

流行病学

epidemiology

队列研究

Cohort Study

 金岳龙 教授

 公共卫生学院

TEL:13965178350

引言

病因和危险因素研究是流行病学研究的重要任务，病因研究的逻辑顺序应该是先有病因存在，然后有疾病发生，即先因后果。

那么，怎样的研究满足先因后果的逻辑顺序，能更有效地证实因果关系呢？

队列研究

✚ **队列研究 (cohort study)** 是分析流行病学研究中两大重要方法之一。

✚ 它通过直接观察暴露于某因素不同状况人群的结局来探讨该因素与所观察结局的关系。

目录

ontents



概 述



队列研究的设
计与实施



队列研究资料
的分析



常见偏倚及其
控制



队列研究的优
点与局限性

第一节 概述

- ✚ 概念及基本原理

- ✚ 特 点

- ✚ 研究类型

- ✚ 用 途

概念

➤ 暴露(exposure)

指研究对象接触过某种待研究的物质或具有某种研究的特征或行为。

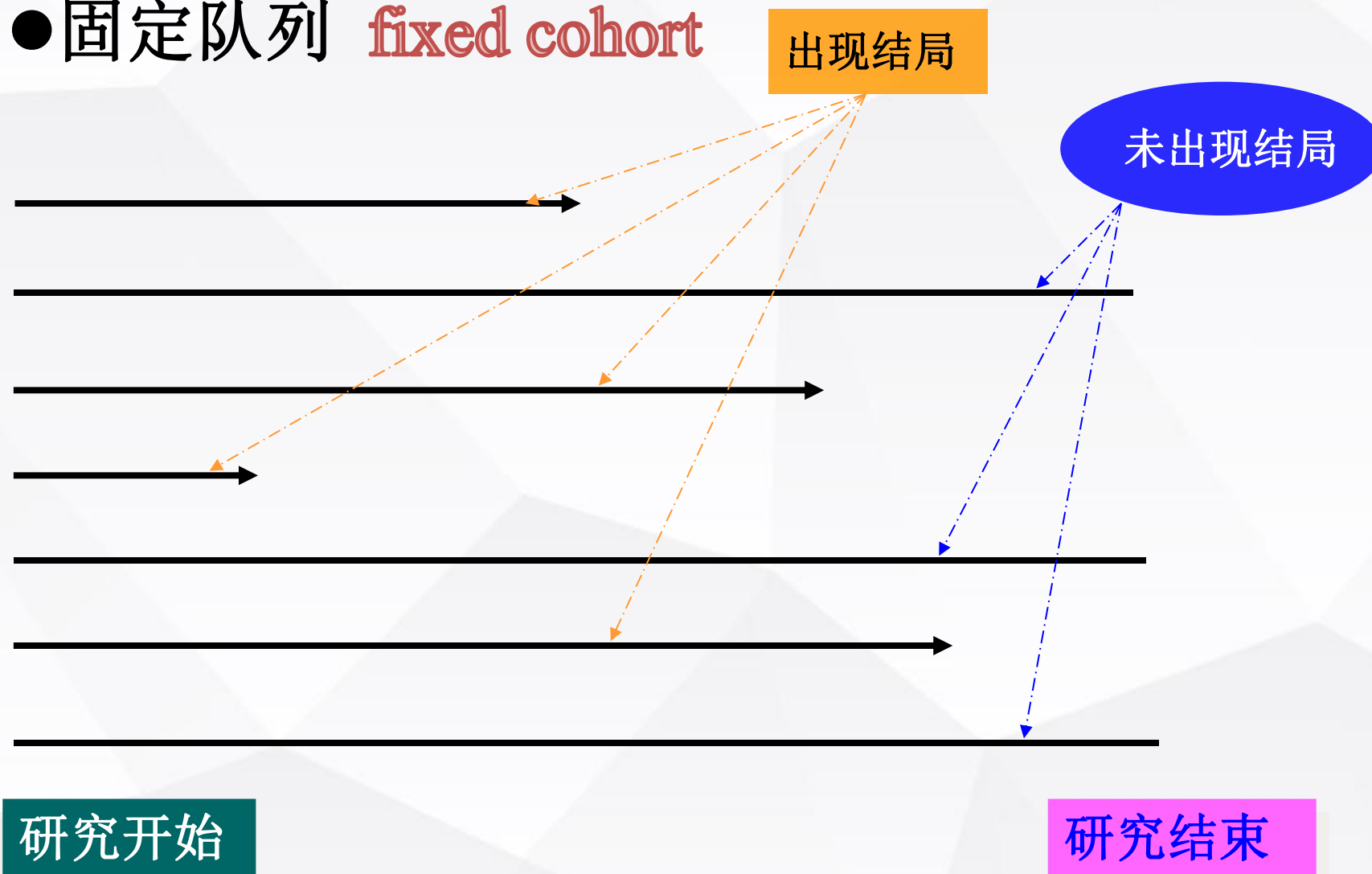
是本研究需要探讨的因素。

➤ 队列(cohort)

