EVALUATING A FACILITY-PROFILING METRIC BASED ON SURVIVAL PROBABILITY APPLICATION TO U.S. TRANSPLANT CENTERS

Amelia H. Tran

A THESIS

in

Epidemiology and Biostatistics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Masters of Science

2023

Supervisor of Thesis

Douglas E. Schaubel, Professor of Biostatistics

Graduate Group Chair

Russell T. Shinohara, Professor of Biostatistics

Thesis Committee

Douglas E. Schaubel, Professor of Biostatistics

Peter P. Reese, Professor of Medicine and Epidemiology

Ian J. Barnett, Assistant Professor of Biostatistics

ACKNOWLEDGEMENT

First and foremost, my sincere appreciation goes to my academic and thesis advisor, Dr. Doug Schaubel, for his dedicated mentorship and encouragement throughout my graduate studies. Doug has always been there to support and look out for me every step of the way. I am grateful for his kindness and patience. My passion for survival analysis and causal inference gets bigger because of all the exciting projects I worked on under his mentorship. Many thanks to Drs. Peter Reese and Ian Barnett for kindly volunteering their time to read and give thoughtful and informative feedback on my work.

I would like to express my gratitude to Dr. Kay-See Tan from Memorial Sloan Kettering Cancer Center for her constant emotional support. Kay-See has been very sweet to let me bother, especially when I was going through my worst. The encouragement and patience that she gave was what kept me going thus far. I would like to thank Dr. Jacek Urbanek for his guidance and mentorship during my internship at Regeneron Pharmaceuticals, where I was first introduced to accelerometry data. I thank Jacek for his encouragement and words of inspiration to help me shift my perspectives, and look at any setback as a learning opportunity.

I would like to thank my 'big brothers' Dane Isenberg, Ian Delorey, JJ Zhang, and Haotian Zheng in the Department of Biostatistics at Penn. I thank Dane for his valuable advice and perspectives on academic matters or otherwise, and for tolerating my occasional last-minute request for help on homework assignments. I thank Ian for letting me bother with emergency phone calls to discuss my quarter life crisis. JJ has been kind to give me valuable perspectives on my career decisions, and to share with me his knowledge about the field. Haotian has been generous with his time to explore restaurants in Philly with me, and never fail to help me find humour in my grad school experiences. I also would like to thank my

graduate fellows in the department.

I would like to thank my good friend Hoang-Anh Phan in the Department of Chemistry at Penn, and my buddies Meng Xu and Quang Nguyen at Regeneron. I thank Hoang-Anh for being there for me despite her fully packed schedules, and for introducing me to Penn GAPSA where I got to interact and make friends with so many interesting peeps. Our mundane conversations have brought me so much joy and positivity! I thank Meng and Quang for giving me advice and perspectives on my career options, and making time to chat about any biostatistics and philosophy topics that I came up with throughout my summer internship and afterwards.

Last but certainly not least, I am thankful for my undergraduate advisor, Professor Margaret Robinson, and my Mount Holyoke friends. I appreciate Margaret for always giving me a safe haven to come back to, for either motherly advice or a guiding force. I thank my friends Heidi Zhang and Maria Maria Castillo for always being there for me, and for emboldening me to make no apologies or regrets on any decision that I make, dust myself off, and move forward. I stay brave and optimistic because of you all, and am grateful for you all because of that!

ABSTRACT

EVALUATING A FACILITY-PROFILING METRIC BASED ON SURVIVAL PROBABILITY APPLICATION TO U.S. TRANSPLANT CENTERS

Amelia H. Tran

Douglas E. Schaubel

The performance of health care providers and medical centers is of great interest to patients, surgeons, insurance companies, and regulatory organizations. In evaluation of different facilities or centers on survival outcomes, the standardized mortality ratio (SMR), which compares the observed to expected mortality, is arguably most widely used, especially for kidney transplant centers. Despite its wide acceptance, the SMR has certain limitations with respect to estimation stability and clinical interpretability which act as motivation for alternative evaluation metrics. In particular, we consider a novel prognostic score-based weighting approach. In this project, we use data from United Network for Organ Sharing (UNOS) to evaluate kidney transplant centers using the afore-mentioned prognostic score-based method along with the SMR. We detail the limitations of the SMR for center evaluation in the kidney transplant setting and how these shortcomings can be overcome by the prognostic score based weighting approach. Finally, we discuss the potential reasons for the discrepancy between the two evaluation metrics and recommend clinical settings where the latter prognostic score approach is most appropriate.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1

INTRODUCTION

1.1. Background

The last two decades have witnessed unprecedented increases in the availability of data on patient outcomes that provide essential information to plan the most optimal treatment and improve quality of care. Examples of patient-reported outcomes include description of their symptoms; satisfaction with care; disease progression; and physical or mental well-being. Other examples of patient outcomes include information about side effects of treatment under investigation in clinical trials. This has fueled the increased scrutiny and performance monitor of medical providers for providing highest standard care to patients and better management in healthcare facilities.

Evaluations of healthcare providers and medical centers are of great interest to different parties since they provide critical information for decision making to patients, transplant professionals and medical practitioners, as well as insurance companies and regulatory bodies (Dickinson et al., 2006). For example, patients and families may use these findings to decide between transplant centers with positive experience among patients of similar conditions, whereas transplant surgeons may use them quality control mechanisms. In other instances, insurance companies or the Centers for Medicare and Medicaid Services (CMS) may use center evaluations to ensure quality care for patients, while regulatory bodies within transplantation networks may use center-specific evaluations to direct remedial actions or further investigation of current policies.

1.2. Motivation

In the setting of transplantation, medical researchers can use a simple framework of quality in two dimensions: whether the center or "system" delivers the care to patients in need the most and whether these patients achieve good outcomes after the care is delivered. There is concern about transplant centers having the incentives to deny care to patients more likely to have unfavorable outcomes after transplant, and work that discusses the necessity to improve the validity of profiling metrics based on pre-transplant outcomes or the combination of both outcomes prior to and after transplant (Jay and Schold, 2017). However, our work focuses on evaluating center' performance upon the receipt of kidney transplants, mainly determined by post-transplant outcomes.

In the end-stage renal disease setting, post-transplant outcomes by transplant center is an important factor of ensuring highest-quality care for patients. One common type of outcome used for such evaluation is the time until events, also known as survival time. In the evaluation of kidney transplant centers on survival outcomes, the most frequently used measure of mortality among existing facility profiling methods is the standardized mortality ratio (SMR) which compares the center's observed to expected mortality (Wolfe, 1994). In particular, kidney transplant centers in the U.S. are subject to two evaluations: one by the Organ Procurement and Transplantation Network (OPTN) and one by the Centers for Medicare and Medicaid Services (CMS). In both cases, the most scrutinized metric is standardized mortality ratio (SMR) for one-year graft failure (GF), with GF defined as the earliest of death and transplant failure (which is said to occur when the transplanted organ is not functioning well enough to sustain life).

