus on modeling survival at the population level (e.g., standard parametric models) or among subsets of the population (e.g., cure models and landmark models). Given the level of heterogeneity of treatment response associated with these therapies, it is imperative to develop models that are capable of flexibly estimating individual-level survival functions that predict survival beyond available follow up of a clinical trial. Such models can inform precision and personalized medicine strategies, and serve as the foundation for individual-level simulation of these strategies' relative economic value and efficiency.

As the choice of extrapolation models could substantially impact survival estimates,[10] more accurate alternative modeling approaches are needed to extrapolate survival beyond trial follow-up that can accommodate the level of heterogeneity and survival dynamics present among treatments like checkpoint immunotherapy.[7,10] In this tutorial, we provide a novel method that leverages survival forest and kernel-weighted parametric extrapolation to estimate patient-level survival functions to predict survival beyond trial follow up. More specifically, this tutorial aims to: (1) describe, in detail, the kernel-weighted survival forest (KWSF) model as a novel individual-level survival extrapolation method for right-censored observations; and, (2) implement and compare the accuracy of KWSF with population-level standard parametric models, using patient-level data from a checkpoint immunotherapy trial in patients with advanced melanoma.

This tutorial is organized as follows: In the next section, we provide a step by step guide to understand and apply a survival forest algorithm in addition to individual-level kernel-weighted parametric extrapolation (together comprising the KWSF); Next, we compare the results of the KWSF extrapolation with standard population-level parametric models, using an illustrative example from a checkpoint immunotherapy trial in advanced

melanoma patients. Finally, we discuss the advantages and limitations of KWSF survival extrapolation.

Survival Forests and Kernel-Weighted Extrapolation

Below, we describe two main components of the KWSF algorithm, specifically survival forests and kernel-weighted parametric extrapolation.

Survival Forest

Survival forests are a category of tree-based models that have been suggested for imputing failure times (e.g., death in our case) for right-censored observations. Tree-based models form a broad class of nonparametric estimators for regression and classification and have evolved to some of the most popular machine learning tools with applications in survival analysis.[11,12] These models provide a powerful tool to classify observations into homogenous groups, which can be used for extrapolating survival beyond trial follow up. Although other tree-based methods have been suggested for modeling survival in right-censored observations, survival forest models based on extremely randomized trees (ERTs) tend to be more accurate than other tree-based models[13] and were used in this tutorial. The survival ERT model was introduced by Zhu and Kosorok in 2012 as part of the recursively-imputed survival trees (RIST) algorithm.[13]

Extremely randomized trees

To better understand the ERT algorithm, a sophisticated tree-based prediction model, we first review a few basics on decision trees and the process of building them. Of note, decision trees discussed here should not be confused with decision tree modeling, a modeling framework commonly used in cost-effectiveness models as a way to lay out decision alternatives (decision nodes) and chance events (event nodes) culminating in outcomes to support decision making. Tree-based methods have become increasingly popular statistical tools since Breiman and colleagues introduced the classification and regression tree (CART) algorithm in 1984.[14] Early tree-based methods, including CART,

were based on a single tree structure, and the prediction rules were easy to interpret. Although the CART algorithm is one of the better-known tree-based methods, more sophisticated and accurate methods have been introduced since.[15]

Fitting a Single Decision Tree

A decision tree is the building block of tree-based prediction models. The term *decision tree* is used to describe a set of splitting rules, summarized in a tree structure, that are used to segment the predictor space of a given dataset, where predictor space is defined as a set of possible values for different predictors in a dataset.[15]

Before starting to build a decision tree, the data need to be partitioned into the training and test datasets. Training data are the observations that are used to develop the predictive model (i.e., decision tree) that estimates the outcome based on available predictors.[15] Different methods of partitioning data to training and test sets exist; however, the appropriateness of the partitioning method depends on the research question and the type of prediction model.[16]

With training data in hand, building a decision tree requires a set of binary questions, the answer to each resulting in a split of the predictor space. The goal is to identify a series of splits that lead to the most accurate predictions. Each binary question partitions the predictor space into two distinct and non-overlapping regions, typically based on the value of only a single predictor at a time. A “goodness-of-split criterion” is required to assess the ability of competing splits to create the most distinct daughter nodes with respect to the outcome that the decision tree is trying to predict. In other words, goodness-of-split criteria are determined based on the outcome that the model is striving to predict. Every allowable split on each and every predictor is examined against the selected goodness-of-split criterion, and the best of these splits is chosen to partition the observations into two daughter nodes. Notably, a single predictor with more than two distinct values (i.e., potential splitting points) could be used to create multiple splits of the predictor space corresponding

to each splitting point. This process of splitting is repeated for each of the resulting daughter nodes, until a stopping criterion (e.g., a minimum number of observations in each daughter node) is reached. At this point, daughter nodes are referred to as terminal nodes.

As for any prediction model, tree-based methods attempt to minimize the prediction error and predict the target outcome as accurately as possible. Learning algorithms achieve this goal by increasing a model's complexity, for example, through adding more and more predictors to the model. This tendency of the algorithm, although leading to steady fall in bias, might result in "overfitting" of the model. Overfitting happens when the prediction algorithm models the random noise in the training data rather than the relevant relations between predictors and target outcomes. While an over-simplistic model that fails to use all relevant data for prediction (under-fitted model) can increase prediction error via increasing the bias, overfitting the model may increase the prediction variance (i.e., the amount by which the prediction for a given observation would change if we estimated the target outcome using a different training dataset). In prediction models, bias and variance are both important and one should not be improved at an excessive expense of the other.[15] The bias-variance tradeoff in the context of the proposed model is explained below.

Single tree models like CART, if grown deeply enough, can minimize bias at the expense of increased variance (overfitting). As many other tree-based methods, CART deals with overfitting by pruning the decision tree – a technique that reduces the size of decision trees by removing sections of the tree that provide little power to classify observations,[15] hence decreasing the variance without a significant increase in bias. Once the tree is built and pruned, it can be used to predict the outcome of interest for a given observation based on the corresponding predictor values.

Ensembles and Randomization

More recent tree-based methods address overfitting differently, for example by training numerous decision trees, introducing randomization in the tree building process, or

a combination of both. By creating many decision trees – an ensemble – and then averaging them, the variance of the final model can be greatly reduced over that of a single tree. However, the bias of the full model is equivalent to the bias of a single decision tree (which itself has low bias but high variance).[17]

Implementing and interpreting a single decision tree model is fairly straightforward. However, accuracy has been shown to improve through implementation of methods that use ensembles and randomization.[18] A common example, Breiman introduced a framework for tree ensembles called "Random Forests.[19]" Random forests make use of a process called bootstrap aggregation ("bagging") to create an ensemble of decision trees. In bagging, numerous replicates of the original dataset are created using random sampling with replacement from the training dataset (bootstrapping) and a decision tree is fitted to each replicate dataset. Averaged predictions across the ensemble of trees are used to predict the target outcome (in the form of the predicted probability) for a given observation. In addition, rather than trying every possible predictor at each split of the tree, the random forest algorithm randomly selects a subset of predictors to be considered. More recently, Geurts and colleagues introduced the ERT method, implementing randomization at multiple levels, but without bagging.[18] In the next section, we explain the features and process of fitting ERTs to training data to predict patient survival, as one of the two components of the KWSF method introduced in this tutorial.

Fitting a Survival ERT

The KWSF process starts by fitting an extremely randomized tree to the entire training dataset. Note that unlike random forest, ERTs are not built on bootstrap replications of the original training dataset but use the entire training dataset (Figure 2.1). Assume that the training dataset includes P predictors (where P = 10 in the Figure 2.1 example). For each predictor in the training dataset, a number of possible splitting points exist that correspond to the distinct values of that predictor. For example, a splitting point of 56 years

for the predictor AGE means that one can divide the training dataset into two groups: people younger than 56 and people who are 56 or older. Many other possible splitting points likely exist for AGE in the dataset, which will be considered in the ERT process. The ERT algorithm handles continuous and ordinal variables with ease. However, for nominal variables, we recommend creating binary variables corresponding to the number of distinct values similar to an indicator (dummy) variable approach in the context of regression models.

For each split of the tree, the ERT algorithm randomly picks K predictors (where K=5 in Figure 2.1 example) from the list of P possible predictors in the training dataset along with one randomly selected splitting point for each selected predictor. Randomly selecting the splitting point, a feature of ERT, adds another level of randomization to the fitted trees, hence the name extremely randomized tree. Randomization both at the *predictor* and the *splitting point* level allows the ERT algorithm to build distinct trees despite starting with the same dataset, as every tree is fitted to the entire training dataset. Once the K 'predictor-splitting point' pairs are selected, the ERT algorithm tries each of the K candidate pairs in the split and picks the pair that maximizes the log-rank statistic (i.e., the goodness-of-split criterion for survival ERT) between the two resulting daughter nodes, where the Kaplan-Meier estimator is used to calculate the survival function within each node. Identified this way, the predictor-splitting point pair will provide the most distinct daughter nodes in terms of survival among the K candidate pairs. The tree growing process continues until the model exhausts all possible predictor-splitting point pairs or until a node has no less than a user-determined minimum number of observed events (e.g., deaths). The minimum number of observed events is a tuning parameter of the algorithm (see *Setting Tuning Parameters*). Of note, unlike CART models, pruning is not used in building ERTs, as ensemble methods and multi-level randomization are used to prevent overfitting. This process allows the ERT

algorithm to minimize bias by building deeper trees while avoiding the problem of high variance.

Fitting M Survival ERTs

The survival forest algorithm uses the above procedure M times to generate M ERTs, each starting with the entire training dataset. The number of trees M is a tuning parameter (see *Setting Tuning Parameters* below). Although each tree is built on the same dataset (i.e., the entire training dataset), randomization at predictor and splitting point levels helps generate M distinct trees, illustrated as differently shaped trees in Figure 2.2.

Estimating Survival function

For each terminal node of a fitted ERT, the Kaplan-Meier (KM) survival function is calculated. Having several events in each terminal node allows the algorithm to calculate the KM survival function within each terminal node. For any particular individual, that person eventually falls into only one terminal node per each fitted ERT (Illustrated as red paths in Figure 2.2). The algorithm assumes that every individual who falls in a terminal node has the corresponding survival function of that node, which is referred to as within-terminal node homogeneity. Each individual will have a tree-specific survival function denoted $\hat{S}_m^i$ in Figure 2.2, which is the survival function of the corresponding terminal node that the individual falls into within that tree. Averaging over M trees, the forest-level survival function for an individual can be estimated using the following equation:

$$\hat{S}_i = \frac{1}{M}\sum_{m=1}^{M} \hat{S}_m^i$$

Individual-Level Kernel-Weighted Survival Extrapolation

To extrapolate survival for an individual, we use the forest-level survival function that was estimated by averaging the survival functions across all terminal nodes that include the particular individual ($\hat{S}_i$). Multiple parametric distributions e.g., exponential, Weibull, log logistic, log normal, gamma, and generalized gamma are fitted to the forest-averaged

survival function to extrapolate survival beyond trial follow-up for that individual. The distribution with the lowest Akaike Information Criterion (AIC) is chosen to estimate the extrapolated survival function. The same process is repeated for every individual in the trial.

Setting tuning parameters

The ERT model offers several tuning parameter adjustments. For example, the number of predictors considered at each split, K, can be adjusted by the user. By default, K is set to the integer part of $\sqrt{P}$, where P is the number of predictors in the dataset. Increasing the value of K could result in reducing bias but at the cost of increasing the calculation burden. The user can also determine the minimal number of observed events (i.e., deaths) in each terminal node, n_{min}. As n_{min} gets smaller, deeper trees can be fitted, which can result in decreased bias but increased likelihood of overfitting. Similarly, the user can determine the number of trees, M, that are fitted to form the survival forest. We expect that increasing the number of trees would improve the prediction accuracy albeit at the cost of increasing the calculation burden.

Illustrative Example

As an illustration, we implement and compare the performance of KWSF extrapolation with standard parametric models using clinical trial data from the CA184-024 Trial. This trial evaluated the efficacy of ipilimumab (a checkpoint immunotherapy drug) combined with Dacarbazine (a chemotherapy drug) compared to the standard of care at the time of the trial in advanced melanoma patients. Advanced melanoma is the most aggressive form of skin cancer and is associated with poor prognosis with median OS ranging from 5.1 to 22.3 months.[20] Ipilimumab is a monoclonal antibody that attaches to cytotoxic T-lymphocyte associated protein 4 (CTLA-4), a protein on some T cells that acts as a type of “off switch” to keep the immune system in check and to stop it from working.[21,22] By inhibiting CTLA-4, Ipilimumab can boost the body’s immune response against cancer cells. Because this trial has longer follow-up period than most checkpoint immunotherapy trials,

the data allow to assess the prediction accuracy of proposed extrapolation models using varied amounts of follow-up data (i.e., two and three years).

Dataset

The CA184-024 Study is "A Multi-Center, Randomized, Double-Blind, Two-Arm, Phase III Study in Patients with Untreated Stage III (Unresectable) or IV Melanoma Receiving Dacarbazine Plus 10 mg/kg of Ipilimumab vs. Dacarbazine with Placebo". A total of 502 subjects were randomized to ipilimumab plus DTIC (n=250), hereafter referred to as immunotherapy arm and to DTIC monotherapy (n=252), hereafter referred to as chemotherapy arm. [23] Ipilimumab is the first FDA-approved checkpoint immunotherapy for advanced melanoma, which makes CA184-024 one of the longest running phase III trials of any checkpoint immunotherapy for advanced melanoma.[24-26] We identified 19 predictive variables based on the literature and data availability; predictors with any missing values were excluded from the analysis. All analyses were conducted based on an intention-to-treat framework i.e., trial data were analyzed assuming that subjects received the randomly assigned treatment.

Partitioning the data

For each arm of the CA184-024 trial, we construct two longitudinally-partitioned subsets of the data that include 2-year and 3-year follow up (the training datasets). The two and three years of follow up were selected because trial data are typically reported within these time frames for regulatory submission and obtaining reimbursement. For the 2-year partitioning, subjects who did not experience the event and were not censored before 2 years were assumed to be censored at 2 years. Hence, in the corresponding training dataset, these subjects will have survival time (Y_{il2}) of 2 years and censorship status (δ_{il2}) of 0 (censored). Subscript *l2* indicates 2-year longitudinal partition. For subjects who experienced the event or were censored before 2 years, their survival time (Y_{il2}) and censorship status (δ_{il2}) remain the same as the values in the original dataset (Y_i and δ_i,

respectively). Similar process was applied to 3-year partitioned data (see the equations below). This process creates two training datasets per trial arm for a total of four training datasets.

For the 2-year training dataset:

$$Y_{il2} = \begin{cases} 2\ years, & Y_i > 2\ years \\ Y_i, & Y_i \leq 2\ years \end{cases} \quad \delta_{il2} = \begin{cases} 0, & Y_i > 2\ years \\ 0, & Y_i \leq 2\ years, \delta_i = 0 \\ 1, & Y_i \leq 2\ years, \delta_i = 1 \end{cases}$$

For the 3-year training dataset (subscript *l3* indicates 3-year longitudinal partition):

$$Y_{il3} = \begin{cases} 3\ years, & Y_i > 3\ years \\ Y_i, & Y_i \leq 3\ years \end{cases} \quad \delta_{il3} = \begin{cases} 0, & Y_i > 3\ years \\ 0, & Y_i \leq 3\ years, \delta_i = 0 \\ 1, & Y_i \leq 3\ years, \delta_i = 1 \end{cases}$$

Data Analysis

We estimated the "true" individual-level survival functions by applying the survival forest algorithm as explained above using data from the full duration of available follow-up. The survival forest provides a nonparametric estimate of the survival function for each individual ($\hat{S}_{ic}$) where subscript *c* indicates complete dataset. Using these survival functions for each individual, we calculated survival percent for 12 time points corresponding to 6-month intervals from 6 to 72 months ($\hat{S}_{ic}(t_j)$). For each individual, we compared the survival percent for *j=1, …, 12* (number of timepoints considered) with corresponding estimates from the two models described below:

Standard parametric survival models

We fit multiple parametric survival distributions including exponential, Weibull, log logistic, log normal, gamma, and generalized gamma to all four training datasets. Based on statistical metrics of goodness-of-fit (i.e., AIC), the best parametric fit was selected for each training dataset.[1] For each selected model, we calculate percent survival for each of the 12 time points. Since this model only generates population-level survival percent at the arm level (i.e., Dacarbazine + 10 mg/kg of Ipilimumab versus Dacarbazine + placebo), we

assume the same estimate applies to each subject in that arm. In other words, every individual in each arm is assumed to have the same survival percent at each time point as the arm-level estimate ($\tilde{S}(t_j)$).