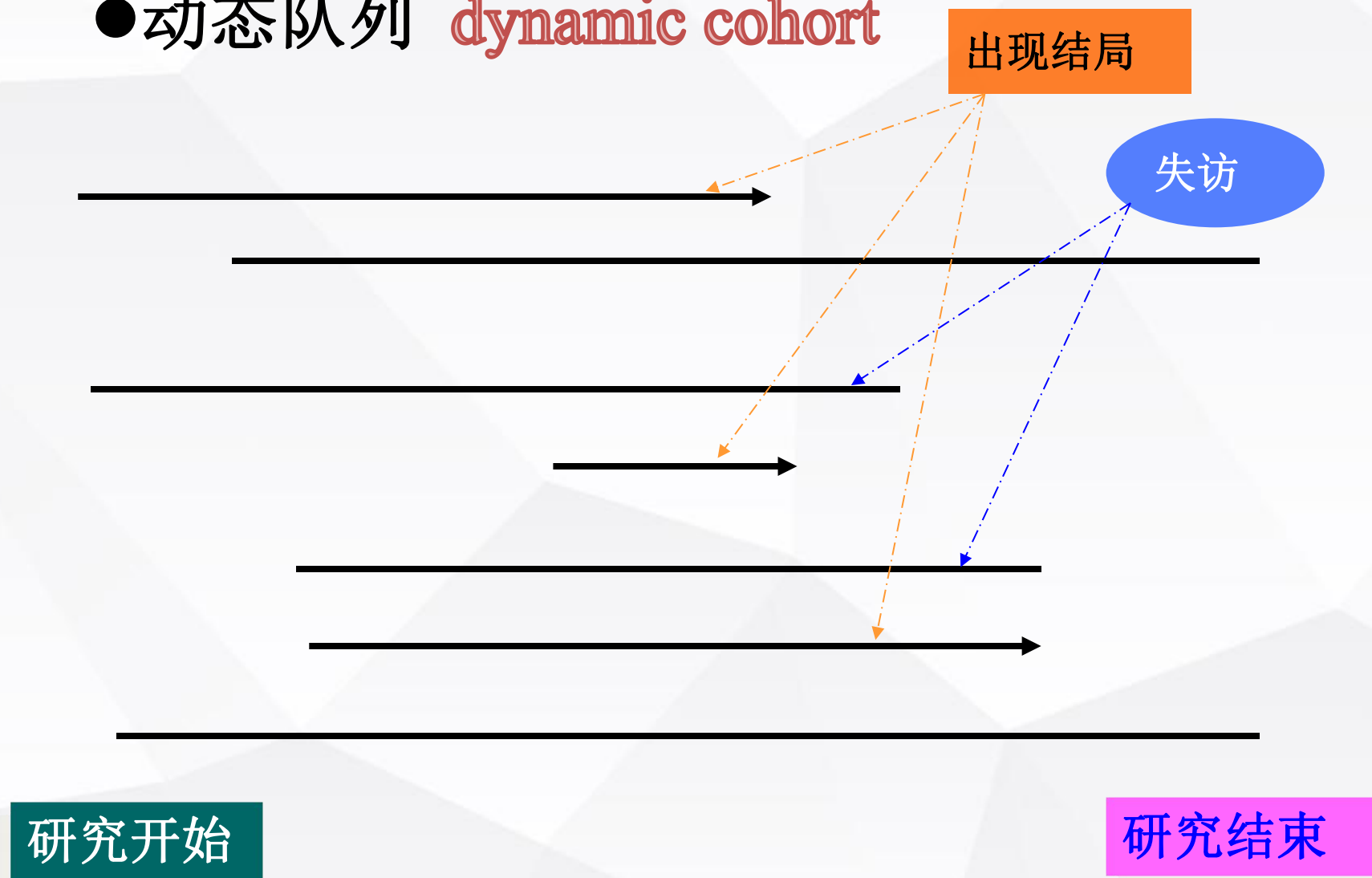
有共同经历或有共同暴露特征的一组人群（如出生队列和暴露队列）。

根据其变化分为固定队列和动态队列。

● 固定队列 *fixed cohort*



● 动态队列 dynamic cohort

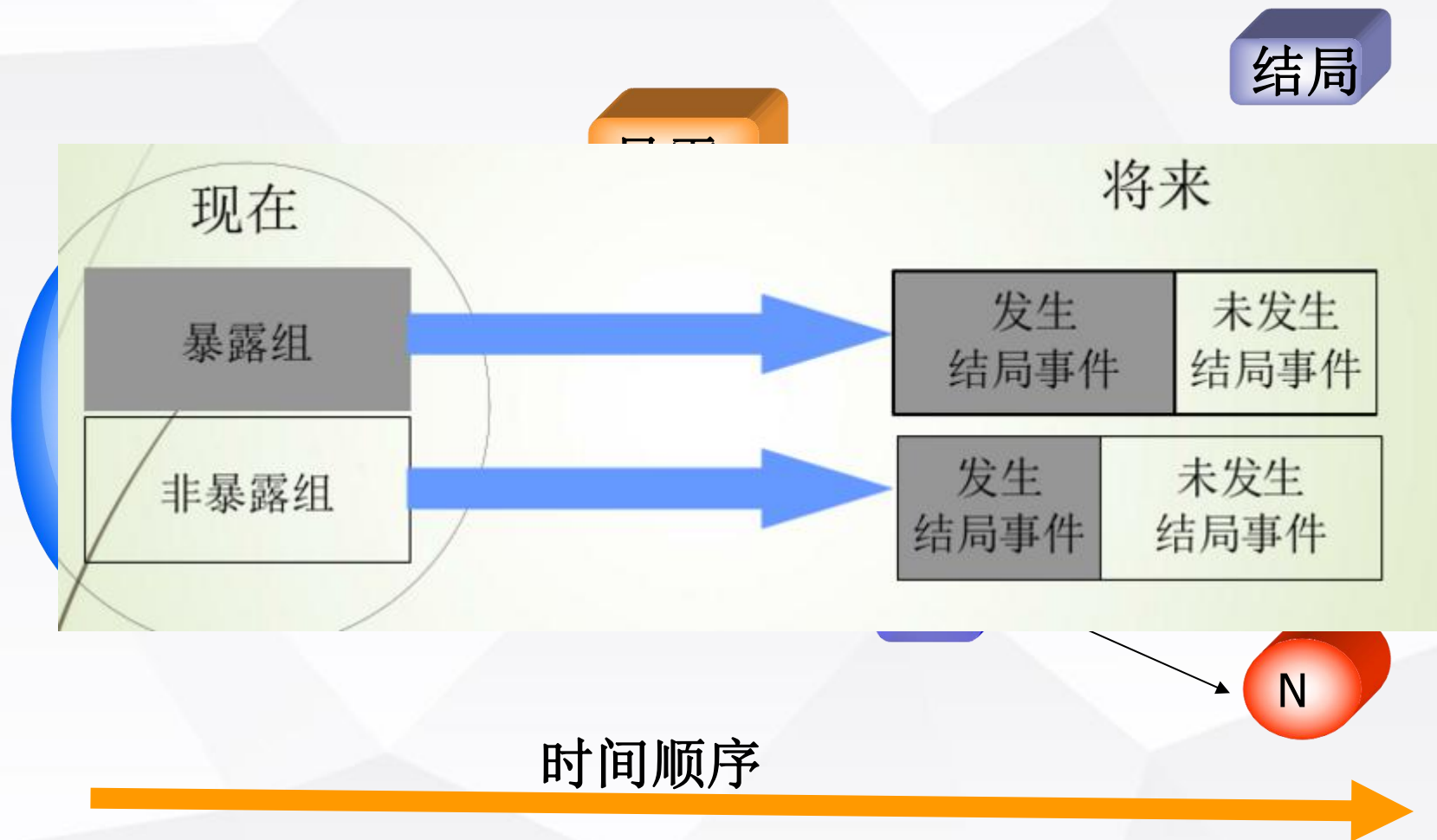


队列研究——概念

- 将特定人群按照是否暴露于某因素分为暴露组与非暴露组，或不同暴露水平的亚组；
- 随访观察一定时期各组人群某种结局事件（发病/死亡）的发生情况，通过比较各组人群结局的发生率；
- 以检验该因素与结局时间之间有无因果关联及关联强度的一种观察性研究方法。

基本原理

- 根据研究对象是否暴露于某研究因素或其不同水平将研究对象分成暴露组（**E**）与非暴露组（ **$\bar{E}$** ）；
- 随访一定时间，比较两组之间所研究结局(outcome)发生率的差异，以分析暴露因素与研究结局之间的关系。



➤ 主要特点

- 观察法
