Journal of Health Informatics and Statistics

50주년 특별호
Special Issue

J Health Info Stat 2022;47(Suppl 3):S51-S60 • https://doi.org/10.21032/jhis.2022.47.S3.S51

pISSN 2465-8014 | eISSN 2465-8022

경쟁위험을 가진 생존자료의 분석에 대한 고찰: R 패키지 함수를 중심으로

김진흠

수원대학교 데이터과학부 교수

Analyzing Survival Data with Competing Risks Based on R-packages

Jinheum Kim

Professor, Department of Applied Statistics, University of Suwon, Hwaseong, Korea

When it comes to data in survival analysis, it is easy to think of one censored or observed survival time of an individual. In the longitudinal study such as a clinical trial, an individual may be simultaneously exposed to different types of events. It is easy to think of competing risks data as an extension of univariate survival data, but since competing risks events are dependent on each other, if the analysis method of univariate survival data is applied, inferences with bias can be made. In this paper, we briefly introduce the statistical methods developed to analyze competing risks data and apply them to the real examples using the built-in functions in the R package.

Key words: Cause-specific hazard, Sub-distribution hazard, Left truncation, Time-dependent covariate, Multiple imputation

서 론

생존분석에서 자료라고 하면 한 개체의 중도절단되었거나 실제 관측된 생존시간 하나를 떠올리기 쉽다. 실제로 카플란-마이어(Kaplan-Meier, KM) 플롯이나[1] 콕스(Cox) 분석을[2,3] 포함해서 대부분의 통계적 방법들은 이와 같은 일변량 생존자료를 분석하는 데 초점을 맞추고 있다. 하지만 예방의학, 의학 등에서는 다변량 생존자료도 자주 접하게 된다. 다변량 생존자료란 한 개체가 동일한 유형의 이벤트를 연속적으로 경험하거나, 쌍둥이, 가계, 코호트 등과 같이 서로 다른 개체들이 동일한 이벤트를 동시에 경험하여 관측된 자료를 말한다[4]. 흔히 전자를 재발 혹은 반복 이벤트 자료라고 하고, 후자를 군집 혹은 그룹 혹은 상관된 자료라고 한다. 한편 임상시험과 같은 경시적 연구에서는 한 개체가 서로 다른 유형의 이벤트에 동시에 노출될 수 있는데 이와 같은 자료도 다변량 생존자료라고 말할 수 있다[5,6]. 이는 서로 다른 유형의 이벤트들이 모두 치명적이라면 가장 먼저 발생한 이벤트만 관측되지만, 치명적이지 않은 이벤트가 포함되어 있다면 두 개 이상의 이벤트가 관측될 수 있기 때문이다. 흔히 전자를 경쟁위험(competing risks, CR) 자료라고 하고, 후자를 다중상태(multi-state) 자료라고 한다. CR 자료를 일변량 생존자료의 확장으로 생각하기 쉬우나 CR 이벤트가 서로 종속되어 있기 때문에 일변량 생존자료의 분석방법을 그대로 적용하면 편의를 가진 추론을 할 수 있다[7,8]. 본 논문에서는 CR 자료를 분석하기 위해 개발된 통계적 방법을 간략히 소개하고, R 패키지에 내장된 함수를 이용하여 실제 자료에 적용하는 방법을 소개하고자 한다.

2절에서는 본 논문에서 예시로 사용할 자료와 용어를 소개하고, 3절에서는 CR 이벤트의 누적위험함수(cumulative hazard function)와 누적발생함수(cumulative incidence function, CIF)에 대한 비모수적인

Corresponding author: Jinheum Kim
17 Wquan-gil, Bongdam-eup, Hwaseong 18323, Korea
E-mail: jkimdt65@gmail.com

Received: November 7, 2022 **Accepted:** November 27, 2022 **Published:** November 30, 2022

*This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. NRF-2020R1F1A1A01048397).

No potential conflict of interest relevant to this article was reported.

How to cite this article:
Kim J. Analyzing survival data with competing risks based on R-packages. J Health Info Stat 2022;47(Suppl 3):S51-S60. Doi: https://doi.org/10.21032/jhis.2022.47.S3.S51

추정량을 소개하고, 4절에서는 비례위험모형(proportional hazards model, PHM)을 이용하여 위험인자가 CR 이벤트에 미치는 효과에 대한 통계적 추론방법을 소개하고, 5절에서는 4절에서 다룬 PHM을 좌절단된(left-truncated, LT) 자료와 시간가변(time-varying, TV) 공변량을 가진 자료로 확장하고, 6절에서는 본 연구에 대해 간략히 고찰하고자 한다.

자료 및 용어 소개

피부암 및 병원 자료

피부암 자료는 덴마크 오덴세(Odense)대학병원에서 1962-1977년까지 피부암으로 종양 절제 수술을 받은 환자 256명을 추적조사한 자료이다[9]. 이 자료는 수술을 받은 후부터 피부암으로 사망하거나 피부암이 아닌 다른 원인으로 사망할 때까지 시간(단위: 일)과 사망 원인, 성별, 수술 당시 연령, 종양 두께(단위: mm), 궤양 유무 등을 포함하고 있다. 256명 중에서 57명이 피부암으로 사망하였고, 14명이 피부암이 아닌 다른 원인으로 사망하였으며, 134명이 중도절단되었다. 이 자료는 MASS 패키지에서 얻었으며 다음 R 코드와 같이 자료를 변환하였다. 앞으로 이 자료를 'melanoma data'로 부른다.

```
library(MASS)
data(Melanoma)
Melanoma$sex = factor(Melanoma$sex, levels = c(0, 1), labels =
  c("Female","Male"))
Melanoma$ulcer = factor(Melanoma$ulcer, levels = c(0, 1), labels =
  c("No","Yes"))
Melanoma$cause = ifelse(Melanoma$status = = 2, 0,
  ifelse(Melanoma$status = = 3, 2, 1))
Melanoma$mel = ifelse(Melanoma$status = = 1, 1, 0)
Melanoma$oth = ifelse(Melanoma$status = = 3, 1, 0)
Melanoma$pid = 1:dim(Melanoma)[1]
Melanoma$time = Melanoma$time / 365.25
str(Melanoma)
```

