HHS Public Access

Author manuscript

Ann Appl Stat. Author manuscript; available in PMC 2019 January 07.

Published in final edited form as:

Ann Appl Stat. 2018 September ; 12(3): 1773–1795. doi:10.1214/17-AOAS1130.

ESTIMATING AND COMPARING CANCER PROGRESSION RISKS UNDER VARYING SURVEILLANCE PROTOCOLS

Jane M. Lange, PhD[*], **Roman Gulati, MS**[*], **Amy S. Leonardson, MS**[*], **Daniel W. Lin, MD**[†], **Lisa F. Newcomb, PhD**[*], **Bruce J. Trock, PhD**[‡], **H. Ballentine Carter, MD**[‡], **Matthew R. Cooperberg, MD**[§], **Janet E. Cowan, MA**[§], **Lawrence H. Klotz, MD**[¶], and **Ruth Etzioni, PhD**[*,†]

[*] Fred Hutchinson Cancer Research Center

[†] University of Washington

[‡] Johns Hopkins University

[§] University of California, San Fransisco

[¶] University of Toronto

Abstract

Outcomes after cancer diagnosis and treatment are often observed at discrete times via doctor-patient encounters or specialized diagnostic examinations. Despite their ubiquity as endpoints in cancer studies, such outcomes pose challenges for analysis. In particular, comparisons between studies or patient populations with different surveillance schema may be confounded by differences in visit frequencies. We present a statistical framework based on multistate and hidden Markov models that represents events on a continuous time scale given data with discrete observation times. To demonstrate this framework, we consider the problem of comparing risks of prostate cancer progression across multiple active surveillance cohorts with different surveillance frequencies. We show that the different surveillance schedules partially explain observed differences in the progression risks between cohorts. Our application permits the conclusion that differences in underlying cancer progression risks across cohorts persist after accounting for different surveillance frequencies.

1 Introduction.

Many outcomes after cancer diagnosis and treatment are observed at discrete times via doctor-patient encounters or specialized diagnostic examinations (Sridhara, Mandrekar and Dodd, 2013). For example, prostate cancer progression following primary surgery is typically asymptomatic and is identified by a high or rising prostate-specific antigen (PSA) level on follow-up testing (Stephenson et al., 2006). Similarly, breast cancer recurrence after an initial diagnosis of in-situ disease is generally identified by surveillance mammography screening (Narod and Rakovitch, 2014). A surveillance-dependent random variable is a continuous-time failure outcome that is detected by exams or biomarker measurements that occur at discrete times (e.g., patient visits). A key feature of studies with surveillance-

Fred Hutchinson Cancer Reesarch Center 1100 Fairview Avenue North Seattle, WA 98109-1024, jlange@fredhutch.org.

dependent outcomes is that the observed time of an event is sensitive to the frequency of patient visits: patients with more frequent visits will have an event detected earlier (Gignac et al., 2008; Zeng et al., 2015). Other differences across surveillance studies include inconsistent definitions of failure and variable frequencies of dropout, each of which may affect estimated risks of the event.

In this article, we consider the problem of comparing risks of cancer progression across multiple cohorts with different surveillance frequencies, where inconsistent definitions of progression were involved and variable frequencies of dropout were observed. We present a statistical framework based on multistate models (Andersen and Keiding, 2002) that considers a standardized definition of cancer progression as an event that occurs on a continuous time scale and accounts for dependent censoring due to the variable dropout.

Our application focuses on the setting of prostate cancer grade progression among patients on active surveillance. In active surveillance studies, patients with low-risk prostate cancer do not undergo active treatment at the time of diagnosis but rather are assigned to a schedule of regular biopsies and PSA measurements to monitor disease progression. Patients are generally referred to treatment if progression is detected, but they may also initiate treatment at any time for other reasons, including rising PSA, fatigue with serial biopsies, or anxiety about forgoing treatment (Penson, 2012; Dall'Era, 2015).

Active surveillance is the preferred approach for managing newly diagnosed, low-risk prostate cancer (Tosoian et al., 2016). However, there have been no randomized trials comparing prostate cancer mortality or other long-term outcomes under different active surveillance protocols. At present, information about active surveillance outcomes is based on prospective cohorts with limited follow-up. Among cohorts with the longest follow-up, reported risks of disease progression have been highly variable. However, it is unclear whether differences in the reported risks are due to underlying differences in participant selection or to differences in active surveillance implementation, including surveillance schedules, definitions of progression, and rates of dropout to treatment without progression. Clarifying whether risks of progression observed across cohorts persist after accounting for differences in implementation will provide valuable information about the representativeness of individual cohorts and about uncertainty of expected outcomes for newly diagnosed prostate cancer patients who are considering active surveillance.

Most active surveillance cohorts define prostate cancer progression as an increase in tumor grade or volume on biopsy. Grade refers to the degree of cellular differentiation in the tumor, i.e., the degree to which cancer cells resemble ordinary prostate cells. Prostate cancer grade is quantified by Gleason score, which reflects the degree of differentiation of the majority of tissue and the rest of the tissue in the tumor (Humphrey, 2004). Gleason score is an established predictor of progression risk in treated and untreated prostate cancer patients (Popiolek et al., 2013). Low-grade cancer refers to Gleason score ≤6. In this article, we define progression on active surveillance as an increase in Gleason score because this is a common component of the definition of progression in all active surveillance cohorts, whereas an increase in tumor volume is not consistently defined. Treatment for reasons other

than tumor upgrading is considered to be a competing risk in that it prevents observing tumor upgrading in the absence of treatment.

We describe a method that accommodates differing biopsy frequencies across active surveillance cohorts and produces comparable, continuous-time projections of the risk of tumor upgrading in each cohort. In practice, the risks of upgrading and treatment are likely to be correlated: factors such as a rising PSA, which may induce a decision to initiate treatment, may also be related to the risk of progression (Ross et al., 2010). Our method accommodates this by modeling both the time to progression and the time to treatment as dependent on baseline PSA level and PSA velocity. In addition, the method accommodates as a fixed input the misclassification that occurs when biopsy grade is not an accurate reflection of the true, underlying grade (i.e., the pathological grade that would be assessed following surgery). Since biopsies only sample a limited portion of the prostate, they may be subject to misrepresentation of the true tumor biology. It has been estimated that in active surveillance cohorts, high-grade cancers are misclassified as low grade 10–50% of the time (Palisaar et al., 2012; Inoue et al., 2014; Pinsky, Parnes and Ford, 2008). Misclassification of a low-grade tumor as a high-grade tumor may occur up to 15% of the time (Inoue et al., 2014). We present results with and without misclassification of specified magnitudes.

Our approach considers serial biopsies to reflect discrete views of a continuous-time stochastic process with a discrete state space corresponding to low-grade (Gleason score ≤ 6 and high-grade (Gleason score ≥ 7) cancer while capturing early treatment as a competing risk state. The underlying disease progression is modeled as occurring according to a latent continuous-time Markov chain (CTMC), and we assume that the biopsies consist of discrete, possibly misclassified observations of the underlying process (Titman and Sharples, 2010; Lange and Minin, 2013). Latent CTMCs offer much more flexibility than standard CTMCs, which, due to their tractability, are frequently used to characterize discretely observed, multistate processes describing disease progression (Mandel, 2010; Jackson et al., 2003). Further, the likelihood of latent CTMC models is analogous to a hidden Markov model likelihood, and therefore this framework naturally is able to incorporate misclassified disease outcomes. We link risk of progression and competing treatment via baseline PSA and PSA velocity, enabling us to estimate the risk of upgrading in the absence of treatment. We use our modeling framework to analyze individual-level data from four of the largest and most prominent North American active surveillance studies.

2. Methods.

2.1. Overview.

In this section we describe the model framework that allows us to compare the four active surveillance cohorts in terms of their continuous-time underlying risks of progression. First we describe each of the active surveillance datasets. Next we detail the latent CTMC model used to describe the underlying upgrades: the competing risks structure for disease progression, the dependence of the event times on the evolving PSA, the mechanism for incorporating misclassification error in the biopsy observations and the likelihood function. Finally, we describe a hypothesis test for comparing the risk of upgrading across cohorts.