- 设立对照
- 由因到果，符合时间顺序（前瞻性；一因多果）
- 可验证暴露和结局之间的因果关系（可计算发生率或死亡率）

研究类型

- 前瞻性队列研究(prospective cohort study)
- 历史性队列研究(historical cohort study)
- 双向性队列研究(ambispective cohort study)

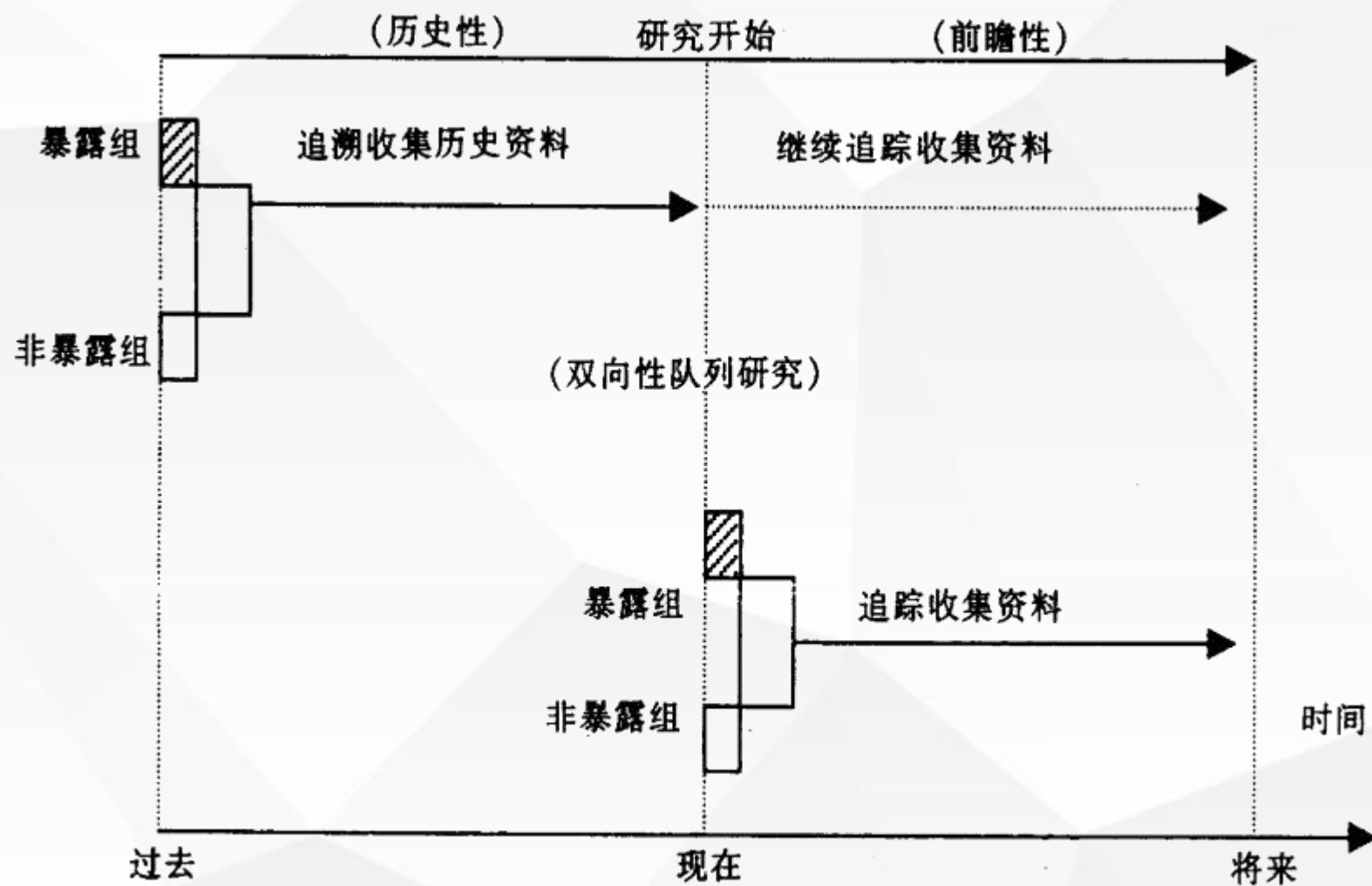


图 6-2 三类队列研究示意图

用途

- **检验病因假设** **最主要的目的和用途**
- **评价预防效果** **戒烟者与继续吸烟者进行比较分析，可评价戒烟对肺癌的预防效果，为吸烟与肺癌病因学关系的判断提供可靠依据**
- **研究疾病的自然史** **队列研究从暴露开始观察，可以了解到疾病或健康等结局事件从发生、发展到结局的全过程**
- **预测疾病发展趋势** **通过队列研究长期随访观察积累的资料，可以分析某种疾病发生、发展的变化特点与规律，预测其未来发展动态与趋势**

