Stata在医疗健康领域生存分析中的应用

高　培

教授、博士生导师
北京大学公共卫生学院

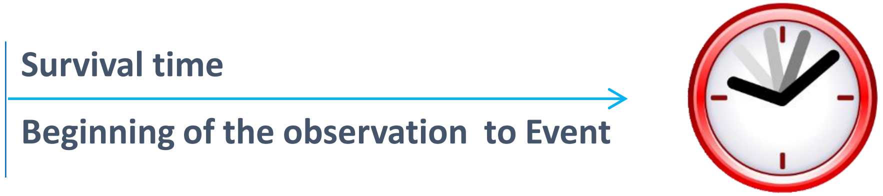
2021年Stata洞察数据科学大会

什么是生存分析？(Survival analysis)

What is survival?

Life span or living process before change of the status, i.e. event.

- Life span: time-to-event
- Event: disease, deaths …


Survival time

Beginning of the observation to Event

生存分析，即描述、测量和分析事件的特征，寻找其发生的原因，并对生存以及到事件发生的时间进行预测的分析方法

生存数据的结构特征

Event 结局事件

- change in status as the underlying outcome measure。例如死亡，特定疾病的发生，婚姻状态的改变，汽车产品的break down等等。

Time-to-event 生存时间

- Time-to-event process。到事件发生的时间被认为是一个随机变量

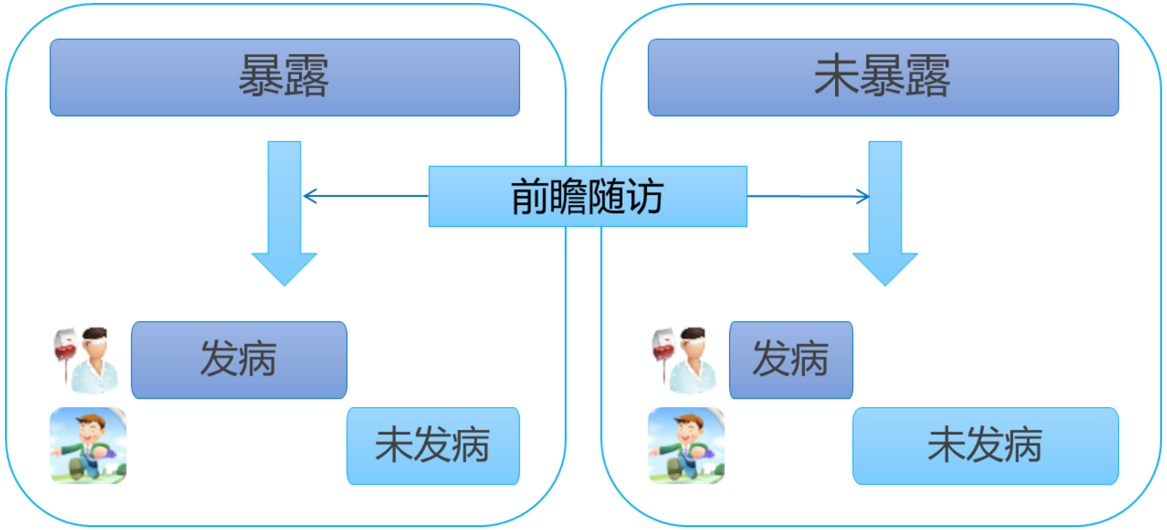
Censoring 截尾

- 研究者只能观察调查窗的两个时间节点之间的相关信息，造成了观察对象在时间窗口外的信息缺失，被称为截尾

Predictors 预测变量（Exposure & Covariates）

- 绝大多数生存分析数据的要素之一：生存时间、截尾状态、和预测变量。

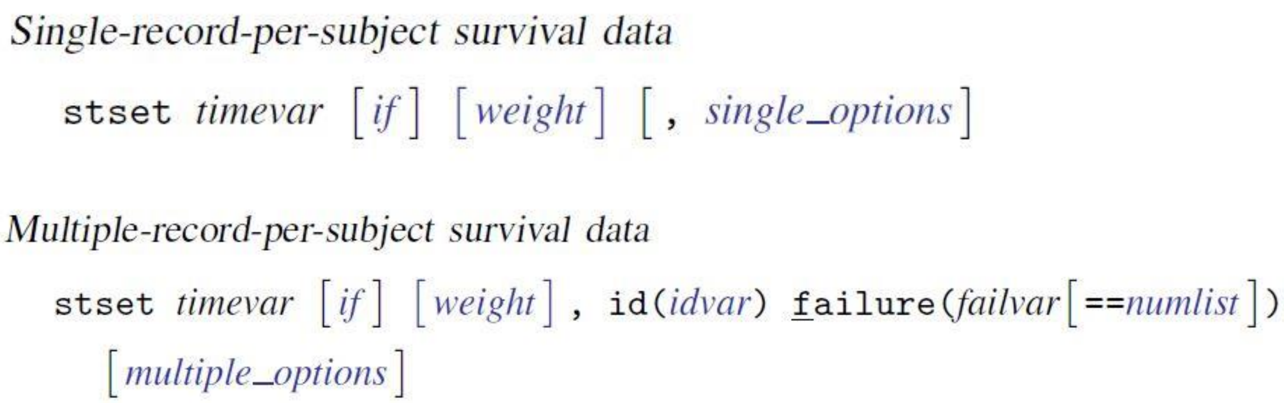
队列研究 (Cohort study)



$$\text{相对危险度RR} = \frac{\text{暴露组发病率}}{\text{非暴露组发病率}}$$

Step 1: stset生存数据的设置

st: survival-time data. **stset**用来告诉stata内存中读入的数据为生存数据，设定重要的生存数据变量：生存时间，结局时间，ID信息等。



Single-record-per-subject survival data

```
stset timevar [if] [weight] [, single_options]
```

Multiple-record-per-subject survival data

```
stset timevar [if] [weight] , id(idvar) failure(failvar[==numlist])
    [multiple_options]
```

failure(failvar[==numlist])	结局事件
origin(time exp)	定义观测样本becomes at risk的时间起始
enter(time exp)	观测样本第一次进入研究的时间节点
exit(time exp)	观测样本离开研究的时间节点
Scale(#)	Rescale时间

Step 1: stset生存数据的设置

- ***At risk***: 样本成为有概率可以发生事件的时间窗。例如，如果结局时间为失业的发生，那么样本成为有概率发生失业的时候是观测样本现在有工作的时候。

- ***Under observation***: 一旦在观测期发生结局事件，该事件将会被观测和记录。有时样本只有在他们at risk之后才会被观测，如某临床试验是针对癌症患者的。

Origin

样本at risk
的时间

Entry

样本进入观察期
的起始时间点

Exit

样本在观察期的
最后时间点

分析时间t

$$t = \frac{time - \texttt{origin()}}{\texttt{scale()}}$$

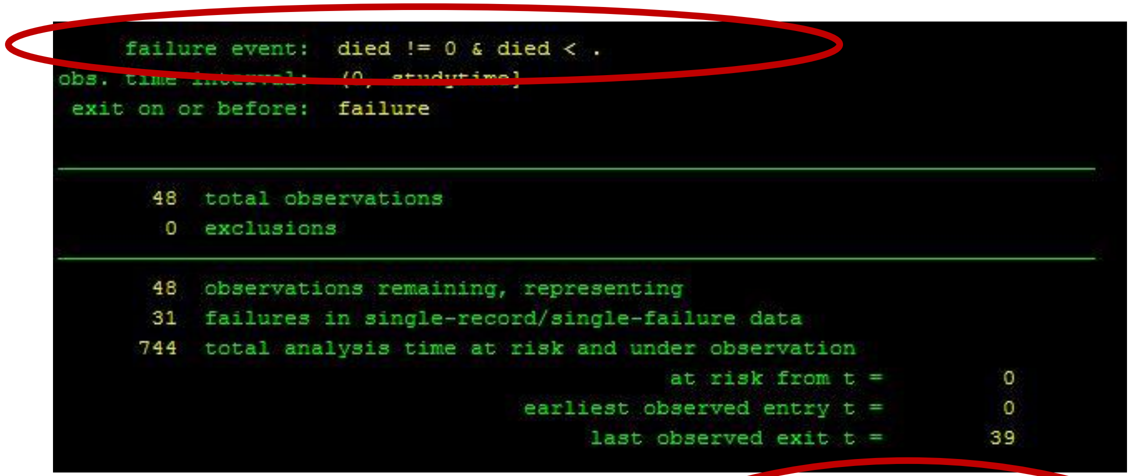
- 一般情况下，entry time = 0, i.e. t=0. 因为time = origin，研究观测开始时样本开始at risk。

- Delayed entry: entry time corresponds to t>0。样本在进入观测前exposed at risk。

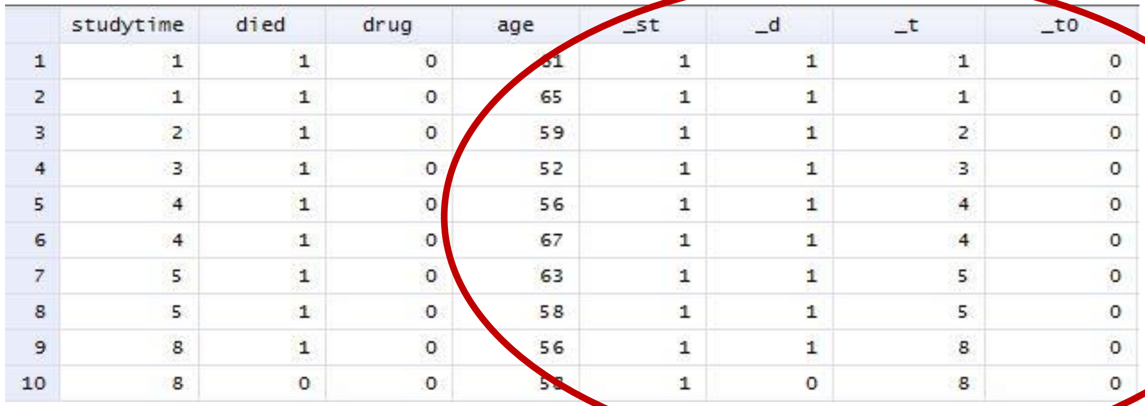
Stata的生存分析

Step 1: stset生存数据的设置

```
webuse drugtr
stset studytime, failure(died)
```



```
     failure event:  died != 0 & died < .
obs. time interval:  (0, studytime]
 exit on or before:  failure

------------------------------------------------------------------------------
         48  total observations
          0  exclusions
------------------------------------------------------------------------------
         48  observations remaining, representing
         31  failures in single-record/single-failure data
        744  total analysis time at risk and under observation
                                                at risk from t =         0
                                     earliest observed entry t =         0
                                          last observed exit t =        39
```



	studytime	died	drug	age	_st	_d	_t	_t0
1	1	1	0	61	1	1	1	0
2	1	1	0	65	1	1	1	0
3	2	1	0	59	1	1	2	0
4	3	1	0	52	1	1	3	0
5	4	1	0	56	1	1	4	0
6	4	1	0	67	1	1	4	0
7	5	1	0	63	1	1	5	0
8	5	1	0	58	1	1	5	0
9	8	1	0	56	1	1	8	0
10	8	0	0	58	1	0	8	0

stset之后，Stata根据命令生成4个系统变量，用于生存分析：

- _st : 符合生存分析设定的样本标志
- _d：生存分析的结局变量，如0或1
- _t：生存时间(time-to-event/censoring)
- _t0: 起始时间

生存分析的几个重要概念

- **Survival function**

$$S(t) = \Pr(T > t)$$

$$\boxed{S(t) = \Pr(T > t) = \int_t^{\infty} f(u)\,du = 1 - F(t).}$$

- **Lifetime distribution function (Probability of event)**

$$F(t) = \Pr(T \le t) = 1 - S(t).$$

- **Event density: rate of failure event per unit of time**

$$f(t) = F'(t) = \frac{d}{dt}F(t).$$

- **Hazard function: event rate per unit time by the number at risk**

$$\lambda(t) = \lim_{dt \to 0} \frac{\Pr(t \le T < t + dt)}{dt \cdot S(t)} = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)}.$$