KWSF survival extrapolation

We use the KWSF model (as described above) to estimate percent survival for the same 12 time points for each trial subject ($\overline{S_i}(t_j)$). Of note, the same parametric models as above were fitted to the individual-level survival functions (i.e., the forest-level survival function for an individual) and the distribution with the lowest AIC was selected for survival percent estimations for that individual. This process allows for potentially different distributions to be selected for different subjects within a given arm.

Comparison of Predictive Performance

We assess predictive performance of each model by estimating mean squared error. We compare the predicted survival percent at each time point from the standard parametric survival models (MSE_1) and KWSF survival extrapolation (MSE_2) with the corresponding "true" survival percent estimates ($\hat{S}_{ic}(t_j)$). For *i=1, …, n* (number of subjects in each training dataset and *j=1, …, 12* (number of timepoints), we calculated MSE_1 and MSE_2 for each of the four training datasets using the below equations[12,27]:

$$MSE_{1j} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{S}_{ic}(t_j) - \tilde{S}(t_j)\right)^2$$

$$MSE_{2j} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{S}_{ic}(t_j) - \bar{S}_i(t_j)\right)^2$$

All analyses were run using R version 4.0.2.

Results

The KM curves using the complete dataset are presented in Figure 2.3, where the dashed vertical lines indicate 24-month and 36-month longitudinally partitioned data. For the complete dataset, median survival was 11.17 and 9.07 months for patients who received

immunotherapy and patients who received chemotherapy, respectively. For the 2-year training dataset, 2-year survival was 28.78% and 17.77% for the immunotherapy and chemotherapy arms, respectively. Similarly, for the 3-year training dataset, 3-year survival was 21.17% and 12.12% for patients who received immunotherapy and patient who received chemotherapy, respectively.

The estimated MSE for the selected models using 2-year and 3-year partitioned data for immunotherapy and chemotherapy arms are presented in Table 2.1. As expected, using the training dataset with longer follow-up period improves the prediction performance, as indicated by the lower MSE estimates in 3-year datasets compared to 2-year datasets across both arms. The KSWF model consistently outperforms the standard parametric model across all time points for both study arms, regardless of the duration of follow up. The sum of MSEs for survival projections beyond 30 months are 1.88% (using 2-year dataset) and 1.21% (using 3-year dataset) when estimated using the KWSF model in the chemotherapy arm, while the corresponding numbers for the immunotherapy arm are 4.41% (using 2-year dataset) and 3.40% (using 3-year dataset). Similarly, the estimated sum of MSEs across all time points in the chemotherapy arm is lower than the corresponding estimate for the immunotherapy arm, which signifies the challenges of extrapolating survival in the Immunotherapy arm of the trial.

Figure 2.4 presents the estimated MSE and the associated error bars by trial arm and estimation method for 2 and 3-year partitioned datasets. Although the MSE is lower for KWSF method for both arms across all timepoints, the difference in predictive performance of the two models is more pronounced when the datasets with longer follow-up (3-year) are used, which might indicate that KWSF method is associated with more efficient use of additional data. Further, Figures 2.4C and 2.4D show a monotonous increase in MSE estimates after 30 months in the immunotherapy arm, while the corresponding rate of MSE increase is not as notable in the chemotherapy arm. This finding is consistent with the notion

that extrapolating survival for time points that are further away is more challenging in the immunotherapy arm.

Discussion

This tutorial presents a novel survival extrapolation method that can accommodate potential individual-level treatment response heterogeneity and survival dynamics of checkpoint immunotherapy treatments. The implementation of the proposed method on data from an immunotherapy clinical trial indicates that compared to standard parametric models, KWSF can more accurately predict survival beyond the available trial follow up.

Currently more than 1,000 clinical trials are being conducted on the use of checkpoint immunotherapy drugs for numerous cancers and many of these trials have overall survival as the primary endpoint.[28-30] Methods that can accurately estimate long-term survival impact earlier in the trial are necessary as such trials typically face data limitations due to time and/or budget constraints, which can prohibit longer follow-up durations. Despite the fact that the treatment regimens tested in the illustrative example of this tutorial are no longer considered as the standard of care for advanced melanoma, we believe the proposed extrapolation method can be used as a prototype for any randomized controlled trial of checkpoint immunotherapy or similar treatments in the context of a limited follow-up and known heterogeneity of treatment response.

For the illustrative example, we selected the best fitting standard parametric models as the KWSF's comparator, because such models are commonly used in extrapolating survival in the context of checkpoint immunotherapy drugs. However, other methods such as piecewise or spline-based models have been suggested for extrapolating survival beyond trial follow-up in this context.[31] It is worth noting that these models can be similarly incorporated in the KWSF model and might further reduce prediction errors.

In addition to the piecewise models, other extrapolation models have been suggested that capture heterogeneity by partitioning the trial population into more

homogenous groups. A number of such models that have been suggested for checkpoint immunotherapy drugs include cure fraction models, parametric mixture models, and landmark models.[32,33] Although these models can account for the complexities of the hazard functions associated with checkpoint immunotherapy drugs, they are not designed to capture potential individual-level heterogeneity. The survival forest model presented in this tutorial allows for estimating a non-parametric individual-level survival function by leveraging patient-level characteristics.[13] The survival forest combined with parametric extrapolation allows for estimating an individual-level survival function that can predict survival beyond the available follow up. We believe this individual-level extrapolation is an effective way to model the heterogeneity of treatment effects in checkpoint immunotherapy.

Although KWSF shows reduction in prediction error (MSE) across all time points for both trial arms, the MSE of KWSF method seems to be getting closer to the MSE of standard parametric extrapolation for time points that are further away in the future particularly in the immunotherapy arm. We believe this trend happens as a result of the parametric extrapolation that was used in KWSF. Incorporating other survival extrapolation models such as cure fraction models might further improve the prediction accuracy of KWSF. Further, limited follow up from the trial and lack of external data (e.g., real-world data from registries) makes it difficult to validate the results of the extrapolation beyond the available trial data.[33] This limitation is particularly important for recently approved immunotherapy drugs, where real-world data have not accumulated. In addition, all the calculations in the proposed method are based on the original randomly assigned interventions, and KWSF does not explicitly model potential survival impacts of the second- and third-line treatments.

Although the proposed method does not require the use of any external data, access to patient-level trial data is necessary to develop the survival forest, which can be a limitation. However, we believe with the recent trend towards improved data sharing and trial

transparency, patient-level data from clinical trials will continue to become more accessible in the future, making the proposed method more feasible.

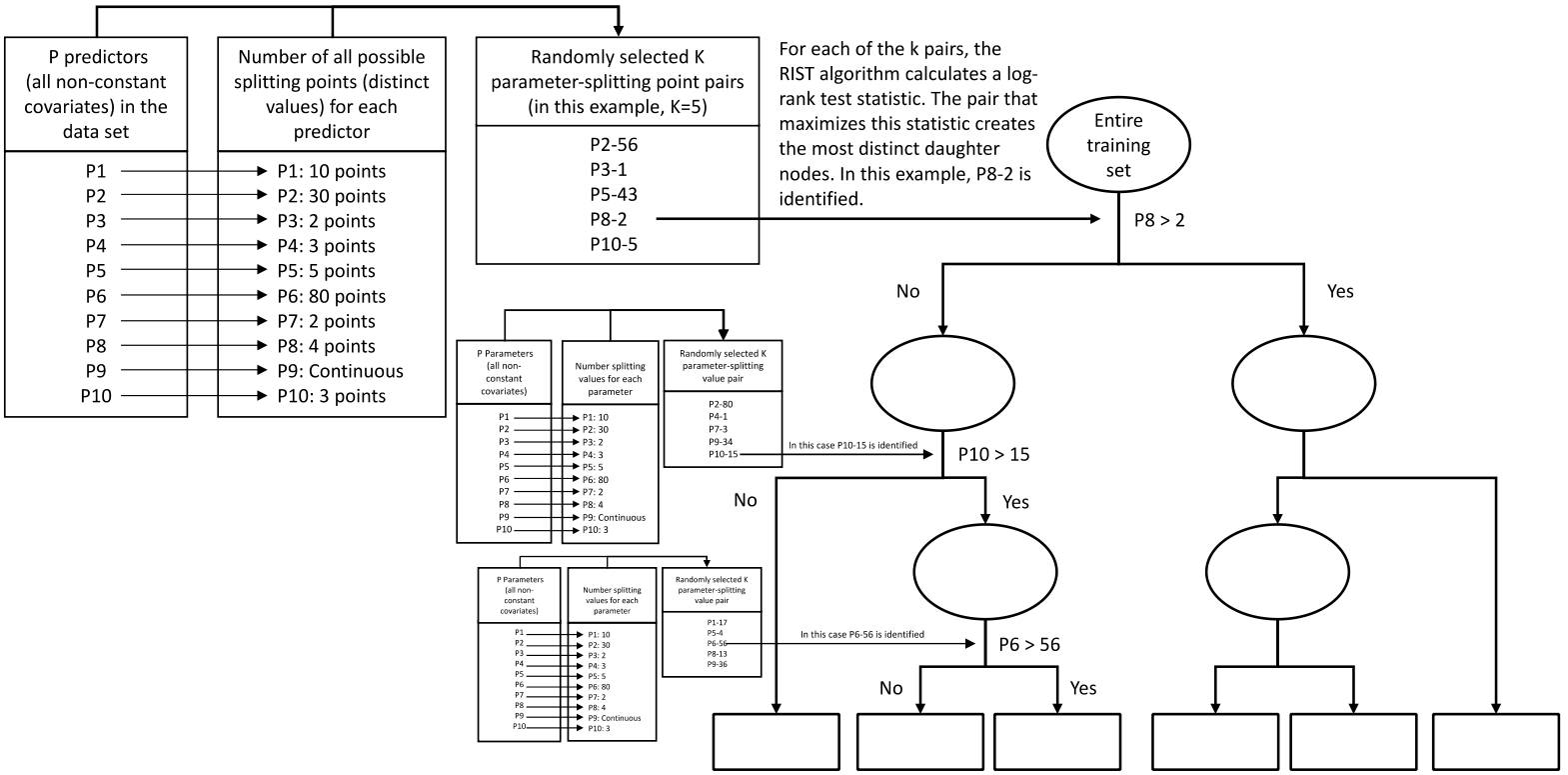
This tutorial introduces the KWSF model as a novel survival extrapolation method that uses patient-level characteristics to estimate individualized survival function, which can then be used for individual-level survival extrapolation. The KWSF algorithm, as described above, can be used to develop an application that inputs the characteristics of a given patient (outside the clinical trial) who received similar interventions to estimate their individualized survival function. Such application can help develop microsimulation models for economic evaluation of checkpoint immunotherapy drugs as well as informing the estimation of individualized treatment rules. Future studies are needed to further evaluate the performance of the KWSF models in predicting survival beyond trial follow up.

Table 2.1. Percent MSE using 2-year and 3-year partitioned data, by arm and estimation methods

	Chemotherapy 2-year follow up		Chemotherapy 3-year follow up		Immunotherapy 2-year follow up		Immunotherapy 3-year follow up	
Time Periods (months)	**KWSF**	**SPM***	**KWSF**	**SPM***	**KWSF**	**SPM****	**KWSF**	**SPM****
6	0.19%	0.27%	0.16%	0.27%	0.08%	0.25%	0.11%	0.23%
12	0.19%	0.31%	0.24%	0.33%	0.19%	0.26%	0.18%	0.31%
18	0.13%	0.27%	0.16%	0.27%	0.12%	0.23%	0.11%	0.27%
24	0.10%	0.24%	0.09%	0.23%	0.09%	0.16%	0.07%	0.18%
30	0.17%	0.26%	0.09%	0.23%	0.10%	0.14%	0.05%	0.13%
36	0.21%	0.30%	0.09%	0.28%	0.14%	0.18%	0.07%	0.14%
42	0.25%	0.33%	0.13%	0.29%	0.24%	0.27%	0.15%	0.19%
48	0.25%	0.32%	0.14%	0.29%	0.42%	0.46%	0.29%	0.33%
54	0.29%	0.35%	0.18%	0.32%	0.63%	0.66%	0.47%	0.50%
60	0.36%	0.41%	0.24%	0.38%	0.82%	0.87%	0.64%	0.69%
66	0.34%	0.38%	0.24%	0.35%	1.04%	1.11%	0.84%	0.91%
72	0.39%	0.42%	0.28%	0.39%	1.25%	1.33%	1.02%	1.13%
Total after 3 years	1.88%	2.21%	1.21%	2.02%	4.41%	4.69%	3.40%	3.75%
Total	2.88%	3.87%	2.02%	3.62%	5.13%	5.91%	3.99%	5.01%

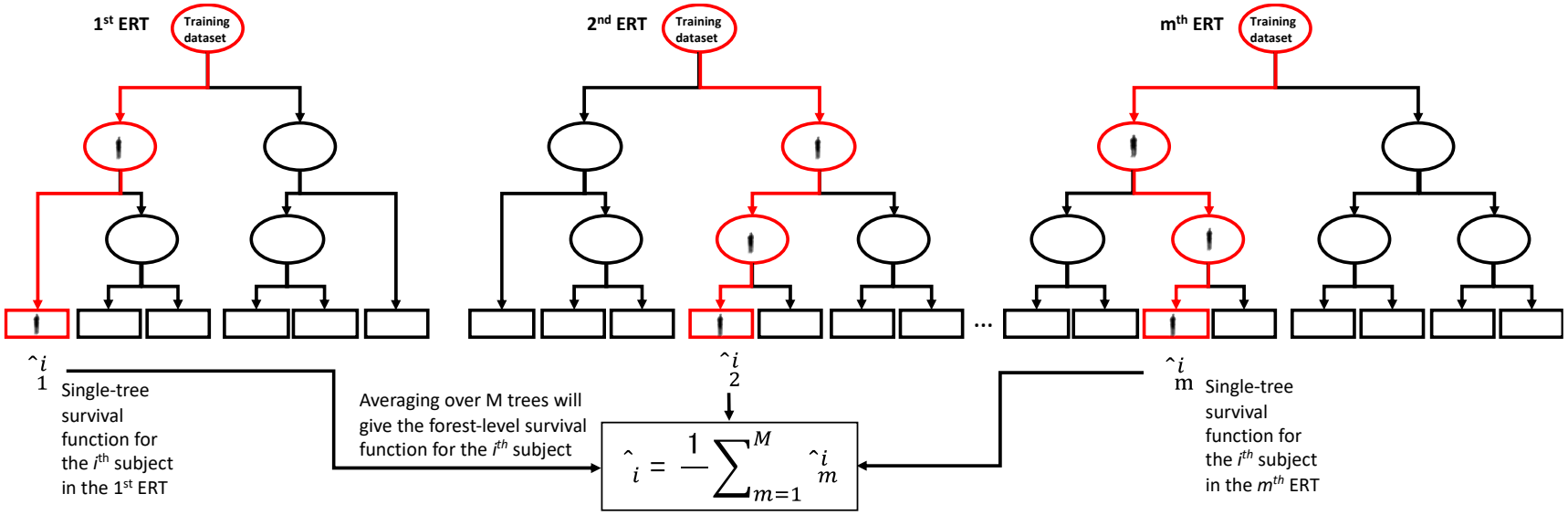
MSE: mean squared error. KWSF: kernel-weighted survival forest. SPM: standard parametric model
* Log normal distribution had the lowest AIC and was used for this estimation
** Log logistic distribution had the lowest AIC and was used for this estimation

Figure 2.1. Fitting an extremely randomized tree (ERT)



The tree fitting process starts with a training dataset, which includes P predictors (P=10 in this example). Each predictor has a number of distinct values that can be used to split the predictor space (For example, P1 has 10 distinct values (splitting points) and P9 is a continuous variable). The algorithm randomly selects K predictor-splitting point pairs (K=5 in this example) and tries each pair in the split, evaluating each based on a goodness-of-split criterion (in this example log rank test statistic). In this example, predictor P8 and splitting point 2 is identified as the best split in the first split of the tree. Similar splitting process is repeated for each daughter node until no further splitting can be done without a node having fewer than nmin events, at which point daughter nodes are referred to as terminal nodes, depicted as rectangles at the bottom of the tree.

