《R 语言》课程报告

题目：基于 Cox 回归的动态预测模型

子题目：界标在生存分析中的应用

成员：李林吉（3217042035）（组长）

李　坦（3227042001）

丘逸风（3227042005）

陈奕昕（3227042017）

钟端端（3227042036）

2022 级应用统计学专业（生物统计方向）

完成时间：2024 年 6 月 15 日

授课教师：陈征

目录

摘要

提供预后的动态评估对于改善个性化医疗至关重要，常用的一般Cox比例风险模型则是一种“静态预测”模型，不考虑时间效应，且需满足等比例假定，然而这一假定往往是不能满足的。因此，生存数据的界标模型为疾病进展的动态预测提供了一个潜在的强大解决方案。对于此目的，我们构建了创建界标数据集$LM()$函数以及时间交互$define_time()$函数，以至于更加简便地去构建动态预测模型。最后，我们通过对实例数据的分析与建模，证明了所构建函数的可用性，以及界标在处理动态模型中的实用性。

Abstract

Providing a dynamic assessment of prognosis is essential to improve personalized medicine, and the commonly used Cox proportional risk model is a "static prediction" model that does not consider the time effect and needs to satisfy the assumption of equal proportions, which is often not satisfied. Thus, the landmark model of survival data offers a potentially powerful solution for dynamic prediction of disease progression. For this purpose, we built the LM() function for creating the boundary data set and the define_time() function for time interaction to make it easier to build dynamic prediction models. Finally, through the analysis and modeling of the example data, we prove the usability of the constructed functions and the practicability of the boundary markers in dealing with dynamic models.

1 方法理论介绍

1.1 COX 比例风险模型及其局限性

目前对于癌症或肿瘤患者的预后研究，主要使用的是 Cox 比例风险模型。Cox 比例风险模型，是一种生存分析的半参数模型。它的因变量包括生存时间与生存结局，自变量包括可能的多个协变量，例如年龄、性别、相关临床指标等，该模型用来预测在这些协变量的影响下发生生存结局的时间。Cox 比例风险模型具有不需要估计资料生存分布类型的优点，适用于长期追踪研究与随机对照试验。其风险函数的计算公式为：

$$h(t|X) = h_0(t)\exp\sum_{i=1}^{m}(\beta_i X_i)$$

其中，$h(t|X)$是在给定协变量X时的风险函数，$h_0(t)$是基线风险函数，代表在$X=0$时的风险，β_k是回归系数，表示协变量X_k的影响。

若两组的协变量分别为X和X^*，则风险比HR为：

$$\frac{h(t|X)}{h(t|X^*)} = \frac{h_0(t)exp\sum_{i=1}^{m}(\beta_i X_i)}{h_0(t)exp\sum_{i=1}^{m}(\beta_i X_i^*)} = \exp\left[\sum_{i=1}^{m}\beta_i(X_i - X_i^*)\right]$$

在利用 Cox 比例风险模型时，它必须满足 PH 假定，即各协变量风险比 HR 在时间尺度上是恒定不变的，但在长期随访研究期间这一假定常常是无法成立的。此外，Cox 比例风险模型不能在随访期间的特定预测时间 s 使用，它仅使用基线时可用的信息来预测患者未来不发生感兴趣事件的概率，例如预测诊断后 3 年的累计生存率。

然而，在实际临床诊断及预后中，癌症或肿瘤的治疗是一个长期的过程，患者更希望了解的是每一个阶段的预后，而 Cox 模型仅用于预测从基线开始未来的累计生存率，不一定适用于初次诊断后已存活一段时间的患者，这是因为患者在初次诊断后的一段时间内存活可能会影响其未来的预后，这样的患者会对未来再存活几年的概率更感兴趣[1]。因此，对于以上 Cox 比例风险模型的两大局限性，我们构建了时间效应的 Cox 动态预测模型并利用 R 4.4.0 加以实现，以至于更好地反映预后如何随时间的推移而变化等相关问题。

1.2 时变效应的 Cox 动态预测模型的建立

1.2.1 动态 Cox 模型的建立

“动态预测”是根据事件的历史和不同的协变量[2]，从基线和随后的时间点获得预测概率的一种预测方法，它能对患者在任何预测时间点$s(s_0, s_1, \ldots, s_{LM})$进行预测，当需要考虑基线特

征X时，它包括了动态生存率和动态风险率，其动态生存率表达式为：

$$P(T > s_{LM} + w | T \geq s_{LM})$$

其中，w为预测窗口，它取决于疾病病程的长短；s_{LM}为预测时间点，且$s_{LM} \in [0, s_L]$。基于上述表达式，我们可以从全部数据集中取出一个子集，该数据集包含了已存活了s_{LM}年的受试者。这个子集称为界标数据集，指在时间点s_{LM}的风险人群，即已经存活了s_{LM}年但仍处于风险人群中的受试者，同时忽略所有在$s_{LM} + w$年之后的终点事件和删失的一类数据形式。

基于每个界标时间点所对应的界标数据集R_{LM}，我们为其拟合单独的 Cox 模型，从而估计从s_{LM}到$s_{LM} + w$的任何时间点处的动态危险率，其数学表达式为：

$$h(t|X, s_{LM}) = h_0(t|X, s_{LM}) \exp \sum_{i=1}^{m} \left(X_i \beta_i (s_{LM}) \right), s_{LM} \leq t \leq s_{LM} + w$$

注意上述基本形式，实际上相当于构建多个独立的 Cox 模型，这显然并不便于模型的解释，且当界标点过多，即界标数据集R_{LM}过多时，模型的建立较为复杂，因此我们将所有界标数据集进叠加，构成一个“超预测数据集”[3]，并且在此基础上将界标时间点s_{LM}作为分层变量纳入 Cox 模型，构建 Cox 动态预测模型，其数学表达式为：

$$h(t|X, s_{LM}) = h_0(t) \exp \left[\sum_{i=1}^{m} (X_i \beta_i) + \theta(s_{LM}) \right], s_{LM} \leq t \leq s_{LM} + w$$

其中，$\theta(s_{LM}) = \sum_{i=1}^{m} \omega_i g_i(t_{LM})$，$g_i(t_{LM})$是一系列光滑函数，其作用等同于将多个独立的 Cox 模型组合成一个整体，用来预测在不同时间点未来w年的生存率，达到动态预测的目的。

1.2.2 时间效应协变量

由于“静态模型”如 Cox 比例风险模型需满足比例风险(PH)假定，即预测因素对结局的影响随时间保持恒定，当PH假定不成立时可能产生具有误导性或错误的结论。在实际应用中，PH假定不成立的原因主要有两个方面：1、时间相依：协变量随时间的变化而变化，从而使协变量对生存时间分布的影响发生改变。2、时间效应：协变量不随时间的变化而变化，只是同样的协变量对生存时间分布的影响效应发生变化。在本次报告中，我们仅考虑时间效应协变量对生存时间分布的影响。

为了解决协变量时间效应的影响，对 Cox 动态预测模型的回归系数β同样引入一个光滑函数，描述协变量的时间效应，其数学表达式为：

$$h(t|X,s_{LM}) = h_0(t)\exp\left[\sum_{i=1}^{m}(X_i\beta_i(s_{LM})) + \theta(s_{LM})\right], s_{LM} \le t \le s_{LM} + w$$

其中，$\beta(s_{LM}) = \sum_{j=1}^{n}\beta_j f_j(s_{LM}), f_j(s_{LM})$是一系列光滑函数，通过引入协变量与时间的交互，从而解决原 Cox 比例风险模型所不能解决的非等比例假定的问题。

2 与上对应的各部分程序介绍

2.1 Cox 回归模型的建立

在构建Cox比例风险模型时，我们使用了'survival'包所带的coxph()函数，Surv()函数进行模型的建立，并利用cox.zph()函数检验是否满足等比例假定，其代码如下：

```
#建立多元 Cox 模型
model1 <- coxph(Surv(time,status)~rx+obstruct+adhere+differ+extent+node4+
                      cluster(id),data = data)
cox.zph(model1) #检验是否满足等比例风险假定
```

2.2 时变效应的 Cox 动态预测模型的建立

2.2.1 界标数据集及超预测数据集的构建

在建立 Cox 动态预测模型之前，需要构建界标数据集，并将其叠加成超预测数据集，基于上述目的，我们搭建了构建界标数据集并叠加成超预测数据集的函数$LM()$代码如下：

```
LM <- function(w,by,data,time = 'time',status = 'status'){
  # 构建综合数据集的计算函数
  # 输入变量:
  #    w:预测时间(基于疾病病程长短决定)
  #    by:界标步长
  #    data:基线生存数据集
  #    time:数据集中生存时间变量名
  #    status:数据集中结局状态变量名
  #
  # 输出变量:
  #    LMdata:各界标数据集叠加所得的综合数据集

  sL <- max(data$time) - w
  sl <- seq(0,sL,by = by)
  nsl <- length(sl) #界标数据集的个数
  sldata <- list() #将得到的界标数据集放入列表中,便于后续提取叠加数据集
  for(i in 1:nsl){
    data <- data %>% filter(time > sl[i]) %>% mutate(LM = sl[i])
    if (nrow(data) > 0) {
      sldata[[i]] <- data #如果界标数据集不为空,则添加到列表中
```

```
    } else {}
  }

  #界标数据集的叠加,构建超预测数据集
  LMdata <- data.frame(array(,dim = c(0,0))) #定义超预测数据集
  for (i in 1:nsl) {
    LMdata = rbind(LMdata,sldata[[i]]) #将所有界标数据集叠加得到超预测数据集
  }
  LMdata1 <- LMdata %>% filter(LM+w <= time) %>%
    mutate(time = LM+w,status = 0)
  LMdata <- LMdata %>% filter(LM+w > time)
  LMdata <- rbind(LMdata1,LMdata)
}
```

