Topics on survival analysis

Konan Hara

University of Arizona

December 27, 2021

Outline

1. Introduction

2. Definition

3. Statistical Methods

4. Related Topics

5. Applications

Outline

1. Introduction

2. Definition

3. Statistical Methods

4. Related Topics

5. Applications

Introduction[1]

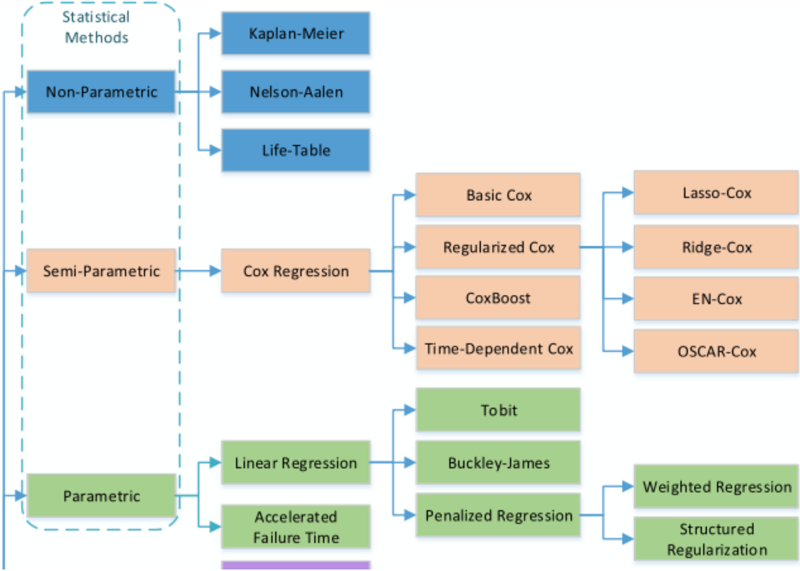
- The goal of survival analysis is to analyze and model data where the outcome is the time until an event of interest occurs, T.
- T can always be transformed into an occurrence of the event within a specified period.
- The advantages and disadvantages of using either survival analysis methods or other rudimentary regression methods should be weighed.

[1] The following contents are primarily based on Wang, Li, and Reddy (2019, ACM Computing Surveys).

Pros and Cons

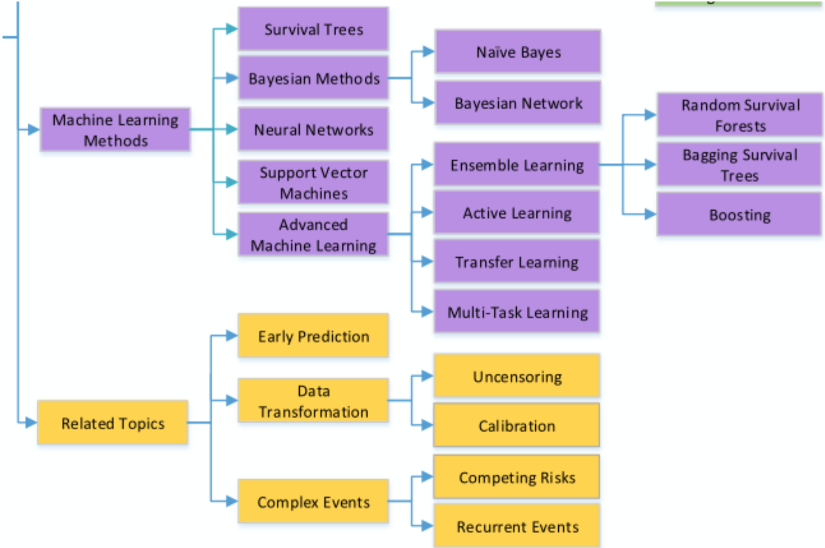
- Survival analysis can deal with
  - Censoring
  - Time-dependent covariates
  - Competing risks
- However, the costs are that
  - Assumptions for causal inference are strong
  - Models are complicated and not so intuitive
  - Machine learning methods are difficult to apply

Topics



Statistical Methods
Non-Parametric
Kaplan-Meier
Nelson-Aalen
Life-Table
Semi-Parametric
Cox Regression
Basic Cox
Regularized Cox
CoxBoost
Time-Dependent Cox
Lasso-Cox
Ridge-Cox
EN-Cox
OSCAR-Cox
Parametric
Linear Regression
Accelerated Failure Time
Tobit
Buckley-James
Penalized Regression
Weighted Regression
Structured Regularization

Topics


Survival Trees
Bayesian Methods
Naïve Bayes
Bayesian Network
Machine Learning Methods
Neural Networks
Support Vector Machines
Advanced Machine Learning
Ensemble Learning
Active Learning
Transfer Learning
Multi-Task Learning
Random Survival Forests
Bagging Survival Trees
Boosting
Early Prediction
Data Transformation
Uncensoring
Calibration
Related Topics
Complex Events
Competing Risks
Recurrent Events

Outline

1. Introduction

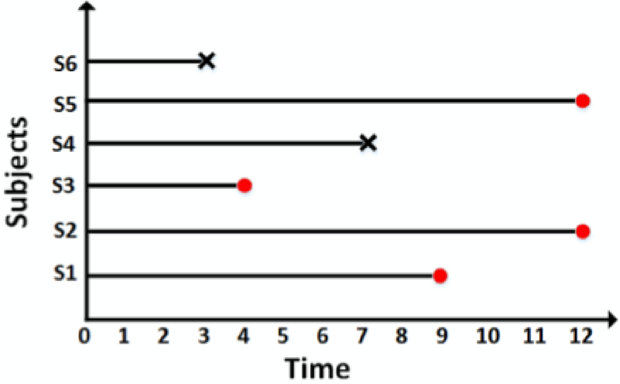
2. Definition

3. Statistical Methods

4. Related Topics

5. Applications

Survival Data and Censoring



X marks: event occurrence

Red dots: censored

Problem Statement

- Data: a triplet (X_i, y_i, δ_i)
  - $i \in \{1 \dots N\}$: individual
  - $X_i \in \mathbb{R}^p$: independent variables
  - δ_i: 0 if censored and 1 if otherwise
  - y_i: T_i if $\delta_i = 1$ and C_i if otherwise
  - T_i: survival time
  - C_i: censored time
- Goal: estimate the effect of some elements of X on T or predict T with X

Survival and Hazard Function

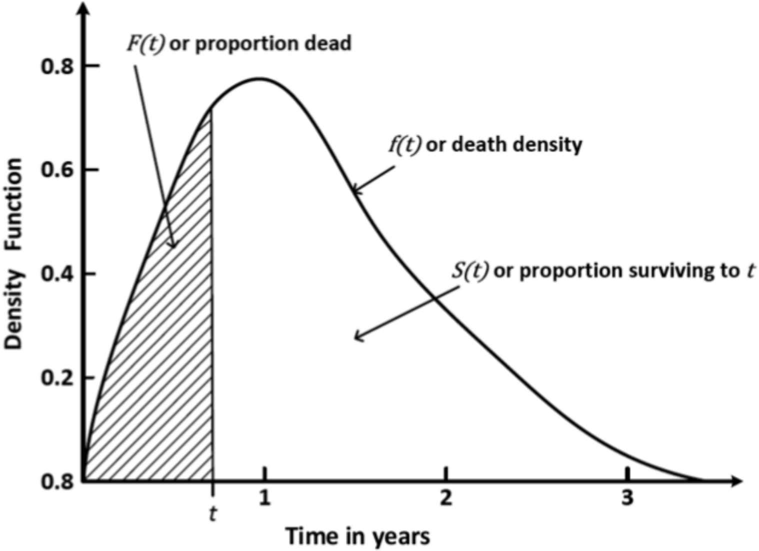
- Survival function: $S(t) = \Pr(T \geq t)$
- Cumulative distribution function: $F(t) = 1 - S(t)$
- Density function:

$$f(t) = \frac{d}{dt}F(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t)}{\Delta t}$$

- Hazard function:

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \frac{f(t)}{S(t)}$$

Survival and Hazard Function



Identification

- Consider a simple nonparametric estimator of $f(t)$:

$$\begin{aligned} \hat{f}(t) &= \frac{1}{N}\sum_{i=1}^{N} I(t \leq T_i < t + \Delta t) \\ &\rightarrow \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t, \delta = 1)}{\Delta t} \end{aligned}$$

- Unfortunately,

$$\Pr(t \leq T < t + \Delta t, \delta = 1) < \Pr(t \leq T < t + \Delta t)$$

unless $\delta_i = 1$ for all i, no censoring.