Figure 2.2. Fitting M independently generated extremely randomized trees



The trees start with the entire training dataset (same dataset across all trees). For each tree, a particular subject will end up in a single terminal node, for example the path for the ith subject in the 1st, 2nd, …, and the mth ERT are depicted in red. Each terminal node contains a predetermined minimum number of events (nmin). Note that the shape of the trees can be different due to the randomization process used in fitting them. Once the survival function for the ith subject is estimated at node level within each tree ($\hat{S}_m^i$), the algorithm averages the survival function over all trees and estimates the pooled survival function ($\hat{S}_i$).

Figure 2.3. KM curves for the complete dataset

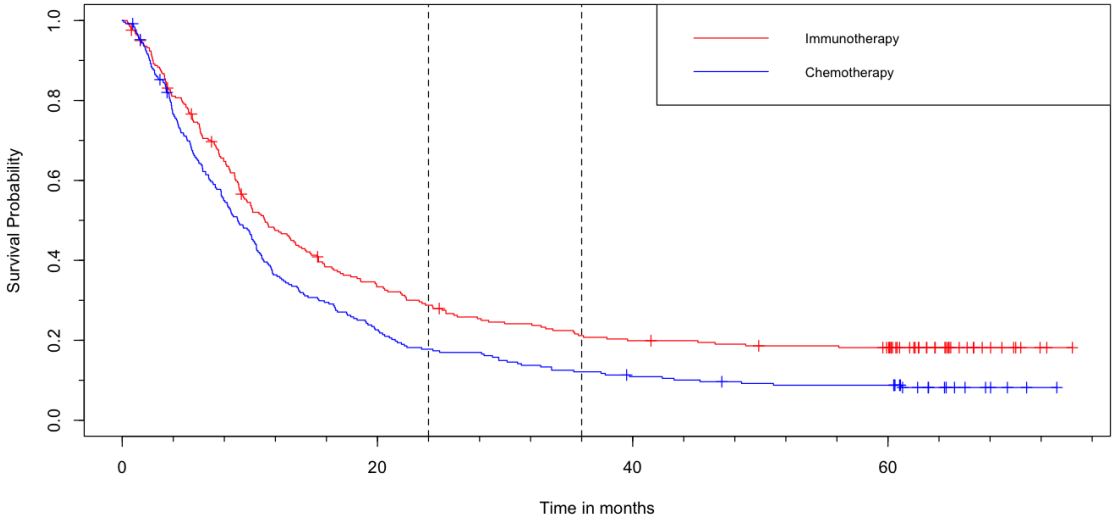
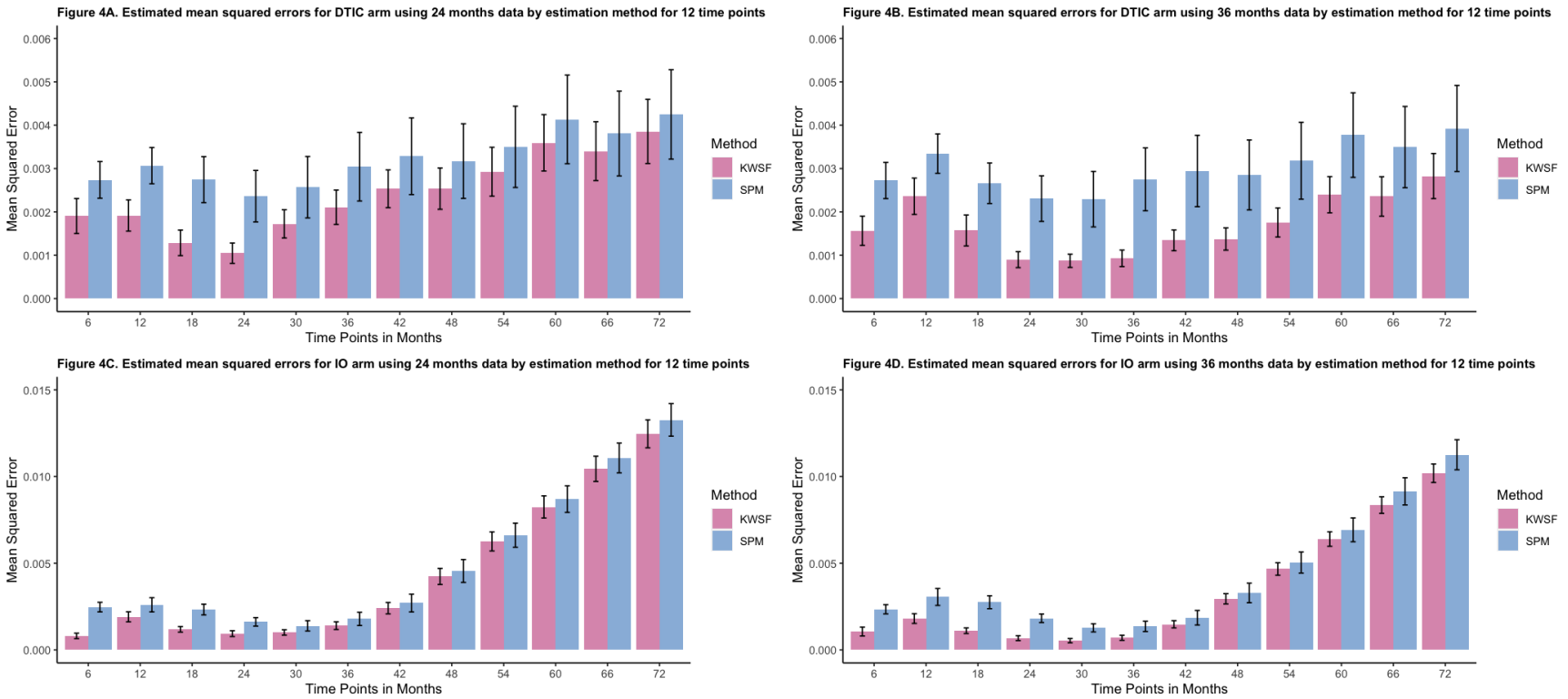


Figure 2.4. Estimated MSE by trial arm and estimation method for 2 and 3-year partitioned datasets



MSE: mean squared errors. DTIC: dacarbazine. IO: immune-oncology (i.e., immunotherapy). KWSF: kernel-weighted survival forest. SPM: standard parametric model.

REFERENCES

1. Latimer NR. Survival analysis for economic evaluations alongside clinical trials--extrapolation with patient-level data: inconsistencies, limitations, and a practical guide. Med Decis Mak Int J Soc Med Decis Mak. 2013 Aug;33(6):743–54.

2. Azoury SC, Straughan DM, Shukla V. Immune Checkpoint Inhibitors for Cancer Therapy: Clinical Efficacy and Safety. Curr Cancer Drug Targets. 2015;15(6):452–62.

3. He X, Xu C. Immune checkpoint signaling and cancer immunotherapy. Cell Res. 2020 Aug;30(8):660–9.

4. Gemmen E, Parmenter L. Special Considerations for the Analysis of Patient-Level Immuno-Oncology Data. 2018;4(1):23–4.

5. Guyot P, Ades AE, Beasley M, Lueza B, Pignon JP, Welton NJ. Extrapolation of Survival Curves from Cancer Trials Using External Information. Med Decis Mak Int J Soc Med Decis Mak. 2017 May;37(4):353–66.

6. Kim H, Goodall S, Liew D. Health Technology Assessment Challenges in Oncology: 20 Years of Value in Health. Value Health J Int Soc Pharmacoeconomics Outcomes Res. 2019 May;22(5):593–600.

7. Huang M, Latimer N, Zhang Y, Mukhopadhyay P, Ouwens M, Briggs A. Estimating the Long-term outcomes associated with Immuno-oncology therapies: challenges and approaches for overall survival extrapolations. Value Outcomes Spotlight. 2018;4(1):28–30.

8. Connock M, Hyde C, Moore D. Cautions regarding the fitting and interpretation of survival curves: examples from NICE single technology appraisals of drugs for cancer. PharmacoEconomics. 2011 Oct;29(10):827–37.

9. Hawkins N, Grieve R. Extrapolation of Survival Data in Cost-effectiveness Analyses: The Need for Causal Clarity. Med Decis Mak Int J Soc Med Decis Mak. 2017 May;37(4):337–9.

10. Bagust A, Beale S. Survival analysis and extrapolation modeling of time-to-event clinical trial data for economic evaluation: an alternative approach. Med Decis Mak Int J Soc Med Decis Mak. 2014 Apr;34(3):343–51.

11. Bou-Hamad I, Larocque D, Ben-Ameur H. A review of survival trees. Stat Surv. 2011;5:44–71.

12. Cui Y, Zhu R, Zhou M, Kosorok M. Consistency of survival tree and forest models: splitting bias and correction. ArXiv Prepr ArXiv170709631. 2017;

13. Zhu R, Kosorok MR. Recursively Imputed Survival Trees. J Am Stat Assoc. 2012;107(497):331–40.

14. Breiman L, Friedman JH, Stone CJ, Olshen RA. Classification and Regression Trees. New York; 1984. 368 p.

15. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning. Vol. 112. Springer; 2013.

16. Liu H, Cocea M. Semi-random partitioning of data into training and test sets in granular computing context. Granul Comput. 2017;2(4):357–86.

17. Hastie T, Tibshirani R, Friedman J. Random Forests. In: The Elements of Statistical Learning. 2nd ed. 2009. p. 587–604.

18. Geurts P, Ernst D, Wehenkel L. Extremely Randomized Trees. Mach Learn. 2006;36:3–42.

19. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.

20. Song X, Zhao Z, Barber B, Farr AM, Ivanov B, Novich M. Overall survival in patients with metastatic melanoma. Curr Med Res Opin. 2015 May;31(5):987–91.

21. Sharma P, Wagner K, Wolchok JD, Allison JP. Novel cancer immunotherapy agents with survival benefit: recent successes and next steps. Nat Rev Cancer. 2011 Oct 24;11(11):805–12.

22. Woo SR, Turnis ME, Goldberg MV, Bankoti J, Selby M, Nirschl CJ, et al. Immune inhibitory molecules LAG-3 and PD-1 synergistically regulate T-cell function to promote tumoral immune escape. Cancer Res. 2012 Feb 15;72(4):917–27.

23. Robert C, Thomas L, Bondarenko I, O'Day S, Weber J, Garbe C, et al. Ipilimumab plus dacarbazine for previously untreated metastatic melanoma. N Engl J Med. 2011 Jun 30;364(26):2517–26.

24. Maio M, Grob JJ, Aamdal S, Bondarenko I, Robert C, Thomas L, et al. Five-year survival rates for treatment-naive patients with advanced melanoma who received ipilimumab plus dacarbazine in a phase III trial. J Clin Oncol Off J Am Soc Clin Oncol. 2015 Apr 1;33(10):1191–6.

25. Wolchok JD, Hodi FS, Weber JS, Allison JP, Urba WJ, Robert C, et al. Development of ipilimumab: a novel immunotherapeutic approach for the treatment of advanced melanoma. Ann N Y Acad Sci. 2013 Jul;1291:1–13.

26. Alexander W. The Checkpoint Immunotherapy Revolution: What Started as a Trickle Has Become a Flood, Despite Some Daunting Adverse Effects; New Drugs, Indications, and Combinations Continue to Emerge. P T Peer-Rev J Formul Manag. 2016 Mar;41(3):185–91.

27. Cui Y, Hannig J. Nonparametric generalized fiducial inference for survival functions under censoring. Biometrika. 2019 Sep 1;106(3):501-18.

28. Franklin MR, Platero S, Saini KS, Curigliano G, Anderson S. Immuno-oncology trends: preclinical models, biomarkers, and clinical development. Journal for Immunotherapy of Cancer. 2022;10(1).

29. Upadhaya S, Neftelinov ST, Hodge J, Campbell J. Challenges and opportunities in the PD1/PDL1 inhibitor clinical trial landscape. Nat Rev Drug Discov. 2022 Feb 10;10.

30. Baik CS, Rubin EH, Forde PM, Mehnert JM, Collyar D, Butler MO, Dixon EL, Chow LQ. Immuno-oncology clinical trial design: limitations, challenges, and opportunities. Clinical Cancer Research. 2017 Sep 1;23(17):4992-5002.

31. Gibson E, Koblbauer I, Begum N, Dranitsaris G, Liew D, McEwan P, Tahami Monfared AA, Yuan Y, Juarez-Garcia A, Tyas D, Lees M. Modelling the survival outcomes of immuno-oncology drugs in economic evaluations: a systematic approach to data analysis and extrapolation. Pharmacoeconomics. 2017 Dec;35(12):1257-70.

32. Bullement A, Meng Y, Cooper M, Lee D, Harding TL, O'Regan C, Aguiar-Ibanez R. A review and validation of overall survival extrapolation in health technology assessments of cancer immunotherapy by the National Institute for Health and Care Excellence: how did the initial best estimate compare to trial data subsequently made available?. Journal of Medical Economics. 2019 Mar 4;22(3):205-14.

33. Bullement A, Latimer NR, Gorrod HB. Survival extrapolation in cancer immunotherapy: a validation-based case study. Value in Health. 2019 Mar 1;22(3):276-83.

CHAPTER 3: COMPARING THE ACCURACY OF KERNEL-WEIGHTED SURVIVAL FOREST AND EXTRAPOLATION METHODS THAT DIRECTLY MODEL HETEROGENEITY IN PREDICTING SURVIVAL USING DATA FROM A CHECKPOINT IMMUNOTHERAPY TRIAL

Introduction

Checkpoint immunotherapy drugs are approved or under investigation in many cancers.[1] Currently hundreds of clinical trials are being conducted on the use of checkpoint immunotherapy drugs for numerous indications and many of these trials have overall survival (OS) as their primary endpoint.[2-4] Methods that can accurately estimate long-term survival impact earlier in the trial are necessary as such trials typically face data limitations due to time and/or budget constraints, which can prohibit longer follow-up durations.[5]

Checkpoint immunotherapy drugs typically trigger a durable response in a subset of patients, which translates to long-term survival for some – but not all – patients.[6] This heterogenous survival effect for a subset of patients makes extrapolation beyond trial follow-up particularly challenging for checkpoint immunotherapy interventions, as such extrapolation models are required to accommodate the complex survival dynamics of these treatment.[7,8]

Numerous methods have been suggested and used to extrapolate survival beyond clinical trial follow-up for checkpoint immunotherapy drugs.[8-10] These extrapolation methods can be categorized into three main groups: (1) Models that are built based on an entire survival curve as one unit; (2) Models that are built based on partitioning the survival curve into different time periods and modeling each period separately; (3) Models that are built based on partitioning the trial population into more homogenous groups and separately extrapolating the survival for each group.

Models built based on entire survival curve as one unit include standard parametric survival models and more flexible models such as fractional polynomials that can model complex hazard functions but still use the entire trial survival curve as one unit. Such methods are more flexible than standard parametric models and thus more capable of capturing complex hazard functions associated with checkpoint immunotherapy's unique survival dynamics.[11]

Models built based on partitioning the survival curve into pieces include piecewise and spline-based (e.g., restricted cubic splines) models. Such models offer higher degree of flexibility and are suitable to perform extrapolation when hazard rates are not constant over time.[9,10] Piecewise and spline-based models have been commonly used for extrapolating survival beyond trial follow-up in the checkpoint immunotherapy trials.[10]

Models built based on segmenting the trial population into homogenous groups assume that the trial population is composed of groups that respond differently to treatments and may have distinct survival curves, hence require different modeling techniques to produce an unbiased OS estimation.[12] Cure fraction models and landmark models are examples of this category that have been suggested and used for checkpoint immunotherapy drugs.[12,13] Although these methods offer more flexibility, they typically require longer follow-up to detect populations with markedly different survival. Further, such models are designed to extrapolate survival at population level or for segments of the population.[12]

Given the level of heterogeneity associated with checkpoint immunotherapy treatments, we believe models that are capable of flexibly estimating individual-level survival functions can provide an effective way to model the heterogeneity of treatment response in survival extrapolation for checkpoint immunotherapy recipients. Such models can inform precision and personalized medicine strategies, and serve as the foundation for individual-level simulation of these strategies' relative economic value and efficiency. The Aim 1

tutorial presented a novel method to estimate individual-level survival functions that can predict survival beyond trial follow up for right-censored observations. This novel model leverages survival forest based on extremely randomized trees[14] and kernel-weighted parametric extrapolation, together called kernel-weighted survival forest (KWSF).

