Published in final edited form as:

J Am Stat Assoc. 2017 ; 112(518): 534–546. doi:10.1080/01621459.2016.1240078.

Latent Class Survival Models Linked by Principal Stratification to Investigate Heterogenous Survival Subgroups Among Individuals With Early-Stage Kidney Cancer

Brian L. Eggleston¹, Robert G. Uzzo², and Yu-Ning Wong³

²Chairman of Surgery, Fox Chase Cancer Center, Temple University Health System

³Medical Oncology, Fox Chase Cancer Center, Temple University Health System

Abstract

Rates of kidney cancer have been increasing, with small incidental tumors experiencing the fastest growth rates. Much of the increase could be due to increased use of CT scans, MRIs, and ultrasounds for unrelated conditions. Many tumors might never have been detected or become symptomatic in the past. This suggests that many patients might benefit from less aggressive therapy, such as active surveillance by which tumors are surgically removed only if they become sufficiently large. However, it has been difficult for clinicians to identify subgroups of patients for whom treatment might be especially beneficial or harmful. In this work, we use a principal stratification framework to estimate the proportion and characteristics of individuals who have large or small hazard rates of death in two treatment arms. This allows us to assess who might be helped or harmed by aggressive treatment. We also use Weibull mixture models. This work differs from much previous work in that the survival classes upon which principal stratification is based are latent variables. That is, survival class is not an observed variable. We apply this work using Surveillance Epidemiology and End Results-Medicare claims data. Clinicians can use our methods for investigating treatments with heterogeneous effects.

2 Introduction

The incidence of localized kidney cancer has been increasing for the past several decades (Chow *et al.* 1999, Hock *et al.* 2002). Some of this increase may be due to the detection of small tumors found incidentally after ultrasound, CT scans, or MRI scans ordered for non-kidney cancer related reasons. This suggests two possibilities 1) many tumors that would previously have been found at a later stage are simply being found at earlier stages, and 2) tumors that would never have progressed to a symptomatic or lethal stage are now being found. The possibility that biologically inconsequential small renal masses are now being detected has opened a debate concerning the benefits of active surveillance prior to or in lieu of surgical or other interventions for the treatment of small localized renal masses (Kunkle *et al.* 2008, Hollingsworth *et al.* 2006, Sun *et al.* 2014). Active surveillance consists of

¹Corresponding author Biostatistics and Bioinformatics Facility, Fox Chase Cancer Center, Temple University Health System, 333 Cottman Ave., Philadelphia, PA 19111, ph. 215-214-3917, Brian.Eggleston@fccc.edu.

monitoring tumors over time and surgically intervening only after tumors demonstrate problematic growth.

One way to compare the utility of intervention for early-stage cancer would be to compare outcomes in those who did have and those who did not have surgery. However, even in a well-designed study that adjusts for differences between surgical groups, such an approach might not identify subgroups for whom intervention is particularly beneficial or harmful. Surgical intervention here generally means partial or radical nephrectomy (i.e. removal of the kidney) but also can mean other forms of intervention such as ablation.

Latent class analysis (Bandeian Roche *et al.* 1997) provides an exploratory approach to investigate whether there is evidence of subgroups for whom a treatment is particularly effective. One difference between latent class analysis and conventional subgroup analyses is that the subgroups of interest do not need to be prespecified in latent class analyses. Instead, latent class analysis, like cluster analytic methods, can use covariates to identify groups that have different hazards of death. Cluster analytic methods in general are of increasing interest in oncologic research (Kristensen *et al.* 2014). Latent class methods in particular are being more commonly used in high impact clinical journals (e.g. Desai *et al.* 2015, Viatte *et al.* 2015). This suggests that a current dissemination of latent class methods from the statistics literature (e.g. Qu *et al.* 2015) to the substantive literature. In this work, we use Weibull survival mixture models (Marin *et al.* 2005, Farewell 1982, Mair and Hudec 2009) to investigate whether there is evidence of latent survival classes among those with localized kidney cancer treated with and without surgery. The Weibull mixture models assume that observed survival times come from a mixture of Weibull distributions. These mixture models are different from the discrete time survival mixture models of Muthén and Masyn (2005). Weibull mixture models can be used to estimate the separate parameters from each model in the mixture, and also estimate the proportion of the population that has each mixing characteristic (i.e. the latent survival class to which people belong).

While traditional latent class models for binary data have been used to explore subgroups that may moderate treatment effects (Lanza and Rhoades 2013), the assumptions needed to identify such latent subgroup treatment effects have not been extensively elaborated, particularly with respect to survival outcomes. Here, we use principal stratification (Frankagis and Rubin 2002) to link the potential latent survival classes across treatment modalities to provide estimates of expected survival outcomes that individuals can expect if they do and do not have surgical treatment. In particular, we identify four potential survival rate groups in the investigation of clinical effectiveness of surgery 1) those that have long expected survival with or without treatment, 2) those that have long expected survival without treatment but short expected survival with treatment due to complications such as end stage renal failure, 3) those that have short expected survival without treatment but long expected survival with treatment who should undergo immediate intervention, and 4) those that have short expected survival with or without treatment. Like latent class methods, principal stratification is increasingly being used in the clinical literature (e.g. Halpern *et al.* 2015).

This research differs from much other work on principal stratification in that the principal strata are based on latent survival class membership. Generally, principal strata are based on partially observed variables, such as death by a certain time point (Egleston *et al.* 2007), or HIV infection (Shepherd *et al.* 2007). Our work is most similar to that of Jo and colleagues (2009) who examined principal strata defined by latent trajectories of attention deficit disorder over time. While Jo *et al.* defined classes based on repeated measures data, our latent classes are defined by time-to-event survival time and our methods account for censoring. Previously, Egleston *et al.* (2009) defined principal strata by observable survival to a fixed time, such as one year survival that could be modeled by a logistic regression. Here we define principal strata based on classes with different underlying survival models which are not fully observed (i.e. latent).

We apply our methods to examine the effectiveness of nephrectomy for localized kidney cancer using data from the Surveillance Epidemiology and End Results (SEER) database linked to Medicare claims data (<http://appliedresearch.cancer.gov/seermedicare/>).

3 Estimand

Notationally, let Z be a surgery indicator ($Z = 0$ if no surgery, $Z = 1$ otherwise). Let $\mathbf{X}$ represent a vector of baseline covariates. Let $T(z)$ represent the potential survival time under $Z = z$. Similarly, let $C(z)$ be a potential censoring time, with $Y(z)$ equal to the minimum of $T(z)$ or $C(z)$. Let $D(z)$ be death within the time frame of the study indicator such that $D(z) = 1$ if $T(z) < C(z)$, $D(z) = 0$ otherwise. The joint set of potential outcomes is $\{T(0), T(1), C(0), C(1), Y(0), Y(1), D(0), D(1)\}$. However, in the course of a study, we only observe the outcome variables $Y = Y(Z)$, $D = D(Z)$, so that the full set of observed variables is $\{Y, D, Z, \mathbf{X}\}$. For ease of presentation, we generally omit subscripts $i = \{1, \dots, n\}$ to denote individual level variables, except where necessary. Note that the variables $Y = Y(z)$, $D = D(z)$, and $\mathbf{X}$ are observed for a person when $Z = z$. Only the minimum of $T(z)$ and $C(z)$ are observed through $Y = Y(z)$. In our example, we use Weibull mixture models to identify the latent survival classes, as we detail in Section 4.3.

In addition to the observable variables, let $R(z)$ indicate a latent survival class indicator. $R(z) = 1$ if a person is in the class that generally has shorter survival (i.e. a generally rapid dier), $R(z) = 0$ if a person is in the class with generally longer survival. Even if a person is observed to die during the observation period, we would not know with probability one which latent survival model generated the survival time. Mapping back to the four principal strata of interest, those with 1) $\{R(0) = R(1) = 0\}$ generally have the longest survival, 2) $\{R(0) = 0, R(1) = 1\}$ have longer survival without treatment, 3) $\{R(0) = 1, R(1) = 0\}$ have longer survival with treatment, and 4) $\{R(0) = R(1) = 1\}$ generally have the shortest survival.

Our causal estimand of interest is $E[T(z)|R(z) = r, R(1 - z) = r']$ for $z, r, r' \in \{0, 1\}$. This is the expected survival time when receiving treatment $Z = z$ given the principal strata defining survival class membership under both treatment arms. We are also interested in differences in survival comparing surgery to no-surgery among the four principal strata.

4 Identification and Estimation

In order to identify our estimand, we need to make untestable assumptions, although we will relax these assumptions using two sensitivity analyses. Our first two assumptions are the standard consistency (Pearl 2009, Cole and Frangakis 2009) and Stable Unit Treatment Value Assumptions (SUTVA, Rubin 1986). We believe that consistency is reasonable as surgical removal of a tumor is a well-defined and manipulable treatment. We believe that SUTVA is reasonable as kidney cancer is relatively rare so that one patient's treatment decision is unlikely to affect a typically unrelated person's decision. We also make the assumption that censoring (loss to follow-up) is non-informative. Our censoring assumption is realistic in the case of SEER data, that we use, since the reason for loss to follow-up is generally that a person has not yet died. When a person dies, the death information will be reported to the cancer registrar. Below, we detail the assumptions more unique to our approach.

$$\{T(0), T(1), C(0), C(1), Y(0), Y(1), D(0), D(1), R(0), R(1)\} \perp Z | \mathbf{X} \quad \text{A.1)}$$

This is a pseudo-randomization (strong ignorability, Rosenbaum and Rubin 1983) assumption common in observational studies that states that the potential outcomes are independent of treatment given the set of confounding covariates. We believe that assumption A.1 is plausible since we have measured tumor characteristics and comorbidities available in the linked SEER-Medicare database.

$$R(1) \perp R(0) | \mathbf{X} \quad \text{A.2)}$$

This assumes the cross-world independence (Imai and Yamamoto 2013) of the latent variables conditional on $\mathbf{X}$. This is a strong assumption that we relax through the sensitivity analysis assumption A.4 described below. However, A.2 may be a useful anchoring point to provide a unique solution to clinicians; the sensitivity analyses can be used to evaluate deviations from this anchor (Egleston *et al.* 2007).

$$T(z) \perp R(1 - z) | R(z), \mathbf{X} \quad \text{A.3)}$$

This assumes that the survival time under $Z = z$ is independent of the cross-world latent survival class given the survival class under $Z = z$ and the set of confounding covariates. This assumption is akin to components of Hayden *et al.*'s (2005) Explainable Nonrandom Survival Assumption, but less strong in that we condition on the survival latent class, $R(z)$. This assumption is also similar to Jo and Stuart's (2009) principal ignorability, although the authors did not cross-classify considering $R(0)$ and $R(1)$ as we do this paper. While we relax this through sensitivity analysis assumption A.5 below, this assumption may not need relaxation. The reason that A.3 may be a realistic assumption is that both $T(z)$ and $R(z)$ are

directly related; $R(z)$ is the latent survival class that specifies the parameterization of $T(z)$'s probability distribution. Hence, if we know the latent survival class under z (i.e. $R(z)$) and the confounding covariates, the cross-world latent survival class, $R(1 - z)$, might not provide much additional information about $T(z)$.