2.2. Data sources.

De-identified, individual-level datasets were obtained from four active surveillance cohorts following institutional review board approval.

1. The Johns Hopkins University (JHU) dataset (Tosoian et al., 2015) consists of 913 men enrolled during 1994–2014. This study enrolled men with very low risk prostate cancer (clinical stage T1c, PSA density ≤0.15 ng/mL, Gleason score ≤6, ≤2 positive biopsy cores, and ≤50% involvement of any biopsy core with cancer), as well as older men with low risk disease (clinical stage ≤T2a, PSA <10 ng/mL, and Gleason score ≤6). Men were tracked with PSA tests every 6 months and had annual biopsies. Treatment intervention was recommended for disease reclassification, defined as any adverse grade or volume change detected on biopsy.

2. The Canary Prostate Active Surveillance Study (PASS) dataset (Newcomb et al., 2016) consists of 1,067 men enrolled during 2008–2013. Enrollment criteria included prostate cancer with clinical stage ≤T2 disease with no previous treatment and either a 10-core biopsy within one year before enrollment or ≥2 biopsies with ≥1 in the year before enrollment. Participants were followed with PSA measurements every four months and had repeat biopsies 6–12, 24, 48, and 72 months after enrollment. Treatment intervention was recommended if there was either an increase in biopsy Gleason score or volume detected on biopsy (from <33% to ≥33% of cores positive for cancer).

3. The University of California San Francisco (UCSF) dataset (Welty et al., 2015) consists of 1,319 men enrolled during 1990–2015. Al-though eligibility criteria have evolved over time, the current criteria are PSA ≤10 ng/mL, clinical stage ≤T2, biopsy Gleason score ≤6, ≤33% positive cores, and ≤50% tumor in any single core. However, carefully selected cases who do not satisfy these criteria may be enrolled. Surveillance biopsies are recommended within the first year and every 12–24 months thereafter, sampling at least 12 biopsy cores. The primary trigger for treatment was biopsy grade reclassification.

4. The University of Toronto (UT) dataset (Klotz et al., 2015) consists of 1,104 men enrolled during 1995−2015. During 1995−1999, the study was offered to all low-risk men (Gleason score ≤6 and PSA ≤10 ng/mL) and to men >70 years of age with PSA <15 ng/mL or Gleason score ≤3+4. During 2000−2015, the study also included men with favorable intermediate-risk disease (PSA ≤20 ng/mL and/or Gleason score ≤3+4) with significant comorbidities and a life expectancy <10 years. PSA tests were performed every 3 months for 2 years and then every 6 months in stable men. A confirmatory biopsy was performed within 12 months of the initial biopsy and then every 3−4 years until age 80. Treatment intervention was recommended if there was an upgrade in histology on repeat biopsy or clinical progression between biopsies. Until 2009, intervention was also recommended on the basis of having a PSA doubling time <3 years.

Records from each dataset included diagnosis years, patient ages, clinical and pathologic information at diagnosis, dates and results of all surveillance tests, including PSA values and

biopsy results, dates and types of curative treatment, and vital status. Patients diagnosed before 1995, older than 80 years at enrollment, or with Gleason score ≥7 at diagnosis were excluded from the analysis in order to create a more homogeneous population across the four cohorts.

To enforce consistency, we defined disease progression strictly in terms of grade progression, i.e., the first point at which a Gleason score ≥7 was reached. For our analyses, we differentiate between the time of underlying upgrade (UGC) and observed upgrade, where the former refers to the unobserved time when the cancer progresses from Gleason score ≤6 to Gleason score ≥7 and the latter to the observed time when grade is reclassified on biopsy. We refer to the Gleason score ≤6 as "low grade" and Gleason score ≥7 as "high grade." We also define the event of treatment without grade reclassification as the initiation of treatment in the absence of an observed biopsy upgrade. Volume reclassification or patient choice may also trigger treatment without observed upgrading. We refer to treatment without observed upgrading as "competing treatment" since its initiation precludes observing upgrading on active surveillance. One can view this competing treatment as a form of dependent censoring that may be correlated with the time of underlying upgrade.

2.3. A model for continuous-time prostate cancer progression.

2.3.1. Model overview.—Patients enter the active surveillance cohort just following diagnosis of low-grade disease. Their data consists of a follow-up sequence of biopsy times and Gleason score results, a sequence of PSA test times and results, and the time of competing treatment, if any (Figure 1A). The end of follow-up occurs at the minimum of the time that the first biopsy detects high-grade cancer, the time that competing treatment is initiated, or the time of the last PSA test in the absence of either of these events. Let $t_{k0} = 0$ be the time of the original cancer diagnosis for patient k; $t_{k1}, \ldots, t_{kn}$ be the biopsy times; and $o_{k1}, \ldots, o_{kn}$ be the Gleason score results. Let $t_{k1}^{psa}, \ldots, t_{km}^{psa}$ be the PSA test times and $y_{k1}, \ldots, y_{km}$ be the results (PSA values on the log-scale). Let h_k be an indicator of whether the final observation time corresponds to a biopsy ($h_k = 1$ if yes and $h_k = 0$ otherwise). If $h_k = 0$, let z_k be the end of follow-up, which may be either the time of competing treatment (in which case $c_k = 1$) or the final PSA test (in which case $c_k = 0$).

These data are used to inform a model of continuous-time prostate cancer grade progression involving three components.

1. A random effects model captures log-PSA growth for subject k at time $j \in \{1, 2, \ldots, m\}$ as follows:

$$Y_{kj} = \gamma_{0k} + \gamma_{1k} t_{kj}^{psa} + \epsilon_{kj},$$

where $(\gamma_{0k}, \gamma_{1k})$ are normally distributed subject-level random effects and ϵ_{kj} is a zero-mean, normally distributed, within-subject error term.

2. A competing risks model for underlying upgrading and competing treatment based on a latent continuous-time Markov chain (CTMC), as described in

Section 2.3.2. Transition rates in the model depend on baseline age and PSA intercepts and slopes estimated in advance of fitting the competing risks model. Times of underlying upgrading and competing treatment are assumed to be conditionally independent given baseline age and PSA intercept and slope.

3. A misclassification model governing imperfect sensitivity of biopsies to detect high-grade disease, as described in Section 2.3.6.

2.3.2. A competing risks model for underlying upgrade.—In the competing risks model structure, individuals proceed from low-grade cancer to either high-grade cancer or competing treatment. We can characterize the trajectory through underlying disease states as a multistate process, $W(t)$, with state space $R = \{1, 2, 3\}$, where state 1 is the low-grade state, state 2 is the high-grade state, and state 3 is the competing treatment state. At each biopsy, $W(t)$ corresponds to the true underlying state, so that $W(t_{k0}), \ldots, W(t_{kn})$ reflect discrete snapshots at times $t_{k0}, \ldots, t_{kn}$.

As a first pass, we might specify $W(t)$ as the simplest of multistate models, a time-homogeneous CTMC with state space R. The Markov property of this model means that transitions between states at any given time depend only on the state occupied at that time and not on the history of the process before that time. Time homogeneity means that the probability of transitioning from state i to state j between times s and $s + t$ is the same as the probability of transitioning between these states at times 0 and t. We specify a CTMC $W(t)$ by a transition intensity matrix $\mathbf{\Lambda} = \{\lambda_{ij}\}$, where λ_{ij} refers to the instantaneous transition rate between state i and j, and an initial distribution $\boldsymbol{\pi}$ that specifies the probability of occupying each state at time 0.

Suppose that $\mathbf{P}(t) = \{P_{ij}(t)\}$ is the matrix representing transition probabilities between states i and j in the interval $[s, s+t]$. The transition probability matrices between states are characterized by a matrix exponential of the intensity matrix $\mathbf{\Lambda}$,

$$\mathbf{P}(t) = \exp(\mathbf{\Lambda} t).$$

In general, the density of the time transition to state k at time $s + t$, given that the process is in state i at time s, is:

$$f_{ik}(t) = \sum_j P_{ij}(t)\lambda_{jk}.$$

Transition probabilities can be computed with standard methods (Moler and Loan, 2003).