- **Cumulative hazard function**

$$\Lambda(t) = \int_0^t \lambda(u)\,du$$

$$\boxed{S(t) = \exp(-\Lambda(t))}$$

Kaplan-Meier生存曲线

Step 2: sts描述生存曲线及Hazard function

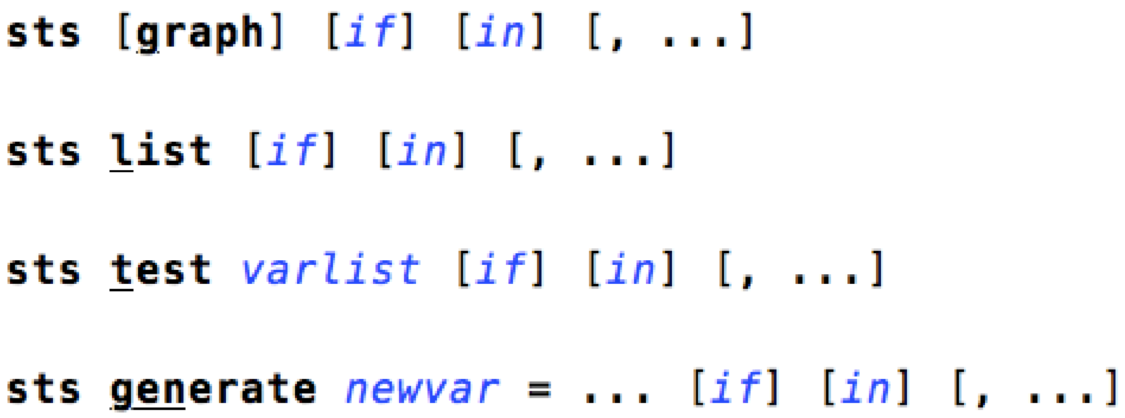
- 乘积极限法（Product-Limit method），基本思想是：将生存时间由小到大依次排列，在每个死亡点上，计算其期初人数、死亡人数、死亡概率、生存概率和生存率。

- 生存率=生存概率的乘积，i.e. $S(t_k) = p_1{*}p_2{*}\ldots{*}p_k$，思想与寿命表法(life table)相同，只不过寿命表法中时间段的划分是人为的、等距的，而乘积极限法划分时间段的分割点是实际死亡发生时间。

- 完全使用经验数据构造生存曲线，是一种非参数方法。

- 既可以适用于小样本，又可以适用于大样本。当然，基于大样本的生存曲线会更合理些，基于小样本的生存曲线的误差可能会比较大。

Step 2: sts描述生存曲线及Hazard function

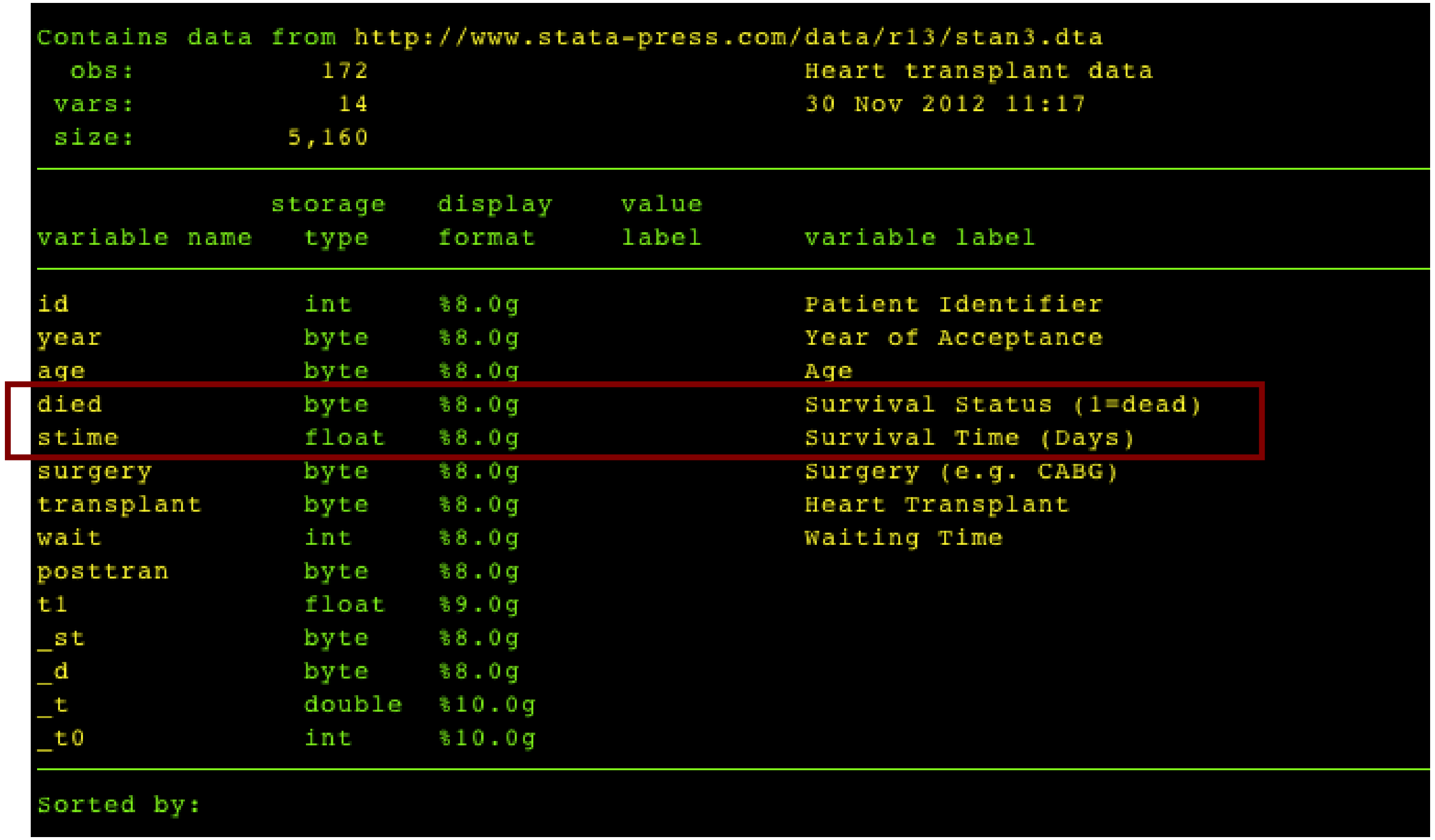


```
sts [graph] [if] [in] [, ...]

sts list [if] [in] [, ...]

sts test varlist [if] [in] [, ...]

sts generate newvar = ... [if] [in] [, ...]
```

- **sts graph** (=**sts**): 绘制生存函数曲线 （Kaplan-Meier）
- **sts list:** 列表生存函数 (或Nelson-Aalen cumulative hazard function)
- **sts test:** 检验生存函数是否相同
- **sts gen:** 生成包含生存函数（或Nelson-Aalen cumulative hazard function)的变量

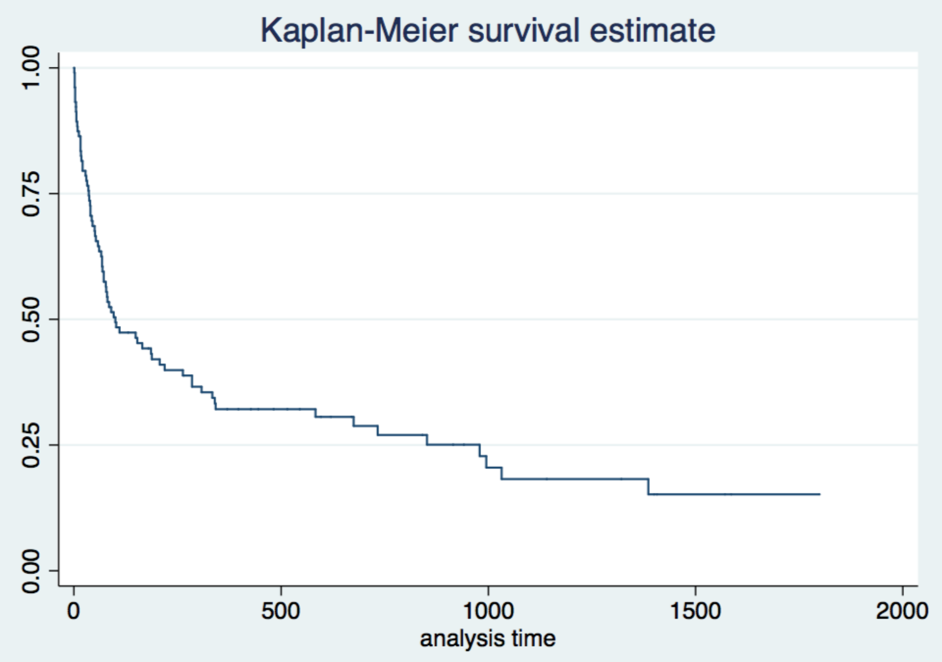
Step 2: sts描述生存曲线及Hazard function

```
webuse stan3, clear
```



```
Contains data from http://www.stata-press.com/data/r13/stan3.dta
  obs:           172                          Heart transplant data
 vars:            14                          30 Nov 2012 11:17
 size:         5,160
--------------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
--------------------------------------------------------------------------------------
id              int     %8.0g                 Patient Identifier
year            byte    %8.0g                 Year of Acceptance
age             byte    %8.0g                 Age
died            byte    %8.0g                 Survival Status (1=dead)
stime           float   %8.0g                 Survival Time (Days)
surgery         byte    %8.0g                 Surgery (e.g. CABG)
transplant      byte    %8.0g                 Heart Transplant
wait            int     %8.0g                 Waiting Time
posttran        byte    %8.0g
t1              float   %9.0g
_st             byte    %8.0g
_d              byte    %8.0g
_t              double  %10.0g
_t0             int     %10.0g
--------------------------------------------------------------------------------------
Sorted by:
```

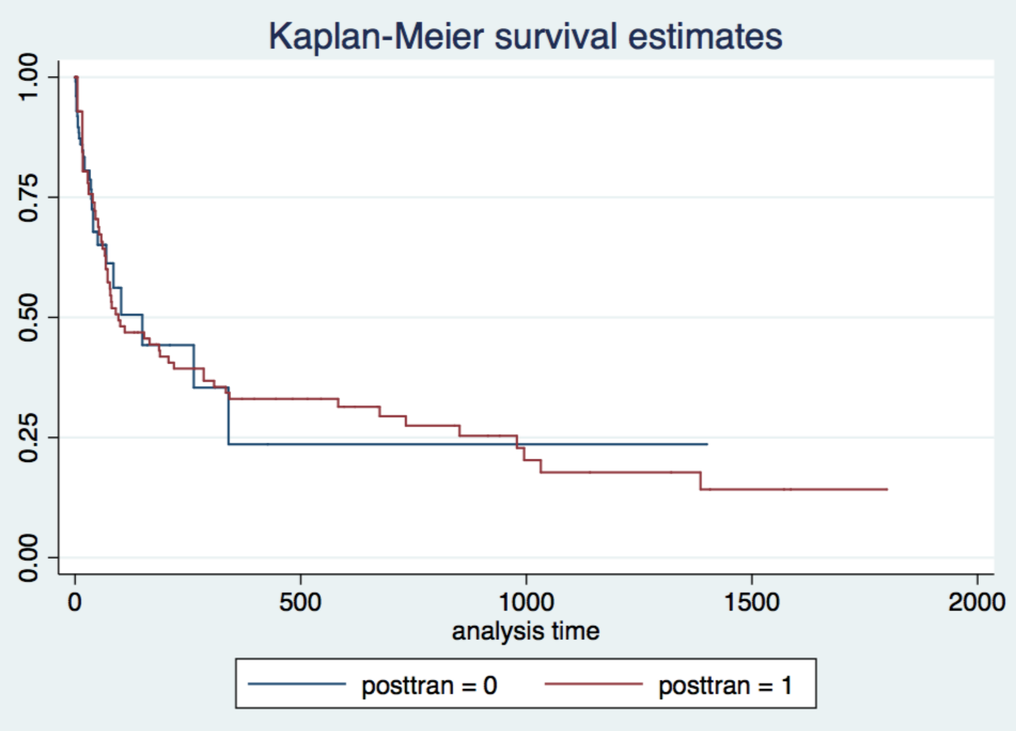
sts graph:绘制生成生存函数等曲线
(Kaplan-Meier or Nelson-Aalen)