Identification

- Thus, we cannot get a consistent estimator for $f(t)$.
- How about $h(t)$?
- $R(t)$: set of individuals considered to be “at risk” at time t
- $r(t)$: number of individuals considered to be “at risk” at time t, $|R(t)|$

Identification

Then, a simple nonparametric estimator of $h(t)$ is

$$\begin{aligned} & \hat{h}(t) \\ = & \frac{1}{r(t)} \sum_{i \in R(t)} I(t \leq T_i < t + \Delta t) \\ \rightarrow & \lim_{\Delta t \rightarrow 0} \frac{\Pr(t \leq T < t + \Delta t, \delta = 1 | T \geq t, \delta = 1)}{\Delta t} \end{aligned}$$

What is the condition to be

$$\begin{aligned} & \Pr(t \leq T < t + \Delta t, \delta = 1 | T \geq t, \delta = 1) \\ = & \Pr(t \leq T < t + \Delta t | T \geq t)? \end{aligned}$$

Identification

If $T \perp\!\!\!\perp \delta$, then

$$\begin{aligned}
& \Pr(t \leq T < t + \Delta t, \delta = 1 | T \geq t, \delta = 1) \\
= \; & \frac{\Pr(t \leq T < t + \Delta t, \delta = 1)}{\Pr(T \geq t, \delta = 1)} \\
= \; & \frac{\Pr(t \leq T < t + \Delta t) \Pr(\delta = 1)}{\Pr(T \geq t) \Pr(\delta = 1)} \\
= \; & \Pr(t \leq T < t + \Delta t | T \geq t).
\end{aligned}$$

Thus, we can get a consistent estimator for $h(t)$.

From $h(t)$ to $f(t)$

As

$$\begin{aligned} & h(t) = \frac{f(t)}{1 - F(t)} = -\frac{d \log\{1 - F(t)\}}{dt} \\ \Leftrightarrow \quad & F(t) = 1 - \exp\left\{-\int_0^t h(\tau) d\tau\right\} \\ \Leftrightarrow \quad & f(t) = h(t) \exp\left\{-\int_0^t h(\tau) d\tau\right\}, \end{aligned}$$

we can calculate $f(t)$ from $h(t)$.

Outline

1. Introduction

2. Definition

3. Statistical Methods

4. Related Topics

5. Applications

Nonparametric Models: Kaplan-Meier Curve

Let event times be $T_1 < T_2 < \cdots < T_K$ for N individuals and consider a specific event time T_j:

- $d(T_j)$: observed events
- $c(T_j)$: censored individuals between T_j and T_{j+1}
- then, $r(T_j) = r(T_{j-1}) - d(T_{j-1}) - c(T_{j-1})$

Kaplan-Meier Curve

- Conditional probability of surviving beyond T_j:

$$p(T_j) = \frac{r(T_j) - d(T_j)}{r(T_j)}$$

- The estimator of $S(t) = \Pr(T \geq t)$ is given as

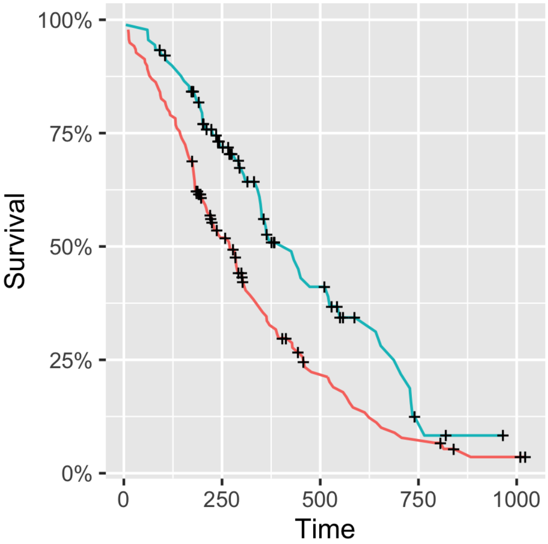
$$\hat{S}(t) = \prod_{j:T_j<t} p(T_j)$$

- Nonparametric: no functional form specification

Kaplan-Meier Curve—Covariates

- The effect of X can be quantified by stratification.
- Time-dependent covariates $X(t)$ can be incorporated as well.
- Set of individuals considered to be “at risk” at time t will be $R(t, X(t))$.

Kaplan-Meier Curve



Semiparametric Models: Cox Model

- Semiparametric models can give more efficient estimates than nonparametric models
- Hazard function $h(t, X_i)$ follows the proportional hazards assumption:

$$h(t, X_i) = h_0(t) \exp(X_i \beta)$$

- Proportional hazards:

$$\frac{h(t, X_i)}{h(t, X_j)} = \frac{h_0(t) \exp(X_i \beta)}{h_0(t) \exp(X_j \beta)} = \frac{\exp(X_i \beta)}{\exp(X_j \beta)}$$

Cox Model

- Semiparametric: baseline hazard function $h_0(t)$ can be left unspecified in the estimation of β
- $h_0(t)$ can be estimated nonparametrically
- Likelihood of event time T_i:

$$\frac{h(T_i, X_i)\Delta t}{\sum_{j \in R(T_i)} h(T_i, X_j)\Delta t}$$

- Maximum partial likelihood estimator:

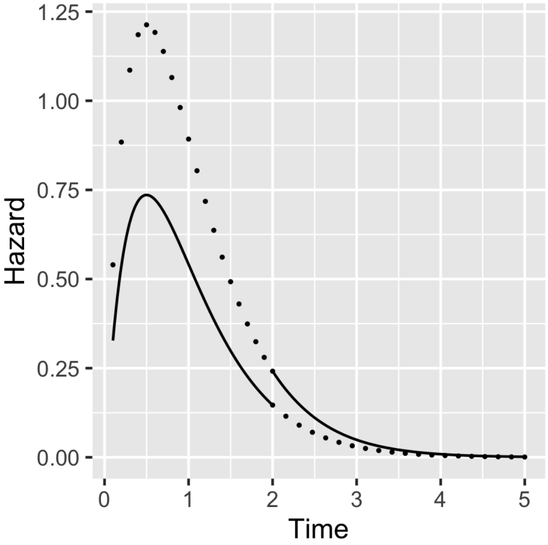
$$\hat{\beta} = \arg\max_{\beta} \prod_{i=1}^{N} \left\{ \frac{\exp(X_i \beta)}{\sum_{j \in R_i} \exp(X_j \beta)} \right\}^{\delta_i}$$

Time-Dependent Cox Model

- Cox model can handle time-dependent covariates
- Hazard function:

$$h(t, X_i(t)) = h_0(t)\exp(X_i(t)\beta)$$

Time-Dependent Cox Model



Parametric Models: Parametric Hazard Function

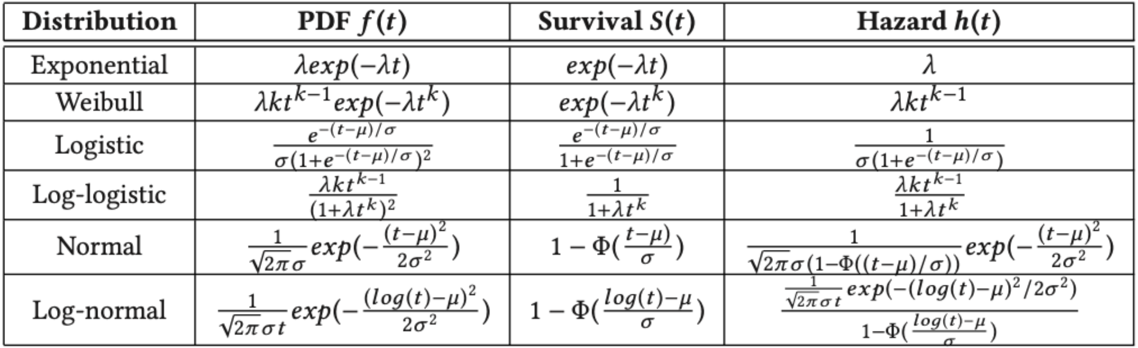
- Parametric baseline hazard function, i.e, $h_0(t;\lambda)$
- Hazard function can be estimated more efficiently than semiparametric models
- Hazard function:

$$h(t, X_i) = h_0(t;\lambda)\exp(X_i\beta)$$

- Likelihood:

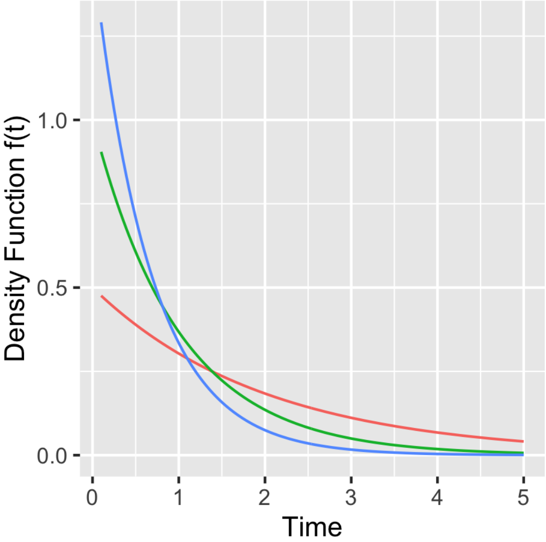
$$\prod_{i:\delta_i=1} f(T_i;\beta,\lambda) \prod_{i:\delta_i=0} S(C_i;\beta,\lambda)$$