While standard CTMCs have appealing analytic tractability, they assume that sojourn times (i.e., durations spent in a state before transitioning) are exponentially distributed and that rates of transitioning between states are constant with respect to sojourn duration. In our setting, this assumption is unrealistic. Therefore, to enable more flexible sojourn time distributions, we assume that the disease process $W(t)$ is based on an underlying latent CTMC $X(t)$ with state space S, where multiple latent states in S may map to each observable

disease state in R. We do not assign biological meaning to the latent states in the model but simply use them as a tool for more flexible sojourn time distributions.

In particular, the state space is $S = \left\{1_1, \ldots, 1_{s_1}, 2, 3\right\}$, where s_1 is the number of latent low-grade states. The mapping of S into R is:

$$W(t) = 1 <=> X(t) \in \left\{1_1, \ldots, 1_{s_1}\right\}$$

$$W(t) = 2 <=> X(t) = 2$$

$$W(t) = 3 <=> X(t) = 3$$

The latent CTMC sojourn time distribution in the low-grade cancer state ($W(t) = 1$) can be thought of a as mixture of all of the possible paths out of that state, which allows for more flexibility than exponential sojourn times.

2.3.3. Model selection.—Figure 1A represents the structure of a standard continuous time Markov chain ($s_1 = 1$), and Figures 1C and 1D depict the more flexible latent structures that we consider for the prostate cancer progression model. In general the coarseness of data provides practical limits on the number of latent states it is advisable to fit, and we recommend people start with smaller models and build up as the data permits, stopping when the model estimation is numerically unstable. For this setting, we considered standard CTMCs ($s_1 = 1$) and latent CTMCs with 2 ($s_1 = 2$) and 3 low-grade states ($s_1 = 3$), which we refer to as $M1$, $M2$, and $M3$, respectively. We use the Bayesian information criterion (BIC) as a means of choosing the number of latent states given its good performance in choosing the number of components in latent models (Steele and Raftery, 2010).

2.3.4. Incorporation of patient covariates.—We incorporate individual age and PSA intercept and slope (γ_{0k}, γ_{1k}) as covariates into $\mathbf{\Lambda} = \{\lambda_{ij}\}$, the transition matrix for the latent CTMC $X(t)$ in the competing risk model. Including these covariates in the model induces a dependence between times of underlying upgrading and competing treatment. To do so, we relate lograte to a linear predictor, $\log(\lambda_{ij}^{(k)}) = \boldsymbol{\xi}_{ij}^T \mathbf{v}^{(k)}$, where $\mathbf{v}^{(k)}$ is the vector of covariates for patient k. We incorporate covariates in this way for each possible transition out of the low-grade state to the high-grade state and to the competing treatment state, i.e., for each λ_{ij} where $i \in \left\{1_1, \ldots, 1_{s_1}\right\}$ and $j \in \{2, 3\}$.

For parameter identifiability, we assume that the covariate effect for transitions from low-grade to high-grade states is the same (i.e., for each λ_{i2} where $i \in \left\{1_1, \ldots, 1_{s_1}\right\}$) and similarly for the transition from low-grade to competing treatment states (i.e., for each λ_{i3} where $i \in \left\{1_1, \ldots, 1_{s_1}\right\}$). Note that unless $W(t)$ is a standard CTMC, this specification does not necessarily imply proportional hazards (e.g., Marshall and Jones (1995)). Therefore, rather

than interpreting the covariate parameter estimates directly, it is more revealing to visualize covariate effects on the risk of underlying upgrading.

2.3.5. Projecting the cumulative distribution of underlying upgrade times in absence of competing treatment.—Assuming that underlying upgrade and competing treatment are conditionally independent given patient covariates, we can project the distribution of time to underlying upgrade in absence of treatment by eliminating competing treatment in the transition intensity matrix. This involves creating a new transition matrix $\widetilde{\Lambda} = \left\{\tilde{\lambda}_{ij}\right\}$, which is identical to $\mathbf{\Lambda}$ except that transition rates to the competing treatment state Aj3, where $i \in (1_1, \ldots, 1_{s1}\}$, are set to zero. The distribution function for underlying upgrade in the absence of competing treatment, starting from the first low-grade state, is thus provided by the row of the transition probability matrix

$$\widetilde{\mathbf{P}}(t) = \exp(\widetilde{\mathbf{\Lambda}}t)$$

corresponding to the transition between state 1_1 and state 2.

2.3.6. Biopsy misclassification.—An emission matrix $\boldsymbol{E} = (e(i, j)\}$ characterizes the relationship between the observed biopsy Gleason score and the underlying state $X(t)$ and has entries $e(i, j) = P(O_t = j \mid X(t) = i)$. If there is no biopsy misclassification, the emission matrix simply maps the state space of $W(t)$ to the state space of $X(t)$. Otherwise, the emission matrix describes the probability of the observed biopsy outcomes given the underlying states in $W(t)$, assuming that the observed biopsies are conditionally independent given the values of the underlying state at each observation time. The emission matrix is given by

$$\boldsymbol{E} = \begin{pmatrix} 1 & 2 & 3 \\ p & 1-p & 0 \\ \vdots & \vdots & \vdots \\ p & 1-p & 0 \\ 1-q & q & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{matrix} 1_1 \\ \vdots \\ 1_{s_1} \\ 2 \\ 3 \end{matrix}$$

where q is biopsy sensitivity (i.e., the probability of observing upgrading if underlying upgrading has occurred) and p is the biopsy specificity (i.e., the probability of not observing upgrading if no upgrading has occurred). In practice it is challenging to estimate biopsy misclassification jointly with the transition rate parameters in a competing risks multistate process without some observations corresponding to a gold standard (Inoue et al., 2014). To investigate the impact of biopsy misclassification, we perform analyses assuming no biopsy misclassification error and compare results when biopsy sensitivity is 60%, 75%, and 90%, and biopsy specificity is 95%, 90%, and 85%, which encompass the range presented in the literature (Inoue et al., 2014).

2.3.7. Initial state distribution.—The initial state distribution $\boldsymbol{\pi}$ with entries $\pi_j = P(X(t_0) = j)$ represents the probabilities of underlying states at diagnosis. In our dataset, the first biopsy occurs ≥6 months after diagnosis, so this initial state distribution is an

extrapolation. If we assume there is no biopsy misclassification, then biopsy Gleason score at diagnosis correctly identifies all patients with low-grade disease. Per the assumptions of the Coxian sojourn time distribution (Cumani, 1982), individuals with low grade disease initially occupy the first latent state: $\pi_{1_1} = 1$. If we allow biopsy misclassification, we assume that individuals start either in state 1_1 or in state 2, and all other initial probabilities are zero. In this case, the initial distribution is estimated from a logistic regression model that depends on patient age and PSA intercept (γ_{0k}):

$$\log\frac{\pi_2}{\pi_{1_1}} = \beta_1 + \beta_2\gamma_{0k} + \beta_3 \mathrm{age}_k .$$

2.3.8. Likelihood of the observed data.—Suppressing the individual subscripts, a patient's observed data vector is $\boldsymbol{o} = (o_1, \ldots, o_n, h, c, z)$, reflecting the observed biopsy results, an indicator for whether his final observation was a biopsy, an indicator for whether he initiated competing treatment, and the end of follow-up (see Section 2.3.1). His underlying disease states corresponding to his status at the entry into active surveillance at diagnosis and each of the follow-up biopsy times is $(x_0, \ldots, x_n)$. If there were no latent states or misclassification error, the likelihood would be a product of conditional probabilities of the observed data at each time, given the previous observed data. In a general case, where we assume a latent CTMC model, or a model with misclassification error, we need to marginalize (sum) the product of conditional probabilities across the hidden states to obtain the likelihood of the observed states. In this sense it resembles a hidden Markov model which marginalizes the joint probability of the underlying disease states $(x_0, \ldots, x_n)$ and the observed data at the corresponding times across $(x_0, \ldots, x_n)$:

$$\mathrm{P}(\boldsymbol{o}) = \sum_{x_0}\sum_{x_1}\cdots\sum_{x_n}\pi_i\prod_{i=0}^{n-1}\mathrm{P}_{x_i x_{i+1}}(t_{i+1}-t_i)\prod_{i=1}^{n}e(x_i, o_i) \times\left(\left[f_{x_n 3}(z-t_n)\right]^c\left[\sum_{j\neq 3}\mathrm{P}_{x_n j}(z-t_n)\right]^{1-c}\right)^{1-h}. \quad (1)$$

Here $f_{x_n 3}(z-t_n)$ is the density function for competing treatment at time z given underlying grade at the final biopsy, and $\sum_{j\neq 3}\mathrm{P}_{x_n j}(z-t_n)$ is the probability of not initiating competing treatment at time z.