4.1 Sensitivity Analyses

$$\frac{\text{odds}P(R(1)=1|R(0)=1, \mathbf{X})}{\text{odds}P(R(1)=1|R(0)=0, \mathbf{X})} = \tau \quad \text{A.4}$$

This is a sensitivity analysis assumption that relaxes assumption A.2, and allows for the condition that all confounders might not have been measured in $\mathbf{X}$ for assumption A.2 to be valid. It links the joint probabilities of the potential latent survival classes using an odds ratio. We can fix τ and estimate our estimands of interest using a sensitivity analysis approach. This is similar to an odds ratio sensitivity parameter used by Dominici and colleagues (2006) to identify the joint distribution of potential outcomes. Roy and colleagues (2008) used such a sensitivity parameter on the linear probability scale, which seemed more appropriate under their more stringent monotonicity assumption that bounded probabilities. Muthén and Brown's (2009) parameter ω_i is akin to our τ , but the authors used stronger assumptions to uniquely identify their parameter rather than use a sensitivity analysis approach. Parameters such as τ relax earlier published assumptions that there are no discordant principal strata such that everyone belongs to the strata $\{R(0) = 1, R(1) = 1\}$ or $\{R(0) = 0, R(1) = 0\}$ and no one belongs to the strata $\{R(0) = 0, R(1) = 1\}$ or $\{R(0) = 1, R(1) = 0\}$ (e.g. Muthén *et al.* 2002).

We believe that $\tau \geq 1$. The justification is that having a poor outcome in one arm predisposes one to a poor outcome in the other arm. That is, if $R(z) = 1$ then $R(1 - z)$ is likely to equal one as well.

$$\text{For } r \text{ and } z \in \{0, 1\}, \quad \text{A.5}$$

$$\frac{\frac{E[T(z)|R(z)=1, R(1-z)=0, \mathbf{X}]}{E[T(z)|R(0)=1, R(1)=1, \mathbf{X}]}}{\frac{E[T(z)|R(z)=0, R(1-z)=1, \mathbf{X}]}{E[T(z)|R(0)=0, R(1)=0, \mathbf{X}]}} = \frac{\kappa_{1z}}{\kappa_{0z}}$$

This is a sensitivity analysis assumption that relaxes assumption A.3, and again allows for a situation in which not all confounders are contained in $\mathbf{X}$ to make A.3 a realistic assumption. As with τ , we can fix each κ_{rz} and estimate our estimands using a sensitivity analysis approach. This is similar to sensitivity analysis parameters used by Egleston and colleagues (2007, 2009) to identify the joint distribution of potential outcomes. One difference between this assumption and Egleston and colleagues' prior assumptions is that we do not make a monotonicity assumption here. As such, there are four sensitivity parameters for the four

combinations of r and z : $\{r=0, z=0\}$, $\{r=0, z=1\}$, $\{r=1, z=0\}$, $\{r=1, z=1\}$. While the formulation requires that the denominator not equal zero, the denominator does not have expectation zero in our case since $T(z)$ represents time greater than zero.

As $R(z) = 1$ indicates the latent survival class under $Z = z$ with worse survival outcomes, we have parameterized κ_{0z} and κ_{1z} such that expected survival would be longer for principal strata with smaller sums of $R(0) + R(1)$ compared to those with larger sums of $R(0) + R(1)$. This assumption plausibly bounds the range of our sensitivity parameters such that $\kappa_{1z} \leq 1$ and $\kappa_{0z} \geq 1$. That is, on average, the more latent survival classes a person is in with poor expected survival outcomes, the worse off a person is in general.

4.2 Identification

The proof below shows how the above assumptions identify the causal estimand. For ease of presentation, we demonstrate identifiability of $E[T(0)|R(0) = 1, R(1) = 1]$ only.

$$\begin{aligned} E[T(0)|R(0)=1, R(1)=1] &= \frac{E[T(0)R(0)R(1)]}{E[R(0)R(1)]} \\ &= \frac{E[E[T(0)R(0)R(1)|\mathbf{X}]]}{E[E[R(0)R(1)|\mathbf{X}]]} \\ &= \frac{E[E[T(0)|R(0)=1, R(1)=1, \mathbf{X}]P(R(1)=1|R(0)=1, \mathbf{X})P(R(0)=1|\mathbf{X})]}{E[P(R(1)=1|R(0)=1, \mathbf{X})P(R(0)=1|\mathbf{X})]} \quad (1) \end{aligned}$$

$$\begin{aligned} &= \frac{E[E[T(0)|R(0)=1, \mathbf{X}]P(R(1)=1|R(0)=1, \mathbf{X})P(R(0)=1|\mathbf{X})]}{E[P(R(1)=1|R(0)=1, \mathbf{X})P(R(0)=1|\mathbf{X})]} \quad \text{by A. 3} \\ &= \frac{E[E[T(0)|R(0)=1, \mathbf{X}]P(R(1)=1|\mathbf{X})P(R(0)=1|\mathbf{X})]}{E[P(R(1)=1|\mathbf{X})P(R(0)=1|\mathbf{X})]} \quad \text{by A. 2} \\ &= \frac{E[E[T|R(0)=1, Z=0, \mathbf{X}]P(R(1)=1|Z=1, \mathbf{X})P(R(0)=1|Z=0, \mathbf{X})]}{E[P(R(1)=1|Z=1, \mathbf{X})P(R(0)=1|Z=0, \mathbf{X})]} \quad \text{by A. 1} \quad (2) \end{aligned}$$

Proofs for $E[T(z)|R(0) = r, R(1) = r']$ for all $r, r' \in \{0, 1\}$ follow similarly. The above proof demonstrates how the estimand of interest, which is an expectation of a potential outcome conditional on latent potential outcomes can be related to quantities estimable using the observed data. Although $R(z)$ is a latent survival indicator that is not observed, $E[T|R(z) = r, Z = z, \mathbf{X}]$ and $P(R(z) = r|Z = z, \mathbf{X})$ can be identified and estimated using survival mixture models as detailed in Section 4.3.

We can also use the sensitivity analysis assumptions to identify quantities in Equation (1), thereby relaxing our assumptions. For example, we can identify quantities such as $P(R(1) = 1|R(0) = 1, \mathbf{X})$ by examination of the following equality:

$$P(R(1)=1|\mathbf{X})=P(R(1)=1|R(0)=1, \mathbf{X})P(R(0)=1|\mathbf{X})+P(R(1)=1|R(0)=1, \mathbf{X})P(R(0)=0|\mathbf{X}) \quad (3)$$

Assumption A.1 ensures that $P(R(z) = 1|\mathbf{X}) = P(R(z) = 1|Z = z, \mathbf{X})$. Further, setting

$$P(R(1)=1|R(0)=1, \mathbf{X}) = \frac{\tau_{\text{odds}} P(R(1)=1|R(0)=0, \mathbf{X})}{1 + \tau_{\text{odds}} P(R(1)=1|R(0)=0, \mathbf{X})}$$

by Assumption A.4 reduces Equation (3) to three identifiable probabilities and one unknown. We can solve for the unknown.

Similarly, by Assumption A.5, $E[T(0)|R(0) = 1, R(1) = 1, \mathbf{X}]$ in Equation (1) is equal to

$$E[T(0)|R(0)=1, R(1)=1, \mathbf{X}] = \frac{E[T(0)|R(0)=1, \mathbf{X}]}{P(R(1)=1|R(0)=1, \mathbf{X}) + \kappa_{10}(1 - P(R(1)=1|R(0)=1, \mathbf{X}))} \quad (4)$$

This equality can be found by noting that we can substitute

$$E[T(0)|R(0)=1, R(1)=0, \mathbf{X}] = \kappa_{10} E[T(0)|R(0)=1, R(1)=1, \mathbf{X}]$$

into the following equation:

$$\begin{aligned} & E[T(0)|R(0)=1, \mathbf{X}] \\ &= E[T(0)|R(0)=1, R(1)=1, \mathbf{X}] P(R(1)=1|R(0)=1, \mathbf{X}) + \\ & E[T(0)|R(0)=1, R(1)=0, \mathbf{X}] P(R(1)=0|R(0)=1, \mathbf{X}) \end{aligned} \quad (5)$$

Equation (4) is identifiable since $E[T(0)|R(0) = 1, \mathbf{X}] = E[T|R(0) = 1, Z = 0, \mathbf{X}]$ by assumption A.1, and $P(R(1) = 1|R(0) = 1, \mathbf{X})$ is identifiable as per the discussion of Equation (3) above. The sensitivity parameter κ_{10} allows Equation (5) to have one unknown quantity for which we can solve, rather than two. The four κ_{iz} parameters hence allow us to reduce an overdetermined system of four equations with eight unknowns to four equations with four unknowns, allowing us to solve for each $E[T(z)|R(z) = r, R(1 - z) = r', \mathbf{X}]$ where $z, r, r' \in \{0, 1\}$.

