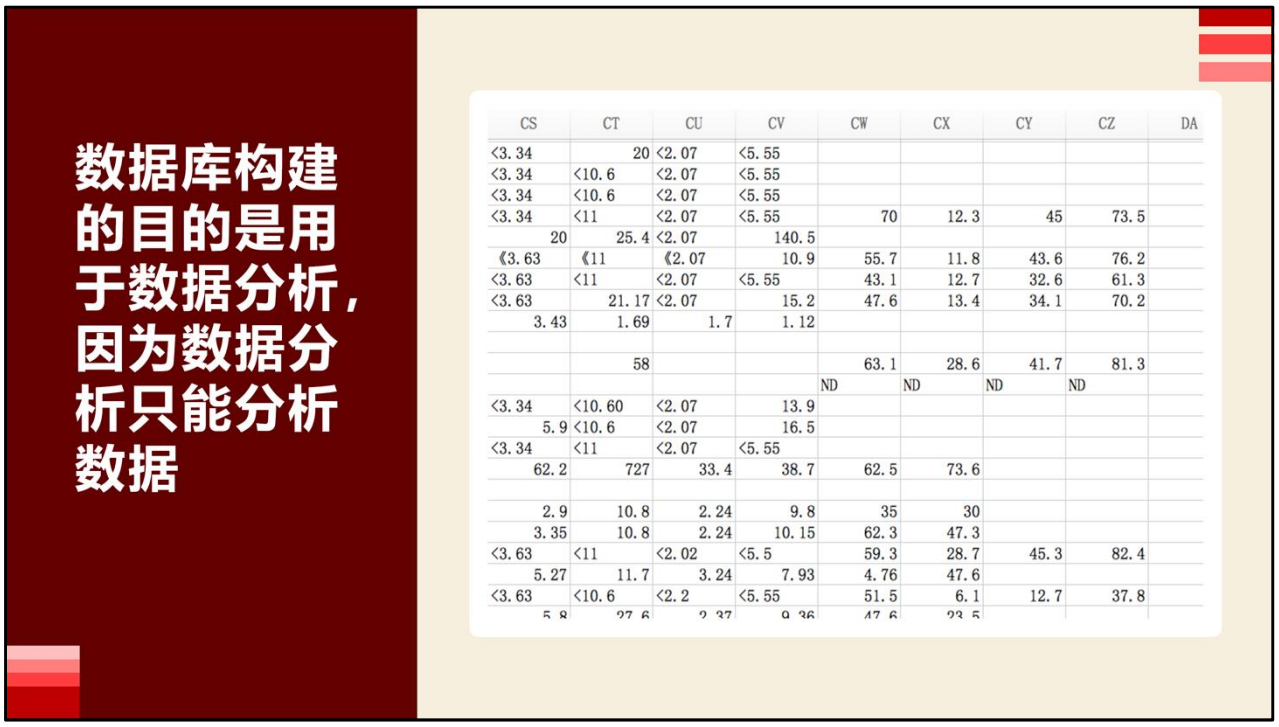
临床科研数据库构建

易佾学院　　陈驰

数据库的基本要求



很多的临床研究者在使用各种工具记录临床信息时，往往通过不规范的记录形式，导致结果中充满字符、不规则的，ID未对齐等。这些，为数据分析带来巨大的麻烦，导致此前的数据采集工作做了无用功。

结构化

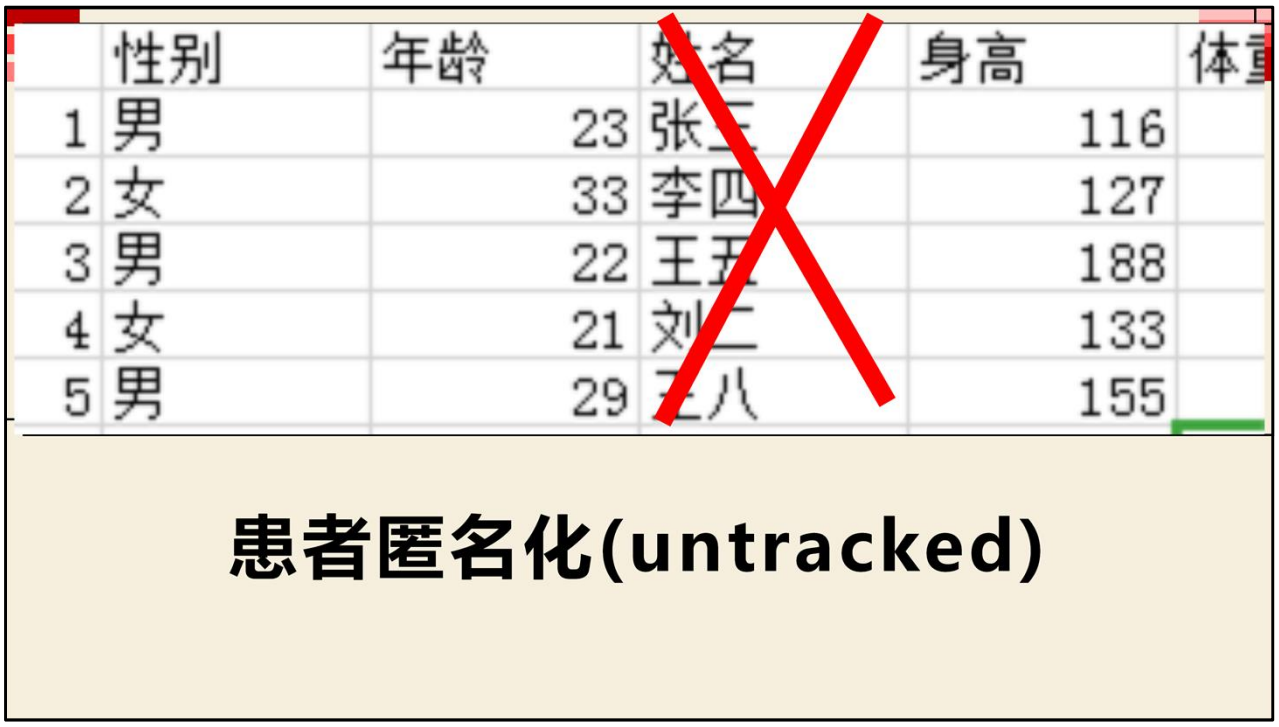
	性别	年龄	姓名	身高	体重
1	男	23	张三	116	
2	女	33	李四	127	
3	男	22	王五	188	
4	女	21	刘二	13	
5	男	29	王八	15	

因此，对于数据库构建的第一个要求，即为结构化。第一行为变量名称，第二行为变量的各种记录信息。

数字化

	性别	年龄	姓名	身高	体重
1	1	23	张三	116	
2	2	33	李四	127	
3	1	22	王五	188	
4	2	21	刘二	133	
5	1	29	王八	155	