Parametric Hazard Function



Distribution	PDF $f(t)$	Survival $S(t)$	Hazard $h(t)$
Exponential	$\lambda exp(-\lambda t)$	$exp(-\lambda t)$	λ
Weibull	$\lambda k t^{k-1} exp(-\lambda t^k)$	$exp(-\lambda t^k)$	$\lambda k t^{k-1}$
Logistic	$\frac{e^{-(t-\mu)/\sigma}}{\sigma(1+e^{-(t-\mu)/\sigma})^2}$	$\frac{e^{-(t-\mu)/\sigma}}{1+e^{-(t-\mu)/\sigma}}$	$\frac{1}{\sigma(1+e^{-(t-\mu)/\sigma})}$
Log-logistic	$\frac{\lambda k t^{k-1}}{(1+\lambda t^k)^2}$	$\frac{1}{1+\lambda t^k}$	$\frac{\lambda k t^{k-1}}{1+\lambda t^k}$
Normal	$\frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(t-\mu)^2}{2\sigma^2})$	$1-\Phi(\frac{t-\mu)}{\sigma})$	$\frac{1}{\sqrt{2\pi}\sigma(1-\Phi((t-\mu)/\sigma))} exp(-\frac{(t-\mu)^2}{2\sigma^2})$
Log-normal	$\frac{1}{\sqrt{2\pi}\sigma t} exp(-\frac{(log(t)-\mu)^2}{2\sigma^2})$	$1-\Phi(\frac{log(t)-\mu}{\sigma})$	$\frac{\frac{1}{\sqrt{2\pi}\sigma t} exp(-(log(t)-\mu)^2/2\sigma^2)}{1-\Phi(\frac{log(t)-\mu}{\sigma})}$

Exponential distribution



Parametric Models: Accelerated Failure Time Model

- Event time is directly parameterized:

$$\log(T_i) = X_i\beta + \sigma\epsilon$$

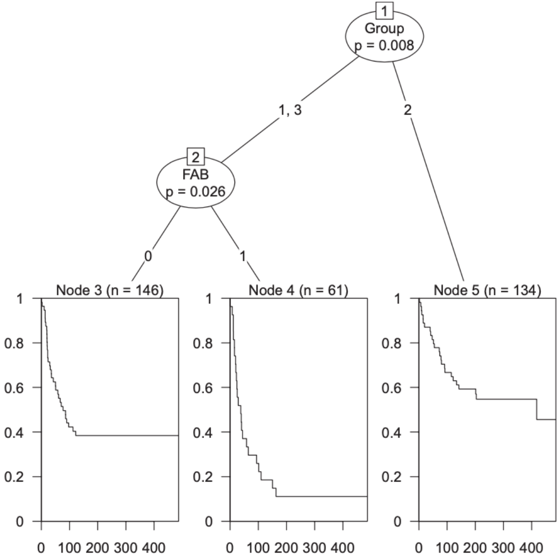
- Likelihood:

$$\prod_{i:\delta_i=1} f(T_i;\beta,\sigma) \prod_{i:\delta_i=0} S(C_i;\beta,\sigma)$$

Machine Learning Methods

- Available methods:
  - Regularized Cox models
  - Survival trees
  - Bayesian methods
  - Neural networks
  - Support vector machines
- Machine learning methods primarily aim to build prediction models
- Causal inference framework is still in its infancy

Survival Trees



Outline

1. Introduction

2. Definition

3. Statistical Methods

4. Related Topics

5. Applications

Performance Measures: C-index

- Compare the rankings of observed and predicted survival times
- Concordance probability:

$$c = \Pr(\hat{y}_i > \hat{y}_j | y_i > y_j)$$

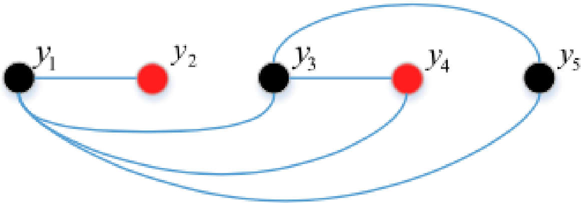
- Estimator of C-index for Cox models:

$$\hat{c} = \frac{1}{num} \sum_{i:\delta_i=1} \sum_{j:y_i<y_j} I(X_i\hat{\beta} > X_j\hat{\beta})$$

  num: number of all comparable pairs
- If y_i is binary, C-index = AUC

C-index

$$y_1 < y_2 < y_3 < y_4 < y_5$$



Black: event observed
Red: censored
Edges: possible ranking comparisons

Performance Measures: Brier Score

- Compare the observed and predicted event occurrence before time t
- $z_i(t)$: indicator of event before t
- $\hat{z}_i(t)$: predicted probability of event before t
- Brier score at t:

$$BS(t) = \frac{1}{N} \sum_{i=1}^{N} \{\hat{z}_i(t) - z_i(t)\}$$

- Censored information can be incorporated

Competing Risks

- $D_i \in \{1 \dots K\}$: i's cause of event
- Cause-specific density:

$$f_k(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t, D = k)}{\Delta t}$$

- Cumulative incidence function (CIF):

$$I_k(t) = \int_0^t f_k(\tau) d\tau = \Pr\{T \leq t, D = k\}$$

- Can we estimate the effect of X on CIF?

Competing Risks

- Consider cause-specific hazard function:

$$h_k^{cs}(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t, D = k | T \geq t)}{\Delta t}$$

- $h_k^{cs}(t)$ can be estimated using the Cox model:

$$h_k^{cs}(t, X_i) = h_{k0}^{cs}(t) \exp(X_i \beta_k^{cs})$$

- Effect of X on CIF cannot be inferred from β_k^{cs}

Competing Risks

- Alternative is subdistribution hazard function[2]:

$$h_k^{sd}(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t, D = k | A)}{\Delta t},$$

  where $A = \{T \geq t$ or $(T < t, D \neq k)\}$
- $h_k^{sd}(t)$ can be estimated using the Cox model:

$$h_k^{sd}(t, X_i) = h_{k0}^{sd}(t) \exp(X_i \beta_k^{sd})$$

[2] Fine and Gray (1999, JASA)

Competing Risks

- Effect of X on CIF can be inferred from β_k^{sd} as

$$\begin{aligned} & \log(-\log(1 - I_k(t, X))) \\ = \ & \log(-\log(1 - I_{k0}(t))) + X\beta_k^{sd} \end{aligned}$$

- When the probability of an event is low, then the logistic link function and the complementary log-log link function are very similar.
- Thus, if this is the case, β_k^{sd} can be interpreted as odds ratios for the CIF.

Competing Risks with Time-Dependent Covariates[3]

- Individuals with time-dependent covariates can be incorporated in a competing risk model.
- Split their observation periods into periods where the covariates do not vary with time.
- Consider left truncation as well as right censoring in the estimation.

[3] Geskus (2011, Biometrics)

Competing Risks with Time-Dependent Covariates

- "Weights" of individuals who experienced competing events need to be adjusted for left truncation and right censoring.
- Each i has weight $w_i(T_j)$ at T_j given by:

$$w_i(T_j) = \begin{cases} 1 & \text{if "at risk" at } T_j \\ \frac{\Pr(C \geq T_j)\Pr(L < T_j)}{\Pr(C \geq T_j')\Pr(L < T_j')} & \text{if } B \\ 0 & \text{otherwise,} \end{cases}$$

where L is left entry time and $B = \{i \text{ had competing events at } T_j' < T_j\}$.

Competing Risks with Time-Dependent Covariates: Estimation Procedure

- Split each observation period into periods where the covariates do not vary with time.
  - Some rows will have left entry time > 0.
- Attach weights to the observations.
  - Individuals who experienced competing events will have many time-varying weights.
- Apply survival analysis program that can handle left truncation in addition to right censoring.

Multi-State Models: Continuous-Time Markov Model[4]

- Survival analysis models usually deal with transition from one state to another.
- In some cases, modeling transitions between multiple states may be valuable.
- Continuous-time Markov model can be used for this purpose.

