フリーソフト R を活用した生存データ解析 (総合報告)

Survival Data Analysis Using R

辻谷将明 [*1], 田中祐輔 [*2]

The Cox's proportional hazards model has been widely used for the analysis of treatment and prognostic effects with censored survival data. The model was developed based on the hazard function for an individual as a fixed-time covariates. When the covariates values change for the duration of the study, however, a number of theoretical problems that are to be solved with respect to the baseline survival function and the baseline cumulative hazard function, are involved. This article presents time-dependent survival data analysis using R which may be evaluated for their impact on survival. We consider how these analyses can affect the prediction of patient outcome using multistate model, stratified Cox' proportional hazard model, and competing risk model.

Key words:Cox's proportional hazard model, Event-history analysis, competing risk, stratified proportional hazard model, cumulative incidence function

キーワード :Cox 比例ハザードモデル , イベント - ヒストリー解析 , 競合リスク , 層別比例ハザードモデル , 累積発生関数

1　はじめに

従来、生存時間解析には Cox 比例ハザードモデルが広範に活用されてきた (大橋 , 浜田 ,1995; 中村 , 2001;Kalbfleish and Prentice, 2002;Klein and Moeschberger, 2003;Lee et al.,2003)。Cox 比例ハザードモデルでは、生存時間 (観測時点) t のハザード関数を

$$h\left(t\,|\boldsymbol{x}\right)=h_0\left(t\right)\exp\left(\sum_{i=1}^{I}\beta_i x_i\right) \tag{1}$$

と表す。ここで、ベースラインハザード関数 $h_0\left(t\right)$ は観測時点の関数であって、共変量 $\boldsymbol{x}=(x_1,\cdots,x_I)$ を含んでいない。このモデルでは、共変量の観測は各患者について 1 回のみで、時間が変化しても不変である。しかし、反復して測定される検査値の生存時間への影響の評価、あるいは観測期間中に薬剤の投与量を変化させたとき、その効果の有無の検証が必要な場面もでてくる (Andersen et al., 1982; Aitkin et al., 1983; Altman et al., 1984; Christensen et al., 1986; Aalen et al., 2008)。

本稿では近年 、広範に活用されているフリー・ソフト R(辻谷 , 外山 , 2007; 辻谷 , 竹澤 , 2009; 辻谷 , 和田 , 2012) を援用し、共変量が時間と共に変動する“時間依存型”の典型的な生存データについて考察する。そして、それを拡張した multi-state モデルに基づくイベントヒストリー解析を取上げる。同一の個体について、イベントが複数回、繰返して発生する場合である。例えば、膀胱癌では再発を繰返すことがよくある。骨髄移植のように移植→再発→死亡と多段階の state を繰返すこともある。次に、競合リスクモデルについて解説する (西川 ,2008)。ある個体について、注目している死亡が発生するまでの観測過程で、それ以外の死亡 (これを競合リスクと呼ぶ) が起こることがある。これは、従来の比例ハザードモデルを採用し、データを時間依存型にして解析できる。累積発生関数に基づく 2 群間の比

[*1] 大阪電気通信大学　情報通信工学部　情報工学科 , 連絡先〒 572-8530 寝屋川市初町 18-8;E-mail:ekaaf900@ricv.zaq.ne.jp
[*2] イーピエス株式会社臨床情報本部データサイエンスセンター , 〒 112-0004 大阪市淀川区宮原 3-4-30 ニッセイ新大阪ビル 11 階

2013 年 4 月 30 日受理

較や有意性検定を行う。更に、共変量を考慮し、部分分布という考えを取込んだ競合リスク回帰モデルを取上げる。ここでも、共変量に時間依存型データの含まれる場面に拡張できる。

2 比例ハザードモデル

生存時間 T を確率変数としたとき，時間 t まで生存する確率が，生存時間関数

$$S(t) = \Pr\{T \geq t\} \tag{2}$$

である．累積率分布関数は

$$F(t) = 1 - S(t) \tag{3}$$

となる．t の直前まで生存した人が，次の Δt の期間に死亡する条件付き確率

$$\Pr\{t \leq T < t + \Delta t \,|T \geq t\} = \frac{\Pr\{T < t + \Delta t\} - \Pr\{T \geq t\}}{\Pr\{T \geq t\}} \tag{4}$$

は，Δt に依存する．このとき，単位時間当たりの平均死亡率

$$\frac{\Pr\{t \leq T < t + \Delta t \,|T \geq t\}}{\Delta t} \tag{5}$$

について，$\Delta t \to 0$ のとき，時間 t におけるハザード関数を

$$h(t\,|\boldsymbol{x}) = \lim_{\Delta t \to 0} \left\{ \frac{S(t) - S(t + \Delta t)}{S(t)\Delta t} \right\} \tag{6}$$

と定義する．すなわち，t まで生存した人のうち，$t + \Delta t$ までに死ぬ人の割合を，単位時間当たりの量に換算し，$\Delta t \to 0$ としたときの極限値である．ハザード関数 $h(t)$，$S(t)$, 確率密度関数 $f(t)$ との間には

$$\begin{aligned} h(t\,|\boldsymbol{x}) &= \lim_{\Delta t \to 0} \left\{ \frac{S(t) - S(t + \Delta t)}{S(t)\Delta t} \right\} \\ &= \frac{1}{S(t)} \lim_{\Delta t \to 0} \left\{ \frac{F(t + \Delta t) - F(t)}{\Delta t} \right\} \\ &= \frac{f(t)}{S(t)} \end{aligned} \tag{7}$$

という関係がある．

個体が生存時間に影響を与える因子（予測変数，背景因子，予後因子）を $\boldsymbol{x} = (x_1, x_2, \cdots, x_I)$ とする．この値は時間に依存しない．このとき，予測変数 $\boldsymbol{x}$ をもつ個体のハザード関数 $h(t\,|\boldsymbol{x})$ を

$$h(t\,|\boldsymbol{x}) = h_0(t)\, r(\boldsymbol{x}) \tag{8}$$

と書く．ここに，$h_0(t)$ をベースラインハザード関数，$r(\boldsymbol{x})$ を相対リスク (relative risk) と呼ぶ．2 つの予測変数 $\boldsymbol{x}, \boldsymbol{x}'$ について

$$h(t\,|\boldsymbol{x}') = h(t\,|\boldsymbol{x}) \left\{ \frac{r(\boldsymbol{x}')}{r(\boldsymbol{x})} \right\} \tag{9}$$

より，$h(t\,|\boldsymbol{x})$ と $h(t\,|\boldsymbol{x}')$ は比例する（比例ハザード性）．更に，対数線形性

$$\ln r(x) = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_I x_I = \boldsymbol{\beta}^t \boldsymbol{x} \tag{10}$$

を仮定すると比例ハザードモデル

$$\begin{aligned} h(t\,|\boldsymbol{x}) &= h_0(t) \exp\left(\boldsymbol{\beta}^T \boldsymbol{x}\right) \\ &= h_0(t) \exp\left(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_I x_I\right) \end{aligned} \tag{11}$$

となる．特に

$$x_1 = x_2 = \cdots = x_I = 0 \tag{12}$$

なら

$$h\left(t\right) = h_0\left(t\right) \tag{13}$$

となる．

(11) 式の未知パラメータ β は部分尤度

$$\begin{aligned} PL\left(\beta\right) &= \prod_{i=1}^{\text{死亡数}} \left\{ \frac{\exp\left(\boldsymbol{\beta}^T \boldsymbol{x}_i\right)}{\sum\limits_{k \in R\left(t_i\right)} \exp\left(\boldsymbol{\beta}^T \boldsymbol{x}_k\right)} \right\} \\ &= \prod_{i=1}^{n} \left\{ \frac{\exp\left(\boldsymbol{\beta}^T \boldsymbol{x}_i\right)}{\sum\limits_{k=1}^{n} Y_k\left(t_i\right) \exp\left(\boldsymbol{\beta}^T \boldsymbol{x}_k\right)} \right\}^{\delta_i} ; Y_k = \begin{cases} 1 : k \in R\left(t\right) \\ 0 : o.w. \end{cases} \end{aligned} \tag{14}$$

を最大化して求める (Cox, 1975)．ここに，$\boldsymbol{x}_i = \left(x_{i1}, x_{i2}, \cdots, x_{iI}\right) = 0$ で、$R\left(t_i\right)$ は時点 i でのリスク集合 (時点 i の直前まで生存した個体から成る集合) である．

表 1 は原発性胆汁性肝硬変 (PBC) データである (Collett,2003)。共変量としてビリルビンの初診時の対数をとる。

表 1　原発性胆汁性肝硬変 (PBC) データ

患者#	生存時間 (日)	打切り (=0)	ln（ビリルビン値）:初診時
1	281	1	3.2
2	604	0	3.1
3	457	1	2.2
・	・	・	・
7	1514	1	2.4
・	・	・	・
12	1071	1	3.1

この部分尤度の計算は、表 2 のようになる。

表 2　部分尤度の計算表

生存時間	打切り *(=0)*	ln(bil)	尤度
182	0	2.4	
281	1	3.2	$e^{3.2\beta}/(e^{3.2\beta} + e^{2.8\beta} + e^{3.9\beta} + \cdot\cdot + e^{2.3\beta} + e^{2.4\beta})$
341	0	2.8	
384	1	3.9	$e^{3.9\beta}/(e^{3.9\beta} + e^{2.2\beta} + e^{3.1\beta} + \cdot\cdot + e^{2.3\beta} + e^{2.4\beta})$
457	1	2.2	$e^{2.2\beta}/(e^{2.2\beta} + e^{3.1\beta} + e^{3.8\beta} + \cdot\cdot + e^{2.3\beta} + e^{2.4\beta})$
604	0	3.1	
814	1	3.8	$e^{3.8\beta}/(e^{3.8\beta} + e^{2.4\beta} + e^{3.1\beta} + \cdot\cdot + e^{2.3\beta} + e^{2.4\beta})$
842	1	2.4	$e2.4\beta/(e2.4\beta + e3.1\beta + e2.5\beta + e2.3\beta + e2.4\beta)$
1071	1	3.1	$e^{3.1\beta}/(e^{3.1\beta} + e^{2.5\beta} + e^{2.3\beta} + e^{2.4\beta})$
1121	1	2.5	$e^{2.5\beta}/(e^{2.5\beta} + e^{2.3\beta} + e^{2.4\beta})$
1411	0	2.3	
1514	1	2.4	$e^{2.4\beta}/e^{2.4\beta}$

通常の Cox の比例ハザードモデルのプログラムは、

```
>library(survival)
>train<-read.csv("F:\\train.csv",header=FALSE)
>fit<-coxph(Surv(V1,V2)~V3,data=train)
>summary(fit)
```

となり、解析結果

```
     coef exp(coef) se(coef)     z Pr(>|z|)
V3 1.4328    4.1904   0.7903 1.813   0.0698 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
   exp(coef) exp(-coef) lower .95 upper .95
V3     4.190     0.2386    0.8904     19.72
Rsquare= 0.25   (max possible= 0.877 )
Likelihood ratio test= 3.46  on 1 df,   p=0.06291
Wald test            = 3.29  on 1 df,   p=0.06983
Score (logrank) test = 3.91  on 1 df,   p=0.04788
```

を得る。ビリルビン (の対数値) は、10 %で有意である。

3　時間依存型モデルとイベントヒストリー解析

3．1 時間依存型モデル

表 1 において、例えば、患者＃ 7 は 1514 日目で死亡するまで、7 回来院 (clinic visit) しており、ビリルビン (の対数) 値は、表 3 のように時間と共に変動している。本稿では、このように個体が複数個の観測値を生成している場合を“時間依存型”と呼ぶ。近年、共変量の非線形性を考慮したニューラルネットワークモデル (Tsujitani and Koshimizu,2000;Tsujitani and Sakon,2009,Tsujitani,Iba, Tanaka,2012)、サポートベクターマシン (Tsujitani and Tanaka,2011)、一般化加法モデル (Tsujitani and Baesens,2012;Tsujitani,Tanaka and Sakon,2012; Tsujitani and Tanaka, 2011; wood, 2000, 2006, 2008) などによる解析も考案されている。