其二，是数字化。譬如性别，如果用“男”、“女”字符表示，则无法用于数据分析。因此，必须将其赋值，数字化，才可用于数据分析。

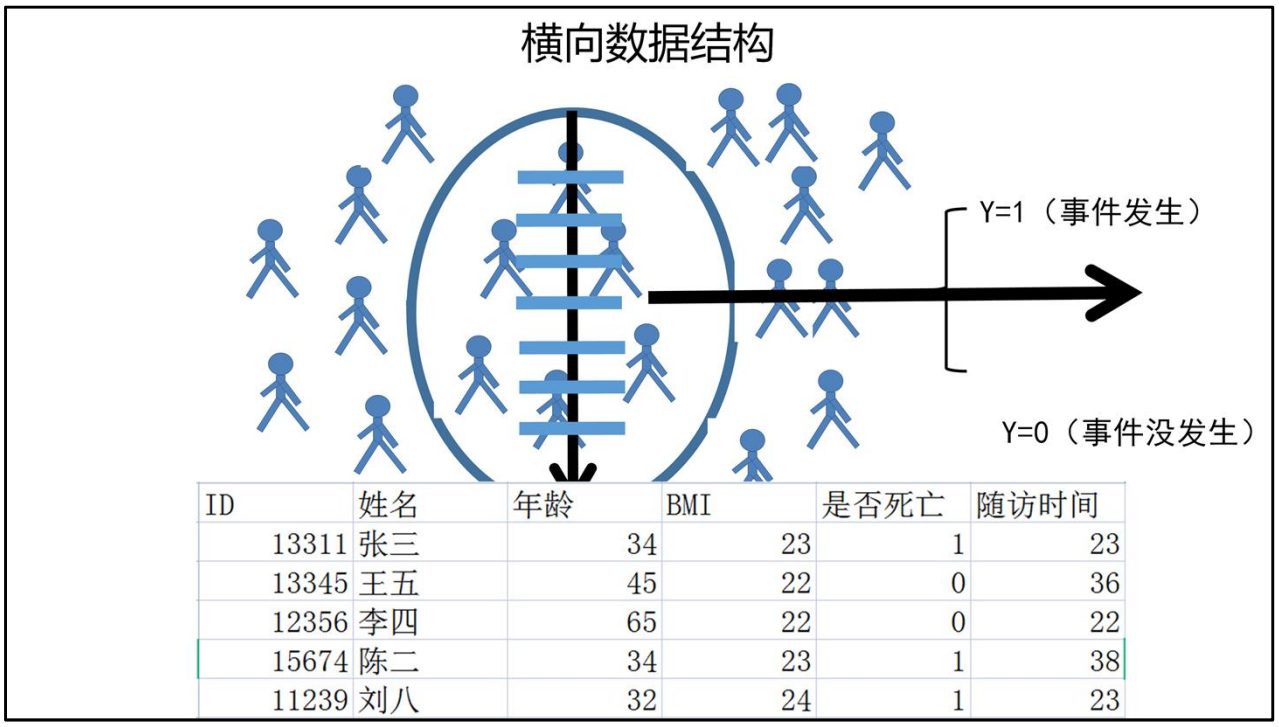


第三个标准是匿名化。在数据库中不可包含患者的个人身份信息，标准的做法是通过untracked的方法，抹去这些信息。

数据库的结构

临床研究数据库常见有两种结构

1. 横向结构
2. 纵向结构



ID	姓名	年龄	BMI	是否死亡	随访时间
13311	张三	34	23	1	23
13345	王五	45	22	0	36
12356	李四	65	22	0	22
15674	陈二	34	23	1	38
11239	刘八	32	24	1	23

横向结构指每一位研究者一条记录。常见于仅采集患者基线资料+结局时间±随访时间。

ID号	性别	年龄	姓名	身高	体重	吸烟状态	记录的时间
1	1	23	张三	116	56	0	2019/3/1
1	1	23	张三	116	56	1	2019/4/1
1	1	23	张三	116	56	1	2019/5/1
1	1	23	张三	116	56	1	2019/6/1
1	1	23	张三	116	56	0	2019/6/1
1	1	23	张三	116	56	0	2019/7/1
2	2	33	李四	127	56	0	2019/3/1
2	2	33	李四	127	56	0	2019/4/1
2	2	33	李四	127	56	0	2019/5/1
2	2	33	李四	127	56	0	2019/6/1
2	2	33	李四	127	56	0	2019/6/1
2	2	33	李四	127	56	0	2019/7/1
2	2	33	李四	127	56	0	2019/8/1

纵向结构数据库

纵向数据结构，常见于重复测量数据。即每一位患者有多条记录。如上图，ID=1的患者有6条记录，ID=2的患者有7条记录。

建议

新手在学习临床研究初期，尽可能的选择横向数据结构，方便记录、理解以及分析。

数据库的变量分类

以变量类型作为分类，可分为数值型、分类型、日期型、字符型变量。

其中，数值型变量为具体数字，常见于实验室数据等。分类变量不可用数字去描述，只能进行分类。日期型变量不可直接用于数据分析，但常用于计算时间，譬如随访时间、住院时间等，用日期相减即可得到。字符型变量，如患者姓名，检查结果的诊断结论即具体描述等。