Despite its utility and wide acceptance among the kidney transplantation community, the SMR is not well-suited for evaluating centers due to certain limitations, such as exaggerat-

ing center effects in settings where survival probability is relatively high. In addition, the SMR can be ill-defined if the underlying assumption for the survival model to obtain expected mortality is violated, i.e., proportionality assumption for Cox regression. Moreover, the SMR is based on an indirect standardization method as it involves averaging across the center-specific covariate distributions. These properties preclude the validity of direct SMR comparison across transplant centers and serve as motivation for a novel approach to overcome these shortcomings.

Previous studies on evaluation of healthcare facilities and medical centers have also mentioned these issues and proposed several alternatives. Issues with covariate case-mix distribution when estimating SMRs have been discussed in studies on the performance of Australian and New Zealand intensive care units and on pediatric intensive care (Kasza et al., 2013; Manktelow et al., 2014). Richardson et al. detailed the interpretation issue with SMRs due to non-comparability of the occupational cohort and reference population (Richardson et al., 2015). Pouw et al. assessed the validity and applicability of hospital standardized hospital mortality ratios and recommended investigating the potential interaction between hospital and case-mix to avoid misinterpretation of SMRs (Pouw et al., 2013).

Lee et al. propose a novel prognostic score-based weighting method for estimating center effects which offer important advantages over SMR (Lee et al., 2023+). Some of them include clinically meaningful interpretation based on the difference in survival compared to the population average, robust estimation to model mis-specification or violation of proportionality assumptions, and direct standardization. In addition, this method also accommodates for unequal covariate distribution across centers which facilitates valid and accurate comparisons.

The remainder of the thesis is organized as follows. In the next chapter, we describe the data

with summary statistics and univariate analysis. We explain the SMR and prognostic score-based weighting methods for center effect measures in Chapter 3. In Chapter 4, we apply the standardized and proposed methods to evaluate the U.S. kidney transplant centers. Chapter 5 concludes the thesis with the discussion on the strengths and weaknesses of the facility-profiling metric based on survival probability with respect to the standardized measure.

CHAPTER 2

DATA DESCRIPTION

2.1. The Data

We used data from the United Network for Organ Sharing (UNOS) with the goal to evaluate U.S. kidney transplant centers with respect to 1-year graft survival. Time-to-event is defined as the time between transplantation and the earliest of death, return to dialysis or repeat transplantation. The study population includes 65,266 patients who received deceased-donor kidney transplants at age >= 18 between 1/1/2016 and 12/31/2020, which is longer than the Scientific Registry of Transplant Recipients (SRTR) uses. Among these patients, there are 6,913 patients experienced graft failures. The censoring rate is around 83%.

There were 242 transplant centers in total. However, we excluded centers with fewer than 25 transplants to avoid instability issues, which left the data set with 201 centers. Center sizes ranged from 25 to 151 patients with the median of 325 patients for each center. We truncated the data at 1-year post-transplant by censoring any patients receiving kidney transplants after the first year. Information on the event occurrence after year 1 is not accounted for in this one year analysis.

The data set includes center indicator, recipient covariates such as recipient age, sex, race, calendar year of transplant, primary renal diagnosis, years between wait-listing and kidney transplant, years on dialysis to wait-listing, diabetes status, body mass index, blood types, hepatitis C virus (HCV) status, chronic obstructive pulmonary disease (COPD), hypertension, malignant tumor indicator, insurance mechanism used to pay for the transplant, and Kidney Donor Risk Index (KDRI) which is a continuous comprehensive score to quantify graft failure risk by combining donor and transplant variables (Rao et al., 2009).

2.2. Data Exploration

Table 2.1 shows the summary statistics of deceased donor characteristics including human leukocyte antigen mismatch, terminal creatinine level, age, gender, race, HCV status, height, weight, cold ischemia time, diabetes and KDRI between patients with and without graft failure. The results show two groups are comparable with respect to deceased donor characteristics.

Summary Statistics of Donor Characteristics			
Characteristics	**N = 65,266**	**Without GF = 58,353**	**With GF = 6,913**
AMIS: 0	6,606 (10%)	5,984 (10%)	622 (9.0%)
1	25,381 (39%)	22,677 (39%)	2,704 (39%)
2	33,279 (51%)	29,692 (51%)	3,587 (52%)
BMIS: 0	3,967 (6.1%)	3,621 (6.2%)	346 (5.0%)
1	15,940 (24%)	14,272 (24%)	1,668 (24%)
2	45,359 (69%)	40,460 (69%)	4,899 (71%)
DRMIS: 0	9,514 (15%)	8,678 (15%)	836 (12%)
1	31,566 (48%)	28,279 (48%)	3,287 (48%)
2	24,186 (37%)	21,396 (37%)	2,790 (40%)
Creatinine level	0.94 (0.70, 1.40)	0.94 (0.70, 1.40)	0.98 (0.70, 1.42)
Age	39 (28, 51)	39 (27, 51)	44 (31, 54)
Sex: male	40,200 (62%)	36,126 (62%)	4,074 (59%)
female	25,066 (38%)	22,227 (38%)	2,839 (41%)
Race: non-black	56,443 (86%)	50,612 (87%)	5,831 (84%)
black	8,823 (14%)	7,741 (13%)	1,082 (16%)
HCV status: No	61,869 (95%)	55,237 (95%)	6,632 (96%)
Yes	3,397 (5.2%)	3,116 (5.3%)	281 (4.1%)
Height (cm)	172 (163, 178)	173 (164, 179)	170 (163, 178)
Weight (kg)	81 (68, 96)	81 (68, 96)	80 (68, 96)
Cold ischemia time	17 (11, 22)	17 (11, 22)	18 (12, 24)
Diabetes: No	60,259 (92%)	54,111 (93%)	6,148 (89%)
Yes	5,007 (7.7%)	4,242 (7.3%)	765 (11%)
KDRI	1.20 (0.98, 1.49)	1.19 (0.98, 1.48)	1.30 (1.06, 1.61)

Table 2.1: Summary Statistics of Donor Characteristics

Table 2.2 shows the summary statistics of recipient characteristics including age, gender, race, HCV status, height in cm, weight in kg, diabetes, blood group, malignancy condition, years on waitlist before kidney transplants, and years on dialysis prior to waitlist between patients experiencing graft failure and ones without the event. Median age of patients with graft failure is higher by 4 years. There are more male and Hispanic patients with graft failure, both by 4%.