在$LM()$函数中，包含以下 5 个参数设置，相关解释如下：

参数	相关解释
w	预测时间
by	界标步长
data	基线生存数据
time	数据集中生存时间变量名
status	数据集中结局状态变量名

为了检验我们所构建的超预测数据集是否正确，我们使用了$'dynpred'$包中的$cutLM()$函数，并利用实例数据在同一预测时间$(w = 3)$与界标步长$(by = 0.25)$下所创建的超预测数据集进行等同性检验。$cutLM()$函数的使用代码如下：

```
#install.packages('dynpred')
library(dynpred)
sL = 6
sl <- seq(0,sL,by=0.25)
nsl <- length(sl)
w <- 3
LMdata1 <- data.frame(array(,dim=c(0,0)))
for (i in 1:nsl){
  LMdata1 <- rbind(LMdata1,cutLM(data=data,
                                 outcome=list(time="time",status="status"),
                                 LM=sl[i],horizon=sl[i]+w,
                                 covs=list(fixed=c('id','obstruct','adhere','node4',
                                                   'rx1','rx2','differ1',
                                                   'differ2','extent1','extent2',
```

```
                                                    'extent3'),varying=NULL)))
}
```

根据二者所构建的超预测数据集，我们分别从两个超预测数据集中筛选出$id = 1$的患者数据进行了等同性对比，代码及结果如下：

```
LMdata %>% filter(id == 1)   #LM()函数构建的数据集
LMdata1 %>% filter(id == 1) #cutLM()函数构建的数据集
```

$LM()$函数构建数据集结果如下：

```
> LMdata %>% filter(id == 1)  #LM()函数构建的数据集
# A tibble: 17 × 14
   obstruct adhere status node4  time    id   rx1   rx2 differ1 differ2 extent1 extent2 extent3
      <dbl>  <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
 1        0      0      0     1 3         1     0     1       0       1       0       1       0
 2        0      0      0     1 3.25      1     0     1       0       1       0       1       0
 3        0      0      0     1 3.5       1     0     1       0       1       0       1       0
 4        0      0      0     1 3.75      1     0     1       0       1       0       1       0
 5        0      0      0     1 4         1     0     1       0       1       0       1       0
 6        0      0      1     1 4.17      1     0     1       0       1       0       1       0
 7        0      0      1     1 4.17      1     0     1       0       1       0       1       0
 8        0      0      1     1 4.17      1     0     1       0       1       0       1       0
 9        0      0      1     1 4.17      1     0     1       0       1       0       1       0
10        0      0      1     1 4.17      1     0     1       0       1       0       1       0
11        0      0      1     1 4.17      1     0     1       0       1       0       1       0
12        0      0      1     1 4.17      1     0     1       0       1       0       1       0
13        0      0      1     1 4.17      1     0     1       0       1       0       1       0
14        0      0      1     1 4.17      1     0     1       0       1       0       1       0
15        0      0      1     1 4.17      1     0     1       0       1       0       1       0
16        0      0      1     1 4.17      1     0     1       0       1       0       1       0
17        0      0      1     1 4.17      1     0     1       0       1       0       1       0
# ℹ 1 more variable: LM <dbl>
```

$cutLM()$函数构建数据集结果如下：

```
> LMdata1 %>% filter(id == 1) #cutLM()函数构建的数据集
# A tibble: 17 × 14
    time status    id obstruct adhere node4   rx1   rx2 differ1 differ2 extent1 extent2 extent3
   <dbl>  <dbl> <dbl>    <dbl>  <dbl> <dbl> <dbl> <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
 1  3         0     1        0      0     1     0     1       0       1       0       1       0
 2  3.25      0     1        0      0     1     0     1       0       1       0       1       0
 3  3.5       0     1        0      0     1     0     1       0       1       0       1       0
 4  3.75      0     1        0      0     1     0     1       0       1       0       1       0
 5  4         0     1        0      0     1     0     1       0       1       0       1       0
 6  4.17      1     1        0      0     1     0     1       0       1       0       1       0
 7  4.17      1     1        0      0     1     0     1       0       1       0       1       0
 8  4.17      1     1        0      0     1     0     1       0       1       0       1       0
 9  4.17      1     1        0      0     1     0     1       0       1       0       1       0
10  4.17      1     1        0      0     1     0     1       0       1       0       1       0
11  4.17      1     1        0      0     1     0     1       0       1       0       1       0
12  4.17      1     1        0      0     1     0     1       0       1       0       1       0
13  4.17      1     1        0      0     1     0     1       0       1       0       1       0
14  4.17      1     1        0      0     1     0     1       0       1       0       1       0
15  4.17      1     1        0      0     1     0     1       0       1       0       1       0
16  4.17      1     1        0      0     1     0     1       0       1       0       1       0
17  4.17      1     1        0      0     1     0     1       0       1       0       1       0
# ℹ 1 more variable: LM <dbl>
```

根据上述检验，可以证明$LM()$函数所构建的超预测数据集与$cutLM()$函数所构建的数据集一致，且在实际应用中，$cutLM()$函数所需要定义的参数过多，当协变量过多时，需要输入每一个协变量的变量名，且在使用前期所需要的定义较多，操作较为复杂；而在我们所搭建的$LM()$函数中所需要定义的参数较少且更容易使用，因此说明我们所搭建的$LM()$函数更能为研

究基线数据的人员提供便利。

2.2.2 协变量与时间的一次交互项以及时间变量的创建

在创建完超预测数据集后，我们需要在数据集中添加协变量与时间的一次交互项以及时间变量，从而去拟合时间效应的 Cox 动态预测模型。因此，我们搭建了$define_time()$函数，用于快速添加协变量与时间交互以及时间变量数据，其代码如下：

```
define_time <- function(data,fixed,time1){
  # 创建各协变量与时间的一次交互项以及时间变量
  # 输入变量:
  #    data:超预测数据集
  #    fixed:需要创建交互项的变量名
  #    time1:所需要与协变量交互的时间值
  #
  # 输出变量:
  #    data:创建好交互项与时间变量的数据集

  for (var in fixed) {
    # 计算交互项并添加到数据框
    interaction_name_time <- paste0(var,'1')
    data <- data %>%
      mutate(!!interaction_name_time := .data[[var]] * time1)
  }
  data$LM1 <- time1
  data$LM2 <- time1^2
  return(data)
}
```

在$define_time()$函数中，包含了以下 3 个参数设置，相关解释如下：

参数	相关解释
data	超预测数据集
fixed	需要创建一次交互项的变量名
time1	与协变量交互的时间值

该函数可以有效迅速创建协变量与时间的一次交互项以及时间变量，为后续模型的建立奠定了基础，也减少了研究人员创建交互项所花费的时间，具有普适性。

3 实例分析

3.1 实例数据介绍

实例分析部分采用的结肠癌患者从患病到死亡的基线数据，该数据来源于 R 软件“survival”

包中的示例数据“colon”，原数据集共 929 例患者数据，对其中治疗方案、年龄、性别、肿瘤阻塞情况、穿孔情况、粘连情况、扩散情况、淋巴结数量部分的缺失值进行了删除处理，并将生存时间由原始的天为单位转化为年为单位（生存时间范围为 0.63—9.12 年），得到有效数据共 906 例，基线统计描述汇总见附录，相关代码如下：

```
#数据导入
data <- read_csv("ex1.csv")

#数据预处理
data <- na.omit(data) #删除缺失值
data <- data %>% select(-id) %>% mutate(id = c(1:906)) #重新定义每个 id
data <- data %>% mutate(time = time/365) #将 time 修改为年份为单位
data$rx = factor(data$rx,levels = c('Lev','Lev+5FU','Obs'),
                 labels = c(1,2,3)) #将变量因子化,便于后续处理
data$age <- ifelse(data$age <= 60,0,1) #将 age 变量二分类
data <- data %>% mutate(age = factor(age),sex = factor(sex),
                        obstruct = factor(obstruct),perfor = factor(perfor),
                        adhere = factor(adhere),differ = factor(differ),
                        extent = factor(extent),
                        node4 = factor(node4)) #变量因子化便于统计描述
```

3.2 变量筛选

为确保无过多无统计学意义的变量混入模型，采取单因素回归且入选阈值为 0.1 进行变量筛选，筛选出不具有统计学意义的指标（sex, age, perfor），将其去除后，其余指标纳入后续模型分析，该部分代码如下：

```
#变量筛选
fit <- coxph(Surv(time,status)~sex+age+rx+obstruct+perfor+adhere+differ+
               extent+node4+cluster(id),data = data) #建立全因子 Cox 比例风险模型
result <- autoReg(fit,uni = T,threshold = 0.1) %>% myft() #利用单因素 Cox 回归进行变量筛选
result
data <- data %>% select(-c(sex,age,perfor)) #根据筛选结果剔除没有统计学意义的变量
```

其结果如下：

Variable	Status =1	HR (univariable)	P	HR (multivariable)	P
性别					
女性	438				
男性	468	1.00 (0.83-1.21)	.984		

年龄					
<=60 岁	431				
>60 岁	475	1.17 (0.97-1.41)	.104		
治疗方案					
观察组	300				
Lev	298	0.73 (0.57-0.92)	.008	0.71 (0.56-0.9)	.005
Lev+5-FU	308	1.04 (0.84-1.30)	.710	1.03 (0.83-1.29	.770
梗阻					
无	731				
有	175	1.30 (1.03-1.63)	.025	1.29 (1.03-1.63)	.027
穿孔					
无	879				
有	27	1.17 (0.70-1.9)	.560		
粘连					
无	775				
有	131	1.35 (1.05-1.73,)	.018	1.21 (0.94-1.55)	.147
肿瘤分化程度					
良好	93				
中等	663	1.05 (0.76-1.4)	.763	0.93 (0.67-1.29)	.656
较差	150	1.70 (1.18-2.4)	.005	1.35 (0.93-1.9)	.115
局部扩散程度					
粘膜下	20				
肌肉	102	1.83 (0.65-5.15)	.253	1.38 (0.49-3.8)	.547
浆膜	745	3.25 (1.21-8.7)	.019	2.23 (0.83-6.0)	.113