4.3 Models

We use Weibull mixture models (Marin *et al.* 2005, Farewell 1982) to estimate each of the inner expectations and probability components of Equation (2). We chose parametric survival models so that we could 1) estimate life expectancies, which may be more readily interpretable by patients than other statistics, such as medians, and 2) reduce the dimension of the problem to improve identifiability of the mixture models. A limitation of a reliance on models in general is that the models may not be correct. To reduce model mis-specification,

we chose Weibull proportional hazards models as they are relatively flexible survival models with parameters (i.e. estimable hazard ratios) that are interpretable by clinicians. We chose two class mixture models based on our prior hypotheses (Kunkle *et al.* 2008) that surgery has little effect for many (the principal strata in which $R(1) = R(0)$), a beneficial effect for some (principal stratum with $R(0) > R(1)$) and a harmful effect for some (stratum $R(0) < R(1)$). We felt that the two class model resulting in four principal strata would provide the clearest way to explore this. The mixture models are used to estimate outcomes and probabilities of latent survival class membership within intervention surgery arms; the mixture models are not used to estimate the probability of principal stratum membership directly. Examples of the mixture formulation of principal stratification include the likelihood based methods of Frumento *et al.* (2012) and equations 10-14 in Gallop *et al.* (2009). Here, we do not estimate the probability of principal stratum membership directly through a model based on mixtures of principal strata. Instead, we estimate Weibull mixture models separately for each treatment group and then use assumptions A.1-A.3 or a sensitivity analysis approach (e.g. Equation 3 and assumptions A.1, A.4, A.5) to link (i.e. jointly identify) cross-world potential outcomes, thereby forming principal stratum specific estimates. In summary, we estimate mixture models using maximum likelihood estimation, and use the estimated parameters from the models to estimate probabilities of latent class membership and survival within latent class. We apply the δ -method to the maximum likelihood asymptotic standard errors to obtain standard errors of the estimates of interest.

Here we apply subscripts to denote individual level data. Let $h_{rZ}(\cdot)$ represent the hazard function for survival class $R_i(z) = r$ and treatment $Z_i = z$. Under a Weibull model, an individual's hazard of death at time t is $h_{rZ}(t|X_i = x_i, R_i(z) = r, Z_i = z) = \exp(x_i \beta_{rZ}) \lambda_{rZ} \gamma_{rZ} t^{\gamma_{rZ}-1}$ for baseline scale parameter λ_{rZ} , shape parameter γ_{rZ} , and parameter vector β_{rZ} . For the probability of class membership, we use a logistic function, $\text{logit } P(R_i(z) = 1 | X_i, Z_i = z) = \mu_z + X_i \alpha_z$ for intercept μ_z and parameter vector α_z . Let $I(\cdot)$ be the indicator function that takes value 1 when the expression inside is true, 0 otherwise. Let $\psi = \{\mu_z, \alpha_z, \beta_{rZ}, \lambda_{rZ}, \gamma_{rZ}\}'$ for $r \in \{0, 1\}$ and $z \in \{0, 1\}$ be the parameter set. The likelihood for the Weibull mixture model that accounts for independent censoring becomes,

$$L(\psi) = \prod_{i=1}^n \prod_{z=0}^1 I(Z_i = z) \left(\frac{1}{1 + \exp(\mu_z + x_i \alpha_z)} (\lambda_{0z} \gamma_{0z} \exp(x_i \beta_{0z}) y_i^{\gamma_{0z}-1})^{d_i} \exp(-\exp(x_i \beta_{0z}) \lambda_{0z} y_i^{\gamma_{0z}}) + \frac{\exp(\mu_z + x_i \alpha_z)}{1 + \exp(\mu_z + x_i \alpha_z)} (\lambda_{1z} \gamma_{1z} \exp(x_i \beta_{1z}) y_i^{\gamma_{1z}-1})^{d_i} \exp(-\exp(x_i \beta_{1z}) \lambda_{1z} y_i^{\gamma_{1z}}) \right)$$

Note that the quantities in Equation (2) are equal to the following.

$$P(R_i(z)=1|Z_i=z, X_i=x_i) = \frac{\exp(\mu_z + x_i \alpha_z)}{1 + \exp(\mu_z + x_i \alpha_z)} \quad (6)$$

$$E[T_i|R_i(z)=r, Z_i=z, \mathbf{X}_i=\mathbf{x}_i] = (\lambda_{rz} \exp(\mathbf{x}_i \boldsymbol{\beta}_{rz}))^{-1/\gamma_{rz}} \Gamma\left(\frac{1}{\gamma_{rz}} + 1\right) \quad (7)$$

where the gamma function is defined as $\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du$ (Abramowitz and Stegun, 1972).

4.4 Estimation

We have developed an R package (R Statistical Foundation, Vienna, Austria) called latentPS that fits our Weibull mixture models and estimates the principal strata effects; it is available from the authors. The program uses a Newton Raphson algorithm (Lehmann 2010) with analytic first and second derivatives that we programmed for the Weibull mixture models, as detailed below. Of note is that the ability to fit the mixture models is highly dependent on starting values. Without appropriate starting values, parameters may diverge when the Newton Raphson algorithm is applied. To overcome this problem, we randomly divide the data into two groups using 40%, 50%, and 60% splits that are not dependent on any variables. Within each split dataset, we fit a logistic regression on the group indicator and separate Weibull proportional hazards regressions for each of the two randomly split groups. Similar to the methods of Razali and Al-Wakeel (2013), these naive regression estimates become the starting values for the Weibull mixture models. We repeat the splitting and fitting algorithm several times to better ensure that the model converges to the global rather than local maximum. In order to refine the naive starting values, we fit the Weibull mixture models first using R's *optim*(·) command. While the *optim*(·) command generally obtains parameter estimates that increase the value of the log likelihood, they are not always the maximum likelihood estimates. We then apply our Newton Raphson algorithm to the *optim*(·) refined estimates; using the refined estimates reduces convergence problems. The set of coefficients chosen after fitting the model with various starting values is the set that maximizes the log likelihood.

As we will further detail in section 6, there are many more people treated for localized kidney cancer than not treated. In our data analysis, we hence matched each untreated case in our data example to two treated cases using propensity score matching (Rosenbaum and Rubin 1983). While our package, latentPS, does not implement the propensity score matching, it allows for the incorporation of the matched design when estimating standard errors. Although propensity score matching balances covariates, we estimate the Weibull survival models separately for the two surgical groups. In this way, we are able to investigate the effect modification (i.e. interaction) that the matched variables have on subclass and treatment effects. Also, estimating the models separately allows for greater flexibility in allowing for mean or median survival differences across treatment groups; those who generally have longer life expectancy regardless of treatment might still have extraordinary long survival in one arm compared to the other.

In order to estimate standard errors for the Weibull survival mixture models, we used robust standard errors that corrected for clustering within matched group (Williams 2000, Huber 1964). Here, $\boldsymbol{\psi}^*$ represents the true parameters and $S = \{X, Z, Y, D\}$ represents the observed

data. Let $l_{jk}(\boldsymbol{\psi})$ represent the contribution to the log likelihood of person number k , $k \in \{1, 2\}$ in propensity score matched group j . Let $l_j(\boldsymbol{\psi}) = \sum_{k=1}^2 l_{jk}(\boldsymbol{\psi})$. Under the appropriate regularity conditions, we can show that

$$\sqrt{n}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}^*) \xrightarrow{D} \text{Normal}(0, \Sigma^*)$$

$$\Sigma^* = E_S \left[\frac{\partial^2 l_{jk}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}^2} \middle| \boldsymbol{\psi} = \boldsymbol{\psi}^* \right]^{-1} E_S \left[\left(\frac{\partial l_j(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}} \left(\frac{\partial l_j(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}} \right)' \middle| \boldsymbol{\psi} = \boldsymbol{\psi}^* \right) E_S \left[\frac{\partial^2 l_{jk}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}^2} \middle| \boldsymbol{\psi} = \boldsymbol{\psi}^* \right]^{-1} \right]' \quad (8)$$

In Equation (8), $E_S[\cdot]$ represents the expectation with respect to the observed data, and $q(\boldsymbol{\psi})|_{\boldsymbol{\psi}=\boldsymbol{\psi}^*}$ represents the evaluation of the function $q(\boldsymbol{\psi})$ when $\boldsymbol{\psi}$ is set to $\boldsymbol{\psi}^*$. We used empirical estimates of the expectations similar to those described by Bieler and Williams (1995) to estimate the covariance matrix, Σ^* .

Let $\hat{\boldsymbol{\psi}} = \{\hat{\mu}_z, \hat{\boldsymbol{\alpha}}_z, \hat{\boldsymbol{\beta}}_{rz}, \hat{\lambda}_{rz}, \hat{\gamma}_{rz}\}'$ for $r \in \{0, 1\}$ and $z \in \{0, 1\}$ represent the estimated parameter set. Our Newton Raphson algorithm iterates $\hat{\boldsymbol{\psi}}_q$ at each step q by updating

$$\hat{\boldsymbol{\psi}}_q = \hat{\boldsymbol{\psi}}_{q-1} - \left[\frac{\partial^2 l_{jk}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}^2} \right]^{-1} \frac{\partial l_{jk}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}} \middle| \boldsymbol{\psi} = \hat{\boldsymbol{\psi}}_{q-1} \text{ until convergence.}$$