Summary Statistics of Recipient Characteristics			
Characteristics	**N = 65,266**	**Without GF = 58,353**	**With GF = 6,913**
Age	56 (45, 64)	56 (45, 64)	60 (50, 67)
Sex: male	39,460 (60%)	35,044 (60%)	4,416 (64%)
female	25,806 (40%)	23,309 (40%)	2,497 (36%)
Race: Asian	4,980 (7.6%)	4,590 (7.9%)	390 (5.6%)
Hispanic	12,706 (19%)	11,548 (20%)	1,158 (17%)
Black	21,704 (33%)	19,137 (33%)	2,567 (37%)
Others	25,876 (40.4%)	23,078 (39.1%)	2,798 (40.4%)
HCV status: No	64,741 (99%)	57,877 (99%)	6,864 (99%)
Yes	525 (0.8%)	476 (0.8%)	49 (0.7%)
Height (cm)	170 (163, 178)	170 (163, 178)	170 (163, 178)
Weight (kg)	81 (69, 95)	81 (68, 95)	83 (70, 97)
Diabetes: No	39,843 (61%)	36,293 (62%)	3,550 (51%)
Yes	25,423 (39%)	22,060 (38%)	3,363 (49%)
Blood: A	22,854 (35%)	20,436 (35%)	2,418 (35%)
AB	3,395 (5.2%)	3,075 (5.3%)	320 (4.6%)
B	9,394 (14%)	8,396 (14%)	998 (14%)
O	29,623 (45.8%)	26,446 (45.7%)	3,177 (46.4%)
Malignancy: No	59,273 (91%)	53,106 (91%)	6,167 (89%)
Yes	5,993 (9.2%)	5,247 (9.0%)	746 (11%)
Years on waitlist	1.51 (0.33, 3.74)	1.52 (0.34, 3.73)	1.46 (0.28, 3.78)
Years on dialysis	1.11 (0.00, 3.20)	1.08 (0.00, 3.14)	1.40 (0.28, 3.64)

Table 2.2: Summary Statistics of Recipient Characteristics

Table 2.3 shows the univariate analysis of deceased donor characteristics with respect to time until graft failure. From the univariate analysis, terminal creatinine level, donor HCV status, and weight are statistically significant at the level $\alpha = 0.05$.

Univariate Analysis of Donor Characteristics			
Characteristics	**Hazard Ratio**	**95% Confidence Interval**	**p-value**
AMIS: 0	*Ref*	-	0.003
1	1.14	(1.05, 1.25)	
2	1.15	(1.06, 1.26)	
BMIS: 0	*Ref*	-	<0.001
1	1.20	(1.07, 1.35)	
2	1.25	(1.12, 1.40)	
DRMIS: 0	*Ref*	-	<0.001
1	1.21	(1.12, 1.31)	
2	1.36	(1.26, 1.47)	
Creatinine level	1.00	(0.98, 1.02)	0.8
Age/5	1.02	(1.02, 1.02)	<0.001
Sex: male	*Ref*	-	<0.001
female	1.11	(1.05, 1.16)	
Race: non-black	*Ref*	-	<0.001
black	1.20	(1.12, 1.28)	
HCV status: No	*Ref*	-	0.2
Yes	0.92	(0.82, 1.04)	
Height/10 (cm)	0.99	(0.99, 1.00)	<0.001
Weight/5 (kg)	1.0	(1.0, 1.0)	0.2
Cold ischemia time	1.01	(1.01, 1.02)	<0.001
Diabetes: No	*Ref*	-	<0.001
Yes	1.64	(1.52, 1.76)	
Log(KDRI)	2.92	(2.70, 3.17)	<0.001

Table 2.3: Univariate Analysis of Donor Characteristics

Table 2.4 shows the univariate analysis of recipient characteristics with respect to time until graft failure. From the univariate analysis, recipient HCV status and blood are statistically significant at the level $\alpha = 0.05$. The summary tables (median and IQR) and univariate analyses for Cox regression on time until graft failure are generated using the R package 'gtsummary' (Sjoberg et al., 2021).

Univariate Analysis of Recipient Characteristics			
Characteristics	**Hazard Ratio**	**95% Confidence Interval**	**p-value**
Age/5	1.02	(1.02, 1.03)	<0.001
Sex: male	*Ref*	-	<0.001
female	0.85	(0.81, 0.89)	
Race: Asian	*Ref*	-	<0.001
Hispanic	1.17	(1.04, 1.31)	
Black	1.51	(1.36, 1.68)	
Others	1.40	(1.26, 1.56)	
HCV status: No	*Ref*	-	0.3
Yes	1.16	(0.88, 1.54)	
Height/10 (cm)	1.90	(1.50, 2.41)	<0.001
Weight/5 (kg)	1.16	(1.12, 1.20)	<0.001
Diabetes: No	*Ref*	-	<0.001
Yes	1.56	(1.49, 1.64)	
Blood: A	*Ref*	-	0.2
AB	1.01	(0.94, 1.09)	
B	0.90	(0.80, 1.01)	
O	1.02	(0.97, 1.08)	
Malignant tumor: No	*Ref*	-	<0.001
Yes	1.23	(1.14, 1.33)	
Years on waitlist	0.99	(0.98, 1.00)	0.032
Years on dialysis	1.03	(1.02, 1.03)	<0.001

Table 2.4: Univariate Analysis of Recipient Characteristics

CHAPTER 3

STATISTICAL METHODS

3.1. Standardized Mortality Rate (SMR)

A widely accepted measure for center profiling, the standardized mortality ratio (SMR) is defined as the ratio of a given center's observed to expected number of events. In the context of our analysis, the observed number of events is the number of GFs at each center, while the expected number of events is obtained under the assumption that the center had experienced standard population average event rates (Borgan and Langholz, 1993).

In practice, the standard population mortality rate is often unknown. To relax this requirement, a statistical model for survival outcome is used to calculate the expected number of mortality events for the SMR from the observed data through Cox regression, which is sometimes referred to as semi-parametric Cox SMR (He and Schaubel, 2015). In our study, the event is defined as one-year graft failure (GF), with GF defined as the earliest of death and transplant failure. We then define survival with a functioning graft as the time between transplantation and GF.

We now establish notation to be used for the remainder of the thesis. Let i be index subject ($i = 1, ..., n$) and let j index center ($j = 1, ..., J$). Let T_i and C_i denote the event and censoring times, respectively. Thus, $\delta_i = I(T_i \leq C_i)$ denotes the observed event indicator and let $U_i = \min(T_i, C_i)$ represent the observed follow-up time for subject i. In addition, let G_i center for subject i, with $G_{ij} = I(G_i = j)$ representing a center j membership indicator. Let $\mathbf{X}_i$ denote a vector of baseline covariates for subject i and $Y_i(t) = I(U_i \geq t)$ denote at-risk indicator. We assume that observed data $(U_i, \delta_i, \mathbf{X}_i, G_i)$ are independent and identically distributed (i.i.d) across $i = 1, ..., n$.

