Comparative Outcomes of Treatment Regimens in Waldenström Macroglobulinemia

By

Chang Chen

BS, Purdue University, 2015

Submitted in partial fulfillment of the requirements for the Master of Science in the Department of Biostatistics at Brown University School of Public Health

PROVIDENCE, RHODE ISLAND

May, 2017

This thesis by Chang Chen is accepted in its present for by the Department of Biostatistics as satisfying the thesis requirements for the degree of Master of Science

Date	Advisor: Roee Gutman
Date	Reader: Adam J. Olszewski
Date	Director: Adam J Sullivan, Ph.D.

Approved by the Graduate Council

Date	Dean of Graduate School: Andrew G. Campbell, Ph.D.

AUTHORIZATION TO LEND AND REPRODUCE THE THESIS

As the sole author of this thesis, I authorize Brown University to lend it to other Institutions or individuals of the purpose of scholarly research.

_____ _____

Date Author: Chang Chen

I further authorize Brown University to reproduce this thesis by photocopying or other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

_____ _____

Date Author: Chang Chen

Acknowledgements

I would first like to thank my thesis advisor Prof. Roee Gutman of the Department of Biostatistics at Brown University. His expertise and patience added considerably to my graduate experience. He consistently steered me in the right direction with infinite patience.

I would also like to thank my thesis reader Dr. Adam J. Olszewski, who is an assistant professor at the Warren Alpert Medical School of Brown University, for providing me with all the data and providing feedback on our work.

Furthermore, I would like to express my gratitude to the faculty and staff at Brown for helping me address any issue I met during my graduate life at Brown University.

Finally, I would like to thank my parents and girlfriend who experienced this long process with me, always offering lots of support and most sincere love.

Table of Content

1. Introduction……..1

2. Research Objectives…….....3

3. Dataset Description……..3

4. Statistical Methodology……...6

    4.1 Estimating the Propensity Score……9

    4.2 Assessing Covariates Balance……..10

5. Multiple Imputation with Two Subclassification Splines (MITSS) …….....10

    5.1 Spline……...11

    5.2 Extending MITSS to Survival Models…….11

        5.2.1 Mixture Cure Fraction Model…….12

        5.2.2 Convergence Diagnostic…….14

        5.2.3 Multiple Imputation……15

        5.2.4 Rubin's Rule……...16

    5.3 Bayesian Logistic Models……16

6. Results…….....17

    6.1 Propensity Score……..17

    6.2 Outcome Analysis……23

7. Discussion……..27

8. Reference……...29

List of Tables

Table 1. Patient characteristics, stratified by upfront regimen……4

Table 2. T-statistics and F test value – Final Propensity Score Specification (regimens with versus without rituximab) …….20

Table 3. T-statistics and F test value – Final Propensity Score Specification (rituximab monotherapy versus combination immunochemotherapy) ……...21

Table 4. T-statistics and F test value – Final Propensity Score Specification (purine analogue versus alkylator-based regimens) ……..….22

Table 5. Difference in proportions (regimen with rituximab versus without rituximab) …….24

Table 6. Difference in proportions (rituximab monotherapy versus combination immunochemotherapy) …….25

Table 7. Difference in proportions (purine analogue versus alkylator-based regimens) …….26

List of Figures

Figure 1. Cohort selection and treatment regimens……4

Figure 2. Trace Plot……...15

Figure 3. Standardized difference plots for regimens with versus without rituximab……..18

Figure 4. Standardized difference plots for rituximab monotherapy versus combination immunochemotherapy……18

Figure 5. Standardized difference plots for purine analogue versus alkylator-based regimens…….19

Figure 6. Kaplan-Meier Plot for regimens with versus without rituximab, after adjusting for covariates. …….24

Figure 7. Kaplan-Meier Plot for rituximab monotherapy versus combination immunochemotherapy, after adjusting for covariates. ……..25

Figure 8. Kaplan-Meier Plot for purine analogue versus alkylator-based regimens, after adjusting for covariates. ……26

1. Introduction

Waldenström Macroglobulinemia (WM) was first reported in 1944 by Jan Gosta Waldenström on two patients with lymphadenopathy, oronasal bleeding, anemia, thrombocytopenia, and infiltration of the BM by lymphoid cells. This rare disorder, which account for less than 2% of lymphoid malignancies, was recognized as a systemic B-cell lymphoma, in which malignant cells produce monoclonal IgM paraprotein, resulting in unique clinical complications. Plasmapheresis was the first successful therapy for WM-related hyperviscosity syndrome manifesting with hemostatic, pulmonary, neurologic and ocular dysfunction. This procedure still remains an important treatment modality for severe cases. Systemic therapies for WM had been developed using case series, single-arm clinical trials, and randomized studies that often pooled various indolent B-cell histologies (Olszewski, Treon, and Castillo 2016). Recent international guidelines recommend upfront rituximab in combination with alkylating agents (bendamustine, cyclophosphamide), purine analogues (fludarabine, cladribine, pentostatin), or as monotherapy, largely on the basis of phase II studies (Kapoor et al. 2017; Owen et al. 2014; Talaulikar et al. 2017; Dimopoulos et al. 2014; Buske et al. 2013).

It is generally difficult to study treatment options for WM in randomized clinical trials because of its rarity, and higher incidence among older patients. Many therapeutic agents are being tested in clinical trials for WM, including second-generation proteasome inhibitors, immunomodulators, mammalian target of rapamycin inhibitors, histone deacetylase inhibitors, and bruton tyrosine kinase inhibitors. However, most of these studies comprise small numbers of patients, and include heterogeneous patient populations, such as untreated or relapsed and refractory patients. Given the lack of large clinical trials in WM, especially phase III clinical

trials, the establishment of a standard treatment regimen that can be used as guideline for assessing the new experimental agents is of critical importance (Ghobrial 2013).
Using Medicare claims linked to the Surveillance, Epidemiology and End Results data, Olszewski et al (2016) showed that within 1 year of diagnosis, 50% of WM patients started chemotherapy, while 10% died without treatment. These estimates were 63% and 21%, respectively, at 5 years, with 16% of patients surviving without any treatment. The 5-year risk of death was 17% after 2005, and the proportion of patients receiving active chemotherapy was 67%. Median time from diagnosis to chemotherapy was 4 months in 2009–2013.

In 2013, more than half of the patients received rituximab as a single agent, and nearly a third received it in combination with chemotherapy. Although the absolute numbers of patients treated with bortezomib or bendamustine were low, these 2 agents surpassed all other classic cytotoxic drugs in 2012/2013 data, indicating that most patients that initiated multi-agent regimens in 2012 –2013, received rituximab with bortezomib or bendamustine (Olszewski, Treon, and Castillo 2016).

Large-scale observational studies that use causal inference methods can fill in gaps in clinical knowledge (Booth and Tannock 2014). Although they lack the benefit of randomization and require extreme scrutiny with regard to potential bias, they enable assessment of toxicities and outcomes in populations that are under-represented in clinical trials (Korn and Freidlin 2012). This may be of particular importance for older patients, who often experience a therapeutic risk/benefit ratio different than the one that is observed in clinical trials that include younger and fitter patients (Talarico, Chen, and Pazdur 2004).

2. Research Objectives

The objective of our study was to compare the overall survival (OS) and the rates of hospitalization, transfusion, and plasmapheresis among US Medicare beneficiaries with WM treated with various upfront chemo-immunotherapy strategies. In this study we performed three sets of comparisons. First, we compared patients receiving rituximab to those who weretreated without rituximab. Second, we compared patients receiving rituximab monotherapy with those receiving rituximab-based combinations. Finally, we compared regimens based on purine analogues with regimens based on alkylating agents while controlling for rituximab use.