表 3　患者＃ 7 に関する時間依存型共変量

Start	Stop	打切り	ln(ビリルビン値)
0	74	0	2.4
74	202	0	2.9
202	346	0	3.0
346	917	0	3.0
917	1411	0	3.9
1411	1514	1	5.1

全ての患者について、最終生存時間の短い順に並べると表 4 のようになる。来院ごとの ln(bil) を示している。最初の死亡例 #281、2 番目の死亡例 #384 に対する部分尤度の計算は表 5 のようになる。最初の死亡例の部分尤度は

$$\text{部分尤度} = \frac{e^{5.0\beta}}{e^{5.0\beta} + e^{2.9\beta} + e^{4.9\beta} + \cdots + e^{3.0\beta}}$$

となる。

表 4　全患者に関する時間依存型共変量

ID	time	δ	Clinic visit					ln(bil)					
8	182	0	90	182	.	.	.	2.4	2.5	2.9	.	.	.
1	281	1	47	184	251	.	.	3.2	3.8	4.9	5.0	.	.
5	341	0	87	192	341	.	.	2.8	2.6	2.9	3.4	.	.
4	384	1	92	194	372	.	.	3.9	4.7	4.9	5.4	.	.
3	457	1	61	97	142	359	440	2.2	2.8	2.9	3.2	3.4	3.8
2	604	0	94	187	321	.	.	3.1	2.9	3.1	3.2	.	.
11	814	1	167	498	.	.	.	3.8	3.9	4.3	.	.	.
6	842	1	94	197	384	795	.	2.4	2.3	2.8	3.5	3.9	.
12	1071	1	108	187	362	694	.	3.1	2.8	3.4	3.9	3.8	.
9	1121	1	101	410	774	1043	.	2.5	2.5	2.7	2.8	3.4	.
10	1411	0	182	847	1051	1347	.	2.3	2.2	2.8	3.3	4.9	.
7	1514	1	74	202	346	917	1411	2.4	2.9	3.0	3.0	3.9	5.1

表 5　最初の死亡例＃ 281 および 2 番目の死亡例＃ 383 に対する部分尤度の計算

$t_{(1)}$=281			$t_{(2)}$=384		
ID	ln(bil)	$exp(x_{(1)}\beta)$	ID	ln(bil)	$exp(x_{(2)}\beta)$
1	5	$exp(5.0\beta)$	4	5.4	$exp(5.4\beta)$
5	2.9	$exp(2.9\beta)$	3	3.4	$exp(3.4\beta)$
4	4.9	$exp(4.9\beta)$	2	3.2	$exp(3.2\beta)$
3	3.2	$exp(3.2\beta)$	11	3.9	$exp(3.9\beta)$
2	3.2	$exp(3.2\beta)$	6	3.9	$exp(3.9\beta)$
11	3.9	$exp(3.9\beta)$	12	3.9	$exp(3.9\beta)$
6	2.8	$exp(2.8\beta)$	9	2.5	$exp(2.5\beta)$
12	3.4	$exp(3.4\beta)$	10	2.2	$exp(2.2\beta)$
9	2.5	$exp(2.5\beta)$	7	3	$exp(3.0\beta)$
10	2.2	$exp(2.2\beta)$			
7	3	$exp(3.0\beta)$			

一般的に尤度は

$$L(\beta) = \prod_{i=1}^{\text{死亡数}} \frac{\Phi(i)}{\sum_{k \in R(t_i)} \Phi_k} = \prod_{i=1}^{\text{死亡数}} \left\{ \frac{\exp\left(\boldsymbol{\beta}^T \boldsymbol{x}_i(t_i)\right)}{\sum_{k \in R(t_i)} \exp\left(\boldsymbol{\beta}^T \boldsymbol{x}_k(\boldsymbol{t_i})\right)} \right\}^{\delta_i}$$
$$= \prod_{i=1}^{n} \left\{ \frac{\exp\left(\boldsymbol{\beta}^T \boldsymbol{x}_i(t_i)\right)}{\sum_{k=1}^{n} Y_k(t_i) \exp\left(\boldsymbol{\beta}^T \boldsymbol{x}_k(t_i)\right)} \right\} ; Y_k = \begin{cases} 1 : k \in R(t) \\ 0 : o.w. \end{cases} \tag{15}$$

から計算される (Lawless,2003,7.1.8 節)。ここに、$\boldsymbol{x}_i(t_i)$は、すべての共変量の値である。

R プログラムは

```
>library(survival)
>train<-read.csv("F:\\train.csv",header=FALSE)
>fit<-coxph(Surv(V1,V2,V3==1)~V4,data=train)
>summary(fit)
```

となり、解析結果

```
     coef exp(coef) se(coef)     z Pr(>|z|)
V4  3.299    27.092    1.734 1.903    0.057 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

   exp(coef) exp(-coef) lower .95 upper .95
V4     27.09    0.03691    0.9062       810
Rsquare= 0.215   (max possible= 0.372 )
Likelihood ratio test= 13.07  on 1 df,   p=0.0002998
Wald test            = 3.62  on 1 df,   p=0.05702
Score (logrank) test = 12.17  on 1 df,   p=0.000486
```

を得る。ビリルビン（の対数値）は、尤度比検定 (Likelihood ration test)、およびスコア検定 (Score test) で 1 %有意となる。

3．2 イベントヒストリー解析

時間依存型の特殊なケースとして、multistate モデルに基づくイベントヒストリー解析がある（赤澤ら，2010，第 9 章）。最も単純な場面は、図 1 のように表せる。観測開始から目標事象（死亡あるいは打切り）発生までに種々のイベント（すなわち，multistate) が起こり、その時間と共変量の値が記録される。

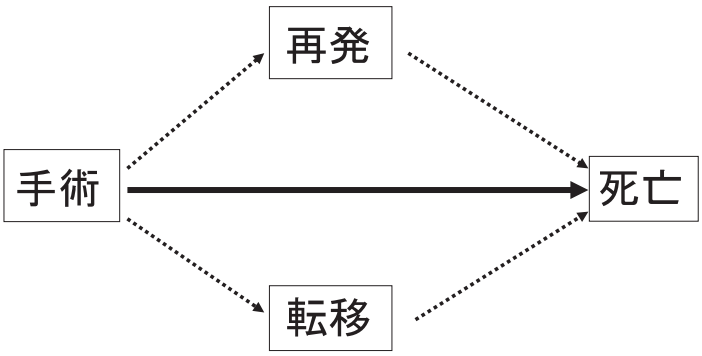

図 1　multistate モデル

膀胱癌データについて、患者 #15，43，86 に関する実際のデータを表 6 に示す。

表 6　患者 #15，43，86 に関する実際のデータ

患者#	追跡時間	個数 (初期値)	サイズ (初期値)	処置	再 1	発 2	時 3	間 4
#15	25	3	1	1	3	15	25	
#43	53	1	3	1	3	15	46	51
#86	59	1	3	2				

(1)Anderson-Gill モデル

Anderson-Gill モデルでは、表 7 のような時間依存型で表す。Num,Size,Treat は時間と共に変動し

ても良い。例えば、患者 #43 は、時間 3,15,46,51 で再発し、観測終了の 53 時間で打切りとなった。

表 7 Anderson-Gill モデル

患者#	Start	Stop	Cens	Num	Size	Treat
・	・	・	・	・	・	・
15	0	3	1	1	1	1
15	3	15	1	1	1	1
15	15	25	1	1	1	1
・	・	・	・	・	・	・
43	0	3	1	3	1	1
43	3	15	1	3	1	1
43	15	46	1	3	1	1
43	46	51	1	3	1	1
43	51	53	0	3	1	1
・	・	・	・	・	・	・

R プログラムは

```
>library(survival)
>train<-read.csv("F:\\train.csv",header=FALSE)
>kfit<-coxph(Surv(V1,V2,V3)~V4+V5+factor(V6),data=train,method="breslow")
>summary(kfit)
```

となり、解析結果

```
                coef exp(coef) se(coef)      z Pr(>|z|)
V4           0.17164   1.18726  0.04733  3.627 0.000287 ***
V5          -0.04256   0.95833  0.06903 -0.617 0.537528
factor(V6)2 -0.45979   0.63142  0.19996 -2.299 0.021481 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

            exp(coef) exp(-coef) lower .95 upper .95
V4             1.1873     0.8423    1.0821    1.3027
V5             0.9583     1.0435    0.8371    1.0972
factor(V6)2    0.6314     1.5837    0.4267    0.9344
Rsquare= 0.09   (max possible= 0.994 )
Likelihood ratio test= 16.77  on 3 df,   p=0.000787
Wald test            = 18.21  on 3 df,   p=0.0003984
Score (logrank) test = 18.57  on 3 df,   p=0.0003355
```

を得る。よって、Num と Treat が高度に有意となる。

(2)PWP モデル

PWP (Prentice,Williams, and Peterson,1981) モデルは、層別解析の援用である。来院ごとに層別し、データを表 8 のように並び替える。

表 8　PWP モデル

ID#	Start	Stop	Cens	Num	Size	Treat	Clinic visit
1	0	1	0	1	3	1	1
・	・	・	・	・	・	・	・
43	0	3	1	1	3	1	1
・	・	・	・	・	・	・	・
85	0	59	0	1	3	2	1
5	6	10	0	4	1	1	2
・	・	・	・	・	・	・	・
43	3	15	1	1	3	1	2
・	・	・	・	・	・	・	・
84	38	54	0	2	1	2	2
・	・	・	・	・	・	・	・
25	12	30	0	2	1	1	5
・	・	・	・	・	・	・	・
43	51	53	0	1	3	1	5
・	・	・	・	・	・	・	・
76	27	44	0	6	1	2	5

同表から分るように、来院ごとに層別されており、その層の中では、観測値は独立になり、対数尤度を

$$\begin{aligned}\ln L = &\ln L_1(\beta) : Clinic\ visit\ 1\\ &+\ln L_2(\beta) : Clinic\ visit\ 2\\ &+\ln L_3(\beta) : Clinic\ visit\ 3\\ &+\ln L_4(\beta) : Clinic\ visit\ 4\\ &+\ln L_5(\beta) : Clinic\ visit\ 5\end{aligned}$$

と分解できる。

R プログラムは

```
>library(survival)
>train<-read.csv("F:\\train.csv",header=FALSE)
>kfit<-coxph(Surv(V1,V2,V3)~V4+V5+factor(V6)+strata(V7),data=train,method="breslow")
>summary(kfit)
```

となり、解析結果

```
                  coef exp(coef)  se(coef)      z Pr(>|z|)
V4            0.115653  1.122606  0.053681  2.154   0.0312 *
V5           -0.008051  0.991982  0.072725 -0.111   0.9119
factor(V6)2 -0.334295  0.715842  0.216087 -1.547   0.1219
---
Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ' '  1
            exp(coef) exp(-coef) lower .95 upper .95
V4             1.1226     0.8908    1.0105     1.247
V5             0.9920     1.0081    0.8602     1.144
factor(V6)2    0.7158     1.3970    0.4687     1.093
Rsquare= 0.034   (max possible= 0.973 )
Likelihood ratio test= 6.11  on 3 df,   p=0.1062
Wald test            = 6.41  on 3 df,   p=0.0934
Score (logrank) test = 6.45  on 3 df,   p=0.09152
```

を得る。

(3)PWP gap モデル

ＰＷＰモデルで生存時間を時間依存型で与えたが、それをギャップ time=stop-start で置換えたモデルである。R プログラムは

```
>library(survival)
>train<-read.csv("F:\\train.csv",header=FALSE)
>kfit<-coxph(Surv(V1,V2)~V3+V4+factor(V5)+strata(V6),data=train,method="breslow")
>summary(kfit)
```