With this setup, the center-specific SMR has the structure

$$\text{SMR}_j = \frac{O_j}{E_j} \tag{3.1}$$

where the observed O_j and expected E_j number of mortality events is given by

$$\begin{aligned} O_j &= \sum_{i=1}^{n} G_{ij} N_i(t) \\ E_j &= \sum_{i=1}^{n} G_{ij} \int_0^t Y_i(u) d\Lambda_{ij}(u) \end{aligned} \tag{3.2}$$

with $Y_i(u) = I(U_i \geq u)$ as the at-risk indicator at time u, $N_i(t) = \delta_i I(U_i \leq t)$ as the observed event by time t, and $d\Lambda_{ij}(u) = P(u \leq T < u + du \mid T \geq u)$ as the cumulative hazard increment at time u for subject i at center j.

Define the Martingale residuals and the corresponding increment as

$$\begin{aligned} M_{ij}(t) &= N_i(t) - \int_0^t Y_i(u) d\Lambda_{ij}(u) \\ dM_{ij}(t) &= dN_i(t) - Y_i(t) d\Lambda_{ij}(t) \\ \sum_{i=1}^{n} G_{ij} M_{ij}(t) &= \sum_{i=1}^{n} G_{ij} N_i(t) - \sum_{i=1}^{n} G_{ij} \int_0^t Y_i(u) d\Lambda_{ij}(u) \end{aligned} \tag{3.3}$$

We obtained the expected number of mortality events as

$$E_j = O_j - \sum_{i=1}^{n} G_{ij} M_{ij}(t) \tag{3.4}$$

where the observed O_j and expected E_j number of mortality events are defined as in 3.2.

The SMR variance could be obtained with a Poisson variance assumption

$$\begin{aligned} V(\text{SMR}_j) &= V\left(\frac{O_j}{E_j}\right) = E_j^{-2}V(O_j) = E_j^{-1} \\ V(\log(\text{SMR}_j) &= \frac{1}{\text{SMR}_j^2}V(\text{SMR}_j) = \frac{1}{E_j\text{SMR}_j^2} \end{aligned} \tag{3.5}$$

We can use this to determine center effects based on normal distribution

$$Z_j = \frac{\log(\text{SMR}_j)}{SE_j} \sim N(0,1) \tag{3.6}$$

where $SE_j = \sqrt{V\{\log(\text{SMR}_j)\}}$.

Despite its wide acceptance, the SMR is not well-suited in kidney transplantation. Some of the limitations include its close correspondence to the modeling assumption of proportionality of center-specific hazard functions. Moreover, the SMR is likely to exaggerate center effects when survival probability is relatively high. For example, in a comparison of two groups with high survival of 0.7 and 0.8 respectively, the two groups would have high SMRs but low difference in survival probability. In other words, small potential bias in the expected mortality could result in a highly biased SMR. Addition, the expected mortality is averaged over the center-specific case mix covariate distribution. This essentially precludes the validity of the SMR as a comparison metric. Finally, the interpretation of the SMR is suboptimal compared to other metrics with survival probabilities which are more intuitive and understandable. These limitations act as motivations for a novel metric to estimating center effects that is more interpretable and clinically meaningful than the SMR, and more robust to model mis-specification.

3.2. Prognostic Score-Based Weighting

Lee et al. propose a novel prognostic based weighting approach to estimate center effects in terms of differences in survival probability, which compares each center versus a reference population (Lee and Schaubel, 2022). The prognostic score was originally established as an alternative to propensity score in observational studies, defined as the association between observed covariates and potential outcome in the placebo or control group (Hansen, 2008). The prognostic score can be used as a balancing score through subclassification, matching, or weighting in similar ways to the propensity score.

The prognostic score is arguably a better choice than the propensity score in settings where the number of treatment groups is large. In such cases, there is little overlap among propensity score distribution among treatment groups, or researchers are interested in removing the systematic association between covariates and the outcome. The use of prognostic score has been studied in previous literature including the joint use of prognostic and propensity score on the estimation of treatment effects (Leacy and Stuart, 2014). Other study focuses on the use of prognostic scores for causal inference with general treatment regimes (Nguyen and Debray, 2019).

To estimate the prognostic score, Hansen suggested modeling only within the placebo or treatment) group (Hansen, 2008). In this context of evaluating center effects, the novel approach considers using the prognostic score which is the association between $\boldsymbol{X}_i$ and T_i, after filtering basic differences due to center effects. With respect to the setup of our interest, there are essentially many 'treatment' groups corresponding to transplant centers. However, under the assumption of equal covariate effects across centers implies that prognostic scores could be estimated using any center.

Lee et al. proposes a more robust approach using stratified partial likelihood to estimate

the prognostic score (Lee et al., 2023+). This way, all patients are leveraged and any center can act as the reference group. This approach focuses on the association between event time and observed covariates, after factoring out baseline differences due to center effects via stratification. While Hansen proposes to use one reference group (control or treatment) to obtain prognostic score to block out bias, the novel prognostic score approach rules out potential bias from center memberships through stratified partial likelihood.

As such, we estimate the prognostic score based on a semi-parametric center-stratified Cox model, where the baseline is unspecified and center-specific

$$\lambda_{ij}(t; \boldsymbol{X}_i) = \lambda_{0j}(t) \exp(\beta^T \boldsymbol{X}_i) \tag{3.7}$$

The estimated prognostic scores $\eta(\mathbf{X}_i) = \beta^T \mathbf{X_i}$ are continuous and can be used to construct in risk classes R through their quantiles. In our study, we construct $R = 5$ risk classes using the quintiles of the prognostic score $\eta(\boldsymbol{X}_i)$. Let $Q_i = r$ denote risk class membership where $r = 1, ..., 5$ and $Pr(Q_i = r) = 0.2$ for all r. Weight for each subject is then constructed

$$\widehat{w}_{ij} = G_{ij} Q_{ir} \frac{n_j}{n_{jr}} \widehat{p}_r \tag{3.8}$$

where $\widehat{p}_r = n^{-1} \sum_{i=1}^{n} Q_{ir}$ and $n_{jr} = \sum_{i=1}^{n} G_{ij} Q_{ir}$ with $Q_{ir} = I(Q_i = r)$. Intuitively, each of the subjects in the n_{jr} group in the risk class r of center j represents $n_j p_r / n_{jr}$ population, where $n_j p_r = n_j$ represents the number of subjects at center j. Consequently, summing over all risk classes r = 1, 2, . . . , R can guarantee the center size of n_j even after weighting.