2.3.9. Parameter estimation and software implementation.—We obtain the maximum likelihood estimates for model parameters $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\Lambda})$ using an expectation-maximization (EM) algorithm that uses quantities from the Baum-Welch algorithm for obtaining estimates for a discrete-time hidden Markov model (Baum et al., 1970). The algorithm is described in Lange and Minin (2013) and is recapped in Supplement A. The algorithm is implemented using our R package, cthmm, available at http://r-forge.r-

project.org/projects/multistate/ . To obtain variance estimates for the model parameters, we use numerical estimation of the observed Fisher information matrix using the R package "NumDeriv" (Gilbert and Varadhan, 2012).

2.4. Comparisons across cohorts.

Our primary goal is to compare rates of underlying upgrading across active surveillance cohorts. To do so, we first combine data from the cohorts and parameterize the rates of underlying upgrade and competing treatment using a dummy variable for cohort as a covariate. Given estimated cohort effects, we use Wald tests to evaluate the statistical significance of differences in rates of underlying upgrading across cohorts.

3. Results.

3.1. Cohort summary and empirical outcomes.

After exclusions, the sample consists of 699, 613, 764, and 421 patients from JHU, PASS, UCSF, and UT, respectively. Distributions of baseline characteristics, surveillance biopsy and PSA test frequencies, and outcomes across the cohorts are shown in Supplement B, Table 1. JHU had the most frequent biopsies (median of 1 per 0.9 years) and UT the least frequent biopsies (median of 1 per 3.8 years); both PASS and UCSF had a median of 1 per 1.8 years.

Figure 2 shows the empirical cumulative incidence curves of observed upgrading and competing treatments across cohorts, each derived using the other event as a competing risk. Figure 2A shows that the PASS and UCSF cohorts had the highest empirical cumulative incidence of observed upgrading, while JHU had the lowest. The 10-year empirical cumulative incidence in PASS and UCSF was 59%, in JHU was 27%, and in UT was 45%. Figure 2B shows that JHU had the highest risk of competing treatment (10-year empirical cumulative incidence is 27%), possibly due to the relatively high incidence of volume-only progression in this cohort, which accounts for about half the cases reclassified, whereas the other cohorts had similar, lower risks of competing treatment (10-year empirical cumulative incidences ranged from 9–12%).

Based on the PSA growth models, median PSA velocity was similar across the cohorts, with 1% annual increase (IQR –3%, 6%) for JHU, 4% (IQR –3%, 11%) for PASS, 4% (IQR –1%, 10%) for UCSF, and 5% (IQR –2%, 11%) for UT.

3.2. Models of continuous-time prostate cancer progression.

3.2.1. No biopsy misclassification.—Table 1 presents the results of fitting standard and latent CTMC models to each cohort. When no biopsy misclassification was assumed, PASS, UCSF, and UT all achieved the best fit (according to the BIC) under model M2 (s_1 = 2); JHU achieved the best fit under model M3 (s_1 = 3). Parameter estimates are presented in Supplement C, Table 1. Using best fitting models for each cohort, Figure 3 presents the predicted cumulative probability of continuous-time underlying upgrading in the absence of competing treatment in each cohort based on averaging individual-level results. Naive Kaplan Meier (KM) based estimates of cumulative distribution functions are included for

comparison purposes. In general, the KM curves are shifted right relative to the continuous-time predictions, consistent with the expectation that underlying upgrading precedes observed upgrading. The magnitude of this shift depends on the biopsy frequency, with more frequent biopsies yielding a smaller shift.

Comparing underlying upgrading in the absence of competing treatment across cohorts indicates that JHU has a markedly lower rate than the other three cohorts. We estimate that 33% of patients in JHU had an underlying upgrade within 10 years, but this probability is 65–73% in the other cohorts. This difference is highly significant ($p < 0.001$ from Wald test for any differences across cohorts based on the model fit to combined cohort data), as are pairwise differences between JHU and the other cohorts (all $p < 0.001$). Pairwise differences between PASS, UCSF, and UT are also significant, even after Bonferroni correction for multiple comparisons, except for the difference between PASS and UT ($p=0.31$). These observations differ considerably from the impression given by the empirical curves. In contrast, a naive cohort comparison based on a Cox regression (which ignores both the differences in surveillance frequency and dependent competing risks) suggests that UT has a significantly lower hazard of upgrading relative to PASS (hazard ratio=.61, $p < .001$). All other pairwise comparisons are also significant, except for PASS and UCSF (hazard ratio= 1.17, $p=.08$).

3.2.2. Allowing for biopsy misclassification.—We first examined how biopsy sensitivity affects estimated probabilities of underlying upgrade, assuming biopsies ere 60%, 75%, and 90% sensitive and 100% specific. The BIC-selected latent structures were the same as for the models fit assuming no misclassification error. With imperfect biopsy sensitivity, some patients may have had high-grade disease that was not detected at the time of diagnosis. We fit these models (1) assuming all patients had low-grade disease at entry and (2) allowing a non-zero probability of high-grade disease at entry. In the standard CTMC model, allowing a non-zero probability of high-grade disease at entry substantially improved the fit. However, in the latent CTMC models, allowing a non-zero probability of high-grade disease at entry did not improve fit. These results are shown in Supplement D, Tables 1–4, where the maximum likelihood values are virtually identical for the latent CTMC models with zero and non-zero probabilities of high-grade disease at entry. Moreover the MLEs for this probability are zero. As discussed in Bladt and Sorensen (2005), the CTMCs under discrete observations may have MLEs that fall on the boundary of the parameter space or may not exist if the observations are not sufficiently frequent. Thus it is likely that the latent CTMCs did not have sufficient data to estimate the probability of high-grade disease at enrollment. To understand this issue further, we plotted the estimated cumulative probability of underlying upgrading assuming a 60% biopsy sensitivity under a model with one and under a model with two or three low-grade states (Supplement E, Figure 1). The figure shows that there is little difference between the two model projections after the median time of first active surveillance biopsy in each cohort, suggesting that the latent CTMC model can explain the observed data equally well with or without a non-zero probability of high-grade disease at enrollment.

Figure 5 presents underlying upgrading incidence for selected biopsy sensitivity rates (and 100% specificity), averaged across the individuals in each cohort. In general, when biopsies

are less sensitive, the predicted underlying upgrade occurs sooner. This is because lower biopsy sensitivity implies a lower detection of the underlying condition so that the true frequency of the underlying condition must be higher than that observed. The impact of different biopsy sensitivities on the estimated probability of high-grade disease was more pronounced at earlier biopsies.The impact of biopsy sensitivity also varied across cohorts, with JHU exhibiting the smallest impact and UT the largest, consistent with the ordering of biopsy frequency across cohorts. Even assuming a relatively low biopsy sensitivity, in JHU less than 10% were estimated to have true high grade disease 1 year after they entered the cohort, in contrast to the other cohorts where 20–45% were estimated to have true high grade disease at the same time.