병원 자료는 독일 샤리테(Charité)대학병원에서 수행한 SIR 3 (Spread of nosocomial Infections and Resistant pathogens) 코호트 연구에서 중환자실(intensive care unit, ICU)에 입원한 환자 1,313명을 추적조사한 자료이다[7,10]. 이 자료는 ICU에 입원한 때부터 ICU에서 사망하거나 퇴원할 때까지 시간(단위: 일)과 중도절단 여부, 이벤트 유형, 폐렴 감염(hospital-acquired pneumonia, HAP) 여부, 성별, 연령 등을 포함하고 있다. 또한 입원 중에 폐렴에 감염된 환자들은 감염 시점에 대한 정보도 가지고 있다. 147명이 ICU에서 사망하였으며, 1,145명이 살아서 퇴원하였고 21명이 중도절단되었다. HAP 환자가 108명이었는데 사망한 환자 중에서는 21명, 퇴원한 환자 중에서는 82명, 중도절단된 환자 중에서는 5명이 각각 포함되었다. 이 자료는 kmi 패키지에서 얻었으며 다음 R 코드와 같이 자료를 변환하였다. 앞으로 이 자료를 'hospital data'로 부른다.

```
library(kmi)
data(icu.pneu)
icu.pneu$pneu = factor(icu.pneu$pneu, levels = c(0, 1), labels =
  c("No","Yes"))
icu.pneu$outcome = icu.pneu$status * icu.pneu$event
str(icu.pneu)
```

용어

다중상태모형(multi-state model, MSM)은 2개 이상의 상태를 가진 모형을 일컫는다. 가장 간단하고 널리 쓰이는 MSM은 Figure 1A와 같이 'Alive'와 'Dead' 상태만 가진 모형이며 일변량 생존자료가 이에 해당한다. Figure 1B는 질병사망모형(illness-death model, IDM)이라고 부르는데, 질병 프로세스가 중간 이벤트에 해당하는 'Diseased' 상태로 전이되었다가 사망할 수 있는 자료에 적합한 모형이다. 본 연구에서 다루고자 하는 모형은 Figure 1C와 같이 흡수(absorbing) 상태가 2개 이상인 CR 모형이다. 이 모형은 한 이벤트가 발생하면 나머지 이벤트들은 중도절단되는 자료에 적합한 모형이다. 2.1절에서 소개한 피부암 자료에서는 피부암으로 인한 사망과 피부암이 아닌 다른 원인으로 인한 사망을 CR이라고 정의할 수 있으며, 병원 자료에서는 사망과 퇴원이 CR이라고 정의할 수 있다. 한편 CR 모형과 자주 비교되는 모형이 준경쟁위험(semi-competing risks, semi-CR)모형이다. CR 관계에 있는 이벤

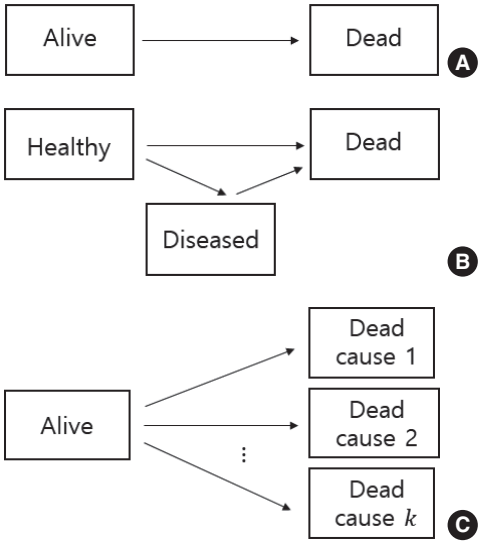

Figure 1. Graphical representation for the survival data (A), for the illness-death model (B), and for the competing risks data (C).

트가 피부암 자료나 병원 자료처럼 2개라고 가정하자. CR 모형처럼 CR 중에서 한 이벤트가 발생하면 다른 이벤트를 중도절단 시키지만 그 역은 성립하지 않는 모형이 semi-CR모형이다. IDM에서 'Dead' → 'Diseased' 전이가 허용되지 않으므로 이 모형은 semi-CR 모형에 해당한다.

CR 관계에 있는 이벤트가 2개라고 가정하면 CR 자료는 다음과 같이 이벤트 발생시간과 이벤트 유형으로 이루어진다.

$$(T = \min(T_1, T_2), e \in \{1,2\}).$$

여기서 T_k는 이벤트 유형 k의 발생 시간이며, e는 이벤트 유형이다. CR 자료에서는 위험함수를 이벤트 유형별 위험함수(cause-specific hazard, CSH)와 하위분포 위험함수(sub-distribution hazard, SH)로 정의한다.

첫째, 이벤트 유형 k의 CSH를 다음과 같이 정의하고,

$$\lambda_k(t) = \lim_{\Delta t \downarrow 0} \frac{P(t \le T < t+\Delta t,\ e=k | T \ge t)}{\Delta t}, k = 1,2, \quad (1)$$

통합 위험함수(overall hazard)를 다음과 같이 정의한다[7,8].

$$\lambda(t) = \sum_{k=1}^{2} \lambda_k(t) = \lim_{\Delta t \downarrow 0} \frac{P(t \le T < t+\Delta t | T \ge t)}{\Delta t}.$$

또한 이벤트 유형 k의 누적위험함수(cause-specific cumulative hazard, CSCH)를 다음과 같이 정의하고,

$$\Lambda_k(t) = \int_0^t \lambda_k(s)ds, k = 1,2 \quad (2)$$

통합 누적위험함수(overall cause-specific cumulative hazard)를 다음과 같이 정의한다[7,8].

$$\Lambda(t) = \int_0^t \lambda(s)ds.$$

이벤트 유형 k가 t 이내에 발생할 확률을 CIF라고 하며 다음과 같이 정의한다[7,8].

$$F_k(t) = P(T \le t, e = k) = \int_0^t S(s-)\lambda_k(s)ds, k = 1,2. \quad (3)$$

여기서 $S(t)$는 통합 생존함수(overall survival, OS)이고, 통합 누적위험함수로 표현하면 다음과 같다.

$$S(t) = P(T \ge t) = \exp(-\Lambda(t)). \quad (4)$$

한편 CIF는 다음과 같은 성질을 만족한다[7].

(i) $\lambda_k(t) = \frac{F_k'(t)}{S(t)}$ 이므로 $F_k'(t) = S(t)\lambda_k(t)$이다.
(ii) 모든 $t > 0$에 대해 $F_1(t) + F_2(t) = 1 - S(t)$이다.
(iii) $F_k(\infty) = P(e_i = k) < 1$ 이므로 F_k를 하위분포함수라고 한다.