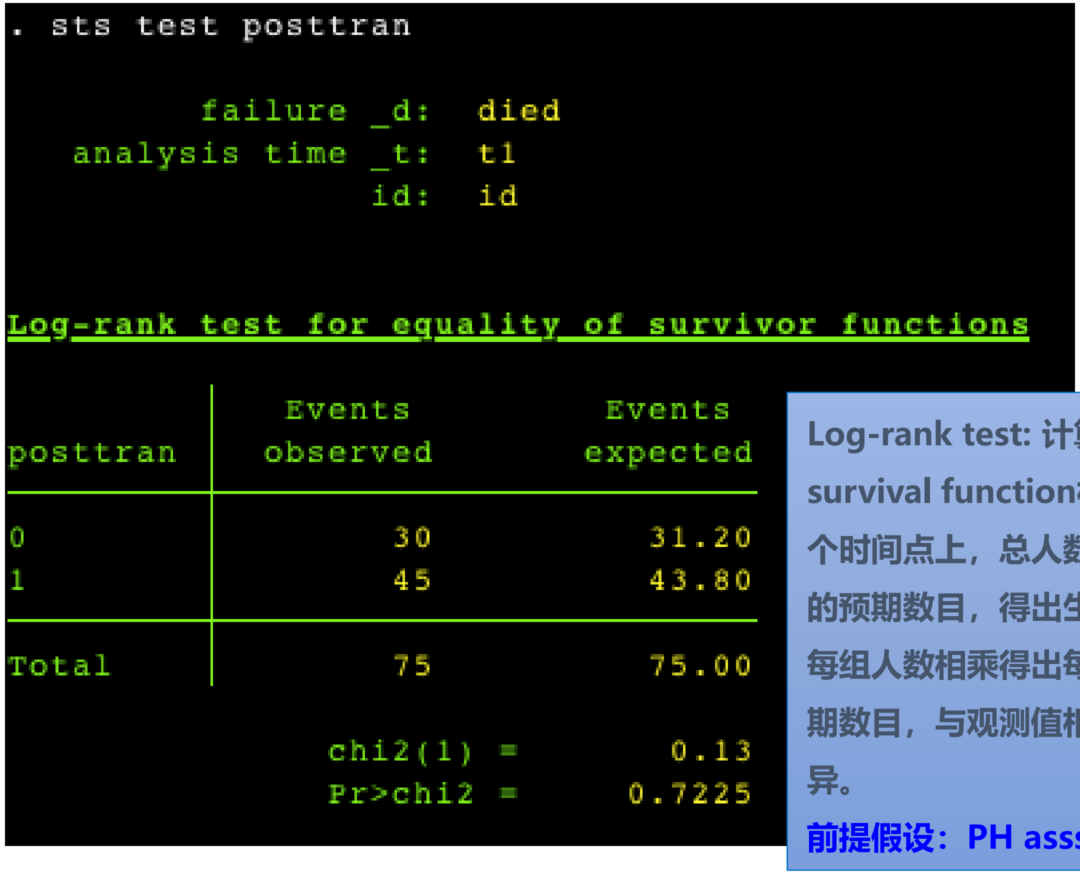
Step 2: sts描述生存曲线及Hazard function

sts graph: by选项分组绘制生成生存函数曲线 (Kaplan-Meier)

```
sts graph, by(posttran)
```



sts test: 检验两个生存函数是否相等（log-rank test）



```
. sts test posttran

         failure _d:  died
   analysis time _t:  t1
                 id:  id


Log-rank test for equality of survivor functions

         |   Events         Events
posttran |  observed       expected
---------+-------------------------
0        |        30          31.20
1        |        45          43.80
---------+-------------------------
Total    |        75          75.00

                   chi2(1) =       0.13
                   Pr>chi2 =     0.7225
```

Log-rank test: 计算如果两组survival function相同时，在每个时间点上，总人数中发生事件的预期数目，得出生存概率，与每组人数相乘得出每组的事件预期数目，与观测值相比是否有差异。

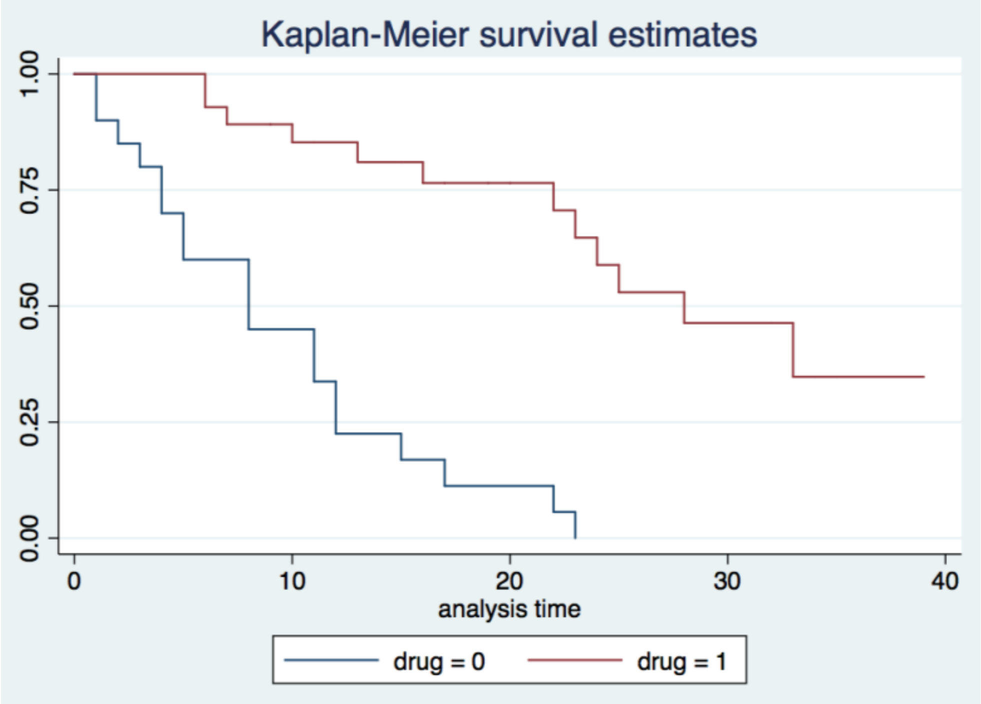
前提假设：PH asssumption

Stata的生存分析

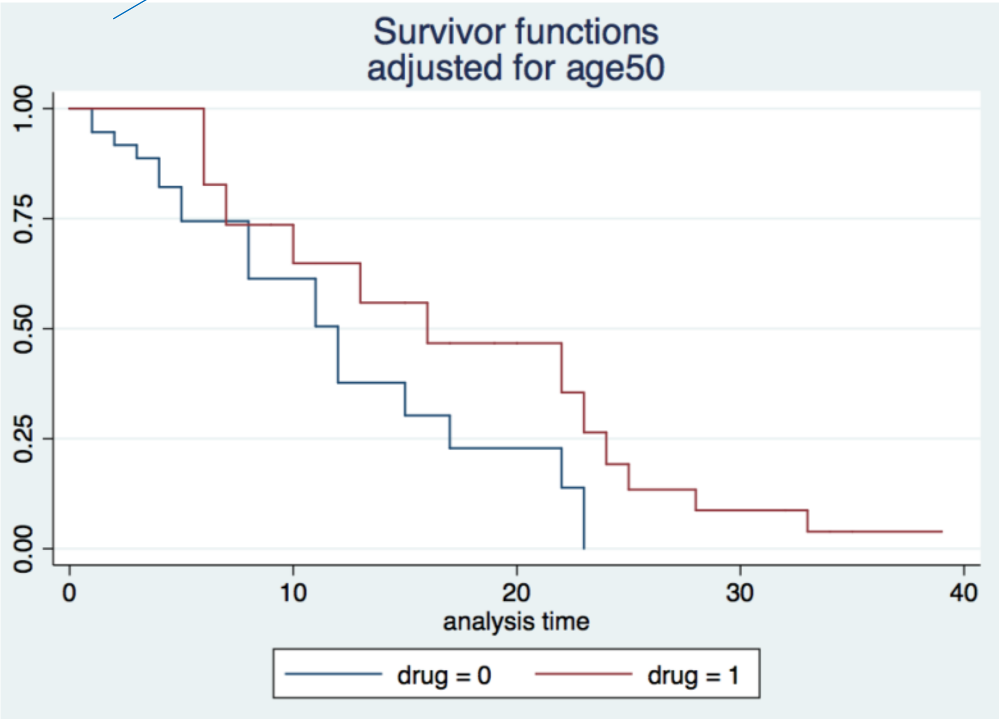
Step 2: sts描述生存曲线及Hazard function

如何调整个别协变量?

生成centered age变量，sts的adjustfor ()选项，默认调整到该变量的0值



```
sts graph, by(drug)
```



```
generate age50 = age-50
sts graph, by(drug) adjustfor(age50)
```

混杂因素

电影的观众有着不同的饮料偏好?

混杂因素

电影的观众有着不同的饮料偏好?

2017.7.27上映

2017.12.15上映

Step 2: sts描述生存曲线及Hazard function

绘制Kaplan-Meier生存曲线	• sts graph • sts graph, by(drug) • sts graph, by(drug) adjustfor(age50)
绘制Cumulative hazard function	• sts graph, cumhaz • sts graph, cumhaz by(drug)
绘制hazard function	• sts graph, hazard • sts graph, hazard by(drug)
列表Kaplan-Meier生存曲线	• sts list • sts list, by(drug) compare
列表Nelson-Aalen cumulative hazard function	• sts list, cumhaz • sts list, cumhaz by(drug) compare
生成KM生存曲线变量	• sts gen surv = s • sts gen surv_by_drug = s, by(drug)
生成NA cumulative hazard function的变量	• sts gen haz = na • sts gen haz_by_drug = na, by(drug)
检验生存曲线是否等同	• sts test drug

比例危险模型Cox Model简介

- ✓ 生存分析中的重要模型之一：用于研究各种因素对于疾病生存期长短的关系，进行多因素分析。
- ✓ 生存期的资料一般不服从正态分布，所以常规的统计方法不使用。

$$h(t,x,c)=h_0(t)\exp(\beta_1 x+\beta_2 c_1+\cdots+\beta_m c_m)$$

Hazard function

Baseline hazard function

Exposure

Confounders

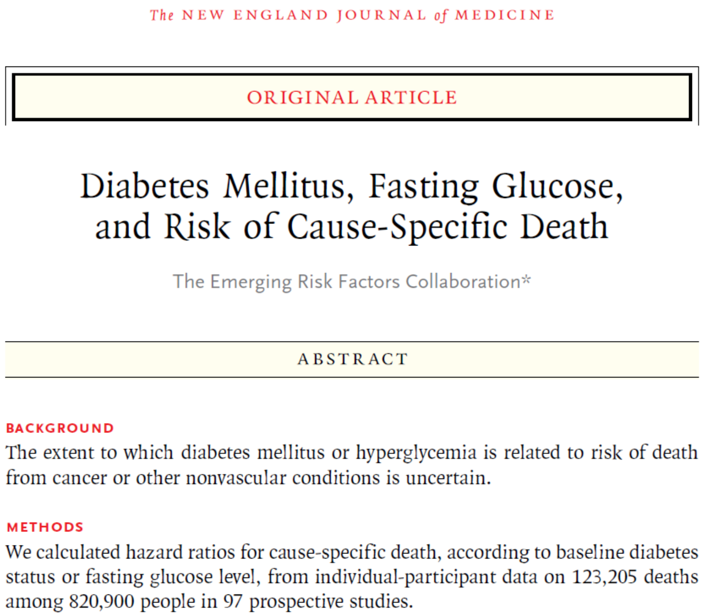
β_i **Log-HR，回归系数，由样本估计而得**

>0表示该协变量是危险因素，越大使生存时间越短

<0表示该协变量是保护因素，越大使生存时间越长

案例

Diabetes mellitus, glycaemia markers and cardiovascular disease



The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Diabetes Mellitus, Fasting Glucose, and Risk of Cause-Specific Death

The Emerging Risk Factors Collaboration*

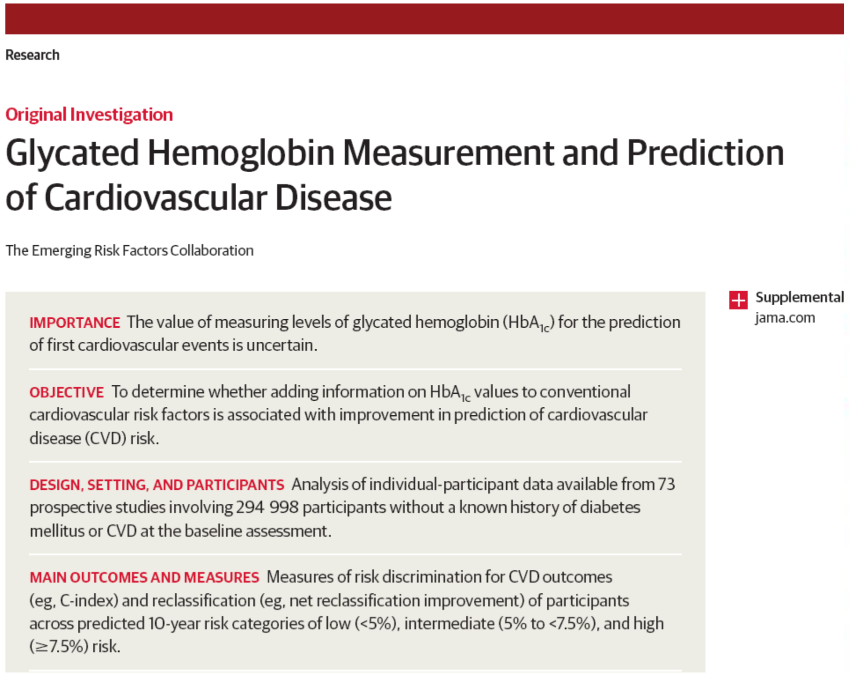
ABSTRACT

BACKGROUND

The extent to which diabetes mellitus or hyperglycemia is related to risk of death from cancer or other nonvascular conditions is uncertain.