3. Dataset Description

Using Medicare claims from 1999-2013, linked to the Surveillance, Epidemiology and End Results registry data, we identified WM patients who are over 65 years old and initiated first-line chemotherapy with or without rituximab, with purine analogues (fludarabine or cladribine) and/or classic alkylating agents (chlorambucil or cyclophosphamide). Patients receiving both fludarabine and cyclophosphamide were classified into the purine analogues group.

All of the patients in the dataset have been observed within 90 days of starting treatment. For each patient we also recorded baseline characteristics (such as age, sex, race, socioeconomic status, comorbidities, performance status, time from diagnosis), and indicators of WM severity previously validated to correlate with outcomes (anemia, transfusions, plasmapheresis). The outcomes for each patient included survival from start of treatment, occurrence of transfusion, hospitalization, and plasmapheresis. Detailed information about the dataset can be found in **Figure 1** and **Table 1**.

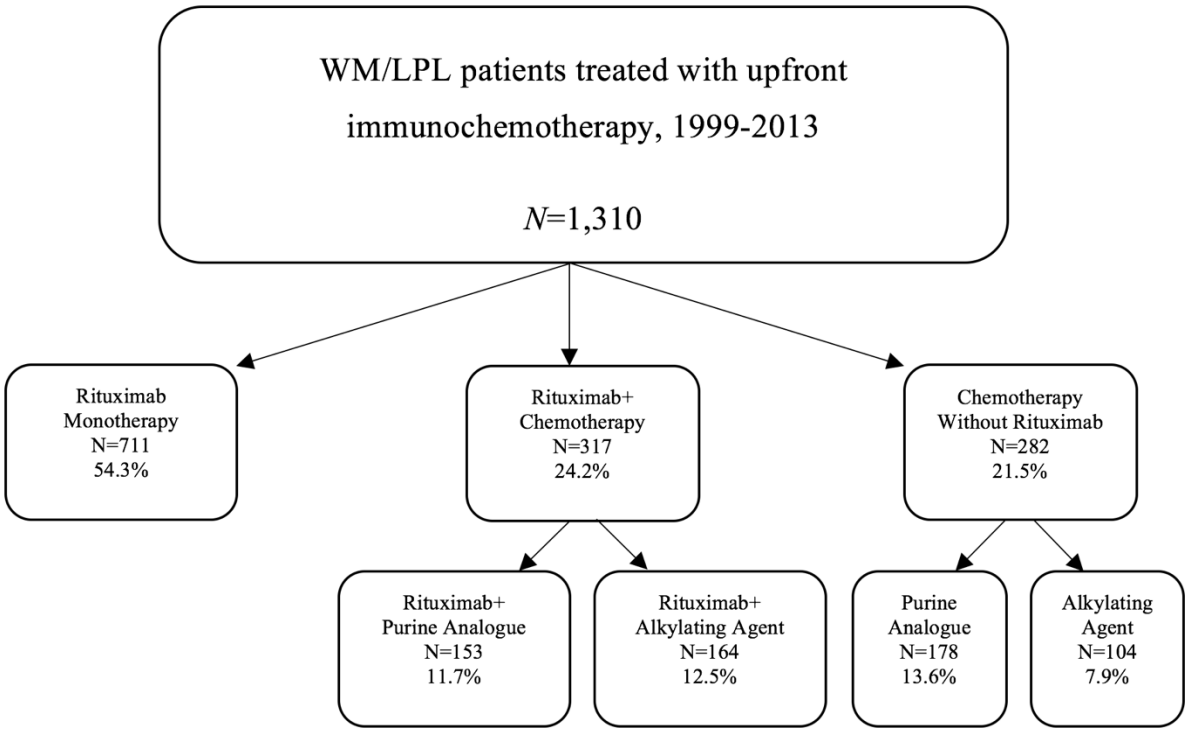

Figure 1. Cohort selection and treatment regimens.

Table 1. Patient characteristics, stratified by upfront regimen.

	Rituximab monotherapy		Rituximab+ chemotherapy		Chemotherapy without rituximab		P
	N	%	N	%	N	%	
Number of patients	711		317		282		
Age, median	**79 years**		**77 years**		**77 years**		**<0.001**
IQR	(75-84)		(72-81.5)		(73-81.5)		
Sex							
Women	325	(45.7)	121	(38.2)	116	(41.1)	0.06
Men	386	(54.0)	196	(61.8)	166	(58.9)	
Race							
White	654	(92.0)	287	(90.5)	263	(93.3)	0.31
Non-white	57	(8.0)	30	(9.5)	19	(6.7)	
Medicaid co-insurance							
No	659	(92.7)	296	(93.4)	263	(93.3)	0.9
Yes	52	(7.3)	21	(6.6)	19	(6.7)	
Marital status							
Married	**381**	**(53.6)**	**195**	**(61.5)**	**179**	**(63.5)**	**0.005**

Other	**330**	**(46.4)**	**122**	**(38.5)**	**103**	**(36.5)**	
Area of residence							
Metropolitan	**630**	**(88.6)**	**273**	**(86.1)**	**232**	**(82.3)**	**0.03**
Urban/rural	**81**	**(11.4)**	**44**	**(13.9)**	**50**	**(17.7)**	
Prior malignancy							
No	494	(69.5)	243	(76.7)	207	(73.4)	0.052
Yes	217	(30.5)	74	(23.3)	75	(26.6)	
Comorbidity index*							
0	373	(52.5)	173	(54.6)	163	(57.8)	0.41
1	156	(21.9)	73	(23.0)	62	(22.0)	
≥2	182	(25.6)	71	(22.4)	57	(20.2)	
Poor performance status*	58	(8.2)	16	(5.0)	13	(4.6)	0.055
WM complications#							
Neuropathy	**58**	**(8.2)**	**18**	**(5.7)**	**<11**	**(<3.0)**	**0.007**
Plasmapheresis	25	(3.5)	17	(5.4)	18	(6.4)	0.11
Transfusion	139	(19.5)	61	(19.2)	43	(15.2)	0.27
Anemia	**441**	**(62.0)**	**198**	**(62.5)**	**166**	**(58.9)**	**<0.001**
AIHA	**27**	**(3.8)**	**<11**	**(<2.0)**	**<11**	**(<2.0)**	**0.04**
Hospitalization	97	(13.6)	41	(12.9)	39	(13.8)	0.94
Time to treatment							
≥12 months	**198**	**(27.8)**	**52**	**(16.4)**	**74**	**(26.2)**	**<0.001**
<12 months	**513**	**(72.2)**	**265**	**(83.6)**	**208**	**(73.8)**	
Histology code							
WM	403	(56.7)	170	(53.6)	175	(62.1)	0.11
LPL	308	(43.3)	147	(46.4)	107	(37.9)	
Year of treatment							
1999/03	**120**	**(16.9)**	**69**	**(21.8)**	**168**	**(59.6)**	**<0.001**
2004/07	**271**	**(38.1)**	**121**	**(38.2)**	**71**	**(25.2)**	
2008/13	**320**	**(45.0)**	**127**	**(40.1)**	**43**	**(15.2)**	

* Based on Medicare claims from 12 months before start of chemotherapy

Based on Medicare claims from 6 months before start of chemotherapy

4. Statistical Methodology

The statistical methods that were used for all three comparisons were similar. Thus, we will describe the methodology used for comparing the effects of regimens with and without rituximab. The methodology used for the two other comparisons, follow similar suit, but with different treatments comparisons.

We consider a population of N units indexed by $i = 1, \ldots, N$. We characterized the i^{th} unit in this population as K–component row vector of covariates, X_i, where $\mathbf{X}$ is the $N \times K$ matrix of covariates. The covariates are assumed to be unaffected by the assignment of treatment (Imbens and Rubin 2015).