둘째, 식 (3)에 대응하는 위험함수를 이벤트 유형 k의 SH로 정의하면 다음과 같다[7,8,11].

$$\lambda_k^*(t) \equiv -(\log(1 - F_k(t)))' = \frac{F_k'(t)}{1-F_k(t)}, k = 1,2. \quad (5)$$

식 (5)는 다음과 같이 표현할 수도 있다[7,8,11].

$$\lambda_k^*(t) = \lim_{\Delta t \downarrow 0} \frac{P(t \le T < t+\Delta t, e=k | T \ge t \cup (T \le t,\ e \ne k))}{\Delta t}. \quad (6)$$

따라서 SH는 CSH와 다르게 t 직전까지 이벤트가 발생하지 않은 개체뿐만 아니라 t 이전에 k가 아닌 다른 이벤트 유형이 발생한 개체도 위험집합에 포함된다. 식 (6)을 이용하여 CIF를 정의하면 다음과 같다[7,8,11].

$$F_k(t) = 1 - \exp(-\Lambda_k^*(t)), \quad k = 1,2. \quad (7)$$

여기서 $\Lambda_k^*(t)$는 하위분포 누적위험함수(sub-distribution cumulative hazard)이고 다음과 같이 정의된다.

$$\Lambda_k^*(t) = \int_0^t \lambda_k^*(s)ds, k = 1,2.$$

한편

$$P(e = k) = F_k(\infty) = 1 - \exp(-\Lambda_k^*(\infty)), k = 1,2$$

이고,

$$P(e = 1) + P(e = 2) = 1$$

이므로

$$\exp(-\Lambda_1^*(\infty)) + \exp(-\Lambda_2^*(\infty)) = 1$$

이다[7,8,11]. 따라서 SH에 기초한 방법은 이벤트에 따라 독립적으로 모형을 만들 수 없다. 한편

$$1 - F_1(t) = 1 - P(T \le t,\ e = 1) = P(T > t) + P(T \le t,\ e = 2) = S(t) + F_2(t)$$

이므로 CIF의 성질 (i)과 식 (5)로부터 이벤트 유형 1에 대한 CSH와 SH는 다음과 같은 관계가 성립한다[7,11].

$$\lambda_1(t) = w(t)\lambda_1^*(t), w(t) = \left(1 + \frac{F_2(t)}{S(t)}\right).$$

따라서 가중값 $w(t)$이 t에 따라 변할 뿐만 아니라 CR에도 의존하지만 모든 $t > 0$에 대해 CSH가 SH 보다 크다.

비모수적 추정

모든 개체는 우중도절단(right censoring, RC)될 수 있으므로 실제 관측되는 자료는 다음과 같다.

$$\{(x_1, e_1\delta_1), \ldots, (x_n, e_n\delta_n)\}.$$

여기서 $x_i = \min(t_i, c_i)$, $\delta_i = I(t_i \le c_i)$, c_i는 중도절단시간이다. 서로 다른 이벤트 발생 시간이 다음과 같다고 가정하자.

$$t_{(1)} < t_{(2)} < \cdots < t_{(r)}.$$

$d_{jk}\ (j = 1,2,\ldots,r)$는 $t_{(j)}$에서 발생한 이벤트 유형 k의 개체수라고 하자. 따라서 $d_j = \sum_{k=1}^{2} d_{jk}$는 $t_{(j)}$에서 발생한 이벤트 개체수이다. n_j는 $t_{(j)}$에서 위험집합(risk set)에 있는 개체수라고 하자.

CSH에 기초한 추정

식 (1)에서 정의한 CSH의 추정량을 다음과 같이 정의하고,

$$\hat{\lambda}_k(t_{(j)}) = \frac{d_{jk}}{n_j}, j = 1,2,\ldots,r,$$

식 (2)에서 정의한 CSCH의 넬슨-알렌(Nelson-Aalen, NA) 추정량을 다음과 같이 정의한다[7,8].

$$\hat{\Lambda}_k(t) = \sum_{j:t_{(j)} \le t} \hat{\lambda}_k(t_{(j)}).$$

식 (3)에서 정의한 CIF의 알렌-조한센(Aalen-Johansen, AJ) 추정량은 다음과 같이 정의된다[7,8].

$$\hat{F}_k(t) = \sum_{j:t_{(j)} \le t} \hat{S}(t_{(j)} -) \frac{d_{jk}}{n_j}. \quad (8)$$

여기서 $\hat{S}(t-)$는 식 (4)에서 정의한 OS에 대한 KM 추정량이며 다음과 같이 정의된다.

$$\hat{S}(t-) = \sum_{l:t_{(l)} < t} \left(1 - \frac{d_l}{n_l}\right).$$

CR 중에서 이벤트 유형 1만 이벤트로 정의하고 이벤트 유형 2는 RC로 간주한다고 하자. 이때 이벤트 유형 1이 t 바로 직전까지 발생하지 않을 확률 $S_1(t)$를 다음과 같이 정의한다.

$$S_1(t) = \exp\left(-\int_0^t \lambda_1(s)ds\right).$$

한편 모든 $t>0$에 대해 $\lambda(t) \ge \lambda_1(t)$이므로 다음과 같은 관계를 만족한다.

$$S_1(t) = \exp\left(-\int_0^t \lambda_1(s)ds\right) \ge \exp\left(-\int_0^t (\lambda_1(s) + \lambda_2(s))ds\right) = S(t)$$

따라서

$$\int_0^t S_1(s-)\lambda_1(s)ds \ge \int_0^t S(s-)\lambda_1(s)ds = F_1(t) \quad (9)$$

이다[7,8]. 그러므로 CR 중에서 한 유형만 이벤트로 정의하고 나머지 이벤트는 RC로 간주하여 CR 자료를 일변량 생존자료처럼 다루면 관심 있는 이벤트의 CIF를 과대 추정하게 된다.