となり、解析結果

```
                coef exp(coef) se(coef)      z Pr(>|z|)
V3           0.15353   1.16595  0.05211  2.947  0.00321 **
V4           0.00684   1.00686  0.07001  0.098  0.92217
factor(V5)2 -0.26952   0.76375  0.20766 -1.298  0.19433
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

            exp(coef) exp(-coef) lower .95 upper .95
V3             1.1659     0.8577    1.0528     1.291
V4             1.0069     0.9932    0.8778     1.155
factor(V5)2    0.7637     1.3093    0.5084     1.147

Rsquare= 0.048   (max possible= 0.984 )
Likelihood ratio test= 8.76  on 3 df,   p=0.03272
Wald test            = 9.46  on 3 df,   p=0.02379
Score (logrank) test = 9.6  on 3 df,   p=0.02231
```

を得る。

4　競合リスクモデル

例えば、喫煙習慣と肺癌の疫学調査において、死因として肺癌，その他の癌，その他の疾患，事故が考えられる。このとき、肺癌がエンドポイントで、その他の癌, その他の疾患, 事故を競合リスクという。4 つの死因は (確率的に) 独立ではない。すなわち、肺癌のリスクの高い人は、他の癌のリスクも高い。打切りのタイプとして

1) 研究 (治験) の終了：打切りと死亡は独立

2) 追跡不能

$$\begin{cases} \text{生存関数を過小推定 (download bias):患者自身で完治したと判断して治療打切り} \\ \text{生存時間を過大評価 (upload bias):余命が少ないため、故郷へ帰郷} \end{cases}$$

3) 競合リスク：研究対象以外の死因が発生

がある (Putter et al., 2007)。

表 9 は、マウス発癌性データ (単位：日数) である (Kalbfleisch et al., 2002,pp.257-259)。死因とし

て、胸腺リンパ腫、網状組織細胞肉腫、その他がある。共変量は 2 値（対照群と，無菌群）である。

表 9　マウス発癌性データ

	死因	(300rad の放射後)	
共変量	胸腺リンパ腫	網状組織細胞肉腫	その他
対照群	159,189,191,. . . ,428,432	317,318,399,. . . , 748,753	40,42,51,. . . ,761,763
無菌群	158,192,193,. . . ,707,800	430,590,606,. . . , 821,986	136,246,. . . , 1015,1019

4．1 周辺モデル

表 9 において、周辺モデルでは胸腺リンパ腫を注目しているイベントとし、他の 2 つの死因を打切りとして比例ハザードモデルを適用する。R プログラムは

```
>library(survival)
>train<-read.csv("F:\\train.csv",header=FALSE)
>kfit<-coxph(Surv(V1,V2)~V3,data=train,method="breslow")
>summary(kfit)
```

となり、解析結果

```
     coef exp(coef) se(coef)     z Pr(>|z|)
V3 0.3019    1.3524   0.2866 1.053    0.292
   exp(coef) exp(-coef) lower .95 upper .95
V3     1.352     0.7394    0.7712     2.371

Rsquare= 0.006   (max possible= 0.935 )
Likelihood ratio test= 1.12  on 1 df,   p=0.2903
Wald test            = 1.11  on 1 df,   p=0.2921
Score (logrank) test = 1.12  on 1 df,   p=0.2903
```

を得る。3 つの死因別の結果を表 10 に与えておく。注目しているイベントを胸腺リンパ腫とした場合のみ、対照群と無菌群の有意差はない。

表 10　死因別の結果

死因	$\hat{\beta}_j$	S.E.$[\hat{\beta}_j]$	p 値
胸腺リンパ腫	0.302	0.287	0.29
網状組織細胞肉腫	-2.030	0.354	$<< 0.01$
その他	-1.107	0.304	< 0.01

これら 3 種類のイベントごとの累積ハザードを求めると図 2 〜図 4 のようになる。

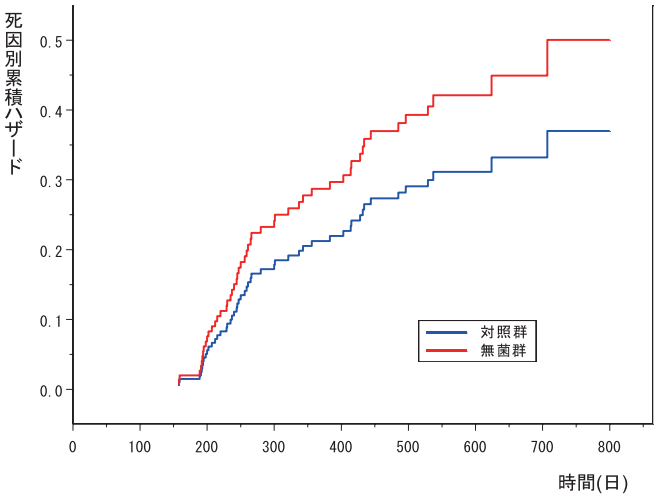

図２　胸腺リンパ腫

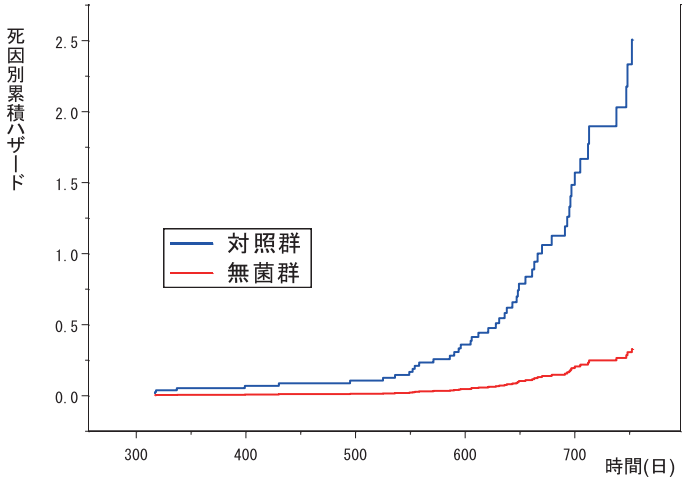

図３　網状組織細胞肉腫

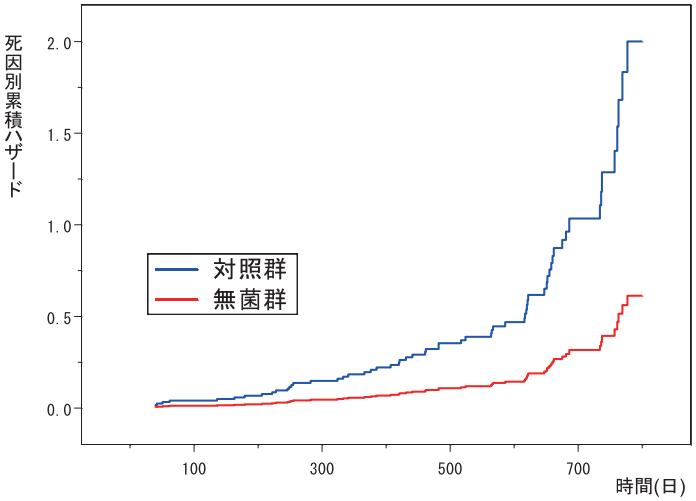

図４　その他

4．2 層別比例ハザードモデル

表 11 は、骨髄腫症データ (Allison,1995,p.31) である。DUR は生存時間、$STATUS = \begin{cases} 0 : 打切り \\ 1 : 死亡 \end{cases}$ で、共変量として薬剤に関する $TREAT = \begin{cases} 1 : 薬剤\ A\ 投与 \\ 2 : 薬剤\ B\ 投与 \end{cases}$ を取上げる。

表 11　骨髄腫症データ

DUR	STATUS	TREAT
8	1	1
180	1	2
.	.	.
852	0	1
.	.	.
23	1	2

生存時間 (dur) と打切り (status) について、比例ハザードモデル

$$h\left(t\,|\boldsymbol{x}\right) = h_0\left(t\right)\exp\left(\beta x\right)$$

を当てはめると

```
       coef   exp(coef) se(coef)  z    Pr(>|z|)
TREAT 0.5611    1.7525   0.5098 1.101    0.271


        exp(coef) exp(-coef) lower .95 upper .95
TREAT    1.753     0.5706    0.6453      4.76
Rsquare= 0.049   (max possible= 0.977 )
Likelihood ratio test= 1.26  on 1 df,   p=0.2609
Wald test            = 1.21  on 1 df,   p=0.2710
Score (logrank) test = 1.24  on 1 df,   p=0.2650
```

となり、共変量 (薬剤) は有意にならない。

しかし、患者に陣機能障害があるか否かによってハザード関数の形状が異なることもある。そこで、層別比例ハザードモデル

$$h\left(t\,|\boldsymbol{x}\right) = \begin{cases} h_{01}\left(t\right)\exp(\boldsymbol{\beta}^T\boldsymbol{x}) \\ h_{02}\left(t\right)\exp(\boldsymbol{\beta}^T\boldsymbol{x}) \end{cases} \tag{16}$$

を採用する。m 個の層がある場合、一般に

$$h_j\left(t\,|\boldsymbol{x}\right) = h_{j0}\left(t\right)\exp\left(\boldsymbol{\beta}^T\boldsymbol{x}\right), j = 1, 2, ..., m \tag{17}$$

と書け、ベースラインハザード関数は j (層) ごとに異なるが、同一の相対リスク関数 $\exp\left(\boldsymbol{\beta}^T x\right)$ をもつ比例ハザードモデルである。

層ごとに求めた部分尤度の積の対数

$$\ln L = \underset{\text{腎機能障害}}{\ln L_1\left(\beta\right)} + \underset{\text{正常}}{\ln L_2\left(\beta\right)} \tag{18}$$

を最大にする β が ML 推定量である。データは、表 12 のように与えられる。ここに、$RENAL = \begin{cases} 1 : \text{腎機能障害} \\ 2 : \text{正常} \end{cases}$
とする。部分尤度の計算は表 13 で与えられる。

表 12　腎機能障害で層別された骨髄腫症データ

DUR	STATUS	TREAT	RENAL
8	1	1	2
180	1	2	1
.	.	.	.
852	0	1	1
.	.	.	.
23	1	2	2

表 13　部分尤度の計算

生存時間	打切り (=0)	TREAT	尤度	生存時間	打切り (=0)	TREAT	尤度
8	1	0	$1/(8+10e^{\beta})$	8	1	0	$1/(4+3e^{\beta})$
70	1	1	$e^{\beta}/(8+9e^{\beta})$	13	1	1	$e\beta/(3+3e^{\beta})$
76	1	1	$e^{\beta}/(8+8e^{\beta})$	18	1	1	$e^{\beta}/(3+2e^{\beta})$
.	.	.	.	23	1	1	$e^{\beta}/(3+e^{\beta})$
365	0	0	.	52	1	0	$1/3$
632	1	1	$e^{\beta}/(5+5e^{\beta})$	63	1	0	$1/2$
.	.	.	.	63	1	0	1
2240	0	0	.				