第二节 设计与实施

- ✚ 暴露因素
- ✚ 研究结局
- ✚ 研究对象
- ✚ 样本含量
- ✚ 资料的收集与随访

一、确定研究因素

主要暴露因素

在描述性研究和病例对照研究的基础上确定

可能影响结局的因素

混杂因素、人口学特征等

暴露测量

性质：定性和定量

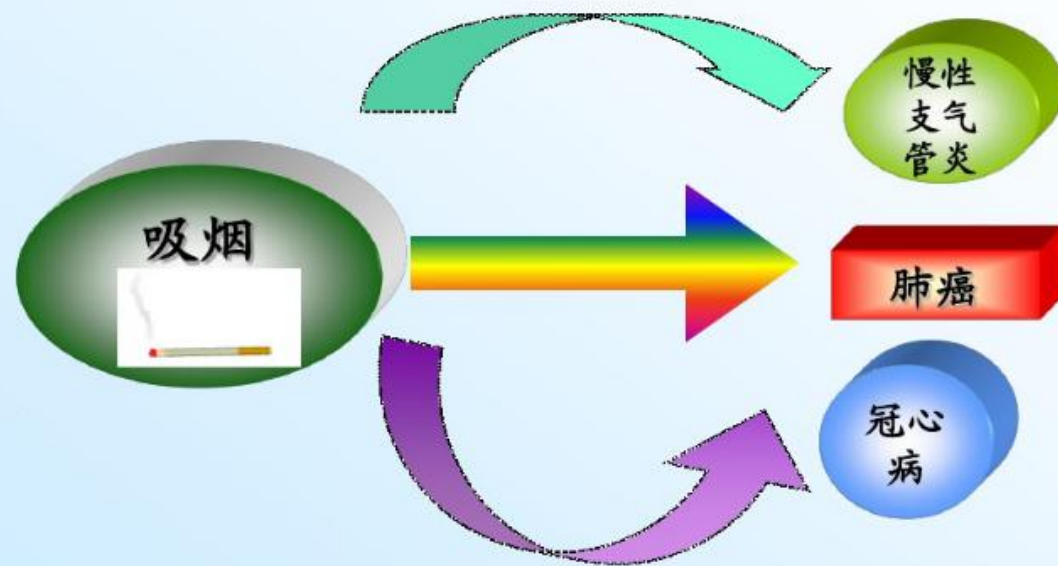
方法：访谈、实验室检查及查阅记录等

二、确定研究结局

结局是研究队列中预期结果事件，是队列研究的观察终点。研究结局的确定应全面、具体、客观。可包括：

- 1.发病、死亡、健康状况和生命质量
- 2.终极结果（发病）或中间结局（血清变化）
- 3.正面（生命延长）和负面（发病）
- 4.定性或定量
- 5.可多个或多方面

◆ 一次研究可有多结局



结局的测量 采用国际或国内通用的标准

三、确定研究现场与研究人群

➤研究现场

- 有足够符合条件的研究对象
- 领导重视、群众支持
- 医疗条件较好，交通较便利
- 发病率较高
- 有代表性

如研究现场是安徽省芜湖市

➤研究人群

- 从目标人群中抽出的具有代表性的人
- 未患所研究疾病
- 分为暴露组和对照组

➤ 暴露人群选择

- 职业人群
- 特殊暴露人群
- 一般人群
- 有组织的人群团体

➤ 对照人群的选择

- **内对照**(internal control)： 一群研究对象内部
- **外对照**(external control)： 一群研究对象外部
- **一般人群对照**(total population control)： 整个地区现成的发病或死亡统计资料
- **多重对照**(multiple controls)： 两种或以上的对照形式

四、确定样本量

➤ 计算样本量时需考虑的问题

1. 暴露组与对照组结局事件的发生率之差 (d)

2. 暴露因素与结局事件的关联强度

3. 显著性水平 (α)

4. 把握度 ($1-\beta$)

$$n = \frac{\left(z_{\alpha} \sqrt{2pq} + z_{\beta} \sqrt{p_0q_0 + p_1q_1} \right)^2}{(p_1 - p_0)^2}$$

为了考察母亲孕期使用雌激素与所生子女患先天性心脏病的联系，拟进行队列研究。已知非暴露孕妇所生婴儿的先天性心脏病发病率 (p_0) 为 2‰，使用雌激素引起子女患该病的 RR 为 2。设 $\alpha=0.05$ (双侧)， $\beta=0.10$ ，求样本大小。

已知： $p_0 = 0.002$ ， $RR=2$ ，则

$$p_1 = RR \times p_0 = 2 \times 0.002 = 0.004, \quad q_1 = 0.996 \quad q_0 = 0.998$$

$$\bar{p} = (0.004 + 0.002) / 2 = 0.003, \quad \bar{q} = 1 - \bar{p} = 0.997$$

查表得： $Z_\alpha = 1.96$ ， $Z_\beta = 1.28$

代入公式 7-1 得：

$$n = \frac{(1.96\sqrt{2 \times 0.003 \times 0.997} + 1.28\sqrt{0.002 \times 0.998 + 0.004 \times 0.996})^2}{(0.004 - 0.002)^2} = 15700$$

即暴露组与非暴露组各需 15700 人年。

资料的收集与随访

- 基线资料
- 随 访
- 资料收集方法

➤ 基线(baseline)资料

- 暴露的资料
- 个体的其他信息

资料收集方式

- 查阅纪录
- 调查询问
- 健康或疾病检查
- 环境监测等

➤ 随访(follow up)