NA 추정량은 survfit{survival} 함수나 mvna{mvna} 함수를 이용하고, AJ 추정량은 survfit{survival} 함수나 cuminc{cmprsk} 함수, etmCIF{etm} 함수, Cuminc{mstate}함수 등을 이용하면 된다[7,8]. 여기서 'ft{pkg}'에서 pkg는 R 패키지명을 의미하고, ft는 함수명을 의미한다. Figure 2는 피부암 자료를 분석한 결과이며 이에 대한 R 코드는 다음과 같다.

```
# Figure 2
library(mstate)
s.all = survfit(Surv(time, cause, type = "mstate") ~ 1, data = Melanoma)
s.mel = survfit(Surv(time, mel, type = "mstate") ~ 1, data = Melanoma)
s.oth = survfit(Surv(time, oth, type = "mstate") ~ 1, data = Melanoma)
par(mfrow = c(1, 2))
plot(s.all, lty = 1:2, lwd = 2, xlab = "Years", ylab = "Cumulative hazard",
  main = "(A)", fun = "cumhaz")
legend("topleft", lty = c(1, 2), lwd = 2, legend = c("Melanoma","Other"),
  bty = "n")
plot(s.all, lty = 1:2, lwd = 2, xlab = "Years", ylab = "CIF", main = "(B)")
lines(s.mel, lty = 3, lwd = 2)
lines(s.oth, lty = 3, lwd = 2)
legend("topleft", lty = 1:3, lwd = 2, legend = c("Melanoma","Other",
  "Ignoring competing risks"), bty = "n")
```

Figure 2A를 보면, 수술 후 204일까지는 피부암으로 인한 사망 위험(실선)보다 다른 원인으로 인한 사망 위험(파선)이 높다가 그 이후로는 서로 뒤바뀌어 피부암으로 인한 사망 위험이 피부암이 아닌 다른 원인으로 인한 사망보다 훨씬 높은 것으로 나타났다. Figure 2B에서 실선(피부암으로 인한 사망)과 파선(피부암이 아닌 다른 원인으로 인한 사망)은 CIF를 나타내고, 점선은 CR 중에서 관심 있는 이벤트만 이벤트로 간주하고 다른 이벤트는 RC로 가정한 경우인데 식 (9)에서 예상했던 것처럼 후자는 CIF를 과대 추정하였다.

SH에 기초한 추정

식 (6)에서 정의한 SH의 추정량을 다음과 같이 정의한다[7,8,11].

$$\hat{\lambda}_k^*(t_{(j)}) = \frac{d_{jk}}{n_j^*}, j = 1,2,\ldots,r.$$

여기서 n_j^*는 $t_{(j)}$에서 각 개체가 위험집합에 기여하는 양의 합인데, 만약 $i-$번째 개체가 $t_{(j)}$ 직전까지 이벤트가 발생하지 않았다면 1만큼 기여하고, 만약 $t_{(l)}(< t_{(j)})$에서 k가 아닌 다른 유형의 이벤트가 발생했다면

$$P(C \ge t_{(i)} | C \ge t_{(l)}), t_{(l)} < t_{(i)}$$

만큼 기여하는데 이에 대한 추정값은 다음과 같다.

$$\frac{\hat{\Gamma}(t_{(i)}-)}{\hat{\Gamma}(t_{(l)}-)}. \quad (10)$$

여기서 $\hat{\Gamma}(t)$는 $\Gamma(t) = P(C > t)$에 대한 KM 추정량이며 T와 C의 역할을 서로 바꾸어 추정한다. 식 (7)에 따른 CIF 추정량을 다음과 같이 정의한다[7,8,11].

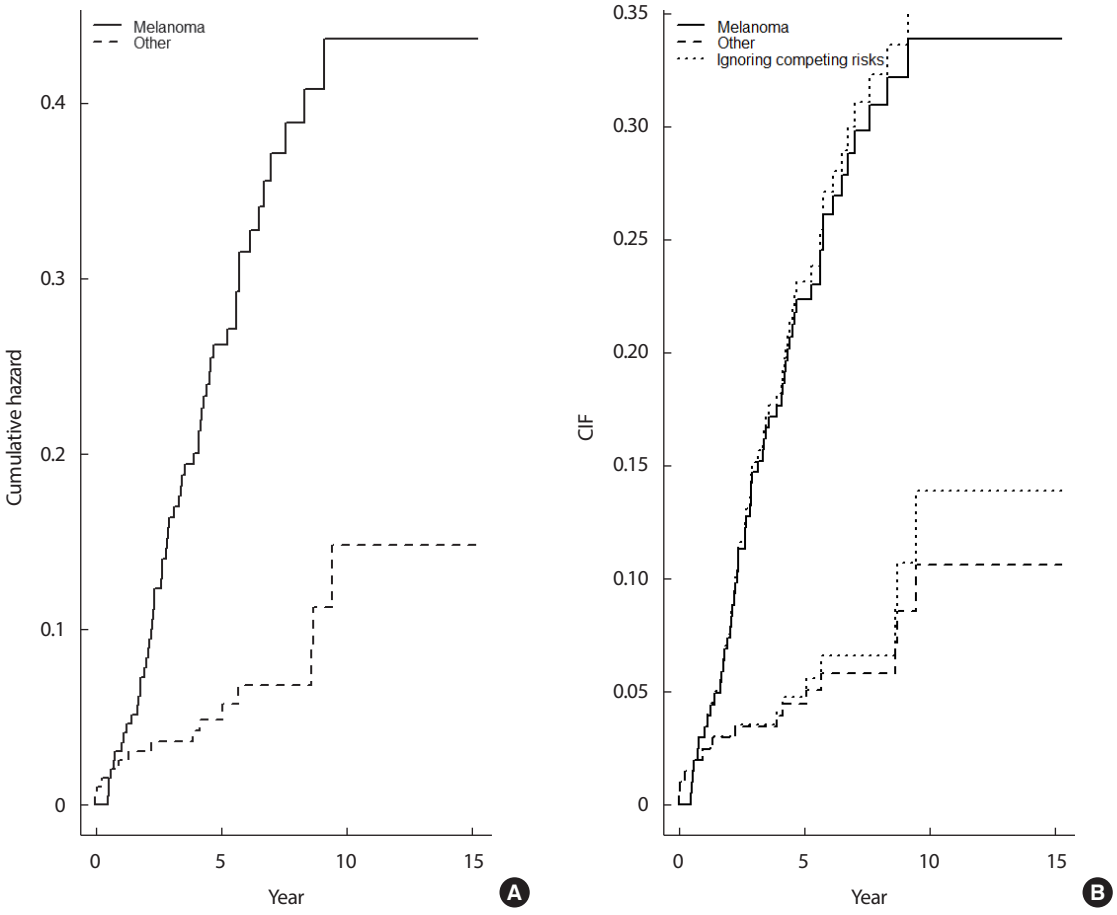

Figure 2. Nelsen-Aalen estimators (A) and Aalen-Johansen estimators (B) for the melanoma data.