毗连结构	39	5.04 (1.75-14.49)	.003	3.08 (1.06-8.9)	.039
4 个及以上阳性淋巴结					
无	654				
有	252	2.58 (2.13-3.12)	<.001	2.46 (2.03-2.99)	<.001

3.3 Cox 回归模型各协变量 PH 假定检验

对于已经筛选出的指标，建立Cox比例风险模型，并对其各协变量进行等比例假设检验，该部分代码及PH假定结果如下：

```
#建立多元 Cox 模型
model1 <- coxph(Surv(time,status)~rx+obstruct+adhere+differ+extent+node4+
                  cluster(id),data = data)
cox.zph(model1) #检验是否满足等比例风险假定
```

各协变量 PH 假定结果如下：

协变量	χ^2	df	P
rx	2.4213	2	0.298
ostruct	6.3335	1	<0.05
ahere	0.0934	1	0.75882
differ	15.8333	2	<0.01
extent	7.4802	3	0.05807
node4	5.9329	1	<0.05

从输出结果结果我们可以得到，obstruct, differ, node4三个协变量P值均小于 0.05，说明不满足等比例风险假定，即协变量风险比随时间的变化而改变，因此说明Cox模型没有很好的反映风险比随时间的变化效应。

3.4 各协变量随时间的风险比变化

为构建界标数据集，本次将数据依据生存时间区间进行子集分割，并确定预测窗口$w = 3$，分割点即界标点以 0.25 为间隔构成形如$[s_0, s_1, \dots, s_{24}]$的界标点集，将各界标数据集进行叠加构成“超预测数据集”（共 15899 例数据），并构建时间效应的Cox动态预测模型及求得各协变量系数、标准误以及P值，该部分代码及其结果如下：

```
model2 <- coxph(Surv(LM,time,status)~obstruct+obstruct1+adhere+adhere1+node4+
```

```
                node41+rx1+rx11+rx2+rx21+differ1+differ11+differ2+differ21+
                extent1+extent11+extent2+extent21+extent3+extent31+LM1+LM2+
                cluster(id),data = LMdata,method="breslow")

#输出各协变量的回归系数,标准误以及 P 值
output<-(cbind(coef = model2$coefficients,
               SE = sqrt(diag(model2$var)),
               PVal = 2-2*pnorm(abs(model2$coefficients/
                                    sqrt(diag(model2$var))))))
round(output,3)
```

模型各协变量系数，标准误以及P值汇总如下：

Variable	Time function	Coefficient	SE	P value
梗阻（Ref:无）				
有	1	0.317	0.187	0.090
	$s_{LM}/6$	-1.182	0.797	0.138
粘连（Ref:无）				
有	1	0.209	0.182	0.251
	$s_{LM}/6$	0.319	0.752	0.671
四个及以上淋巴结（Ref:无）				
有	1	1.020	0.141	<0.001
	$s_{LM}/6$	-0.993	0.615	0.106
处理组（Ref：安慰剂组）				
阿咪唑	1	0.108	0.157	0.492
	$s_{LM}/6$	-1.098	0.648	0.090
阿咪唑+5-氟尿嘧				
	1	-0.422	0.176	0.016
	$s_{LM}/6$	-0.109	0.664	0.869
肿瘤分化程度（Ref:较差）				
良好	1	-0.266	0.225	0.238
	$s_{LM}/6$	1.527	1.059	0.149

适中	1	-0.682	0.159	<0.001
	$s_{LM}/6$	2.916	0.846	0.001
局部扩散部位（Ref:肌肉）				
粘膜下	1	0.476	0.684	0.487
	$s_{LM}/6$	-4.446	3.061	0.146
浆膜	1	0.920	0.291	0.002
	$s_{LM}/6$	-2.169	0.943	0.021
毗连结构	1	1.222	0.466	0.009
	$s_{LM}/6$	-2.651	1.763	0.133
$\boldsymbol{\theta(s_{LM})}$	$s_{LM}/6$	1.079	1.205	0.371
	$(s_{LM}/6)^2$	-1.202	0.397	0.002

注：$Time\ function: \beta(s_{LM}) = \beta_0 + \beta_1(s_{LM}/6), \theta(s_{LM}) = \theta_1(s_{LM}/6) + \theta_2(s_{LM}/6)^2$

根据拟合好的时间效应Cox动态预测模型，需要计算各协变量随时间的风险比，以用于更好地反映协变量时间效应对生存时间分布的影响，其风险比计算公式如下：

$$HR^w = \exp(\beta_0 + \beta_1 \bar{s}_{LM}) = \exp[\beta_0 + \beta_1 \times (s_{LM}/6)], s_{LM} \in [s_0, s_L]$$

因此我们构建了计算各协变量随时间变化的$\log(HR)$值及其 95%可信区间函数$HR()$，以至于便捷计算各协变量的相关数据，其代码如下：

```
HR <- function(w,data,by,model,time = 'time'){
  # 协变量随时间风险比的变化以及 95%可信区间计算函数
  # 输入变量:
  #    w:预测窗口
  #    data:综合数据集
  #    by:界标划分步长
  #    model:时间效应的 Cox 动态预测模型
  #    time:数据集中生存时间变量名
  #
  # 输出变量:
  #    HR:各协变量在各界标时间点的 log(HR)及其 95%置信区间列表

  sL = max(data$time) - w
  sl = seq(0,sL,by = 0.25)
  bet <- matrix(c(rep(1,length(sl)),sl/sL),length(sl),2)
  HR <- list() #将各协变量风险率随时间变化的 HR 值放入列表
  i = 1
```

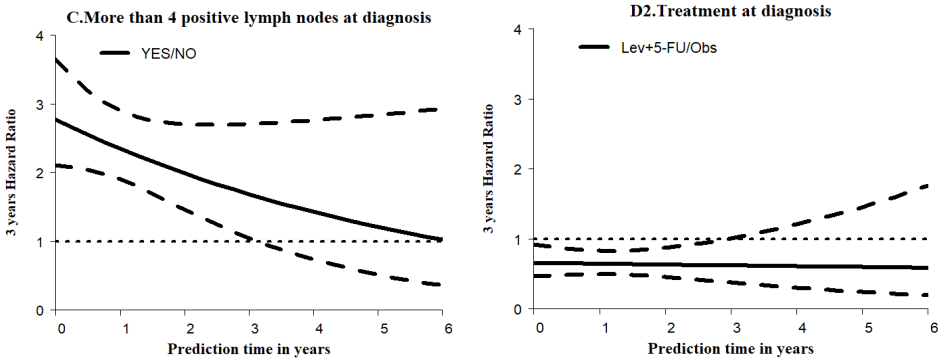
```
  j = 1
  while(i <= (length(model$coef)-2)){
    Hr <- data.frame(sl = sl,logHR = as.numeric(bet %*% model$coef[i:(i+1)])) #计算随时间变化的 log(HR)值
    se <- sqrt(diag(bet %*% model$var[i:(i+1),i:(i+1)] %*% t(bet))) #计算标准误
    Hr$lower <- exp(Hr$logHR-qnorm(0.975)*se) #计算 log(HR)95%下限
    Hr$upper <- exp(Hr$logHR+qnorm(0.975)*se) #计算 log(HR)95%上限
    HR[[j]] <- Hr
    j = j+1
    i = i+2
  }
  return(HR)
}
```

随后我们根据计算出的log (HR)及其 95%置信区间，对各协变量绘制其随时间风险比变化图，由于绘制曲线图代码较复杂且单一，因此只展示其中一个协变量风险比曲线图绘制代码如下：

```
sL <- 6
sl <- seq(1,sL,by = 0.25)
nsl <- length(sl)

#obstruct
par(xaxs = 'i',yaxs = 'i',mar = c(5,5,3,2))
plot(HR[[1]]$sl,exp(HR[[1]]$logHR),type='l',lwd=6,xlim=c(0,sL),ylim=c(0,4),
     bty='l',xaxt='n',yaxt='n',xlab='',ylab='')
axis(2,las=1,pos=0,cex.axis=1.3,tcl=-0.4,hadj=0.9,lwd=1.6)
axis(1,las=1,pos=0,cex.axis=1.3,tcl=-0.4,hadj=-0.3,lwd=1.6)
title(main = list('A.Obstruction of colon by tumour at diagnosis',font=7,cex=1.6),line=1)
title(xlab = 'Prediction time in years',font.lab=7,cex.lab=1.5,line=2.6)
title(ylab = '3 years Hazard Ratio',font.lab=7,cex.lab=1.3,line=3)
lines(HR[[1]]$sl,HR[[1]]$lower,type='l',lty=2,lwd=6)
lines(HR[[1]]$sl,HR[[1]]$upper,type='l',lty=2,lwd=6)
abline(h=1,lwd=4,lty=3)
legend(0.5,4,c('YES/NO'),col=1,cex=1.4,bty='n',lwd=6)
```