METHODS

We calculated hazard ratios for cause-specific death, according to baseline diabetes status or fasting glucose level, from individual-participant data on 123,205 deaths among 820,900 people in 97 prospective studies.



Research

Original Investigation

Glycated Hemoglobin Measurement and Prediction of Cardiovascular Disease

The Emerging Risk Factors Collaboration

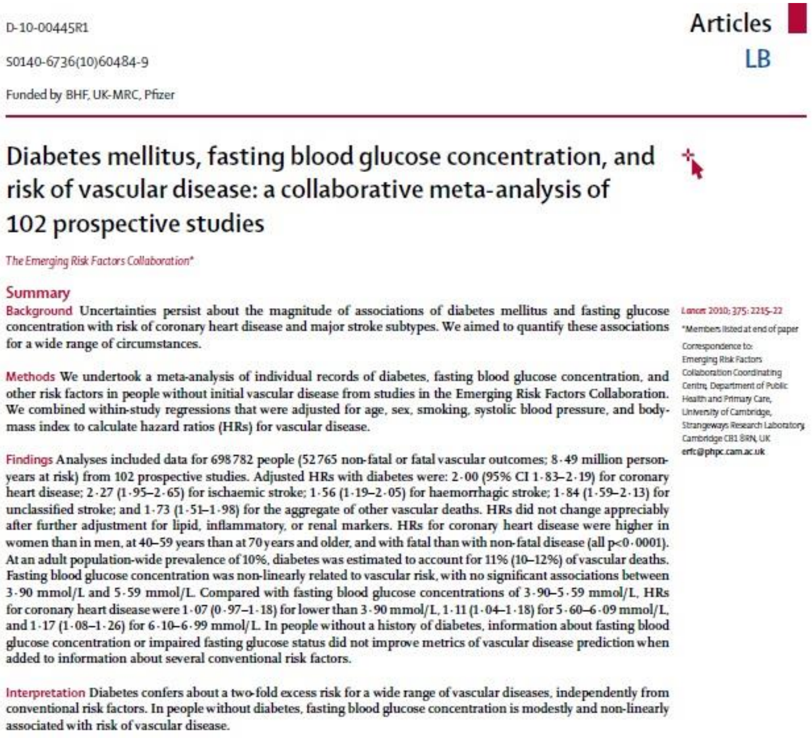
Supplemental
jama.com

IMPORTANCE The value of measuring levels of glycated hemoglobin (HbA_{1c}) for the prediction of first cardiovascular events is uncertain.

OBJECTIVE To determine whether adding information on HbA_{1c} values to conventional cardiovascular risk factors is associated with improvement in prediction of cardiovascular disease (CVD) risk.

DESIGN, SETTING, AND PARTICIPANTS Analysis of individual-participant data available from 73 prospective studies involving 294 998 participants without a known history of diabetes mellitus or CVD at the baseline assessment.

MAIN OUTCOMES AND MEASURES Measures of risk discrimination for CVD outcomes (eg, C-index) and reclassification (eg, net reclassification improvement) of participants across predicted 10-year risk categories of low (<5%), intermediate (5% to <7.5%), and high (≥7.5%) risk.



D-10-00445R1

S0140-6736(10)60484-9

Funded by BHF, UK-MRC, Pfizer

Articles

LB

Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies

*The Emerging Risk Factors Collaboration**

Summary

Background Uncertainties persist about the magnitude of associations of diabetes mellitus and fasting glucose concentration with risk of coronary heart disease and major stroke subtypes. We aimed to quantify these associations for a wide range of circumstances.

Methods We undertook a meta-analysis of individual records of diabetes, fasting blood glucose concentration, and other risk factors in people without initial vascular disease from studies in the Emerging Risk Factors Collaboration. We combined within-study regressions that were adjusted for age, sex, smoking, systolic blood pressure, and body-mass index to calculate hazard ratios (HRs) for vascular disease.

Findings Analyses included data for 698 782 people (52 765 non-fatal or fatal vascular outcomes; 8·49 million person-years at risk) from 102 prospective studies. Adjusted HRs with diabetes were: 2·00 (95% CI 1·83–2·19) for coronary heart disease; 2·27 (1·95–2·65) for ischaemic stroke; 1·56 (1·19–2·05) for haemorrhagic stroke; 1·84 (1·59–2·13) for unclassified stroke; and 1·73 (1·51–1·98) for the aggregate of other vascular deaths. HRs did not change appreciably after further adjustment for lipid, inflammatory, or renal markers. HRs for coronary heart disease were higher in women than in men, at 40–59 years than at 70 years and older, and with fatal than with non-fatal disease (all $p<0·0001$). At an adult population-wide prevalence of 10%, diabetes was estimated to account for 11% (10–12%) of vascular deaths. Fasting blood glucose concentration was non-linearly related to vascular risk, with no significant associations between 3·90 mmol/L and 5·59 mmol/L. Compared with fasting blood glucose concentrations of 3·90–5·59 mmol/L, HRs for coronary heart disease were 1·07 (0·97–1·18) for lower than 3·90 mmol/L, 1·11 (1·04–1·18) for 5·60–6·09 mmol/L, and 1·17 (1·08–1·26) for 6·10–6·99 mmol/L. In people without a history of diabetes, information about fasting blood glucose concentration or impaired fasting glucose status did not improve metrics of vascular disease prediction when added to information about several conventional risk factors.

Interpretation Diabetes confers about a two-fold excess risk for a wide range of vascular diseases, independently from conventional risk factors. In people without diabetes, fasting blood glucose concentration is modestly and non-linearly associated with risk of vascular disease.

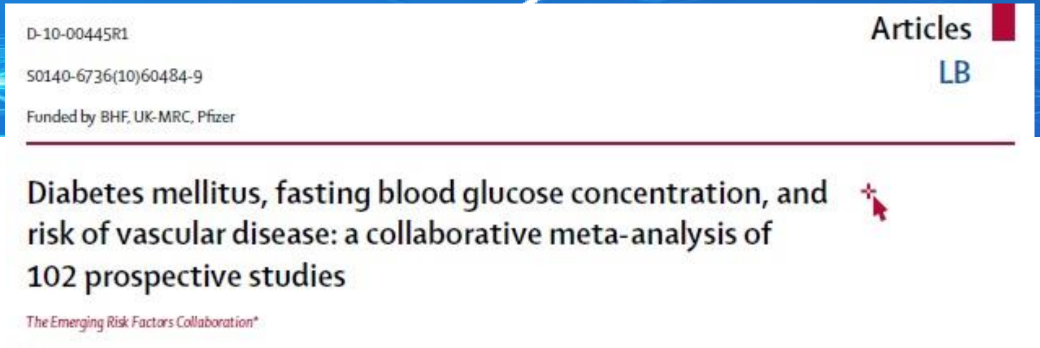
Lancet 2010; 375: 2215–22

*Members listed at end of paper

Correspondence to:
Emerging Risk Factors Collaboration Coordinating Centre, Department of Public Health and Primary Care, University of Cambridge, Strangeways Research Laboratory, Cambridge CB1 8RN, UK
erfc@phpc.cam.ac.uk

Diabetes mellitus, glycaemia markers and CVD

利用重复测量值估计血糖，血脂及血压的长期期望值，并利用长期期望值直接估计危害比(HR)



D-10-00445R1

S0140-6736(10)60484-9

Funded by BHF, UK-MRC, Pfizer

Articles

LB

Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies

*The Emerging Risk Factors Collaboration**

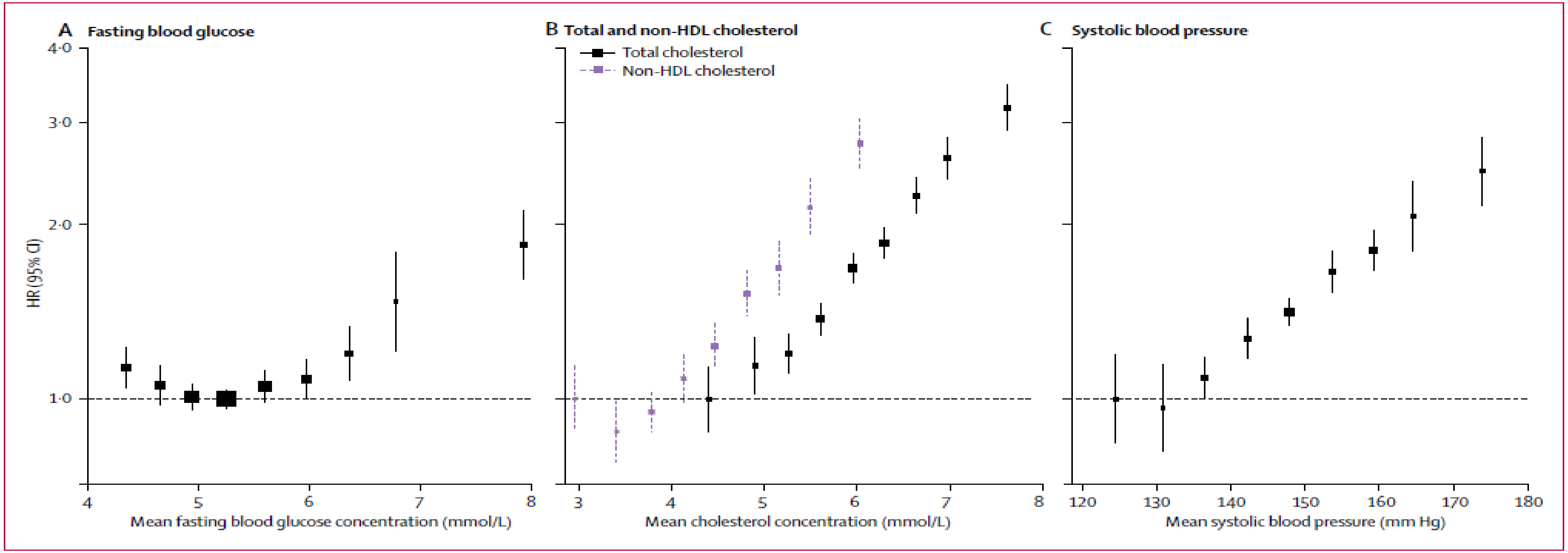
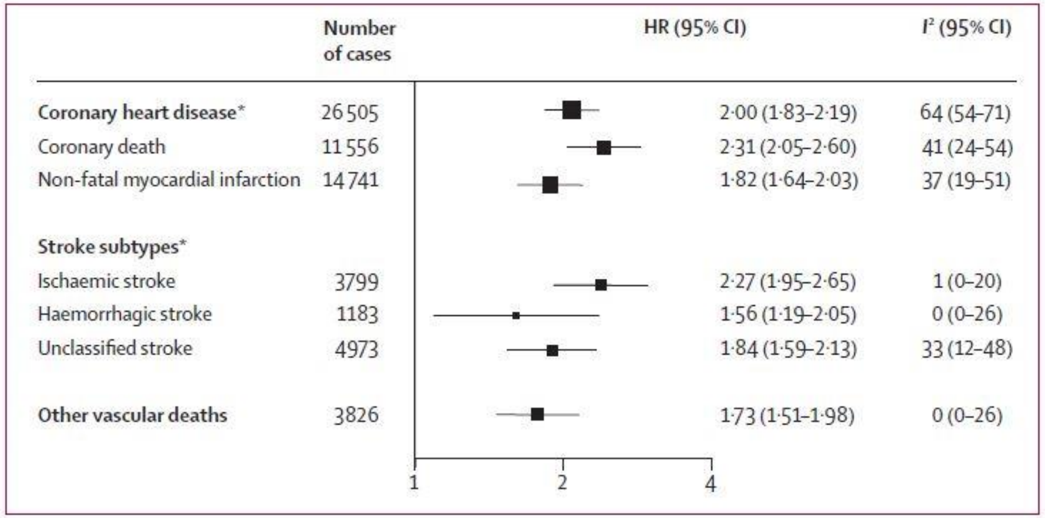