Next we detail estimation of the principal strata effects using $\hat{\boldsymbol{\psi}}$. With one of the life expectancy quantities as an example, we can combine Equations (2), (6), and (7) with assumptions A.1, A.2, and A.3 to obtain,

$$\begin{aligned} & E[T(0)|R(0)=1, R(1)=1] \\ &= \frac{E[E[T|R(0)=1, Z=0, \mathbf{X}]P(R(1)=1|Z=1, \mathbf{X})P(R(0)=1|Z=0, \mathbf{X})]}{E[P(R(1)=1|Z=1, \mathbf{X})P(R(0)=1|Z=0, \mathbf{X})]} \\ &= \frac{E\left[(\lambda_{10}\exp(\mathbf{X}\boldsymbol{\beta}_{10}))^{-1/\gamma_{10}}\Gamma\left(\frac{1}{\gamma_{10}}+1\right)\frac{\exp(\mu_1+\mathbf{X}\boldsymbol{\alpha}_1)}{1+\exp(\mu_1+\mathbf{X}\boldsymbol{\alpha}_1)}\frac{\exp(\mu_0+\mathbf{X}\boldsymbol{\alpha}_0)}{1+\exp(\mu_0+\mathbf{X}\boldsymbol{\alpha}_0)}\right]}{E\left[\frac{\exp(\mu_1+\mathbf{X}\boldsymbol{\alpha}_1)}{1+\exp(\mu_1+\mathbf{X}\boldsymbol{\alpha}_1)}\frac{\exp(\mu_0+\mathbf{X}\boldsymbol{\alpha}_0)}{1+\exp(\mu_0+\mathbf{X}\boldsymbol{\alpha}_0)}\right]} \quad (9) \end{aligned}$$

Let $E_n[U_i] = \sum_{i=1}^n U_i/n$ denote the empirical expectation operator for random variable U_i on which we have n i.i.d. observations. We can estimate Equation (9) by using $\hat{\boldsymbol{\psi}}$ in the following equation,

$$\frac{E_n \left[(\hat{\lambda}_{10}\exp(\mathbf{x}_i\hat{\boldsymbol{\beta}}_{10}))^{-1/\hat{\gamma}_{10}}\Gamma\left(\frac{1}{\hat{\gamma}_{10}}+1\right)\frac{\exp(\hat{\mu}_1+\mathbf{x}_i\hat{\boldsymbol{\alpha}}_1)}{1+\exp(\hat{\mu}_1+\mathbf{x}_i\hat{\boldsymbol{\alpha}}_1)}\frac{\exp(\hat{\mu}_0+\mathbf{x}_i\hat{\boldsymbol{\alpha}}_0)}{1+\exp(\hat{\mu}_0+\mathbf{x}_i\hat{\boldsymbol{\alpha}}_0)} \right]}{E_n \left[\frac{\exp(\hat{\mu}_1+\mathbf{x}_i\hat{\boldsymbol{\alpha}}_1)}{1+\exp(\hat{\mu}_1+\mathbf{x}_i\hat{\boldsymbol{\alpha}}_1)}\frac{\exp(\hat{\mu}_0+\mathbf{x}_i\hat{\boldsymbol{\alpha}}_0)}{1+\exp(\hat{\mu}_0+\mathbf{x}_i\hat{\boldsymbol{\alpha}}_0)} \right]} \quad (10)$$

The denominator in Equation (10) can be used for an estimate of $P(R(0) = 1, R(1) = 1)$. The other seven expectations and probabilities of joint latent class membership can be estimated analogously. We use the δ -method using the covariance matrix from the estimated survival mixture models to calculate the standard errors of the estimated survival time.

We can estimate the probabilities of characteristics for each covariate, X_j by class membership analogously using the following estimator for stratum $\{R(0)=1, R(1)=1\}$ under assumptions A.1 - A.3:

$$\frac{E_n \left[x_i \frac{\exp(\hat{\mu}_1 + x_i \hat{\alpha}_1)}{1 + \exp(\hat{\mu}_1 + x_i \hat{\alpha}_1)} \frac{\exp(\hat{\mu}_0 + x_i \hat{\alpha}_0)}{1 + \exp(\hat{\mu}_0 + x_i \hat{\alpha}_0)} \right]}{E_n \left[\frac{\exp(\hat{\mu}_1 + x_i \hat{\alpha}_1)}{1 + \exp(\hat{\mu}_1 + x_i \hat{\alpha}_1)} \frac{\exp(\hat{\mu}_0 + x_i \hat{\alpha}_0)}{1 + \exp(\hat{\mu}_0 + x_i \hat{\alpha}_0)} \right]}$$

4.4.1 Sensitivity analyses—Above, we detailed estimators under assumptions A.1 through A.3. Here, we detail how we can estimate our estimand under assumptions A.4 and A.5.

Let $\hat{P}(\cdot)$ and $\hat{E}[\cdot]$ denote estimated probabilities and estimated expectations, respectively. Under our sensitivity analysis assumption, A.4, we would replace

$\hat{P}(R_i(1)=1|Z_i=1, \mathbf{X}_i=\mathbf{x}_i) = \frac{\exp(\hat{\mu}_1 + x_i \hat{\alpha}_1)}{1 + \exp(\hat{\mu}_1 + x_i \hat{\alpha}_1)}$ in Equation (10) with an estimate of $P(R_i(1)=1|Z=1, R_i(0)=1, \mathbf{X}_i=\mathbf{x}_i)$, obtained by solving Equation (3) with the help of assumption A.1:

$$\frac{\exp(\hat{\mu}_1 + x_i \hat{\alpha}_1)}{1 + \exp(\hat{\mu}_1 + x_i \hat{\alpha}_1)} = \frac{\hat{P}(R_i(1)=1|R_i(0)=1, \mathbf{X}_i=\mathbf{x}_i) \frac{\exp(\hat{\mu}_0 + x_i \hat{\alpha}_0)}{1 + \exp(\hat{\mu}_0 + x_i \hat{\alpha}_0)} + \frac{\text{odds} \hat{P}(R_i(1)=1|R_i(0)=1, \mathbf{X}_i=\mathbf{x}_i)}{\tau + \text{odds} \hat{P}(R_i(1)=1|R_i(0)=1, \mathbf{X}_i=\mathbf{x}_i)} \times \frac{1}{1 + \exp(\hat{\mu}_0 + x_i \hat{\alpha}_0)}}{(11)}$$

Under our second sensitivity analysis assumption, A.5, combined with Assumption A.1, we

would replace $\hat{E}[T_i|R_i(0)=1, Z_i=0, \mathbf{X}_i=\mathbf{x}_i] = (\hat{\lambda}_{10} \exp(x_i \hat{\beta}_{10}))^{-1/\hat{\gamma}_{10}} \Gamma(\frac{1}{\hat{\gamma}_{10}} + 1)$ in Equation (10) with the following estimate of $E[T_i(0)|R_i(0)=1, R_i(1)=1, Z_i=0, \mathbf{X}_i=\mathbf{x}_i]$, which includes $\hat{P}(R_i(1)=1|R_i(0)=1, \mathbf{X}_i=\mathbf{x}_i)$ from Equation (11):

$$\begin{aligned} & \hat{E}[T_i(0)|R_i(0)=1, R_i(1)=1, Z_i=0, \mathbf{X}_i=\mathbf{x}_i] \\ &= \frac{\hat{E}[T_i(0)|R_i(0)=1, Z_i=0, \mathbf{X}_i=\mathbf{x}_i]}{\hat{P}(R_i(1)=1|R_i(0)=1, \mathbf{X}_i=\mathbf{x}_i) + \kappa_{10}(1 - \hat{P}(R_i(1)=1|R_i(0)=1, \mathbf{X}_i=\mathbf{x}_i))} \\ &= \frac{(\hat{\lambda}_{10} \exp(x_i \hat{\beta}_{10}))^{-1/\hat{\gamma}_{10}} \Gamma(\frac{1}{\hat{\gamma}_{10}} + 1)}{\hat{P}(R_i(1)=1|R_i(0)=1, \mathbf{X}_i=\mathbf{x}_i) + \kappa_{10}(1 - \hat{P}(R_i(1)=1|R_i(0)=1, \mathbf{X}_i=\mathbf{x}_i))} \end{aligned}$$

5 Simulations

We performed a simulation study to investigate the finite sample size properties of our method and the properties of our R package. In order to do this, we generated four covariates, $\mathbf{X} = \{X_1, X_2, X_3, X_4\}$, from a multivariate normal distribution with means all

$$\Sigma = \begin{pmatrix} 1 & -.1 & -.3 & -.5 \\ -.1 & 1 & .4 & .6 \\ -.3 & .4 & 1 & .7 \\ -.5 & .6 & .7 & 1 \end{pmatrix}$$
 equal to 2 and covariance matrix following parameters for the Weibull mixture model from Equation 6: $\mu_0 = .7$, $\mu_1 = -1.1$, $\gamma_{00} = 2.1$, $\gamma_{10} = 2.5$, $\gamma_{01} = 2.2$, $\gamma_{11} = 2.4$, $\lambda_{00} = .008$, $\lambda_{10} = .005$, $\lambda_{01} = .006$, $\lambda_{11} = .004$, and

$$\alpha_0 = \begin{pmatrix} -.3 \\ -.3 \\ -.2 \\ .5 \end{pmatrix} \quad \alpha_1 = \begin{pmatrix} .5 \\ -.2 \\ -.1 \\ -.2 \end{pmatrix}$$

$$\beta_{00} = \begin{pmatrix} -.2 \\ -.1 \\ .3 \\ -.1 \end{pmatrix} \quad \beta_{10} = \begin{pmatrix} .4 \\ .2 \\ .2 \\ .3 \end{pmatrix} \quad \beta_{01} = \begin{pmatrix} -.2 \\ .1 \\ .1 \\ -.2 \end{pmatrix} \quad \beta_{11} = \begin{pmatrix} .3 \\ .4 \\ .3 \\ .2 \end{pmatrix}$$

We assumed a censoring time, $C(z) = c$, that was exponentially distributed with $f(c) = .07\exp(-.07c)$. This censored approximately 39% of observations. We examined simulations under two scenarios: when $\tau = \kappa_{rz} = 1$ and $\tau = \kappa_{rz} = 2$ for all $r, z \in \{0, 1\}$. We empirically derived the probability of principal strata membership and expected survival within principal strata as described in Table 1. Under our assumptions, treatment on average is always beneficial in every strata except for $\{R(0) = 0, R(1) = 1\}$, in which treatment is harmful. When $\tau = \kappa_{rz} = 1$, life expectancy under no surgery is shorter for those with $\{R(0) = 0, R(1) = 0\}$ compared to those with $\{R(0) = 0, R(1) = 1\}$ due to different characteristics, X , between the two groups.