Figure 5 presents underlying upgrading incidence for selected biopsy specificity rates (and 100% sensitivity), averaged across the individuals in each cohort. In general, lower specificity implies a higher detection of underlying upgrading (more false positives) so that the true frequency of underlying upgrading must be lower than that observed. In JHU, under 90% and 85% biopsy specificity, the incidence of true underlying upgrade is estimated to be 0% across the follow-up period. Under a yearly schedule a 10% or higher false positive rate on each biopsy is substantial enough to account for all of the observed biopsy upgrades in this cohort.

Finally, we estimated the probability of underlying upgrade in each cohort assuming 60% biopsy sensitivity and 85% specificity (Inoue et al., 2014). Results of these analyses are shown in Figure 6 along with naive KM based estimates of biopsy upgrades. Notably, JHU is still substantially lower than the other cohorts, which are relatively concordant.

4. Discussion.

The problem of comparing results from multiple studies in which observation schedules differ has been referred to as a “Twenty first century Tower of Babel” since studies with different surveillance schema are very difficult to compare (Gignac et al., 2008). In the case of active surveillance for low-risk prostate cancer, a number of studies have reported disease progression risks for different cohorts with different definitions of progression and different biopsy schedules (Tosoian et al., 2016). To adequately compare the reported results, it is necessary to derive estimates under a consistent definition of progression and on a time scale that is the same for all studies. The latent CTMC multistate modeling approach presented here does exactly this, allowing us to characterize the risk of underlying upgrade, an outcome that means the same thing across studies. The method represents a novel alternative to a more standard meta-analysis of the empirical time to biopsy upgrade. The results are a proof-of-principle that the underlying, continuous-time patterns of biopsy upgrade across cohorts are not the same as the ones suggested by the empirical results.

Our results have implications for both clinical practice and policy development because they mandate caution when basing clinical predictions and/or recommendations on any single study. It appears that even when studies are made comparable in terms of outcomes and time scales, they do not always concur. The lower risk of underlying upgrade among JHU participants may be due to the fact that they were also selected on the basis of low PSA

density (i.e., PSA level relative to prostate volume) and tended to have low volumes of cancer in their biopsy specimens. UCSF did not have a PSA density threshold for entry and included some cases with relatively high-volume disease in terms of the percent of biopsy cores with cancer. Thus, differences in inclusion criteria not captured in the available data could explain the persistent differences across studies.

A feature of the latent CTMC modeling approach is that it accommodates biopsy misclassification. We examine the impact of selected biopsy sensitivities and specificities on our results rather than simultaneously estimating these parameters since they are not identifiable without a gold standard result (Inoue et al., 2014). Our results demonstrate that the estimated rates of underlying upgrading can depend on biopsy accuracy. We found that assuming a low biopsy sensitivity was most compatible with the conclusion that a high fraction of patients have misclassified high-grade disease at the first active surveillance biopsy, but there is little change in underlying grade over time. In contrast, if we assume a high biopsy sensitivity, we estimate a low fraction of patients with high-grade disease at the first active surveillance biopsy and a faster rate of grade progression over time. Additionally, lower specificity (i.e., a higher false positive rate), leads to lower estimated rates of grade progression. Data sets that include surgical grade results will provide a gold standard and enable estimation of sensitivity and specificity in surveillance cohorts, which should provide better information on the true rates of grade progression over time.

By decoupling the underlying process of disease progression from the observed outcomes, our model accommodates both the biological process of underlying upgrading and the biopsy misclassification that contributes to the observed risk of upgrading. This property of our model contrasts with a recent study (Coley et al., 2016) that effectively assumes non-changing underlying grade and imputes this grade on the basis of a Bayesian predictive model based on the subset of patients who went on to have surgery. Like Coley et al. (2016) our modeling approach also has the potential for use as a dynamic prediction tool; given accumulated observed data, one can calcu-late the underlying probability of having high-grade disease at a particular time and then predict state transitions beyond that time.

While our method provides novel insights, it is subject to some limitations. We assume that the events of treatment and underlying upgrade are correlated only via patient age and PSA intercept and slope. Failure to account for other factors associated with both events may lead to bias in estimates of the upgrading distribution in absence of competing treatment (Huang and Wolfe, 2002). In particular, if we are ignoring factors that are positively correlated with upgrading and entering competing treatment, then our method would underestimate the upgrading risks, since men with higher treatment rates would enter competing treatment before their upgrading event was observed. Furthermore, rather than using joint modeling we incorporate PSA via a two-step approach: first estimating random effects and then using them as covariates in our disease progression model. This method may under-represent uncertainty and lead to some bias in projections of underlying upgrade, but has been used by others with latent CTMC models (Donnelly et al., 2017). Another implicit assumption is that that individuals who undergo treatment without upgrading do not have an undetected grade progression before treatment. While this is not an issue for patients treated due to increases in tumor volume on their final biopsy, it could be relevant for cases treated between

scheduled biopsies. Failure to account for progression in these patients could lead to an underestimate of the risk of underlying upgrade. A further limitation is that we were not able to estimate the misclassification probabilities jointly with the disease upgrading rates. Estimation of misclassification rates jointly with other parameters has been successful in other uses of the latent CTMC model under panel observation, albeit with substantially lower rates of misclassification than are likely in the prostate cancer biopsy setting Titman and Sharples (2010). Despite this limitation, being able to project results under different misclassification probabilities clearly provides useful information and high-lights the importance of this information in interpreting empirical results from active surveillance studies.

Finally, we do not consider the possibility of informative observation times, which could be a feature of studies with a more relaxed biopsy protocol, such as UT. In a discretely observed multistate model, observation times are ignorable if they correspond to visits that are scheduled in advance, including those based on a patient's prior history of observed data (Gruger, Kay and Schumacher, 1991). However, if visits are patient-initiated and depend on the underlying disease status, visit times are non-ignorable. Our prior simulation work has shown that when observations are more frequent if patients are in diseased states, ignoring informative visit times will lead to overestimates of the rates of transitioning between healthy and diseased states (Lange et al., 2015). In the active surveillance setting, it is unlikely that patients are symptomatic even if they do upgrade. Adherence or lack of adherence to protocol biopsies may be related to the prior history of biopsy and PSA results, but it is unlikely to additionally depend on the underlying grade. In future work, it may be worth investigating whether allowing for informative visit times changes our conclusions about times of upgrading, perhaps using the expansion of the model we previously developed (Lange et al., 2015).

We note that others have developed methods that could be applied in the analysis of active surveillance data. Most relevantly, Mao, Lin and Zeng (2017) recently extended the non-parametric current status data methods (Hudgens, Satten and Longini, 2001) for an arbitrary sequence of examination times and time varying external covariates. In addition, Rouanet et al. (2016) provided a latent class model approach for longitudinal data and interval censored competing events. However, neither of these methods incorporates misclassification error as an input parameter.

More broadly, there are other statistical methods that can be used to model discretely observed continuous-time multistate processes with non-constant hazard functions for transitions between states. Semi-Markov multistate models represent one such approach, but options for these models are limited by the structure of the model and completeness of the observations. Data that are interval censored permit semi-Markov models that can be estimated via either parametric (Foucher et al., 2007) or non-parametric (Frydman and Szarek, 2009) means. Panel data, for which both transition times and states are unknown in the inter-observation interval, present more difficulties since calculating transition probabilities requires integrating over all possible trajectories connecting states i and j on a given time interval. In particular, semi-Markov models for panel data for a general multistate model with reversible transitions are only feasible if one assumes that some of the transitions

are Markov (Kang and Lagakos, 2007) or that the trajectories follow the most parsimonious path between observed states (Aralis, 2016). In contrast, the latent CTMC modeling approach has no trouble computing transition probabilities in a wide variety of models, given that it is based on an underlying standard CTMC model. Ultimately, the latent CTMC approach offers both flexibility in terms of sojourn times and broad applicability to a variety of settings, including but not limited to the active surveillance context.

In conclusion, our harnessing of the latent CTMC approach allows us to compare cohorts with diverse surveillance schema. The power of our approach is that it enables us to coherently use discretely observed data to investigate an underlying process that occurs in continuous time. This has widespread application in cancer research, including the study of disease recurrence following primary treatment. Using this approach, we find that perceptions of cross-cohort differences may be revealed to be artifactual, and real differences in the underlying disease progression process may be different from those assessed on the basis of empirical data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements.