For each unit let $Y_i(0)$ and $Y_i(1)$ denote the potential outcomes under the two values of the treatment: the outcomes under the control treatment are denoted by $Y_i(0)$, and the outcomes under the active treatment are denoted by $Y_i(1)$. The use of this notation tacitly accepts the Stable Unit Treatment Value Assumption (SUTVA) that treatment assignments for other units do not affect the outcomes for unit i, and that there is only one version of the treatments (Imbens and Rubin 2015). We use $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$ to denote the N–component vectors of the potential outcomes. In our study, we consider the situation where the potential outcomes are scalars.

We also denote the N-component vector of treatment assignments by $\mathbf{W}$. The i-th element of this vector, $W_i \in \{0, 1\}$ represent the observed assignment of unit i. $W_i = 0$ denotes that unit i received the control treatment, and $W_i = 1$ denotes unit i received the active treatment. Let $N_c = \sum_{i=1}^{N}(1 - W_i)$ and $N_t = \sum_{i=1}^{N} W_i$ be the number of units assigned to the control and active treatment respectively, with $N_c + N_t = N$. We define the observed outcomes:

$$Y_i^{obs} = Y_i(W_i) = \begin{cases} Y_i(0), & \text{if } W_i = 0 \\ Y_i(1), & \text{if } W_i = 1 \end{cases}$$

and the missing potential outcomes:

$$Y_i^{mis} = Y_i(1 - W_i) = \begin{cases} Y_i(1), & \text{if } W_i = 0 \\ Y_i(0), & \text{if } W_i = 1 \end{cases}$$

where $\mathbf{Y}^{obs}$ and $\mathbf{Y}^{mis}$ are the corresponding N-vectors. These relations can be inverted and the potential outcomes can be characterized in terms of the observed and missing potential outcomes:

$$Y_i(0) = \begin{cases} Y_i^{mis}, & \text{if } W_i = 1 \\ Y_i^{obs}, & \text{if } W_i = 0 \end{cases}$$

$$Y_i(1) = \begin{cases} Y_i^{mis}, & \text{if } W_i = 0 \\ Y_i^{obs}, & \text{if } W_i = 1 \end{cases}$$

Causal inference problem can be considered as a missing data problem from the characterization above (Rubin 1978). In our analysis, we explicitly impute the missing outcomes, and create a complete dataset with all the potential outcomes observed so that we can estimate the causal effect for each unit.

The assignment mechanism is the process that determines which units receive which treatments, hence determining which potential outcomes are observed and are missing. Formally, the assignment mechanism describes, as a function of all covariates and of all potential outcomes, the probability of any vector of assignments. We consider two basic restrictions on the assignment mechanism: positivity and unconfoundedness:

1. Positivity requires that every unit has a non-zero probability to receive each of the treatments.

2. Unconfounded disallows dependence of the assignment mechanism on the potential outcomes.

Positivity is essential for inference but is often overlooked in practice. Often in many cases we find that somestrata defined by unique values of the covariates contain only a single unit, so that the proportion of treated units within the stratum is either zero or one. Such an occurrence makes many of the methods that rely on the estimated propensity score either inappropriate or suffer from significant extrapolation (Imbens and Rubin 2015). In our study, we attempt to ensure the positivity assumption by removing units that have covariates' values beyond the range of the observed values in the opposite treatment group.

Because the assignment of patients to regimens is not controlled by the investigators, there are significant differences in patients' characteristics between the regimens. Assuming that the treatment assignment is unconfounded, subclassification is a commonly used technique to remove bias in observational studies (Cochran, 1968, Rosenbaum & Rubin 1983). This method creates subgroups of patients that are relatively similar in terms of the covariates. Creating subgroups using a single covariate is relatively straight forward; however, when there are many covariates this task is more complex because finding patients that received different treatments and are exactly similar on all of the covariates has a very small probability (Stuart 2010). Possible techniques that were shown in theory (Rubin & Thomas 1992) and in applications (Hirano & Imbens 2001) to provide statistically valid inference, are matching and subclassification on the propensity score (PS) (Rosenbaum & Rubin 1983). The propensity score is the probability of being assigned to one treatment over the other given the covariates. Formally, let W_i be an indicator variable that is equal to one if patient *i* received rituximab and zero otherwise, and let X_i be the set of pre-treatment covariates. The propensity score is defined as $P(W_i|X_i)$. Using the propensity score, the sample is partitioned into subclasses, such that

within each subclass the covariate distributions in the treated and control subsamples are relatively similar. Commonly, the PS is unknown and requires estimation. Here, we relied on the logistic regression model to estimate the PS:

$$P(W_i|X_i) = \frac{\exp(X_i\beta)}{1+\exp(X_i\beta)} = e(X_i),$$

where β are a set of unknown parameters.

4.1 Estimating the Propensity Score

To estimate the propensity score we use a two-step iterative algorithm (Imbens and Rubin 2015). First, we specify an initial model, motivated by substantive knowledge. Second, we assess the adequacy of the estimate of that initial model, by checking whether the covariates are balanced within strata defined by the estimated propensity score. We iterate between these two steps, each time refining the specification of the model by including other covariates, quadratic terms and second order interactions based on the likelihood ratio test. The final propensity score model was decided based on covariate balance assessment. Individuals with propensity scores outside the region of common support in the opposite treatment group were removed from analysis (Imbens and Rubin 2015). The goal of the procedure is to find a specification for the propensity score that leads to adequate balance between covariate distributions in treatment and control groups in our sample, and not the model with the best fit for the data. (Imbens and Rubin 2015) It is important to note that this procedure is implemented without involving any outcome data, thus maintaining separate design and analysis stages (Rubin 2008).

4.2 Assessing Covariates Balance

Given the estimated propensity score, $\hat{e}(\boldsymbol{X}_i)$, we partitioned the sample into 6 blocks based on the quantiles. We assessed the balance between the two treatment groups using two sets of statistics and visualization methods. The first set of statistics were the t-statistic for the null hypothesis that the block-adjusted average difference in average covariate values is equal to zero for each of the covariates. This leads to 23 t-statistics or z-values (in comparative analysis 3 we have 24 t-statistics). Then we carry out the F-test which is based on comparing each of the within-stratum pseudo causal effects to zero. The F-test leads to a second set of statistics (F statistics) for assessing the null hypothesis that the difference in average covariate's values is zero in each block (Imbens and Rubin 2015). The love plot was used to display covariates balance conditional on the estimated propensity score.

5. MITSS

To adjust for residual imbalances and to obtain more precise estimates, we applied the multiple imputation with two subclassification splines (MITSS) (Gutman and Rubin 2015b). This procedure explicitly views causal effect estimation as a missing data problem, and was demonstrated to be superior to other commonly used causal effects estimation procedures. MITSS models the response surface in each treatment group as an additive model that combines a spline along the propensity score and linear model adjustments on all other orthogonalized covariates $\mathbf{X}^{ort}$ (Gutman and Rubin 2015b). Formally, MITSS assumes that

$$E(Y_i(W)|\boldsymbol{X_i}) = s\big(\hat{e}(\boldsymbol{X_i})\big) + \beta^T \boldsymbol{X_i^{ort}},$$

where *s* is a cubic spline with knots placed at the boundaries of the subclasses, and $\mathbf{X}^{ort}$ are the

covariates orthogonal to the propensity score. The intuition behind MITSS is to first subclassify on the estimated propensity score, so that within each subclass, $\mathbf{X}^{ort}$ is well balanced across treatment groups; second, to fit a spline across propensity subclasses, because Bayesian models based on $e(\mathbf{X}_i)$, are likely to be well calibrated (Gutman and Rubin 2015b). Third, adjust for minor imbalances with linear additive modeling.