We used 1,500 iterations for each simulation. We used sample sizes of 5,364 (the same size as our sample) and 9,000, with 1/3 of the sample in the control arm and 2/3 in the treatment arm. In estimating empirical coverage, we excluded those iterations for which the model did not converge. We examined simulations in which we set the assumed sensitivity parameters to the truth and also when the assumed sensitivity parameters were incorrectly specified. We used incorrect specification to investigate how estimates might be affected if assumptions are not correct.

We present the results of our simulations in Tables 2. The empirical 95% confidence intervals are over 90% and very close to 95% when the sample size is very large, suggesting that our algorithms are appropriate. Our program converged over 99.4% of the time. When τ and κ_{rz} are incorrectly specified, the estimates are biased and the empirical coverage decreases to 0% in some cases. This demonstrates the utility of taking a sensitivity analysis approach to examine how inferences could change if one's assumptions are not correct. Of note, is that the empirical coverage was not as good (results not shown) when we examined simulations in which more than two of the scale or shape parameters were set equal to each other (e.g. $\lambda_{10} = \lambda_{01} = \lambda_{11}$). Hence, the method may not work well when the parameters among the mixture classes are not different enough.

The importance of obtaining appropriate starting values for Weibull mixture models has been noted previously (Mair and Hudec 2009). Anecdotally, we have found that it might take several attempts with different starting values to achieve convergence to a global maxima when using the *strsmix* macro in STATA (Lambert 2007). We let our program, *latentPS* empirically estimate starting values (rather than specifying starting values), and the excellent coverage suggests that the program is successful in doing so.

6 SEER-Medicare Data Example

We applied this methodology to investigate whether there are subgroups who might benefit or be harmed by surgical treatment for early-stage kidney cancer (Stage T1 or T2, node negative, metastatic negative tumors of 20 cm or less in diameter). We used the Surveillance Epidemiology and End Results (SEER) database which is maintained by the National Cancer Institute. SEER has long-term data on tumor characteristics and demographics information about incident cancers in over 14% of the United States, and more recently has been expanded to cover over 25% of the United States population. The National Cancer Institute has linked SEER data to Medicare claims data; we used the Medicare claims data to identify patient comorbidities. Medicare covers almost all individuals over 65 years old in the SEER database. Fee-for-service claims from Medicare part A and part B provide a record of treatments and services obtained before and after cancer diagnosis. For this study, we included all of those with kidney cancer over 66 years of age who had Medicare part A and B (non-HMO) coverage. Including only those 66 years and older ensured that we had a year of pre-diagnosis comorbidity claims data on all patients. Cases were diagnosed between 1995 and 2009 with Medicare claims follow-up through 2010. We defined surgery as having treatment within 6 months of diagnosis.

We initially identified 2,206 (12%) cases with localized kidney cancer not treated and 16,173 (88%) treated with some form of surgery within six months of diagnosis (i.e. partial/radical nephrectomy or ablation). In order to better approximate an intention-to-treat approach, we assigned to the surgery group the small proportion of SEER records indicating that death occurred before recommended surgery was performed.

Because those who had surgery were very different from those who did not have surgery, we used propensity score matching to select our sample, implemented via the *matchit* (Ho *et al.* 2007) package in R (R Statistical Foundation, Vienna, Austria). We used a logistic regression to model the propensity score. To account for pre-cancer diagnosis comorbidity differences between the groups, we used the Charlson Comorbidity Index (Charlson *et al.* 1987, Deyo *et al.* 1992, Romano *et al.* 1993) and the Elixhauser Score (Elixhauser *et al.* 1998, van Walraven *et al.* 2009). Comorbidities were identified in the one-year period prior to cancer diagnosis. Additional variables included in the propensity score model were sex, race, age, size (diameter) of tumor, marital status, and history of prior cancers (non-kidney cancer). We entered the continuous variables into the propensity score and Weibull mixture models using restricted cubic splines with 3 knots (Harrell 2001). We added the binary (sex, marital status, history of prior cancer) and categorical variable (i.e. race) to the model using indicator variables. For cases with multiple kidney cancer diagnoses, we selected the first kidney cancer diagnosis for analysis. Table 3 displays the characteristics of the sample. The

table indicates that the average sample characteristics were generally balanced between the two treatment arms.

When examining cause-specific death via unadjusted cumulative incidence functions in this propensity score matched sample, those who did not have surgery had 5 year mortality rates of 16% from kidney cancer, 19% from other cancers, and 33% from non-cancer causes. Those who had surgery had 5 year mortality rates of 6% from kidney cancer, 7% from other cancers, and 21% from non-cancer causes. The difference in cumulative incidence functions using Gray's test (Gray 1988) for all three outcomes was statistically significant between the surgical status groups ($p < 0.0001$ in all three cases).

In Table 4, we present our survival class estimates when the sensitivity parameters are set to $\tau = 1$ and $\kappa_{rz} = 1$ for all r and z .

The point estimates in Table 4 suggest that most people would potentially benefit from intervention, but a small group in one class could expect very poor outcomes. The treatment effects are statistically significant in three of the four classes; the effect is not statistically significant in class 4 in which $R(0) = 1$ and $R(1) = 1$. In Figure 1, we show the marginal (i.e. population-level) Weibull-mixture model estimated survival curves for each of the principal strata.

More specifically, Class 1 depicts a group that comprises a large majority (74.9%) of the total that has relatively long life expectancy (LE) without surgery, but still has a substantial surgical benefit of 3.6 years ($p < 0.001$). This group is notable in that it has the smallest average tumor size and lowest Charlson comorbidity burden of any group. In addition, the group has a relatively large proportion of white and married cases. Although not statistically significant, the larger percentage of married people in class 1 may be a proxy measure for better social support that can improve outcomes.

Also shown in Table 4, there are 5.1% in class 2 for whom intervention seems particularly harmful, reducing life expectancy by over 4 years ($p < 0.0001$). This is a group that has relatively small tumors but relatively worse comorbidities. It is conceivable that removing the kidney causes treatment related problems in this group. Indeed, although included in the model only through the Elixhauser and Charlson scores, class 2 had a substantial proportion with pre-cancer diagnosis renal failure (21.3%); it is plausible that surgical removal of a kidney would be especially harmful for this group.

Class 3 represents an estimated 18.0% of the group in which intervention is particularly beneficial, raising average life expectancy by 8.3 years ($p < 0.001$). This is a group with medium size tumors and relatively low Charlson comorbidity burdens.

Class 4 represents 1.9% of the sample which has particularly bad outcomes without treatment, and does not receive a statistically significant life expectancy increase from intervention (0.5 years, $p = 0.40$). This class with relatively poor outcomes has the largest tumors, on average, and relatively high Charlson comorbidity burdens. A significant proportion of class 4 is black (44.3%).

As everyone in our sample had Medicare coverage for at least a year prior to diagnosis, it is unlikely that differences in patients' current insurance status were having a major impact on outcomes.

We performed likelihood ratio tests, as well as examined AIC and BIC statistics to assess our two-class interaction model choice. In terms of the likelihood ratio tests, both of our Weibull mixture models fit in the separate treatment arms better than the traditional Weibull proportional hazards regression in both cases. ($\chi^2 = 134,29$ d.f., $p < 0.0001$ for those in the sample who did not have intervention. $\chi^2 = 147,29$ d.f., $p < 0.0001$ for those who did have intervention). Estimating Weibull mixture models separately for the intervention arms provided a better fit than a single Weibull mixture model on the whole sample (i.e. a main effects only model versus a full interaction model $\chi^2 = 768,44$ d.f., $p < 0.0001$). In terms of AIC, a 3-class interaction model performed the best (AIC=13,075) compared to the 2-class interaction model (AIC=13,084). Weibull regressions with or without interactions and mixture models without interactions had larger AIC statistics (AIC 13,250). In terms of BIC, the 2-class interaction model was superior (BIC=13,664) compared to all other models which all had BIC statistics above 14,000. The totality of the evidence suggests that the two-class mixture model provides an appropriate balance of a good model fit with a number of classes that most readily addresses our prior hypotheses of the existence of four principal strata (those who benefit and do not benefit from intervention cross-classified by those who have generally longer or shorter survival). While the AIC statistic suggested that a 3-class mixture model might be better (which would result in $3 \times 3 = 9$ principal strata), this was not supported by the BIC statistic.

6.1 Sensitivity Analyses

Of note is that the previous findings relied on the relatively strong assumptions that our sensitivity parameters, τ and κ_{TZ} were equal to one. In Table 5, we present the results of the sensitivity analyses. In the table, we present estimates at sensitivity parameter values of 1 and 2, and also at extreme values of 5 and 20. Although the life expectancies are presented in rows in which the κ_{TZ} are all equal, the survival estimates can be obtained in the table within values of τ for a mixture of κ_{TZ} values as the κ_{TZ} parameters independently impact each of four pairs of survival estimates.

In the sensitivity analysis, the prior probabilities of survival class membership are not substantially affected by the sensitivity parameter τ . The κ_{TZ} parameters have no impact on the probability of class membership as the κ_{TZ} parameters are only used to assess relationships of average survival outcomes across principal strata. However, the τ parameters do affect the relationships of the probabilities, and the effect on the relationships is modest. This is a strength of the analysis because it demonstrates that the estimates of the probability of class membership are robust to deviations from our assumptions.

In terms of the expected survival within each class, point estimates indicate that surgery is always better than non-surgery in three of the four groups (Classes 1, 3, and 4). The direction of the surgical effect is harmful in class 2 for all τ when $\kappa_{TZ} = 2$. The effect of surgery is never statistically significant in class 4, and becomes non-statistically significant in class 2 when $\kappa_{TZ} = 5$ for all τ .

The clinical and demographic characteristics of each class, as shown in Table 4, change with respect to τ but are independent of the κ_{TZ} sensitivity parameters. Hence, the general characteristics of the classes do not change substantially with modest changes in τ since the probabilities of class membership do not change very much with changes in τ (results not shown).