This work was supported by the National Cancer Institute Award Number R01 CA183570 for the Prostate Modeling to Identify Surveillance Strategies (PROMISS) consortium (all authors), P50 CA097186 as part of the Pacific Northwest Prostate Cancer Specialized Program in Research Excellence (SPORE) (ASL), and U01 CA199338 as part of the Cancer Intervention and Surveillance Modeling Network (CISNET) (JML, RG, RE); the Canary Foundation (DWL, LFN); Genomic Health Inc. and the US Department of Defense Translational Impact Award for Prostate Cancer Award Number W81XWH-13–2-0074 (PRC, MRC, JEC); and Prostate Cancer Canada (LHK).

References.

Andersen PK and Keiding N (2002). Multi-state models for event history analysis. Statistical Methods in Medical Research 11 91–115. [PubMed: 12040698]

Aralis HJ (2016). Modeling Multistate Processes with Back Transitions: Statistical Challenges and Applications PhD thesis, UCLA.

Baum LE, Petrie T, Soules G and Weiss N (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Annals of Mathematical Statistics 41 164–171.

Bladt M and Sorensen M (2005). Statistical inference for discretely observed Markov jump processes. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67 395–410.

Coley RY, Zeger SL, Mamawala M, Pienta KJ and Carter HB (2016). Prediction of the Pathologic Gleason Score to Inform a Personalized Management Program for Prostate Cancer. European Urology 3–9.

Cumani A (1982). On the canonical representation of homogeneous Markov processes modelling failure-time distributions. Microelectronics and Reliability 22 583–602.

Dall'Era MA (2015). Patient and disease factors affecting the choice and adherence to active surveillance. Current opinion in urology 25 272–6. [PubMed: 25692724]

Donnelly C, McFetridge LM, Marshall AH and Mitchell HJ (2017). A two-stage approach to the joint analysis of longitudinal and survival data utilising the Coxian phase-type distribution. Statistical Methods in Medical Research 0 0962280217706727. PMID: . [PubMed: 28633604]

Foucher Y, Giral M, Soulillou JP and Daures JP (2007). A semi-Markov model for multistate and interval-censored data with multiple terminal events. Application in renal transplantation. Statistics in Medicine 26 5381–5393. [PubMed: 17987670]

Frydman H and Szarek M (2009). Nonparametric estimation in a Markov “illness-death” process from interval censored observations with missing intermediate transition status. Biometrics 65 143–51. [PubMed: 18505421]

Gignac GA, Morris MJ, Heller G, Schwartz LH and Scher HI (2008). Assessing outcomes in prostate cancer clinical trials: a twenty-first century tower of Babel. Cancer 113 966–974. [PubMed: 18661513]

Gilbert p. and Varadhan R (2012). numDeriv: Accurate Numerical Derivatives R package version 2012.9–1.

Gruger J, Kay R and Schumacher M (1991). The validity of inferences based on incomplete observations in disease state models. Biometrics 47 595–605. [PubMed: 1912263]

Huang X and Wolfe RA (2002). A Frailty Model for Informative Censoring. Biometrics 58 510–520. [PubMed: 12229985]

Hudgens MG, Satten GA and Longini IM (2001). Nonparametric Maximum Likelihood Estimation for Competing Risks Survival Data Subject to Interval Censoring and Truncation. Biometrics 57 74–80. [PubMed: 11252621]

Humphrey, p. A. (2004). Gleason grading and prognostic factors in carcinoma of the prostate. Modern Pathology 17 292–306. [PubMed: 14976540]

Inoue LYT, Trock BJ, Partin AW, Carter HB and Etzioni R (2014). Modeling grade progression in an active surveillance study. Statistics in medicine 33 930–939. [PubMed: 24123208]

Jackson CH, Sharples LD, Thompson SG and Duffy SW (2003). Multi-state Markov models for disease progression with classification error. The Statistician 52 193–209.

Kang M and Lagakos SW (2007). Statistical methods for panel data from a semi-Markov process, with application to HPV. Biostatistics 8 252–64. [PubMed: 16740624]

Klotz L, Vesprini D, Sethukavalan P, Jethava V, Zhang L, Jain S, Yamamoto T, Mamedov A and Loblaw A (2015). Long-term follow-up of a large active surveillance cohort of patients with prostate cancer. Journal of Clinical Oncology 33 272–277. [PubMed: 25512465]

Lange JM and Minin VN (2013). Fitting and interpreting continuous-time latent Markov models for panel data. Statistics in Medicine 32 4581–95. [PubMed: 23740756]

Lange JM, Hubbard RA, Inoue LYT and Minin VN (2015). A joint model for multistate disease processes and random informative observation times, with applications to electronic medical records data. Biometrics 71 90–101. [PubMed: 25319319]

Mandel M (2010). Estimating disease progression using panel data. Biostatistics 11 304–16. [PubMed: 20064845]

Mao L, Lin D-Y and Zeng D (2017). Semiparametric regression analysis of interval-censored competing risks data. Biometrics n/a-n/a.

Marshall G and Jones RH (1995). Multi-state models and diabetic retinopathy. Statistics in Medicine 14 1975–1983. [PubMed: 8677398]

Moler C and Loan CV (2003). Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. SIAM Review 45 801–836.

Narod SA and Rakovitch E (2014). A comparison of the risks of in-breast recurrence after a diagnosis of DCIS or early invasive breast cancer. Current Oncology 21 119–124. [PubMed: 24940092]

Newcomb LF, Jr IMT, Boyer HD, Brooks JD, Carroll PR, Cooper-berg MR, Dash A, Ellis WJ, Fazli L, Feng Z, Martin E, Kunju P, Lance RS, Mckenney JK, Meng MV, Marlo M, Sanda MG, Simko J, So A, Tretiakova MS, Troyer D. a., True LD, Vakar-lopez F, Virgin J, Wagner A. a., Wei JT, Nelson PS, Lin DW, Prostate C and Surveillance A (2016). Outcomes of active surveillance for the management of clinically localized prostate cancer in the prospective, multi-institutional Canary PASS cohort. Journal of Urology 195 206–221. [PubMed: 26259991]

Palisaar JR, Noldus J, Löppenberg B, Von Bodman C, Sommerer F and Eggert T (2012). Comprehensive report on prostate cancer misclassification by 16 currently used low-risk and active surveillance criteria. BJU International 110.

Penson DF (2012). Factors influencing patients' acceptance and adherence to active surveillance. Journal of the National Cancer Institute - Monographs 45 207–212.

Pinsky P, Parnes H and Ford L (2008). Estimating rates of true high-grade disease in the Prostate Cancer Prevention Trial. Cancer Prevention Research 1 182–186. [PubMed: 19138954]

Popiolek M, Rider JR, AndrÉn O, Andersson S-O, Holmberg L, Adami H-O and Johansson J-E (2013). Natural History of Early, Localized Prostate Cancer: A Final Report from Three Decades of Follow-up. European Urology 63 428–435. [PubMed: 23084329]

Ross AE, Loeb S, Landis P, Partin AW, Epstein JI, Kettermann A, Feng Z, Carter HB and Walsh PC (2010). Prostate-specific antigen kinetics during follow-up are an unreliable trigger for intervention in a prostate cancer surveillance program. Journal of Clinical Oncology 28 2810–2816. [PubMed: 20439642]

Rouanet A, Joly P, Dartigues J-F, Proust-Lima C and Jacqmin-Gadda H (2016). Joint latent class model for longitudinal data and interval-censored semi- competing events: Application to dementia. Biometrics 72 1123–1135. [PubMed: 27123856]

Sridhara R, Mandrekar SJ and Dodd LE (2013). Missing data and measurement variability in assessing progression-free survival endpoint in randomized clinical trials. Clinical Cancer Research 19 2613–2620. [PubMed: 23669421]

Steele RJ and Raftery AE (2010). Performance of Bayesian Model Selection Criteria for Gaussian Mixture Models In Frontiers of Statistical Decision Making and, Bayesian Analysis (Chen M-H, Muller P, Sun D, Ye K and Dey DK, eds.) 113–130. Springer.