- 随访对象、内容和方法
- 随访问隔
- 随访者
- 观察终点（**研究对象出现了预期的结果**）
- 观察终止时间（**整个研究工作截止的时间**）
- 失访

失访及其处理:

- 1.由于队列研究观察对象多、时间长，失访难免。
- 2.对于失访者应尽可能进行补访，未能追访到者，应尽可能取得失访者结局的信息，对失访原因做进一步的分析，与在访人群的结局进行比较，以推测失访对结果的可能影响。
- 3.如果暴露组与未暴露组的失访率相似，失访者与未失访者的结局发生率也相似，则失访对研究结果的影响作用较小。

比较现实可行的方法:

把失访者与未失访者基线资料中的一些特征加以比较，如差别不大，则可假定结局发生率的差别可能也不大，否则，对选择偏倚可能产生的影响应有充分估计。

➤ 质量控制

同任何流行病学调查研究一样，队列研究中资料收集的质量控制尤为重要，关乎研究结果的真实性，因此，在资料收集的各个环节均应予以足够的重视。

- 调查员的选择
- 调查员培训
- 制定调查员手册
- 监督与管理

第三节 资料分析

- ✚ 资料的整理与描述

- ✚ 率的计算

- ✚ 效应估计

- ✚ 分级分析

- ✚ 分层分析

- ✚ 多因素分析

✚ 基本整理模式

表4-1 队列研究资料归纳整理表（累积发病率资料）

	病例	非病例	合计
暴露组	a (1079)	b (13259)	$a+b=n_1$
非暴露组	c (387)	d (75511)	$c+d=n_0$
合计	$a+c=m_1$	$b+d=m_0$	$a+b+c+d=t$

非暴露组发病率= $c/n_0=75.25\%$

暴露组发病率= $a/n_1=5.10\%$

率的计算

● 累积发病率 (cumulative incidence)

变化范围 0~1

$$CI = \frac{\text{观察期内发病（或死亡）人数}}{\text{观察开始时的人口数}}$$

适用条件 样本大

人口稳定

整齐的资料

报告时必须注明时间长短

● 适用于固定队列

率的计算

● 发病密度 (incidence density)

变化范围 $0 \sim \infty$

$$ID = \frac{\text{观察期内发病（或死亡）人数}}{\text{观察人时}}$$

适用条件

- 观察时间长
- 人口不稳定
- 存在失访
- 资料不很整齐

● 适用于动态队列

人时的计算

- 精确法： 逐人逐天计算
- 近似法： 按年平均人口数计算
- 寿命表法： 当年内进入或退出队列的个人均作 $1/2$ 人年计算



Jan, July, Jan, July Jan, July, Jan, July person
1998 1988 1999 1999 2000 2000 2001 2001 time

A _____ 2
B _____ 2.5
C _____ 1.5
D _____ 4

Total person-time 10

观察时间 (年数)	年初人数	年内进人数	年内发病数	年内失访人数	暴露人年数
x (1)	I _x (2)	N _x (3)	D _x (4)	W _x (5)	L _x (6)
1	1403	79	4	30	1425.5
2	1448	45	2	11	1464.0
3	1480	60	3	8	1504.5
4	1529	5	2	19	1521.0
5	1513	10	7	25	1502.0
6	1491	18	8	29	1481.5
7	1472	13	3	73	1440.5
8	1409	12	5	74	1375.5
9	1342	9	2	467	1112.0
10	882	3	1	819	473.5
11	65	0	0	57	36.5
			37		13336.5 人年

年龄组	对象1 1927年3月21日出生	对象2 1935年4月9日出生	对象3 1942年4月12日出生	暴露人年
25~		61.11.11~65.04.08 共3 年4个月27天, 合3.41 人年	70.02.01~72.11.11 共2 年9个月10天, 合2.78 人年	6.19
30~		65.04.09~70.04.08 共 5.00人年	72.11.12~77.11.11 共 5.00人年	10.00
35~	66.07.19~67.03.20 共8 个月, 合0.67人年	70.04.09~73.12.01 共3 年7个月22天, 合3.65 人年	77.11.12~81.01.01 共3 年1个月20天, 合3.14 人年	7.46
40~	67.03.21~72.03.20 共 5.00人年			5.00
45~	72.03.21~77.03.20 共 5.00人年			5.00
50~54	77.03.21~77.09.14 共5 个月24天, 合0.48人年			0.48
累计	66.07.19~77.09.14 共 11.15人年	61.11.11~73.12.01 共 12.06人年	70.02.01~81.01.01 共 10.92人年	34.13 人年