5.1 Spline

Following Gutman & Rubin 2015, we partition the n units into six subclasses based on the quantiles of $\hat{e}(\mathbf{X}_i)$. We examined that there are at least three units from each treatment group in each of the subclasses. The number 3 was chosen so that all the parameters of the spline model can be estimated even when using an improper prior distribution (Gutman and Rubin 2015b). MITSS use regression splines with knots located at the borders of the subclasses and a linear combination of $\mathbf{X}_i^{ort}$, the location of the knots is guided by covariates balance assessment in section 4.2. If the covariates are not well balanced after partitioning based on propensity score, we adjust the propensity score model and relocate the knots until the covariates are well balanced within each subclass, so that the distribution of $\mathbf{X}_i^{ort}$ in each subclass are relatively similar in the control and treatment groups.

5.2 Extending MITSS to Survival Models

MITSS was previously described for continuous and binary outcomes (Gutman and Rubin 2015a; Gutman and Rubin 2015b). Here, we apply it to binary outcomes and extend it to estimate survival curves.

We use a Bayesian approach to sample from the posterior predictive distribution of the survival models, and multiply imputed both the missing potential outcomes and the censored

observation. To compare the two survival models, we plotted survival curves for each treatment group, and estimated the hazard ratio as a summary statistic. For binary outcomes, we used the absolute risk as a summary statistic. These plots, summary statistics and their corresponding standard errors would be calculated for each imputed dataset and then combined using Rubin's Rule for MI to obtain point and interval estimate.

5.2.1 Mixture Cure Fraction Model

Causal effects estimation should be composed of two distinct phases: design and analysis. The design phase includes contemplating, collecting, organizing, and analyzing data without seeing any outcome data (Rubin 1974). The analysis phase should proceed with analysis of the data according to a predetermined analysis plan (Rubin 2007). In observational studies, these two phases should be performed sequentially to approximate a randomized experiment (Rubin 2008). When the approximate functional form of survival is known, it can be used to estimate the response surfaces. However, when it is not known it is useful to examine several possible survival models on the control group in order to preserve objectivity of the analysis. We have estimated the Weibull model and the log-Normal model.

The Weibull model had the form:

$$Y_i^t(0) \sim \text{Weibull}(\lambda\ ,\ k),$$

$$\lambda = s(\hat{e}\ (\boldsymbol{X}_i)) + \beta^T \boldsymbol{X}_i^{ort},$$

where $Y_i^t(0)$ is the survival time for an individual *i* receiving the control, λ is reparameterized in terms of covariates and regression adjustment, *k* is the shape parameter.

The log-Normal model had the form:

$$Y_i^t(0) \sim \ln \mathcal{N}(\mu, \sigma^2),$$

$$\mu = s(\hat{e}\ (\boldsymbol{X}_i)) + \beta^T \boldsymbol{X}_i^{ort},$$

where $Y_i^t(0)$ is the survival time for an individual *i* receiving the control, μ is reparameterized in terms of covariates and regression adjustment, σ is the scale parameter.

WM is an incurable indolent disease, and some patients may survive many years with the disease (Olszewski, et al., 2016). To examine the fit of common parametric survival models, we estimated the distribution of survival times for censored control units using samples from the posterior predictive distribution. Both parametric models tend to have a long tail in their posterior predictive distributions for some individuals, which resulted in improbable survival times (e.g. for some individuals who are older than 70, the probability to live for 40 additional years isestimated to be 20%). Thus, we concluded that these two models are inappropriate for this data.

Mixture cure models have been originally proposed to analyze survival data when a fraction of the patients are cured from the disease, and the failure time for the uncured patients follows a proper survival distribution, referred to as latency distribution (Boag 1949). These models can also be used to investigate the heterogeneity between patients with indolent cancer who are long-term survivors and those who are not (Othus, et al., 2012). The cure model is formally described using the following survival function:

$$S(t) = p + (1-p)S_0(t),$$

where $0<p<1$ is a parameter which represents the proportion of "long-term survivors", and $S_0(t)$ is the baseline survival function for the susceptible individuals. Based on this model, the probability density function for the lifetime T is

$$f(t) = \frac{dF(t)}{dt} = (1-p)f_0(t),$$

where $F(t) = 1 - S(t)$ and $f_0(t)$ is the baseline probability density function for the susceptible individuals. Let Y_i^{obs} i=1, …, n, be the survival time for unit i and let δ_i be an indicator that is equal to 1 for observed lifetime and 0 for censored lifetime. The contribution of the *i*th subject for the likelihood function is given by

$$L_i = [f(Y_i^{obs})]^{\delta_i}[S(Y_i^{obs})]^{1-\delta_i} = [(1-p)f_0(Y_i^{obs})]^{\delta_i}[p + (1-p)S_0(Y_i^{obs})]^{1-\delta_i},$$

We assumed that S_0 follows the Weibull distribution (Martinez et al. 2013), such that

$$S_0\big(Y_i(0)\big) = \exp\left(-\frac{Y_i(0)}{\lambda}\right)^k,$$

$$\lambda = s(\hat{e}\,(\boldsymbol{X}_i)) + \beta^T \boldsymbol{X}_i^{ort},$$

and $\log\left(\frac{p_W}{1-p_W}\right) = s(\hat{e}\,(\boldsymbol{X}_i)) + \beta^T \boldsymbol{X}_i^{ort}$, where p_W is the proportion of long-term survivors in treatment group W.

This mixture cure Weibull model was found to provide better fit to the data.

5.2.2 Convergence Diagnostic

MCMC algorithms are used for obtaining samples from complex posterior distributions. A major consideration when using MCMC algorithms to sample from a complex distribution is the assessment of its convergence to the stationary distribution. One approach to assess MCMC convergence is based on analysis of multiple chains with different initial values. The Gelman–Rubin diagnostic (Gelman and Rubin 1992) evaluates MCMC convergence by analyzing the ratio between the overall variability of the sampled parameter to the variability within the chains. Large ratios indicate non-convergence (Balov 2016). In addition to the Gelman-Rubin statistic we have also examined the chains' trace plot.

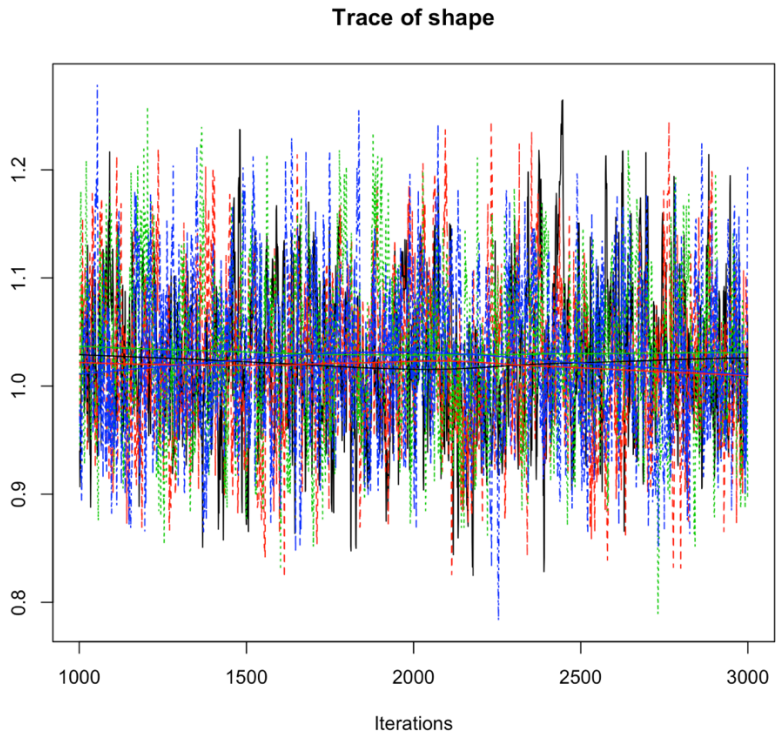

Figure 2. Trace Plot

Figure 2 shows trace-plots for 4 different chains. The chains seem to be mixing well. The Gelman-Rubin Statistics for all of the parameters were smaller than 1.05, and each of the corresponding upper confidence interval was less than 1.1. These results support the convergence of the chains to the stationary distribution.