R プログラム

```
>library(survival)
>train<-read.csv("F:\\train.csv",header=TRUE)
>kfit<-coxph(Surv(DUR,STATUS)~TREAT+strata(RENAL),data=train)
summary(kfit)
```

を用いると

```
       coef    exp(coef) se(coef)    z    Pr(>|z|)
TREAT 1.4640    4.3232   0.6596   2.219   0.0265 *
---
Signif. codes:  0  ‘***’  0.001  ‘**’  0.01  ‘*’  0.05  ‘.’  0.1  ‘ ’  1

      exp(coef) exp(-coef) lower .95 upper .95
TREAT     4.323     0.2313     1.187     15.75

Rsquare= 0.216   (max possible= 0.933 )
Likelihood ratio test= 6.07  on 1 df,   p=0.01372
Wald test            = 4.93  on 1 df,   p=0.02646
Score (logrank) test = 5.79  on 1 df,   p=0.01611
```

が得られる。層別比例ハザードモデルを採用すると薬剤は有意になる。

次に、共変量が連続値である表 14 のデータを解析する。腎障害，骨病変，タンパク尿などを伴わない良性単クローン性 γ グロブリン血症 (MGUS) に関するデータで、一部が多発性骨髄腫に移行するため，定期的な経過観察が必要である (Therneau et. al., 2000,8.4.1 節)。共変量として、性別、年齢、ヘモグロビンレベル (hgb)、単クローン性蛋白ピーク (mspike) をもつ。死因としては、MGUS, 骨髄腫、その他の 3 つがある。この死因 (競合リスク) を層別因子とし、表 15 のように書き換える。これは、

$$\ln L = \underset{MGUS}{\ln L_1(\beta)} + \underset{\text{骨髄腫}}{\ln L_2(\beta)} + \underset{\text{その他}}{\ln L_3(\beta)} \tag{19}$$

と書ける。通常の層別比例ハザードモデルとの相違点は、患者が必ず層別されたすべての対数尤度に入ることである (通常の層別比例ハザードモデルでは、患者を層ごとに分類する)。

表 14　良性単クローン性 γ グロブリン血症

id	time	Endpoint (status)	age	sex	hgb	mspike
1	760	MGUS:死亡 (1)	79	2	11.5	2.0
2	2160	その他 (1)	76	2	13.3	1.8
3	277	MGUS:死亡 (1)	87	1	11.2	1.3
・	・	・	・	・	・	・
14	7807	打切り (0)	66	1	15.3	1.9
・	・	・	・	・	・	・
17	3590	骨髄腫 (1)	53	2	11.1	2.0
・	・	・	・	・	・	・

表 15　競合リスクを層別因子

id	time	status	age	sex	hgb	mspike	endpoint
1	760	1	79	2	11.5	2.0	MGUS
1	760	0	79	2	11.5	2.0	骨髄腫
1	760	0	79	2	11.5	2.0	その他
・	・		・	・	・	・	・
17	3590	0	53	2	11.1	2.0	MGUS
17	3590	1	53	2	11.1	2.0	骨髄腫
17	3590	0	53	2	11.1	2.0	その他
・	・		・	・	・	・	・

R プログラムは、

```
>library(survival)
>train<-read.csv("F:\\train.csv",header=FALSE)
>kfit<-coxph(Surv(V1,V2)~V3+V4+V5+V6+strata(V7),data=train,method="breslow")
>summary(kfit)
```

となり、解析結果

```
        coef exp(coef)  se(coef)      z Pr(>|z|)
V3  0.051574  1.052927  0.007316  7.050 1.79e-12 ***
V4 -0.348884  0.705475  0.156116 -2.235  0.02543 *
V5 -0.166800  0.846369  0.044188 -3.775  0.00016 ***
V6 -0.094699  0.909646  0.186743 -0.507  0.61208
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

   exp(coef) exp(-coef) lower .95 upper .95
V3    1.0529     0.9497    1.0379     1.068
V4    0.7055     1.4175    0.5195     0.958
V5    0.8464     1.1815    0.7762     0.923
V6    0.9096     1.0993    0.6308     1.312

Rsquare= 0.102   (max possible= 0.919 )
Likelihood ratio test= 76.76  on 4 df,   p=8.882e-16
Wald test            = 74.74  on 4 df,   p=2.220e-15
Score (logrank) test = 76.68  on 4 df,   p=8.882e-16
```

を得、死因（競合リスク）をすべてイベント (status=1) として解析した結果と同一になる。

4．3 累積発生関数

(1) ハザード関数

m 個の競合リスクの中で、タイプ $k(=1,2,...,m)$ のイベントに対する累積発生関数 CIF (Cumulative Incidence Function) は

$$F_k(t) = \Pr\{T \le t, C = k\} \tag{20}$$

と定義される。すなわち、時間 t 以前に、タイプ k のイベントが発生する同時確率である。累積発生関数は、部分分布 (subdistribution) とも呼ばれている。競合リスクイベント $k(=1,2,...,m)$ は互いに独立である必要はない。

時間 t 以前に、いずれかのイベントが起こる確率は、(20) 式から

$$F(t) = \Pr\{T \le t\} = \sum_{k=1}^{m} \Pr\{T \le t, C = k\} = \sum_{k=1}^{m} F_k(t) \tag{21}$$

となる。時間 t までにタイプ k のイベントが起こらない確率（部分生存関数という）は、(20) 式から

$$S_k(t) = \Pr\{T > t, C = k\} \tag{22}$$

で与えられる。タイプ k のイベントに対する部分密度 (subdensity) 関数は、(20) 式から

$$f_k(t) = \frac{\partial F_k(t)}{\partial t} = \lim_{\Delta t \to 0} \frac{\Pr\{t < T \le t + \Delta t, C = k\}}{\Delta t} \tag{23}$$

となる。

(23) 式から、死因別 (cause-specific) ハザード関数を

$$h_k(t) = \lim_{\Delta t \to 0} \frac{\Pr\{t < T \le t + \Delta t, C = k \,|T > t\}}{\Delta t} \tag{24}$$

と定義する。(24) 式から $h_k(t)$ と $f_k(t), S(t)$ との間には (7) 式と類似の関係

$$\begin{aligned} h_k(t) &= \lim_{\Delta t \to 0} \frac{\Pr\{t<T\leq t+\Delta t, C=k | T>t\}}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{\Pr\{t<T\leq t+\Delta t, C=k\}}{\Delta t \Pr\{T>t\}} \\ &= [\Pr\{T>t\}]^{-1} \lim_{\Delta t \to 0} \frac{\Pr\{t<T\leq t+\Delta t, C=k\}}{\Delta t} \\ &= \frac{f_k(t)}{S(t)} \neq \frac{f_k(t)}{S_k(t)} \end{aligned} \tag{25}$$

$$h_k(t) \neq \frac{f_k(t)}{S_k(t)} \tag{26}$$

があることに留意されたい。死因別ハザード関数に基づくモデル

$$h_k(t|\boldsymbol{x}) = h_{k0}(t) exp\left(\boldsymbol{\beta}_k^T \boldsymbol{x}\right), k = 1, 2, ..., m \tag{27}$$

が、4.1 節の周辺モデルである。(24) 式から

$$\begin{aligned} &\text{全 } (overall) \text{ 死因別ハザード関数：} h(t) = \sum_{k=1}^{m} h_k(t) \\ &\text{累積死因別ハザード関数：} H_k(t) = \int_0^t h_k(x)dx = \int_0^t \left\{\frac{f_k(x)}{S(x)}\right\}dx \\ &\text{全累積死因別ハザード関数：} H_k(t) = \sum_{k=1}^{m} H_k(t) \end{aligned} \tag{28}$$

と定義する。(20),(25) 式から

$$\begin{aligned} F_k(t) &= \Pr\{T \leq t, C=k\} \\ &= \int_0^t f_k(x)dx = \int_0^t h_k(x) S(x) dx \end{aligned} \tag{29}$$

となり、$h_k(t)$ と $S(t)$ を推定すれば、$F_k(t)$ が得られる。ちなみに

$$1 - S_k(t) = \int_0^t h_k(x) S_k(x) dx \neq F_k(t)$$

であることに留意されたい。

全死因での確率密度関数

$$\begin{aligned} f(t) &= \lim_{\Delta t \to 0} \frac{\Pr\{t<T\leq t+\Delta t\}}{\Delta t} \\ &= h(t) S(t) \\ &= \{h_1(t) + h_2(t) + \cdots + h_m(t)\} S(t) \\ &= f_1(t) + f_2(t) + \cdots + f_m(t) \end{aligned} \tag{30}$$

から

$$\sum_{k=1}^{m} F_k(t) = \sum_{k=1}^{m} \int_0^t f_k(x)dx = \int_0^t \sum_{k=1}^{m} f_k(x)dx = \int_0^t f(x)dx = 1 - S(x) \tag{31}$$

を得る。$F_k(t) < 1$ より、$F_k(t)$ は分布関数にならない。そのため部分分布と呼ぶ。

部分分布のハザード関数を

$$\begin{aligned} \gamma_k(t) &= \lim_{\Delta t \to 0} \frac{\Pr\{t<T\leq t+\Delta t, C=k | (T>t) \cup (T\leq t \cap C\neq k)\}}{\Delta t} \\ &= \frac{\partial \ln\{1-F_k(t)\}}{\partial t} \\ &= \frac{f_k(t)}{1-F_k(t)} \end{aligned} \tag{32}$$

と定義する (Gray,1988;Pintilie,2006,p.46)。注目しているイベントが、その時間までに起こらないか、あるいは競合リスクイベントが観測された条件のもとで、注目しているイベントが、次の時間 Δt に起こる確率を単位時間当りの量に換算し、$\Delta t \to 0$ としたときの極限値が部分分布のハザード関数である。競合リスクが無ければ (注目しているイベントのみ)、死因別ハザードと部分分布のハザードは同一に

なる。

(2) 死因別ハザード関数に基づく CIF のノンパラメトリック推定

イベントが発生した時間を $t_1, t_2, \cdots, t_r (t_1 < t_2 < \cdots < t_r)$ としたとき、時間 t_j で起きたタイプ k のイベント数を d_{kj}、t_j でのリスク集合の個体数を n_j、時間 t までにいかなるイベントも起きない確率の Kaplan-Meier 推定量を $\hat{S}(t)$ とする。t_j の直前までにはいかなるイベントも起こらず、t_j でタイプ k のイベントを経験する同時確率である CIF は

$$\hat{F}_k(t) = \sum_{\text{すべての } j, t_j \leq t} \hat{h}_{kj} \hat{S}(t_{j-1}) \tag{33}$$

から推定される (Pintilie,2006,4.2.1 節)。ここに、$\hat{h}_{kj}$ は、時間 t_j でのイベント k に対する死因別ハザード関数の推定量である。

区間 $[t_{j-1}, t_j)$ で C_j 個の打切りが $t_{j1} < t_{j2} < \cdots < t_{jC_j}$ で起きたとき、尤度関数は

$$\begin{aligned} L &= \prod_{j=1}^{r} \prod_{k=1}^{m} \{F_i(t_j) - F_i(t_{j-1})\}^{d_{kj}} \left[\prod_{j=1}^{r+1} \prod_{l=1}^{C_j} S(t_{jl}) \right] \\ &= \prod_{j=1}^{r} \prod_{k=1}^{m} h_{kj}^{d_{kj}} (1 - h_j)^{n_j - d_j} \end{aligned} \tag{34}$$

で与えられる (Kalbfleisch&Prentice,2002,(8.9),(8.10) 式)。ただし、h_k は時間 t_j における任意のタイプのイベントに対するハザード関数で、$d_j = \sum_{k=1}^{m} d_{kj}$ とする。(34) 式から

$$\begin{aligned} \ln L &= \sum_{j=1}^{r} \sum_{k=1}^{m} \{d_{kj} \ln h_{kj} + (n_j - d_j) \ln(1 - h_j)\} \\ &= \sum_{j=1}^{r} \sum_{k=1}^{m} \left\{ d_{kj} \ln h_{kj} + (n_j - d_j) \ln \left(1 - \sum_{k=1}^{m} h_{kj} \right) \right\} \end{aligned} \tag{35}$$

を得る。よって

$$\frac{\partial \ln L}{\partial h_{kj}} = \frac{d_{kj}}{h_{kj}} + (n_j - d_j) \frac{(-1)}{1 - h_j} = 0$$

より

$$\frac{d_{kj}}{h_{kj}} = \frac{n_j - d_j}{1 - h_j}$$

を得、

$$h_{kj} = \frac{d_{kj}(1 - h_j)}{n_j - d_j} \tag{36}$$

が求まる。$h_j = \sum_{k=1}^{m} h_{kj}$、$d_j = \sum_{k=1}^{m} d_{kj}$ より、(36) 式は

$$h_j = \sum_{k=1}^{m} h_{kj} = \frac{\sum_{k=1}^{m} d_{kj}(1 - h_j)}{n_j - d_j} = \frac{d_j(1 - h_j)}{n_j - d_j} \tag{37}$$

となる。$\hat{h}_j = d_j / n_j$ より、(36)、(37) 式を用いると

$$\hat{h}_{kj} = \frac{d_{kj}}{n_j} \tag{38}$$

を得る。(33) 式へ (38) 式を代入すると、$F_k(t)$ は

$$\hat{F}_k(t) = \sum_{t_j \leq t} \frac{d_{kj}}{n_j} \hat{S}(t_{j-1}) \tag{39}$$