Stephenson AJ, Kattan MW, Eastham JA, Dotan ZA, Bianco FJ, Lilja H and Scardino PT (2006). Defining biochemical recurrence of prostate cancer after radical prostatectomy: A proposal for a standardized definition. Journal of Clinical Oncology 24 3973–3978. [PubMed: 16921049]

TITMAN AC and SHARPLES LD (2010). Semi-Markov models with phase-type sojourn distributions. Biometrics 66 742–752. [PubMed: 19912172]

Tosoian JJ, Mamawala M, Epstein JI, Landis P, Wolf S, Trock BJ and CARTER HB (2015). Intermediate and longer-term outcomes from a prospective active-surveillance program for favorable-risk prostate cancer. Journal of Clinical Oncology 33 3379–3385. [PubMed: 26324359]

Tosoian JJ, Carter HB, Lepor A and Loeb S (2016). Active surveillance for prostate cancer: contemporary state of practice. Nature Reviews Urology 116 1477–1490.

Welty CJ, Cowan JE, Nguyen H, Shinohara K, Perez N, Greene KL, Chan JM, Meng MV, Simko JP, Cooperberg MR and Carroll PR (2015). Extended followup and risk factors for disease reclassification in a large active surveillance cohort for localized prostate cancer. Journal of Urology 193 807–811. [PubMed: 25261803]

Zeng L, Cook RJ, Wen L and Boruvka A (2015). Bias in progression-free survival analysis due to intermittent assessment of progression. Statistics in Medicine 34 3181–3193. sim.6529. [PubMed: 26011411]

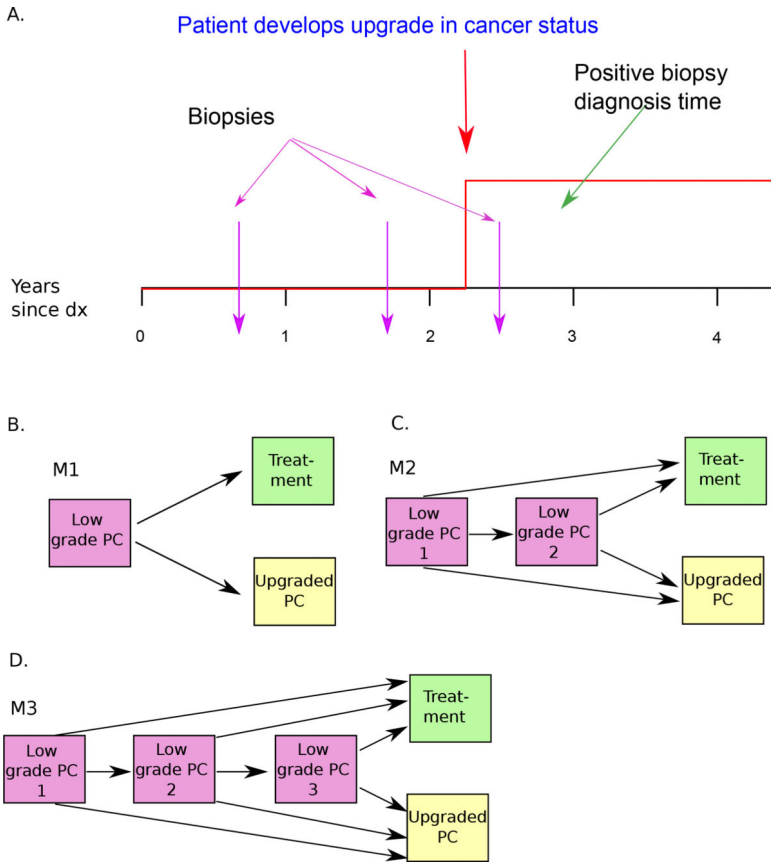

Fig 1.

A. Sample trajectory for a patient who upgrades during the observation period. The biopsies detect the underlying process at discrete times, and the time of diagnosis occurs after the time of underlying upgrade. B. Competing risks structure of the model describing cancer grade progression or competing treatment modeled as a standard CTMC, denoted M1. C. Latent continuous time Markov chain (CTMC) structure with two low-grade cancer states that generates non-constant hazard rates for transitions, denoted M2. D. Latent CTMC structure with three low-grade cancer states, denoted M3.

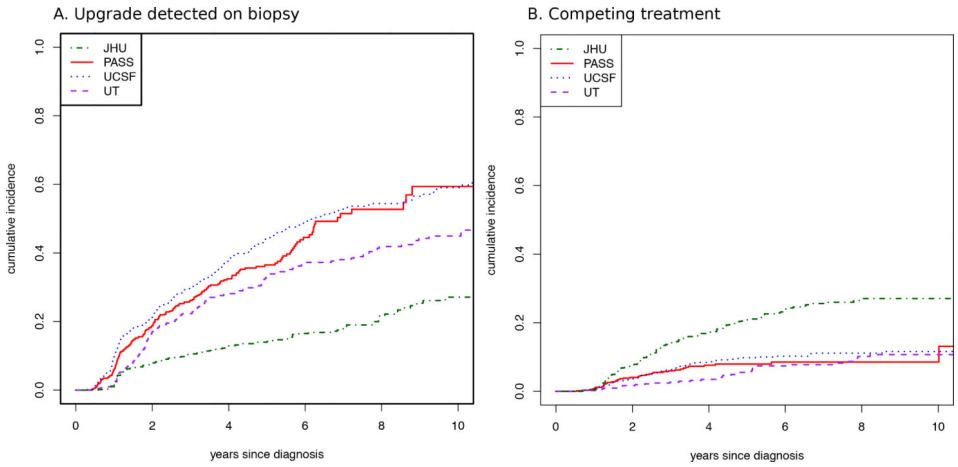

Fig 2.

A. Empirical cumulative incidence of observed upgrading in the presence of competing treatment. B. Empirical cumulative incidence of competing treatment in the presence of observed upgrading.

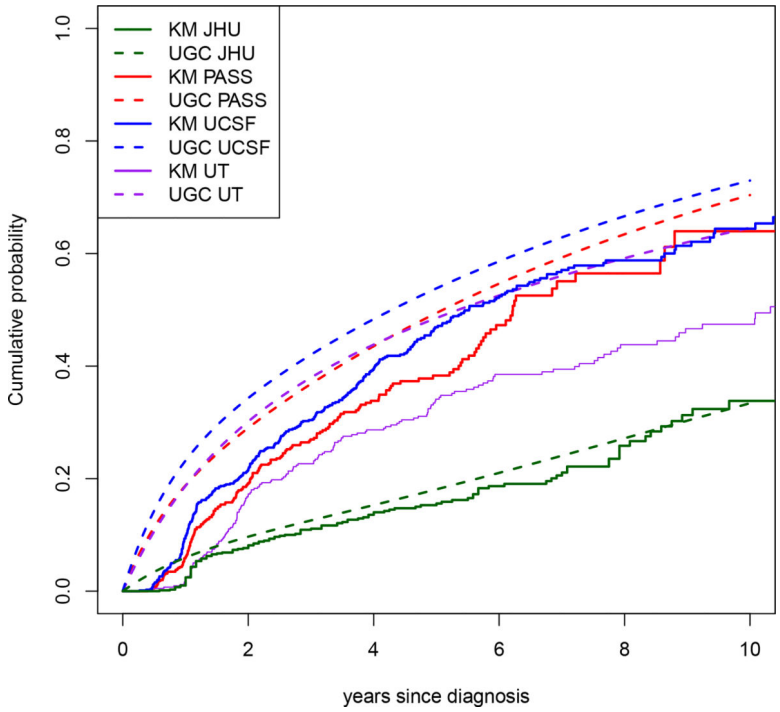

Fig 3.

Comparison of observed upgrading (solid Kaplan-Meier curves) and underlying upgrading (dashed continuous-time model predictions) in the absence of competing treatments assuming no biopsy misclassification.