5.2.3 Multiple Imputation

Using samples from the posterior predictive distribution of the Weibull cure model estimated in the treatment group, we multiply imputed the missing control and the censored outcomes in the active control group. Similarly, using samples from the posterior predictive distribution of the Weibull cure model estimated from the control group, we multiply imputed the missing control potential outcomes, and the censored outcomes in the control group. We repeat the imputation process M=20 times.

5.2.4 Rubin's Rule

Within each imputed dataset we estimate the treatment effects using different statistics, γ, and their corresponding sampling variance. γ, included the mean difference in survival years between the treatment group and the control group and its paired standard error. In addition, it included the hazard ratio for matched paired data based on the Holt and Prentice stratified Cox model (Savignoni et al. 2014).

Let $\hat{\gamma}^{(m)}$ be the estimated treatment effect from dataset m=1, …, M, and let $\hat{V}^{(m)}$ be its corresponding estimated sampling variance. We used Rubin's Rule for MI (Little and Rubin 1987), to obtain a point estimate of γ across imputed dataset is obtained by $\hat{\gamma} = \frac{1}{M}\sum_{m=1}^{M}\hat{\gamma}^{(m)}$, and its corresponding sampling variance estimate is obtained by:

$$\sqrt{\frac{1}{M-1}\sum_{m=1}^{M}(\hat{\gamma}^{(m)}-\hat{\gamma})^2+\frac{1}{M}\sum_{m=1}^{M}\hat{V}^{(m)}}$$

Interval estimate for γ is calculated using the t approximation to the posterior distribution of γ given by Barnard and Rubin (Barnard and Rubin 1999).

5.3 Bayesian Logistic Models

The Bayesian logistic model is applied to analyze the remaining three binary outcome variables, which are the occurrence of hospitalization, transfusion, or plasmapheresis within 3 months after starting chemotherapy—ascertained from specific procedure codes (HCPCS or International Classification of Diseases, 9th Revision, Clinical Modification) in inpatient and outpatient Medicare claims. We estimate each of these outcomes separately in the control and treatment groups using the following logistic regression models:

$$logit(E(Y_i(W)|\boldsymbol{X}_i)) = s(\hat{e}\,(\boldsymbol{X}_i)) + \beta^T \boldsymbol{X}_i^{ort},$$

where $Y_i(W)$ is a binary outcome of individual *i* under treatment *W*.

These models were used to impute the missing potential outcomes by sampling from the posterior predictive distribution for each individual. We imputed Y_i^{mis} M=20 times.

The imputation process creates M matched pairs of completed binary outcomes. We estimated the proportion difference between the treatment and control arm, $p_1 - p_0$, where

$$p_1 = \frac{N_{b1}}{N_1},$$

$$p_0 = \frac{N_{b0}}{N_0},$$

N_{b1} is the number of individuals who have occurrence of hospitalization, transfusion, or plasmapheresis within 3 months after starting chemotherapy in the treatment group. N_{b0} is the number of individuals who have occurrence of hospitalization, transfusion, or plasmapheresis within 3 months after starting chemotherapy in the control group. N_1 is the total number of individuals in the treatment group. N_0 is the total number of individuals in the control group. The standard error for the difference in proportions for matched samples was used to obtain interval estimates (Fleiss, Levin, and Paik 2003).

6. Results

6.1 Propensity Score

Figures 3-5 summarize the standardized diffrences before subclassification (empty circles), as well as the average within class standardized differences (black triangles). All of the covariates display t statistics that are smaller than 1, suggesting good balance for all of the covariates.

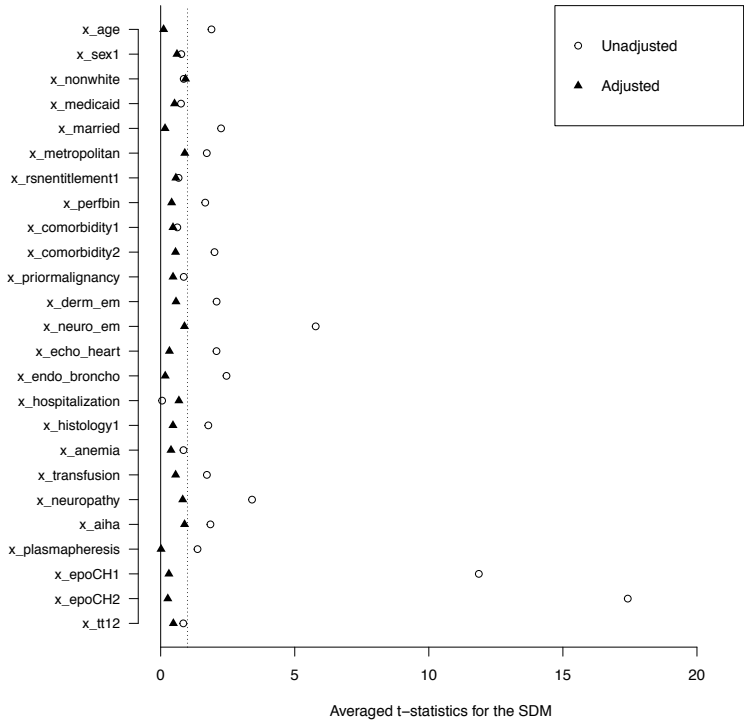

Figure 3. Standardized difference plots for regimens with versus without rituximab

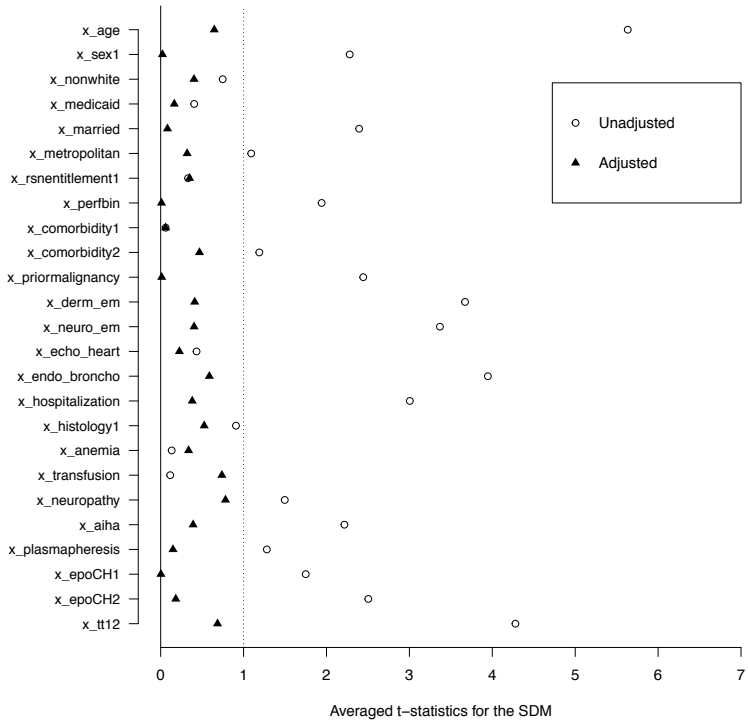

Figure 4. Standardized difference plots for rituximab monotherapy versus combination immunochemotherapy.