から推定される。ここに、$\hat{S}(t_{j-1})$ は、すべてのイベントを同じタイプと見なして得られる Kaplan-Meier 推定量である。

(39) 式から

$$\begin{aligned}&\widehat{Var}\left[\hat{F}_k(t)\right]\\&=\sum_{t_j\le t}\left\{\left[\hat{F}_k(t)-\hat{F}_k(t_j)\right]^2\frac{d_j}{(n_j-1)(n_j-d_j)}\right\}+\sum_{t_j\le t}\hat{S}(t_{j-1})^2\frac{d_{kj}(n_j-d_j)}{n_j^2(n_j-1)}\\&-2\sum_{t_j\le t}\left[\hat{F}_j(t)-\hat{F}_k(t_j)\right]\hat{S}(t_{j-1})\frac{d_{kj}(n_j-d_{kj})}{n_j(n_j-d_j)(n_j-1)}\end{aligned}\tag{40}$$

を得る (Aalen,1978;Pintilie,2006,4.2.4 節)。よって、$\hat{F}_k$ の $100(1-\alpha)\%$ 信頼区間は

$$\hat{F}_k(t)\pm Z_{1-\alpha/2}\sqrt{\widehat{Var}\left[\hat{F}_k(t)\right]}\tag{41}$$

で与えられる。ここに、$Z_{1-\alpha/2}$ は標準正規分布の $(1-\alpha/2)\%$ 点である。

計算手順を示すため、数値例として表 16 のデータを取上げる。ただし、イベントのタイプは、0= 打切り、1= 注目しているイベント、2= 競合リスクイベントとする。$\hat{S}(t),\hat{F}_1(t),\hat{F}_2(t)$ は、表 17 のように算出され、$1-\hat{S}(t)=\hat{F}_1(t)+\hat{F}_2(t)$ となることを確認できる。

表 16　数値例

ID	1	2	3	4	5	6	7	8	9	10
最初のイベントが起きた時間	130	12	203	67	160	145	22	89	12	203
イベントのタイプ	2	0	0	1	1	0	2	0	1	0

表 17 計算手順

発現時間	イベントのタイプ	リスク集合	$\hat{S}(t)$	$\hat{F}_1(t)$	$\hat{F}_2(t)$
12	0	10	1.00	0.00	0.00
22	2	9	1.00×(1-1/9)=0.89	0.00	0.00+1×1/9=0.11
45	1	8	0.89×(1-1/8)=0.78	0.00+0.89×1/8=0.11	0.11
67	1	7	0.78×(1-1/7)=0.67	0.11+0.78×1/7=0.22	0.11
89	0	6	0.67	0.222	0.11
112	1	5	0.67×(1-1/5)=0.53	0.22+0.67×1/5=0.36	0.11
130	2	4	0.53×(1-1/4)=0.40	0.36	0.11+0.53×1/4=0.24
145	0	3	0.40	0.36	0.24
160	1	2	0.40×(1-1/2)=0.20	0.36+0.40×1/2=0.56	0.24
203	0	1	0.20	0.56	0.24

R プログラム (Gray,2010) は

```
>library(cmprsk)
>train <- read.csv("G:\\train.csv",header=FALSE)
>ss<-train$V1 #この行が必要
>cc<-train$V2 #この行が必要
># 打ち切りは 0、死亡のみの場合は、1,2,3,... とし 0 は用いない
>fit<-cuminc(ss,cc)
>fit2<-timepoints(fit,ss) #点 (日) ごとに打ち出し
>fit2$"est"
>plot(fit,xlab="Day",ylab="Probability",
curvlab=c("Event","Competing risk"))
```

と書け、図 5 の CIF が得られる。

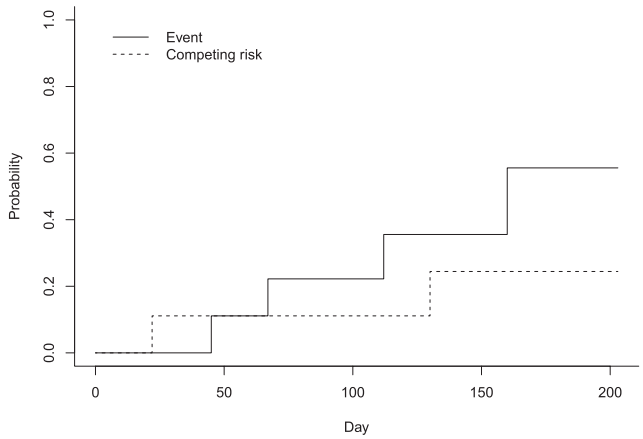

図 5　CIF の推定

ちなみに、競合リスクを従来のように打切りとして Kaplan-Meier 推定を行うと表 18 のようになる。図 5 に表 18 の $1-\hat{S}(t)$ を加えると図 6 のようになる。同図から競合リスクを従来のように打切りとした K-M 推定はナイーブになっていることが分かる。

表 18　競合リスクを打切りとした Kaplan-Meier 推定

発現時間	イベントのタイプ	リスク集合	$\hat{S}(t)$	$\hat{S}(t)$
12	0	10	1	0
22	0	10	1	0
45	1	9	1×(1-1/8)=0.875	0.125
67	1	8	0.875×(1-1/7)=0.75	0.25
89	0	8	0.75	0.25
112	1	5	0.75×(1-1/5)=0.60	0.4
130	0	4	0.6	0.4
145	0	3	0.6	0.4
160	1	2	0.60×(1-1/2)=0.30	0.7
203	0	1	0.3	0.7

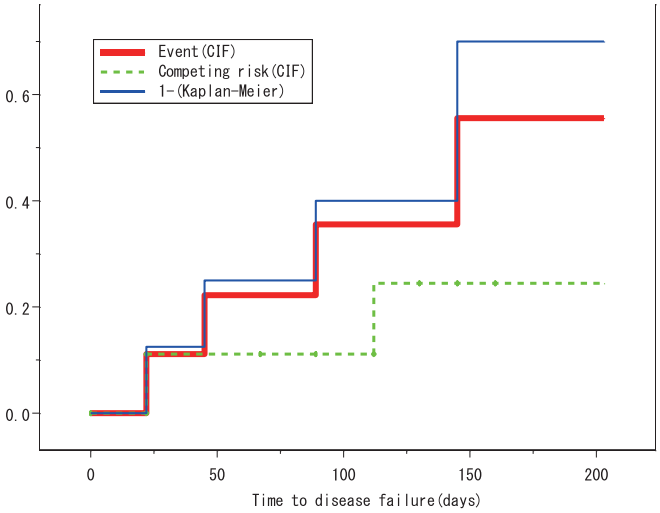

図 6　図 5 に表 18 の$1 - \hat{S}(t)$を加えた CIF

次に、表 9 のマウス発癌性データを解析する。R プログラムは

```
>library(cmprsk)
>train <- read.csv("F:\\train.csv",header=FALSE)
>ss<-train$V1
>gg<-train$V3
>cc<-train$V2      # 打ち切りは 0、死亡のみの場合は、1,2,3,... とし 0 は用いない
>xx<-cuminc(ss,cc,gg)
>plot(xx,lty=1,color=1:6)
>#死因ごとの CIF をプロットする場合、以下の命令を追加
>xx1=list(list(xx$'0 1'$time,xx$'0 1'$est),list(xx$'1 1'$time,xx$'1 1'$est))
>plot.cuminc(xx1,curvlab=c('Germ-free group','Control group'),lty=c(1,2),
xlab='Time to disease failure(days)',ylab='Cumulative incidence')
>title(main='CIF for thymic lymphoma')
>text(0,0.8,adj=0,paste("Gray's test:p-value=",round(xx$Tests[1,2],3)))
>xx2=list(list(xx$'0 2'$time,xx$'0 2'$est),list(xx$'1 2'$time,xx$'1 2'$est))
>plot.cuminc(xx2,curvlab=c('Germ-free group','Control group'),lty=c(1,2),
xlab='Time to disease failure(days)',ylab='Cumulative incidence')
>title(main='CIF for reticulum cell sarcoma')
>text(0,0.8,adj=0,paste("Gray's test:p-value=",round(xx$Tests[2,2],3)))
>xx3=list(list(xx$'0 3'$time,xx$'0 3'$est),list(xx$'1 3'$time,xx$'1 3'$est))
>plot.cuminc(xx3,curvlab=c('Germ-free group','Control group'),lty=c(1,2),
xlab='Time to disease failure(days)',ylab='Cumulative incidence')
>title(main='CIF for other cause')
>text(0,0.8,adj=0,paste("Gray's test:p-value=",round(xx$Tests[3,2],3)))
```

と書ける。イベントごとの CIF は図 7 〜 図 9 のようになる。

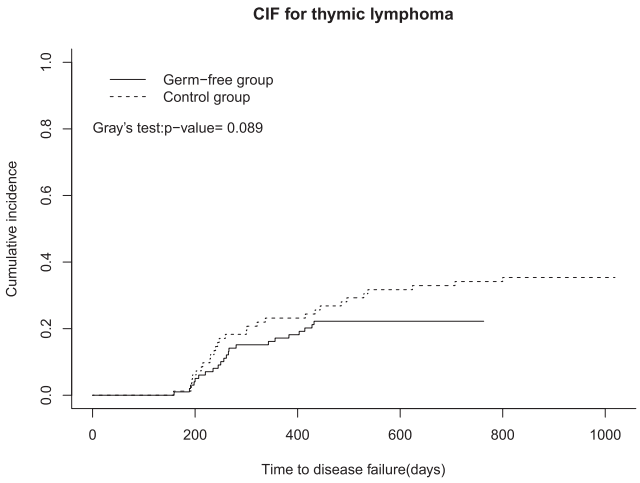

図 7　胸腺リンパ腫

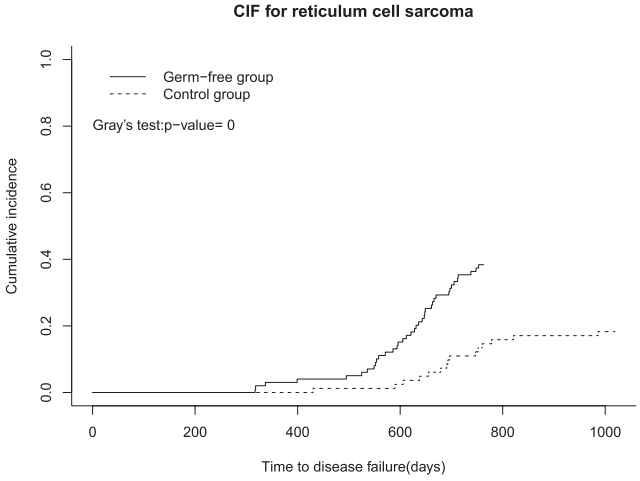

図 8　網状組織細胞肉腫

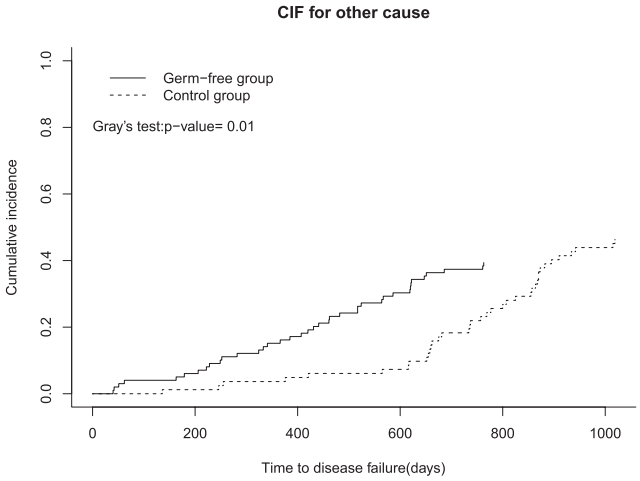

図 9　その他

(3)CIF に関する 2 群間の差の Gray 検定

Gray(1988) は CIF に関する 2 群間の差を検定するため、部分分布ハザードに関する重み付き平均

$$z_i = \int_0^\tau W_i(t)\{\gamma_i(t) - \gamma_0(t)\}dt \tag{42}$$

を採用した。ここに、

τ：すべての群の中の最大生存時間
$\gamma_i(t)$：群 i の部分分布ハザード
$\gamma_0(t)$：すべての群を合併した部分分布ハザード

とする。

2 群での生存時間を $t_1 < t_2 < \cdots < t_n$ 、時間 t_j における群 $k(=1,2)$ の注目しているイベント数を d_{kj} 、時間 t_j における群 $k(=1,2)$ のリスク集合の大きさを n_{kj} とする。