建议：

一．尽可能减少字符型变量，通过赋值，将其变为分类型变量。

二．有的变量，可以录入为连续变量，也可以录入为分类变量，请尽可能录入连续变量形式。

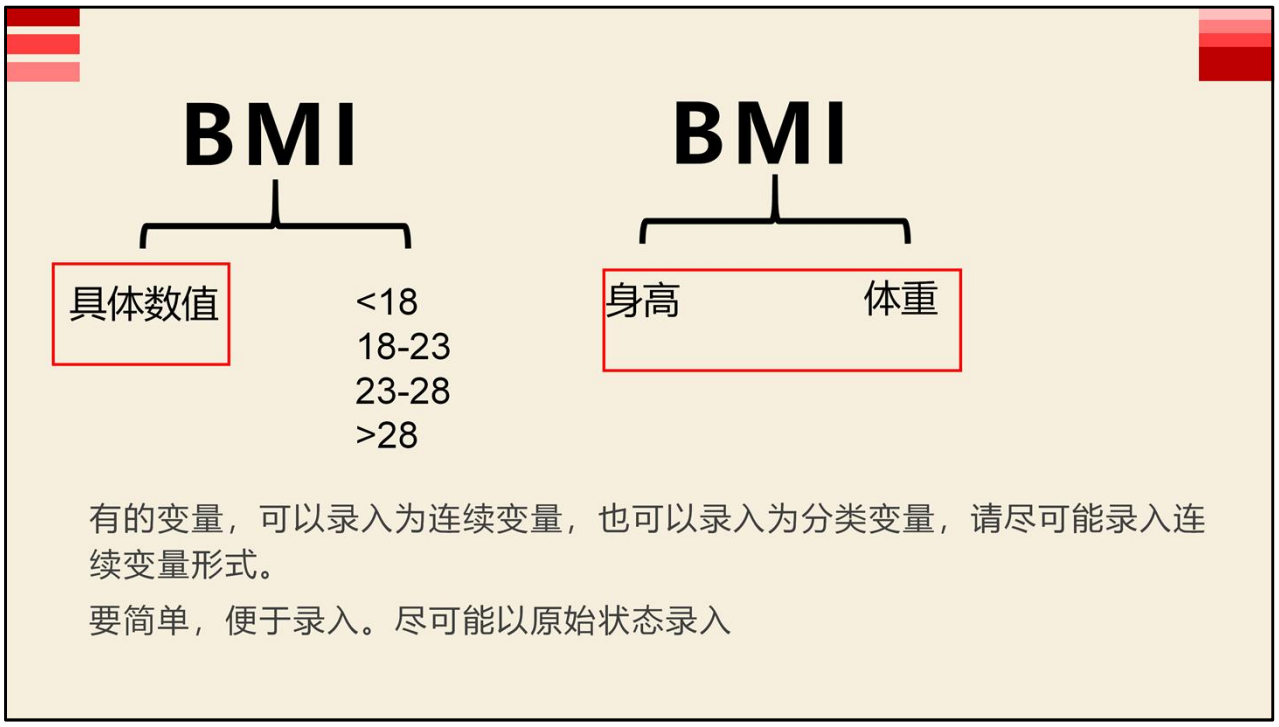
三．要简单，便于录入。尽可能以原始状态录入

以胸片为例

患者肋膈角锐利，肺纹理清晰，胸膜增厚

肋膈角是否锐利	肺纹理是否清晰	胸膜是否增厚
0=是	0=是	0=是
1=否	1=否	1=否

尽可能减少字符型变量，通过赋值，将其变为分类型变量



在录入变量时，以记录人方便为要务。
如果变量可记录为连续变量，亦可记录为分类变量，则以原始的连续变量形式录入为主。如果变量为几个变量计算所得，则只录入用于计算的变量。
以BMI为例，可录入为具体的连续数值变量，也可录入为分类变量（正常，超重，肥胖）。则只录入连续变量。而BMI为身高和体重计算生成。在设计数据库时，则只录入身高和体重即可。