Figure 6: Comparison of hazard ratios (HRs) for coronary heart disease by long-term average concentrations of fasting blood glucose concentration, total (and non-HDL) cholesterol, and systolic blood pressure, in a common set of participants

Cox model for cohorts studies

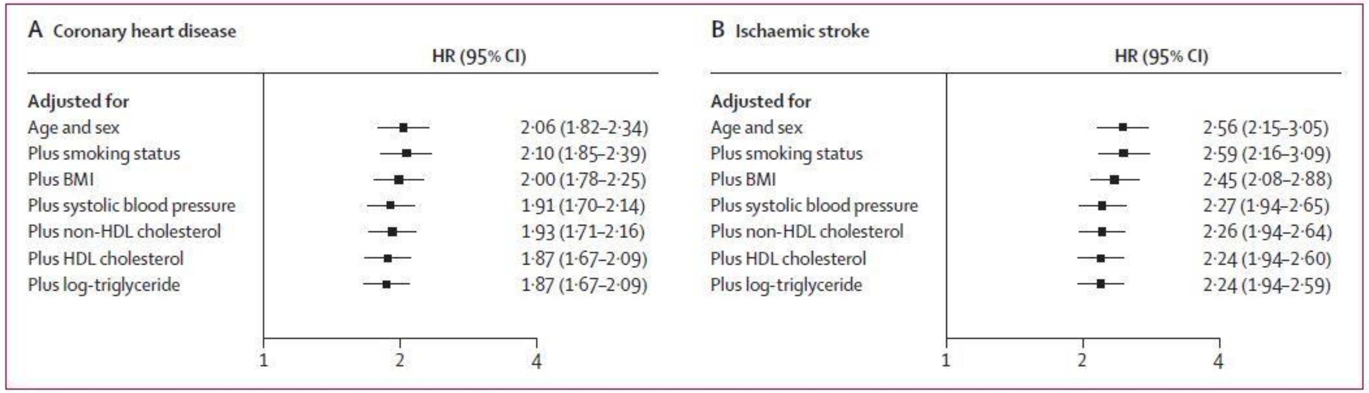
Diabetes mellitus, glycaemia markers and CVD

HRs (95% CI) for different outcomes in people with vs without diabetes



	Number of cases	HR (95% CI)	I^2 (95% CI)
Coronary heart disease*	26 505	2·00 (1·83–2·19)	64 (54–71)
Coronary death	11 556	2·31 (2·05–2·60)	41 (24–54)
Non-fatal myocardial infarction	14 741	1·82 (1·64–2·03)	37 (19–51)
Stroke subtypes*			
Ischaemic stroke	3799	2·27 (1·95–2·65)	1 (0–20)
Haemorrhagic stroke	1183	1·56 (1·19–2·05)	0 (0–26)
Unclassified stroke	4973	1·84 (1·59–2·13)	33 (12–48)
Other vascular deaths	3826	1·73 (1·51–1·98)	0 (0–26)

Figure 1: Hazard ratios (HRs) for vascular outcomes in people with versus those without diabetes at baseline
Analyses were based on 530 083 participants. HRs were adjusted for age, smoking status, body-mass index, and systolic blood pressure, and, where appropriate, stratified by sex and trial arm. 208 coronary heart disease outcomes that contributed to the grand total could not contribute to the subtotals of coronary death or non-fatal myocardial infarction because there were fewer than 11 cases of these coronary disease subtypes in some studies. *Includes both fatal and non-fatal events.



A Coronary heart disease

Adjusted for	HR (95% CI)
Age and sex	2·06 (1·82–2·34)
Plus smoking status	2·10 (1·85–2·39)
Plus BMI	2·00 (1·78–2·25)
Plus systolic blood pressure	1·91 (1·70–2·14)
Plus non-HDL cholesterol	1·93 (1·71–2·16)
Plus HDL cholesterol	1·87 (1·67–2·09)
Plus log-triglyceride	1·87 (1·67–2·09)

B Ischaemic stroke

Adjusted for	HR (95% CI)
Age and sex	2·56 (2·15–3·05)
Plus smoking status	2·59 (2·16–3·09)
Plus BMI	2·45 (2·08–2·88)
Plus systolic blood pressure	2·27 (1·94–2·65)
Plus non-HDL cholesterol	2·26 (1·94–2·64)
Plus HDL cholesterol	2·24 (1·94–2·60)
Plus log-triglyceride	2·24 (1·94–2·59)

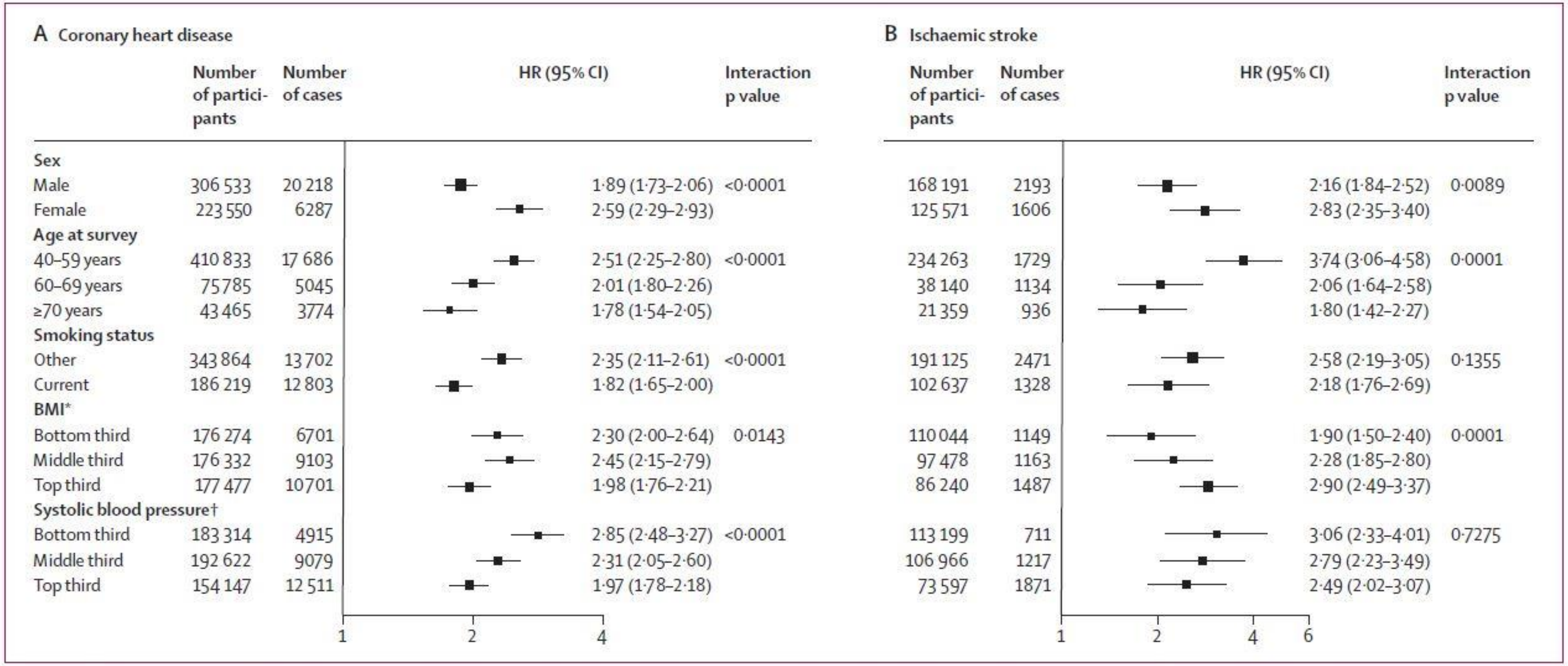
Figure 3: Hazard ratios (HRs) for coronary heart disease and ischaemic stroke in people with versus those without diabetes, progressively adjusted for baseline levels of conventional risk factors
Analyses were based on 264 353 participants (11 848 cases) for coronary heart disease and 157 315 participants (2858 cases) for ischaemic stroke with complete information on all covariates listed. BMI=body-mass index.

HRs (95% CI) for CHD in people with vs without diabetes, progressively adjusted for baseline levels of conventional risk factors

Cox model for cohorts studies, case-cohort (weighted cox) and clinical trials included: HRs for different outcomes on baseline diabetes status

Diabetes mellitus, glycaemia markers and CVD

HRs (95% CI) for CHD in people with vs without diabetes, by individual characteristics



A Coronary heart disease

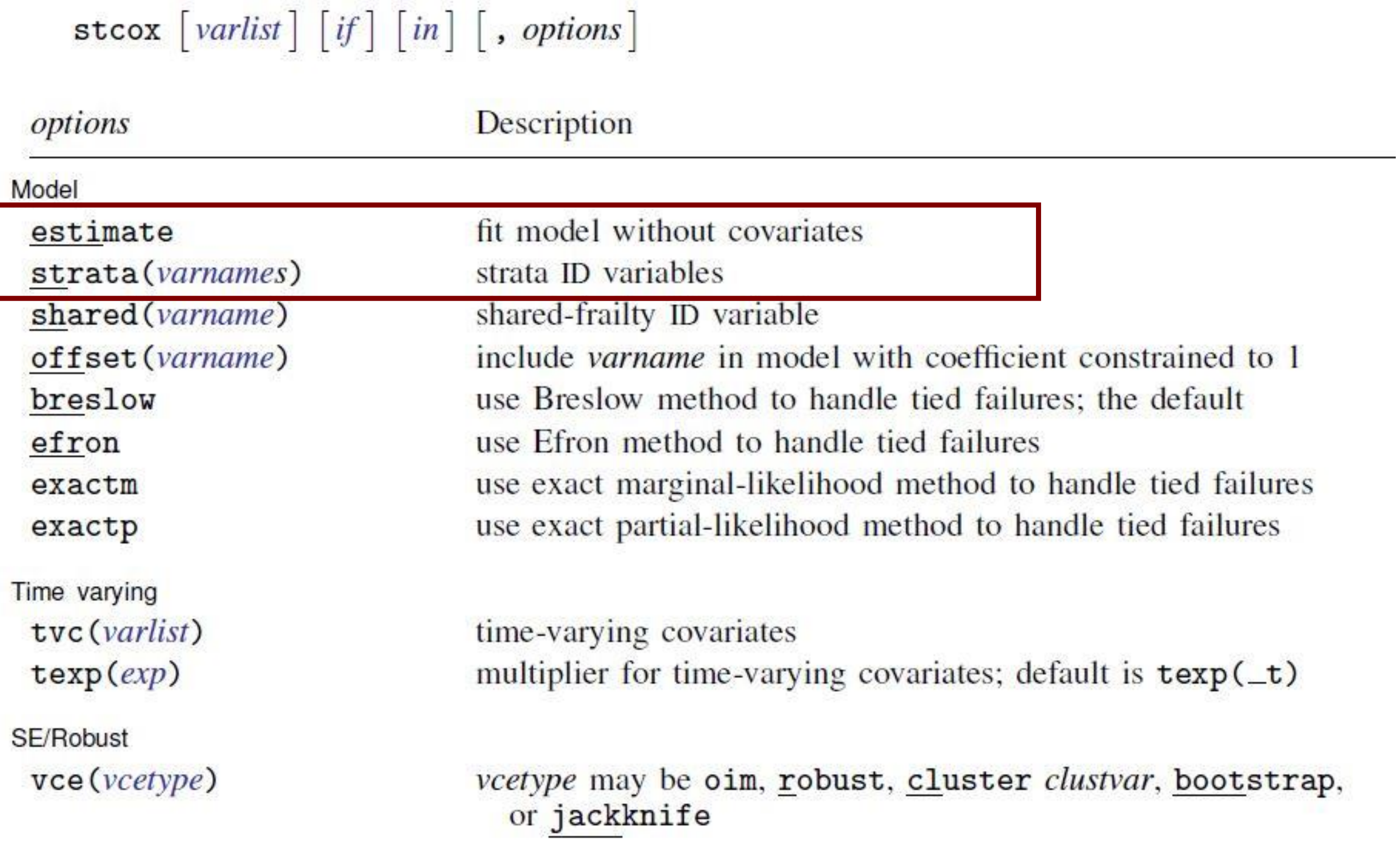
	Number of participants	Number of cases	HR (95% CI)	Interaction p value
Sex				
Male	306 533	20 218	1·89 (1·73–2·06)	<0·0001
Female	223 550	6287	2·59 (2·29–2·93)	
Age at survey				
40–59 years	410 833	17 686	2·51 (2·25–2·80)	<0·0001
60–69 years	75785	5045	2·01 (1·80–2·26)	
≥70 years	43 465	3774	1·78 (1·54–2·05)	
Smoking status				
Other	343 864	13 702	2·35 (2·11–2·61)	<0·0001
Current	186 219	12 803	1·82 (1·65–2·00)	
BMI*				
Bottom third	176 274	6701	2·30 (2·00–2·64)	0·0143
Middle third	176 332	9103	2·45 (2·15–2·79)	
Top third	177 477	10701	1·98 (1·76–2·21)	
Systolic blood pressure†				
Bottom third	183 314	4915	2·85 (2·48–3·27)	<0·0001
Middle third	192 622	9079	2·31 (2·05–2·60)	
Top third	154 147	12 511	1·97 (1·78–2·18)	