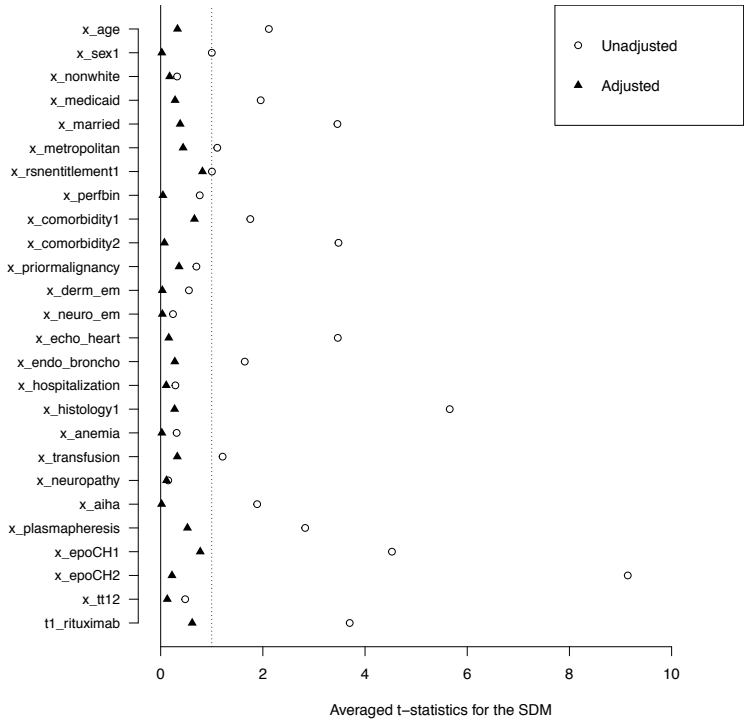

Figure 5. Standardized difference plots for purine analogue versus alkylator-based regimens

Tables 2-4 summarize the t-statistic for assessing global balance for the covariates across subclasses, and the F-Statistic for assessing balance for the covariates within all subclasses for the three different treatment comparisons. The largest of the absolute values of the t-statistics is 0.928, suggesting good balance for all covariates and all treatment comparisons. The largest absolute values of the F-test are 2.27, 2.24, 2.49, 2.04 and 2.15, with all the rest having values below 2.00, again suggesting good balance conditional on the propensity score (Imbens and Rubin 2015). In the first comparative analysis, only 1,301 individuals were included in the analysis, 9 individuals with propensity score greater than 0.995 or less than 0.25 were dropped because of inadequate overlap between the distributions of the covariates in the control and treatment groups. In the second comparative analysis, no individual with propensity scores outside the region of common support were truncated, so all 1,028 individuals were included in

the final analysis. In the third comparative analysis, no individual was truncated as all 599 units in the third comparative analysis have adequate overlap.

covariates	overall t_stat	F test
x_age	-0.110081043	-0.612005614
x_nonwhite	-0.608121988	0.175931733
x_married	-0.927691946	-0.724752165
x_medicaid	-0.517365745	0.809438864
x_perfbin	0.163068751	0.405755104
x_tt12	0.899631309	-0.894442336
x_metropolitan	0.566973672	0.590122203
x_priormalignancy	0.409495623	1.332055742
x_anemia	-0.460223095	-1.270395266
x_transfusion	-0.554485221	0.285849677
x_neuropathy	-0.461508194	0.418047036
x_aiha	-0.572270581	-0.546595167
x_plasmapheresis	0.88812943	1.64460369
x_hospitalization	-0.326719107	-1.83352944
x_derm_em	0.170136905	-0.374583769
x_neuro_em	0.680128848	-2.26644277
x_echo_heart	-0.458670628	-0.358318444
x_endo_broncho	-0.385704228	-0.122889766
x_histology1	-0.557783228	-2.238789223
x_sex1	0.822502194	0.385540809
x_comorbidity1	0.891347857	-0.212737345
x_comorbidity2	-0.014451729	-0.607452338
x_epoCH1	0.306798213	-0.93113532
x_epoCH2	0.266388908	0.514413674

x_rsnentitlement1	0.473525018	-0.849578084

Table 2. T-statistics and F test value – Final Propensity Score Specification (regimens with versus without rituximab)

covariates	overall t_stat	F test
x_age	0.646669927	-0.792365533
x_nonwhite	0.021898807	0.094002687
x_married	-0.402245994	0.774377904
x_medicaid	-0.164325036	-0.203075619
x_perfbin	0.08213384	-0.578053717
x_tt12	0.319289249	-1.579895988
x_metropolitan	-0.348448206	-0.730379345
x_priormalignancy	0.010326668	-0.703887177
x_anemia	-0.059517062	0.286938118
x_transfusion	0.468180448	0.855014809
x_neuropathy	0.012738014	-0.653471473
x_aiha	0.410076827	-2.491764691
x_plasmapheresis	0.402818365	-1.192525257
x_hospitalization	-0.225936397	0.284824705
x_derm_em	-0.587197611	-0.51037452
x_neuro_em	-0.381227734	-0.079951195
x_echo_heart	0.525353044	-1.474856674
x_endo_broncho	0.336629084	2.043343761
x_histology1	0.738672383	0.448765081
x_sex1	0.781663314	-0.649161752
x_comorbidity1	0.391004079	-0.313751195
x_comorbidity2	-0.148745123	0.430185275
x_epoCH1	0.003155045	-0.215881095

x_epoCH2	-0.182009858	0.124362802
x_rsnentitlement1	-0.685669875	0.025795923

Table 3. T-statistics and F test value – Final Propensity Score Specification (rituximab monotherapy versus combination immunochemotherapy)

covariates	**overall t_stat**	**F test**
t1_rituximab	-0.329396181	1.644315994
x_age	0.022743315	-1.629535061
x_nonwhite	-0.175474929	-0.027553104
x_married	-0.282035793	-0.749529113
x_medicaid	-0.381766137	-1.018148861
x_perfbin	-0.438975263	-0.200348663
x_tt12	-0.81824544	-0.850831958
x_metropolitan	0.044325511	-0.284620061
x_priormalignancy	-0.660872685	-1.116502959
x_anemia	0.073429816	-0.504683814
x_transfusion	0.360359405	-0.846060044
x_neuropathy	-0.032701842	-1.65677355
x_aiha	0.033338522	-1.026415693
x_plasmapheresis	-0.158866244	0.937035086
x_hospitalization	0.276759824	0.441076909
x_derm_em	-0.109045765	-0.27618413
x_neuro_em	0.272953225	-0.756620416
x_echo_heart	-0.023498452	-0.666739547
x_endo_broncho	0.325612455	-0.968289779
x_histology1	0.114745573	0.470612561
x_sex1	0.01849116	-2.153368459
x_comorbidity1	0.524317121	-0.876625681

x_comorbidity2	0.774865057	-0.629824612
x_epoCH1	0.221448365	-1.116294334
x_epoCH2	-0.129393265	-1.744560216
x_rsnentitlement1	-0.61688567	0.903487042

Table 4. T-statistics and F test value – Final Propensity Score Specification (purine analogue versus alkylator-based regimens)

6.2 Outcome Analysis

When comparing a regimen with rituximab to a regimen without rituximab, the mean difference in survival years between treatment group (with rituximab) and control group (without rituximab) is 2.60 (95% CI, 1.91-3.28); the paired hazard ratio between these groups is 0.62 (95% CI, 0.55-0.71). **Figure 6** displays the Kaplan-Meier survival curves for each of the regimens. Longer survival time and lower hazard rate have been observed when using a regimen with rituximab. No significant differences in the proportions of plasmapheresis and hospitalizations between two treatment groups was observed. Significantly lower proportion of transfusions was observed when using regimen with rituximab in comparison to a regimen without rituximab (**Table 5**).