由于曲线图结果较多，此处仅对部分图进行结果解读，其余输出图均可见附录。



由图 C 可得，患阳性淋巴结 4 个以上的患者均比患阳性淋巴结少于 4 个的患者死亡风险比高，且随着时间的推移该风险比在逐年降低。然而，在 3 年之后，两者风险比差异没有统计学意义，说明在患结肠癌 3 年之后，是否患有超过 4 个阳性淋巴结的患者其死亡风险比与该因素没有关系。由图 D2 可得，接受$Lev+5-FU$治疗的患者比观察组(Obs)的患者死亡风险比低，且在 3 年之后，二者风险比差异没有统计学意义，说明在患结肠癌 3 年之后，接受$Lev+5-FU$治疗的患者与观察组(Obs)的患者死亡风险比与治疗方式的不同没有关系。

4 讨论

4.1 模拟预测

为验证时间效应的Cox动态预测模型的性能优势，我们从 906 例受试者中随机抽取两名患者分别使用Cox模型与时间效应的Cox动态模型预测各自未来 6 年内的生存率，进一步说明两者的差异程度与优势比较。两名患者相关基线数据如下：

Patient	Variables					
	obstruct	adhere	node4	rx	differ	extent
A	0	0	1	3	3	2
B	0	0	0	3	3	2

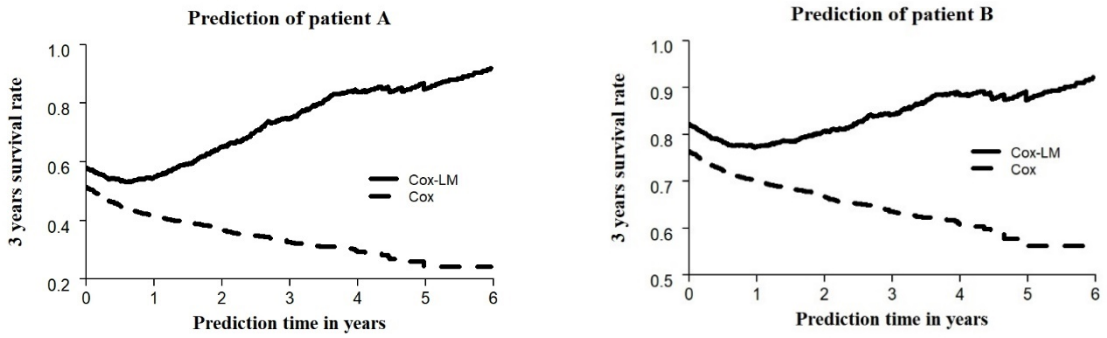
由于二者预测代码相似，在此只展示 A 患者预测代码如下：

```
#install.packages('dynpred')
library(dynpred)
w = 3
sL = 6
tt<-data$time[data$status==1]
tt<-sort(unique(c(0,tt,tt-w)))
tt<-tt[tt>=0 & tt<= sL]
```

```
#利用哑变量进行 Cox 回归
model1 <- coxph(Surv(time,status) ~ obstruct+adhere+node4+rx1+rx2+differ1+differ2+
                        extent1+extent2+extent3+cluster(id),data = data)

#Patient A
dataA <- data[459,]
#Cox 模型的预测
predict_modelA1<-data.frame(time=tt,surv=NA)
out<-summary(survfit(model1,newdata=dataA, type="aalen"))
for (i in 1:length(tt)){
   tmp<-evalstep(out$time,out$surv,c(tt[i],tt[i]+w),subst=1)
   predict_modelA1[i,2]<-tmp[2]
}
#动态 Cox 模型的预测
predict_modelA2<-data.frame(time=tt,surv=NA)
for (i in 1:length(tt)) {
   dt<-define_time(dataA,time1=tt[i]/sL,fixed=c('obstruct','adhere','node4',
                                                'rx1','rx2','differ1','differ2',
                                                'extent1','extent2','extent3'))
   out<-summary(survfit(model2,newdata=dt))
   tmp<-evalstep(out$time,out$surv,c(tt[i],tt[i]+w),subst=1)
   predict_modelA2[i,2]<-tmp[2]/tmp[1]
}
#绘制 COX 与动态 COX 预测曲线图
par(xaxs="i",yaxs="i",mar=c(5,6,3,2))
plot(predict_modelA1$time,predict_modelA1$surv,type='s',lwd=6,lty=2,
       xlim=c(0,sL),ylim=c(0.2,1),bty='l',xaxt ='n',yaxt='n',xlab ='',ylab='')
axis(1,las=1,pos=0.2,cex.axis=1.3,tcl=-0.4,padj=-0.3,lwd=1.6)
axis(2,las=1,pos=0,cex.axis=1.3,tcl=-0.4,padj=0.9,lwd=1.6)
title(main=list('Prediction of patient A',font=7,cex=1.6),line=1.3)
title(xlab='Prediction time in years',font.lab=7,cex.lab=1.5,line=2.4)
title(ylab='3 years survival rate',font.lab=7,cex.lab=1.5,line=4)
lines(predict_modelA2$time,predict_modelA2$surv,type='s',lwd=6,lty=1)
legend(4,0.6,c('Cox-LM','Cox'),lty=1:2,bty='n',lwd=6,xpd=TRUE,cex=1.1)
```

图中实线为时间效应的Cox动态预测结果，虚线为Cox比例风险模型预测的由基线开始的累计生存概率曲线，我们可以明显看出实质上仅有 1 年内二者的趋势基本相同，而后续动态生存率的波动实际上反映了由于患有 4 个阳性淋巴结导致的死亡风险比的下降，其生存率反而在不断上升，这一方面和Cox比例风险模型的结果也是暗中呼应的。在动态预测曲线中生存率提高较快的部分对应的虚线下降趋势也同样较缓，但其比Cox比例风险模型更加直观、精准且灵活，更能动态反映患者生存率随时间的变化，在实际中更加具有应用意义。



4.2 所构建函数不足点

在所构建的函数$LM()$中，目前该函数仅限于对基线数据进行处理，对于纵向数据的处理尚不能解决；在$define_time()$函数中，对于各协变量与时间的交互项仅考虑了一次交互，且时间函数仅考虑二次函数形式，因此对于此函数可以考虑增加参数设置，满足不同使用者对不同交互项构建的需求。

4.3 模型方法不足点

对于界标模型，该方法存在一定的局限性[4]：1、界标时间点选择随意。目前关于界标时间点的设定缺少统一的标准，实际中往往根据研究实际情况（如有临床意义的时间）设置。如：根据高危人群的变化节点来设定界标时间或将随访时间点设定为界标时间，这使得界标法预测结果存在数据驱动的可能性。同时，对于没有明确定义基线时间或左删失的数据信息，界标时间点的选择和模型的建立都存在挑战。2、可能导致偏倚的因素较多，如界标模型不考虑测量误差，由于观察值和真实值存在差异从而导致偏倚。

5 参考文献

[1] Yang Z, Hou Y, Lyu J, Liu D, Chen Z. Dynamic prediction and prognostic analysis of patients with cervical cancer: a landmarking analysis approach. Ann Epidemiol. 2020;44:45-51. doi:10.1016/j.annepidem.2020.01.009.

[2] 杨锋,陈欣,尤东方,等. 基于现实世界研究中临床随访数据的两种动态预测建模方法的实证研究 [J]. 中国临床医学, 2021, 28 (05): 751-756.

[3] 周江杰,王胜锋.界标模型介绍及在动态预测中的应用[J].中华流行病学杂志,2022,43(1):112-117.

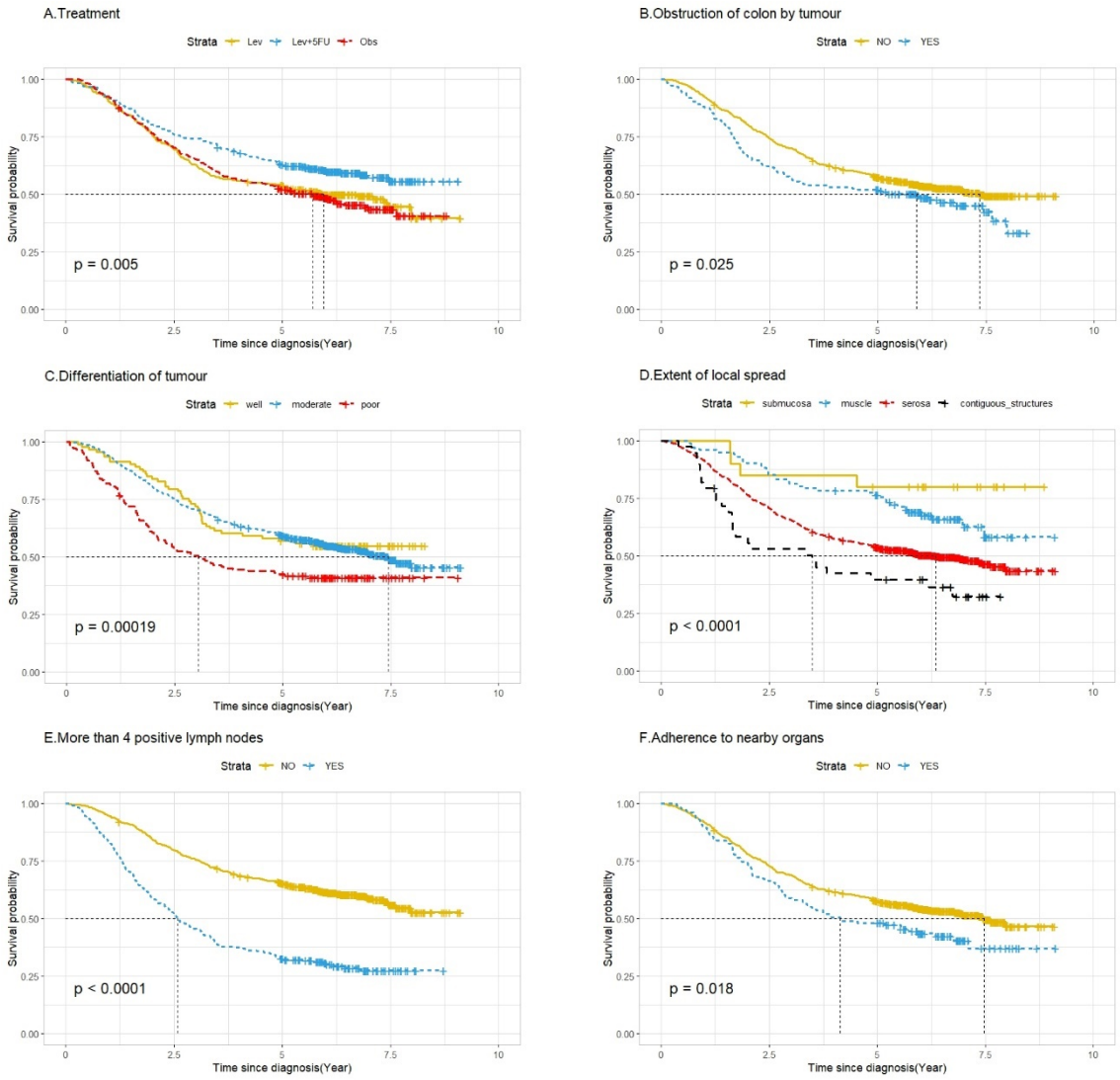
[4] 宋若齐,吴疏桐,王闯世.医学研究中常见动态预测模型方法介绍[J].中国循证医学杂志,2022,22(10):1224-1232.