B Ischaemic stroke

	Number of participants	Number of cases	HR (95% CI)	Interaction p value
Sex				
Male	168 191	2193	2·16 (1·84–2·52)	0·0089
Female	125 571	1606	2·83 (2·35–3·40)	
Age at survey				
40–59 years	234 263	1729	3·74 (3·06–4·58)	0·0001
60–69 years	38 140	1134	2·06 (1·64–2·58)	
≥70 years	21 359	936	1·80 (1·42–2·27)	
Smoking status				
Other	191 125	2471	2·58 (2·19–3·05)	0·1355
Current	102 637	1328	2·18 (1·76–2·69)	
BMI*				
Bottom third	110 044	1149	1·90 (1·50–2·40)	0·0001
Middle third	97 478	1163	2·28 (1·85–2·80)	
Top third	86 240	1487	2·90 (2·49–3·37)	
Systolic blood pressure†				
Bottom third	113 199	711	3·06 (2·33–4·01)	0·7275
Middle third	106 966	1217	2·79 (2·23–3·49)	
Top third	73 597	1871	2·49 (2·02–3·07)	

Figure 2: Hazard ratios (HRs) for coronary heart disease and ischaemic stroke in people with versus those without diabetes at baseline, by individual characteristics
HRs were adjusted as described in figure 1. BMI=body-mass index. *Bottom third=<23·8 kg/m² (mean 21·7 kg/m²); middle third=23·8–<27 kg/m² (mean 25·3 kg/m²); and top third=≥27 kg/m² (mean 30·7 kg/m²). †Bottom third=<123 mm Hg (mean 113 mm Hg); middle third=123–<141 mm Hg (mean 132 mm Hg); and top third=≥141 mm Hg (mean 157 mm Hg).

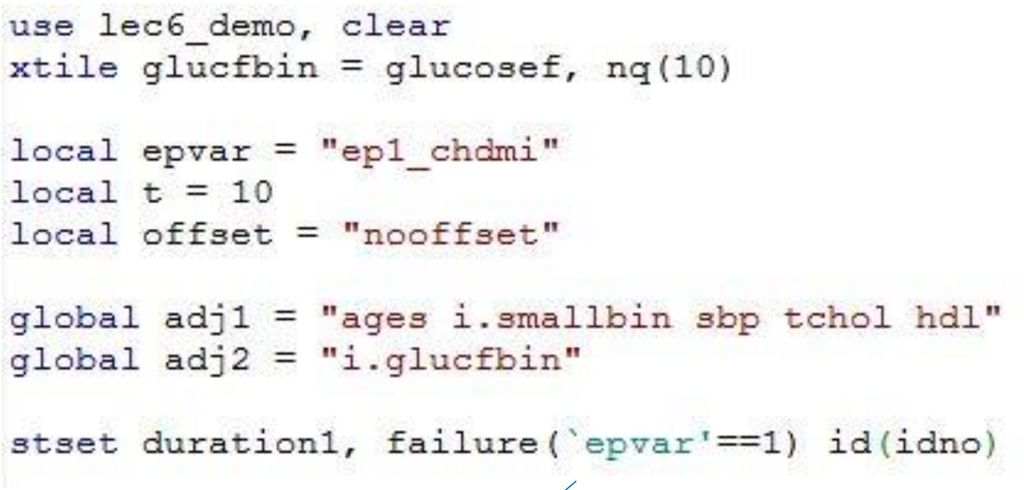
Step 3: stcox实现Cox回归模型



```
stcox [varlist] [if] [in] [, options]
```

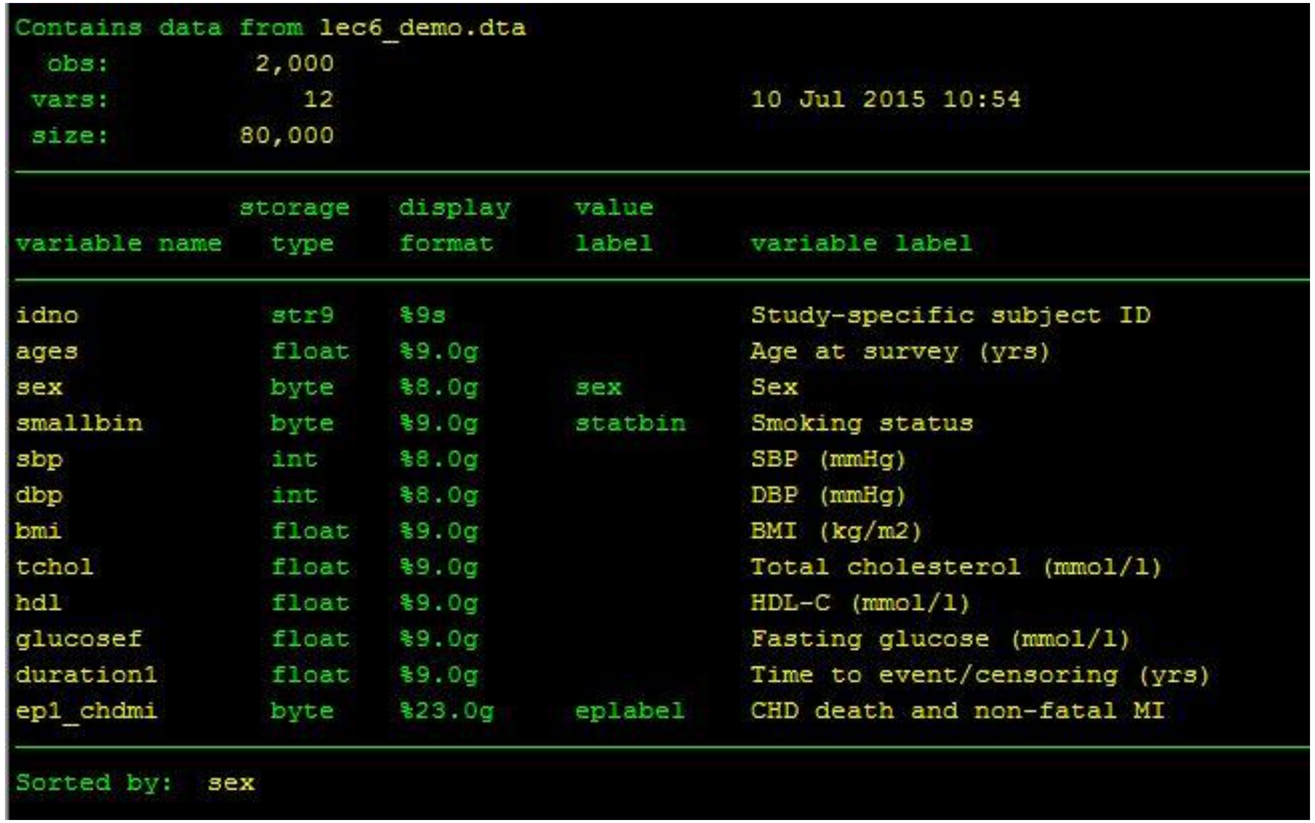
options	Description
Model	
estimate	fit model without covariates
strata(*varnames*)	strata ID variables
shared(*varname*)	shared-frailty ID variable
offset(*varname*)	include *varname* in model with coefficient constrained to 1
breslow	use Breslow method to handle tied failures; the default
efron	use Efron method to handle tied failures
exactm	use exact marginal-likelihood method to handle tied failures
exactp	use exact partial-likelihood method to handle tied failures
Time varying	
tvc(*varlist*)	time-varying covariates
texp(*exp*)	multiplier for time-varying covariates; default is texp(_t)
SE/Robust	
vce(*vcetype*)	*vcetype* may be oim, robust, cluster *clustvar*, bootstrap, or jackknife

Step 3: stcox实现Cox回归模型



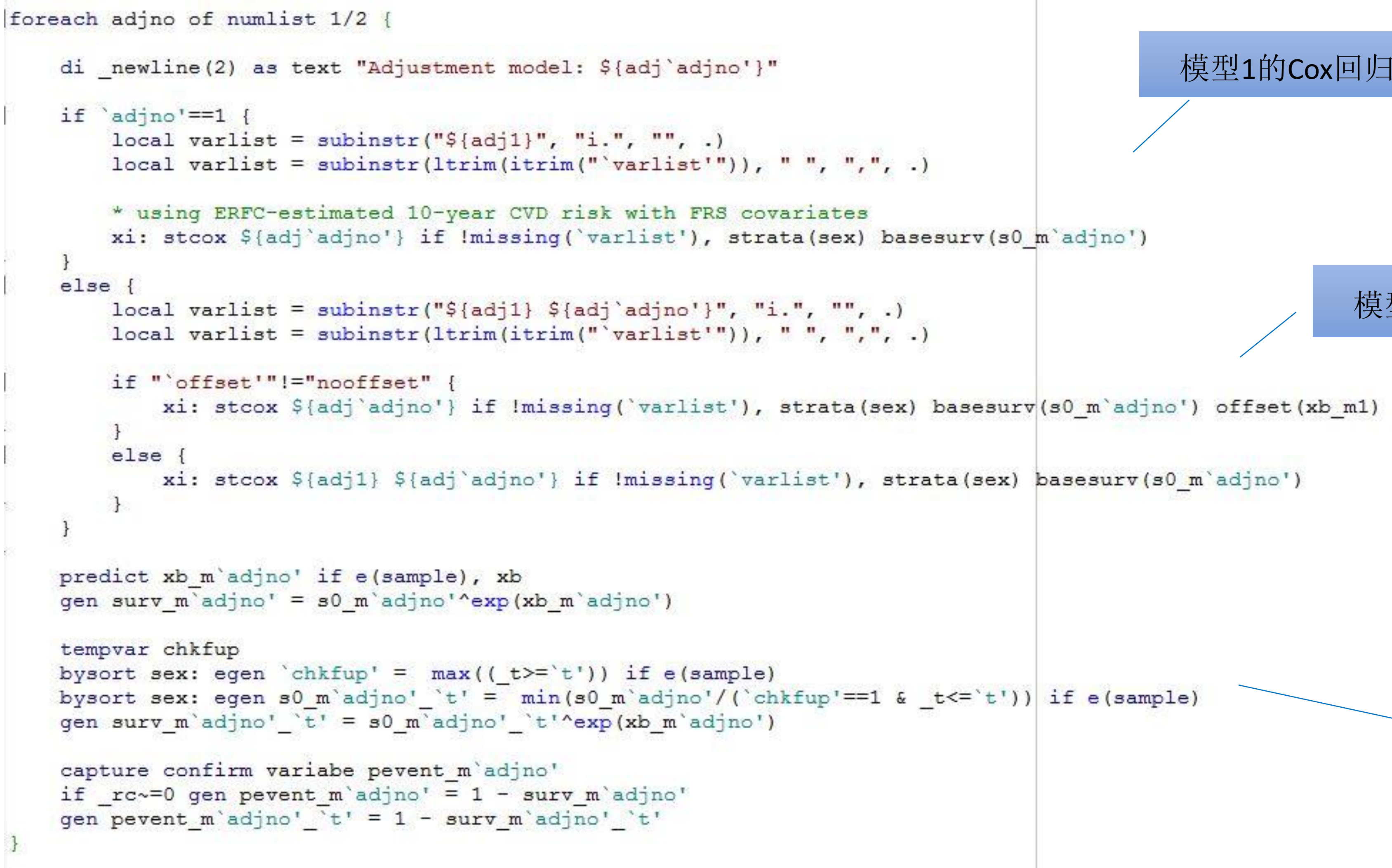
```
use lec6_demo, clear
xtile glucfbin = glucosef, nq(10)

local epvar = "ep1_chdmi"
local t = 10
local offset = "nooffset"

global adj1 = "ages i.smallbin sbp tchol hdl"
global adj2 = "i.glucfbin"

stset duration1, failure(`epvar'==1) id(idno)
```