The group 3 models explained above can directly model the heterogeneity of treatment response and have shown favorable accuracy in extrapolating survival for checkpoint immunotherapy interventions.[9,13] Since the KWSF model has the capability to utilize different extrapolation functions, we hypothesized that using similarly flexible extrapolation functions such as cure fraction models can further improve the prediction accuracy of the KWSF model. However, it is not clear whether a KWSF model that uses cure fraction survival function confers similar or greater accuracy than population-based cure fraction models in estimating individual-level long-term survival effects.

The aim of this study is to compare the accuracy of survival extrapolation models that are designed to directly model heterogeneity of treatment response (e.g., cure fraction models and response-based/landmark models) with the accuracy of the KWSF model that utilizes cure fraction models at individual level. Models are trained using the 2- and 3-year patient-level data from a checkpoint immunotherapy trial in patients with advanced melanoma, and predicted survival estimates from different models are compared with corresponding estimated true survival using maximum follow up data available.

Methods

Dataset

For this study we used individual-level data from the CA184-024 trial. This trial is “A Multi-Center, Randomized, Double-Blind, Two-Arm, Phase III Study in Patients with Untreated Stage III (Unresectable) or IV Melanoma Receiving Dacarbazine Plus 10 mg/kg of Ipilimumab vs. Dacarbazine with Placebo”. A total of 502 subjects were randomized to ipilimumab plus DTIC (n=250) and to DTIC monotherapy (n=252).[15] We identified 18

predictive variables based on the literature and data availability; predictors with missing values were excluded from the analysis. All analyses were conducted based on an intention-to-treat framework i.e., trial data were analyzed assuming that subjects received the randomly assigned treatment.

For each arm of the CA184-024 trial, we constructed two longitudinally-partitioned subsets of the data that include 2-year and 3-year follow up (the training datasets). The two and three years of follow up were selected because trial data are typically reported within these time frames for regulatory submission and other applications. The partitioning process has been previously described in more detail in the Aim 1 paper.

Candidate Models

Population-level models

Cure fraction models

Cure fraction models assume that a fraction of the population will be "cured" and thus the survival curve will eventually reach a plateau. In the context of cancer trials, by definition, cure happens when the hazard rate of death in the cancer patients returns to the same level as that expected in the general population.[16] Parametric cure models can be used to estimate the cure fraction, modeling 'cured' and 'uncured' with different distributions.[13] The most popular framework for cure models is to assume that the study population is a mixture of patients who are cured and patients who are not cured and to explicitly model this mixture (cure fraction mixture models).[13] In a mixture cure model, these 'cured' and 'uncured' subjects are modeled separately, with the cured individuals subject to no excess risk and the uncured individuals subject to excess risk modeled using a parametric survival distribution.[16] In a non-mixture model, a parametric survival distribution is scaled in a way that survival asymptotically approaches the cure fraction.[16,17] For this study we used Weibull distribution for both mixture and non-mixture cure models.

Landmark models

In these models, patients are split into response groups, based on their status at a pre-specified time point (landmark).[12] Standard parametric survival models are fitted to extrapolate response-specific OS curves from landmark. These curves are then weighted by the observed response distribution at the landmark. Using the weighted sum of survival curves, the landmark model calculates a single composite curve to extrapolate survival beyond the key trial follow up.[12] The following response categories were included in trial data: (1) Responders defined as patients who have complete or partial response; (2) Stable disease defined as patients who remain progression free 3 months or more from the start of treatment; (3) Progressive disease defined as patients who progress or are censored prior to 3 months. For each response group a number of standard parametric distributions (similar to standard parametric model estimation below) were tested and the best fitting distribution (i.e., lowest AIC) was selected.

Standard parametric models

Similar to Aim 1 analysis, we included standard parametric models as a baseline comparison because such models are commonly used in extrapolating survival in the context of checkpoint immunotherapy drugs. Briefly, we fit multiple parametric survival distributions including exponential, Weibull, log logistic, log normal, gamma, and generalized gamma to all four training datasets. Based on statistical metrics of goodness-of-fit (i.e., AIC), the best parametric fit was selected for each training data set.[7]

Individual-level models

KWSF using standard parametric extrapolation function

We use the KWSF model (as described in the Aim 1 paper) to estimate percent survival for each trial subject. Of note, the same parametric models as above were fitted to the individual-level survival functions (i.e., the forest-level survival function for an individual) and the distribution with the lowest AIC was selected for survival percent estimations for that

individual. This process allows for potentially different distributions to be selected for different subjects within a given arm.

KWSF using cure fraction extrapolation function

We implemented variations of the KWSF model described in Aim 1 with mixture and non-mixture cure models as the extrapolation functions. Of note, the same mixture and non-mixture cure models as described in population-level models above (i.e., parametric cure model with Weibull distribution) were fit to the individual-level survival functions.

Comparison of Predictive Accuracy

As described in Aim 1, we estimated the "true" individual-level survival functions by applying the survival forest algorithm using data from the full duration of available follow-up. The survival forest provides a nonparametric estimate of the survival function for each individual ($\hat{S}_{ic}$) where subscript *c* indicates complete data set. Using these survival functions for each individual, we calculated survival percent for 12 time points corresponding to 6-month intervals from 6 to 72 months ($\hat{S}_{ic}(t_j)$).

For each selected population-level model above, we calculated percent survival for each of the above 12 time points. Since these models only generate population-level survival percent at the trial arm level, we assumed the same estimate applies to each and every subject in that arm. In other words, every individual in each arm was assumed to have the same survival percent at each time point as the arm-level estimate ($\tilde{S}(t_j)$). For the individual-level models, we calculated percent survival for each subject for the same 12 time points ($\bar{S}_i(t_j)$).

We assess predictive accuracy of each model by estimating mean squared error (MSE). To calculate the MSE, we compared the predicted survival percent at each time point with the corresponding "true" survival percent estimates ($\hat{S}_{ic}(t_j)$). MSE_1 indicates the MSE calculated for the population-level models and MSE_2 indicates the MSE calculated for

individual-level models i.e., KWSF variations (MSE_2). For *i=1, …, n* (number of subjects in each training dataset and *j=1, …, 12* (number of timepoints), we calculated MSE_1 and MSE_2 for each of the four training datasets using the below equations[18,19]:

$$MSE_{1j} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{S}_i(t_j) - \tilde{S}(t_j)\right)^2$$

$$MSE_{2j} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{S}_i(t_j) - \bar{S}_i(t_j)\right)^2$$

All analyses were conducted using R version 4.0.2.

Results

Figure 3.1 shows the distribution of the response status at week 12 after receiving trial treatments for the immunotherapy (Ipilimumab + DTIC) and the chemotherapy (DTIC + placebo) arms. As shown in the figure, the response status for 89 (36%) and 72 (29%) subjects were unknown for the immunotherapy and chemotherapy arms, respectively. Because of the high percentage of unknown responses, landmark/response-based models were not included in model comparison presented in this paper.

The model comparison results are presented as estimated MSEs for 12 time points comparing the predictive accuracy of 3 individual-level models including KWSF with standard parametric extrapolation functions, KWSF with mixture cure model, and KWSF with non-mixture cure model as well as 3 population-level models including standard parametric model, mixture cure model, and non-mixture cure model. Note that the MSEs associated with KWSF model and population-level standard parametric model were previously reported in the Aim 1 paper and are included as a baseline to show potential improvement in the model accuracy when more flexible extrapolation functions are used.

Table 3.1 shows the estimated MSEs associated with the selected models for the immunotherapy arm using 2-year data cut. For this training data set, non-mixture cure models both at individual-level and population-level were associated with lower prediction

errors, with total MSE of 2.48% and 2.24%, respectively. The estimated total MSE after 2 years, representing the MSE of testing data set, indicates that the population-level non-mixture cure model was slightly more accurate than the individual-level non-mixture cure model with estimated MSE of 1.46% and 2.09%, respectively. The estimated MSEs for the immunotherapy arm using 3-year data cut, show that similar to 2-year data cut, both individual- and population-level cure models perform better than models that use standard parametric extrapolation (Table 3.2). Among the selected models, the individual-level mixture cure models had the best accuracy with the estimated total MSE and total MSE after 3 years of 1.07% and 0.56%, respectively.

Tables 3.3 shows the estimated MSEs associated with the selected models for the chemotherapy arm using 2-year data cut. For this training data set, the KWSF model with standard parametric extrapolation function was associated with the lowest total MSE (2.88%) and total MSE after 2 years (2.26%). When using 3-year data cut, the individual-level non-mixture cure model was associated with the lowest MSEs (i.e., highest accuracy) in the chemotherapy arm (Table 3.4).

As expected, the MSE associated with both individual- and population-level cure models decreased when using longer follow up i.e., a 3-year data cut to train the models (Figure 3.2). Additionally, when 3-year training data sets were used, it appears that the MSEs for selected models tend to converge when projecting survival for time points that are further away in the future (Figure 3.2B and 3.2D).

Discussion

Developing and evaluating more flexible and accurate models for survival extrapolation for checkpoint immunotherapy drugs is an active area of research[9,20] and while a selection of these methods is discussed here, this study is not intended to provide a comprehensive review of all suggested methods. We limited the scope of the candidate models to those that directly model the heterogeneity of treatment response by segmenting

the trial population into seemingly more homogenous groups. Cure fraction models and landmark models are two examples of such models that have been used for survival modeling of checkpoint immunotherapy drugs.[12]

Introduced in Aim 1 tutorial, the KWSF model is a novel survival extrapolation method that can accommodate potential individual-level treatment response heterogeneity and unique survival dynamics of checkpoint immunotherapy treatments. The findings from Aim 1 illustrated that the KWSF model that uses standard parametric distributions as the extrapolation function had higher prediction accuracy than standard parametric models. Additionally, the modular feature of KWSF model allows for using a variety of extrapolation functions that are more flexible and may improve the prediction accuracy of the model. This study compared the accuracy of survival extrapolation estimates from cure fraction models with estimates from KWSF with cure fraction extrapolation function.

The results of this study show that compared to models that use standard parametric extrapolation, cure fraction models and KWSF with cure fraction extrapolation function were more accurate in predicting survival in the immunotherapy arm. This finding is aligned with previous literature on cure fraction models for survival extrapolation in checkpoint immunotherapy drugs.[9,13] Our findings also provide further evidence illustrating the utility of cure fraction models for survival extrapolation both at individual and population level. The difference between accuracy of cure models and standard parametric models were less noticeable for the chemotherapy arm, potentially indicating that cure fraction might not be as effective for survival modeling of traditional cancer treatments such as chemotherapy.

Although cure fraction models produced more accurate survival predictions both at population level and when used as an extrapolation function in individual-level models, these models are subject to several limitations. Specifically, applying cure models requires a certain “maturity” of data such that the differences between cured and non-cured subgroups can be identified. However, follow-up times for clinical trials that include checkpoint

immunotherapy treatments are typically insufficient to detect populations with markedly improved outcomes.[13] Similarly, the landmark models assume that patients who respond to treatment are prognostically different from non-responders.[12] However, verifying that the response measure is a reliable predictor of OS can prove difficult particularly for checkpoint immunotherapy.[21]

Although the KWSF model with standard parametric extrapolation function is capable of capturing potential individual-level heterogeneity, our findings illustrated that in the immunotherapy arm, the accuracy of KWSF prediction were further improved when cure fraction models were used as the extrapolation function. That said, the individual-level extrapolation (KWSF + cure fraction extrapolation) did not demonstrate marked improvement in accuracy (as indicated by lower MSEs) compared to the population-level cure models. We believe this finding might be the result of inherent characteristics of the survival forest model used in KWSF. The survival forest algorithm calculates the tree-based survival function for each individual based on limited number of subjects who experienced the event (i.e., death) and share a terminal node with a given subject.[14] Having such small number of events may decrease the effectiveness of cure fraction models in distinguishing cure mixtures for each individual.

Despite the fact that the treatment regimens tested in the clinical trial used for this paper are no longer considered as the standard of care for advanced melanoma,[22-24] we used this trial because it offers a relatively long follow-up period compared to most checkpoint immunotherapy trials.[25,26] Such data allow for assessing the prediction accuracy of proposed extrapolation models using varied amounts of follow-up duration (i.e., two and three years). Additionally, the selection of the 12 timepoints allows for more granular comparison of the model accuracy and makes it possible to demonstrate the longitudinal change in prediction accuracy for the time points that are increasingly further away from the available follow up in the training data sets.

External data with a longer follow up would allow to validate the results of the survival extrapolation models further in the future. However, such data are not available for recently approved immunotherapy drugs. In addition, the possibility of subsequent treatments for patients who did not respond to the first-line immunotherapy treatments makes it more challenging to model survival further in the future as the KWSF model is not designed to explicitly model potential survival impacts of the subsequent treatments. Additionally, none of the variations of KWSF require the use of any external data; however, access to patient-level trial data is necessary to develop the survival forest, which can be a limitation. We believe with the recent trend towards improved data sharing and trial transparency, patient-level data from clinical trials will continue to become more accessible in the future, making the implementation of the proposed methods more feasible.

Although cure fraction models demonstrated reasonably accurate survival predictions for the immunotherapy arm, they are not designed to generate individual-level extrapolated survival functions. The KWSF model with a cure fraction survival extrapolation function demonstrated comparable accuracy with cure fraction models, while uniquely allowing for estimating individual-level survival functions that can be used to inform precision and personalized medicine strategies, and serve as the foundation for individual-level simulation of checkpoint immunotherapy drugs' relative economic value and efficiency. Future studies are needed to further evaluate the accuracy of the KWSF models and its variations in predicting survival beyond trial follow up.

Table 3.1. MSE using 2-year partitioned data in immunotherapy arm by estimation methods

	Estimation Method					
Time Period (months)	**KWSF**	**SPM***	**P-MCM**	**I-MCM**	**P-NMCM**	**I-NMCM**
6	0.08%	0.25%	0.16%	0.06%	0.17%	0.07%
12	0.19%	0.26%	0.23%	0.14%	0.23%	0.13%
18	0.12%	0.23%	0.22%	0.09%	0.22%	0.10%
24	0.09%	0.16%	0.16%	0.09%	0.16%	0.09%
30	0.10%	0.14%	0.16%	0.14%	0.14%	0.11%
36	0.14%	0.18%	0.26%	0.28%	0.18%	0.19%
42	0.24%	0.27%	0.33%	0.37%	0.19%	0.24%
48	0.42%	0.46%	0.37%	0.43%	0.19%	0.28%
54	0.63%	0.66%	0.39%	0.46%	0.19%	0.30%
60	0.82%	0.87%	0.40%	0.48%	0.19%	0.32%
66	1.04%	1.11%	0.40%	0.48%	0.19%	0.32%
72	1.25%	1.33%	0.40%	0.48%	0.19%	0.33%
Total	5.13%	5.91%	3.49%	3.50%	2.24%	2.48%
Total after 2 years	4.64%	5.02%	2.71%	3.12%	1.46%	2.09%

MSE: mean squared error. KWSF: kernel-weighted survival forest. SPM: standard parametric model. P-MCM: population-level mixture cure model. I-MCM: individual-level mixture cure model. P-NMCM: population-level non-mixture cure model. I-NMCM: individual-level non-mixture cure model.