6 附录

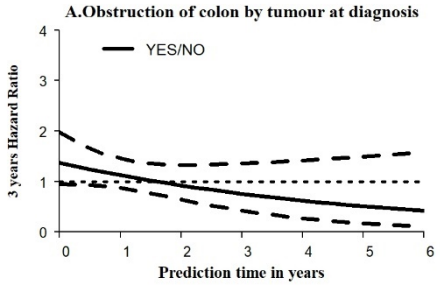
6.1 基线统计表

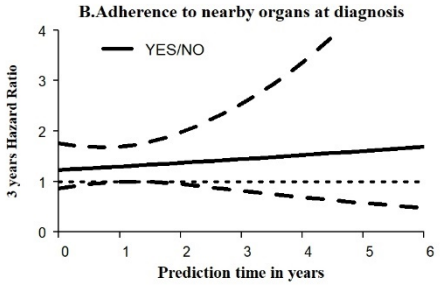
协变量	协变量解释与赋值	样本例数	存活例数 (N=465)	死亡例数 (N=441)
治疗方案	Lev = 1	300	146 (31.4%)	154 (34.9%)
	Lev+5FU = 2	298	176（37.8%）	122（27.7%）
	Obs	308	143（30.8%）	165（37.4%）
性别	女 = 0	438	225 (48.4%)	213 (48.3%)
	男 = 1	468	240 (51.6%)	228 (51.7%)
年龄	≤60 岁= 0	431	232 (49.9%)	199 (45.1%)
	＞60 岁= 1	475	233 (50.1%)	242 (54.9%)
肿瘤阻塞结肠	无 = 0	731	385 (82.8%)	346 (78.5%)
	有 = 1	175	80 (17.2%)	95 (21.5%)
穿孔	无 = 0	879	453 (97.4%)	426 (96.6%)
	有 = 1	27	12 (2.6%)	15 (3.4%)
粘连	无 = 0	775	410 (88.2%)	365 (82.8%)
	有 = 1	131	55 (11.8%)	76 (17.2%)
分化程度	良好 = 1	93	51 (11%)	42 (9.5%)
	中等 = 2	663	352 (75.7%)	311 (70.5%)
	较差 = 3	150	62 (13.3%)	88 (20%)
扩散程度	粘膜下 = 1	20	16 (3.4%)	4 (0.9%)
	肌肉 = 2	102	67 (14.4%)	35 (7.9%)
	浆膜 = 3	745	368 (79.1%)	377 (85.5%)
	毗连结构 = 4	39	14 (3%)	25 (5.7%)
阳性淋巴结个数	＜4 = 0	654	391 (84.1%)	263 (59.6%)
	≥4 = 1	252	74 (15.9%)	178 (40.4%)