利用Cox model计算HR以及个体10年CVD预测风险



```
Contains data from lec6_demo.dta
  obs:         2,000
 vars:            12                          10 Jul 2015 10:54
 size:        80,000
-------------------------------------------------------------------------------
              storage  display    value
variable name   type   format     label      variable label
-------------------------------------------------------------------------------
idno            str9   %9s                   Study-specific subject ID
ages            float  %9.0g                 Age at survey (yrs)
sex             byte   %8.0g      sex        Sex
smallbin        byte   %9.0g      statbin    Smoking status
sbp             int    %8.0g                 SBP (mmHg)
dbp             int    %8.0g                 DBP (mmHg)
bmi             float  %9.0g                 BMI (kg/m2)
tchol           float  %9.0g                 Total cholesterol (mmol/l)
hdl             float  %9.0g                 HDL-C (mmol/l)
glucosef        float  %9.0g                 Fasting glucose (mmol/l)
duration1       float  %9.0g                 Time to event/censoring (yrs)
ep1_chdmi       byte   %23.0g     eplabel    CHD death and non-fatal MI
-------------------------------------------------------------------------------
Sorted by:  sex
```

Step 3: stcox实现Cox回归模型



```
foreach adjno of numlist 1/2 {

    di _newline(2) as text "Adjustment model: ${adj`adjno'}"

    if `adjno'==1 {
        local varlist = subinstr("${adj1}", "i.", "", .)
        local varlist = subinstr(ltrim(itrim("`varlist'")), " ", ",", .)

        * using ERFC-estimated 10-year CVD risk with FRS covariates
        xi: stcox ${adj`adjno'} if !missing(`varlist'), strata(sex) basesurv(s0_m`adjno')
    }
    else {
        local varlist = subinstr("${adj1} ${adj`adjno'}", "i.", "", .)
        local varlist = subinstr(ltrim(itrim("`varlist'")), " ", ",", .)

        if "`offset'"!="nooffset" {
            xi: stcox ${adj`adjno'} if !missing(`varlist'), strata(sex) basesurv(s0_m`adjno') offset(xb_m1)
        }
        else {
            xi: stcox ${adj1} ${adj`adjno'} if !missing(`varlist'), strata(sex) basesurv(s0_m`adjno')
        }
    }

    predict xb_m`adjno' if e(sample), xb
    gen surv_m`adjno' = s0_m`adjno'^exp(xb_m`adjno')

    tempvar chkfup
    bysort sex: egen `chkfup' =  max((_t>=`t')) if e(sample)
    bysort sex: egen s0_m`adjno'_`t' =  min(s0_m`adjno'/(`chkfup'==1 & _t<=`t')) if e(sample)
    gen surv_m`adjno'_`t' = s0_m`adjno'_`t'^exp(xb_m`adjno')

    capture confirm variabe pevent_m`adjno'
    if _rc~=0 gen pevent_m`adjno' = 1 - surv_m`adjno'
    gen pevent_m`adjno'_`t' = 1 - surv_m`adjno'_`t'
}
```

模型1的Cox回归模型

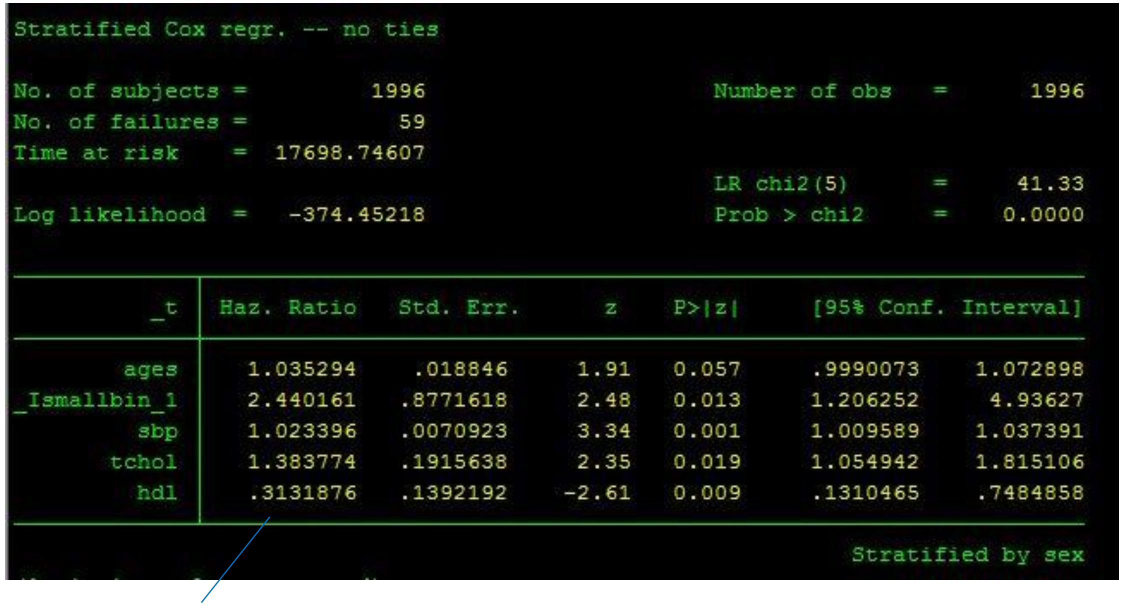
模型2的Cox回归模型

生成预测值

Stata的生存分析

Diabetes mellitus, glycaemia markers and CVD

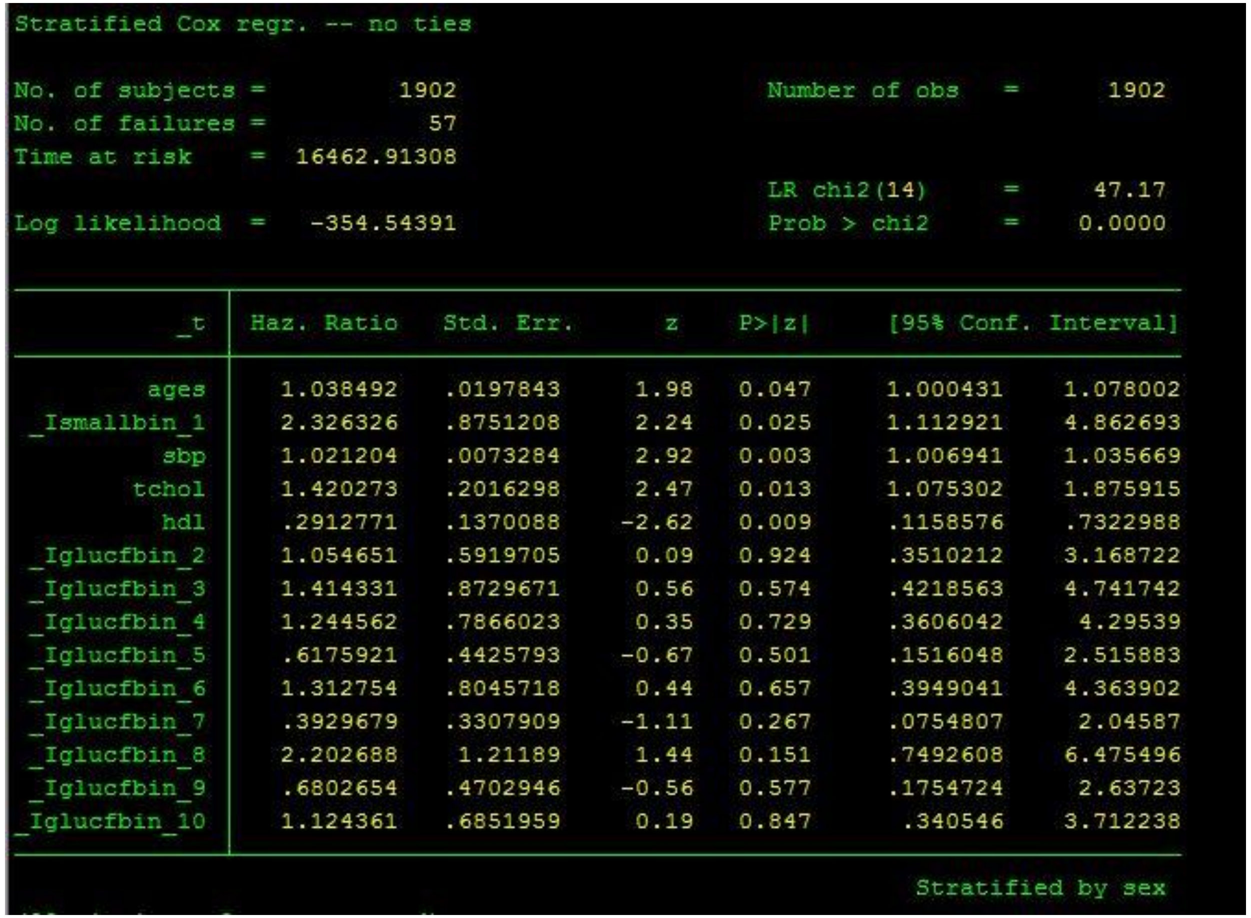
```
xi: stcox ${adj`adjno'} if
!missing(`varlist'), strata(sex)
basesurv(s0_m`adjno')
```



```
Stratified Cox regr. -- no ties

No. of subjects =         1996                  Number of obs   =      1996
No. of failures =           59
Time at risk    =  17698.74607
                                                LR chi2(5)      =     41.33
Log likelihood  =   -374.45218                  Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        ages |   1.035294    .018846     1.91   0.057     .9990073    1.072898
_Ismallbin_1 |   2.440161   .8771618     2.48   0.013     1.206252     4.93627
         sbp |   1.023396   .0070923     3.34   0.001     1.009589    1.037391
       tchol |   1.383774   .1915638     2.35   0.019     1.054942    1.815106
         hdl |   .3131876   .1392192    -2.61   0.009     .1310465    .7484858
------------------------------------------------------------------------------
                                                             Stratified by sex
```

模型1的Cox model的结果

```
xi: stcox ${adj1} ${adj`adjno'} if !missing(`varlist'),
strata(sex) basesurv(s0_m`adjno')
```



```
Stratified Cox regr. -- no ties

No. of subjects =         1902                  Number of obs   =      1902
No. of failures =           57
Time at risk    =  16462.91308
                                                LR chi2(14)     =     47.17
Log likelihood  =   -354.54391                  Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        ages |   1.038492   .0197843     1.98   0.047     1.000431    1.078002
_Ismallbin_1 |   2.326326   .8751208     2.24   0.025     1.112921    4.862693
         sbp |   1.021204   .0073284     2.92   0.003     1.006941    1.035669
       tchol |   1.420273   .2016298     2.47   0.013     1.075302    1.875915
         hdl |   .2912771   .1370088    -2.62   0.009     .1158576    .7322988
_Iglucfbin_2 |   1.054651   .5919705     0.09   0.924     .3510212    3.168722
_Iglucfbin_3 |   1.414331   .8729671     0.56   0.574     .4218563    4.741742
_Iglucfbin_4 |   1.244562   .7866023     0.35   0.729     .3606042     4.29539
_Iglucfbin_5 |   .6175921   .4425793    -0.67   0.501     .1516048    2.515883
_Iglucfbin_6 |   1.312754   .8045718     0.44   0.657     .3949041    4.363902
_Iglucfbin_7 |   .3929679   .3307909    -1.11   0.267     .0754807     2.04587
_Iglucfbin_8 |   2.202688    1.21189     1.44   0.151     .7492608    6.475496
_Iglucfbin_9 |   .6802654   .4702946    -0.56   0.577     .1754724     2.63723
_Iglucfbin_10 |   1.124361   .6851959     0.19   0.847      .340546    3.712238
------------------------------------------------------------------------------
                                                             Stratified by sex
```

Step 4: predict – make predictions

predict是**stcox**命令的后续命令，和regress、logistic一样，当**stcox**命令完成Cox回归之后，**predict**命令可以用来得到HR，Baseline hazard，拟合值和残差。