$$\hat{F}_k(t) = 1 - \prod_{j:t_{(j)} \le t} (1 - \hat{\lambda}_k^*(t_{(j)})). \quad (11)$$

SDH 추정은 먼저 crprep{mstate} 함수를 이용하여 자료를 셈과정(counting process) 포맷으로 변환한 후 survfit{survival} 함수 등을 이용하면 된다[7,8]. Tables 1-3은 피부암 자료를 분석한 결과이며 이에 대한 R 코드는 다음과 같다.

```
# Table 1
crp.m = crprep(Tstop = "time", status = "cause", data = Melanoma,
  trans = 1, cens = 0, id = "pid")
WT = crp.m[crp.m$pid = = 114, ]
print(WT)

# Table 2
crp = crprep(Tstop = "time", status = "cause", data = Melanoma, trans
  = c(1, 2), cens = 0, id = "pid")
str(crp)print(Comp)
csPL = survfit(Surv(Tstart, Tstop, status = = failcode) ~ failcode, data =
  crp, weight = weight.cens)
r.CS = cbind(Time = s.all[[2]], nrisk.CS = s.all[[3]][, 1]) # cause-specific
nrisk.mel.SD = csPL[[3]][1:194] # subdisribution-melanoma
nrisk.oth.SD = csPL[[3]][195:388] # subdisribution-other
Comp = cbind(r.CS, nrisk.mel.SD, nrisk.oth.SD)

# Table 3
m.AJ = cbind(Time = s.all[[2]], CIF.mel.AJ = s.all[[6]][, 2]) # melanoma
  (CSH)
CIF.oth.AJ = s.all[[6]][, 3] # other causes (CSH)
CIF.mel.FG = 1 - csPL[[6]][1:194] # melanoma (SH)
CIF.oth.FG = 1 - csPL[[6]][195:388] # other causes (SH)
Equiv = cbind(m.AJ, CIF.mel.FG, CIF.oth.AJ, CIF.oth.FG)
print(Equiv)
```

Table 1은 한 예로 pid = 114인 개체가 위험집합에 기여하는 양을 정리한 것이다. pid = 114인 개체는 수술 후 2,085일에 피부암이 아닌 다른 원인으로 사망하였으며 이 환자가 사망한 이후 피부암으로 인한 사망 이벤트가 9개 시점에서 발생하였다. Table 1에서 3열 'Weight'는 식 (10)을 이용하여 pid = 114인 개체가 각 시점에서 위험집합에 기여하는 양을 계산한 값이다. Table 2는 CSH와 SH에 따른 위험집합을 비교하기

Table 1. Contribution to the risk set at each of nine competing risks event times since the time of death (208t days) for the subject with pid=114 who died from causes other than melanoma in the melanoma data

Start	Stop	Weight
0	2,085	1.000
2,085	2,103	0.989
2,103	2,108	0.978
2,108	2,256	0.899
2,256	2,388	0.854
2,388	2,467	0.796
2,467	2,565	0.738
2,565	2,782	0.678
2,782	3,042	0.630
3,042	3,338	0.452

Table 2. Comparison of risk sets according to cause-specific hazard and sub-distribution hazard at the specific event times for the melanoma data

Time	Cause-specific	Sub-distribution	
		Melanoma	Other causes
10	205	205.0	205.0
529	189	195.0	198.0
967	174	181.0	197.0
1,506	159	166.9	194.1
1,634	144	152.1	177.1
1,793	125	131.7	160.7
1,942	109	115.7	139.1
2,062	94	100.7	125.1
2,339	78	84.5	98.0
2,570	62	67.2	84.1
3,152	47	50.4	70.9
3,385	32	32.0	48.5
3,968	14	14.0	14.0

위해 일부 결과만 제시하였는데, 예상했던 대로 모든 시점에서 n_j (2열 'Cause-specific')는 n_j^* (3열 'Melanoma'는 피부암으로 인한 사망 이벤트에 해당하고, 4열 'Other causes'는 피부암이 아닌 원인으로 인한 사망 이벤트에 해당)보다 항상 크거나 같았다. Table 3은 CSH와 SH에 따른 CIF의 추정량 (8)과 (11)을 비교하기 위한 일부 결과인데, 예상했던 대로 두 추정량의 값이 정확히 일치하였다(피부암으로 인한 사망은 2열 'Melanoma'의 하위 두 열을 비교, 피부암이 아닌 원인으로 인한 사망은 3열 'Other causes'의 하위 두 열을 비교).

비례위험모형

성별, 처치 여부, 진단시점 연령, 병기(clinical stage) 등과 같은 시간불변(time-invariant) 공변량을 가진 CR 자료가 다음과 같다고 하자.

Table 3. Comparison of the cumulative incidence estimators according to cause-specific hazard and sub-distribution hazard at the specific event times for the melanoma data

Time	Melanoma		Other causes	
	Cause-specific	Sub-distribution	Cause-specific	Sub-distribution
10	0.000	0.000	0.005	0.005
529	0.049	0.049	0.029	0.029
967	0.118	0.118	0.034	0.034
1,506	0.181	0.181	0.039	0.039
1,634	0.207	0.207	0.044	0.044
1,793	0.224	0.224	0.044	0.044
1,942	0.230	0.230	0.050	0.050
2,062	0.245	0.245	0.050	0.050
2,339	0.269	0.269	0.058	0.058
2,570	0.298	0.298	0.058	0.058
3,152	0.322	0.322	0.058	0.058
3,385	0.339	0.339	0.085	0.085
3,968	0.339	0.339	0.106	0.106

$$\{(x_1, e_1\delta_1, z_1), (x_2, e_2\delta_2, z_2), \ldots, (x_n, e_n\delta_n, z_n)\}.$$

여기서 z_i는 p −차원 공변량(위험인자) 벡터이다.