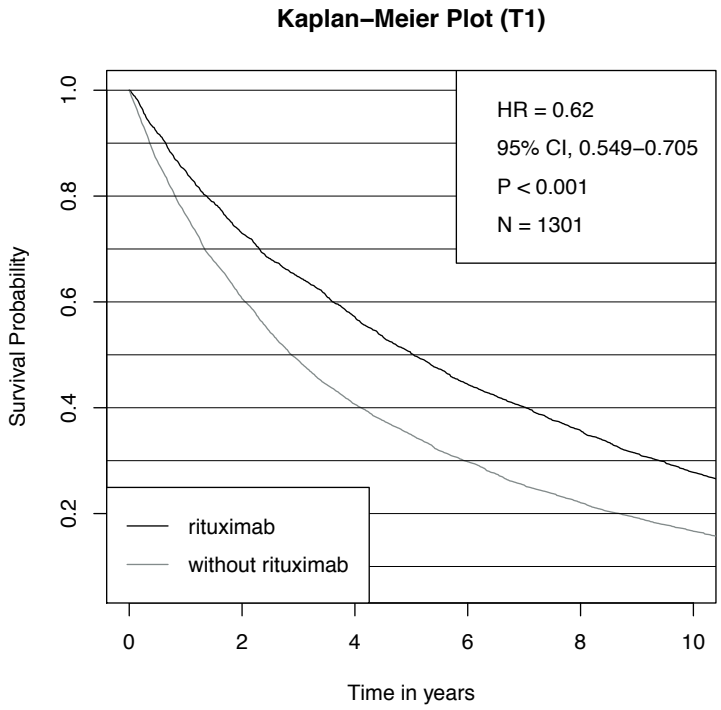

Figure 6. Kaplan-Meier Plot for regimens with versus without rituximab, after adjusting for covariates.

Outcome	with rituximab	without rituximab	risk difference	p-value
transfusion	0.195	0.228	-0.033 (95% CI: -0.063, -0.003)	0.022
plasmapheresis	0.048	0.040	0.008 (95% CI: -0.007, 0.022)	0.270
hospitalization	0.290	0.318	-0.028 (95% CI: -0.063, 0.008)	0.101

Table 5. Difference in proportions (regimen with rituximab versus without rituximab)

When comparing the use of rituximab alone to a combination with chemotherapy, the mean difference in survival years between the treatment group (rituximab alone) and the control group (rituximab in combination with chemotherapy) is -0.34 (95% CI: -1.39, 0.71); the paired hazard ratio between treatment group and control group is 0.91 (95% CI, 0.79-1.04). **Figure 7** depicts the Kaplan-Meier curves for both regimens. There are no significant differences in survival time and hazard rate between the two groups. Significantly lower proportion of transfusions, plasmapheresis and hospitalizations was observed when using a regimen with rituximab

monotherapy in comparison to a regimen composed of combination immunochemotherapy (**Table 6**).

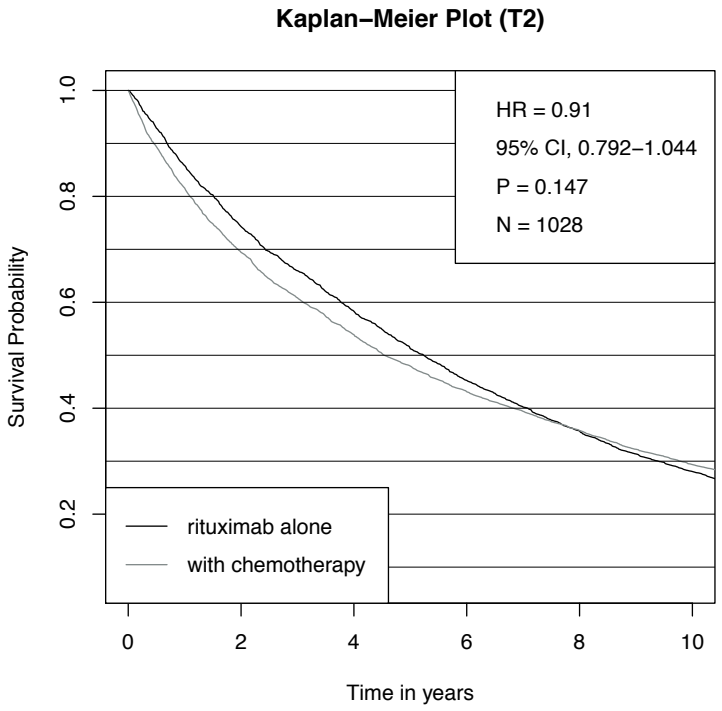

Figure 7. Kaplan-Meier Plot for rituximab monotherapy versus combination immunochemotherapy, after adjusting for covariates.

Outcome	rituximab alone	with chemotherapy	risk difference	p-value
transfusion	0.173	0.267	-0.094 (95% CI: -0.129, -0.059)	<0.001
plasmapheresis	0.028	0.061	-0.033 (95% CI: -0.050, -0.016)	<0.001
hospitalization	0.247	0.382	-0.135 (95% CI: -0.176, -0.094)	<0.001

Table 6. Difference in proportions (rituximab monotherapy versus combination immunochemotherapy)

When comparing a combination therapy regimen that is based on alkylators alone to a regimen based on purine, the mean difference in survival years between the purine group and the alkylators group is -1.10 (95% CI: -2.41, 0.21); the paired hazard ratio between treatment group and control group is 1.10 (95% CI, 0.92-1.32). **Figure 8** displays the Kaplan-Meier survival curves for the two treatments. No significant differences in survival time and hazard rate has

been observed between two groups. No significant differences in the proportions of transfusion, plasmapheresis and hospitalization between two treatment groups (**Table 7**).

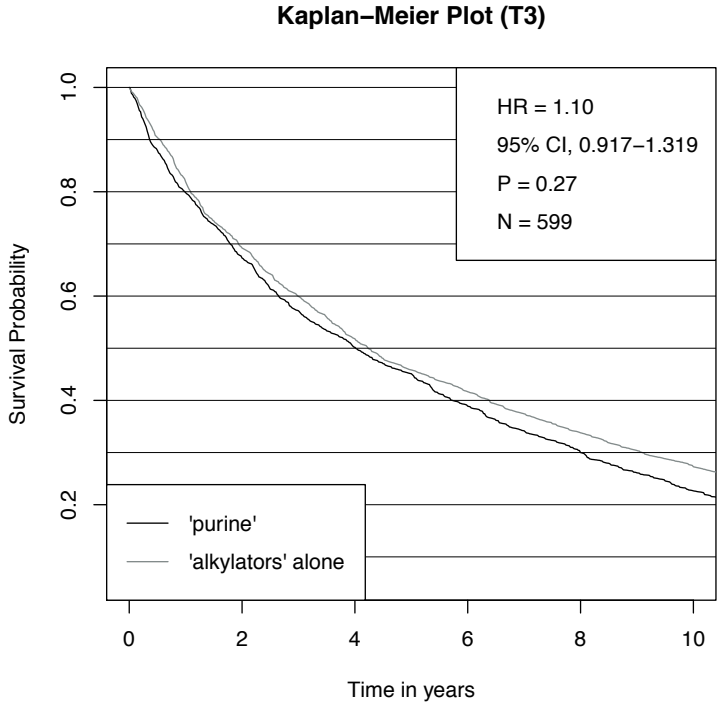

Figure 8. Kaplan-Meier Plot for purine analogue versus alkylator-based regimens, after adjusting for covariates.