The weighted Nelson-Aalen estimator of the cumulative hazard for center j is

$$\begin{aligned} \widehat{\Lambda}_j^w(t) &= \sum_{r=1}^{R}\sum_{i=1}^{n}\int_0^t \widehat{\pi}_j(u)^{-1}\widehat{w}_{ij}dN_{ijr}(u) \\ \text{where } \widehat{\pi}_j(u) &= \sum_{r=1}^{R}\sum_{i=1}^{n}\widehat{w}_{ij}Y_{ijr}(u) \end{aligned} \tag{3.9}$$

Group-specific weighted survival function is then given by

$$\widehat{S}_j^w(t) = \exp(-\widehat{\Lambda}_j^w(t)) \tag{3.10}$$

For $j = 1, ...J$, the estimator for difference in survival probability is as follows

$$\widehat{\tau}_j(t) = \widehat{S}_j^w(t) - J^{-1}\sum_{m=1}^{J}\widehat{S}_m^w(t) \tag{3.11}$$

The property of our interest, the variance of the weighted survival, $V\{\widehat{S}_j^w(t)\}$ can be approximated by the variance of the weighted Nelson-Aalen estimator $V\{\widehat{\Lambda}_j^w(t)\}$ through the Delta method

$$V\{\widehat{S}_j^w(t)\} = \widehat{S}_j^2(t)V\{\widehat{\Lambda}_j^w(t)\} \tag{3.12}$$

where $\widehat{S}_j(t) = \exp\{-\widehat{\Lambda}_j^w(t)\}$ in (3.10).

There are multiple variance estimators for $V\{\widehat{\Lambda}_j^w(t)\} = \sigma_j^2(t)$. The Martingale-based variance estimator is

$$\widehat{\sigma}_j^2(t) = \sum_{i=1}^{n}\int_0^t \left\{\sum_{l=1}^{n}\widehat{w}_{lj}Y_{lj}(u)\right\}^{-2}\widehat{w}_{ij}dN_{ij}(u) \tag{3.13}$$

The Huffer's adaptation of the Martingale-based variance estimator is

$$\widehat{\sigma}_j^2(t) = \sum_{i=1}^{n} \int_0^t \left\{ \sum_{l=1}^{n} \widehat{w}_{lj} Y_{lj}(u) \right\}^{-3} \left\{ \sum_{k=1}^{n} \widehat{w}_{kj}^2 Y_{kj}(u) \right\} \widehat{w}_{ij} dN_{ij}(u) \tag{3.14}$$

Finally, the estimating equations based variance estimator can be derived as

$$\widehat{\sigma}_j^2(t) = \sum_{i=1}^{n} \left[\int_0^t \left\{ \sum_{l=1}^{n} \widehat{w}_{lj} Y_{lj}(u) \right\}^{-1} \widehat{w}_{ij} dM_{ij}(u) \right]^2 \tag{3.15}$$

where $dM_{ij}(u) = dN_{ij}(u) - Y_{ij}(u) d\Lambda_j(u)$.

In our study, we use Huffer's formula to obtain the variance estimator $\widehat{\sigma}_j^2(t)$ because it produces 95% confidence intervals with the most optimal coverage rate.

Through simulation studies, prognostic scores estimated from a Cox model exhibit robustness against model misspecification since the assumed model is used to generate risk classes as opposed to fitted-value based 'expected' counts (Lee et al., 2023+). Thus, the prognostic score-based weighting method is less reliant on the modeling assumption of proportionality. The prognostic risk score is obtained based on the covariate distribution of the study population, making it comprehensive and valid for comparison purposes. Moreover, the interpretation of estimator $\widehat{\tau}_j(t)$ is intuitive and straightforward as the difference in survival probability compared to the population-averaged survival.

CHAPTER 4

RESULTS

4.1. Standardized Mortality Rate (SMR)

We also used the standard method standardized mortality ratio (SMR) to obtain mortality rate for each transplant center, as a means to determine better or worse centers. Figure 4.1 shows the histogram of center-specific expected survival in percentage which ranges from 0.8 to 1.0. The majority of expected survivals range from 0.92 to 0.96.

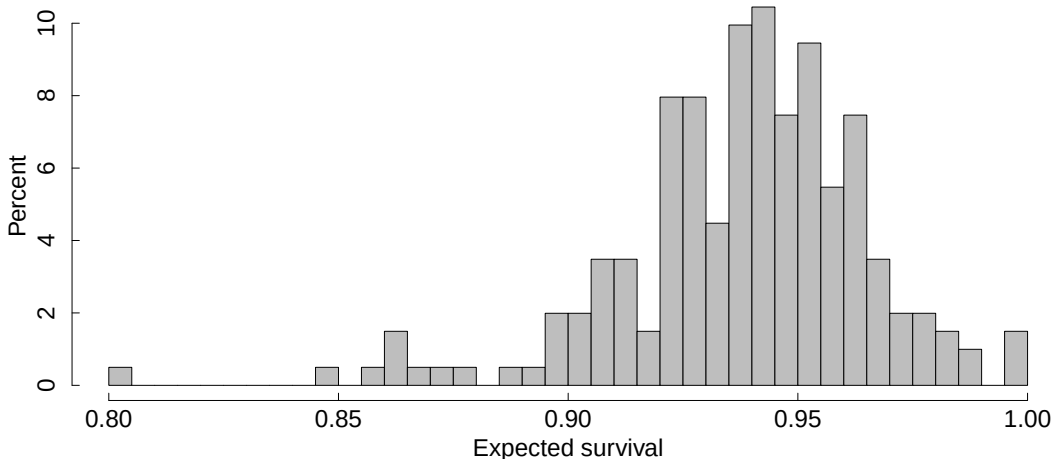

Figure 4.1: Histogram of expected survival in percentage

Figure 4.2 shows the histogram of log SMR with many centers having log SMR of 0. Similarly to τ, the tails of log SMR distribution indicate centers being significantly different from the overall population, with the left end (smaller than 0) representing better centers and the opposite direction representing worse.