CSH에 기초한 회귀계수 추정

공변량 z가 주어졌을 때 이벤트 유형 k에 대한 PHM은 다음과 같다[7,8].

$$\lambda_k(t|z) = \lambda_{0k}(t)\exp(\beta_k' z),\ k = 1,2.$$

여기서 λ_{0k}는 이벤트 유형 k에 대한 기저(baseline) 위험함수이고, β_k는 이벤트 유형 k에 대한 회귀계수이다. β_k에 대한 최대우도추정량(maximum likelihood estimator)은 부분우도(partial likelihood, PL) 함수를 이용하여 정의할 수 있는데 PL은 다음과 같다.

$$L(\beta_1, \beta_2) = \prod_{k=1}^{2} L_k(\beta_k).$$

여기서

$$L_k(\beta_k) = \prod_{i=1}^{n}\left(\frac{\exp(\beta_k' z_i)}{\sum_{l=1}^{n} I(x_l \geq x_i)\exp(\beta_k' z_l)}\right)^{I(\delta_i e_i = k)},\ k = 1,2 \quad (12)$$

이다. PL이 이벤트 유형별로 분해되므로 이벤트 유형별로 회귀계수를 추정할 수 있는 장점이 있다.

회귀계수 추정은 이벤트 유형별로 표준 포맷의 자료를 생성하여 coxph{survival} 함수를 이용하거나 msprep{mstate} 함수 등을 이용하여 길쭉하게 쌓은(stacked-long) 포맷의 자료를 생성한 후 coxph{survival} 함수 등을 한 번만 이용하면 된다[7,8]. CR 자료를 길쭉하게 쌓은 포맷으로 정리하면 모든 개체는 CR 이벤트의 개수만큼 중복된 행으로 표

Table 4. Estimated proportional cause-specific hazards model for the melanoma data

Covariate	Estimate	HR	SE	*p*-value
Age	0.01	1.01	0.01	0.141
Ulceration: Yes	1.16	3.20	0.31	0.000
Tumor thickness	0.11	1.12	0.04	0.004
Sex: Male	0.43	1.54	0.27	0.106
Age × Other causes	0.06	1.06	0.02	0.009
Ulceration: Yes × Other causes	-1.06	0.35	0.67	0.114
Tumor thickness × Other causes	-0.06	0.94	0.10	0.535
Sex: Male × Other causes	-0.07	0.93	0.61	0.902

HR, hazard ratio; SE, standard error.

현된다. 다만 CR 이벤트에 해당하는 행에만 1의 값을 갖고 나머지 행에는 0의 값을 갖는 지시자(indicator) 열이 추가된다. Table 4는 피부암 자료를 분석한 결과이며 이에 대한 R 코드는 다음과 같다.

```
# Table 4
tmat = transMat(x = list(c(2, 3), c(), c()), names = c("operation","melanoma","other"))
M.s = msprep(time = c(NA,"time","time"), status = c(NA,"mel","oth"),
  data = Melanoma, trans = tmat, keep = c("age","sex","ulcer","thickness"))
M.s$failcode = ifelse(M.s$trans == = 1,"Melanoma","Other")
str(M.s)
CS = coxph(Surv(time, status) ~ age * strata(failcode) + ulcer *
  strata(failcode) + thickness * strata(failcode) + sex * strata(failcode),
  data = M.s)
summary(CS)
```

궤양이 있는 환자가 궤양이 없는 환자에 비해 피부암으로 인한 사망 위험이 3.2배 높았으며(95% confidence interval, CI: 1.75, 5.88; $p<0.001$), 종양 두께가 1 mm 증가하면 피부암으로 인한 사망 위험이 1.1배 높았고(95% CI: 1.04, 1.20; $p=0.004$), 진단 시점 연령이 한 살 증가하면 피부암 아닌 다른 원인으로 인한 사망 위험이 1.1배 높았다(95% CI: 1.03, 1.12; $p<0.001$). 남녀 간 사망 위험은 이벤트에 관계없이 서로 다르지 않았다.

SH에 기초한 회귀계수 추정

공변량 z가 주어졌을 때 이벤트 유형 k에 대한 CIF를 다음과 같이 정의하고,

$$F_k(t;z) = P(T \le t,\ e = k|z), k = 1,2,$$

SH에 기초한 PHM을 다음과 같이 정의한다[7,8,11].

$$\lambda_k^*(t|z) = \lambda_{0k}^*(t)\exp(\gamma_k^{'}z), k = 1,2.$$

여기서 λ_{0k}^*는 이벤트 유형 k에 대한 기저 하위분포 위험함수이고, γ_k는 이벤트 유형 k에 대한 회귀계수이다. 관리자에 의한 중도절단(administrative censoring)처럼 모든 개체의 중도절단시간 c_i $(i = 1,2,\dots,n)$를 안다면 PL은 다음과 같이 정의된다[7,8,11].

$$L_k(\gamma_k) = \prod_{i=1}^{n}\left(\frac{\exp(\gamma_k^{'}z_i)}{\sum_{l=1}^{n}(I(x_l\ge x_i)+I(t_l<x_i<c_l,\delta_l e_l=3-k)\frac{\hat{\Gamma}(x_i-)}{\hat{\Gamma}(t_l-)})\exp(\gamma_k^{'}z_l)}\right)^{I(\delta_i e_i=k)}, k = 1,2. \quad (13)$$

관리자에 의한 중도절단이란 연구 종료 시점에 한 번만 중도절단 시키거나(유형 I 중도절단), 사전에 계획된 시점마다 중도절단 시키는(전진적 유형 I 중도절단) 것을 의미한다. 그러나 대부분 CR에 의해 중도절단 되는 개체의 가능(potential) 중도절단시간을 알 수 없기 때문에 중도절단 시간을 알고 있는 개체들을 이용하여 가능 중도절단시간을 다중대체(multiple imputation)하는 방법을 활용해야 한다[12]. 식 (12)와 (13)를 비교하면, x_i 직전까지 이벤트가 발생하지 않은 개체는 두 식 모두 동일하게 위험집합에 1만큼 기여하는 것으로 정의한다. 하지만 x_i 이전에 이벤트 유형 2가 발생한 개체는 서로 다르게 정의하는데, 식 (12)는 위험집합에 기여하지 않는 것으로 정의하는 반면에 식 (13)은 중도절단 되기 전 까지만 조건부 확률의 추정값 만큼 기여하는 것으로 정의한다.