Outcome	purines	alkylators	risk difference	p-value
transfusion	0.257	0.223	0.034 (95% CI: -0.013, 0.082)	0.131
plasmapheresis	0.071	0.063	0.008 (95% CI: -0.019, 0.035)	0.556
hospitalization	0.389	0.338	0.051 (95% CI: -0.006, 0.108)	0.061

Table 7. Difference in proportions (purine analogue versus alkylator-based regimens)

7. Discussion

Our analyses show that the median overall survival from the start of first-line chemotherapy for Medicare beneficiaries with WM/LPL is 4.72 years, 19.5% are transfused, 4.8% are treated with plasmapheresis, 29.0% are hospitalized within the first 3 months of treatment. Upfront rituximab (as monotherapy or combination immunochemotherapy) provides an overall survival benefit by increasing the average survival by 2.6 years. There was no significant difference in the studied toxicities of plasmapheresis and hospitalization between regimens with or without rituximab, but significantly lower proportion of transfusions is observed in the group treated with rituximab. In addition, the rituximab monotherapy provides an overall risk benefit in terms of plasmapheresis, hospitalization, and transfusions, though without a significant difference in overall survival. There was no significant difference in overall survival or in the studied toxicities between purine analogue-based and alkylator-based regimens.

We have extended MITSS to estimate treatment effects with time to event data. In our extension, we have imputed both the unobserved potential outcomes and the censored observations. This type of analysis, allows us to report any estimates of the treatment effect, and it allows investigators to examine potential moderators. To identify parametric survival model that provides a good fit for the data we examined the posterior predictive distribution for censored observations. This analysis was only performed on the control population to maintain objectivity of the analysis. Common parametric survival models do not provide a good fit for WM patients. The cure model that is composed from a mixture of long-term survivors and advanced disease patients modeled using the Weibull distribution provides a better fit. The validity of MITSS is based on the unconfounded assumption. Although, significant effort was employed to collect

many covariates or their proxies that are part of the assignment mechanism, some may have been left out. Future research should extend MITSS to provide sensitivity analysis for confounded assignment mechanism. In addition, future research should examine the use of other semi-parametric or nonparametric survival models.

In conclusion, we have extended MITSS to address time to event data. Our analysis shows that using rituximab extends WM patients survival without significant side effects.

Reference

Booth, C. M., and I. F. Tannock. "Randomised Controlled Trials and Population-Based Observational Research: Partners in the Evolution of Medical Evidence." *Br J Cancer* 110.3 (2014): 551-5. Print.

Buske, C., et al. "Waldenstrom's Macroglobulinaemia: Esmo Clinical Practice Guidelines for Diagnosis, Treatment and Follow-Up." *Ann Oncol* 24 Suppl 6 (2013): vi155-9. Print.

Dimopoulos, M. A., et al. "Treatment Recommendations for Patients with Waldenstrom Macroglobulinemia (Wm) and Related Disorders: Iwwm-7 Consensus." *Blood* 124.9 (2014): 1404-11. Print.

Fleiss, Joseph L., Bruce Levin, and Myunghee Cho Paik. *Statistical Methods for Rates and Proportions*. Wiley Series in Probability and Statistics. 3rd ed. Hoboken, N.J.: J. Wiley, 2003. Print.

Ghobrial, I. M. "Choice of Therapy for Patients with Waldenstrom Macroglobulinemia." *J Clin Oncol* 31.3 (2013): 291-3. Print.

Gutman, R., and D. Rubin. "Estimation of Causal Effects of Binary Treatments in Unconfounded Studies with One Continuous Covariate." *Stat Methods Med Res* (2015). Print.

Gutman, R., and D. B. Rubin. "Estimation of Causal Effects of Binary Treatments in Unconfounded Studies." *Stat Med* 34.26 (2015): 3381-98. Print.

Kapoor, P., et al. "Diagnosis and Management of Waldenstrom Macroglobulinemia: Mayo Stratification of Macroglobulinemia and Risk-Adapted Therapy (Msmart) Guidelines 2016." *JAMA Oncol* (2017). Print.

Korn, E. L., and B. Freidlin. "Methodology for Comparative Effectiveness Research: Potential and Limitations." *J Clin Oncol* 30.34 (2012): 4185-7. Print.

Martinez, E. Z., et al. "Mixture and Non-Mixture Cure Fraction Models Based on the Generalized Modified Weibull Distribution with an Application to Gastric Cancer Data." *Comput Methods Programs Biomed* 112.3 (2013): 343-55. Print.

Olszewski, A. J., S. P. Treon, and J. J. Castillo. "Evolution of Management and Outcomes in Waldenstrom Macroglobulinemia: A Population-Based Analysis." *Oncologist* (2016). Print.

Owen, R. G., et al. "Guidelines on the Diagnosis and Management of Waldenstrom Macroglobulinaemia." *Br J Haematol* 165.3 (2014): 316-33. Print.

Savignoni, A., et al. "Matching Methods to Create Paired Survival Data Based on an Exposure Occurring over Time: A Simulation Study with Application to Breast Cancer." *BMC Med Res Methodol* 14 (2014): 83. Print.

Stuart, E. A. "Matching Methods for Causal Inference: A Review and a Look Forward." *Stat Sci* 25.1 (2010): 1-21. Print.

Talarico, L., G. Chen, and R. Pazdur. "Enrollment of Elderly Patients in Clinical Trials for Cancer Drug Registration: A 7-Year Experience by the Us Food and Drug Administration." *J Clin Oncol* 22.22 (2004): 4626-31. Print.

Talaulikar, D., et al. "Treatment of Patients with Waldenstrom Macroglobulinaemia: Clinical Practice Guidelines from the Myeloma Foundation of Australia Medical and Scientific Advisory Group." *Intern Med J* 47.1 (2017): 35-49. Print.

Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983; 70:41–55.

Hastie T, Tibshirani R, Friedman JJ. The Elements of Statistical Learning, Vol. 1. Springer: New York, 2001.

Rubin DB, Thomas N. Af nely invariant matching methods with ellipsoidal distributions. The Annals of Statistics 1992:1079–1093.

Imbens, Guido W., and Donald B. Rubin. Causal inference in statistics, social, and biomedical sciences. Cambridge University Press, 2015.

Barnard J, Rubin DB. Small-sample degrees of freedom with multiple imputation. Biometrika 1999; 86(4):948–955.

Balov, Nikolay. "Gelman–Rubin convergence diagnostic using multiple chains." The Stata Blog. N.p., 26 May 2016. Web.

Hirano, Keisuke, and Guido W. Imbens. "Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization." Health Services and Outcomes research methodology 2.3 (2001): 259-278.

Rubin, Donald B. "For objective causal inference, design trumps analysis." The Annals of Applied Statistics (2008): 808-840.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. The Annals of Statistics, 6, 34–58.

Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol 1974; 66: 688.

Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. Statistics in Medicine 2007; 26:20–36.

Rubin DB. For objective causal inference, design trumps analysis. The Annals of Applied Statistics 2008; 2(3):808–840.

Othus, Megan, et al. "Cure models as a useful statistical tool for analyzing survival." *Clinical*

Cancer Research 18.14 (2012): 3731-3736.

Gelman, Andrew, and Donald B. Rubin. "Inference from iterative simulation using multiple sequences." *Statistical science* (1992): 457-472.

Little, R. and D. Rubin. 1987. Statistical Analysis with Missing Data. New York: John Wiley & Sons.

Boag, J. W.: Maximum likelihood estimates of the proportion of patients cured by cancer therapy, J. R. Stat. Soc. Ser. B Stat. Methodol. 11 (1949), 15–44.