手順 1 群 k について、時間 t_{j-1} における注目しているイベントの CIF 推定量 $\hat{F}_k(t_{j-1})$ を算出する。

手順 2 群 k について、時間 t_{j-1} における注目しているイベントあるいは競合リスクを共にイベントとした Kaplan-Meier 推定量 $\hat{S}_k(t_{j-1}) :, t_{j-1}$ を算出する。

手順 3 $n_{1j} : R_{1j} = \hat{S}_1(t_{j-1}) : \left\{1 - \hat{F}(t_{j-1})\right\}$ から、修正されたリスク集合の大きさを $R_{kj} = n_{1j}\frac{1-\hat{F}_k(t_{j-1})}{\hat{S}_k(t_{j-1})}, k = 1,2$ とする。

手順 4 スコア $z_1 = \sum_{\text{すべての } t_j} R_{1j}\left(\frac{d_{1j}}{R_{1j}} - \frac{d_{1j}+d_{2j}}{R_{1j}+R_{2j}}\right)$ を算出する。このとき、$z_1^2/\sqrt{\widehat{Var}(z_1)}$ は漸近的に自由度 1 のカイ二乗分布

$$\frac{z_1^2}{\sqrt{\widehat{Var}(z_1)}} \sim \chi_1^2 \tag{43}$$

に従う。

手順 4 のスコアは

$$z_1 = \sum_{\text{すべての } t_j}\left(d_{1j} - R_{1j}\frac{d_{1j}+d_{2j}}{R_{1j}+R_{2j}}\right)$$

のように考えると log-rank 検定と類似の形式になる。

$$\begin{aligned} z_1 &= \sum_{\text{すべての } t_j}\left(d_{1j} - R_{1j}\frac{d_{1j}+d_{2j}}{R_{1j}+R_{2j}}\right) \\ &= \sum_{\text{すべての } t_j} R_{1j}\left(\frac{d_{1j}}{R_{1j}} - \frac{d_{1j}+d_{2j}}{R_{1j}+R_{2j}}\right) \end{aligned}$$

と変形すると R_{1j} が重みで、log-rank 検定の一般化のように他の重みを採用することもできる。手順 4 の z_1 の分散の導出は、Pintilie(2006, 付録 A.3) に詳しい。

数値例として、表 19 を取上げる。なお、同表は正しい結果が得られるために Pintilie(2006) の表 5.2 に最後の 2 行を加えた。イベントの 0 は打切り、1 は注目しているイベント、2 は競合リスクである。共変量は A,B の 2 群からなる。

表 19　数値例

時間	イベントのタイプ	群	n_A	n_B	F_A	F_B	S_A	S_B	R_A	R_B	z_A の成分
1	2	B	13	6	0	0	1	0.85714	13	6.002401	0
2	2	A	13	6	0	0	0.92308	0.85714	13	7.000023	0
3	0	A	12	6	0	0	0.92308	0.85714	12.99996	7.000023	0
4	1	B	11	6	0	0.14286	0.92308	0.71429	11.91663	7.000023	-0.629954393
5	1	A	11	5	0.08392	0.14286	0.83916	0.71429	11.91663	5.999944	0.334882385
6	0	A	10	5	0.08392	0.14286	0.83916	0.71429	10.91663	5.999944	0
7	2	B	9	5	0.08392	0.14286	0.83916	0.57143	9.824968	5.999944	0
8	0	A	9	4	0.08392	0.14286	0.83916	0.57143	9.824968	5.999965	0
9	1	B	8	4	0.08392	0.28571	0.83916	0.42857	8.733305	5.999965	-0.592760799
10	0	A	8	3	0.08392	0.28571	0.83916	0.42857	8.733305	5.000047	0
11	2	A	7	3	0.08392	0.28571	0.71928	0.42857	7.641642	5.000047	0
12	1	A	6	3	0.2038	0.28571	0.5994	0.42857	7.641642	5.000047	0.395520483
13	1	A	5	3	0.32368	0.28571	0.47952	0.42857	6.641642	5.000047	0.429494978
14	2	A	4	3	0.32368	0.28571	0.35964	0.42857	5.641642	5.000047	0
15	2	B	3	3	0.32368	0.28571	0.35964	0.28571	5.641642	5.000047	0
16	1	A	3	2	0.44356	0.28571	0.23976	0.28571	5.641642	5.000105	0.46985755
17	1	B	2	2	0.44356	0.42857	0.23976	0.14286	4.641642	5.000105	-0.48141087
18	1	A	2	1	0.56344	0.42857	0.11988	0.14286	4.641642	3.99993	0.462870664
19	0	A	1	1	0.28172	0.42857	0.11988	0.14286	3.641642	3.99993	0
20	0	B	0	1	0	0.42857	0.11988	0.14286	0	3.99993	0

R プログラム

```
>library(cmprsk)
>train <- read.csv("F:\\train.csv",header=FALSE)
>ss<-train$V1
>gg<-train$V3
>cc<-train$V2     # 打ち切りは 0、死亡は、1,2,3,... とし 0 は用いない
>xx<-cuminc(ss,cc,gg,cencode=0)
>xx
>xx2<-timepoints(xx,ss)   #時間ごとの推定値を打ち出し
>xx2
```

を採用すると

```
Tests:
          stat       pv      df
1     0.07160819 0.7890095   1
2     0.30092485 0.5833032   1
```

を得る。すなわち、

$$\frac{z}{\sqrt{\widehat{Var}(z)}} = \frac{0.3885}{5.4260} = 0.0716 \quad (p\text{ 値} = 0.7890)$$

より、注目するイベントの 2 群間には有意差がない。

(4)CIF に関する 2 群間の差の Pepe&Mori(1993) 検定

競合リスクを伴う 2 群間の差を検定しよう。群 $i(i=1,2)$ の CIF を $F_i(t)$、個体数を N_i とする。2 群を合併した観測値 $t_1, t_2, ..., t_n\,(t_1 < t_2 < \cdots < t_n)$ について、

$$\text{帰無仮説 } H_0 : F_1(t) = F_2(t)$$

を検定しよう。打切り、あるいは競合リスクをイベントとした生存関数の Kaplan-Meier 推定量を $C(t)$ としたとき、

$$\begin{array}{l} W(t_j) = \frac{(N_1+N_2)\hat{c}_1(t_{j-1})\hat{c}_2(t_{j-1})}{N_1\hat{c}_1(t_{j-1})+N_2\hat{c}_2(t_{j-1})} \\ s = \sqrt{\frac{N_1N_2}{N_1+N_2}} \sum \left\{ W(t_j) \left[\hat{F}_1(t_j) - \hat{F}_2(t_j) \right] (t_{j+1} - t_j) \right\} \end{array} \tag{44}$$

について、s は漸近的に正規分布 $N(0, \sigma^2)$、すなわち、$s^2/\hat{\sigma}^2$ は自由度 1 のカイ二乗分布

$$s^2/\hat{\sigma}^2 \sim \chi_1^2 \tag{45}$$

に従う。$\hat{\sigma}^2$ は

$$\hat{\sigma}^2 = \frac{N_1N_2\left(\tilde{\sigma}_1^2 + \tilde{\sigma}_2^2\right)}{N_1+N_2} \tag{46}$$

で推定される。各群 i に関する分散 $\hat{\sigma}_1^2, \hat{\sigma}_2^2$ は

$$\tilde{\sigma}^2 = \sum \frac{\left\{ v_1(t_j) - \hat{F}_{cr}(t_j)\, v_2(t_j) \right\}^2 d_{ev_j} + v_2^2(t_j)\left(d_j - d_{ev_j}\right)}{n_j(n_j-1)} \tag{47}$$

となる。ここに

$$\begin{array}{l} v_1(t_j) = \sum\limits_{t_k \geq t_j} W(t_k)(t_{k+1} - t_k)\left(1 - \hat{F}(t_k)\right) \\ v_2(t_j) = \sum\limits_{t_k \geq t_j} W(t_k)(t_{k+1} - t_k) \end{array} \tag{48}$$

で、群 i の時間 t_j でのリスク集合の個体数を n_j、群 i の時間 t_j の注目している、あるいは競合リスクのイベント数を d_j、群 i の時間 t_j での注目しているイベント数を dev_j、群 i における注目しているイベントの CIF を F、群 i の競合リスクに対する CIF を F_{cr} とする。Lunn(1988) は、3 群以上の場合へ拡張している。

表 9 のマウス発癌性データについて、Peto&Mori 検定を行うと

$$\frac{\hat{s}^2}{\hat{\sigma}^2} = 12.39 (p = 0.00043)$$

より、胸腺リンパ腫を注目するイベントとしたとき、2 群間の CIF は高度に有意になる。R プログラムは Pintilie(2006, 付録 B.3.2) に公開されている。

(5) 部分分布に基づく競合リスク回帰モデル

部分分布のハザード関数 (32) 式に基づく競合リスク回帰モデル (Competing risks regression) は、共変量が 1 個の場合、

$$\gamma(t, x) = \gamma_0(t) \exp(\beta x) \tag{49}$$

で与えられる (Fine&Gray,1999;Pintilie,2006,6.2,6.3 節)。ここに γ は部分分布のハザード関数、γ_0

は部分分布のベースラインハザード関数である。このとき、尤度は

$$L(\beta)=\prod_{j=1}^{r}\frac{\exp(\beta x_j)}{\sum_{i\in R_j} w_{ji}\exp(\beta x_i)} \tag{50}$$

となる。ここに、R_j は時間 t までに、あるイベントが起きていない個体、および時間 t までに競合リスクイベントが起きた個体の集合で$R_j(t)=\{i;T_i>t$ あるいは ($T_i<t$ かつ競合リスクイベントが起きた個体)$\}$ となる。他のイベントが起きた個体は、すべての時間でリスク集合に残る。重み w_{ji} は

$$w_{ji}=\frac{\hat{G}(t_j)}{\hat{G}(\min\{t_j,t_i\})} \tag{51}$$

から算出する。ただし、

$$c_i=\begin{cases}1:\text{打切り}\\ 0:o.w.\end{cases}$$

としたとき、$\hat{G}$ は打切り分布 (T_i,C_i) の生存時間に対する Kaplan-Meier 推定量である。注目しているイベント (j とする) が起きた時間でリスクに入る個体 (i とする) の集合は、t_j より前に競合リスクイベントが起きた (この場合、重みは 1) 個体、および時間 t_j までにいかなるタイプのイベントも起きていない (この場合、重みは 1 より小さい) 個体から成る。競合リスクイベントが起きた個体は、部分尤度に加味されない。

重み w_{ji} の算出法は下記の通りである。

手順 1: 縦軸に注目しているイベントのみの時点 (j)、横軸にすべての個体 (i) の番号を記入し、各セルの重みを w_{ji} とする。

手順 2: 各 (j) について、$i\geq j$ となる w_{ji} を 1 とする。

手順 3: 注目しているイベントが起きた個体、および打切り例について、$w_{ji}=1$ 以外のセルは、部分尤度に加味しない ($\times$ を付ける)。

手順 3: 残りのセルには

$$w_{ji}=\frac{\hat{G}(t_j)}{\hat{G}(\min\{t_j,t_i\})} \tag{52}$$

を割付ける。

(49) 式から、スコア統計量

$$U(\beta)=\sum_{j=1}^{r}\left\{x_j-\frac{\sum_{i\in R_j} w_{ji}x_i\exp(\beta x_i)}{\sum_{i\in R_j} w_{ji}\exp(\beta x_i)}\right\} \tag{53}$$