分类2：变量的录入次数

01

只需要录入一次的变量，譬如性别，身高（成年人）

02

需多次录入的变量随时间变化的变量（血常规，肝肾功能、电解质、年龄等）

建议

一．将只录入一次和需要多次反复录入的变量从表单上就进行区分。

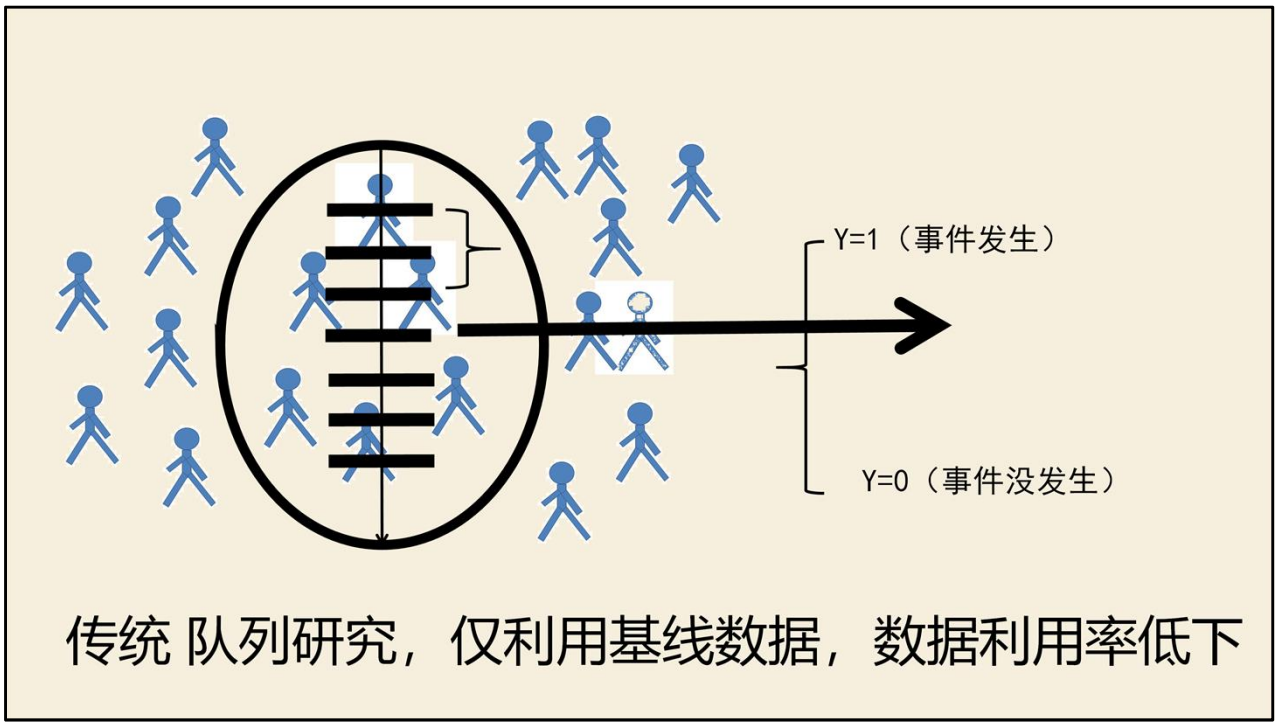
二．新手将变量均设为只需要录入一次即可。

三．但将变量（生化标志物，血清学检查）设置为多次测量并在数据库中进行录入，好处多多。

四. 需多次录入的变量（生化标志物，血清学检查）可不放入数据库内，最后一次性从医院数据库中导出。

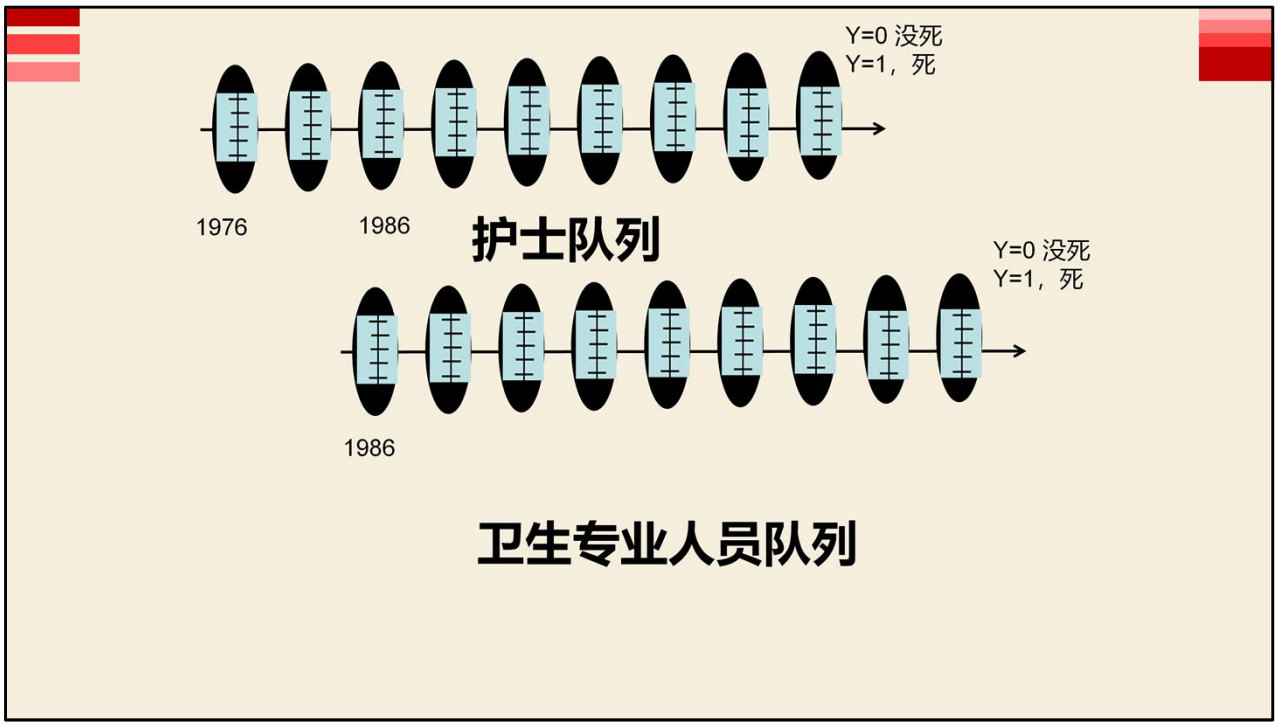
随时间变化的变量数据库设计

只录入一次， 不随时间变化 的变量（横向）	不只录入一次， 随时间变化的 变量（纵向）



然而，临床实践中

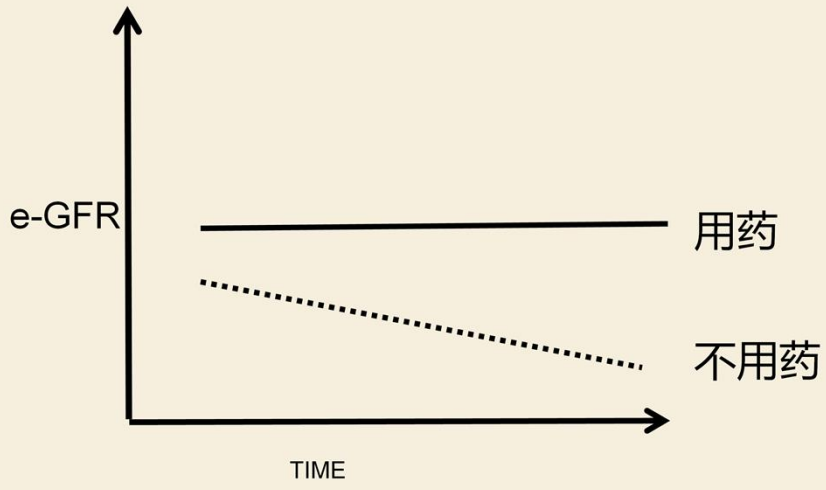
导致结局发生/不发生的原因，往往是不同因素长期，反复，动态变化所造成的。所以，这种以基线和结局的关联，往往无法代表临床的真实情况。所以，研究随时间动态变化的指标与结局 的关联，是未来 的趋势。



例如目前很多顶级的队列数据库，大多为纵向格式的数据库。譬如著名的护士队列及卫生专业人员队列，即记录为纵向数据格式。

因此，我们要设立随时间变化的变量，其原因之一在于能更好的符合临床的真实情况。

譬如：用/不用某药物，对于糖尿病肾病是否进展为ERSD

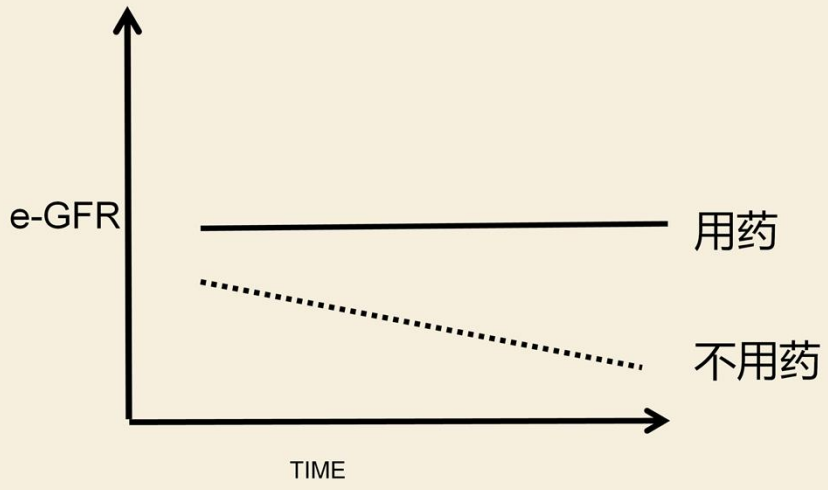


姓名	日期	Cr
张三	1.7	44
张三	1.9	19
张三	1.13	113
张三	1.22	122
张三	1.27	127
李四	2.8	280
李四	2.14	214
李四	2.23	223

此外，纵向数据格式增加了对数据的利用率。譬如，当我们研究某药物是否对肾功能有保护作用时，通过eGFR在不同用药状态下的变化趋势，即可得出该药物是否保护肾功能。而本例中的eGFR只有通过纵向数据格式才可实现对其利用。