Although we do not know the true values of the sensitivity parameters, and the observed data cannot directly identify them without additional assumptions, it may be possible to use scientific reasoning and observed data statistics (e.g. Shepherd *et al.* 2006, pp. 338-339) to place reasonable upper bounds on them. Kaplan-Meier survival estimates suggest that ten-year survival is 14.2% and median survival is 2.7 years for those without surgery and 34.1% and 7.3 years for those with surgery. Such ten-year and median survival estimates do not seem consistent with mixture populations in which 74.9% or more of the population (Class 1) has life expectancy with surgery 10.7 years or more when $\kappa_{TZ} = 5$. The life expectancies with smaller values of κ_{TZ} may be more valid.

7 Conclusion

For medical treatments with heterogeneous effects, researchers may find our principal strata-latent class survival methods useful for identifying possible subgroups who may particularly benefit or be harmed by treatment. Clinical trials could be designed specifically for those subgroups for which treatment seems to have the most clinical equipoise. The benefits of surgical treatment for localized kidney cancer, for example, have been questioned (Kunkle *et al.*, 2008). Some of the evidence concerning the necessity of surgical treatment comes from tumor observation studies that find that many localized kidney tumors are slow growing (Smaldone *et al.* 2012). However, there are few randomized clinical trials concerning the effectiveness of active surveillance compared to surgical treatment for management of cancer.

We have demonstrated how Weibull survival mixture models combined with principal stratification can be used for inferential purposes in evaluating whether there are heterogeneous subgroups in which the effectiveness of treatment varies. In our propensity score matched sample, we found a group that comprises approximately 5% of the population in which partial or complete removal of the kidney might be particularly harmful. It is clinically plausible that such a group exists since removal of a kidney can lead to chronic kidney disease or operative complications in patients. The data suggests that those with smaller tumors and a relatively large number of comorbidities might not be good candidates for surgery. Further research into biomarkers that could more uniquely identify the patients who comprise this group would be worthwhile. Contrary to our expectations that many kidney cancer are indolent, the number of people for whom intervention had no effect on outcomes was very small. For example, while Class 1 was generally longer lived and comprised approximately three-quarters of the population, there seemed to be a clear surgery benefit in this group. Although the benefit of intervention varied by class, the vast majority of patients seemed to benefit from surgery.

There are limitations to our substantive findings. It is possible that those who did not receive surgery were so different from those who did receive surgery that there was unmeasured confounding affecting our inferences. The fact that only 12% of the sample did not have surgery and the Kaplan-Meier median survival estimates were so different, even after propensity score matching, does provide some evidence that residual confounding could still be a problem. Still, it is unlikely that the mortality differences are driven solely by unmeasured comorbidity differences between the samples that did and did not have surgery. Differences in cancer-specific mortality rates between the two groups suggest that kidney cancer itself could be driving the difference. When examining cumulative incidence functions, 19% of the group that did not have intervention died from kidney cancer within 10 years compared with 9% of those who did have intervention ($p < 0.0001$). Further, 21% of those who did not have intervention died from another type of cancer within 10 years compared with 13% of those who did have intervention ($p < 0.0001$). Because of potential sources of error in death records, it is possible that many of the other cancer deaths were related to distant metastatic disease originating from the primary kidney cancer tumor.

Our simulations suggest that our method is a valid method if all of the confounders are known and the target population matches the sample population used for estimation. However, as with most scientific endeavors using observational (i.e. non-randomized) data, it is important that results be confirmed in an independent external sample before being implemented in clinical practice. In our specific case, confirmation could first include replicating the findings in a non-SEER-Medicare observational dataset, and then designing a clinical trial in which treatment is randomized within the subgroups for which the principal stratification results suggest that surgery might not be beneficial. For a clinical trial, an individual's predicted probability of principal stratum membership could be used to determine whether the individual is assumed to be in a relevant subgroup. Higher probabilities of membership would give more assurance that the subgroup assignment is correct. Clinical trial estimated treatment effects within the subgroups that are similar to those presented in this paper would provide confirmation of our findings.

While we used propensity score matching to reduce differences between the two groups and hence improve the internal validity of our study, matching reduces the generalizability of our findings compared to what might be expected if active surveillance were used in a broader group of people (i.e. more diverse in terms of demographics and health status). In our sample, propensity scores suggested that many of those who did not have surgery were too different from those who had surgery to include in the sample. Using the language of clinical trials, our matched analysis was designed primarily to assess the efficacy of the intervention, which is related to the “true biological effect” (Piantadosi 1997). That is, our results are valid for making inferences with respect to how treatment has historically been used, and with respect to the historic subpopulation likely to receive observation rather than active treatment. As active surveillance (i.e. observation) becomes a more common management strategy for indolent cancers in general, more diverse groups of patients will be managed using active surveillance, which will allow for better assessment of its effectiveness. Effectiveness is related to the effect when a therapeutic strategy is used in the general population (Piantadosi 1997). Efficacy is related to the internal validity of a study, while effectiveness is related to the generalizability and external validity of a study.

A strength of our analytic method was that our inferences were not heavily dependent on the sensitivity parameter τ . While the expected survival estimates were more sensitive to changes in the κ_{TZ} parameters, we do not believe that the value of κ_{TZ} was likely to be large enough to alter our general conclusions, as discussed in section 6.1.

The benefit of our method is two fold. One, we can estimate the proportion in each class, and two, the mixture models provide greater flexibility in modeling survival data. Our inferences would not have been as rich had we simply used a Cox or Weibull proportional hazards regression. While we estimated life expectancy conditional on principal strata, our methods can be extended to investigate median survival or the probability of survival to a fixed time point (e.g. 10 year survival rates).

Our work is methodologically innovative since we used principal stratification to link latent survival classes. While principal stratification itself is a form of latent class modeling (Grilli 2011), it is different from the traditional latent class modeling framework. In this work, we have provided assumptions for the identification of principal strata-latent classes of latent survival classes. While there have been a plethora of studies on principal stratification (Egleston *et al.* 2010, Wang et al. *et al.* 2014), defining principal strata based on latent potential outcomes (e.g. our latent survival classes) is still novel. Our results provide preliminary evidence of the proportion and types of people with localized kidney cancer who may be substantially harmed by surgical treatment for their condition. A limitation of our findings is that we estimated effects in a propensity score matched sample which may not be representative of the broader population. Future methodological research can focus on extrapolating estimates to an out-of-sample population. Such extrapolation would allow researchers to maintain internal validity by choosing an appropriately balanced sample, yet still estimate generalizable effects in a larger population. Overall, our methods can be useful for clinicians who seek to identify subgroups who might have very heterogeneous responses to a treatment.

Acknowledgments

This work was supported by the National Institutes of Health, National Cancer Institute, grants P30CA006927 and R03CA152388. We thank Drs. Samuel Litwin, Elizabeth A. Handorf, and Eric A. Ross for their comments.

References

- Abramowitz, M., Stegun, IA. National Bureau of Standards Applied Mathematics Series. Vol. 55. Washington, D.C: U.S. Government Printing Office; 1972. Handbook of mathematical functions with formulas, graphs, and mathematical tables.
- Bandeem-Roche K, Miglioretti DL, Zeger SL, Rathouz PJ. Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association*. 1997; 92(440):1375–1386.
- Bieler GS, Williams RL. Cluster sampling techniques in quantal response teratology and developmental toxicity studies. *Biometrics*. 1995; 51:764–776. [PubMed: 7662858]
- Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis*. 1987; 40(5):373–383. [PubMed: 3558716]
- Chow WH, Devesa SS, Warren JL, Fraumeni JF. Rising incidence of renal cell cancer in the United States. *JAMA*. 1999; 281(17):1628–1631. [PubMed: 10235157]

- Cole SR, Frangakis CE. The Consistency Statement in Causal Inference. A Definition or an Assumption? *Epidemiology*. 2009; 20(1):3–5. [PubMed: 19234395]
- Desai NR, Bradley SM, Parzynski CS, Nallamothu BK, Chan PS, Spertus JA, Patel MR, Ader J, Soufer A, Krumholz HM, Curtis JP. Appropriate Use Criteria for Coronary Revascularization and Trends in Utilization, Patient Selection, and Appropriateness of Percutaneous Coronary Intervention. *JAMA*. 2015; 314(19):2045–2053. [PubMed: 26551163]
- Deyo RA, Cherkin DC, Ciol MA. Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. *J Clin Epidemiol*. 1992; 45(6):613–619. [PubMed: 1607900]
- Dominici F, Zeger SL, Parmigiani G, Katz J, Christian P. Estimating Percentile-Specific Effects in Counterfactual Models: A Case Study of Micronutrient Supplementation, Birth Weight, and Infant Mortality. *Journal of the Royal Statistical Society, Series C*. 2006; 55:261–280. DOI: 10.1111/j.1467-9876.2006.00533.x
- Egleston BL, Cropsey KL, Lazev AB, Heckman CJ. Tutorial on principal stratification-based sensitivity analysis: Application to smoking cessation studies. *Clinical Trials*. 2010; 7(3):286–98. DOI: 10.1177/1740774510367811 [PubMed: 20423924]
- Egleston BL, Scharfstein DO, Freeman EE, West SK. Causal inference for non-mortality outcomes in the presence of death. *Biostatistics*. 2007; 8(3):526–545. [PubMed: 16980696]
- Egleston BL, Scharfstein DO, MacKenzie E. On estimation of the survivor average causal effect in observational studies when important confounders are missing due to death. *Biometrics*. 2009; 65(2):497–504. DOI: 10.1111/j.1541-0420.2008.01111.x [PubMed: 18759833]
- Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. *Medical Care*. 1998; 36(1):8–27. [PubMed: 9431328]
- Farewell VT. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*. 1982; 38:1041–1046. [PubMed: 7168793]
- Frangakis CE, Rubin DB. Principal stratification in causal inference. *Biometrics*. 2002; 58(1):21–9. [PubMed: 11890317]
- Frumento P, Mealli F, Pacini B, Rubin DB. Evaluating the effect of training on wages in the presence of noncompliance, nonemployment, and missing outcome data. *Journal of the American Statistical Association*. 2012; 107(498):450–466.
- Gallop R, Small DS, Lin JY, Elliott MR, Joffe M, Ten Have TR. Mediation analysis with principal stratification. *Statistics in Medicine*. 2009; 28:11081130.
- Gray RJ. A class of K-sample tests for comparing the cumulative incidence of a competing risk. *Annals of Statistics*. 1988; 16(3):1141–1154.
- Grilli, L. Causal inference through principal stratification: a special type of latent class modelling. In: Fichet, B., Piccolo, D., Verde, R., Vichi, M., editors. *Classification and Multivariate Analysis for Complex Data Structures*. Springer-Verlag; Berlin Heidelberg: 2011. p. 265–270.
- Halpern SD, French B, Small DS, Saulsger K, Harhay MO, Audrain-McGovern J, Loewenstein G, Brennan TA, Asch DA, Volpp KG. Randomized trial of four financial-incentive programs for smoking cessation. *New England Journal of Medicine*. 2015; 372:2108–2117. DOI: 10.1056/NEJMoa1414293 [PubMed: 25970009]
- Harrell, FE, Jr. *Regression Modeling Strategies*. Vol. Chapter 2. New York: Springer; 2001. p. 11–40.
- Hayden D, Pauler DK, Schoenfeld D. An estimator for treatment comparisons among survivors in randomized trials. *Biometrics*. 2005; 61(1):305–310. [PubMed: 15737107]
- Ho D, Imai K, King G, Stuart E. Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis*. 2007; 15(3):199–236. DOI: 10.1093/pan/mpi013
- Hock LM, Lynch J, Balaji KC. Increasing incidence of all stages of kidney cancer in the last 2 decades in the United States: An analysis of Surveillance, Epidemiology, and End Results Program Data. *Journal of Urology*. 2002; 167:57–60. [PubMed: 11743275]
- Hollingsworth JM, Miller DC, Daignault S, Hollenbeck BK. Rising incidence of small renal masses: A need to reassess treatment effect. *Journal of the National Cancer Institute*. 2006; 98(18):1331–1334. [PubMed: 16985252]
- Huber PJ. Robust estimation of a location parameter. *Annals of Mathematical Statistics*. 1964; 35(1):73–101.