を得る。CIF の予測は

$$F(t)=1-\exp(-H(t)) \tag{54}$$

から算出する。ここに、$H(t)$ は部分分布の累積ハザードで、予測したい共変量の値を x_0 としたとき、Breslow 型推定量

$$\hat{H}\left(t,x_0,\hat{\beta}\right)=\sum_{t_j\leq t}\left\{\frac{\exp\left(\hat{\beta}x_0\right)}{\sum_{i\in R_j} w_{ji}\exp\left(\hat{\beta}x_i\right)}\right\} \tag{55}$$

を採用する。

表 15 の数値例に、共変量 x を加えた表 20 のデータを解析する。$\hat{G}$ は表 20 のように算出される。

w_{ji} は、表 21 のようになる。よって、(53) 式の $U(\beta)$ は

$$
\begin{aligned}
U(\beta) = x_3 &- \frac{w_{32}x_2\exp(\beta x_2)+w_{33}x_3\exp(\beta x_3)+w_{34}x_4\exp(\beta x_4)+\cdots+w_{3,10}x_{10}\exp(\beta x_{10})}{w_{32}\exp(\beta x_2)+w_{33}\exp(\beta x_3)+w_{34}\exp(\beta x_4)+\cdots+w_{3,10}\exp(\beta x_{10})} \\
+ x_4 &- \frac{w_{42}x_2\exp(\beta x_2)+w_{44}x_4\exp(\beta x_4)+w_{45}x_5\exp(\beta x_5)+\cdots+w_{4,10}x_{10}\exp(\beta x_{10})}{w_{42}\exp(\beta x_2)+w_{44}\exp(\beta x_4)+w_{45}\exp(\beta x_5)+\cdots+w_{4,10}\exp(\beta x_{10})} \\
+ x_6 &- \frac{w_{62}x_2\exp(\beta x_2)+w_{66}x_6\exp(\beta x_6)+\cdots+w_{6,10}x_{10}\exp(\beta x_{10})}{w_{62}\exp(\beta x_2)+w_{66}\exp(\beta x_6)+\cdots+w_{6,10}\exp(\beta x_{10})} \\
+ x_9 &- \frac{w_{92}x_2\exp(\beta x_2)+w_{97}x_7\exp(\beta x_7)+w_{98}x_8\exp(\beta x_8)+w_{99}x_9\exp(\beta x_9)+w_{9,10}x_{10}\exp(\beta x_{10})}{w_{92}\exp(\beta x_2)+w_{97}\exp(\beta x_7)+w_{98}\exp(\beta x_8)+w_{99}\exp(\beta x_9)+w_{9,10}\exp(\beta x_{10})} \\
= 12 &- \frac{4\exp(4\beta)+12\exp(12\beta)+21\exp(21\beta)+5\exp(5\beta)+\cdots+8\exp(8\beta)+7\exp(7\beta)}{\exp(4\beta)+\exp(12\beta)+\exp(21\beta)+\exp(5\beta)+\cdots+\exp(8\beta)+\exp(7\beta)} \\
+21 &- \frac{4\exp(4\beta)+21\exp(21\beta)+5\exp(5\beta)+\cdots+8\exp(8\beta)+7\exp(7\beta)}{\exp(4\beta)+\exp(21\beta)+\exp(5\beta)+\cdots+\exp(8\beta)+\exp(7\beta)} \\
+6 &- \frac{0.82\times4\exp(4\beta)+6\exp(6\beta)+11\exp(11\beta)+13\exp(13\beta)+8\exp(8\beta)+7\exp(7\beta)}{0.82\times\exp(4\beta)+\exp(6\beta)+\exp(11\beta)+\exp(13\beta)+\exp(8\beta)+\exp(7\beta)} \\
+8 &- \frac{0.55\times4\exp(4\beta)+0.666\times11\exp(11\beta)+13\exp(13\beta)+8\exp(8\beta)+7\exp(7\beta)}{0.55\exp(4\beta)+0.666\exp(11\beta)+\exp(13\beta)+\exp(8\beta)+\exp(7\beta)}
\end{aligned}
$$

と書ける。

表 20　共変量 x を加えたデータ

個体#	時間	イベントのタイプ	x	$\hat{G}$
1	12	0	3	0.9
2	22	2	4	0.9
3	45	1	12	0.9
4	67	1	21	0.9
5	89	0	5	0.9(1-1/6)=0.75
6	112	1	6	0.75
7	130	2	11	0.75
8	145	0	13	0.75(1-1/3)=0.5
9	160	1	8	0.5
10	203	0	7	

表 21　重みの計算

時点	1	2	3	4	5	6	7	8	9	10
3	×	$\frac{\hat{G}(3)}{\hat{G}(2)}=1$	1	1	1	1	1	1	1	1
4	×	$\frac{\hat{G}(4)}{\hat{G}(2)}=1$	×	1	1	1	1	1	1	1
6	×	$\frac{\hat{G}(6)}{\hat{G}(2)}=1$	×	×	×	1	1	1	1	1
9	×	$\frac{\hat{G}(9)}{\hat{G}(2)}=1$	×	×	×	×	$\frac{\hat{G}(9)}{\hat{G}(7)}=1$	$\frac{\hat{G}(9)}{\hat{G}(8)}=1$	1	1

R プログラム

```
>library(cmprsk)
>train <- read.csv("G:\\train.csv",header=FALSE)
>ss<-train$V1    #生存時間
>gg<-train$V3    # 共変量の値
>cc<-train$V2    # 打ち切りは 0、死亡のみの場合は、1,2,3,... とし 0 は用いない
>xx<-crr(ss,cc,gg)
>xx
>xx$score
```

から

```
convergence:  TRUE
coefficients:
   gg1
0.1448
standard errors:
[1] 0.04541
two-sided p-values:
   gg1
0.0014
xx$score
[1] 1.614071e-10
```

が求まり

$$\hat{\beta} = 0.1448(p\text{ 値} = 0.0014)$$

より、共変量は高度に有意である。

$$U\left(\hat{\beta}\right) = 1.61 \times 10^{-10}$$

となる。

次に、表 14 のデータを解析する。MGUS を注目しているイベント、骨髄腫とその他を競合リスクとして再解析する。R プログラム

```
>library(cmprsk)
>mgus2<- read.csv("F:\\train.csv",header=TRUE)
>cens=mgus2$endpoint
>x=cbind(mgus2$sex,mgus2$age,mgus2$hgb,mgus2$mspike)
>fit=crr(mgus2$time,cens,x)
>fit
```

を採用すると

```
convergence:  TRUE
coefficients:
      x1       x2       x3       x4
-0.31320  0.07775 -0.00765  0.02976
standard errors:
[1] 0.188600 0.009581 0.006711 0.021340
two-sided p-values:
     x1      x2      x3      x4
9.7e-02 4.4e-16 2.5e-01 1.6e-01
```

が得られる。表 22 に周辺モデルおよび層別比例ハザードモデルの結果も載せておく。

表 22　競合リスク回帰モデル、周辺モデルおよび層別比例ハザードモデルとの比較

共変量	競合リスク回帰 (p 値)	周辺モデル			層別比例ハザード
		死亡	多発性骨髄腫	その他	
sex	-0.3132 (0.097)	0.0753 (<0.0001)	0.0079 (0.6006)	-0.0024 (0.907)	-0.3493 (0.0253)
age	0.0778 (<0.0001)	-0.4525 (0.0176)	-0.0761 (0.8231)	0.0043 (0.993)	0.0516 (<0.0001)
hgb	-0.00765 (0.250)	-0.2055 (<0.0001)	-0.0998 (<0.3329)	0.0573 (0.727)	-0.1669 (0.002)
mspike	0.02976 (0.160)	0.1244 (0.579)	-0.6917 (0.0995)	-0.3071 (0.603)	-0.0946 (0.257)

競合リスクが存在しない場合、死因別ハザードと部分分布のハザードに基づく回帰モデルが一致することを確認しよう。表 20 のデータで、競合リスクイベントをすべて注目するイベントとした表 23 を解析する。

表 23　表 20 のデータで、競合リスクイベントをすべて注目しているイベントとしたデータ

個体#	時間	イベントのタイプ	x
1	12	0	3
2	22	1	4
3	45	1	12
4	67	1	21
5	89	0	5
6	112	1	6
7	130	1	11
8	145	0	13
9	160	1	8
10	203	0	7

従来の比例ハザードモデル (死因別ハザード) の場合、R プログラム

```
>library(survival)
>train<-read.csv("F:\\train.csv",header=FALSE)
>fit<-coxph(Surv(V1,V2)~V3,data=train)
>summary(fit)
```

を採用すると

```
      coef exp(coef) se(coef)     z Pr(>|z|)
V3 0.05639   1.05801  0.09569 0.589    0.556

   exp(coef) exp(-coef) lower .95 upper .95
V3     1.058     0.9452     0.877     1.276

Rsquare= 0.032   (max possible= 0.862 )
Likelihood ratio test= 0.33  on 1 df,   p=0.5684
Wald test            = 0.35  on 1 df,   p=0.5557
Score (logrank) test = 0.35  on 1 df,   p=0.5519
```

を得る。一方、部分分布のハザードに基づく競合リスク回帰モデルの場合、R プログラム

```
>library(cmprsk)
>train <- read.csv("F:\\train.csv",header=FALSE)
>ss<-train$V1
>gg<-train$V3
>cc<-train$V2     # 打ち切りは 0、死亡のみの場合は、1,2,3,... とし 0 は用いない
>xx<-crr(ss,cc,gg)
>xx
```

を採用すると、

```
convergence:  TRUE
coefficients:
    gg1
0.05639
standard errors:
[1] 0.07987
two-sided p-values:
 gg1
0.48
```

を得、両者が一致することが分かる。

(6)CIF の比例ハザード性の検定

CIF の比例ハザード性の検定を行うことができる (Pintilie,2006,6.2.3 節)。表 14 のデータについて、TIME 関数を用い、性別との交互作用を新たな共変量にすれば良い。R プログラム

```
>iid=function(x)
>{y=x
>return(y)
>}
>library(cmprsk)
>mgus2<- read.csv("G:\\train.csv",header=TRUE)
>cens=mgus2$endpoint
>x=cbind(mgus2$sex,mgus2$age,mgus2$hgb,mgus2$mspike)
>fit=crr(mgus2$time,cens,x,mgus2$sex,iid)
>fit
```

を採用すれば

```
convergence:  TRUE
coefficients:
            x1             x2             x3             x4 mgus2$sex1*tf1
    -5.365e-01      7.720e-02     -7.317e-03      3.015e-02      6.453e-05
standard errors:
[1] 3.299e-01 9.559e-03 6.690e-03 2.149e-02 7.768e-05
two-sided p-values:
       x1            x2            x3            x4            mgus2$sex1*tf1
       1.0e-01       6.7e-16       2.7e-01       1.6e-01       4.1e-01
```

が得られる。すなわち、性別× time の交互作用は有意にならないため、比例ハザード性は妥当といえる。また、ハザード比は t と共に$exp(-0.5365+0.0000645t)$ で変動する。

次に、注目しているイベントの CIF を F としたとき、$(ln\,(time)\,, \ln\{-\ln(1-F)\})$ の 2 次元プロットで視覚的にチェックすることもできる。R プログラム

```
>library(cmprsk)
>mgus2 <- read.csv("G:\\train.csv",header=TRUE)
>sex=(mgus2$sex==2)+0
>cens=mgus2$endpoint
>fit=crr(mgus2$time,cens,sex)
>fit
>fit=cuminc(mgus2$time,cens,sex)
>a=timepoints(fit,times=mgus2$time)
>cif=t(a$est[1:2,])
>llcif=log(-log(1-cif))
>matplot(log(unique(sort(mgus2$time))),llcif,
>pch=c(1,3),col=1,xlab='Time to disease fairure',ylab='log(-lig(1-CIF))')
>legend(locator(1),pch=c(1,3),c("男性","女性"))
># locator(1) は凡例を書く場所をマウスで指定
```