* Log logistic distribution had the lowest AIC and was used for this estimation

Table 3.2. MSE using 3-year partitioned data in immunotherapy arm by estimation methods

	Estimation Method					
Time Period (months)	**KWSF**	**SPM***	**P-MCM**	**I-MCM**	**P-NMCM**	**I-NMCM**
6	0.11%	0.23%	0.16%	0.06%	0.16%	0.06%
12	0.18%	0.31%	0.29%	0.20%	0.27%	0.14%
18	0.11%	0.27%	0.22%	0.10%	0.22%	0.09%
24	0.07%	0.18%	0.17%	0.07%	0.16%	0.06%
30	0.05%	0.13%	0.13%	0.05%	0.13%	0.05%
36	0.07%	0.14%	0.13%	0.04%	0.13%	0.04%
42	0.15%	0.19%	0.11%	0.06%	0.11%	0.06%
48	0.29%	0.33%	0.11%	0.08%	0.11%	0.08%
54	0.47%	0.50%	0.10%	0.09%	0.10%	0.11%
60	0.64%	0.69%	0.10%	0.10%	0.10%	0.12%
66	0.84%	0.91%	0.10%	0.11%	0.10%	0.14%
72	1.02%	1.13%	0.10%	0.11%	0.10%	0.15%
Total	3.99%	5.01%	1.71%	1.07%	1.69%	1.12%
Total after 3 years	3.40%	3.75%	0.62%	0.56%	0.62%	0.66%

MSE: mean squared error. KWSF: kernel-weighted survival forest. SPM: standard parametric model. P-MCM: population-level mixture cure model. I-MCM: individual-level mixture cure model. P-NMCM: population-level non-mixture cure model. I-NMCM: individual-level non-mixture cure model.

* Log logistic distribution had the lowest AIC and was used for this estimation

Table 3.3. MSE using 2-year partitioned data in chemotherapy arm by estimation methods

	Estimation Method					
Time Period (months)	**KWSF**	**SPM***	**P-MCM**	**I-MCM**	**P-NMCM**	**I-NMCM**
6	0.19%	0.27%	0.32%	0.27%	0.29%	0.23%
12	0.19%	0.31%	0.31%	0.18%	0.30%	0.16%
18	0.13%	0.27%	0.29%	0.13%	0.29%	0.13%
24	0.10%	0.24%	0.23%	0.09%	0.23%	0.09%
30	0.17%	0.26%	0.21%	0.15%	0.21%	0.14%
36	0.21%	0.30%	0.29%	0.25%	0.25%	0.20%
42	0.25%	0.33%	0.34%	0.31%	0.25%	0.24%
48	0.25%	0.32%	0.41%	0.41%	0.26%	0.30%
54	0.29%	0.35%	0.50%	0.48%	0.31%	0.34%
60	0.36%	0.41%	0.50%	0.48%	0.30%	0.35%
66	0.34%	0.38%	0.58%	0.59%	0.34%	0.43%
72	0.39%	0.42%	0.58%	0.59%	0.34%	0.43%
Total	2.88%	3.87%	4.57%	3.93%	3.36%	3.04%
Total after 2 years	2.26%	2.77%	3.41%	3.26%	2.26%	2.43%

MSE: mean squared error. KWSF: kernel-weighted survival forest. SPM: standard parametric model. P-MCM: population-level mixture cure model. I-MCM: individual-level mixture cure model. P-NMCM: population-level non-mixture cure model. I-NMCM: individual-level non-mixture cure model.

* Log normal distribution had the lowest AIC and was used for this estimation

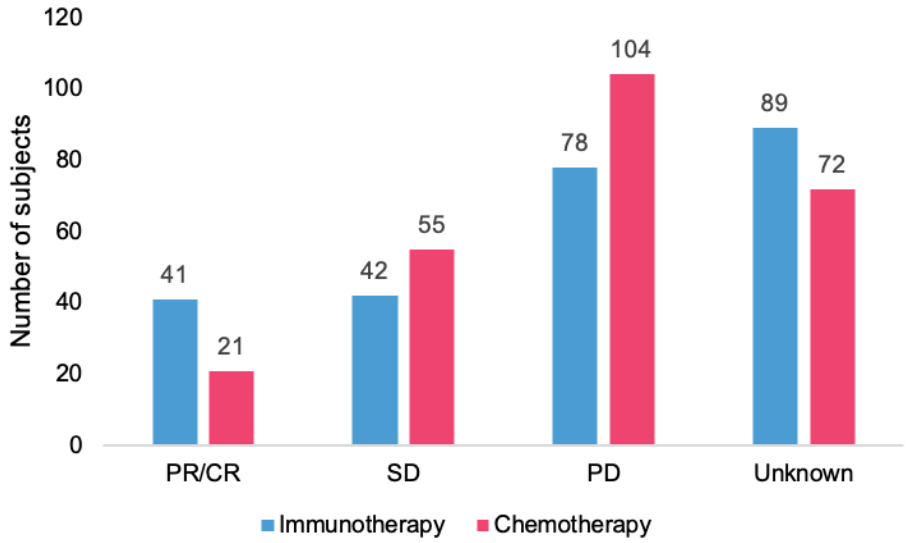
Table 3.4. MSE using 3-year partitioned data in chemotherapy arm by estimation methods

	Estimation Method					
Time Period (months)	**KWSF**	**SPM***	**P-MCM**	**I-MCM**	**P-NMCM**	**I-NMCM**
6	0.16%	0.27%	0.33%	0.25%	0.30%	0.21%
12	0.24%	0.33%	0.40%	0.28%	0.32%	0.18%
18	0.16%	0.27%	0.27%	0.12%	0.28%	0.13%
24	0.09%	0.23%	0.24%	0.07%	0.23%	0.08%
30	0.09%	0.23%	0.23%	0.08%	0.22%	0.08%
36	0.09%	0.28%	0.24%	0.05%	0.24%	0.05%
42	0.13%	0.29%	0.23%	0.08%	0.23%	0.08%
48	0.14%	0.29%	0.23%	0.12%	0.22%	0.10%
54	0.18%	0.32%	0.27%	0.16%	0.24%	0.14%
60	0.24%	0.38%	0.26%	0.17%	0.23%	0.14%
66	0.24%	0.35%	0.29%	0.23%	0.24%	0.19%
72	0.28%	0.39%	0.29%	0.23%	0.24%	0.19%
Total	2.02%	3.62%	3.29%	1.84%	3.00%	1.56%
Total after 3 years	1.21%	2.02%	1.58%	0.98%	1.41%	0.85%

MSE: mean squared error. KWSF: kernel-weighted survival forest. SPM: standard parametric model. P-MCM: population-level mixture cure model. I-MCM: individual-level mixture cure model. P-NMCM: population-level non-mixture cure model. I-NMCM: individual-level non-mixture cure model.

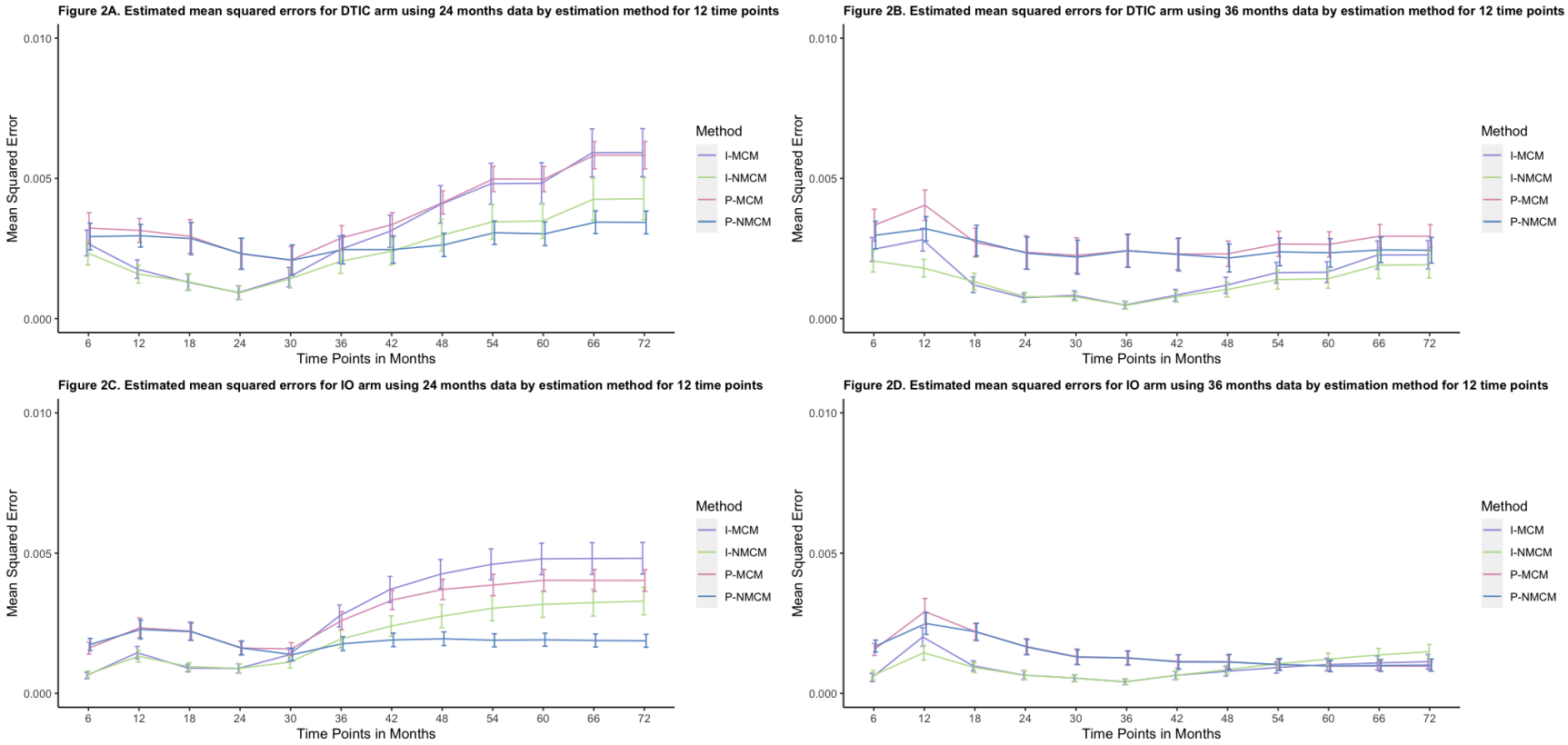
* Log normal distribution had the lowest AIC and was used for this estimation

Figure 3.1. Distribution of objective response at 12 weeks by trial arm



PR/CR: Partial/complete response defined as patients who have complete or partial response. SD: Stable disease defined as patients who remain progression free 3 months or more from start of treatment. PD: Progressive disease defined as patients who progress or are censored prior to 3 months.

Figure 3.2. Estimated mean squared errors for 12 time points by trial arm and estimation method



REFERENCES

1. Hafeez U, Gan HK, Scott AM. Monoclonal antibodies as immunomodulatory therapy against cancer and autoimmune diseases. Current opinion in pharmacology. 2018 Aug 1;41:114-21.

2. Franklin MR, Platero S, Saini KS, Curigliano G, Anderson S. Immuno-oncology trends: preclinical models, biomarkers, and clinical development. Journal for Immunotherapy of Cancer. 2022;10(1).

3. Upadhaya S, Neftelinov ST, Hodge J, Campbell J. Challenges and opportunities in the PD1/PDL1 inhibitor clinical trial landscape. Nat Rev Drug Discov. 2022 Feb 10;10.

4. Baik CS, Rubin EH, Forde PM, Mehnert JM, Collyar D, Butler MO, Dixon EL, Chow LQ. Immuno-oncology clinical trial design: limitations, challenges, and opportunities. Clinical Cancer Research. 2017 Sep 1;23(17):4992-5002.

5. Bullement A, Meng Y, Cooper M, Lee D, Harding TL, O'Regan C, Aguiar-Ibanez R. A review and validation of overall survival extrapolation in health technology assessments of cancer immunotherapy by the National Institute for Health and Care Excellence: how did the initial best estimate compare to trial data subsequently made available?. Journal of Medical Economics. 2019 Mar 4;22(3):205-14.

6. Azoury SC, Straughan DM, Shukla V. Immune Checkpoint Inhibitors for Cancer Therapy: Clinical Efficacy and Safety. Curr Cancer Drug Targets. 2015;15(6):452–62.

7. Latimer NR. Survival analysis for economic evaluations alongside clinical trials--extrapolation with patient-level data: inconsistencies, limitations, and a practical guide. Med Decis Mak Int J Soc Med Decis Mak. 2013 Aug;33(6):743–54.

8. Bullement A, Meng Y, Cooper M, Lee D, Harding TL, O'Regan C, Aguiar-Ibanez R. A review and validation of overall survival extrapolation in health technology assessments of cancer immunotherapy by the National Institute for Health and Care Excellence: how did the initial best estimate compare to trial data subsequently made available?. Journal of Medical Economics. 2019 Mar 4;22(3):205-14.

9. Bullement A, Latimer NR, Gorrod HB. Survival extrapolation in cancer immunotherapy: a validation-based case study. Value in Health. 2019 Mar 1;22(3):276-83.

10. Gibson E, Koblbauer I, Begum N, Dranitsaris G, Liew D, McEwan P, Tahami Monfared AA, Yuan Y, Juarez-Garcia A, Tyas D, Lees M. Modelling the survival outcomes of immuno-oncology drugs in economic evaluations: a systematic approach to data analysis and extrapolation. Pharmacoeconomics. 2017 Dec;35(12):1257-70.

11. Jansen JP. Network meta-analysis of survival data with fractional polynomials. BMC medical research methodology. 2011 Dec;11(1):1-4.

12. Ouwens MJ, Mukhopadhyay P, Zhang Y, Huang M, Latimer N, Briggs A. Estimating lifetime benefits associated with immuno-oncology therapies: challenges and approaches for overall survival extrapolations. Pharmacoeconomics. 2019 Sep;37(9):1129-38.

13. Othus M, Bansal A, Koepl L, Wagner S, Ramsey S. Accounting for cured patients in cost-effectiveness analysis. Value in Health. 2017 Apr 1;20(4):705-9.

14. Zhu R, Kosorok MR. Recursively imputed survival trees. Journal of the American Statistical Association. 2012 Mar 1;107(497):331-40.

15. Robert C, Thomas L, Bondarenko I, O'Day S, Weber J, Garbe C, et al. Ipilimumab plus dacarbazine for previously untreated metastatic melanoma. N Engl J Med. 2011 Jun 30;364(26):2517–26.

16. Lambert PC. Modeling of the cure fraction in survival studies. The Stata Journal. 2007 Sep;7(3):351-75.

17. Martinez EZ, Achcar JA, Jácome AA, Santos JS. Mixture and non-mixture cure fraction models based on the generalized modified Weibull distribution with an application to gastric cancer data. Computer methods and programs in biomedicine. 2013 Dec 1;112(3):343-55.

18. Cui Y, Zhu R, Zhou M, Kosorok M. Consistency of survival tree and forest models: splitting bias and correction. arXiv preprint arXiv:1707.09631. 2017 Jul 30.

19. Cui Y, Hannig J. Nonparametric generalized fiducial inference for survival functions under censoring. Biometrika. 2019 Sep 1;106(3):501-18.

20. Cooper M, Smith S, Williams T, Aguiar-Ibáñez R. How accurate are the longer-term projections of overall survival for cancer immunotherapy for standard versus more flexible parametric extrapolation methods?. Journal of Medical Economics. 2022 Dec 31;25(1):260-73.

21. Adunlin G, Cyrus JW, Dranitsaris G. Correlation between progression-free survival and overall survival in metastatic breast cancer patients receiving anthracyclines, taxanes, or targeted therapies: a trial-level meta-analysis. Breast cancer research and treatment. 2015 Dec;154(3):591-608.

22. Rozeman EA, Dekker TJ, Haanen JB, Blank CU. Advanced melanoma: current treatment options, biomarkers, and future perspectives. American Journal of Clinical Dermatology. 2018 Jun;19(3):303-17.

23. Jenkins RW, Fisher DE. Treatment of advanced melanoma in 2020 and beyond. Journal of Investigative Dermatology. 2021 Jan 1;141(1):23-31.

24. Niezgoda A, Niezgoda P, Czajkowski R. Novel approaches to treatment of advanced melanoma: a review on targeted therapy and immunotherapy. BioMed research international. 2015 Oct;2015.