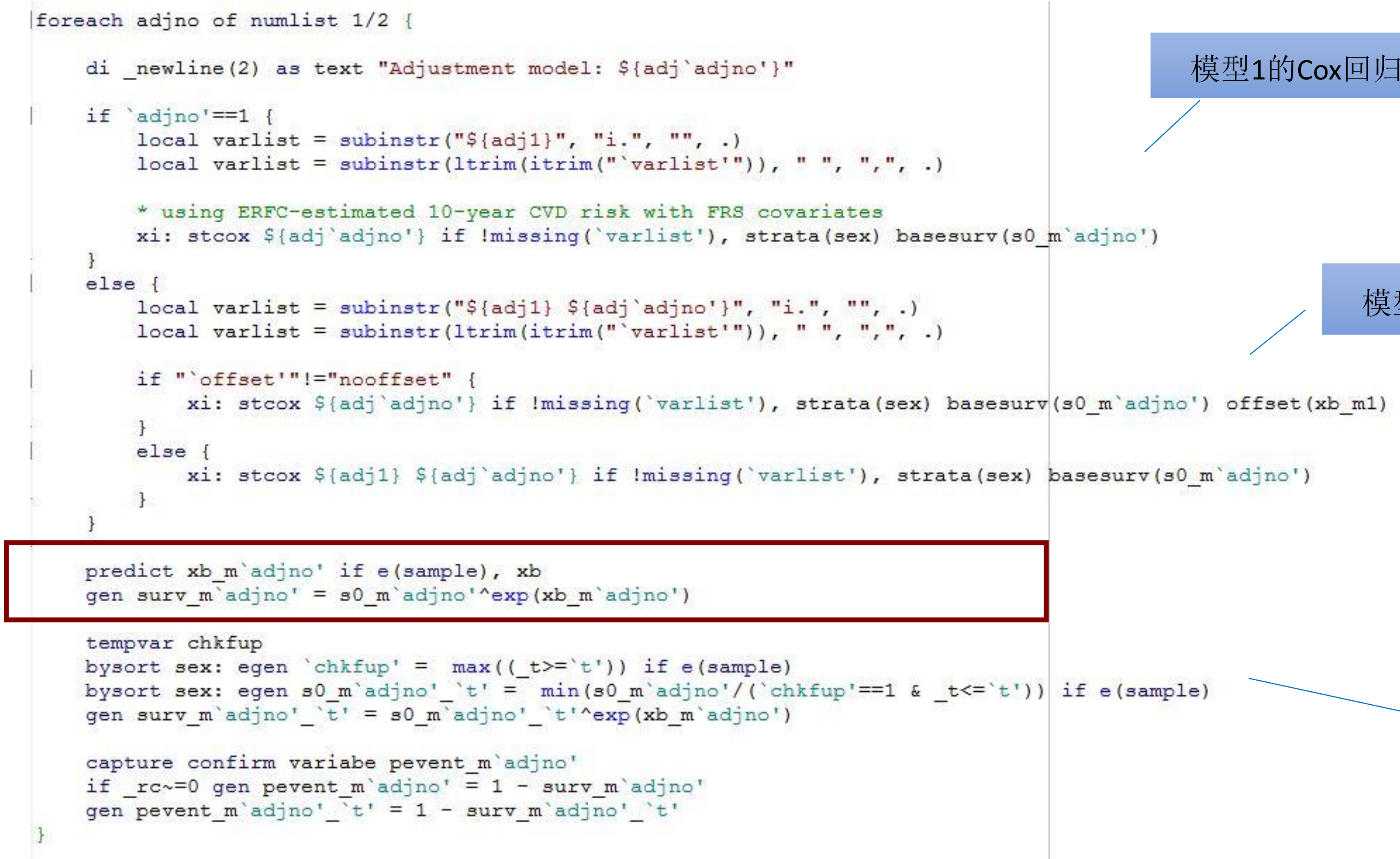
```
stcox first
predict [type] newvar [if] [in] [, option]
```

新生成变量的变量名

指定估计值的选项：

- hr或空缺：predicted hazard ratio
- xb: linear predictor
- stdp: SE of the liniear predictior xb
- basesurv: baseline survivor function
- basechazard: baseline cumulative hazard function

Step 4: predict – make predictions



```
foreach adjno of numlist 1/2 {

    di _newline(2) as text "Adjustment model: ${adj`adjno'}"

    if `adjno'==1 {
        local varlist = subinstr("${adj1}", "i.", "", .)
        local varlist = subinstr(ltrim(itrim("`varlist'")), " ", ",", .)

        * using ERFC-estimated 10-year CVD risk with FRS covariates
        xi: stcox ${adj`adjno'} if !missing(`varlist'), strata(sex) basesurv(s0_m`adjno')
    }
    else {
        local varlist = subinstr("${adj1} ${adj`adjno'}", "i.", "", .)
        local varlist = subinstr(ltrim(itrim("`varlist'")), " ", ",", .)

        if "`offset'"!="nooffset" {
            xi: stcox ${adj`adjno'} if !missing(`varlist'), strata(sex) basesurv(s0_m`adjno') offset(xb_m1)
        }
        else {
            xi: stcox ${adj1} ${adj`adjno'} if !missing(`varlist'), strata(sex) basesurv(s0_m`adjno')
        }
    }

    predict xb_m`adjno' if e(sample), xb
    gen surv_m`adjno' = s0_m`adjno'^exp(xb_m`adjno')

    tempvar chkfup
    bysort sex: egen `chkfup' =  max((_t>=`t')) if e(sample)
    bysort sex: egen s0_m`adjno'_`t' =  min(s0_m`adjno'/(`chkfup'==1 & _t<=`t')) if e(sample)
    gen surv_m`adjno'_`t' = s0_m`adjno'_`t'^exp(xb_m`adjno')

    capture confirm variabe pevent_m`adjno'
    if _rc~=0 gen pevent_m`adjno' = 1 - surv_m`adjno'
    gen pevent_m`adjno'_`t' = 1 - surv_m`adjno'_`t'
}
```

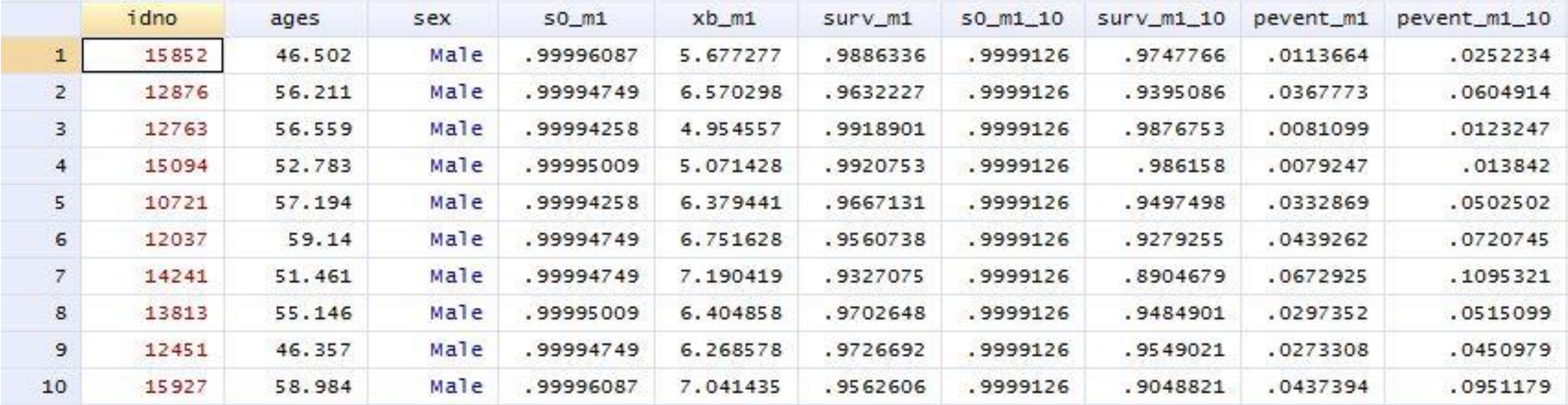
模型1的Cox回归模型

模型2的Cox回归模型

生成预测值

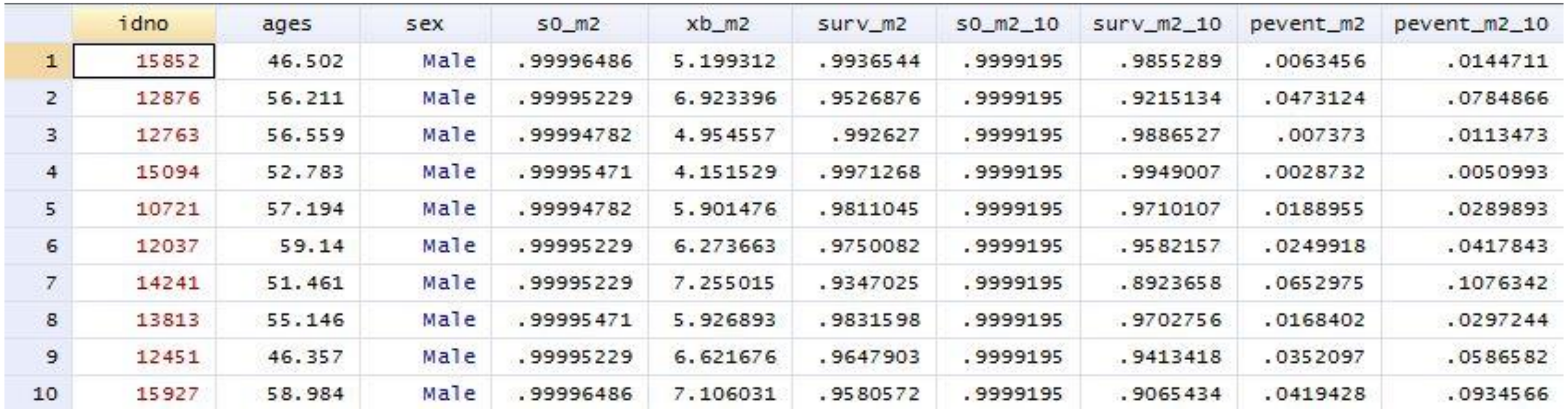
$$S(t) = {S_0}^{\exp[\sum \beta x]}$$

Step 4: predict – make predictions



	idno	ages	sex	s0_m1	xb_m1	surv_m1	s0_m1_10	surv_m1_10	pevent_m1	pevent_m1_10
1	15852	46.502	Male	.99996087	5.677277	.9886336	.9999126	.9747766	.0113664	.0252234
2	12876	56.211	Male	.99994749	6.570298	.9632227	.9999126	.9395086	.0367773	.0604914
3	12763	56.559	Male	.99994258	4.954557	.9918901	.9999126	.9876753	.0081099	.0123247
4	15094	52.783	Male	.99995009	5.071428	.9920753	.9999126	.986158	.0079247	.013842
5	10721	57.194	Male	.99994258	6.379441	.9667131	.9999126	.9497498	.0332869	.0502502
6	12037	59.14	Male	.99994749	6.751628	.9560738	.9999126	.9279255	.0439262	.0720745
7	14241	51.461	Male	.99994749	7.190419	.9327075	.9999126	.8904679	.0672925	.1095321
8	13813	55.146	Male	.99995009	6.404858	.9702648	.9999126	.9484901	.0297352	.0515099
9	12451	46.357	Male	.99994749	6.268578	.9726692	.9999126	.9549021	.0273308	.0450979
10	15927	58.984	Male	.99996087	7.041435	.9562606	.9999126	.9048821	.0437394	.0951179

模型1的预测值



	idno	ages	sex	s0_m2	xb_m2	surv_m2	s0_m2_10	surv_m2_10	pevent_m2	pevent_m2_10
1	15852	46.502	Male	.99996486	5.199312	.9936544	.9999195	.9855289	.0063456	.0144711
2	12876	56.211	Male	.99995229	6.923396	.9526876	.9999195	.9215134	.0473124	.0784866
3	12763	56.559	Male	.99994782	4.954557	.992627	.9999195	.9886527	.007373	.0113473
4	15094	52.783	Male	.99995471	4.151529	.9971268	.9999195	.9949007	.0028732	.0050993
5	10721	57.194	Male	.99994782	5.901476	.9811045	.9999195	.9710107	.0188955	.0289893
6	12037	59.14	Male	.99995229	6.273663	.9750082	.9999195	.9582157	.0249918	.0417843
7	14241	51.461	Male	.99995229	7.255015	.9347025	.9999195	.8923658	.0652975	.1076342
8	13813	55.146	Male	.99995471	5.926893	.9831598	.9999195	.9702756	.0168402	.0297244
9	12451	46.357	Male	.99995229	6.621676	.9647903	.9999195	.9413418	.0352097	.0586582
10	15927	58.984	Male	.99996486	7.106031	.9580572	.9999195	.9065434	.0419428	.0934566

模型2的预测值

Summary : Survival analysis

a) Introduction to Survival analysis

b) Cohort studies

c) Setting up for the survival analysis in Stata: **stset**

d) Describe the survival curve and relative functions in Stata: **sts**

e) Cox model

f) Cox model in Stata： **stcox** (& **predict**)

北京大学
PEKING UNIVERSITY

CREATION：Call from REsearch to AcTION 知行合一

Thank You!

高 培

北京大学公共卫生学院流行病与卫生统计学系

北京大学临床研究所真实世界证据评价中心

（**C**entre for **R**eal-world **E**vidence evalu**ATION**，**CREATION中心**）