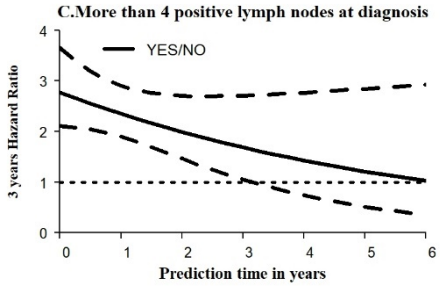
6.2 具有统计学意义的各协变量生存曲线及$Log-rank$检验

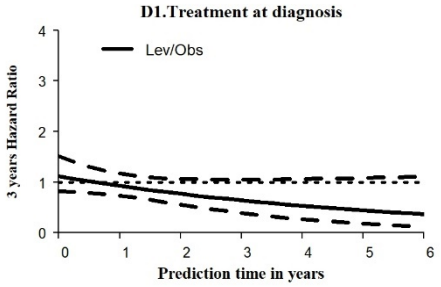


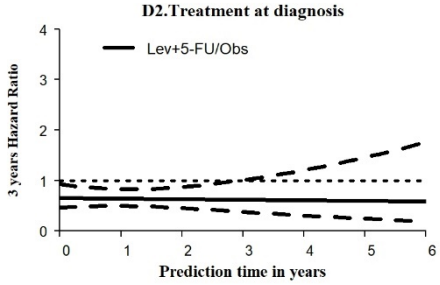
6.3 各协变量随时间风险比变化曲线图

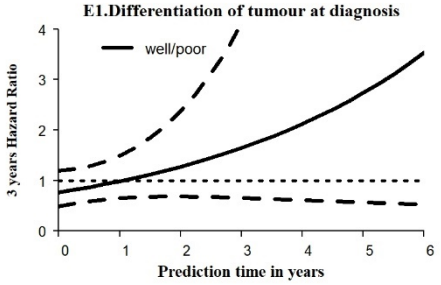


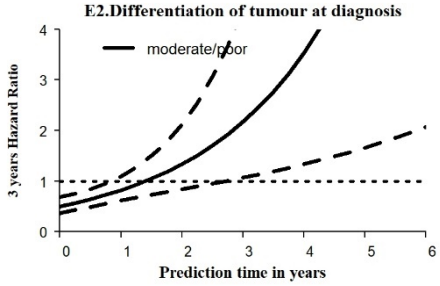


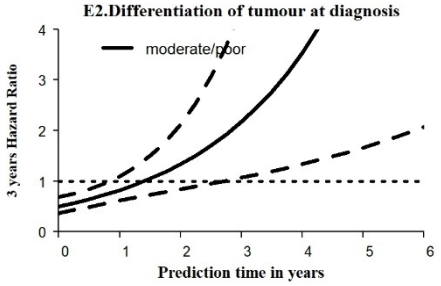


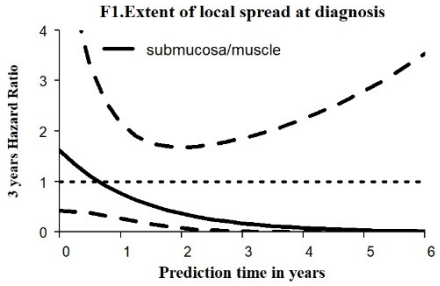


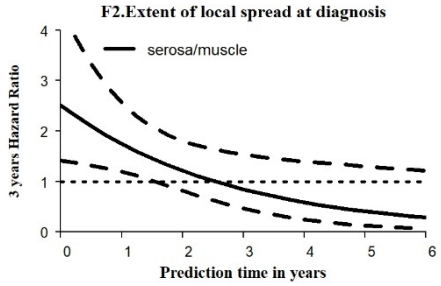


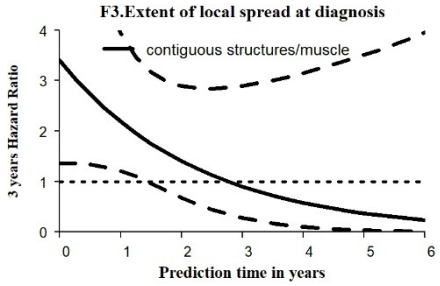












6.4 程序

```
########################################
# 项目名称：基于 Cox 回归的动态预测模型
# 软件版本：R 4.4.0
# 撰写日期：2024 年 6 月 14 日
########################################

#install.packages('survival')
#install.packages('survminer')
#install.packages('tidyverse')
#install.packages('autoReg')
#install.packages('dplyr')
#install.packages('readr')
library(survival)   #用于生存分析
library(survminer) #用于生存分析作图
library(tidyverse) #用于数据处理
library(autoReg)    #用于表格可视化
library(dplyr)
library(readr)       #用于导入数据

#------------------------------
#   PART 1
#
#   数据导入与预处理
#------------------------------

#数据导入
data <- read_csv("ex1.csv")

#数据预处理
data <- na.omit(data) #删除缺失值
data <- data %>% select(-id) %>% mutate(id = c(1:906)) #重新定义每个 id
data <- data %>% mutate(time = time/365) #将 time 修改为年份为单位
data$rx = factor(data$rx,levels = c('Lev','Lev+5FU','Obs'),
```

```
                        labels = c(1,2,3)) #将变量因子化,便于后续处理
data$age <- ifelse(data$age <= 60,0,1) #将 age 变量二分类
data <- data %>% mutate(age = factor(age),sex = factor(sex),
                        obstruct = factor(obstruct),perfor = factor(perfor),
                        adhere = factor(adhere),differ = factor(differ),
                        extent = factor(extent),
                        node4 = factor(node4)) #变量因子化便于统计描述


#---------------------------------------
#   PART 2
#
#   基线统计描述、变量筛选
#---------------------------------------
#基线统计描述
data1 <- data %>% select(-c(time,id)) #由于 time,id 两变量不做统计描述而剔除
table = gaze(status ~.,data = data1) %>% myft() #基于是否发生事件结局分类描述
table

#变量筛选
fit <- coxph(Surv(time,status)~sex+age+rx+obstruct+perfor+adhere+differ+
                 extent+node4+cluster(id),data = data) #建立全因子 Cox 比例风险模型
result <- autoReg(fit,uni = T,threshold = 0.1) %>% myft() #利用单因素 Cox 回归进行变量筛选
result
data <- data %>% select(-c(sex,age,perfor)) #根据筛选结果剔除没有统计学意义的变量


#---------------------------------------------------------------
#   PART 3
#
#   多元 Cox 模型的建立以及各协变量生存曲线图
#---------------------------------------------------------------

#建立多元 Cox 模型
model1 <- coxph(Surv(time,status)~rx+obstruct+adhere+differ+extent+node4+
                    cluster(id),data = data)
cox.zph(model1) #检验是否满足等比例风险假定

#绘制具有统计学意义的协变量生存曲线图(包含 log-rank 检验)
#rx:治疗方案
fit_rx <- survfit(Surv(time, status) ~ rx, data = data)
ggsurvplot(fit_rx,data = data,
           pval = TRUE, #  添加 log-rank 检验的 p 值
```

```
                linetype = 'strata', #  改变不同组别的生存曲线的线性
                surv.median.line = 'hv', #  标注出中位生存时间
                ggtheme = theme_light(),
                legend.labs = c('Lev', 'Lev+5FU','Obs'), #  改变图例标签
                palette = c('#E7B800', '#2E9FDF','red'),
                title = 'A.Treatment',xlab = 'Time since diagnosis(Year)')

#obstruct:肿瘤是否阻塞结肠
fit_obstruct <- survfit(Surv(time, status) ~ obstruct, data = data)
ggsurvplot(fit_obstruct,data = data,
                pval = TRUE,
                linetype = 'strata',
                surv.median.line = 'hv',
                ggtheme = theme_light(),
                legend.labs = c('NO', 'YES'),
                palette = c('#E7B800', '#2E9FDF'),
                title = 'B.Obstruction of colon by tumour',
                xlab = 'Time since diagnosis(Year)')

#differ:肿瘤分化程度
fit_differ <- survfit(Surv(time, status) ~ differ, data = data)
ggsurvplot(fit_differ,data = data,
                pval = TRUE,
                linetype = 'strata',
                surv.median.line = 'hv',
                ggtheme = theme_light(),
                legend.labs = c('well', 'moderate','poor'),
                palette = c('#E7B800', '#2E9FDF','red'),
                title = 'C.Differentiation of tumour ',xlab = 'Time since diagnosis(Year)')

#extent:局部扩散范围
fit_extent <- survfit(Surv(time, status) ~ extent, data = data)
ggsurvplot(fit_extent,data = data,
                pval = TRUE,
                linetype = 'strata',
                surv.median.line = 'hv',
                ggtheme = theme_light(),
                legend.labs = c('submucosa', 'muscle','serosa','contiguous_structures'),
                palette = c('#E7B800', '#2E9FDF','red','black'),
                title = 'D.Extent of local spread',
                xlab = 'Time since diagnosis(Year)')

#node4:是否有 4 个以上阳性淋巴结
```

```
fit_node4 <- survfit(Surv(time, status) ~ node4, data = data)
ggsurvplot(fit_node4,data = data,
           pval = TRUE,
           linetype = 'strata',
           surv.median.line = 'hv',
           ggtheme = theme_light(),
           legend.labs = c('NO','YES'),
           palette = c('#E7B800', '#2E9FDF'),
           title = 'E.More than 4 positive lymph nodes',
           xlab = 'Time since diagnosis(Year)')

#adhere:是否与附近器官粘连
fit_adhere <- survfit(Surv(time, status) ~ adhere, data = data)
ggsurvplot(fit_adhere,data = data,
           pval = TRUE,
           linetype = 'strata',
           surv.median.line = 'hv',
           ggtheme = theme_light(),
           legend.labs = c('NO','YES'),
           palette = c('#E7B800', '#2E9FDF'),
           title = 'F.Adherence to nearby organs',
           xlab = 'Time since diagnosis(Year)')


#----------------------
#   PART 4
#
#   创建哑变量
#----------------------
#变量数值化,便于后续处理
data <- data %>% mutate(obstruct = as.numeric(obstruct),
                        adhere = as.numeric(adhere),node4 = as.numeric(node4),
                        rx = as.numeric(rx),differ = as.numeric(differ),
                        extent = as.numeric(extent),id = as.numeric(id))

#将多分类变量哑变量化
data$rx1 <- ifelse(data$rx == 1,1,0)
data$rx2 <- ifelse(data$rx == 2,1,0) #以'Obs'治疗方式为对照
data$obstruct <- ifelse(data$obstruct == 1,0,1) #以'NO'为对照
data$adhere <- ifelse(data$adhere == 1,0,1) #以'NO'为对照
data$differ1 <- ifelse(data$differ == 1,1,0)
data$differ2 <- ifelse(data$differ == 2,1,0) #以'poor'为对照
data$extent1 <- ifelse(data$extent == 1,1,0)
```

```
data$extent2 <- ifelse(data$extent == 3,1,0)
data$extent3 <- ifelse(data$extent == 4,1,0) #以'muscle'为对照
data$node4 <- ifelse(data$node4 == 1,0,1) #以'NO'为对照
data <- data %>% select(-c(rx,differ,extent))


#---------------------------------
#   PART 5
#
#   超预测数据集的构建
#---------------------------------

LM <- function(w,by,data,time = 'time',status = 'status'){
  # 构建综合数据集的计算函数
  # 输入变量:
  #     w:预测时间(基于疾病病程长短决定)
  #     by:界标步长
  #     data:基线生存数据集
  #     time:数据集中生存时间变量名
  #     status:数据集中结局状态变量名
  #
  # 输出变量:
  #     LMdata:各界标数据集叠加所得的综合数据集

  sL <- max(data$time) - w
  sl <- seq(0,sL,by = by)
  nsl <- length(sl) #界标数据集的个数
  sldata <- list() #将得到的界标数据集放入列表中,便于后续提取叠加数据集
  for(i in 1:nsl){
    data <- data %>% filter(time > sl[i]) %>% mutate(LM = sl[i])
    if (nrow(data) > 0) {
      sldata[[i]] <- data #如果界标数据集不为空,则添加到列表中
    } else {}
  }

  #界标数据集的叠加,构建超预测数据集
  LMdata <- data.frame(array(,dim = c(0,0))) #定义超预测数据集
  for (i in 1:nsl) {