[4] Kalbfleisch and Lawless (1985, JASA)

Continuous-Time Markov Model

- Continuous-time Markov model can be viewed as an extension of a parametric Cox model.
- $S_i(t) \in \{1 \ldots R\}$: i's state at time t

Continuous-Time Markov Model

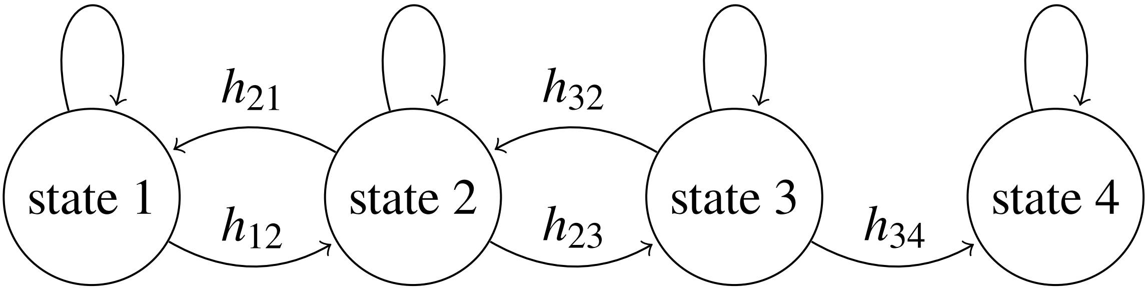
- Hazard (instantaneous risk) of moving from state r to state r' is assumued to be:

$$h_{rr'}(t) = \lim_{\Delta t \to 0} \frac{\Pr(S(t+\Delta t) = r | S(t) = r')}{\Delta t}$$

- $h_{rr'}(t)$ is estimated using the same parameterization as the parametric Cox model:

$$h_{rr'}(t, X_i(t)) = h_{rr'0}(t; \lambda_{rr'}) \exp(X_i(t)\beta_{rr'})$$

Continuous-Time Markov Model



Outline

1. Introduction

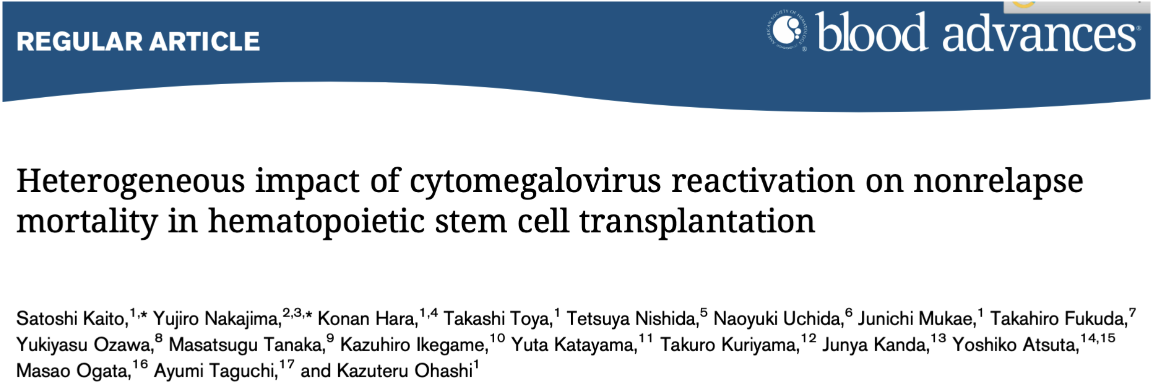
2. Definition

3. Statistical Methods

4. Related Topics

5. Applications

Kaito et al. (2020, Blood Adv.)



REGULAR ARTICLE

blood advances

Heterogeneous impact of cytomegalovirus reactivation on nonrelapse mortality in hematopoietic stem cell transplantation

Satoshi Kaito,[1,*] Yujiro Nakajima,[2,3,*] Konan Hara,[1,4] Takashi Toya,[1] Tetsuya Nishida,[5] Naoyuki Uchida,[6] Junichi Mukae,[1] Takahiro Fukuda,[7] Yukiyasu Ozawa,[8] Masatsugu Tanaka,[9] Kazuhiro Ikegame,[10] Yuta Katayama,[11] Takuro Kuriyama,[12] Junya Kanda,[13] Yoshiko Atsuta,[14,15] Masao Ogata,[16] Ayumi Taguchi,[17] and Kazuteru Ohashi[1]

Introduction

- Investigated the heterogeneous impact of CMV reactivation on NRM in hematopoietic stem cell transplantation.
- Used time-dependent Cox model considering competing risks.
- Heterogeneous impact was quantified by interaction terms.

Heterogeneous Treatment Effect

- Treatment effects may vary across the levels of baseline characteristics.
- Subgroup analysis is one way to investigate heterogeneous treatment effects (HTEs).
- However, subgroups can be very small if there are a lot of baseline characteristics.
- Statistical tests of interactions between the treatment and baseline characteristics can be a solution.

Heterogeneous Treatment Effect

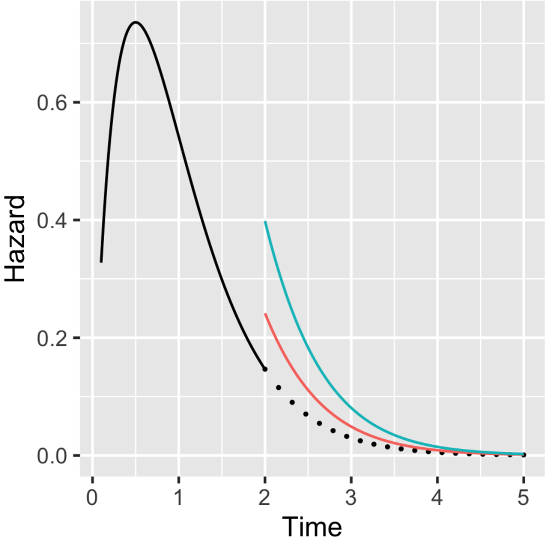
- D: treatment
- X: baseline characteristics
- Statistical tests of interactions:

$$y = X\beta + D * X\gamma$$

- Significant γ indicates the HTE.
- In this study,

$$\text{NRM} = X\beta + \text{CMV} * X\gamma.$$

Heterogeneous Treatment Effect



Scoring model for CMV reactivation

- Cumulative incidence of CMV reactivation was evaluated considering relapse and NRM as competing risks.
- Scoring model for CMV reactivation was developed and assessed by landmark analysis at day 100.

Scoring model for CMV reactivation

Significant factors (HR > 1):

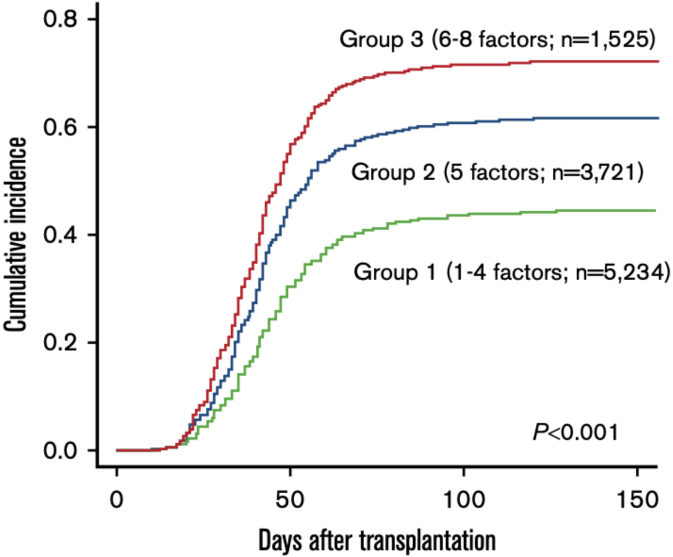
- Recipient positive/donor negative CMV serology
- Recipient positive/donor positive CMV serology
- TCD *in vivo*
- HLA disparity
- $\geq$ 50 years
- Transplant from an unrelated donor
- TBI
- Older transplant year

Scoring model for CMV reactivation

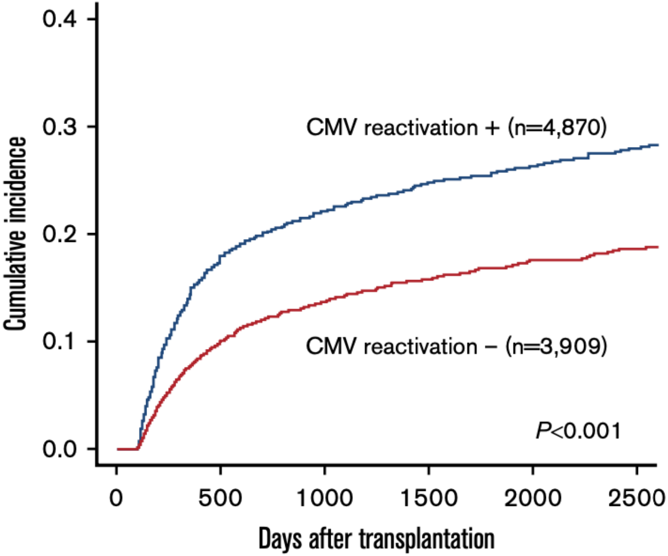
Significant factors ($\text{HR} < 1$):