因此，我们要设立随时间变化的变量，其原因之二在于能更好的提高数据的利用率。

譬如：用/不用某药物，对于糖尿病肾病是否进展为ERSD



姓名	日期	Cr
张三	1.7	44
张三	1.9	19
张三	1.13	113
张三	1.22	122
张三	1.27	127
李四	2.8	280
李四	2.14	214
李四	2.23	223

此外，纵向数据格式还有第三个好处。即可以提高统计学效能。该方法连RCT都在应用（n of 1设计）。本例中，张三有5条机记录，李四有3条记录，尽管只有两位患者，但其统计学效能得到提高（n=8）

因此，我们要设立随时间变化的变量，其原因之三在于能更好的提高统计学效能。

如何建立随时间变化的纵向数据库，其原则即为“流水账”。

姓名	住院号	年龄	身高	性别	PLT	记录时间
张三	11021	36	177	1	113	2021.07.06
张三	11021	36	177	1	89	2021.07.09
张三	11021	36	177	1	112	2021.07.12

譬如，对于张三这名患者，记录了三条血小板数据。但对于每次录入时，录入者可不需要动“脑子”去思考谁是第一次测量，谁是第二次测量，而仅需要记录录入时间，易侕的dataweb可自动转变为纵向数据格式，并排序。

其他

SEX

DRINK

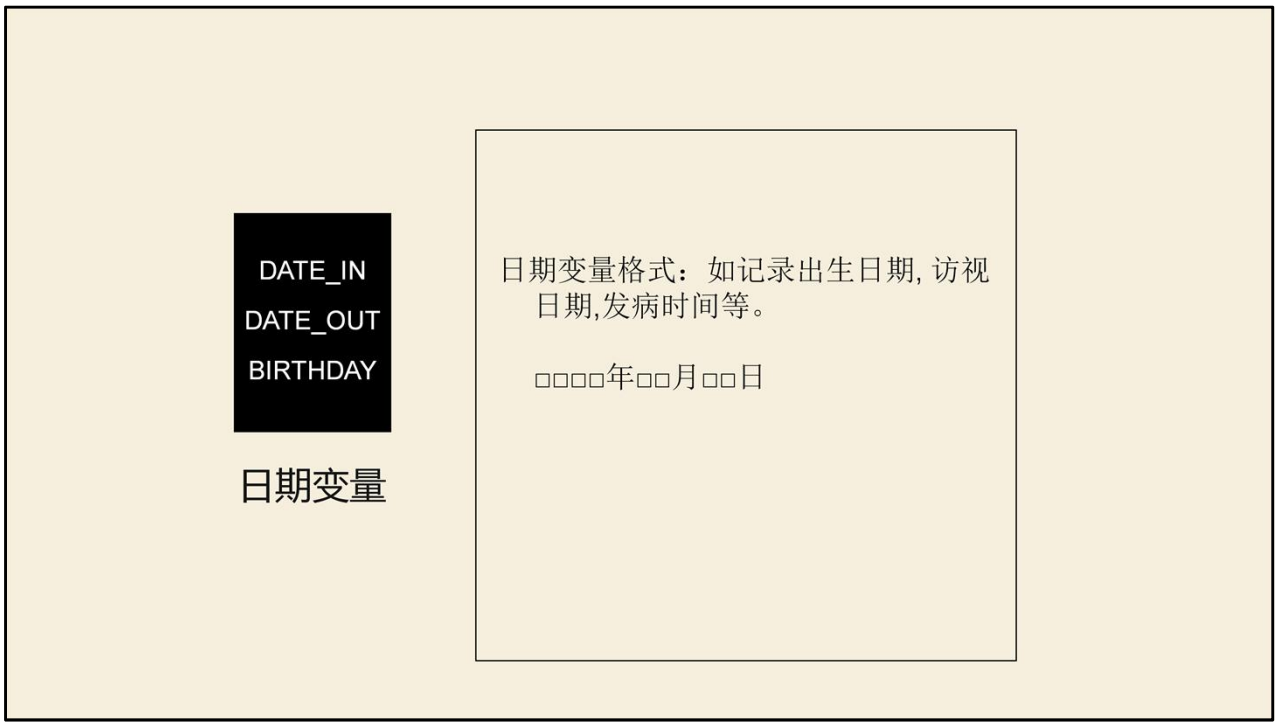
TYPE

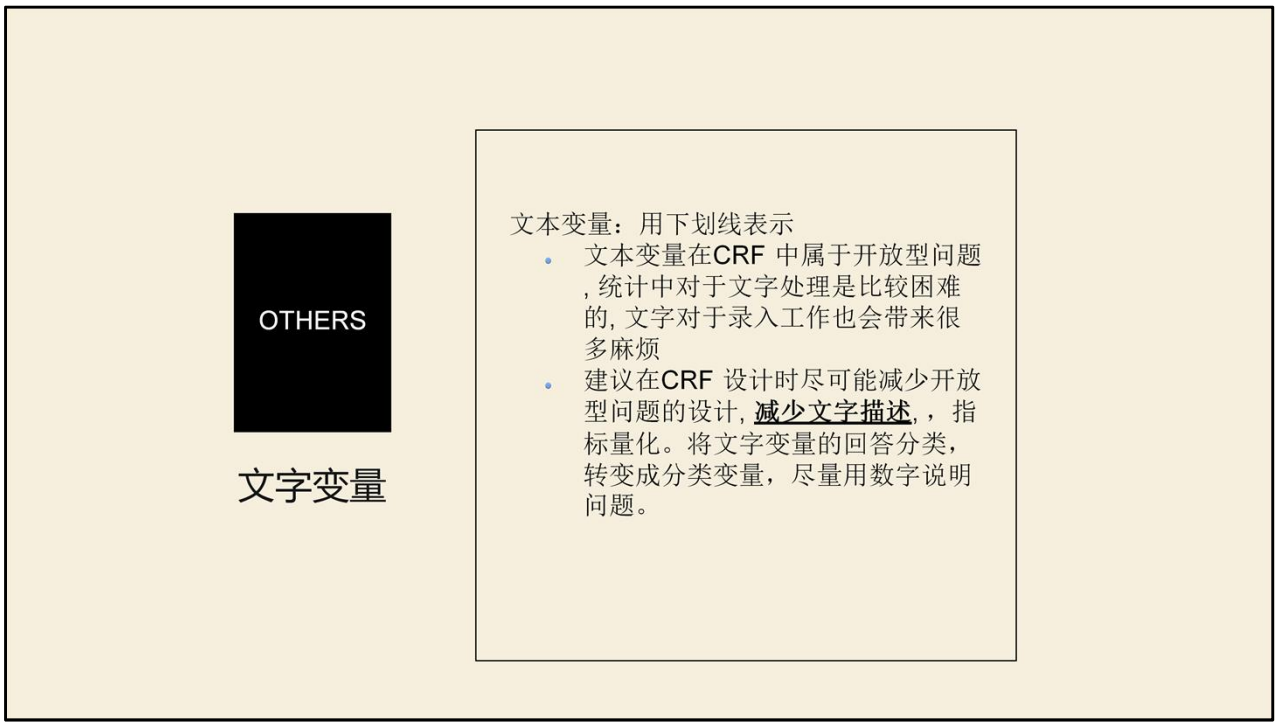
分类变量

多值指标：将所有问题的备选答案提供给研究者，各选项间互相排斥不重叠：

如“疼痛 无□ 轻□ 中□ 重□”，

不要设计成，“疼痛 有□ 无□，
轻□ 中□ 重□”