    LMdata = rbind(LMdata,sldata[[i]]) #将所有界标数据集叠加得到超预测数据集
  }
  LMdata1 <- LMdata %>% filter(LM+w <= time) %>%
    mutate(time = LM+w,status = 0)
  LMdata <- LMdata %>% filter(LM+w > time)
```

```
  LMdata <- rbind(LMdata1,LMdata)
}

#利用示例数据创建超预测数据集
LMdata <- LM(w = 3,by = 0.25,data,time = 'time',status = 'status')


#-----------------------------------------------------------------
#   PART 6
#
#   'dynpred'包中创建界标数据集函数的对比检验
#-----------------------------------------------------------------

#install.packages('dynpred')
library(dynpred)
sL = 6
sl <- seq(0,sL,by=0.25)
nsl <- length(sl)
w <- 3
LMdata1 <- data.frame(array(,dim=c(0,0)))
for (i in 1:nsl){
  LMdata1 <- rbind(LMdata1,cutLM(data=data,
                                 outcome=list(time="time",status="status"),
                                 LM=sl[i],horizon=sl[i]+w,
                                 covs=list(fixed=c('id','obstruct','adhere','node4',
                                                   'rx1','rx2','differ1',
                                                   'differ2','extent1','extent2',
                                                   'extent3'),varying=NULL)))
}

LMdata %>% filter(id == 1)   #LM()函数构建的数据集
LMdata1 %>% filter(id == 1) #cutLM()函数构建的数据集


#-----------------------------------------------------------------
#   PART 7
#
#   各协变量与时间的一次交互项以及时间变量的创建
#-----------------------------------------------------------------

define_time <- function(data,fixed,time1){
  # 创建各协变量与时间的一次交互项以及时间变量
  # 输入变量:
```

```
  #     data:数据集
  #     fixed:需要创建交互项的变量名
  #     time1:与协变量交互的时间值
  #
  # 输出变量:
  #     data:创建好交互项与时间变量的数据集

  for (var in fixed) {
    # 计算交互项并添加到数据框
    interaction_name_time <- paste0(var,'1')
    data <- data %>%
      mutate(!!interaction_name_time := .data[[var]] * time1)
  }
  data$LM1 <- time1
  data$LM2 <- time1^2
  return(data)
}

fixed=c('obstruct','adhere','node4','rx1','rx2','differ1','differ2','extent1',
        'extent2','extent3')
LMdata <- define_time(LMdata,fixed=fixed,time1=LMdata$LM/(max(LMdata$time)))


#-----------------------------------------------------
#   PART 8
#
#   建立时间效应的 Cox 动态预测模型
#-----------------------------------------------------

model2 <- coxph(Surv(LM,time,status)~obstruct+obstruct1+adhere+adhere1+node4+
                node41+rx1+rx11+rx2+rx21+differ1+differ11+differ2+differ21+
                extent1+extent11+extent2+extent21+extent3+extent31+LM1+LM2+
                cluster(id),data = LMdata,method="breslow")

#输出各协变量的回归系数,标准误以及 P 值
output<-(cbind(coef = model2$coefficients,
               SE = sqrt(diag(model2$var)),
               PVal = 2-2*pnorm(abs(model2$coefficients/
                                    sqrt(diag(model2$var))))))
round(output,3)


#-----------------------------------------------------------------
```

```
#   PART 9
#
#   协变量随时间 log(HR)的变化以及 95%可信区间
#----------------------------------------------------------------

HR <- function(w,data,by,model,time = 'time'){
  # 协变量随时间风险比的变化以及 95%可信区间计算函数
  # 输入变量:
  #     w:预测窗口
  #     data:综合数据集
  #     by:界标划分步长
  #     model:时间效应的 Cox 动态预测模型
  #     time:数据集中生存时间变量名
  #
  # 输出变量:
  #     HR:各协变量在各界标时间点的 log(HR)及其 95%置信区间列表

  sL = max(data$time) - w
  sl = seq(0,sL,by = 0.25)
  bet <- matrix(c(rep(1,length(sl)),sl/sL),length(sl),2)
  HR <- list() #将各协变量风险率随时间变化的 HR 值放入列表
  i = 1
  j = 1
  while(i <= (length(model$coef)-2)){
    Hr <- data.frame(sl = sl,logHR = as.numeric(bet %*% model$coef[i:(i+1)])) #计算随时间变化
的 log(HR)值
    se <- sqrt(diag(bet %*% model$var[i:(i+1),i:(i+1)] %*% t(bet))) #计算标准误
    Hr$lower <- exp(Hr$logHR-qnorm(0.975)*se) #计算 log(HR)95%下限
    Hr$upper <- exp(Hr$logHR+qnorm(0.975)*se) #计算 log(HR)95%上限
    HR[[j]] <- Hr
    j = j+1
    i = i+2
  }
  return(HR)
}

HR <- HR(w=3,LMdata,0.25,model2,time = 'time')


#----------------------------------------------------------------
#   PART 10
#
#   绘制变量风险比随时间变化图(含 95%可信区间)
```

```
#-------------------------------------------------------------------

sL <- 6
sl <- seq(1,sL,by = 0.25)
nsl <- length(sl)

#obstruct
par(xaxs = 'i',yaxs = 'i',mar = c(5,5,3,2))
plot(HR[[1]]$sl,exp(HR[[1]]$logHR),type='l',lwd=6,xlim=c(0,sL),ylim=c(0,4),
     bty='l',xaxt='n',yaxt='n',xlab='',ylab='')
axis(2,las=1,pos=0,cex.axis=1.3,tcl=-0.4,hadj=0.9,lwd=1.6)
axis(1,las=1,pos=0,cex.axis=1.3,tcl=-0.4,hadj=-0.3,lwd=1.6)
title(main = list('A.Obstruction of colon by tumour at diagnosis',font=7,cex=1.6),line=1)
title(xlab = 'Prediction time in years',font.lab=7,cex.lab=1.5,line=2.6)
title(ylab = '3 years Hazard Ratio',font.lab=7,cex.lab=1.3,line=3)
lines(HR[[1]]$sl,HR[[1]]$lower,type='l',lty=2,lwd=6)
lines(HR[[1]]$sl,HR[[1]]$upper,type='l',lty=2,lwd=6)
abline(h=1,lwd=4,lty=3)
legend(0.5,4,c('YES/NO'),col=1,cex=1.4,bty='n',lwd=6)

#adhere
par(xaxs = 'i',yaxs = 'i',mar = c(5,5,3,2))
plot(HR[[2]]$sl,exp(HR[[2]]$logHR),type='l',lwd=6,xlim=c(0,sL),ylim=c(0,4),
     bty='l',xaxt='n',yaxt='n',xlab='',ylab='')
axis(2,las=1,pos=0,cex.axis=1.3,tcl=-0.4,hadj=0.9,lwd=1.6)
axis(1,las=1,pos=0,cex.axis=1.3,tcl=-0.4,hadj=-0.3,lwd=1.6)
title(main = list('B.Adherence to nearby organs at diagnosis',font=7,cex=1.6),line=1)
title(xlab = 'Prediction time in years',font.lab=7,cex.lab=1.5,line=2.6)
title(ylab = '3 years Hazard Ratio',font.lab=7,cex.lab=1.3,line=3)
lines(HR[[2]]$sl,HR[[2]]$lower,type='l',lty=2,lwd=6)
lines(HR[[2]]$sl,HR[[2]]$upper,type='l',lty=2,lwd=6)
abline(h=1,lwd=4,lty=3)
legend(0.5,4,c('YES/NO'),col=1,cex=1.4,bty='n',lwd=6)

#node4
par(xaxs = 'i',yaxs = 'i',mar = c(5,5,3,2))
plot(HR[[3]]$sl,exp(HR[[3]]$logHR),type='l',lwd=6,xlim=c(0,sL),ylim=c(0,4),
     bty='l',xaxt='n',yaxt='n',xlab='',ylab='')
axis(2,las=1,pos=0,cex.axis=1.3,tcl=-0.4,hadj=0.9,lwd=1.6)
axis(1,las=1,pos=0,cex.axis=1.3,tcl=-0.4,hadj=-0.3,lwd=1.6)
title(main = list('C.More than 4 positive lymph nodes at diagnosis',font=7,cex=1.6),line=1)
title(xlab = 'Prediction time in years',font.lab=7,cex.lab=1.5,line=2.6)
title(ylab = '3 years Hazard Ratio',font.lab=7,cex.lab=1.3,line=3)
```

```
lines(HR[[3]]$sl,HR[[3]]$lower,type='l',lty=2,lwd=6)
lines(HR[[3]]$sl,HR[[3]]$upper,type='l',lty=2,lwd=6)
abline(h=1,lwd=4,lty=3)
legend(0.5,4,c('YES/NO'),col=1,cex=1.4,bty='n',lwd=6)

#rx1
par(xaxs = 'i',yaxs = 'i',mar = c(5,5,3,2))
plot(HR[[4]]$sl,exp(HR[[4]]$logHR),type='l',lwd=6,xlim=c(0,sL),ylim=c(0,4),
     bty='l',xaxt='n',yaxt='n',xlab='',ylab='')
axis(2,las=1,pos=0,cex.axis=1.3,tcl=-0.4,hadj=0.9,lwd=1.6)
axis(1,las=1,pos=0,cex.axis=1.3,tcl=-0.4,hadj=-0.3,lwd=1.6)
title(main = list('D1.Treatment at diagnosis',font=7,cex=1.6),line=1)
title(xlab = 'Prediction time in years',font.lab=7,cex.lab=1.5,line=2.6)
title(ylab = '3 years Hazard Ratio',font.lab=7,cex.lab=1.3,line=3)
lines(HR[[4]]$sl,HR[[4]]$lower,type='l',lty=2,lwd=6)
lines(HR[[4]]$sl,HR[[4]]$upper,type='l',lty=2,lwd=6)
abline(h=1,lwd=4,lty=3)
legend(0.5,4,c('Lev/Obs'),col=1,cex=1.4,bty='n',lwd=6)

#rx2
par(xaxs = 'i',yaxs = 'i',mar = c(5,5,3,2))
plot(HR[[5]]$sl,exp(HR[[5]]$logHR),type='l',lwd=6,xlim=c(0,sL),ylim=c(0,4),
     bty='l',xaxt='n',yaxt='n',xlab='',ylab='')
axis(2,las=1,pos=0,cex.axis=1.3,tcl=-0.4,hadj=0.9,lwd=1.6)
axis(1,las=1,pos=0,cex.axis=1.3,tcl=-0.4,hadj=-0.3,lwd=1.6)
title(main = list('D2.Treatment at diagnosis',font=7,cex=1.6),line=1)