- Tacrolimus-based GVHD prophylaxis regimen
- CB

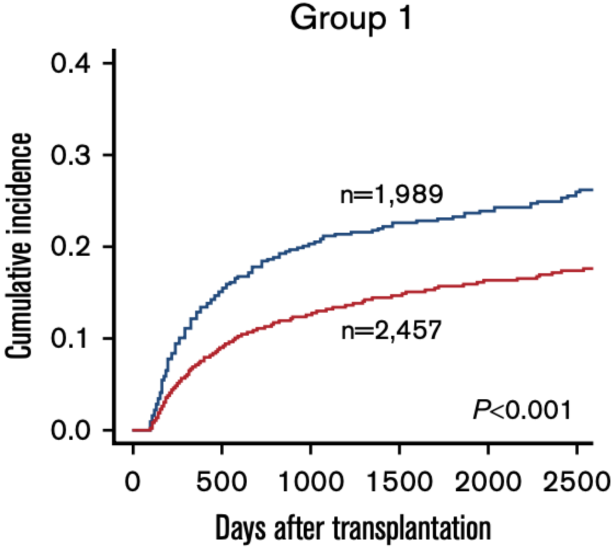
Scoring model for CMV reactivation



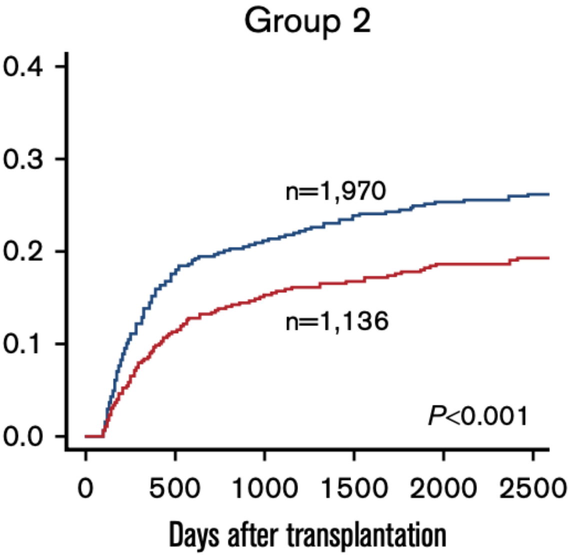
Impact of CMV reactivation on NRM



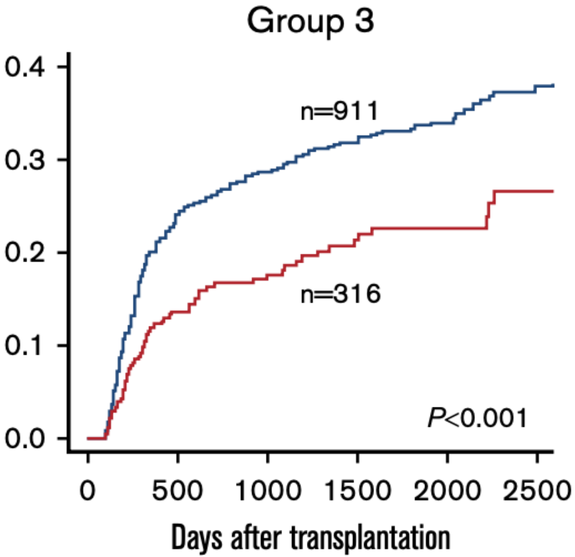
Impact of CMV reactivation on NRM



Impact of CMV reactivation on NRM



Impact of CMV reactivation on NRM



Heterogeneous impact

- Cumulative incidence of NRM was evaluated considering relapse as a competing risk and CMV reactivation as a time-dependent covariate.
- Interaction terms between CMV reactivation and baseline characteristics were included in the model.
- Scoring model for the heterogeneous impact of CMV reactivation on NRM was developed and assessed by landmark analysis at day 100.

Heterogeneous impact

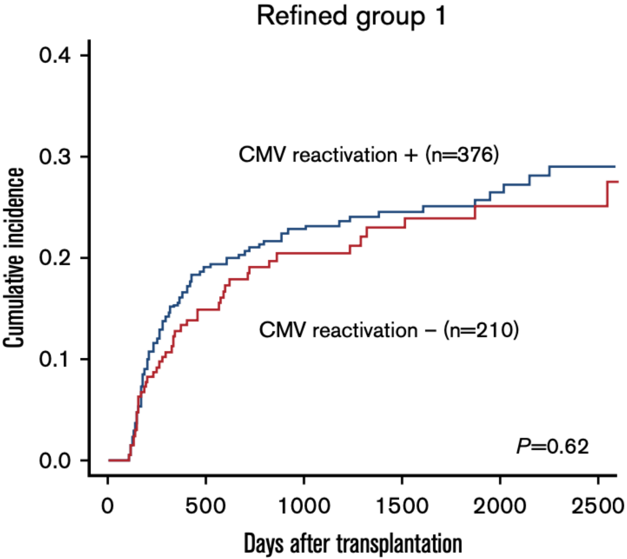
Significant factors ($\mathrm{HR} > 1$):

- CML

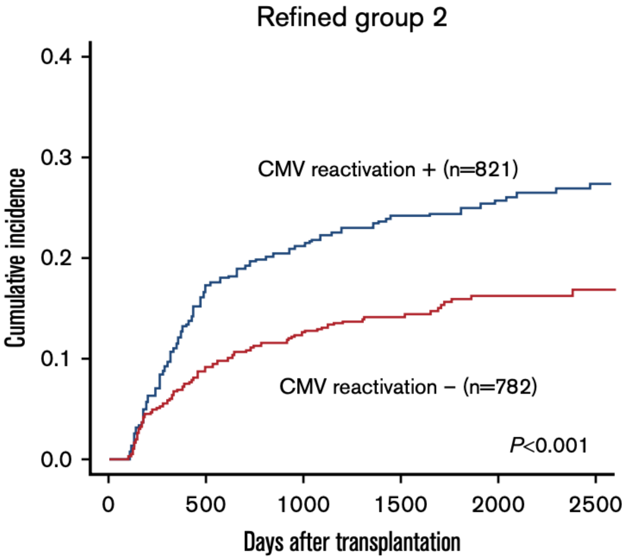
Significant factors ($\mathrm{HR} < 1$):

- Poor PS
- Transplantation from HLA-mismatched donors
- High disease risk

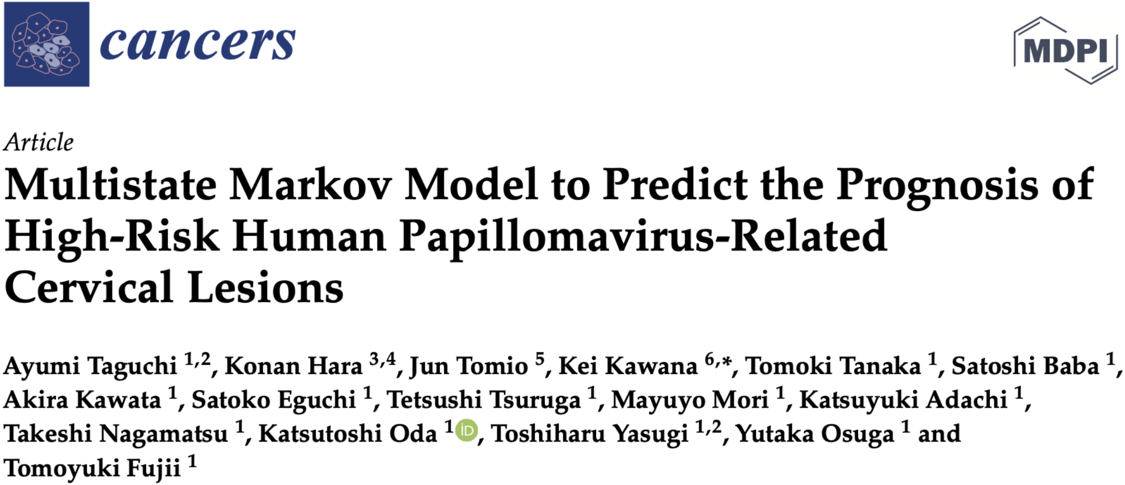
Heterogeneous impact



Heterogeneous impact



Taguchi et al. (2020, Cancers)



cancers

MDPI

Article

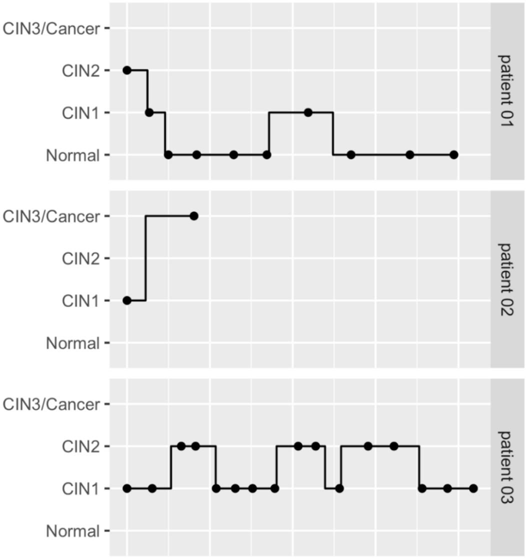
Multistate Markov Model to Predict the Prognosis of High-Risk Human Papillomavirus-Related Cervical Lesions