25. Maio M, Grob JJ, Aamdal S, Bondarenko I, Robert C, Thomas L, Garbe C, Chiarion-Sileni V, Testori A, Chen TT, Tschaika M. Five-year survival rates for treatment-naive patients with advanced melanoma who received ipilimumab plus dacarbazine in a phase III trial. Journal of clinical oncology. 2015 Apr 4;33(10):1191.

26. Schadendorf D, Hodi FS, Robert C, Weber JS, Margolin K, Hamid O, Patt D, Chen TT, Berman DM, Wolchok JD. Pooled analysis of long-term survival data from phase II and phase III trials of ipilimumab in unresectable or metastatic melanoma. Journal of clinical oncology. 2015 Jun 6;33(17):1889.

CHAPTER 4: ESTIMATING INDIVIDUALIZED TREATMENT RULES TO MAXIMIZE OVERALL SURVIVAL USING AN IMMUNOTHERAPY TRIAL FOR ADVANCED MELANOMA

Introduction

Currently hundreds of clinical trials are being conducted globally to evaluate the effectiveness of checkpoint immunotherapy drugs. These drugs target the immune system checkpoints – molecules on certain immune cells that need to be activated or inactivated to start an immune response – and are considered the standard of care in many cancers.[1] Checkpoint immunotherapy drugs typically trigger a durable response in a subset of patients, which translates to long-term survival for some – but not all – patients.[2] However, the predictors of long-term survival are yet to be fully understood and long-term follow-up is the ultimate way to determine whether distinct subpopulations truly exists in terms of response to checkpoint immunotherapy.[3]

Although checkpoint immunotherapy has revolutionized cancer treatment in many ways, these drugs have the potential to cause life-threating side effects and financial toxicity due to high cost. The cost of immunotherapy can vary widely based on factors such as treatment duration and the type and staging of cancer; however, evidence shows that on average the treatment cost of immunotherapy drugs is significantly higher than chemotherapy alternatives.[4] Potential side effects of checkpoint immunotherapy usually result from an overstimulated or misdirected immune response, and can range from mild to moderate or severe.[1] In more serious cases, checkpoint immunotherapy can cause the immune system to attack vital organs, which can lead to severe and sometimes life-threatening side effects in the lungs, intestines, liver, kidneys, or other organs. Serious side

effects could lead to treatment being stopped and might require suppressing the immune system.[5-7]

Because of the known heterogeneity in treatment response, as well as high cost and potential for severe adverse events associated with checkpoint immunotherapy drugs, identifying patients who will benefit from these drugs has become increasingly critical. Therefore, percision medicine strategies that can help individualize treatments for patients (or subsets of patients) have grown in popularity both in clinical practice and medical research.[8] An individualized treatment rule (ITR) is a precision medicine tool that can be used for this purpose. ITR is a data-driven decision algorithm that recommends an optimal treatment according to patient characteristics in a way that, if implemented in practice, can maximize the health outcome of interest.[9] ITRs can be particularly beneficial in the context of known heterogeneity of treatment response, for instance, cancer treatment with checkpoint immunotherapy drugs.

Clinical trial data are commonly used to construct ITRs; however, complexity of disease mechanism, individual heterogeneity, and presence of numerous known and unknown potential outcome predictors make constructing ITRs using trial data difficult, particularly in trials with limited follow-up. In this paper we used outcome-weighted learning (OWL) to estimate ITRs that can maximize individual- and population-level survival gains.[10] OWL is a novel classification approach that has been suggested for constructing ITRs using clinical trial data. This approach directly estimates the decision rule that maximizes outcome of interest and is robust to the model misspecification.[10]

To estimate ITRs via OWL using clinical trial data, the reward value (i.e., the health outcome of interest) needs to be known to calculate the individual weights. However, clinical trials can be limited in the data they provide on the outcome of interest as a result of time and/or budget constraints. Specifically, trial results are commonly published before the outcome of interest i.e., overall survival (OS) is reached for all participants (i.e., some

individuals are right-censored).[11] Cui and colleagues extend the OWL framework to right-censored survival data using the recursively-imputed survival trees (RIST), a tree-based approach that nonparametrically imputes the survival time for right-censored observations.[12] However, the RIST algorithm caps the imputed failure time at the end of the follow-up duration of the clinical trial.[13] Since checkpoint immunotherapy drugs typically result in a long-term survival in a subset of population that goes well beyond typical trial follow-up,[2] using RIST alone for imputation might underestimate the survival impact of these treatments. To extrapolate the survival time beyond trial follow-up, we used an individual-level extrapolation model that was introduced in the Aim1 paper. The proposed method uses kernel-weighted survival forest (KWSF) to estimate failure time in a manner that is suitable for implementation within OWL.

Additionally, the literature is scarce when it comes to the evidence that demonstrates the potential economic and clinical benefits of implementing ITRs in real-world practice as the majority of the ITR-related literature focuses on the methodological aspect of estimating ITRs.[8] Generating evidence that shows the clinical and economic impacts associated with implementing ITRs is critical to convince patients, healthcare providers, and healthcare systems to adopt these precision medicine strategies. Specifically, more evidence is needed to compare individualizing treatment strategies with strategies that allocate treatments based on the average treatment effect observed in clinical trials as typically recommended in evidence-based guidelines and oncology value frameworks.[14,15]

Using patient-level data from a checkpoint immunotherapy clinical trial in advanced melanoma, this study aims to estimate ITRs using most accurate survival projections from a novel individual-level extrapolation method proposed in the Aim1 paper and calculate survival and cost impacts associated with implementing these ITRs in the trial cohort compared to the survival and cost impacts associated with universal use of the trial-

recommended treatment, with the goal of maximizing OS among patients with advanced melanoma.

Methods

Dataset

We trained and tested the ITRs in this study using patient-level data from the CA184-024 trial, a multi-center, randomized, double-blind, two-arm, phase III study in patients with untreated stage III (unresectable) or IV melanoma receiving dacarbazine (DTIC) plus ipilimumab vs. DTIC with placebo.[16] In this trial, a total of 502 subjects were randomized to ipilimumab plus DTIC (n=250) and to DTIC monotherapy (n=252). The CA184-024 study has a relatively long follow-up duration, and available trial data includes minimum follow-up of 5 years.[17,18] In this trial, ipilimumab was found to produce long-term survival results in a fraction of subjects treated; similar impact has been seen in other checkpoint immunotherapy trials.[2] We identified 18 predictive variables based on data availability; predictors with missing values were excluded from the analysis. All analyses were conducted based on an intention-to-treat framework i.e., trial data were analyzed assuming that subjects received the randomly assigned treatment.

Estimating Extrapolated Failure Times for Right-Censored Subjects

We used KWSF, a novel method to estimate individual-level survival functions (introduced in Aim 1 paper), to predict survival beyond trial follow up for right-censored subjects. This novel model leverages survival forest based on extremely randomized trees[13,19] and kernel-weighted parametric extrapolation. For the parametric extrapolation, multiple parametric survival distributions including exponential, Weibull, log logistic, log normal, gamma, and generalized gamma were fitted to each individual-level survival function (i.e., the forest-level survival function for an individual) and the distribution with the lowest AIC was selected for survival extrapolation for that individual. This process allows for

potentially different distributions to be selected for different subjects within the trial. KWSF was used to estimate failure time only for the censored subjects.

Estimating ITRs Using Extrapolated Failure Times

The individual-level extrapolated survival estimates were used as inputs in the OWL algorithm, an innovative classification approach that uses support vector machine techniques, to develop ITRs that maximize patient survival.[10] The OWL method can directly estimate a decision rule that, if implemented in practice, will maximize the clinical output i.e., OS.[10] Using clinical trial data, information about the optimal IRT is available only indirectly through the observed and extrapolated reward (OS). To better utilize this indirect information, the OWL algorithm assigns differential weights to each individual based on their observed/extrapolated OS. More specifically, for subjects observed to have a large reward (i.e., longer OS) this rule is apt to recommend the same treatment assignments that the subject has actually received; however, for individuals with small rewards (i.e., shorter OS), the rule is more likely to give the opposite treatment assignment to what they received. Therefore, the optimal IRT misclassifies less individuals with high reward as compared to the individuals with low reward.[10] The OWL uses support vector machines to solve this weighted classification problem.

In this analysis, we randomly divided the 502 subjects into four groups and use three parts as training data to estimate the optimal rule and calculate the empirical value of the reward function based on the remaining part. We then permute the training and testing groups and average the four results. This procedure is then repeated 100 times and averaged to obtain the empirical value of reward i.e., average OS resulting from the treatment allocations based on ITR recommendations. Additionally, we characterized subjects who were recommended to receive either treatments by the ITR algorithm.

Estimating the Economic Impact of Each Treatment Allocation Strategy

To estimate the economic impact of implementing the ITRs, direct treatment cost (payer perspective) for each trial intervention were estimated using a set of parameters extracted from the literature. For Ipilimumab, we assumed a four-dose regimen at a dose of 3 mg per kilogram of body weight.[20] The unit cost of Ipilimumab was estimated to be $6,659.07 per 50 mg vial.[21] Assuming average weight of 80 kilogram, 5 vials will be needed for each dose, which results in $33,295 per dose and $133,181 per treatment course. The cost of one dose of DTIC was estimated to be $989.[22] Assuming a 6-dose treatment for a typical course of DTIC,[16] the total treatment cost was estimated to be $5,933. Since the same dose of DTIC was used in both arms, we estimated per person cost of treatment in the immunotherapy arm (ipilimumab + DTIC) to be $139,115. For the chemotherapy arm (DTIC monotherapy), the per person treatment cost was estimated to be $5,933. The drug administration costs were assumed to be equal between the two arms of treatment and was not factored in the calculations.

Overall survival (life years gained) and direct treatment cost (measured in 2020 US dollars) were assessed for four distinct treatment allocations strategies in the cohort of subjects studied in this advanced melanoma trial: (1) immunotherapy all: treatment allocation based on the average treatment effect of the clinical trial (i.e., every subjects receives immunotherapy); (2) ITR: treatment allocation based on the estimated ITRs; (3) chemotherapy all: a hypothetical treatment allocation strategy where every subject receives the less costly, and less effective treatment (i.e., every subject receives chemotherapy); and (4) RCT: treatment allocation based on the randomization implemented in the clinical trial. The last two strategies were selected to provide baseline estimates for comparison as they are unlikely to be used in practice.

To compare the cost and benefits of these four treatment allocation strategies, we calculated the incremental cost effectiveness ratio (ICER) as defined by ($Cost_{strategy1}$ −

$Cost_{strategy2}$)/ ($LY_{strategy1} - LY_{strategy2}$), with LY indicating life years gained. Additionally, the net monetary benefit of each strategy was estimated for three levels of willingness to pay (WTP) threshold per life-year gained: $50,000, $100,000, and $150,000 as defined by ($LY_{strategy1}$ * WTP) - $Cost_{strategy1}$.[23] All analyses were conducted using R version 4.0.2.

Results

The mean survival time using the available trial follow up data was 466.76 and 630.89 days in chemotherapy and immunotherapy arms, respectively. The corresponding numbers after survival extrapolation for right-censored subjects were 477.33 and 690.38 days for the chemotherapy and immunotherapy arms, respectively.

Among the 502 trial subjects, the ITR algorithm allocated 273 subjects to immunotherapy and 231 subjects to chemotherapy. Table 4.1 shows the distribution of a select demographic and prognostic factors[24,25] among the subjects who were randomized to receive immunotherapy and chemotherapy in trial compared to subjects who were allocated to the same interventions by the ITR algorithm. Compared to the RCT treatment assignment, the ITR algorithm recommended slightly older subjects to receive immunotherapy (mean age of 59.11 years in ITR vs. 57.52 years in RCT). Similarly, higher proportion of female subjects were recommended to receive immunotherapy under ITR treatment allocation (43% in ITR vs. 39% in RCT). Further, the ITR algorithm recommended higher proportion of subjects with concomitant use of steroid to receive immunotherapy (72% in ITR vs. 63% in RCT).

Compared to other strategies, the ITR treatment allocation resulted in the highest average survival time estimate (879 days), while the “chemotherapy all” strategy resulted in the lowest average survival time (509 days). The discounted survival time were estimated to be 2.35, 1.83, 1.38, and 1.58 years for ITR, “immunotherapy all”, “chemotherapy all”, and RCT treatment allocation strategies, respectively (Table 4.2).

The strategies that allocate immunotherapy to a higher proportion of subjects were associated with higher average treatment costs with $139,115, $77,830, $72,259, and $5,933 for “immunotherapy all”, ITR, RCT, and “chemotherapy all” strategies, respectively. Compared to “immunotherapy all”, the ITR strategy was associated with lower treatment cost and higher life years gained with calculated ICER of -$118,236/LY (Table 4.3). The “immunotherapy all” strategy was associated with the lowest estimated net benefit across all WTP thresholds considered, ranging from -$47,371 to $136,116 for WTP of $50,000/LY and $150,000/LY, respectively (Table 4.4). The ITR strategy resulted in the highest net monetary benefit for WTP of $100,000/LY and WTP of $150,000/LY with $157,490 and $275,149, respectively, while the “chemotherapy all” strategy provided the highest net monetary benefit at the WTP of $50,000. (Figure 4.1)

Discussion

For this study, we used a combination of two machine learning methods together forming a novel approach to ITR estimation using clinical trial data that has a built-in mechanism to extrapolate survival. The KWSF algorithm provides individual-level extrapolated survival estimates that projects survival beyond the available trial follow up for right-censored subjects. This feature is particularly important for checkpoint immunotherapy drugs as such drugs typically result in long-term survival in a subset of patients and the extrapolation can provide a more accurate representation of survival effects of the checkpoint immunotherapy treatments.

We opted to use the OWL approach for constructing ITRs because of its robustness to model misspecification and ability to incorporate extrapolated survival estimates. OWL uses a weighted classification framework to directly estimate the decision rule that maximizes outcome of interest. Robustness to model misspecification combined with OWL’s ability to incorporate support vector machine makes OWL an ideal candidate for

constructing ITR using patient-level clinical trial data. Further, this algorithm provides a platform that can be used for trial or real-world data for other immunotherapy treatments.

Recently, value frameworks have been proposed for use in the clinical settings to facilitate individual treatment discussions between physicians and their patients. For instance, the ASCO Value Framework[14,15] and the NCCN Evidence Blocks[26] aim to assist providers and patients to make informed decisions about the value of oncology treatment regimens. These value frameworks typically assess the survival impact as well as other outcomes associated with an intervention based on the observed average treatment effect in a clinical trial. Our findings indicate that such use of average treatment effect produces less survival gains compared to what can be achieved using ITR recommendations. In fact, in our analysis, the “immunotherapy all” strategy that represents the recommendation based on the average treatment effect was associated with higher direct treatment costs, while resulting in lower survival gains compared to treatment allocation based on the estimated ITRs. Additionally, the “immunotherapy all” strategy resulted in the lowest estimated net monetary benefit across all WTP thresholds considered.

This study is subject to a number of limitations both in terms of the methodology and data availability. The OWL algorithm is based on support vector machine estimator which is primarily designed for binary classification e.g., comparing two treatments at a time[10,12]; however, similar methods can be expanded for multicategory classification.[27-29] For example, in a multi-arm clinical trial, the algorithm can be used to compare two arms at a time or to compare a given arm vs. all others. Further, the reward function for the OWL method was defined as OS; therefore, the OWL algorithm optimizes survival gains but does not account for quality of life, an important factor in treatment selection in oncology settings.

Additionally, patients have heterogenous preferences with regard to outcomes such as a given intervention’s safety and tolerability, quality of life impacts, and financial affordability that will need to be factored in the decision-making process.[30] While the OWL

approach provides individualized treatment recommendation that can maximize OS, it is not designed to incorporate heterogeneity in patient preferences for different health and economic outcomes.[31] Furthermore, the net monetary benefit calculations in this study only include direct treatment costs, not accounting for cost associated with adverse events and other treatment- and cancer-associated healthcare resource utilization.