title(xlab = 'Prediction time in years',font.lab=7,cex.lab=1.5,line=2.6)
title(ylab = '3 years Hazard Ratio',font.lab=7,cex.lab=1.3,line=3)
lines(HR[[5]]$sl,HR[[5]]$lower,type='l',lty=2,lwd=6)
lines(HR[[5]]$sl,HR[[5]]$upper,type='l',lty=2,lwd=6)
abline(h=1,lwd=4,lty=3)
legend(0.5,4,c('Lev+5-FU/Obs'),col=1,cex=1.4,bty='n',lwd=6)

#differ1
par(xaxs = 'i',yaxs = 'i',mar = c(5,5,3,2))
plot(HR[[6]]$sl,exp(HR[[6]]$logHR),type='l',lwd=6,xlim=c(0,sL),ylim=c(0,4),
     bty='l',xaxt='n',yaxt='n',xlab='',ylab='')
axis(2,las=1,pos=0,cex.axis=1.3,tcl=-0.4,hadj=0.9,lwd=1.6)
axis(1,las=1,pos=0,cex.axis=1.3,tcl=-0.4,hadj=-0.3,lwd=1.6)
title(main = list('E1.Differentiation of tumour at diagnosis',font=7,cex=1.6),line=1)
title(xlab = 'Prediction time in years',font.lab=7,cex.lab=1.5,line=2.6)
title(ylab = '3 years Hazard Ratio',font.lab=7,cex.lab=1.3,line=3)
lines(HR[[6]]$sl,HR[[6]]$lower,type='l',lty=2,lwd=6)
```

```
lines(HR[[6]]$sl,HR[[6]]$upper,type='l',lty=2,lwd=6)
abline(h=1,lwd=4,lty=3)
legend(0.5,4,c('well/poor'),col=1,cex=1.4,bty='n',lwd=6)

#differ2
par(xaxs = 'i',yaxs = 'i',mar = c(5,5,3,2))
plot(HR[[7]]$sl,exp(HR[[7]]$logHR),type='l',lwd=6,xlim=c(0,sL),ylim=c(0,4),
     bty='l',xaxt='n',yaxt='n',xlab='',ylab='')
axis(2,las=1,pos=0,cex.axis=1.3,tcl=-0.4,hadj=0.9,lwd=1.6)
axis(1,las=1,pos=0,cex.axis=1.3,tcl=-0.4,hadj=-0.3,lwd=1.6)
title(main = list('E2.Differentiation of tumour at diagnosis',font=7,cex=1.6),line=1)
title(xlab = 'Prediction time in years',font.lab=7,cex.lab=1.5,line=2.6)
title(ylab = '3 years Hazard Ratio',font.lab=7,cex.lab=1.3,line=3)
lines(HR[[7]]$sl,HR[[7]]$lower,type='l',lty=2,lwd=6)
lines(HR[[7]]$sl,HR[[7]]$upper,type='l',lty=2,lwd=6)
abline(h=1,lwd=4,lty=3)
legend(0.5,4,c('moderate/poor'),col=1,cex=1.4,bty='n',lwd=6)

#extent1
par(xaxs = 'i',yaxs = 'i',mar = c(5,5,3,2))
plot(HR[[8]]$sl,exp(HR[[8]]$logHR),type='l',lwd=6,xlim=c(0,sL),ylim=c(0,4),
     bty='l',xaxt='n',yaxt='n',xlab='',ylab='')
axis(2,las=1,pos=0,cex.axis=1.3,tcl=-0.4,hadj=0.9,lwd=1.6)
axis(1,las=1,pos=0,cex.axis=1.3,tcl=-0.4,hadj=-0.3,lwd=1.6)
title(main = list('F1.Extent of local spread at diagnosis',font=7,cex=1.6),line=1)
title(xlab = 'Prediction time in years',font.lab=7,cex.lab=1.5,line=2.6)
title(ylab = '3 years Hazard Ratio',font.lab=7,cex.lab=1.3,line=3)
lines(HR[[8]]$sl,HR[[8]]$lower,type='l',lty=2,lwd=6)
lines(HR[[8]]$sl,HR[[8]]$upper,type='l',lty=2,lwd=6)
abline(h=1,lwd=4,lty=3)
legend(0.5,4,c('submucosa/muscle'),col=1,cex=1.4,bty='n',lwd=6)

#extent2
par(xaxs = 'i',yaxs = 'i',mar = c(5,5,3,2))
plot(HR[[9]]$sl,exp(HR[[9]]$logHR),type='l',lwd=6,xlim=c(0,sL),ylim=c(0,4),
     bty='l',xaxt='n',yaxt='n',xlab='',ylab='')
axis(2,las=1,pos=0,cex.axis=1.3,tcl=-0.4,hadj=0.9,lwd=1.6)
axis(1,las=1,pos=0,cex.axis=1.3,tcl=-0.4,hadj=-0.3,lwd=1.6)
title(main = list('F2.Extent of local spread at diagnosis',font=7,cex=1.6),line=1)
title(xlab = 'Prediction time in years',font.lab=7,cex.lab=1.5,line=2.6)
title(ylab = '3 years Hazard Ratio',font.lab=7,cex.lab=1.3,line=3)
lines(HR[[9]]$sl,HR[[9]]$lower,type='l',lty=2,lwd=6)
lines(HR[[9]]$sl,HR[[9]]$upper,type='l',lty=2,lwd=6)
```

```
abline(h=1,lwd=4,lty=3)
legend(0.5,4,c('serosa/muscle'),col=1,cex=1.4,bty='n',lwd=6)

#extent3
par(xaxs = 'i',yaxs = 'i',mar = c(5,5,3,2))
plot(HR[[10]]$sl,exp(HR[[10]]$logHR),type='l',lwd=6,xlim=c(0,sL),ylim=c(0,4),
     bty='l',xaxt='n',yaxt='n',xlab='',ylab='')
axis(2,las=1,pos=0,cex.axis=1.3,tcl=-0.4,hadj=0.9,lwd=1.6)
axis(1,las=1,pos=0,cex.axis=1.3,tcl=-0.4,hadj=-0.3,lwd=1.6)
title(main = list('F3.Extent of local spread at diagnosis',font=7,cex=1.6),line=1)
title(xlab = 'Prediction time in years',font.lab=7,cex.lab=1.5,line=2.6)
title(ylab = '3 years Hazard Ratio',font.lab=7,cex.lab=1.3,line=3)
lines(HR[[10]]$sl,HR[[10]]$lower,type='l',lty=2,lwd=6)
lines(HR[[10]]$sl,HR[[10]]$upper,type='l',lty=2,lwd=6)
abline(h=1,lwd=4,lty=3)
legend(0.5,4,c('contiguous structures/muscle'),col=1,cex=1.4,bty='n',lwd=6)


#-----------------------------------------------------
#   PART 11
#
#   Cox 模型与动态 Cox 模型的模拟预测
#-----------------------------------------------------

#install.packages('dynpred')
library(dynpred)
w = 3
sL = 6
tt<-data$time[data$status==1]
tt<-sort(unique(c(0,tt,tt-w)))
tt<-tt[tt>=0 & tt<= sL]

#利用哑变量进行 Cox 回归
model1 <- coxph(Surv(time,status) ~ obstruct+adhere+node4+rx1+rx2+differ1+differ2+
                        extent1+extent2+extent3+cluster(id),data = data)

#Patient A
dataA <- data[459,]
#Cox 模型的预测
predict_modelA1<-data.frame(time=tt,surv=NA)
out<-summary(survfit(model1,newdata=dataA, type="aalen"))
for (i in 1:length(tt)){
  tmp<-evalstep(out$time,out$surv,c(tt[i],tt[i]+w),subst=1)
```

```
  predict_modelA1[i,2]<-tmp[2]
}
#动态 Cox 模型的预测
predict_modelA2<-data.frame(time=tt,surv=NA)
for (i in 1:length(tt)) {
  dt<-define_time(dataA,time1=tt[i]/sL,fixed=c('obstruct','adhere','node4',
                                               'rx1','rx2','differ1','differ2',
                                               'extent1','extent2','extent3'))
  out<-summary(survfit(model2,newdata=dt))
  tmp<-evalstep(out$time,out$surv,c(tt[i],tt[i]+w),subst=1)
  predict_modelA2[i,2]<-tmp[2]/tmp[1]
}
#绘制 COX 与动态 COX 预测曲线图
par(xaxs="i",yaxs="i",mar=c(5,6,3,2))
plot(predict_modelA1$time,predict_modelA1$surv,type='s',lwd=6,lty=2,
     xlim=c(0,sL),ylim=c(0.2,1),bty='l',xaxt ='n',yaxt='n',xlab ='',ylab='')
axis(1,las=1,pos=0.2,cex.axis=1.3,tcl=-0.4,padj=-0.3,lwd=1.6)
axis(2,las=1,pos=0,cex.axis=1.3,tcl=-0.4,padj=0.9,lwd=1.6)
title(main=list('Prediction of patient A',font=7,cex=1.6),line=1.3)
title(xlab='Prediction time in years',font.lab=7,cex.lab=1.5,line=2.4)
title(ylab='3 years survival rate',font.lab=7,cex.lab=1.5,line=4)
lines(predict_modelA2$time,predict_modelA2$surv,type='s',lwd=6,lty=1)
legend(4,0.6,c('Cox-LM','Cox'),lty=1:2,bty='n',lwd=6,xpd=TRUE,cex=1.1)

#Patient B
dataB <- data[192,]
#Cox 模型的预测
predict_modelB1<-data.frame(time=tt,surv=NA)
out<-summary(survfit(model1,newdata=dataB, type="aalen"))
for (i in 1:length(tt)){
  tmp<-evalstep(out$time,out$surv,c(tt[i],tt[i]+w),subst=1)
  predict_modelB1[i,2]<-tmp[2]
}
#动态 Cox 模型的预测
predict_modelB2<-data.frame(time=tt,surv=NA)
for (i in 1:length(tt)) {
  dt<-define_time(dataB,time1=tt[i]/sL,fixed=c('obstruct','adhere','node4',
                                               'rx1','rx2','differ1','differ2',
                                               'extent1','extent2','extent3'))
  out<-summary(survfit(model2,newdata=dt))
  tmp<-evalstep(out$time,out$surv,c(tt[i],tt[i]+w),subst=1)
  predict_modelB2[i,2]<-tmp[2]/tmp[1]
}
```

```
#绘制 COX 与动态 COX 预测曲线图
par(xaxs="i",yaxs="i",mar=c(5,6,3,2))
plot(predict_modelB1$time,predict_modelB1$surv,type='s',lwd=6,lty=2,
     xlim=c(0,sL),ylim=c(0.5,1),bty='l',xaxt ='n',yaxt='n',xlab ='',ylab='')
axis(1,las=1,pos=0.5,cex.axis=1.3,tcl=-0.4,padj=-0.3,lwd=1.6)
axis(2,las=1,pos=0,cex.axis=1.3,tcl=-0.4,padj=0.9,lwd=1.6)
title(main=list('Prediction of patient B',font=7,cex=1.6),line=1.3)
title(xlab='Prediction time in years',font.lab=7,cex.lab=1.5,line=2.4)
title(ylab='3 years survival rate',font.lab=7,cex.lab=1.5,line=4)
lines(predict_modelB2$time,predict_modelB2$surv,type='s',lwd=6,lty=1)
legend(4,0.8,c('Cox-LM','Cox'),lty=1:2,bty='n',lwd=6,xpd=TRUE,cex=1.1)
```