회귀계수 추정은 이벤트 유형별로 crr{cmprsk} 함수를 이용하거나 crprep{mstate} 함수를 이용하여 길쭉한 포맷의 자료를 생성한 후 coxph{survival} 함수 등을 한 번만 이용하면 된다[7,8]. Tables 5, 6은 피부암 자료를 분석한 결과이며 이에 대한 R 코드는 다음과 같다.

```
# Table 5
crp.r = crprep(Tstop = "time", status = "cause", data = Melanoma, trans
  = c(1, 2), cens = 0, keep = c("age","sex","ulcer","thickness"), id = "pid")
str(crp.r)
RKs = crp.r[crp.r$pid %in% c(162, 171), ]
print(RKs)

# Table 6
FG = coxph(Surv(Tstart, Tstop, status = = failcode) ~ age *
  strata(failcode) + ulcer * strata(failcode) + thickness * strata(failcode)
  + sex * strata(failcode), data = crp.r, weight = weight.cens)
summary(FG)
```

Table 5는 피부암이 아닌 다른 원인으로 사망한 환자와(pid = 162) 피부암으로 사망한 환자가(pid = 171) 식 (13)의 위험집합에 기여하는 양(5열 'Weight')을 계산한 것이다. SH에 기초한 회귀계수 추정의 결과는 CSH에 기초한 회귀계수 추정의 결과와 유사하였다(Table 6). 궤양이 있는 환자가 궤양이 없는 환자에 비해 피부암으로 인한 사망 위험이 3.1배 높았으며(95% CI: 1.71, 5.57; $p<0.001$), 종양 두께가 1 mm 증가하면 피부암으로 인한 사망 위험이 1.1배 높았고(95% CI: 1.02, 1.17;

Table 5. Contribution of patients who died from causes other than melanoma (pid=162) and patients who died from melanoma (pid=171) to the risk set of the partial likelihood in the proportional sub-distribution hazards model

pid	Start	Stop	Status	Weight	Failcode
162	0	3,182	2	1.000	1
162	3,182	3,338	2	0.814	1
171	0	3,338	1	1.000	1
162	0	3,182	2	1.000	2
171	0	3,338	1	1.000	2
171	3,338	3,458	1	0.824	2

$p = 0.019$), 진단 시점 연령이 한 살 증가하면 피부암 아닌 다른 원인으로 인한 사망 위험이 1.1배 높았다(95% CI: 1.03, 1.09; $p < 0.001$). 남녀 간 사망 위험은 이벤트에 관계없이 서로 다르지 않았다.

비례위험모형의 확장

좌절단 자료

LT 자료는 연구시작 시간이 지연된 자료라고 하는데, 4절에서 다룬 PHM을 LT 자료에 적용하기 위해서는 위험집합만 다시 정의하면 된다. 4절에서 정의한 자료에 지연된 시작(late entry) 정보가 추가된 자료가 다음과 같다고 하자.

$$\{(x_i, e_i\delta_i, z_i, l_i), i = 1,2, \dots, n\}.$$

여기서 l_i는 i-번째 개체의 지연된 시작 시간이다. CSH에 기초한 추정에서는 식 (12)에서

$$I(x_i \geq t)$$

를

$$I(x_i \geq t > l_i)$$

로 대체하고, SH에 기초한 추정에서는 식 (13)에서

$$I(x_i \geq t) + I(t_i < t < c_i, \delta_i e_i = 3 - k) \times \frac{\Gamma(t-)}{\Gamma(t_i -)}$$

를

$$I(x_i \geq t > l_i) + I(l_i \vee t_i < t < c_i, \delta_i e_i = 3 - k) \times \frac{\Gamma(t-)}{\Gamma(t_i -)} \times \frac{\Phi(t -)}{\Phi(t_i -)}$$

로 대체하면 된다. 여기서 $\Phi(t) = P(L_i \leq t)$는 관측 시간을 뒤집어 정의한 분포함수이다[7].

시간가변 공변량

2.1절에서 소개한 병원 자료에서 환자들은 ICU 입원 중에 폐렴에 감염될 수 있다. 이와 같이 연구 참여 기간 중에 공변량의 값이 변하는 공변량을 TV 공변량이라고 한다. TV 공변량은 외적 공변량과 내적 공변량으로 구분된다. 특히 내적 공변량은 측정오차를 포함하고 있기 때문에 내적 공변량을 처리하려면 결합 모형 등을 도입해야 하는데, 본 논문에서는 외적 공변량처럼 처리하고자 한다. 따라서 모니터링 시점 바로 직전까지는 이전 시점에서 관측한 공변량의 값이 그대로 유지한다고 가정한다.

Table 6. Estimated proportional sub-distribution hazards model for the melanoma data

Covariate	Estimate	HR	SE	*p*-value
Age	0.01	1.01	0.01	0.511
Ulceration: Yes	1.13	3.09	0.31	0.000
Tumor thickness	0.09	1.09	0.04	0.009
Sex: Male	0.41	1.50	0.27	0.129
Age × Other causes	0.05	1.05	0.02	0.001
Ulceration: Yes × Other causes	-1.24	0.29	0.67	0.054
Tumor thickness × Other causes	-0.08	0.92	0.09	0.371
Sex: Male × Other causes	-0.14	0.87	0.61	0.823

HR, hazard ratio; SE, standard error.

TV 공변량을 가진 자료에서 회귀계수 추정은 TV 공변량에 대해 tmerge{survival} 함수를 이용하여 길쭉한 포맷으로 자료를 생성한 후, CSH에 기초한 추정에서는 coxph{survival} 함수를 이용하고[7,8], SH에 기초한 추정에서는 kmi{kmi} 함수를 이용하여 CR 이벤트에 의해 중도절단된 중도절단시간을 다중대체한 후 cox.kmi{kmi} 함수 등을 이용한다[7,12]. Table 7은 병원 자료를 분석한 결과이며 특히 SH에 기초한 경우는 다중대체를 100번 반복한 결과이다. 이에 대한 R 코드는 다음과 같다.

```
# Table 7
cs.death = coxph(Surv(start, stop, outcome == 2) ~ pneu + sex + age,
  data = icu.pneu)
cs.disch = coxph(Surv(start, stop, outcome == 3) ~ pneu + sex + age,
  data = icu.pneu)
summary(cs.death) # dead in ICU
summary(cs.disch) # discharged
imp.death = kmi(Surv(start, stop, outcome != 0) ~ 1, data = icu.pneu,
  etype = outcome, id = id, failcode = 2, nimp = 100) # dead in ICU
str(imp.death)
imp.disch = kmi(Surv(start, stop, outcome != 0) ~ 1, data = icu.pneu,
  etype = outcome, id = id, failcode = 3, nimp = 100) # discharged
str(imp.disch)
sh.death = cox.kmi(Surv(start, stop, outcome == 2) ~ pneu + sex +
  age, imp.death)
sh.disch = cox.kmi(Surv(start, stop, outcome == 3) ~ pneu + sex +
  age, imp.disch)
summary(sh.death) # dead in ICU
summary(sh.disch) # discharged
```