Ayumi Taguchi [1,2], Konan Hara [3,4], Jun Tomio [5], Kei Kawana [6,*], Tomoki Tanaka [1], Satoshi Baba [1], Akira Kawata [1], Satoko Eguchi [1], Tetsushi Tsuruga [1], Mayuyo Mori [1], Katsuyuki Adachi [1], Takeshi Nagamatsu [1], Katsutoshi Oda [1], Toshiharu Yasugi [1,2], Yutaka Osuga [1] and Tomoyuki Fujii [1]

Markov model for HPV-related CIN prognosis

- Investigated the prognosis of hrHPV-related cervical lesions.
- Cox proportional hazards model is most frequently used to predict the risk of transition from one state to another.
- CIN has a natural history of bidirectional transition between different states.
- Cox models assuming a unidirectional disease progression oversimplify CIN fate.
- Application of continuous-time Markov model to this situation may be of interest.

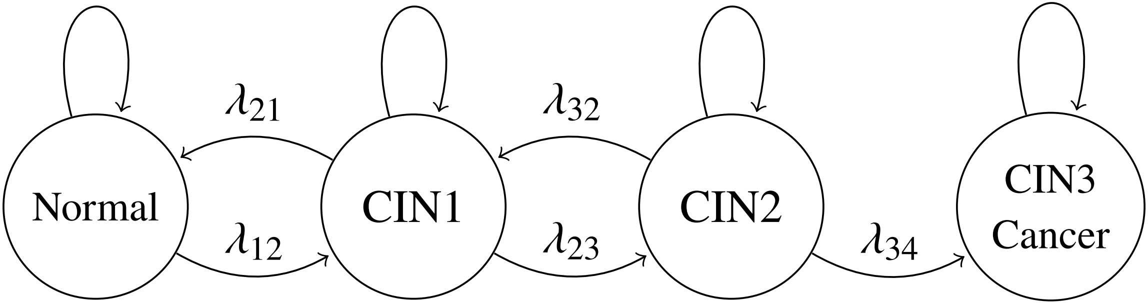
Samples


CIN3/Cancer
CIN2
CIN1
Normal
patient 01
CIN3/Cancer
CIN2
CIN1
Normal
patient 02
CIN3/Cancer
CIN2
CIN1
Normal
patient 03

Summary of transitions

		Diagnosis at t^{th} Visit				
Diagnosis at $(t\text{-}1)^{th}$ Visit	HPV Category	Normal	CIN1	CIN2	CIN3	Cancer
Normal	HPV 16	206 (84.4)	13 (5.3)	21 (8.6)	4 (1.6)	0 (0.0)
	HPV 18	89 (81.6)	12 (11.0)	8 (7.3)	0 (0.0)	0 (0.0)
	HPV 52	277 (75.8)	54 (14.7)	32 (8.7)	2 (0.5)	0 (0.0)
	HPV 58	230 (80.4)	33 (11.5)	21 (7.3)	2 (0.6)	0 (0.0)
	Other hrHPVs	611 (86.1)	72 (10.1)	23 (3.2)	3 (0.4)	0 (0.0)
	No hrHPVs	1289 (90.2)	109 (7.6)	26 (1.8)	3 (0.2)	1 (0.0)
CIN1	HPV 16	29 (28.9)	34 (34.0)	35 (35.0)	2 (2.0)	0 (0.0)
	HPV 18	18 (38.2)	19 (40.4)	8 (17.0)	2 (4.2)	0 (0.0)
	HPV 52	80 (35.0)	90 (39.4)	53 (23.2)	5 (2.1)	0 (0.0)
	HPV 58	51 (31.6)	68 (42.2)	40 (24.8)	2 (1.2)	0 (0.0)
	Other hrHPVs	132 (40.7)	143 (44.1)	45 (13.8)	4 (1.2)	0 (0.0)
	No hrHPVs	203 (54.5)	132 (35.4)	34 (9.1)	3 (0.8)	0 (0.0)
CIN2	HPV 16	31 (12.1)	37 (14.4)	147 (57.4)	40 (15.6)	1 (0.3)
	HPV 18	10 (12.9)	8 (10.3)	51 (66.2)	8 (10.3)	0 (0.0)
	HPV 52	41 (13.8)	53 (17.9)	168 (56.9)	33 (11.1)	0 (0.0)
	HPV 58	32 (10.2)	45 (14.4)	210 (67.5)	24 (7.7)	0 (0.0)
	Other hrHPVs	49 (16.7)	52 (17.8)	166 (56.8)	25 (8.5)	0 (0.0)
	No hrHPVs	58 (27.2)	31 (14.5)	114 (53.5)	10 (4.6)	0 (0.0)

Continuous-time Markov model



Continuous-time Markov Model

- Parameterization of the hazard function:

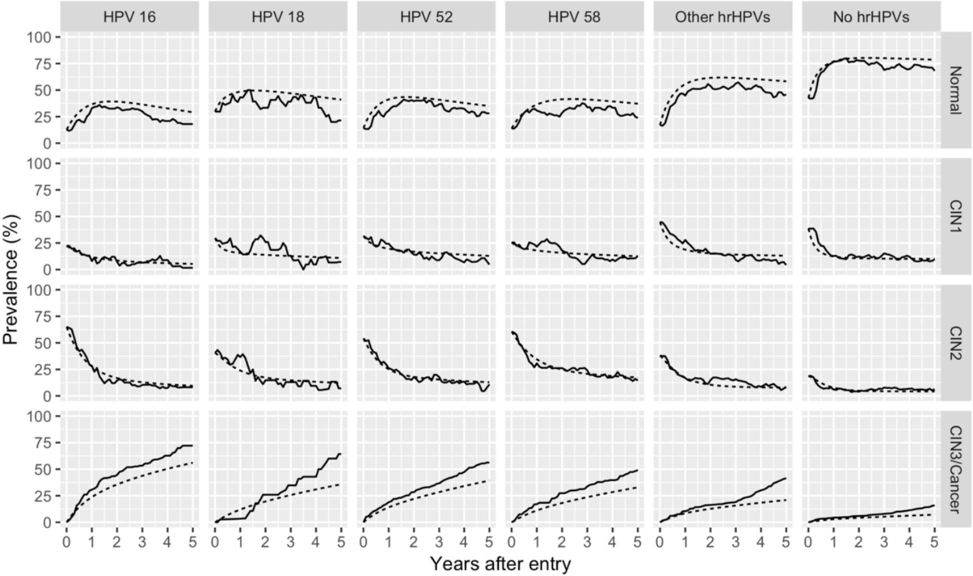
$$\lambda_{rr'}(t, \mathsf{HPV}) = \lambda_{rr'0}^{\mathsf{HPV}}$$

  - Hazards were assumed to be constant across time.
  - Hazards were estimated for each HPV category.
- It was challenging to find a parameterization where the estimation converges.