The CA184-024 trial is one of the first phase III trials that assessed a checkpoint immunotherapy in advanced melanoma comparing Ipilimumab + DTIC with DTIC monotherapy.[32] However, Ipilimumab plus DTIC or DTIC monotherapy that were evaluated in this trial are not considered standard of care in advanced melanoma as several safer and more effective options have since become available for this indication.[33-35] While this study estimated ITRs for Ipilimumab plus DTIC vs. DTIC monotherapy, it prototypes approaches that can be used to inform additional treatment decisions especially in the context of checkpoint immunotherapy clinical trials.

The OWL algorithm, as described above, can be used to develop an application that inputs the characteristics of a given patient (outside the clinical trial) and provides the optimum treatment that can maximize OS. Implementing such tools and applications in practice, however, is subject to many challenges. Unlike traditional forms of statistical analysis, the machine learning algorithms has many features and components with little meaning to a human observer, making the explanation of the modeling technique difficult or impossible to interpret, which in turn makes it more difficult to convince patients and physicians to use these tools.[36]

Additionally, embedding treatment recommendations from such algorithms in clinical protocols and electronic health records (EHR) systems can be challenging.[36] Such integration issues can be a barrier to broader implementation of these precision medicine tools.[37] For widespread adoption of ITRs to take place, factors such as appropriate regulatory approvals, integration with EHR systems, quality standards to ensure the

reliability of these tools, training and educating clinicians, and financial resources for implementation and updating these tools need to be in place.

The findings of this study show that compared to allocating treatment based on the average treatment effect from a clinical trial, treatment allocation based on the estimated ITRs resulted in higher survival gains and lower direct treatment costs, which is likely to persist even when considering the cost of implementing individualized treatment and willingness to pay for life-years gained. We also demonstrated that maximizing survival using ITRs have the potential to create the highest net monetary benefit. This happens as a result of the ITR's ability to allocate the checkpoint immunotherapy treatment to patients who will benefit from it, which can make a big difference in cost for patients and healthcare system. Future studies are needed to test the capabilities of the proposed models in creating ITRs using more recent checkpoint immunotherapy trials.

Table 4.1. Distribution of a select prognostic characteristics of the trial subjects by allocation strategy (n=502)

Variables	RCT allocation		ITR allocation	
	Immunotherapy (n=250)	Chemotherapy (n=252)	Immunotherapy (n=271)	Chemotherapy (n=231)
Age (years)	57.52 (13.18)	56.42 (13.26)	59.11 (12.40)	54.46 (13.74)
Sex				
Female	98 (39%)	103 (41%)	116 (43%)	85 (37%)
Male	152 (60%)	149 (59%)	155 (57%)	146 (63%)
ECOG*				
Grade 0	177 (70%)	179 (71%)	189 (70%)	167 (72%)
Grade 1	73 (29%)	73 (29%)	82 (30%)	64 (28%)
Current tumor stage**				
Stage III	6 (2%)	12 (5%)	4 (1%)	14 (6%)
Stage IV	244 (97%)	240 (95%)	267 (99%)	217 (94%)
Metastasis stage**				
M0	6 (2%)	8 (3%)	4 (1%)	10 (4%)
M1A	37 (15%)	43 (17%)	47 (17%)	33 (14%)
M1B	64 (25%)	62 (25%)	65 (24%)	61 (26%)
M1C	143 (57%)	137 (55%)	155 (57%)	127 (55%)
Prior Adjuvant therapy				
Interferon	58 (23%)	56 (22%)	77 (28%)	37 (16%)

None	184 (73%)	185 (73%)	185 (68%)	184 (80%)
Other	8 (3%)	11 (4%)	9 (3%)	10 (4%)
Baseline elevated LDH				
Elevated	93 (37%)	110 (44%)	93 (34%)	110 (48%)
Normal	157 (62%)	140 (56%)	178 (66%)	119 (52%)
Not reported	0 (0%)	2 (1%)	0 (0%)	2 (1%)
Concomitant steroid use				
No	92 (37%)	164 (65%)	76 (28%)	180 (78%)
Yes	158 (63%)	88 (35%)	195 (72%)	51 (22%)

RCT: Randomized Controlled Trial. ITR: Individualized Treatment Rules. ECOG: Eastern Cooperative Oncology Group Performance Status. LDH: Lactate Dehydrogenase

*ECOG performance status grade 0 indicates 0—Fully active, able to carry on all pre-disease performance without restriction, and grade 1 indicates Restricted in physically strenuous activity but ambulatory and able to carry out work of a light or sedentary nature[38]

**Classified according to the tumor–node–metastasis categorization for melanoma of the American Joint Committee on Cancer[20,39]

Table 4.2. Survival gains estimation for four different treatment allocation strategies

	Treatment allocation strategies			
	ITR recommended	**Ipi + DTIC all***	**DTIC all***	**RCT assignment****
Number allocated to Ipi + DTIC	271	250	0	250
Number allocated to DTIC	231	0	252	252
Average survival-days (extrapolated)	879	679	509	583
Average survival-years (extrapolated)	2.41	1.86	1.39	1.60
Average survival-years (discounted*)**	2.35	1.83	1.38	1.58

ITR: Individualized Treatment Rules. Ipi: Ipilimumab. DTIC: Dacarbazine. RCT: randomized controlled trial.
*Based on the corresponding results from the trial after survival extrapolation
**Based on the original allocation in the trial after survival extrapolation
***At 3% rate

Table 4.3. Incremental cost estimation for 4 different treatment allocation strategies

	Treatment allocation strategies			
	ITR recommended	**Ipi + DTIC all***	**DTIC all***	**RCT assignment****
Average cost per person	$77,830	$139,115	$5,933	$72,259
Average survival-years	2.35	1.83	1.38	1.58
Cost/LY	$33,074	$75,817	$4,294	$45,737
ICER ($/LY)	($118,236)	NA	$293,932	$262,178

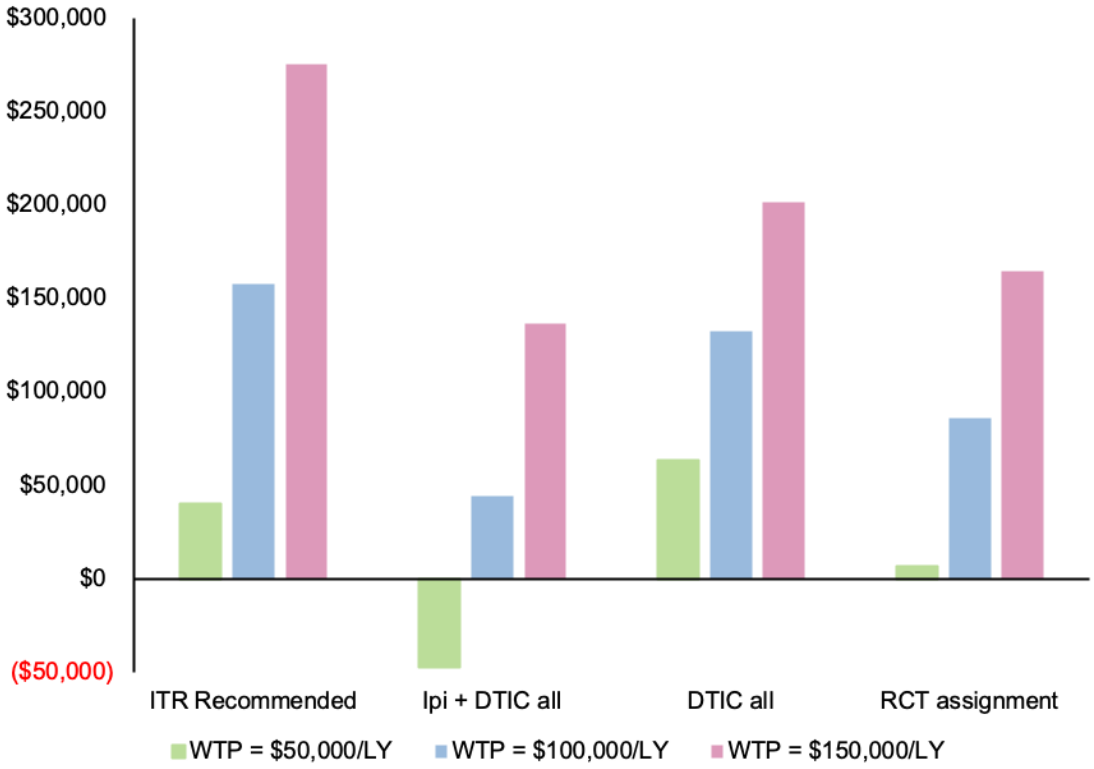
ITR: individualized treatment rules. Ipi: Ipilimumab. DTIC: Dacarbazine. RCT: randomized controlled trial. LY: life years. ICER: incremental cost-effectiveness ratio
*Based on the corresponding results from the trial after survival extrapolation
**Based on the original allocation in the trial after survival extrapolation

Table 4.4. Net monetary benefit estimates for four treatment allocation strategies by different WTP values

Treatment allocation strategies	ICER	Net monetary benefit for different levels of WTP		
		$50,000/LY	**$100,000/LY**	**$150,000/LY**
ITR recommended	($118,236)	$39,830	$157,490	$275,149
Ipi + DTIC all*	NA	($47,371)	$44,372	$136,116
DTIC all*	$293,932	$63,155	$132,243	$201,332
RCT assignment**	$262,178	$6,735	$85,728	$164,721

WTP: willingness to pay. LY: life years. Ipi: Ipilimumab. DTIC: Dacarbazine. ITR: individualized treatment rules. RCT: randomized controlled trial. ICER: incremental cost-effectiveness ratio
*Based on the corresponding results from the trial after survival extrapolation
**Based on the original allocation in the trial after survival extrapolation

Figure 4.1. Net monetary benefit of four treatment allocation strategies by WTP



WTP: willingness to pay. ITR: individualized treatment rule. Ipi: ipilimumab. DTIC: dacarbazine. RCT: randomized controlled trial. LY: life years

REFERENCES

1. Hafeez U, Gan HK, Scott AM. Monoclonal antibodies as immunomodulatory therapy against cancer and autoimmune diseases. Current opinion in pharmacology. 2018 Aug 1;41:114-21.

2. Azoury SC, Straughan DM, Shukla V. Immune Checkpoint Inhibitors for Cancer Therapy: Clinical Efficacy and Safety. Curr Cancer Drug Targets. 2015;15(6):452–62.

3. Jenkins RW, Barbie DA, Flaherty KT. Mechanisms of resistance to immune checkpoint inhibitors. British journal of cancer. 2018 Jan;118(1):9-16.

4. Vasekar MK, Agbese E, Leslie D. The value of immunotherapy: Comparison of annual cost per patient receiving immunotherapy versus chemotherapy in patients with non-small cell lung cancer.

5. American Society of Clinical Oncology (ASCO). *ASCO Annual Meeting 2019: Immunotherapy for lung cancer, gastrointestinal cancers and targeted therapy for breast cancer.* Accessed at https://www.cancer.net/blog/2019-06/asco-annual-meeting-2019-immunotherapy-lung-cancer-gastrointestinal-cancers-and-targeted-therapy on December 19, 2019.

6. American Society of Clinical Oncology (ASCO). *Understanding immunotherapy*. Accessed at https://www.cancer.net/navigating-cancer-care/how-cancer-treated/immunotherapy-and-vaccines/understanding-immunotherapy on December 19, 2019.

7. Bayer VR, Davis ME, Gordan RA, et al. Immunotherapy. In Olsen MM, LeFebvre KB, Brassil KJ, eds. *Chemotherapy and Immunotherapy Guidelines and Recommendations for Practice*. Pittsburgh, PA: Oncology Nursing Society; 2019:149-189.

8. Kosorok MR, Laber EB. Precision medicine. Annual review of statistics and its application. 2019 Mar;6:263.

9. Qian M, Murphy SA. Performance guarantees for individualized treatment rules. Annals of statistics. 2011 Apr 4;39(2):1180.

10. Zhao Y, Zeng D, Rush AJ, Kosorok MR. Estimating individualized treatment rules using outcome weighted learning. Journal of the American Statistical Association. 2012 Sep 1;107(499):1106-18.

11. Latimer NR. Survival analysis for economic evaluations alongside clinical trials--extrapolation with patient-level data: inconsistencies, limitations, and a practical guide. Med Decis Mak Int J Soc Med Decis Mak. 2013 Aug;33(6):743–54.

12. Cui Y, Zhu R, Kosorok M. Tree based weighted learning for estimating individualized treatment rules with censored data. Electronic journal of statistics. 2017;11(2):3927.

13. Zhu R, Kosorok MR. Recursively imputed survival trees. Journal of the American Statistical Association. 2012 Mar 1;107(497):331-40.

14. Schnipper LE, Davidson NE, Wollins DS, Tyne C, Blayney DW, Blum D, Dicker AP, Ganz PA, Hoverman JR, Langdon R, Lyman GH. American Society of Clinical Oncology statement: a conceptual framework to assess the value of cancer treatment options. Journal of Clinical Oncology. 2015 Aug 8;33(23):2563.

15. Schnipper LE, Davidson NE, Wollins DS, Blayney DW, Dicker AP, Ganz PA, Hoverman JR, Langdon R, Lyman GH, Meropol NJ, Mulvey T. Updating the American Society of Clinical Oncology value framework: revisions and reflections in response to comments received. Journal of Clinical Oncology. 2016 Aug 20;34(24):2925-34.

16. Robert C, Thomas L, Bondarenko I, O'Day S, Weber J, Garbe C, et al. Ipilimumab plus dacarbazine for previously untreated metastatic melanoma. N Engl J Med. 2011 Jun 30;364(26):2517–26.

17. Maio M, Grob JJ, Aamdal S, Bondarenko I, Robert C, Thomas L, Garbe C, Chiarion-Sileni V, Testori A, Chen TT, Tschaika M. Five-year survival rates for treatment-naive patients with advanced melanoma who received ipilimumab plus dacarbazine in a phase III trial. Journal of clinical oncology. 2015 Apr 4;33(10):1191.

18. Schadendorf D, Hodi FS, Robert C, Weber JS, Margolin K, Hamid O, Patt D, Chen TT, Berman DM, Wolchok JD. Pooled analysis of long-term survival data from phase II and phase III trials of ipilimumab in unresectable or metastatic melanoma. Journal of clinical oncology. 2015 Jun 6;33(17):1889.

19. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Machine learning. 2006 Apr;63(1):3-42.

20. Hodi FS, O'day SJ, McDermott DF, Weber RW, Sosman JA, Haanen JB, Gonzalez R, Robert C, Schadendorf D, Hassel JC, Akerley W. Improved survival with ipilimumab in patients with metastatic melanoma. New England Journal of Medicine. 2010 Aug 19;363(8):711-23.

21. Wang J, Chmielowski B, Pellissier J, Xu R, Stevinson K, Liu FX. Cost-effectiveness of pembrolizumab versus ipilimumab in ipilimumab-naive patients with advanced melanoma in the United States. Journal of Managed Care & Specialty Pharmacy. 2017 Feb;23(2):184-94.

22. Shih V, Ten Ham RM, Bui CT, Tran DN, Ting J, Wilson L. Targeted therapies compared to dacarbazine for treatment of BRAFV600E metastatic melanoma: a cost-effectiveness analysis. Journal of skin cancer. 2015 Jun 10;2015.

23. Shafrin J, May SG, Skornicki M, Hathway J, Macaulay R, Villeneuve J, Lees M, Hertel N, Penrod JR, Jansen J. Use of net monetary benefit analysis to comprehensively understand the value of innovative treatments. Value in Health. 2016 Nov 1;19(7):A731.

24. Manola J, Atkins M, Ibrahim J, Kirkwood J. Prognostic factors in metastatic melanoma: a pooled analysis of Eastern Cooperative Oncology Group trials. Journal of Clinical Oncology. 2000 Nov 15;18(22):3782-93

25. Ryan L, Kramar A, Borden E. Prognostic factors in metastatic melanoma. Cancer. 1993 May 15;71(10):2995-3005.

26. Carlson RW, Jonasch E. NCCN evidence blocks. Journal of the National Comprehensive Cancer Network. 2016 May 1;14(5S):616-9.

27. Huang X, Goldberg Y, Xu J. Multicategory individualized treatment regime using outcome weighted learning. Biometrics. 2019 Dec;75(4):1216-27.