- Imai K, Yamamoto T. Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments. *Political Analysis*. 2013; 21:141–171.
- Jo B, Stuart EA. On the use of propensity scores in principal causal effect estimation. *Statistics in Medicine*. 2009; 28:2857–2875. [PubMed: 19610131]
- Jo B, Wang CP, Ialongo NS. Using latent outcome trajectory classes in causal inference. *Statistics and Its Interface*. 2009; 2:403–412. [PubMed: 20445809]
- Kristensen VN, Lingjærde OC, Russnes HG, Vollan HKM, Frigessi A, Borrenson-Dale AL. Principles and methods of integrative genomic analyses in cancer. *Nature Reviews Cancer*. 2014; 14:299–313. [PubMed: 24759209]
- Kunkle DA, Egleston BL, Uzzo RG. Excise, ablate or observe: The small renal mass dilemma- A meta-analysis and review. *Journal of Urology*. 2008; 179:1227–1234. [PubMed: 18280512]
- Lambert PC. Modeling the cure fraction in survival studies. *STATA Journal*. 2007; 7(3):351–375.
- Lanza ST, Rhoades BL. Latent Class Analysis: An alternative perspective on subgroup analysis in prevention and treatment. *Prevention Science*. 2013; 14:157–168. [PubMed: 21318625]
- Lehmann, EL. *Elements of large-sample theory*. New York: Springer; 2010.
- Mair P, Hudec M. Multivariate Weibull mixtures with proportional hazard restrictions for dwell-time-based session clustering with incomplete data. *Applied Statistics*. 2009; 58(5):619–639.
- Marin JM, Rodriguez-Bernal MT, Wiper MP. Using Weibull mixture distributions to model heterogeneous survival data. *Communications in Statistics- Simulation and Computation*. 2005; 34(3):673–684. DOI: 10.1081/SAC-200068372
- Muthén B, Brown CH, Masyn K, Jo B, Khoo ST, Yang CC, Wang CP, Kellam SG, Carlin JB, Liao J. General growth mixture modeling for randomized preventive interventions. *Biostatistics*. 2002; 3(4):459–475.
- Muthén B, Masyn K. Discrete-Time Survival Mixture Analysis. *Journal of Educational and Behavioral Statistics*. 2005; 30(1):27–58.
- Muthén B, Brown HC. Estimating drug effects in the presence of placebo response: Causal inference using growth mixture modeling. *Statistics in Medicine*. 2009; 28(27):3363–3385. [PubMed: 19731223]
- Pearl, J. *Causality: Models, Reasoning, and Inference*, Second Edition. Cambridge, UK: Cambridge University Press; 2009.
- Piantadosi, S. *Clinical Trials: A Methodologic Perspective*. New York: John Wiley and Sons; 1997. p. 197–521.
- Qu P, Barlogie B, Crowley J. Using a latent class model to refine risk stratification in multiple myeloma. *Statistics in Medicine*. 2015; 21(20):2971–2980.
- Razali AM, Al-Wakeel AA. Mixture Weibull distributions for fitting failure times data. *Applied Mathematics and Computation*. 2013; 219:11358–11364.
- Romano PS, Roos LL, Jollis JG. Adapting a clinical comorbidity index for use with ICD-9-CM administrative data: differing perspectives. *J Clin Epidemiol*. 1993; 46(10):1075–1079. [PubMed: 8410092]
- Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983; 70(1):41–55.
- Roy J, Hogan JW, Marcus BH. Principal stratification with predictors of compliance for randomized trials with 2 active treatments. *Biostatistics*. 2008; 9(2):277–289. [PubMed: 17681993]
- Rubin DB. Which Ifs have causal answers. Commentary on Holland PW. *Statistics and Causal Inference*. *Journal of the American Statistical Association*. 1986; 81(396):961–962.
- Shepherd BE, Gilbert PB, Jemai Y, Rotnitzky A. Sensitivity analyses comparing outcomes only existing in a subset selected post-randomization, conditional on covariates, with application to HIV vaccine trials. *Biometrics*. 2006; 62:332–342. [PubMed: 16918897]
- Shepherd BE, Gilbert PB, Lumley T. Sensitivity analyses comparing time-to-event outcomes existing only in a subset selected postrandomization. *Journal of the American Statistical Association*. 2007; 102(478):573–582. [PubMed: 19122791]

- Smaldone MC, Kutikov A, Egleston BL, Canter DJ, Viterbo R, Chen DY, Jewett MA, Greenberg RE, Uzzo RG. Small renal masses progressing to metastases under active surveillance: A systematic review and pooled analysis. *Cancer*. 2012; 118(4):997–1006. [PubMed: 21766302]
- Sun M, Becker A, Tian Z, Roghmann F, Abdollah F, Larouche A, Karakiewicz PI, Trinh QD. Management of Localized Kidney Cancer: Calculating Cancer-specific Mortality and Competing Risks of Death for Surgery and Nonsurgical Management. *European Urology*. 2014; 65(1):235–241. DOI: 10.1016/j.eururo.2013.03.034 [PubMed: 23567066]
- van Walraven C, Austin PC, Jennings A, Quan H, Forster AJ. A modification of the Elixhauser comorbidity measures into a point system for hospital death using administrative data. *Medical Care*. 2009; 47:626–633. DOI: 10.1097/MLR.0b013e31819432e5 [PubMed: 19433995]
- Viatte S, Plant D, Han B, Fu B, Yarwood A, Thomson W, Symmons DPM, Worthington J, Yound A, Hyrich KL, Morgan AW, Wilson AG, Isaacs JD, Raychaudhuri S, Barton A. Association of HLA-DRB1 Haplotypes With Rheumatoid Arthritis Severity, Mortality, and Treatment Response. *JAMA*. 2015; 313(16):1645–1656. [PubMed: 25919528]
- Wang CP, Jo B, Brown CH. Causal inference in longitudinal comparative effectiveness studies with repeated measures of a continuous intermediate variable. *Statistics in Medicine*. 2014; 33:3509–3527. [PubMed: 24577715]
- Williams RL. A note on robust variance estimation for cluster-correlated data. *Biometrics*. 2000; 56:645–646. [PubMed: 10877330]

Figure 1.
Estimated survival curves showing survival differences among principal strata after matching by propensity score.

Table 1

True values of the sensitivity parameters, probabilities of principal stratum membership, and expected survival by treatment status within each principal strata.

Characteristic	Simulation 1	Simulation 2
Sensitivity Parameter τ	1	2
Sensitivity Parameters κ_{rz}	1	2
$P(R(0) = 0, R(1) = 0)$	33.4%	36.2%
$P(R(0) = 0, R(1) = 1)$	14.2%	11.4%
$P(R(0) = 1, R(1) = 0)$	39.3%	36.5%
$P(R(0) = 1, R(1) = 1)$	13.0%	15.9%
$E[T(0) R(0) = 0, R(1) = 0]$	9.8	11.2
$E[T(0) R(0) = 0, R(1) = 1]$	10.5	6.5
$E[T(0) R(0) = 1, R(1) = 0]$	3.2	3.7
$E[T(0) R(0) = 1, R(1) = 1]$	3.3	2.1
$E[T(1) R(0) = 0, R(1) = 0]$	10.9	14.2
$E[T(1) R(0) = 0, R(1) = 1]$	3.6	5.0
$E[T(1) R(0) = 1, R(1) = 0]$	10.7	7.4
$E[T(1) R(0) = 1, R(1) = 1]$	3.7	2.7

Table 2

Bias and empirical 95% confidence interval coverage. The true and assumed sensitivity parameters are varied to examine coverage under correct and mis-specified assumptions.