Predicted 2-year transition probabilities

Current State	HPV Category	State after 2 Years			
		Normal	CIN1	CIN2	CIN3/Cancer
Normal	HPV 16	0.598 (0.506–0.684)	0.099 (0.074–0.128)	0.169 (0.127–0.215)	0.132 (0.090–0.183)
	HPV 18	0.610 (0.479–0.719)	0.156 (0.109–0.215)	0.156 (0.093–0.230)	0.076 (0.033–0.148)
	HPV 52	0.533 (0.474–0.593)	0.189 (0.162–0.219)	0.180 (0.146–0.216)	0.096 (0.070–0.130)
	HPV 58	0.559 (0.484–0.627)	0.171 (0.140–0.205)	0.206 (0.162–0.255)	0.062 (0.041–0.089)
	Other hrHPVs	0.723 (0.676–0.766)	0.155 (0.132–0.182)	0.085 (0.066–0.108)	0.034 (0.023–0.050)
	No hrHPVs	0.838 (0.814–0.861)	0.105 (0.090–0.121)	0.042 (0.032–0.054)	0.012 (0.007–0.020)
CIN1	HPV 16	0.434 (0.349–0.512)	0.089 (0.067–0.115)	0.175 (0.134–0.223)	0.300 (0.225–0.378)
	HPV 18	0.535 (0.396–0.652)	0.146 (0.100–0.207)	0.172 (0.102–0.257)	0.146 (0.069–0.267)
	HPV 52	0.473 (0.413–0.529)	0.178 (0.152–0.208)	0.183 (0.150–0.221)	0.164 (0.122–0.219)
	HPV 58	0.469 (0.399–0.535)	0.165 (0.135–0.197)	0.239 (0.192–0.291)	0.126 (0.084–0.181)
	Other hrHPVs	0.656 (0.606–0.702)	0.156 (0.133–0.181)	0.102 (0.079–0.128)	0.084 (0.058–0.119)
	No hrHPVs	0.808 (0.781–0.835)	0.107 (0.091–0.123)	0.049 (0.038–0.065)	0.034 (0.021–0.054)
CIN2	HPV 16	0.335 (0.266–0.404)	0.079 (0.059–0.101)	0.165 (0.121–0.218)	0.418 (0.330–0.512)
	HPV 18	0.373 (0.245–0.501)	0.119 (0.074–0.178)	0.186 (0.099–0.302)	0.320 (0.178–0.507)
	HPV 52	0.381 (0.324–0.434)	0.156 (0.129–0.184)	0.175 (0.138–0.216)	0.286 (0.220–0.367)
	HPV 58	0.356 (0.291–0.419)	0.150 (0.122–0.181)	0.260 (0.209–0.319)	0.232 (0.167–0.307)
	Other hrHPVs	0.518 (0.453–0.571)	0.146 (0.122–0.169)	0.117 (0.089–0.148)	0.218 (0.159–0.299)
	No hrHPVs	0.706 (0.643–0.749)	0.106 (0.090–0.123)	0.063 (0.045–0.089)	0.124 (0.079–0.191)

Observed and simulated prevalence



Ikesu et al. (2021, Cancer Medicine)

DOI: 10.1002/cam4.4470

RESEARCH ARTICLE

Cancer Medicine Open Access WILEY

Prognosis of high-risk human papillomavirus-related cervical lesions: A hidden Markov model analysis of a single-center cohort in Japan

Ryo Ikesu[1] | Ayumi Taguchi[2] | Konan Hara[1,3,4] | Kei Kawana[5] |
Tetsushi Tsuruga[2] | Jun Tomio[1] | Yutaka Osuga[2]

Misclassification concern in CIN diagnosis

- A continuous-time multistate Markov model was also applied to accommodate the bidirectional feature of CIN lesions.
- However, another concern when building structural models of CIN lesion prognosis is diagnostic misclassification.
- Although CIN diagnosis is based on the combination of cytological and histological examinations aided by colposcopy, the accuracy of CIN diagnosis is limited, resulting in the misclassification of the "true" pathology of the lesion.
- Nevertheless, few studies have evaluated HPV pathogenesis, accounting for the probability of diagnostic misclassification.

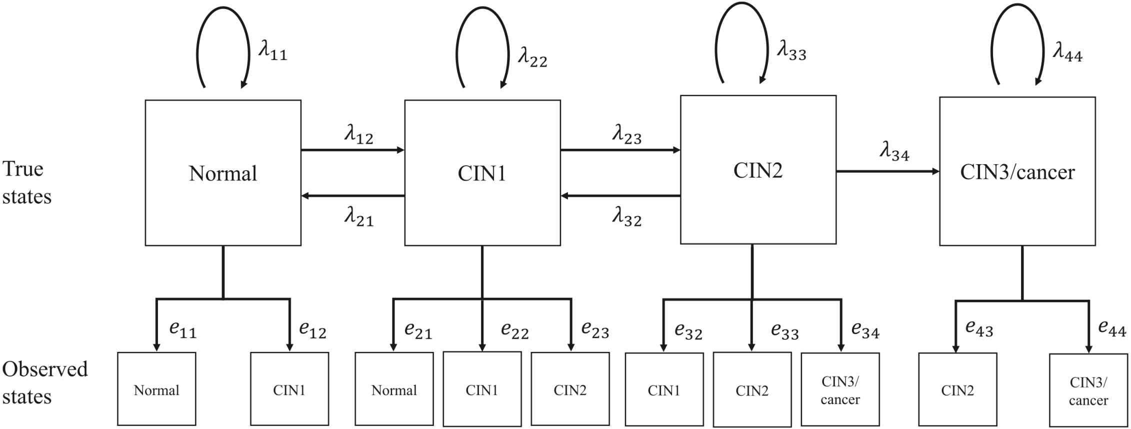
Hidden Markov model

- To accommodate these types of measurement challenges, various latent variable models (e.g., factor models and structural equation models) have been adopted in medical research.
- Latent variable models can manage unobserved random variables.
- Recently, another latent variable model, a hidden Markov model, was applied to model
  1. the transition between the (unobserved) “true” states
  2. the probabilities of the “observed” state conditional on the “true” states (misclassification probabilities).
- Hidden Markov models have been applied in clinical settings with measurement challenges, such as frailty, HIV infection, and diabetic retinopathy.

Hidden Markov model for HPV-related CIN prognosis

- We applied a hidden Markov model to our cohort of HPV-infected patients to clarify the natural course of CIN according to HPV genotype, which accounted for the misclassification probability.
- We aimed to confirm the robustness of the current literature, including a previous study that used a Markov model, on the CIN characteristics according to HPV genotype.
- We also quantified the misclassification probability in CIN diagnosis.

Continuous-time hidden Markov model

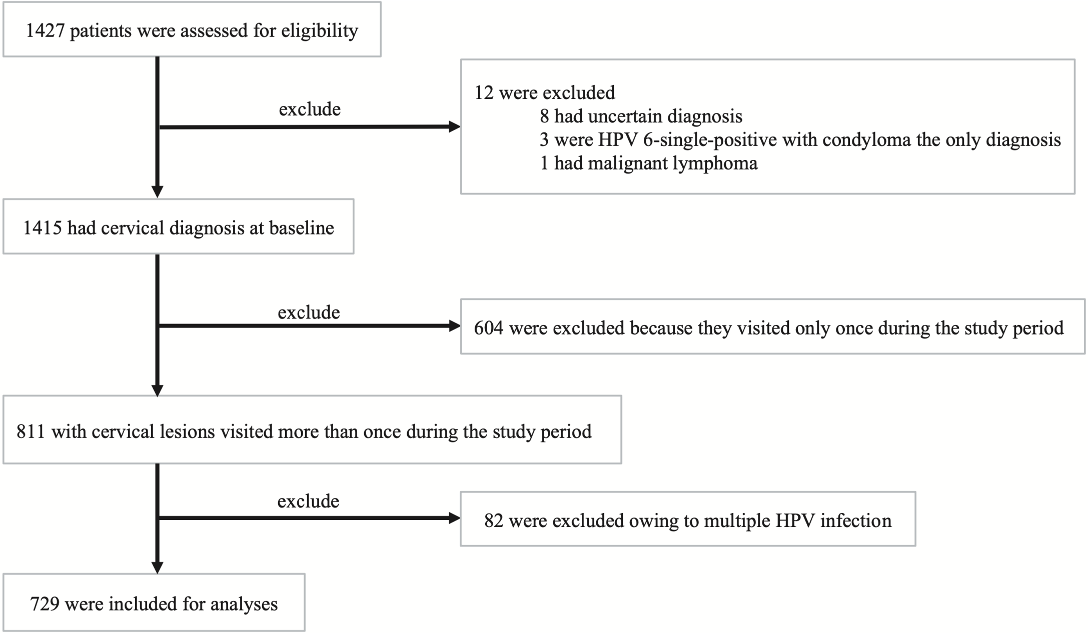


Continuous-time hidden Markov model

- Parameterization of the hazard function:

$$\lambda_{rr'}(t, \mathsf{HPV}) = \lambda_{rr'0}^{\mathsf{HPV}}$$

  - Hazards were assumed to be constant across time.
  - Dummy variables representing each HPV genotype were included in the model as covariates.
- Based on the clinical assumption that adjacent misclassifications could account for most diagnostic misclassifications, we allowed for a "one-step" misclassification adjacent to the true state.
- "True" initial distribution was also estimated in the models.
- It was challenging to find a parameterization where the estimation converges.