一些共性的问题

对于复合问题的设计，应设计成多个变量。

- 如“你是否吸烟喝酒？是□，否□”；
- 应设计成“您是否吸烟？是□，否□”……

问题应使用肯定句，多重否定容易使回答错误。如“”受试者是否没有按方案完成治疗？”等容易答错。

建议

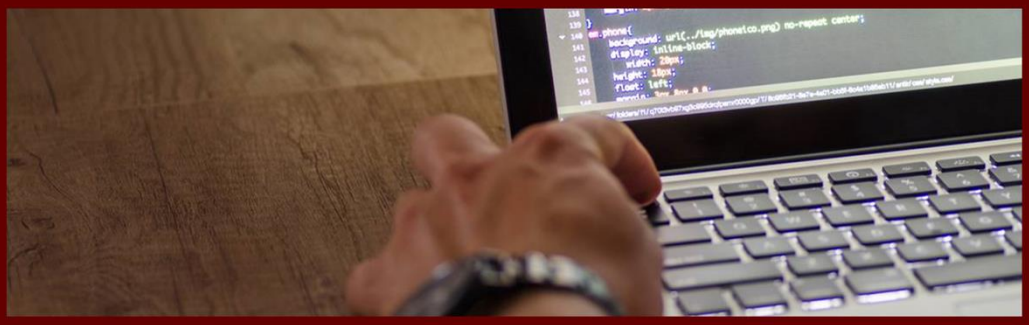
一．CRF 设计不是一项孤立的工作，与前后的方案设计和数据库设计密切相关 。

二．在CRF设计的同时，必须有数据管理/生物统计学者的参与和配合，充分考虑CRF的数据录入、统计分析的需要。

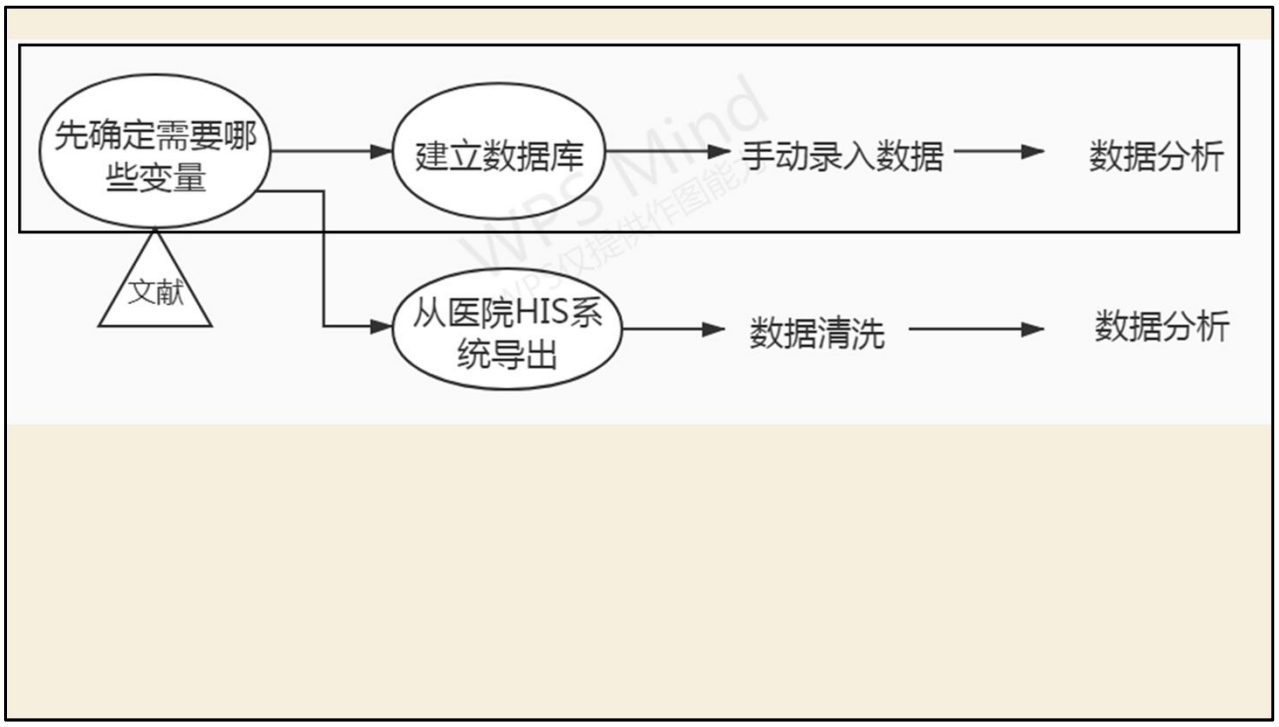
三. 此外，设计后的CRF表格，可先用既往的数据进行试录入，一方面方便改进流程；一方面用于锻炼团队。

四. 完成单中心后，再尝试多中心。

个人/小团队建库策略



盖浇饭策略



何为“盖浇饭”策略？即先通过文献及临床经验，确定需要哪些变量，然后在根据变量库去建立数据库，手动输入数据，最终用于数据分析。该策略适用于单人、小团队建库。因此，盖浇饭之意即为不贪大求全，而仅选择最为需要的变量，从而提高建库的可行性，保证工作量与效率的平衡。