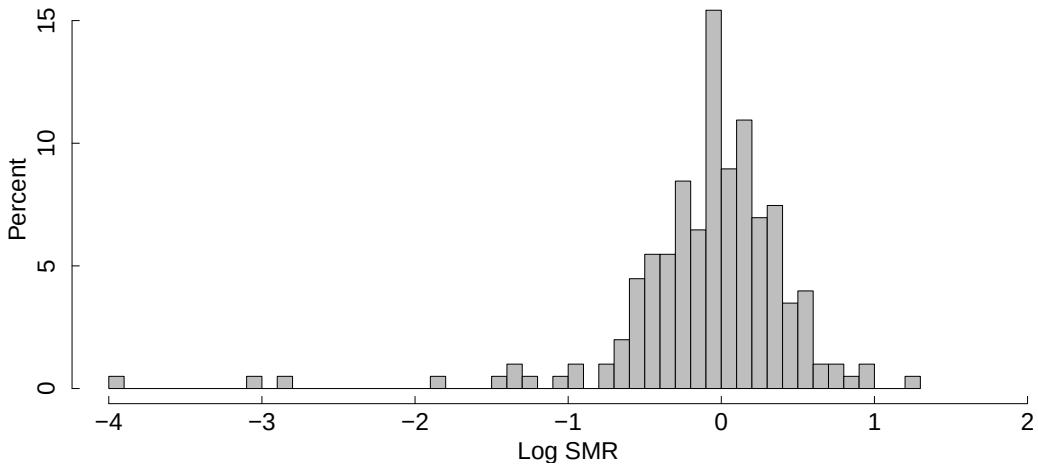

Figure 4.2: Histogram of log SMR in percentage

4.2. Prognostic Score-Based Weighting

In addition to SMR, we also use the prognostic-based weighting method to evaluate center effects. Depending on distinct covariate values, each patient has a corresponding risk score. Below is the histogram of median center-specific risk score, which ranges from -0.2 to 0.7 (Figure 4.3). To put this into perspective, a prognostic score of 0 represents approximately average risk across the study population, with higher prognostic risk scores denoting greater allograft failure risk.

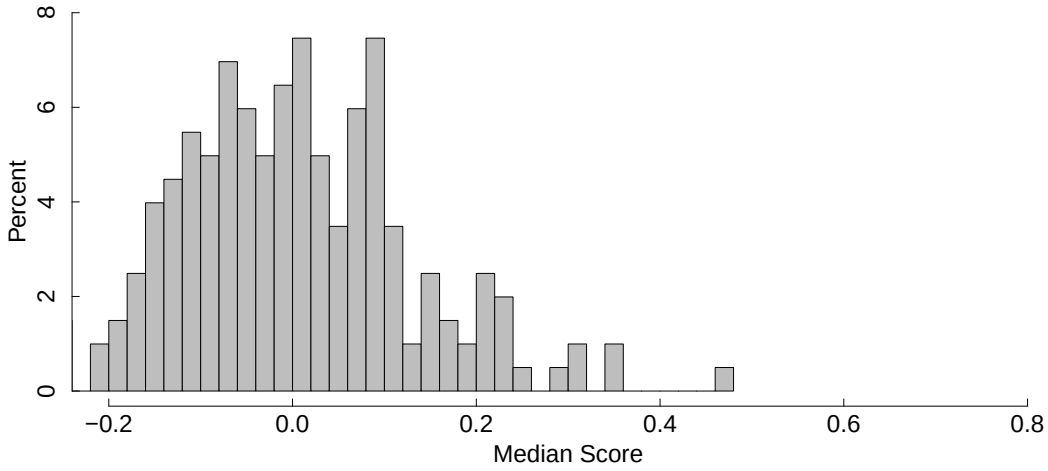

Figure 4.3: Histogram of center-specific median risk score

Quantiles of prognostic scores are used to create stratification for the study population where the proportion of patients in each stratum can be used to obtain individual weights. Weighted center-specific survivals are then computed, with the population average being the standard for comparison. In particular, excess survival τ of 0 represents no difference from the averaged center-specific survival, with higher τ denoting better center.

Figure 4.4 shows the histogram of excess survival of τ ranging from -0.14 to 0.6. There are about 11% of centers with τ equal to 0. The tails of the τ distribution indicate centers being significantly different from the overall population, with the right end (bigger than 0) representing better centers and the opposite direction representing worse.

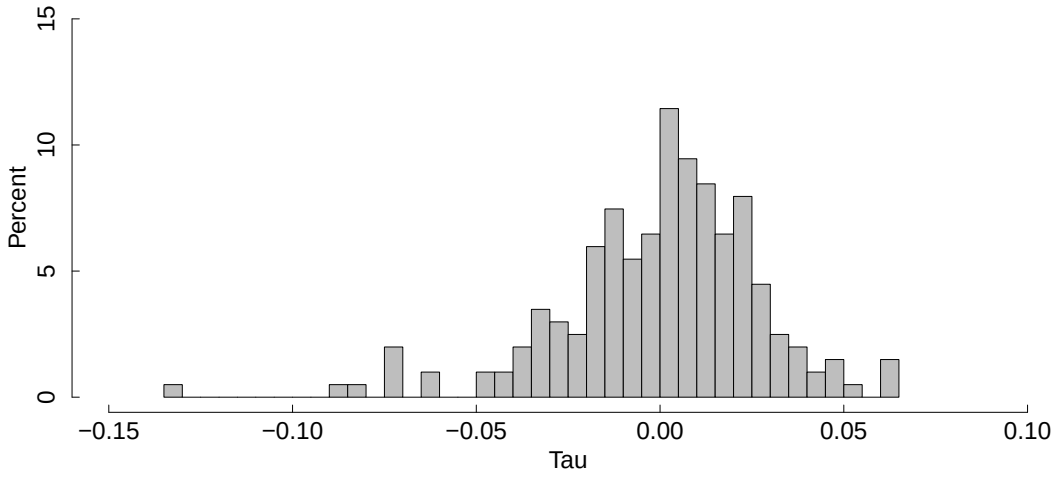

Figure 4.4: Histogram of excess survival in percentage

For ease of comparisons, excess survival probability is visualized in increasing order by 201 centers, in addition to its variance in Figure 4.5. Overall, centers with smaller excess survival tend to have wider variability.

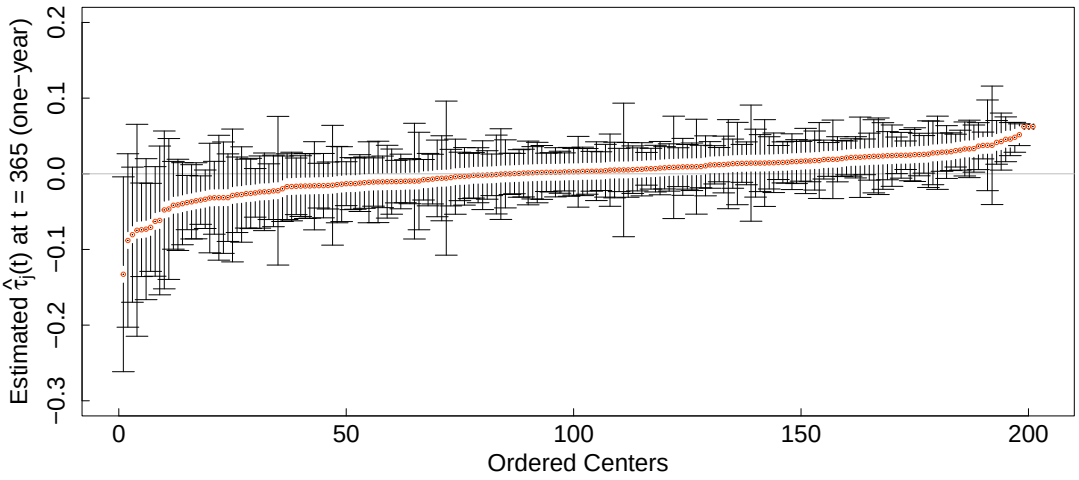

Figure 4.5: Excess survival probability by ordered centers

4.3. Comparison of Transplant Centers Performance

Figure 4.6 shows the scatterplot of τ and log SMR. In this scatterplot, the upper left triangle where τ is smaller than 0 and log SMR is bigger than 0, indicates significantly worse centers, with the lower right triangle indicating significantly better centers. By visualization, there is a strong agreement between the two metrics for facility profiling.