1427 patients were assessed for eligibility

exclude

12 were excluded

- 8 had uncertain diagnosis
- 3 were HPV 6-single-positive with condyloma the only diagnosis
- 1 had malignant lymphoma

1415 had cervical diagnosis at baseline

exclude

604 were excluded because they visited only once during the study period

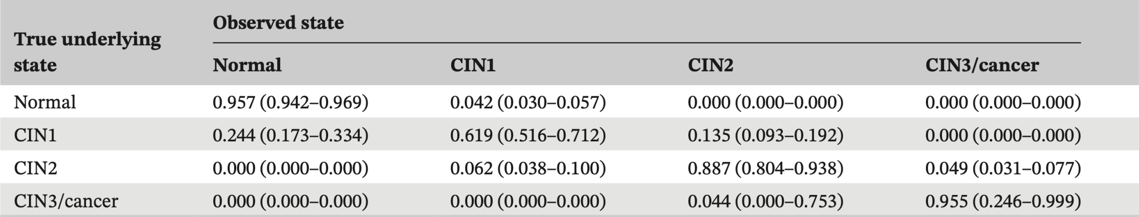
811 with cervical lesions visited more than once during the study period

exclude

82 were excluded owing to multiple HPV infection

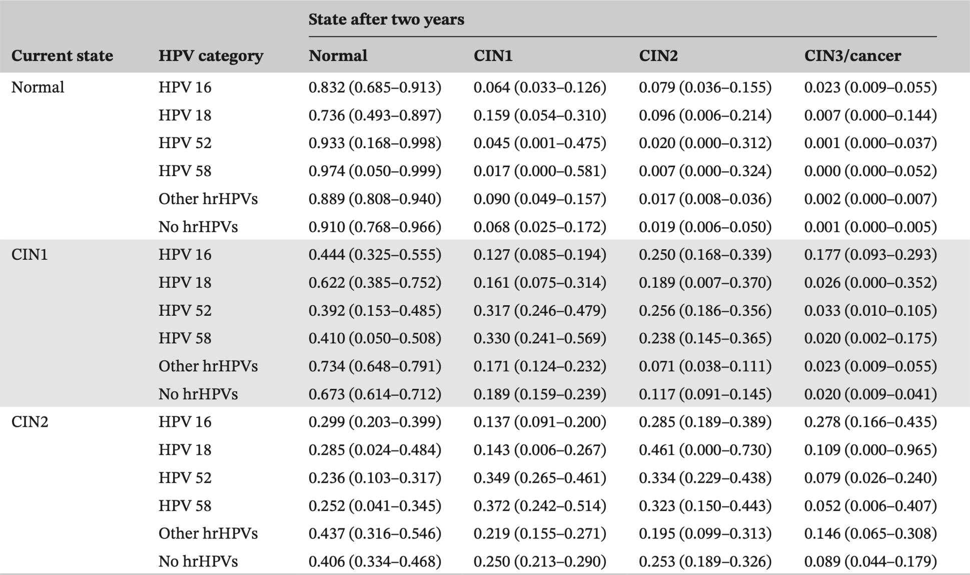
729 were included for analyses

Misclassification probabilities



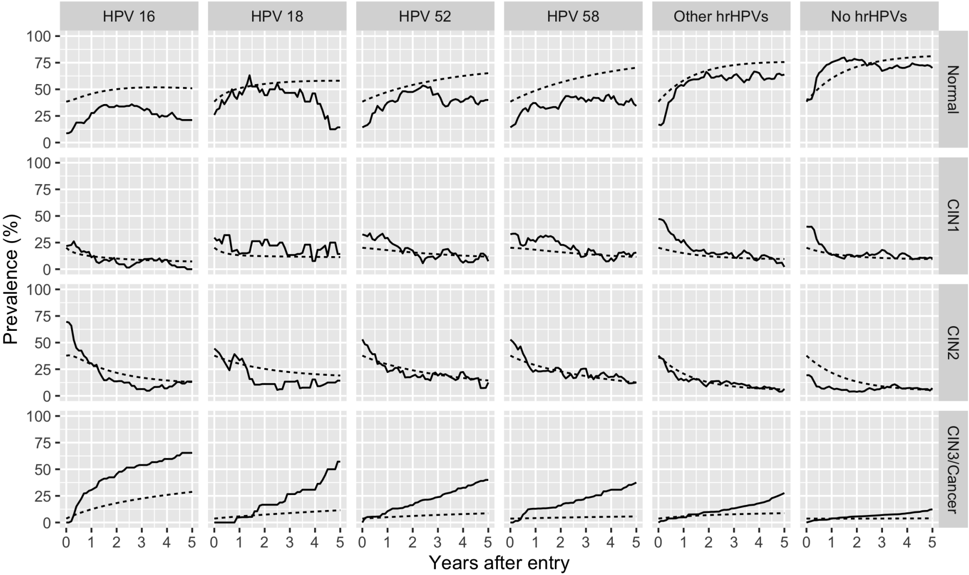
True underlying state	Observed state			
	Normal	CIN1	CIN2	CIN3/cancer
Normal	0.957 (0.942–0.969)	0.042 (0.030–0.057)	0.000 (0.000–0.000)	0.000 (0.000–0.000)
CIN1	0.244 (0.173–0.334)	0.619 (0.516–0.712)	0.135 (0.093–0.192)	0.000 (0.000–0.000)
CIN2	0.000 (0.000–0.000)	0.062 (0.038–0.100)	0.887 (0.804–0.938)	0.049 (0.031–0.077)
CIN3/cancer	0.000 (0.000–0.000)	0.000 (0.000–0.000)	0.044 (0.000–0.753)	0.955 (0.246–0.999)

Predicted 2-year transition probabilities



		State after two years			
Current state	**HPV category**	**Normal**	**CIN1**	**CIN2**	**CIN3/cancer**
Normal	HPV 16	0.832 (0.685–0.913)	0.064 (0.033–0.126)	0.079 (0.036–0.155)	0.023 (0.009–0.055)
	HPV 18	0.736 (0.493–0.897)	0.159 (0.054–0.310)	0.096 (0.006–0.214)	0.007 (0.000–0.144)
	HPV 52	0.933 (0.168–0.998)	0.045 (0.001–0.475)	0.020 (0.000–0.312)	0.001 (0.000–0.037)
	HPV 58	0.974 (0.050–0.999)	0.017 (0.000–0.581)	0.007 (0.000–0.324)	0.000 (0.000–0.052)
	Other hrHPVs	0.889 (0.808–0.940)	0.090 (0.049–0.157)	0.017 (0.008–0.036)	0.002 (0.000–0.007)
	No hrHPVs	0.910 (0.768–0.966)	0.068 (0.025–0.172)	0.019 (0.006–0.050)	0.001 (0.000–0.005)
CIN1	HPV 16	0.444 (0.325–0.555)	0.127 (0.085–0.194)	0.250 (0.168–0.339)	0.177 (0.093–0.293)
	HPV 18	0.622 (0.385–0.752)	0.161 (0.075–0.314)	0.189 (0.007–0.370)	0.026 (0.000–0.352)
	HPV 52	0.392 (0.153–0.485)	0.317 (0.246–0.479)	0.256 (0.186–0.356)	0.033 (0.010–0.105)
	HPV 58	0.410 (0.050–0.508)	0.330 (0.241–0.569)	0.238 (0.145–0.365)	0.020 (0.002–0.175)
	Other hrHPVs	0.734 (0.648–0.791)	0.171 (0.124–0.232)	0.071 (0.038–0.111)	0.023 (0.009–0.055)
	No hrHPVs	0.673 (0.614–0.712)	0.189 (0.159–0.239)	0.117 (0.091–0.145)	0.020 (0.009–0.041)
CIN2	HPV 16	0.299 (0.203–0.399)	0.137 (0.091–0.200)	0.285 (0.189–0.389)	0.278 (0.166–0.435)
	HPV 18	0.285 (0.024–0.484)	0.143 (0.006–0.267)	0.461 (0.000–0.730)	0.109 (0.000–0.965)
	HPV 52	0.236 (0.103–0.317)	0.349 (0.265–0.461)	0.334 (0.229–0.438)	0.079 (0.026–0.240)
	HPV 58	0.252 (0.041–0.345)	0.372 (0.242–0.514)	0.323 (0.150–0.443)	0.052 (0.006–0.407)
	Other hrHPVs	0.437 (0.316–0.546)	0.219 (0.155–0.271)	0.195 (0.099–0.313)	0.146 (0.065–0.308)
	No hrHPVs	0.406 (0.334–0.468)	0.250 (0.213–0.290)	0.253 (0.189–0.326)	0.089 (0.044–0.179)

Observed and simulated prevalence


HPV 16
HPV 18
HPV 52
HPV 58
Other hrHPVs
No hrHPVs
Normal
CIN1
CIN2
CIN3/Cancer
Prevalence (%)
Years after entry