から、図 10 が得られる。

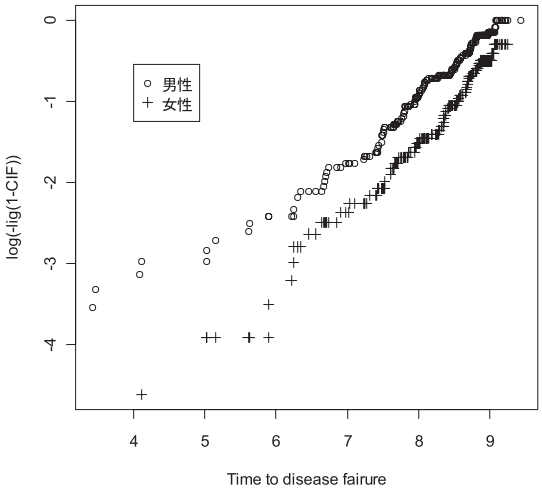

図 10　比例ハザード性の検証

(7) 時間依存型への拡張

時間依存型共変量を伴う場合へ競合リスク回帰モデルを拡張しよう (Beyersmann& Schumacner, 2008)。例えば、図 11 のように“集中治療室 (ICU) に長く留まれば感染症肺炎のため ICU 内での死亡が増えるか”という問題を考える。患者 # 3163 は、ICU 入室中に他の死因 (競合リスク) で死亡した。# 3147 は ICU を退出したが、その後、感染症肺炎で死亡した。# 30148 は、ICU 退出後、入院中 (すなわち、打切り) である。# 30236 は、ICU 入室中の途中までは、異常が無かったが、ICU 退室後、しばらくして感染症肺炎を発症し、入院中である。# 30254 は、ICU 入室中、途中までは異常が無かったが、退室後しばらくして感染症肺炎で死亡した。# 1014378 は、ICU 入室中の途中までは異常は無かったが、しばらくして感染症肺炎のため ICU 内で死亡した (注目しているイベント)。

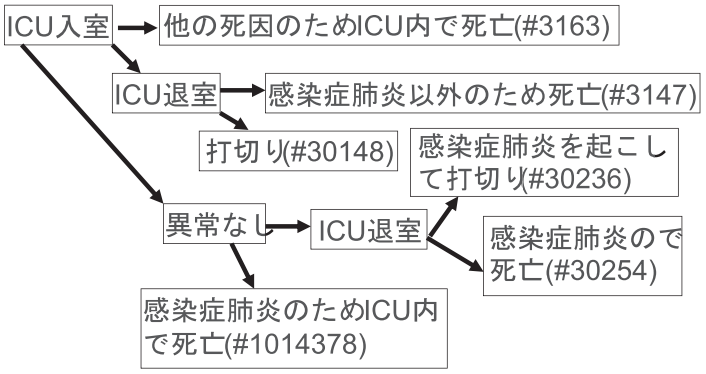

図 11　ICU

表 24 のデータは、それぞれの患者の時間依存型データである。各変数は

$$status = \begin{cases} 0: \\ 1: o.w. \end{cases}$$
$$event = \begin{cases} 2: ICU \\ 3: ICU \end{cases}$$
$$pneu = \begin{cases} 0: \\ 1: \end{cases}$$

を表している。ただし、$status = 0$ の場合は、$event$ の値を無視する。

表 24　ICU データ

id	entry	exit	status	event	pneu
3147	0	11	1	2	0
3163	0	5	1	3	0
30148	0	28	0	2	0
30236	0	6	0	2	0
30236	6	26	0	2	1
30254	0	15	0	2	0
30254	15	18	1	2	1
1014378	0	71	0	2	0
1014378	71	86	1	3	1

R プログラム

```
>library(kmi)
>data(icu.pneu)
>set.seed(1313)
>imp.data<-kmi(Surv(entry,exit,status)~1,data=icu.pneu,
etype=event,id=id,failcode=3,nimp=5)
>fit.kmi<-cox.kmi(Surv(entry,exit,event==3)~pneu,imp.data)
>summary(fit.kmi)
>### Now using the censoring-complete data
>fit<-coxph(Surv(entry,adm.cens.exit,event==3)~pneu,icu.pneu)
>summary(fit)
```

を採用すると、

```
*****************
Pooled estimates:
*****************
              coef     exp(coef)  se(coef)       t            Pr(>|t|)
pneu1  1.1038     3.0157     0.2399     4.602    4.19e-06 ***
---
Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ' '  1

      exp(coef) exp(-coef) lower .95 upper .95
pneu1     3.016     0.3316     1.885     4.826
```

を得る。"ICU 内感染症肺炎" 発症の有無は高度に有意で、院内肺炎のハザード比は 3.016 となる。

R のパッケージ 'kmi' を確認するため、3.1 節の PBC データを解析してみよう (この R パッケージは http://cran.r-project.org/web/packages/kmi を参照)。このデータは時間依存型であるが、注目しているイベント (死亡) のみで、競合リスクは存在しない。よって、表 25 のように書ける。

表 25　PBC データ

id	entry	exit	status	event	bilirubin
1	0	47	0	3	3.2
1	47	184	0	3	3.8
1	184	251	0	3	4.9
1	251	281	1	3	5.0
2	0	94	0	3	3.1
2	94	187	0	3	2.9
2	187	321	0	3	3.1
2	321	604	0	3	3.2
.	.	.	.	.	.

R プログラム

```
>library(kmi)
>PBC<-read.csv("E:\\train.csv",header=T)
>imp.data<-kmi(Surv(entry,exit,status)~1,data=PBC,
etype=event,id=id,failcode=3,nimp=5)
>fit.kmi<-cox.kmi(Surv(entry,exit,event==3)~bilirubin,imp.data)
>summary(fit.kmi)
```

を採用すると

```
*****************
Pooled estimates:
*****************
                    coef  exp(coef)    se(coef)     t        Pr(>|t|)
bilirubin    3.299    27.092       1.734     1.903    0.057 .
---
Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ' '  1

          exp(coef) exp(-coef) lower .95 upper .95
bilirubin     27.09    0.03691    0.9062       810
```

となり、3.1 節の結果と一致する。

参考文献

Aalen,O., Borgan, Φ , and Gjessing,H.K. (2008): Survival and Event History Analysis. Springer.

Aitkin, M., Laird, N., and B. Francis, B.(1983). A reanalysis of the Stanford heart transplant data, *Journal of the American Statistical Association* **78**, 264-292.

赤澤宏平, 柳川堯 (2010): サバイバルデータの解析 —生存時間とイベントヒストリーデータ—、近代科学社.

Allison,P.D.(1995):Survival Analysis Using the SAS System, A Practical Guide. SAS Publisher

Altman, D.G., & De Stavola, B.L.(1984). Practical problems in fitting a proportional hazards model to data with updated measurements of the covariates, *Statistics in Medicine*, **13**, 301-341.

Andelsen, P.K., and Gill, R.D.(1982). Cox's regression model for counting process: A large sample study, *Annals of Statistics*, **10**, 1100-1120.

Beyersmann,J. and Schumacher,M.(2008):Time-depependent covariates in the proportional subdistribution hazards model for competing risks, *Biostatistics*, **9**, 765-776.

Collet,D.(2003):Modelling Survival Data in Medical Research, 2nd edi. Chapman & Hall.

Cox,D.R.(1975): Partial likelihood, *Biometrika*, **62**,269-276.

Christensen, E., Schlichting, P., Andersen, P.K., Fauerholdt, L., Schou, G., Pedersen, B.V., Juhl, E., Poulsen, H., and Tygstrup, N., Copenhagen Study Group for Liver Disease (1986). Updating prognosis and therapeutic effect evaluation in cirrhosis with Cox's multiple regression model for time-dependent variables, *Scandinavian Journal of Gastroenterology*, **21**, 163-174.

Fine,J.P., and Gray,R.J.(1999): A proportional hazards model for the subdistribution of a compet-

ing risk, *Journal of the American Statistical Association*, **94**, 496-509.

Gray, R.J.(1988): A class of k-sample tests for comparing the cumulative incidence of a competing risk, *Annals of Statistics*, **16**, 1141-1154.

Gray,R.J.(2010):cmprsk:Subdistribution analysis of competing risks. R package version 2.2-1, URL http://cran.r-project.org/web/packages/cmprsk.

Kalbfleisch,J.D.,and Prentice,R.L.(2002):The Statistical Analysis of Failure Time Data. John Wiley.

Klein,J.P.,and Moeschberges,M.L.(2003): Survival Analysis, Springer.

Lee,E.M., et al.,(2003):Statistical Methods for Survival Data Analysis. John Wiley.

Lunn,M.(1998): Applying k-sample tests to conditional probabilities for competing risks in a clinical trial, *Biometrics*, **54**, 1662-1672.

中村剛 (2001):Cox 比例ハザードモデル , 朝倉書店

西川正子 (2008): 生存時間解析における競合リスクモデル , 計量生物学 , **29**,141-170.

大橋靖雄 , 浜田知久馬 (1995): 生存時間解析 , 東大出版

Pepe, M.S., and Mori,M.(1993): Kaplan-Meier, Marginal or conditional probability curves in summarizing competing risks failure time data, *Statistics in Medicine*, **12**, 737-751.

Pintilie,M.(2006):Copeting Risks, a Practical Perspective. John Wiley.

Prentice,R.L., Williams,B.J., and Peterson, A.V.(1981): On the regression analysis of multivariate failure time data, *Biometrika*, **68**, 373-379.

Putter,H., Fiacco, M., and Geskus, R.B.(2007): Tutorial in Biostatistics; Competing risks and multi-state models, *Statistics in Medicine*, **26**, 2389-2430.

Therneau,T., and Grambsch,P.(2000): Modeling Survival Data: Extending the Cox Model. Springer-Verlag.

Tsujitani, M. and Koshimizu,T(2000): Neural Discriminant Analysis, *IEEE Transaction on Neural Networks*, **11**, 1394-1401.

辻谷将明 , 外山信夫 (2007):R によろ GAM 入門 , 行動計量学 ,**34**,111-131.

辻谷将明 , 竹澤邦夫 (2009):R で学ぶデータサイエンス 6 マシンラーニング , 共立出版 .

辻谷将明 , 和田武夫 (2012):R で学ぶ確率・統計 , 共立出版 .

Tsujitani,M. and Sakon,M.(2009): Analysis of survival data having time-dependent covariates, *IEEE Transaction on Neural Networks,* **20**,389-394.

Tsujitani, M., and Tanaka, Y.(2011):Cross-validation,Bootstrap, and support vector machines, *Advances in Articial Neural Systems*, **2011**, Article ID302572, 6pages.

Tsujitani, M., and Baesens, B.(2012) : Survival analysis for personal loan data using generalized additive models, *Behaviormetrika*, **39**,9-23.

Tsujitani,M., Tanaka,Y., and Sakon,M.(2012):Survival analysis with time-dependent covariates using generalized additive models, *Computational and Mathematical Methods in Medicine*, **2012**, Article ID986176, 9 pages.

Tsujitani, M.,Iba,K., and Tanaka,Y.(2012):Neural discriminant models, bootstrapping and simulation, *Computational and Mathematical Methods in Medicine*, **2012**, Article ID820364, 12 pages.

Tsujitani, M., and Tanaka,Y.(2013):Analysis of heart transplant survival data using generalized additive models, *Computational and Mathematical Methods in Medicine*, **2013**, Article ID609857, 7 pages.

Wood, S. N. (2000): Modelling and smoothing parameter estimation with multiple quadratic penalties. *Journal of the Royal Statistical Society, Series B*, **62**, 413-428.
Wood, S. N. (2006): Generalized additive models: An Introduction with R. Chapman & Hall / CRC.
Wood, S. N. (2008): Fast stable direct fitting and smoothness selection for generalized additive models, *Journal of the Royal Statistical Society Series B*, **70**, 495-518.