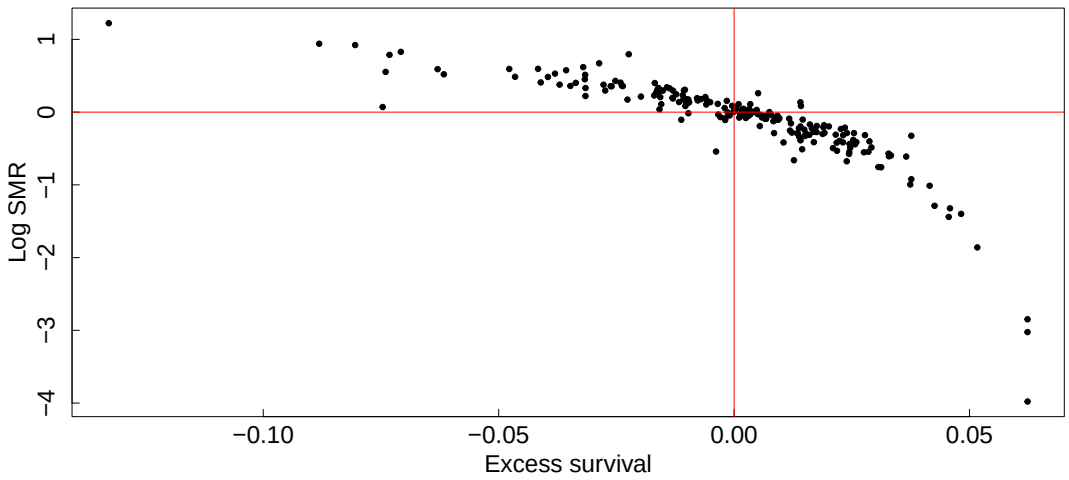

Figure 4.6: Scatterplot of excess survival probability and log SMR

Table 4.1 cross-classifies centers into three categories: better, null, worse based on results from each of the two evaluation metrics. Among 201 kidney transplant centers in total, there are 19, 177, and 5 centers in better, null, worse categories respectively by the prognostic score based method. There are 1, 177, and 23 centers in better, null, worse categories respectively by the SMR. Overall, there are 165 centers in agreement and 36 in disagreement between the two metrics. No centers show opposite results from the two metrics: better versus worse or vice versa.

Cross classification		τ**(1)**			
		Better Center	**Null Center**	**Worse Center**	**Total**
SMR(1)	**Better**	**1**	0	0	1
	Null	18	**159**	0	177
	Worse	0	18	**5**	23
	Total	19	177	5	**201**

Table 4.1: Number of centers in each stratum

Table 4.2 shows the center-specific median SMR and median $\tau(1)$ for each of the subclassification between the two metrics. The center-specific median SMR for the better, null, and worse centers categorized in agreement by the two metrics are 0.37, 0.97, and 1.74 respectively. Similarly, the center-specific median $\tau(1)$ for better, null, worse centers in agreement are 0.04, 0.003, and -0.07 respectively.

Median SMR ***Median τ (1)***		τ**(1)**		
		Better Center	**Null Center**	**Worse Center**
SMR(1)	**Better**	**0.37** **0.04**	NA NA	NA NA
	Null	**0.416** **0.04**	**0.97** **0.003**	NA NA
	Worse	NA NA	**1.70** **-0.03**	**1.74** **-0.07**

Table 4.2: Center-specific median SMR and median τ

Table 4.3 shows the center-specific median number of patients, median risk score, and median KDRI for each of the subclassification group between the two metrics. The center-specific median number of patients for better, null, and worse centers categorized in agreement by the two metrics are 646, 288, and 190 patients respectively. Similarly, the center-specific median linear predictor risk score for better, null, and worse centers are 0.14, 0.22, and 0.11 respectively. The center-specific median KDRI for better, null, and worse centers are 1.09, 1.21, and 1.12 respectively.

Median # patients ***Median risk score*** ***Median KDRI***		τ**(1)**		
		Better Center	**Null Center**	**Worse Center**
SMR(1)	**Better**	**646** **0.14** **1.09**	0 NA NA	0 NA NA
	Null	**236** **0.21** **1.19**	**288** **0.22** **1.21**	0 NA NA
	Worse	0 NA NA	**272** **0.21** **1.20**	**190** **0.11** **1.12**

Table 4.3: Center-specific median number of patients, median prognostic score, and median KDRI

CHAPTER 5

DISCUSSION

In this project, we use data from United Network for Organ Sharing to evaluate kidney transplant centers with the SMR and prognostic score based method. Despite wide utility, SMR is ill-suited for settings including kidney transplantation for several limitations where the outcome of interest is survival outcome. Its shortcomings include a heavy reliance on hazard proportionality assumptions, unstable estimation when survival rate is high, optimal clinical interpretation, and indirect standardization over case-mix distribution of covariate across centers. The second metric that we use is the prognostic score based weighting approach, designed to overcome these limitations by making use of survival probability to yield a more accurate estimate of center effects.

The correlation between the two metrics is approximately -0.94, with a higher SMR not necessarily implying a lower excess survival probability. From the results, there are 36 medical centers that were categorized differently between the two metrics. We further investigated this with a logistic regression to see how much impact the number of patients, median risk scores, and median KDRI have on the results. The findings in Table 5.1 show that none of these covariates was statistically significant on the discrepancy between the two metrics.

Summary of Logistic Regression			
Characteristics	**Odds Ratio**	**95% Confidence Interval**	**p-value**
Median # patients	1.00	(1.00, 1.00)	0.13
Median risk score	14.8	(0.47, 499)	0.13
Median KDRI	0.17	(0.00, 11.6)	0.4

Table 5.1: Logistic Regression on Metrics Agreement

High values of τ indicate centers better than averaged population, while high prognostic scores indicate patients with worse condition. High correlation between excess survival probability τ and center-specific median prognostic risk scores indicates potential residuals that were not accounted for. In Figure 5.1, the scatterplot shows that there is little correlation between τ and median prognostic risk score. The Spearman's rank correlation between center-specific τ and median prognostic risk score is 0.05, illustrating that the novel prognostic score-based method balances out the confounders and that there is negligible residuals.