28. Zhang C, Chen J, Fu H, He X, Zhao YQ, Liu Y. Multicategory outcome weighted margin-based learning for estimating individualized treatment rules. Statistica sinica. 2020;30:1857.

29. Zhou X, Wang Y, Zeng D. Outcome-weighted learning for personalized medicine with multiple treatment options. In2018 IEEE 5th international conference on data science and advanced analytics (DSAA) 2018 Oct 1 (pp. 565-574). IEEE.

30. Meropol NJ, Egleston BL, Buzaglo JS, Benson III AB, Cegala DJ, Diefenbach MA, Fleisher L, Miller SM, Sulmasy DP, Weinfurt KP, CONNECT Study Research Group. Cancer patient preferences for quality and length of life. Cancer. 2008 Dec 15;113(12):3459-66.

31. Butler EL, Laber EB, Davis SM, Kosorok MR. Incorporating patient preferences into estimation of optimal individualized treatment rules. Biometrics. 2018 Mar;74(1):18-26.

32. Gorry C, McCullagh L, Barry M. Economic evaluation of systemic treatments for advanced melanoma: a systematic review. Value in Health. 2020 Jan 1;23(1):52-60.

33. Rozeman EA, Dekker TJ, Haanen JB, Blank CU. Advanced melanoma: current treatment options, biomarkers, and future perspectives. American Journal of Clinical Dermatology. 2018 Jun;19(3):303-17.

34. Jenkins RW, Fisher DE. Treatment of advanced melanoma in 2020 and beyond. Journal of Investigative Dermatology. 2021 Jan 1;141(1):23-31.

35. Niezgoda A, Niezgoda P, Czajkowski R. Novel approaches to treatment of advanced melanoma: a review on targeted therapy and immunotherapy. BioMed research international. 2015 Oct;2015.

36. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future healthcare journal. 2019 Jun;6(2):94.

37. Low LL, Lee KH, Hock Ong ME, Wang S, Tan SY, Thumboo J, Liu N. Predicting 30-day readmissions: performance of the LACE index compared with a regression model among general medicine patients in Singapore. BioMed research international. 2015 Oct;2015.

38. Oken MM, Creech RH, Tormey DC, Horton J, Davis TE, McFadden ET, Carbone PP. Toxicity and response criteria of the Eastern Cooperative Oncology Group. American journal of clinical oncology. 1982 Dec 1;5(6):649-56.

39. Keung EZ, Gershenwald JE. The eighth edition American Joint Committee on Cancer (AJCC) melanoma staging system: implications for melanoma treatment and care. Expert review of anticancer therapy. 2018 Aug 3;18(8):775-84

CHAPTER 5: CONCLUSIONS

Since the first approval of a checkpoint immunotherapy drug by the US Food and Drug Administration in March 2011, this class of drugs, either individually or in combination with other drugs, have become standard of care for many cancers.[1] The long-term survival impacts of checkpoint immunotherapy drugs on a subset of treated population and their unique survival dynamics have challenged traditional statistical methods for example to model long-term survival impacts and estimate individualized treatment rules.[2] In this dissertation, I proposed novel machine learning techniques that can be used to tackle some of these challenges.

In the first paper of this dissertation, I used the Recursively Imputed Survival Trees (RIST) algorithm, a sophisticated nonparametric survival imputation method,[3] to develop Kernel-Weighted Survival Forest (KWSF) algorithm, a fit-for-purpose algorithm for individual-level survival extrapolation beyond the trial follow up. I tested the KWSF model on patient-level data from a checkpoint immunotherapy trial comparing its prediction accuracy with that of standard parametric models as a proof-of-concept study. Additionally, the first paper has been written in a tutorial format, where I described in details all the innerworkings of this new algorithm to make it more accessible for applied decision modelers.

The second paper of this dissertation expands on the novel method proposed in the first paper by taking advantage of its modular feature that allows for more flexible extrapolation functions to be used at individual level. In this paper, the predictive accuracy of the KWSF variations with cure fraction extrapolation function was compared with survival extrapolation methods that directly model heterogeneity of treatment response.[4]

Finally, the third paper of this dissertation uses the individual-level extrapolated survival estimates from the first two papers as inputs for outcome weighted learning (OWL) algorithm, a novel machine learning algorithm that directly estimates individualized treatment rules (ITRs).[5] The estimated ITRs were designed to allocate treatment based on patient's characteristics in a way that, if implemented in practice, will maximize the overall survival.[6] Additionally, I estimated the survival impact, direct treatment cost, and net monetary benefit associated with using the estimated ITRs and compared these outcomes with treatment allocation strategies that are based on average treatment effect estimate from clinical trials, as commonly recommended in oncology value frameworks.[7,8]

Summary of Results

The first paper provides a tutorial that introduces the KWSF model as a novel survival extrapolation method that uses patient-level characteristics to estimate individualized survival function, which can then be used for individual-level survival extrapolation. This model can accommodate potential individual-level treatment response heterogeneity and survival dynamics of checkpoint immunotherapy treatments. The implementation of the proposed method on data from a checkpoint immunotherapy clinical trial in advanced melanoma showed that compared to standard parametric models, KWSF more accurately predicted survival beyond the available trial follow up. The KSWF model consistently outperformed the standard parametric model across all time points assessed in the study for both chemotherapy and immunotherapy arms, regardless of the duration of follow up available to train the models.

The results of the second paper showed that compared to models that use standard parametric extrapolation, cure fraction models and KWSF with cure fraction extrapolation function were more accurate in predicting survival in the immunotherapy arm of the trial. Our findings also provided further evidence illustrating the utility of cure fraction models for survival extrapolation both at individual and population levels. The difference between

accuracy of cure models and standard parametric models were less noticeable for the chemotherapy arm, potentially indicating that cure fraction might not be as effective for survival modeling of traditional cancer treatments such as chemotherapy. The KWSF model with a cure fraction survival extrapolation function demonstrated comparable accuracy with cure fraction models, while uniquely allowing for estimating individual-level survival functions that can be used to inform precision medicine strategies, and serve as the foundation for individual-level simulation of checkpoint immunotherapy drugs' relative economic value and efficiency.

The findings of third paper showed that compared to allocating treatment based on the average treatment effect from a clinical trial, treatment allocation based on the estimated ITRs resulted in higher survival gains and lower direct treatment costs, which is likely to persist even when considering the cost of implementing individualized treatment and willingness to pay for life-years gained. We also demonstrated that maximizing the overall survival using ITRs has the potential to create the highest net monetary benefit. This happens as a result of the ITR's ability to allocate the checkpoint immunotherapy treatment to patients who will benefit from it, which can make a big difference in cost for patients and healthcare systems.

Limitations

For all three aims of this study, I used patient-level data from CA184-024 Study: a multi-center, randomized, double-blind, two-arm, phase III study in patients with untreated stage III (unresectable) or IV melanoma receiving dacarbazine plus ipilimumab vs. dacarbazine with placebo.[9] The CA184-024 offers a relatively long follow-up duration with minimum follow-up of 5 years.[10,11] The longer-term follow-up duration of this trial allows for assessing the prediction accuracy of proposed extrapolation models using varied amounts of follow-up duration (i.e., two and three years); however, the treatments evaluated in this clinical trial are not considered a standalone treatment options for advanced melanoma.[12-14]

That being said, we believe the proposed extrapolation method can be used as a prototype for any randomized controlled trial of checkpoint immunotherapy or similar treatments in the context of a limited follow-up and known heterogeneity of treatment response. Similarly, in the third paper, although the ITRs were estimated for ipilimumab plus DTIC vs. DTIC monotherapy, the proposed algorithm prototypes approaches that can be used to inform additional treatment decisions especially in the context of checkpoint immunotherapy clinical trials.

Further, limited follow up from the trial and lack of external data (e.g., real-world data from registries) makes it difficult to validate the results of the extrapolation beyond the available trial data. This limitation is particularly important for recently approved checkpoint immunotherapy drugs, where long-term real-world data have not accumulated yet. In addition, the algorithms used in this dissertation are not designed to explicitly model potential survival impacts of the subsequent i.e., second- and third-line treatments.

The OWL algorithm used in the third paper is based on support vector machine estimator which is primarily designed for binary classification[5]; however, similar methods can be expanded for multicategory classification, e.g., clinical trials with more than two arms.[15-17] Further, the reward function for the OWL method was defined as overall survival; therefore, the OWL algorithm optimizes survival gains but does not account for quality of life, an important factor in treatment selection in oncology settings. Additionally, we acknowledge that patients have heterogenous preferences with regard to treatment outcomes such as safety and tolerability, quality of life impacts, and financial affordability that will need to be factored in the decision-making process. While the OWL approach provides individualized treatment recommendation that can maximize overall survival, it is not designed to incorporate heterogeneity in patient preferences for different health and economic outcomes. Furthermore, the net monetary benefit calculations in this third paper only include

direct treatment costs, not accounting for cost associated with adverse events and other treatment- and cancer-associated healthcare resource utilization.

Real-world implementation of machine learning tools such as proposed algorithms is subject to many challenges.[18] Unlike traditional forms of statistical analysis, the machine learning algorithms has many features and components with little meaning to a human observer, making the explanation of the modeling technique difficult or impossible to interpret, which in turn makes it more difficult to convince patients, physicians, and healthcare systems to use these tools.[18] Additionally, embedding treatment recommendations from such algorithms in clinical protocols and electronic health records (EHR) systems can be challenging.[18] We acknowledge that such integration issues can be a barrier to a widespread implementation of these precision medicine tools.[19] For broader adoption of ITRs to take place, factors such as appropriate regulatory approvals, integration with EHR systems, quality standards to ensure the reliability of these tools, training and educating clinicians, and financial resources for implementation and updating these tools need to be established.

Policy Implications and Future Research

Results from the three papers of this dissertation have implications for decision analysis methods, precision medicine and clinical care, and policy development. This work serves as a case example of novel methodologic approaches to predict long-term survival impacts of checkpoint immunotherapy treatments beyond trial follow up that account for potential individual-level heterogeneity in treatment response. The KWSF algorithm can be used to develop an application that inputs the characteristics of a given patient (outside the clinical trial) who received similar interventions to estimate their individualized survival function. Such application can help develop individual-level simulation models for economic evaluation of checkpoint immunotherapy drugs as well as informing the estimation of individualized treatment rules. Lastly, although treatment decisions involve a number of

complex and inter-related factors, application of the proposed predictive models may provide valuable individualized information that can improve decision making in the clinical setting.

Several potential expansions of this work may provide more insights and help facilitate the widespread adoption of these algorithms: First, it is imperative to train and validate the proposed novel methods using patient-level data from more recent and clinically-relevant checkpoint immunotherapy clinical trials as well as external data sources. Second, developing user-friendly applications based on the proposed methods that can input characteristics of a given patient and generate extrapolated survival function and the optimum individualized treatment rule that can be used by researchers and clinicians. Third, evaluating the real-world effectiveness of using the ITR application for allocating treatment and comparing the associated health and economic outcomes with the standard of care treatment allocation can generate more convincing evidence for adoption of these new models. Lastly, embedding the proposed ITR application in electronic health records systems and clinical workflows can facilitate a more widespread use of the proposed application in real-world practice.

The first two papers of this dissertation are methodological, addressing gaps in decision science methods through the use of predictive analytics (machine learning algorithms) and setting the stage for informed decision-making. The third paper builds on this foundation and other novel machine learning methods to inform cancer treatment decision making by assessing the cost and survival impacts of individualized treatment rules versus treatment assignment based on average treatment effect from clinical trial results. Using the novel algorithms described above, this dissertation provides valuable tools for individual-level survival extrapolation and developing individualized treatment rules that has the potential to improve patient outcomes and reduce healthcare costs.

REFERENCES

1. Hafeez U, Gan HK, Scott AM. Monoclonal antibodies as immunomodulatory therapy against cancer and autoimmune diseases. Current opinion in pharmacology. 2018 Aug 1;41:114-21.

2. Annemans L, Asukai Y, Barzey V, Kotapati S, Lees M, Van Baardewijk M, Wang Q, Batty AJ, Fisher D. MO2 Extrapolation in oncology modelling: novel methods for novel compounds. Value in Health. 2011 Nov 1;14(7):A242-3.

3. Zhu R, Kosorok MR. Recursively imputed survival trees. Journal of the American Statistical Association. 2012 Mar 1;107(497):331-40.

4. Ouwens MJ, Mukhopadhyay P, Zhang Y, Huang M, Latimer N, Briggs A. Estimating lifetime benefits associated with immuno-oncology therapies: challenges and approaches for overall survival extrapolations. Pharmacoeconomics. 2019 Sep;37(9):1129-38.

5. Zhao Y, Zeng D, Rush AJ, Kosorok MR. Estimating individualized treatment rules using outcome weighted learning. Journal of the American Statistical Association. 2012 Sep 1;107(499):1106-18.

6. Qian M, Murphy SA. Performance guarantees for individualized treatment rules. Annals of statistics. 2011 Apr 4;39(2):1180.

7. Schnipper LE, Davidson NE, Wollins DS, Tyne C, Blayney DW, Blum D, Dicker AP, Ganz PA, Hoverman JR, Langdon R, Lyman GH. American Society of Clinical Oncology statement: a conceptual framework to assess the value of cancer treatment options. Journal of Clinical Oncology. 2015 Aug 8;33(23):2563.

8. Schnipper LE, Davidson NE, Wollins DS, Blayney DW, Dicker AP, Ganz PA, Hoverman JR, Langdon R, Lyman GH, Meropol NJ, Mulvey T. Updating the American Society of Clinical Oncology value framework: revisions and reflections in response to comments received. Journal of Clinical Oncology. 2016 Aug 20;34(24):2925-34.

9. Robert C, Thomas L, Bondarenko I, O'Day S, Weber J, Garbe C, et al. Ipilimumab plus dacarbazine for previously untreated metastatic melanoma. N Engl J Med. 2011 Jun 30;364(26):2517–26.

10. Maio M, Grob JJ, Aamdal S, Bondarenko I, Robert C, Thomas L, Garbe C, Chiarion-Sileni V, Testori A, Chen TT, Tschaika M. Five-year survival rates for treatment-naive patients with advanced melanoma who received ipilimumab plus dacarbazine in a phase III trial. Journal of clinical oncology. 2015 Apr 4;33(10):1191.

11. Schadendorf D, Hodi FS, Robert C, Weber JS, Margolin K, Hamid O, Patt D, Chen TT, Berman DM, Wolchok JD. Pooled analysis of long-term survival data from phase II and phase III trials of ipilimumab in unresectable or metastatic melanoma. Journal of clinical oncology. 2015 Jun 6;33(17):1889.

12. Rozeman EA, Dekker TJ, Haanen JB, Blank CU. Advanced melanoma: current treatment options, biomarkers, and future perspectives. American Journal of Clinical Dermatology. 2018 Jun;19(3):303-17.

13. Jenkins RW, Fisher DE. Treatment of advanced melanoma in 2020 and beyond. Journal of Investigative Dermatology. 2021 Jan 1;141(1):23-31.

14. Niezgoda A, Niezgoda P, Czajkowski R. Novel approaches to treatment of advanced melanoma: a review on targeted therapy and immunotherapy. BioMed research international. 2015 Oct;2015.

15. Huang X, Goldberg Y, Xu J. Multicategory individualized treatment regime using outcome weighted learning. Biometrics. 2019 Dec;75(4):1216-27.

16. Zhang C, Chen J, Fu H, He X, Zhao YQ, Liu Y. Multicategory outcome weighted margin-based learning for estimating individualized treatment rules. Statistica sinica. 2020;30:1857.

17. Zhou X, Wang Y, Zeng D. Outcome-weighted learning for personalized medicine with multiple treatment options. In2018 IEEE 5th international conference on data science and advanced analytics (DSAA) 2018 Oct 1 (pp. 565-574). IEEE.

18. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future healthcare journal. 2019 Jun;6(2):94.

19. Low LL, Lee KH, Hock Ong ME, Wang S, Tan SY, Thumboo J, Liu N. Predicting 30-day readmissions: performance of the LACE index compared with a regression model among general medicine patients in Singapore. BioMed research international. 2015 Oct;2015.