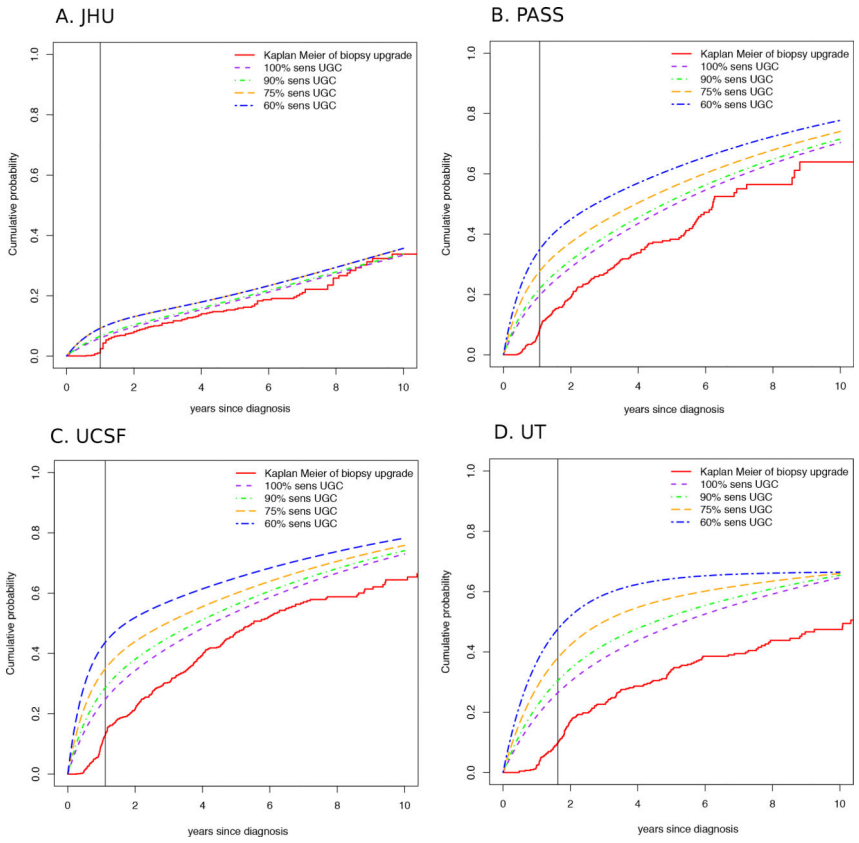

Fig 4.

Underlying upgrading under varying fixed levels of biopsy sensitivity, assuming 100% specificity, averaged across individuals in the cohort. Vertical lines show the median time of the first active surveillance biopsy in each cohort.

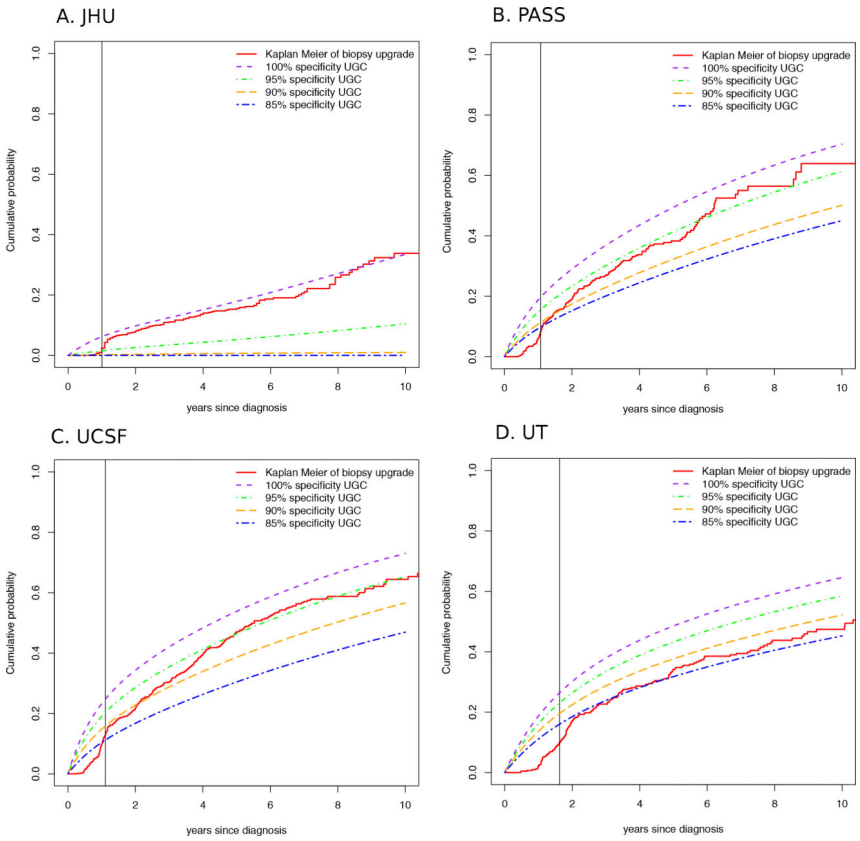

Fig 5.

Underlying upgrading under varying fixed levels of biopsy specificity, assuming 100% sensitivity, averaged across individuals in the cohort. Vertical lines show the median time of the first active surveillance biopsy in each cohort.

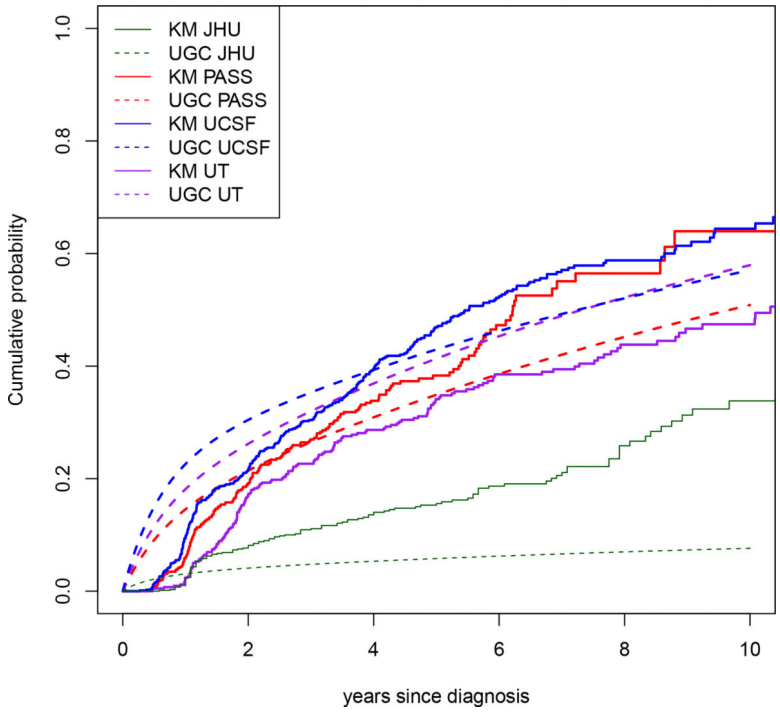

Fig 6.

Comparison of observed upgrading (solid Kaplan-Meier curves) and underlying upgrading (dashed continuous-time model predictions) in the absence of competing treatments assuming biopsies are 60% sensitive and 85% specific at detecting upgraded disease.

Table 1

Model selection assuming no biopsy misclassification. Abbreviations: BIC=Bayesian Information Criterion; M1, M2, M3 refer to standard CTMC and models with 2 and 3 latent states, respectively.

Cohort	No. patients	Latent CTMC model	No. params	Log likelihood	BIC
JHU	699	M1	8	−1086.4	2195.6
		M2	11	−1069.6	2170.5
		M3	14	−1055.8	2151.4
PASS	613	M1	8	−688.1	1398.5
		M2	11	−677.1	1384.9
		M3	14	−674.8	1388.6
UCSF	764	M1	8	−1149.4	2321.8
		M2	11	−1122.0	2275.7
		M3	14	−1122.0	2284.4
UT	421	M1	8	−495.3	1011.6
		M2	11	−480.4	989.7
		M3	14	−479.2	995.1