Table 7. Estimated proportional cause-specific and sub-distribution hazards models for the hospital data

Covariate	Estimate	HR	SE	*p*-value
	Cause-specific hazard, Outcome = dead			
Pneumonia: Yes	-0.09	0.92	0.25	0.724
Sex: Male	-0.02	0.98	0.17	0.902
Age	0.02	1.02	0.01	< 0.001
	Cause-specific hazard, Outcome = discharged alive			
Pneumonia: Yes	-0.50	0.61	0.12	< 0.001
Sex: Male	-0.12	0.89	0.06	0.049
Age	-0.001	1.00	0.002	0.626
	Sub-distribution hazard, Outcome = dead			
Pneumonia: Yes	1.06	2.89	0.24	< 0.001
Sex: Male	0.11	1.12	0.17	0.515
Age	0.02	1.02	0.01	< 0.001
	Sub-distribution hazard, Outcome = discharged alive			
Pneumonia: Yes	-0.24	0.79	0.12	0.045
Sex: Male	-0.11	0.90	0.06	0.077
Age	-0.004	1.00	0.002	0.030

HR, hazard ratio; SE, standard error.

연령과 성별을 고정한 후 HAP 여부가 CR 이벤트에 미치는 효과를 살펴보면 다음과 같다. CSH에 기초한 추정에서는 HAP를 가진 환자가 그렇지 않은 환자보다 퇴원율이 낮았으나(hazard ratio, HR = 0.91, $p < 0.001$) 사망 위험은 HAP 여부에 따라 통계적으로 유의한 차이가 없었다. SH에 기초한 추정에서도 마찬가지로 HAP를 가진 환자가 그렇지 않은 환자보나 퇴원이 낮았으며(HR = 0.79, $p = 0.045$), 사망 위험도 높았다(HR = 2.89, $p < 0.001$). 두 방법에 관계없이 HAP는 입원 기간을 연장시키는 효과가 있었고, SH 결과에 따르면 ICU 입원이 길어지면 사망 위험이 높아지는 경향이 있는 것으로 나타났다.

고 찰

본 논문에서는 CR이 있는 생존자료를 분석하기 위해 개발된 통계적 방법과 R 패키지에 내장된 함수를 이용하여 실제 자료에 적용하는 방법을 소개하였다. CR 자료에서는 위험함수를 CSH와 SH로 정의하는데 비모수적 추정에서 두 방법 간 위험집합의 차이에 대해 살펴보았으며 이벤트 유형별 CIF 추정량을 서로 비교하였다. PHM에서도 두 방법 간 PL의 차이에 대해 살펴보았으며 mstate 패키지에 내장된 msprep 함수와 crprep 함수를 CSH에 기초한 추정과 SH에 기초한 추정에 각각 적용하여 자료를 변환한 후 가장 널리 사용되고 있는 survival 패키지에 내장된 coxph 함수를 이용하여 회귀계수 추정하였다. 마지막으로 연구 시작 시점이 좌절단되었거나 TV 공변량을 가진 자료로 확장하였으며 특히 SH에 기초한 추정에서 CR 이벤트에 의해 중도절단된 중도절단시간을 kmi 패키지에 내장된 kmi 함수를 이용하여 다중대체한 후 TV 공변량이 CR 이벤트의 발생에 미치는 효과를 추정하였다.

2.2절에서 언급한 것처럼 다변량 생존자료분석에서 CR 모형만큼이나 semi-CR 모형도 널리 이용되고 있는데 mstate 패키지에서 상태 간 전이에 대한 행렬을 semi-CR 모형에 부합되게 정의하면 CSH에 기초한 CIF의 비모수적 추정과 PHM에서 회귀계수의 추정이 가능하다. 또한 mstate 패키지는 중간 이벤트가 2개 이상이고 양방향 전이가 가능한 다중상태모형으로도 확장이 가능하다. 다만 CR 모형 이외의 다른 모형에 대해서는 아직 SH 기초한 추정이 어렵기 때문에 이에 대한 패키지 개발이 필요하다고 생각한다.

ORCID

Jinheum Kim *https://orcid.org/0000-0003-3408-900X*

REFERENCES

1. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc 1958;53(282):457-481. DOI: 10.2307/2281868
2. Cox DR. Regression models and life-tables. J R Stat Soc Series B Stat Methodol 1972;34(2):187-220. DOI: 10.1111/j.2517-6161.1972.tb00899.x
3. Cox DR. Partial Likelihood. Biometrika 1975;62(2):269-276. DOI: 10.2307/2335362
4. Aalen OO, Borgan Ø, Gjessing HK. Survival and event history analysis: a process point of view. New York, NY: Springer; 2000, p. 271-272.
5. Hougaard P. Analysis of multivariate survival data. New York, NY: Springer; 2000, p. 11-14.
6. Kalbfleisch JD, Prentice RL. The statistical analysis of failure time data. 2nd ed. Hoboken, NJ: John Wiley & Sons; 2002, p. 247-248.
7. Beyersmann J, Allignol A, Schumacher M. Competing risks and multistate models with R. New York, NY: Springer; 2012.
8. Geskus RB. Data Analysis with competing risks and intermediate states. New York, NY: Chapman and Hall/CRC; 2016.
9. Andersen PK, Borgan O, Gill RD, Keiding N. Statistical models based on counting processes. New York, NY: Springer; 1993, p. 11-14.
10. Beyersmann J, Schumacher M. Time-dependent covariates in the proportional hazards model for competing risks. Biostatistics 2008;9(4):765-776. DOI: 10.1093/biostatistics/kxn009

11. Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. J Am Stat Assoc 1999;94(446):496-509. DOI: 10.1080/01621459.1999.10474144

12. Ruan PK, Gray RJ. Analyses of cumulative incidence functions via non-parametric multiple imputation. Stat Med 2008;27(27):5709-5724. DOI: 10.1002/sim.3402