$$L_x = I_x + 1/2(N_x - D_x - W_x)$$

$$I_{x+1} = I_x + N_x - D_x - W_x$$

率的差异性检验



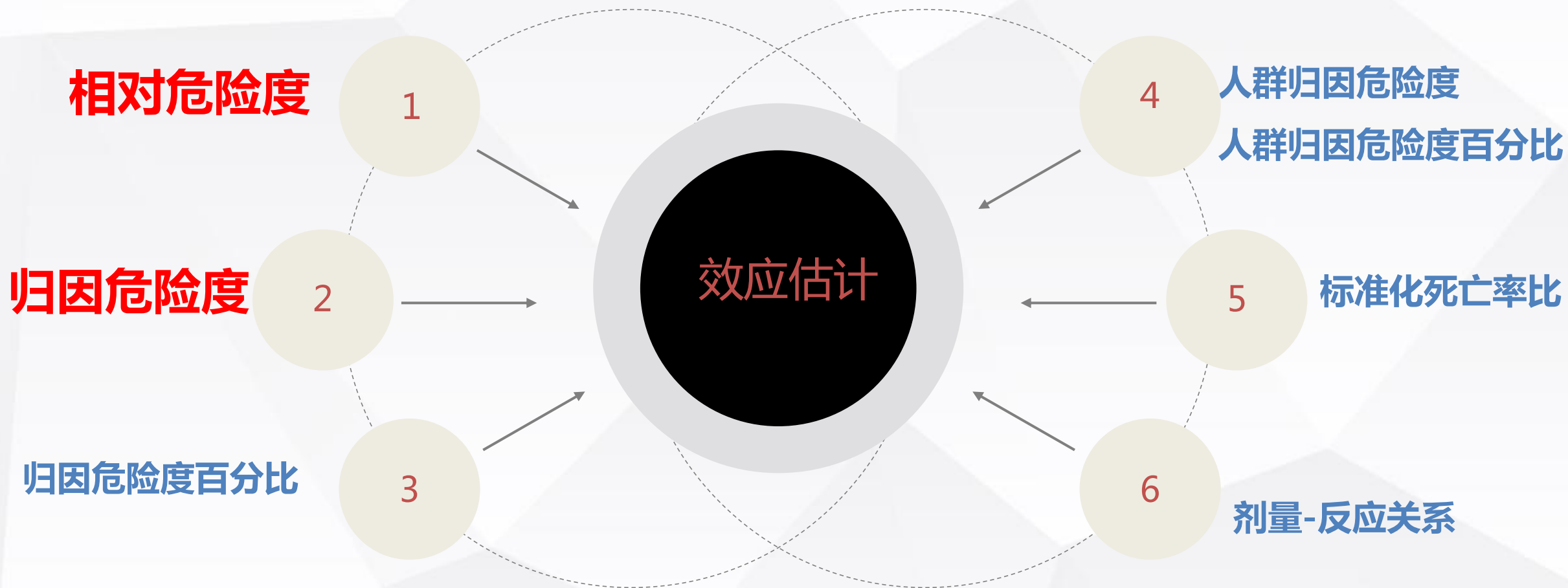
人时率的 χ^2 检验

孕期暴露于X线子女乳腺癌死亡数与观察人年数

	暴露组	非暴露组	合计
死亡数	41 a	15 c	m_1
人年数	28010 n_1	19017 n_0	t

$$\chi^2 = \frac{a - n_1 m_1 / t}{\sqrt{m_1 n_1 n_0 / t^2}} = \frac{41 - 28010(56 / 47027)}{\sqrt{56 \times 28010 \times 19017 / 47027^2}} = \frac{7.65}{3.67} = 2.08$$

三、效应估计



➤ 相对危险度（relative risk, RR）

- 即暴露组与对照组的发病率或死亡率之比。

表示暴露组发病或死亡的危险度
是非暴露组的多少倍。

暴露
组率

$$RR = \frac{I_e}{I_0} = \frac{a/n_1}{c/n_0}$$

非暴露
组率

*RR*与关联强度

<i>RR</i>		关联强度
0.9~1.0	1.0~1.1	无
0.7~0.8	1.2~1.4	弱
0.4~0.6	1.5~2.9	中
0.1~0.3	3.0~9.9	强
<0.1	10~	很强

RR 95%CI

Woolf法

$$Var \left(l_n RR \right) = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$$

l_n RR的95%CI

$$l_n RR \pm 1.96 \sqrt{Var \left(l_n RR \right)}$$

反自然对数即为RR 95%CI

➤ 归因危险度 (attributable risk, AR)

- 即暴露组结局发生率与对照组结局发生率之差

$$AR = I_e - I_0 = \frac{a}{n_1} - \frac{c}{n_0} \quad \text{或}$$

$$AR = RR \times I_0 - I_0 = I_0 (RR - 1)$$

表示暴露组中完全由于暴露所增加或降低的发病率或死亡率。

***RR*与*AR*的区别**

吸烟与肺癌和心血管疾病的RR与AR比较（1/10万人年）

疾病	吸烟者	非吸烟者	<i>RR</i>	<i>AR</i>
肺癌	48.33	4.49	10.76	43.84
心血管疾病	296.75	169.54	1.74	125.13

意义：

✓ *RR* 吸烟对肺癌的**病因学意义**较大

✓ *AR* 戒烟对心血管疾病的预防作用较大，即**公共卫生意义**较大

➤ 归因危险度百分比AR%（病因分值EF）

- 指暴露人群中的发病或死亡归因于暴露的部分占全部发病或死亡的百分比。

$$AR\% = \frac{I_e - I_0}{I_e} \times 100\% \quad \text{或}$$

$$AR\% = \frac{RR - 1}{RR} \times 100\%$$

- 意义：暴露人群中的发病或死亡归因于暴露的部分占全部发病或死亡的百分比

➤ 人群归因危险度百分比 $PAR\%$

$$PAR\% = \frac{I_t - I_0}{I_t} \times 100\% \quad \text{或}$$

$$PAR\% = \frac{P_e(RR - 1)}{P_e(RR - 1) + 1} \times 100\%$$

P_e : 总人群的暴露比例

意义: PAR 占总人群全部发病（或死亡）的百分比

● 标化死亡比

(standardized mortality ratio, SMR)

$$SMR = \frac{\text{研究人群中的观察死亡数}}{\text{以标准人口死亡率计算的预期死亡数}}$$

变化范围 0~∞

适用条件 结局事件的发生率低
样本小，不宜直接计算率时

预期死亡（发病）数的计算：全人口某病的死亡（发病）率×观察人口数

✓SMR的意义

被研究人群死于（发生）某病的危险性是标准人群的多少倍

◆ **SMR=1**

研究人群某病死亡（发病）危险=标准人群

◆ **SMR>1**

研究人群某病死亡（发病）危险>标准人群，是标准人群的SMR倍

◆ **SMR<1**

研究人群某病死亡（发病）危险<标准人群

某工厂30-45岁组工人有1250人，某年有5人死于肺癌。已知当年当地同年龄组人群肺癌的死亡率为2‰，求SMR。

观察死亡数=5；预期死亡数=1250×2‰=2.5

则：SMR=5/2.5=2

结论：该工厂30-45岁工人死于肺癌的危险是一般人群的2倍。

➤ 标化比例死亡比 (standardized proportional mortality ratio, SPMR)

$$SPMR = \frac{\text{实际死亡数}}{\text{预期死亡数}}$$

- **变化范围** $0 \sim \infty$
- **适用条件** 不能得到历年人口资料
仅有死亡人数、原因、日期和年龄

某工厂20-24年龄组工人死亡总数为**292**，其中因结核病死亡**100**人，一般人群中**201--24**岁组结核病死亡占全死因死亡的比例为**6.28%**，试评价该工厂结核病的死亡情况。

$$O=100, E=292*6.28\%=18.34$$

$$\text{则：SPMR} = 100/18.34 = 5.45$$

结论：该工厂20-24岁组工人结核病死亡为一般人群的5.45倍。

➤ 剂量-反应关系(分级分析)