Characteristics				
Sample size	5,364	9,000	5,364	9,000
True τ , κ_{rz}	1,1	1,1	1,1	1,1
Assumed τ , κ_{rz}	1,1	1,1	2,2	2,2
$P(R(0) = 0, R(1) = 0)$	-0.005, 90.5%	-0.002, 93.3%	0.023, 87.6%	0.024, 84.6%
$P(R(0) = 0, R(1) = 1)$	0.001, 93.4%	0.001, 93.4%	-0.027, 61.9%	-0.028, 43.3%
$P(R(0) = 1, R(1) = 0)$	0.000, 93.0%	-0.001, 93.3%	-0.028, 87.8%	-0.026, 84.0%
$P(R(0) = 1, R(1) = 1)$	0.004, 92.7%	0.002, 94.2%	0.032, 60.0%	0.030, 45.4%
$E[T(0) R(0) = 0, R(1) = 0]$	0.079, 92.0%	0.021, 92.5%	1.453, 31.4%	1.429, 13.0%
$E[T(0) R(0) = 0, R(1) = 1]$	0.079, 91.8%	0.013, 92.5%	-4.018, 0.0%	-4.032, 0.0%
$E[T(0) R(0) = 1, R(1) = 0]$	0.012, 94.5%	0.005, 94.3%	0.544, 1.0%	0.538, 0.0%
$E[T(0) R(0) = 1, R(1) = 1]$	0.018, 94.0%	0.008, 93.5%	-1.209, 0.0%	-1.219, 0.0%
$E[T(1) R(0) = 0, R(1) = 0]$	0.047, 93.8%	0.025, 94.3%	3.429, 0.0%	3.430, 0.0%
$E[T(1) R(0) = 0, R(1) = 1]$	0.020, 93.8%	0.007, 93.3%	1.418, 0.1%	1.403, 0.0%
$E[T(1) R(0) = 1, R(1) = 0]$	0.042, 94.9%	0.025, 94.4%	-3.288, 0.0%	-3.307, 0.0%
$E[T(1) R(0) = 1, R(1) = 1]$	0.029, 94.4%	0.007, 94.0%	-0.980, 0.1%	-0.994, 0.0%
True τ , κ_{rz}	2,2	2,2	2,2	2,2
Assumed τ , κ_{rz}	1,1	1,1	2,2	2,2
$P(R(0) = 0, R(1) = 0)$	-0.032, 77.7%	-0.030, 72.7%	-0.004, 92.3%	-0.003, 92.1%
$P(R(0) = 0, R(1) = 1)$	0.029, 58.1%	0.029, 36.3%	0.001, 93.0%	0.000, 95.1%
$P(R(0) = 1, R(1) = 0)$	0.029, 79.7%	0.028, 76.3%	0.001, 93.1%	0.001, 94.2%
$P(R(0) = 1, R(1) = 1)$	-0.026, 74.9%	-0.027, 58.7%	0.002, 93.4%	0.002, 92.9%
$E[T(0) R(0) = 0, R(1) = 0]$	-1.289, 33.9%	-1.307, 9.9%	0.081, 92.2%	0.068, 93.0%
$E[T(0) R(0) = 0, R(1) = 1]$	4.139, 0.0%	4.127, 0.0%	0.056, 93.2%	0.051, 92.9%
$E[T(0) R(0) = 1, R(1) = 0]$	-0.513, 1.1%	-0.523, 0.0%	0.020, 93.3%	0.009, 95.3%
$E[T(0) R(0) = 1, R(1) = 1]$	1.245, 0.0%	1.234, 0.0%	0.019, 93.9%	0.011, 95.1%
$E[T(1) R(0) = 0, R(1) = 0]$	-3.337, 0.0%	-3.350, 0.0%	0.057, 92.9%	0.041, 93.8%
$E[T(1) R(0) = 0, R(1) = 1]$	-1.374, 0.0%	-1.375, 0.0%	0.017, 93.4%	0.016, 93.5%
$E[T(1) R(0) = 1, R(1) = 0]$	3.387, 0.0%	3.378, 0.0%	0.059, 92.8%	0.037, 94.2%
$E[T(1) R(0) = 1, R(1) = 1]$	1.032, 0.1%	1.027, 0.0%	0.024, 94.1%	0.017, 93.8%

Table 3

Characteristics of sample by treatment status after matching by propensity score.

	No Surgery	Surgery	P-value
N	1,788	3,576	
Age (SD)	76.9 (5.9)	76.7 (5.6)	0.23
Size in cm (SD)	4.3 (2.5)	4.3 (2.4)	0.96
Charlson Score ¹ (SD)	1.2 (1.5)	1.2 (1.4)	0.88
Elixhauser Score ¹ (SD)	3.1 (5.1)	3.1 (5.0)	0.98
Female	42.0%	42.6%	0.65
Married	51.5%	52.7%	0.40
Sequence of cancer ²			0.15
First in lifetime	71.9%	73.8%	
Not first in lifetime	28.1%	26.2%	
Race			0.95
Black	13.0%	13.0%	
White	83.5%	83.4%	
Asian/Other	3.5%	3.7%	

¹Higher scores indicate worse comorbidity status.

²Kidney cancer is the first kidney cancer recorded. However, cases may have had a prior cancer in another organ.

Table 4
Characteristics of principal strata survival classes using propensity-score matched sample for estimation (n=5,364). Mean/Percent(Standard Error)

Class	Class 1	Class 2	Class 3	Class 4	P-value
R(0)	R(0)=0	R(0)=0	R(0)=1	R(0)=1	
R(1)	R(1)=0	R(1)=1	R(1)=0	R(1)=1	
Proportion in class	74.9% (0.03)	5.1% (0.02)	18.0% (0.02)	1.9% (0.01)	<0.001
LE ¹ without surgery (years)	5.9 (0.3)	5.2 (0.6)	0.4 (0.1)	0.5 (0.1)	<0.001
LE ¹ with surgery (years)	9.5 (0.3)	0.7 (0.5)	8.7 (0.3)	1.0 (0.6)	<0.001
Benefit of intervention	+3.6(0.4)	-4.4(0.7)	+8.3(0.3)	+0.5(0.6)	<0.001
Age	76.8 (0.1)	76.2 (1.0)	76.7 (0.4)	75.5 (0.9)	0.25
Size in cm	4.0 (0.05)	4.3 (0.2)	5.5 (0.2)	6.1 (0.5)	<0.001
Charlson Score	1.1 (0.04)	2.0 (0.2)	1.2 (0.1)	2.2 (0.2)	<0.001
Elixhauser Score	2.8 (0.1)	5.9 (0.5)	3.4 (0.4)	6.6 (0.5)	<0.001
Renal failure	9.1% (0.01)	21.3% (0.03)	10.7% (0.01)	26.4% (0.04)	<0.001
Female	41.1% (0.01)	31.7% (0.06)	51.0% (0.04)	40.8% (0.06)	0.099
Married	52.3% (0.01)	46.0% (0.09)	54.3% (0.04)	47.9% (0.09)	0.88
Sequence of kidney cancer					
First cancer in lifetime	75.4% (0.01)	75.3% (0.05)	64.0% (0.04)	66.5% (0.07)	0.02
Race					
Black	9.3% (0.01)	21.6% (0.06)	22.5% (0.04)	44.3% (0.08)	0.001
White/Asian/Other ²	91.7%	78.4%	77.5%	55.7%	0.001

¹LE = Life Expectancy

²While these are latent class estimates that do not represent observed sample proportions, we combined the white, Asian, and other-race groups to prevent perceived identification of cases due to small cell sizes.

Table 5

Sensitivity analysis results of principal strata survival classes. The potential outcome variables, $R(z)$, are noted above the appropriate columns.

τ	κ_z	Proportion in Class				Life Expectancy (years)							
		Class 1	Class 2	Class 3	Class 4	Class 1	Class 2	Class 3	Class 4	Surgery	Surgery	Surgery	Surgery
		$R(0)=0$	$R(0)=0$	$R(0)=1$	$R(0)=1$	$R(0)=0$	$R(0)=0$	$R(0)=1$	$R(0)=1$				
		$R(1)=0$	$R(1)=1$	$R(1)=0$	$R(1)=1$	$R(1)=0$	$R(1)=1$	$R(1)=0$	$R(1)=1$				
1	1	0.749	0.051	0.180	0.019	5.9	9.5	5.2	0.7	0.4	8.7	0.45	0.95
1	2	0.749	0.051	0.180	0.019	6.1	10.4	2.8	0.9	0.4	5.1	0.26	0.63
1	5	0.749	0.051	0.180	0.019	6.2	11.0	1.2	1.0	0.4	2.4	0.11	0.33
1	20	0.749	0.051	0.180	0.019	6.2	11.4	0.3	1.1	0.4	0.7	0.03	0.11
2	1	0.756	0.044	0.173	0.026	5.9	9.5	5.1	0.7	0.4	8.7	0.42	0.90
2	2	0.756	0.044	0.173	0.026	6.0	10.3	2.7	0.9	0.4	5.1	0.25	0.63
2	5	0.756	0.044	0.173	0.026	6.1	11.0	1.1	1.1	0.4	2.4	0.11	0.35
2	20	0.756	0.044	0.173	0.026	6.2	11.3	0.3	1.2	0.4	0.7	0.03	0.12
5	1	0.767	0.033	0.163	0.037	5.9	9.5	5.1	0.7	0.4	8.7	0.39	0.85
5	2	0.767	0.033	0.163	0.037	6.0	10.3	2.7	1.0	0.4	5.1	0.24	0.64
5	5	0.767	0.033	0.163	0.037	6.1	10.8	1.1	1.2	0.4	2.3	0.11	0.39
5	20	0.767	0.033	0.163	0.037	6.1	11.2	0.3	1.5	0.5	0.7	0.03	0.15
20	1	0.781	0.019	0.148	0.052	5.9	9.5	5.0	0.8	0.4	8.8	0.36	0.81
20	2	0.781	0.019	0.148	0.052	5.9	10.2	2.6	1.1	0.4	5.1	0.23	0.67
20	5	0.781	0.019	0.148	0.052	6.0	10.7	1.1	1.7	0.5	2.3	0.11	0.48
20	20	0.781	0.019	0.148	0.052	6.0	11.0	0.3	2.3	0.5	0.7	0.03	0.23