实例

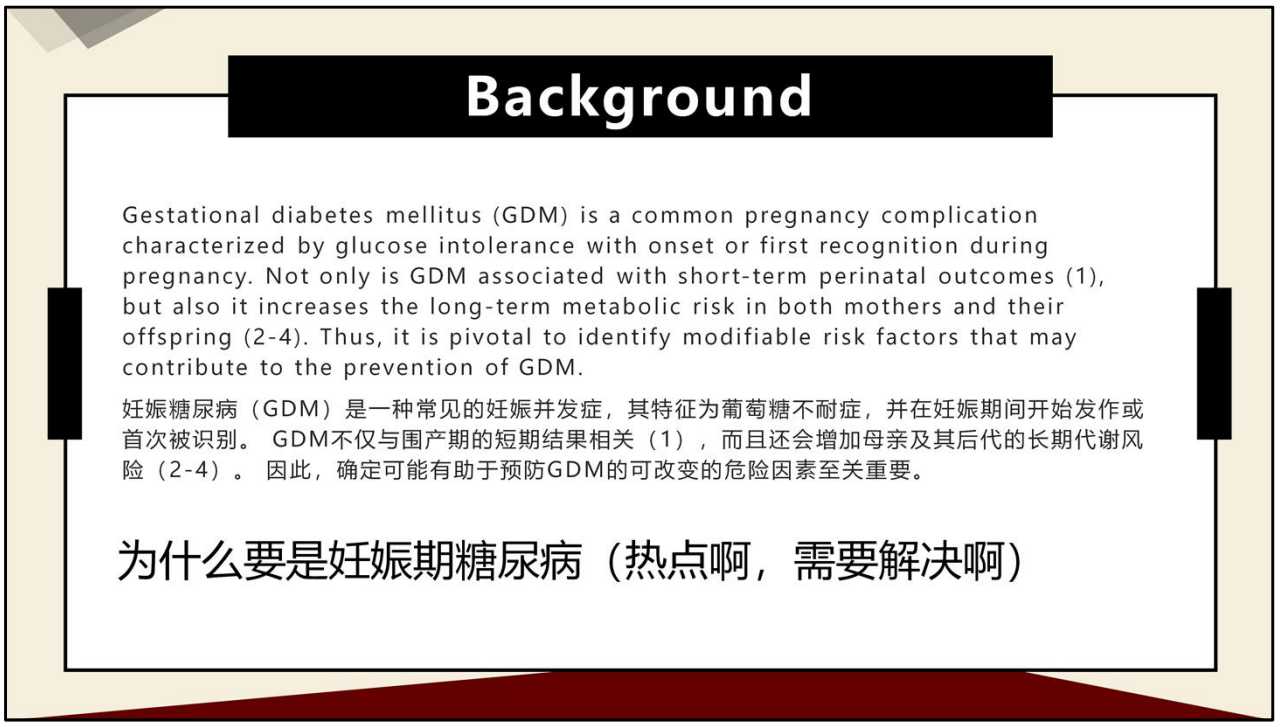
举例：美国护士队列中，有几千个变量，而研究者仅想研究的是土豆摄入与妊娠期糖尿病的关联。（即食堂中有众多食材，我只需要青椒肉丝）。

A prospective cohort study of prepregnancy potato consumption and risk of gestational diabetes

妊娠期土豆消耗和妊娠期糖尿病的风险，前瞻性队列

X=妊娠期土豆消耗　Y=妊娠期是否发生糖尿病

因此，作者提前做好了protocol，进行建库数据申请。



作者先介绍了背景，为何要观察XY的关联。

Background

Potato and potato products are widely consumed in the United States, and they have been continuously included in the vegetable category as a food to be encouraged by the Dietary Guidelines for Americans (5). However, high potato consumption may have detrimental effects on glucose metabolism because they contain large amounts of rapidly absorbed starch and high glycemic index (6). Epidemiological studies examining the association between potato consumption and risk of type 2 diabetes have yielded mixed findings (7-9). The association between potato consumption and risk of GDM remains unknown.

马铃薯和马铃薯产品在美国被广泛消费，并且作为美国人的饮食指南所鼓励的食品，它们已连续被列入蔬菜类别（5）。 但是，大量食用马铃薯可能会对葡萄糖代谢产生不利影响，因为它们含有大量快速吸收的淀粉和高血糖指数（6）。 流行病学研究检查了马铃薯食用量与2型糖尿病风险之间的关系，得出的结论不一（7-9）。 马铃薯食用量与GDM风险之间的关联仍然未知。

Objective
To examine the association between potato consumption and risk of GDM

Study population
The Nurses' Health Study (NHS) II (1991-2001)

- Unit of observation = individual pregnancies (not persons)
- 1991 is the first year in which FFQ was collected
- 2001 is the last year in which GDM was ascertained from the main questionnaire

Exclusion criteria
(1) Prior gestational diabetes
(2) Chronic disease (type 2 diabetes, CVD, cancer) prior to GDM pregnancy
(3) No valid dietary information or missing >70 items
(4) No pre-pregnancy FFQ
(5) Implausible total energy intake (<600, >3500 kcal/d)
(6) Prior twin/multiple births
(7) Missing information on age, parity

然后，设定纳排标准，涉及人群年份等。

Exposure variables
Potato consumption (total; baked, boiled, or mashed; French-fried)

Outcome variable
Incident GDM

Statistical method
Multivariable log-binominal models with generalized estimating equations (GEE)

最终，设定XY

(1) Model 1: adjusted for age (in months) and parity (0, 1, 2, 3+).
(2) Model 2: Model 1 + race/ethnicity (Caucasians, African-American, Hispanic, Asian, others), family history of diabetes (yes, no), cigarette smoking (never, past, current), alcohol intake (0, 0.1-5.0, 5.1-10.0 or > 10 g/day), physical activity (quartiles), and total energy intake (quartiles).
(3) Model 3a: Model 2 + alternate healthy eating index (quartiles).
(4) Model 3b: Model 2 + consumption of red meat (unprocessed and processed), fruits, vegetables, whole grains and sugar-sweetened beverages (all in quartiles).
(5) Model 4a: Model 3a + BMI (nine categories; < 21, 21-22.9, 23.0-24.9, 25.0-26.9, 27.0-28.9, 29.0-30.9, 31.0 -32.9, 33.0-34.9 and ≥ 35.0 kg/m^2).
(6) Model 4b: Model 3b + BMI (nine categories; < 21, 21-22.9, 23.0-24.9, 25.0-26.9, 27.0-28.9, 29.0-30.9, 31.0 -32.9, 33.0-34.9 and ≥ 35.0 kg/m^2).