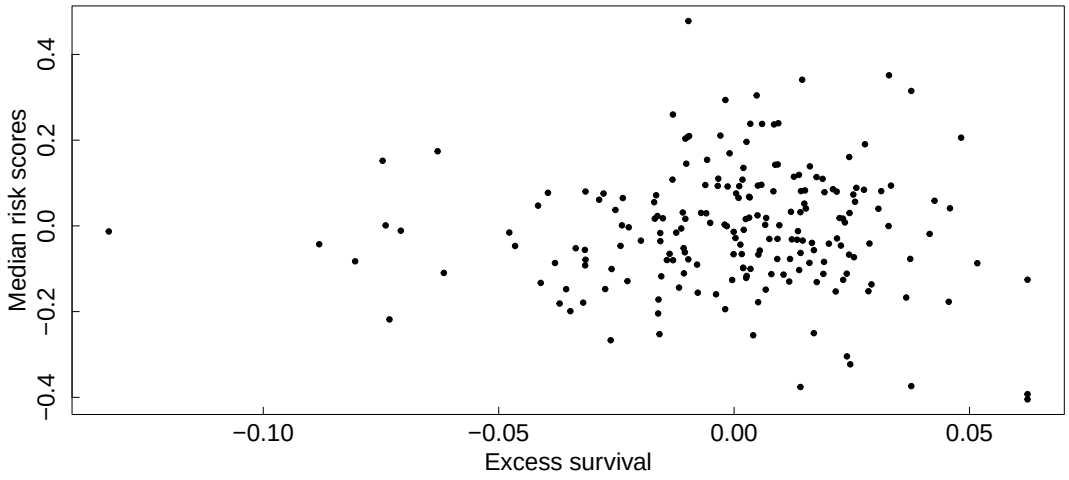

Figure 5.1: Scatterplot of excess survival probability and median prognostic risk score

There has been clinical concerns about the comparability among kidney centers. Given the heterogeneous underlying populations of kidney centers such as demographics or social economics, there are potentially centers that lie in two extreme ends of the comparability spectrum. For example, a center located in a suburban area serve patients mostly elderly White, while another located in an urban area receive patients that are Black and come from high poverty. However, a cross-classification of the 5 risk classes by 201 centers yields 1005 cells. The degree to which these cells are unpopulated provides evidence against the

overlapping assumption. However, all of the 1005 cells are populated, implying no violation of positivity assumption.

We recommend using the prognostic score-based weighting method in clinical settings where the proportionality assumption is not guaranteed to satisfy, or when survival rates are high. However, we also want to note that the prognostic score method was developed under the assumption of independent censoring and no effect modification. In settings where the independent censoring assumption is violated, potential solutions involve employing inverse probability censoring weighting techniques, which lie beyond the scope of this thesis. For future work, clinicians may want to add an interaction effect between center memberships and potential effect modifiers in the model used to generate linear predictors to construct risk classes.

BIBLIOGRAPHY

Ornulf Borgan and Bryan Langholz. Nonparametric estimation of relative mortality from nested case-control studies. *Biometrics*, pages 593–602, 1993.

DM Dickinson, TH Shearon, J O'Keefe, H-H Wong, CL Berg, JD Rosendale, FL Delmonico, RL Webb, and RA Wolfe. Srtr center-specific reporting tools: Posttransplant outcomes. *American Journal of Transplantation*, 6(5):1198–1211, 2006.

Ben B Hansen. The prognostic analogue of the propensity score. *Biometrika*, 95(2):481–488, 2008.

Kevin He and Douglas E Schaubel. Standardized mortality ratio for evaluating center-specific mortality: assessment and alternative. *Statistics in Biosciences*, 7:296–321, 2015.

Colleen Jay and Jesse D Schold. Measuring transplant center performance: the goals are not controversial but the methods and consequences can be. *Current transplantation reports*, 4:52–58, 2017.

J Kasza, John L Moran, PJ Solomon, ANZICS Centre for Outcome, Resource Evaluation (CORE) of the Australian, and New Zealand Intensive Care Society (ANZICS). Evaluating the performance of australian and new zealand intensive care units in 2009 and 2010. *Statistics in Medicine*, 32(21):3720–3736, 2013.

Finbarr P Leacy and Elizabeth A Stuart. On the joint use of propensity and prognostic scores in estimation of the average treatment effect on the treated: a simulation study. *Statistics in medicine*, 33(20):3488–3508, 2014.

Y Lee, PP Reese, AH Tran, and DE Schaubel. Prognostic score-based methods for estimating center effects based on survival probability: Application to post-kidney transplant survival. *Unpublished manuscript*, 2023+.

Youjin Lee and Douglas E Schaubel. Facility profiling under competing risks using multivariate prognostic scores: Application to kidneytransplant centers. *Statistical Methods in Medical Research*, 31(3):563–575, 2022.

Bradley N Manktelow, T Alun Evans, and Elizabeth S Draper. Differences in case-mix can influence the comparison of standardised mortality ratios even with optimal risk adjustment: an analysis of data from paediatric intensive care. *BMJ quality & safety*, 23(9): 782–788, 2014.

Tri-Long Nguyen and Thomas PA Debray. The use of prognostic scores for causal inference with general treatment regimes. *Statistics in medicine*, 38(11):2013–2029, 2019.

Maurice E Pouw, Linda M Peelen, Hester F Lingsma, Daniel Pieter, Ewout Steyerberg, Cor J Kalkman, and Karel GM Moons. Hospital standardized mortality ratio: consequences of adjusting hospital mortality with indirect standardization. *PloS one*, 8(4):e59160, 2013.

Panduranga S Rao, Douglas E Schaubel, Mary K Guidinger, Kenneth A Andreoni, Robert A Wolfe, Robert M Merion, Friedrich K Port, and Randall S Sung. A comprehensive risk quantification score for deceased donor kidneys: the kidney donor risk index. *Transplantation*, 88(2):231–236, 2009.

David B Richardson, Alexander Keil, Tchetgen E Tchetgen, and Glinda S Cooper. Negative control outcomes and the analysis of standardized mortality ratios. *Epidemiology (Cambridge, Mass.)*, 26(5):727, 2015.

Daniel Sjoberg, Karissa Whiting, Michael Curry, Jessica A Lavery, and Joseph Larmarange. Reproducible summary tables with the gtsummary package. *The R Journal*, 13(1):570–580, 2021.

Robert A Wolfe. The standardized mortality ratio revisited: improvements, innovations, and limitations. *American Journal of Kidney Diseases*, 24(2):290–297, 1994.