分析方法

- 列出不同暴露水平下的发病率
- 以最低暴露水平组为对照，计算各暴露水平的 RR 和危险度差(RD)
- 必要时，应对率的变化作率的趋势性检验

40~59岁男子按基线血清胆固醇分组的冠心病6年发生情况

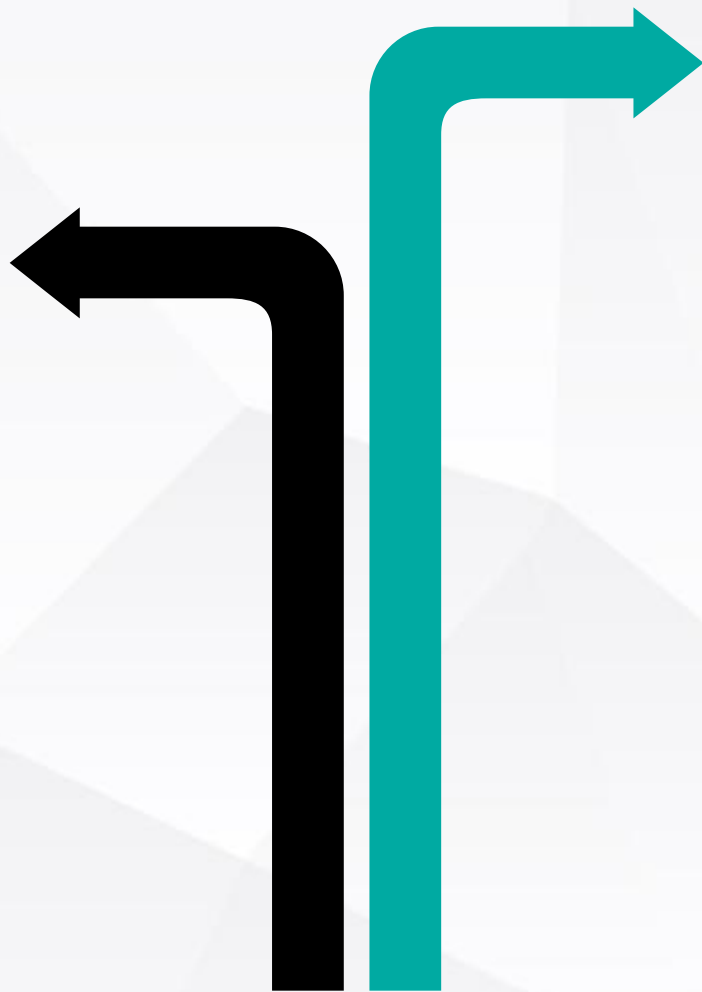
血清胆固醇 (mg/dl)	人数	病例数	危险度	平均年发病率	RR	AR
<210	454	16	0.0352	0.0059	1.00	0.0000
210~	455	29	0.0637	0.0106	1.81	0.0285
>245	424	51	0.1203	0.0200	3.39	0.0851
合计	1333	96	0.0720	0.0120		

结果

血清胆固醇水平↑ 患冠心病的RR↑
表明存在剂量效应关系

多因素分析

分层分析



第四节 偏倚及其控制

- ✚ 选择偏倚

- ✚ 信息偏倚

- ✚ 混杂偏倚

第五节 队列研究的优点与局限性



优点

- 1 符合时间顺序，验证病因的能力较强
- 2 可计算发病率或死亡率，直接估计危险度
- 3 获得一种暴露与多种结局的关系
- 4 收集的资料完整可靠，不存在回忆偏倚
- 5 可研究疾病的自然史

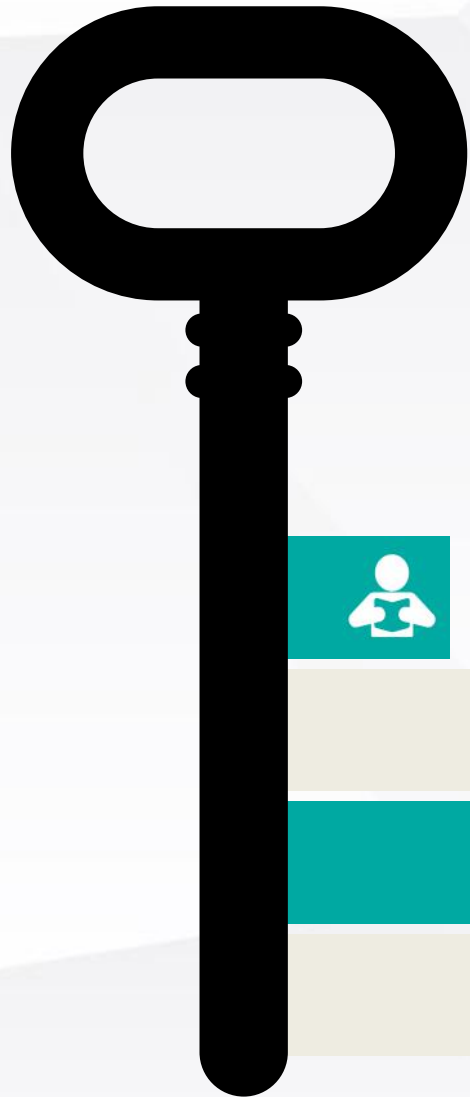


局限性

- 1 不适于发病率很低的疾病
- 2 易发生失访偏倚
- 3 耗时，耗人力、物力、财力
- 4 设计要求严密，资料的收集和分析难度较大
- 5 随访过程中，已知变量的变化或未知变量的引入增加分析难度

大型前瞻性人群队列研 究进展

最经典：FHS（美国弗明汉心脏研究）



历经3代人群队列



1948年美国“国立卫生研究院”（NIH）决定开始在马萨诸塞州 Framingham 镇启动一项心血管病研究；1977年子代；2002年第三代



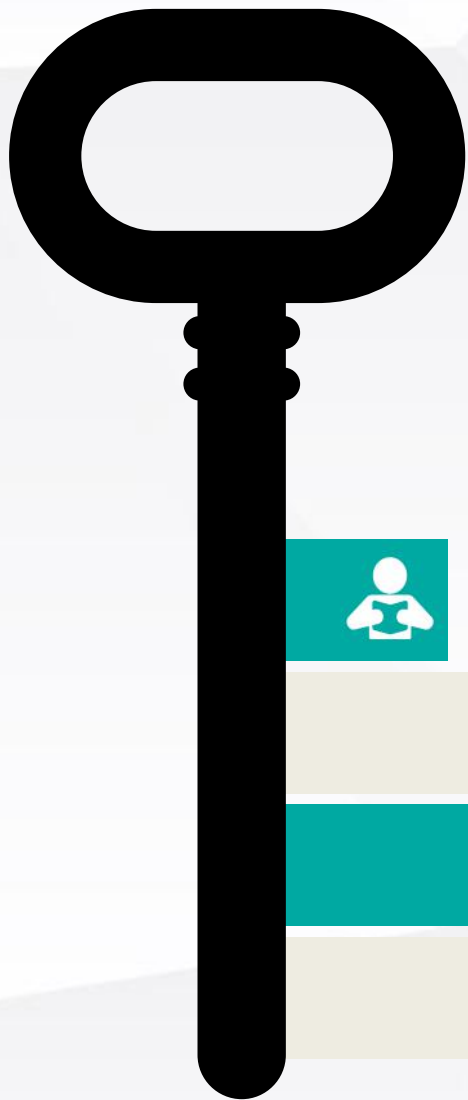
初始队列由 5 2 0 9 名 2 8 ~ 6 2 岁的男女两性组成，包括 1 6 4 4 对夫妻和 5 9 6 个家庭



发表在世界顶尖医学杂志发表 2600 余篇论文

FHS 通过多年监测和随访，鉴别了导致心血管疾病的主要危险因素：高血压、高血脂、吸烟、肥胖、糖尿病和缺乏体育运动，以及相关因素，例如血液中甘油三酯和 HDL 胆固醇水平、年龄、性别和心理因素，以及更进一步的危险因素如同型半胱氨酸等。

最经典：FHS（美国弗明汉心脏研究）



第一期（1948 — 1978 年）致力于传统研究、如临床观察、血液生化、心电图和胸片等



第二期（70 年代后期～80 年代）引进新的技术与方法进行观察，如超声心动图、颈动脉斑块测定、运动试验



第三期（80 年代后期至今）致力于分子遗传学定位与危险因素及 CVD 相关的特殊基因，进而研究其他常见疾病与基因的关系

FHS突出的优点是：前瞻性设计和长期的随访，以及研究结果的实用性

欧美达到50万规模的人群队列



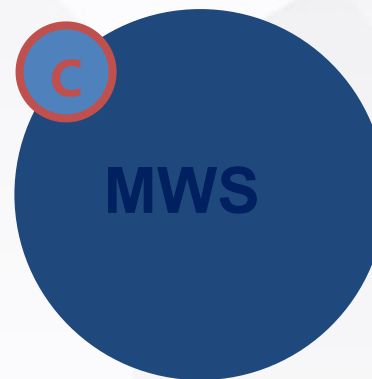
欧洲10国52.1万

European
Prospective
Investigation
into Cancer and
Nutrition



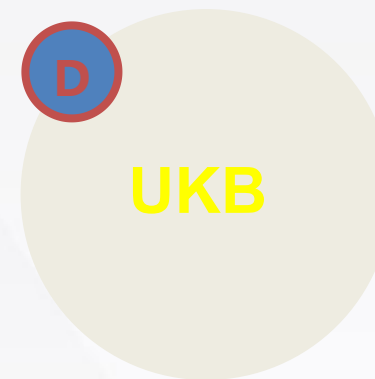
美国56.6万

Diet-Health
Study



英国130万

Million Women
Study



英国49.8万

UK Biobank

我国前瞻性人群队列研究

规模小

起步晚

数量少

缺乏稳定
长期的支持，随访
时间短



李立明教授：CKB项目

3. 中国慢性病前瞻性研究:2004年,中国疾病预防控制中心(项目Ⅱ期改为中国医学科学院)与英国牛津大学合作启动了一项大型的中国慢性病前瞻性研究项目(China Kadoorie Biobank, CKB)^[5-7]。项目目标为调查主要慢性病及相关危险因素状况,通过长期随访,探讨环境、个体生活方式、体格和生化指标、遗传等众多因素对复杂慢性病或症状的发生、发展的影响。

CKB项目具有50余万人的超大规模,建立有存储量世界领先的生物样本库,覆盖我国东北、西北、华东、华南和西南具有不同经济发展水平、社会文化背景以及暴露谱和疾病谱的城乡地区,与当地疾病和死亡监测系统有机整合,极大的提高了疾病监测能力。CKB项目为我国转型期社会人群健康状况的发展和变化、疾病谱的改变及影响因素研究提供了宝贵的人群现场;如果可以得到足够的经费支持、长期维持队列,随着随访时间的延续,必将成为我国人群复杂疾病机制及防治研究不可或缺的重要基础性资源,为制定重大慢性病防治策略和指南提供本土高质量的病因学证据。

THE END
THANKS