通过不同策略，告知需要哪些变量。最终，整个护士队列中的几千个变量，仅需要作者规划的XY，以及模型中需要调整的变量即可。

References

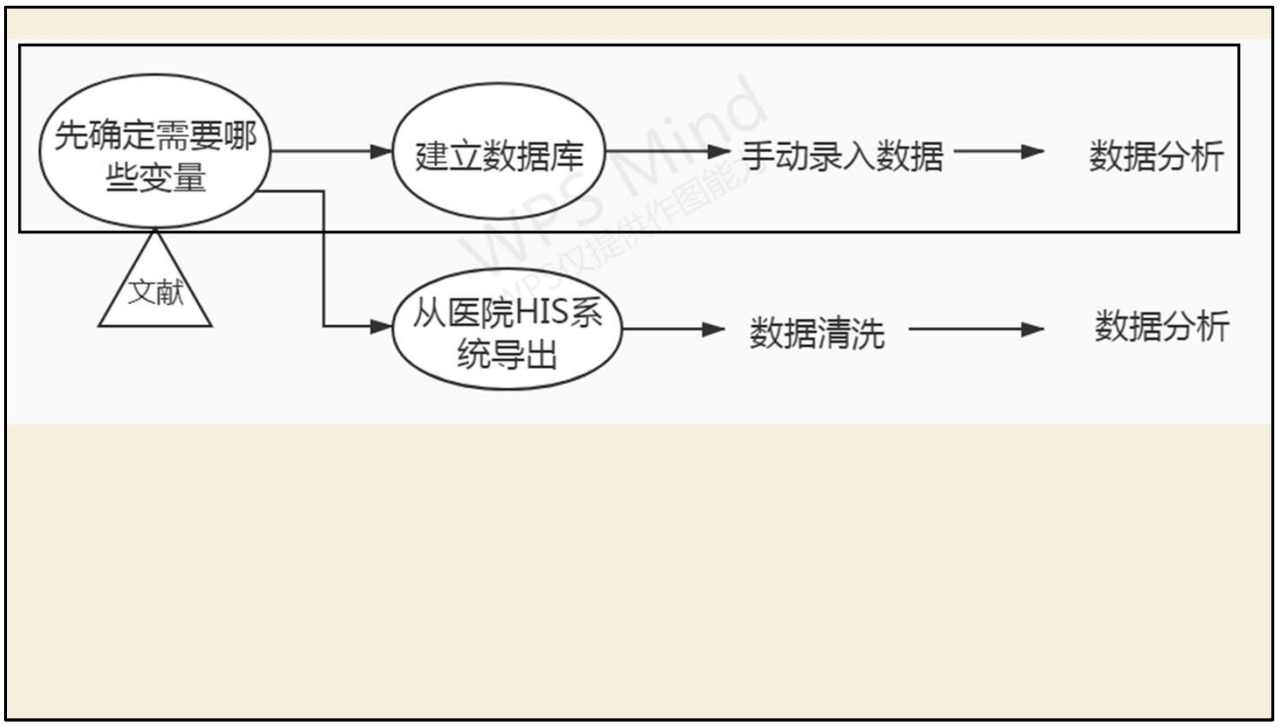
1. Metzger BE, Lowe LP, Dyer AR, et al. Hyperglycemia and adverse pregnancy outcomes. N Engl J Med 2008;358:1991-2002
2. Reece EA, Leguizamon G, Wiznitzer A. Gestational diabetes: the need for a common ground. Lancet 2009;373:1789-1797
3. Bellamy L, Casas JP, Hingorani AD, Williams D. Type 2 diabetes mellitus after gestational diabetes: a systematic review and meta-analysis. Lancet 2009;373:1773-1779
4. American Diabetes Association. Gestational diabetes mellitus. Diabetes Care 2004;27 Suppl 1:S88-90
5. U.S. Department of Agriculture and U.S. Department of Health and Human Services: Dietary Guidelines for Americans, 2010. 7th ed. Washington, DC, December 2010
6. McGill CR, Kurilich AC, Davignon J. The role of potatoes and potato components in cardiometabolic health: a review. Ann Med 2013;45:467-473
7. Halton TL, Willett WC, Liu S, Manson JE, Stampfer MJ, Hu FB. Potato and french fry consumption and risk of type 2 diabetes in women. Am J Clin Nutr 2006;83:284-290
8. Montonen J, Jarvinen R, Heliovaara M, Reunanen A, Aromaa A, Knekt P. Food consumption and the incidence of type II diabetes mellitus. Eur J Clin Nutr 2005;59:441-448
9. Liu S, Serdula M, Janket SJ, et al. A prospective study of fruit and vegetable intake and the risk of type 2 diabetes in women. Diabetes Care 2004;27:2993-2996

从这个实例我们得到的启示是：

一．明确要从总库中，提取哪些变量。一般而言，一篇SCI，顶天25-30个变量，尤其对于样本量较小的研究，变量多了反而是负担和累赘。

二．这个研究假设有没有临床价值？从回答三个问题着手。

三．想清楚要做的事情（数据分析，话术），为未来写文章奠定基础。



从这个实例我们得到的启示是：

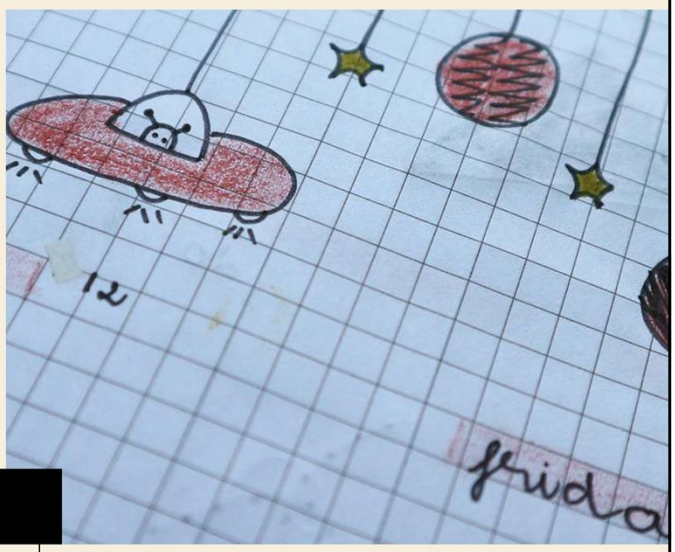
一．明确要从总库中，提取哪些变量。一般而言，一篇SCI，顶天25-30个变量，尤其对于样本量较小的研究，变量多了反而是负担和累赘。

二．这个研究假设有没有临床价值？从回答三个问题着手。

三．想清楚要做的事情（数据分析，话术），为未来写文章奠定基础。

第一步：读文献

以关键词“疾病名称”进行检索，并限定出现范围为TITLE/ABSTRACT或TITLE，然后读高分文章题目，大致判定目前高分的文章主要围绕什么在做，我们能否做到



因此，作者建库首先得要获取变量信息。一般而言，策略是先在pubmed中以所要研究的疾病名称作为关键词，找到同类研究。

- **窍门1：读综述，综述里面有大量参考文献及作者对这些文献 的总结，而且对于这个领域里面有什么不足，也列举的很清楚。**

“

- 窍门2：读protocol。尤其是队列的protocol最好。如果没有队列的，就用RCT的protocol。然后，改造成队列研究（真实 世界）。最后，用话术去弥补证据级别上的差异

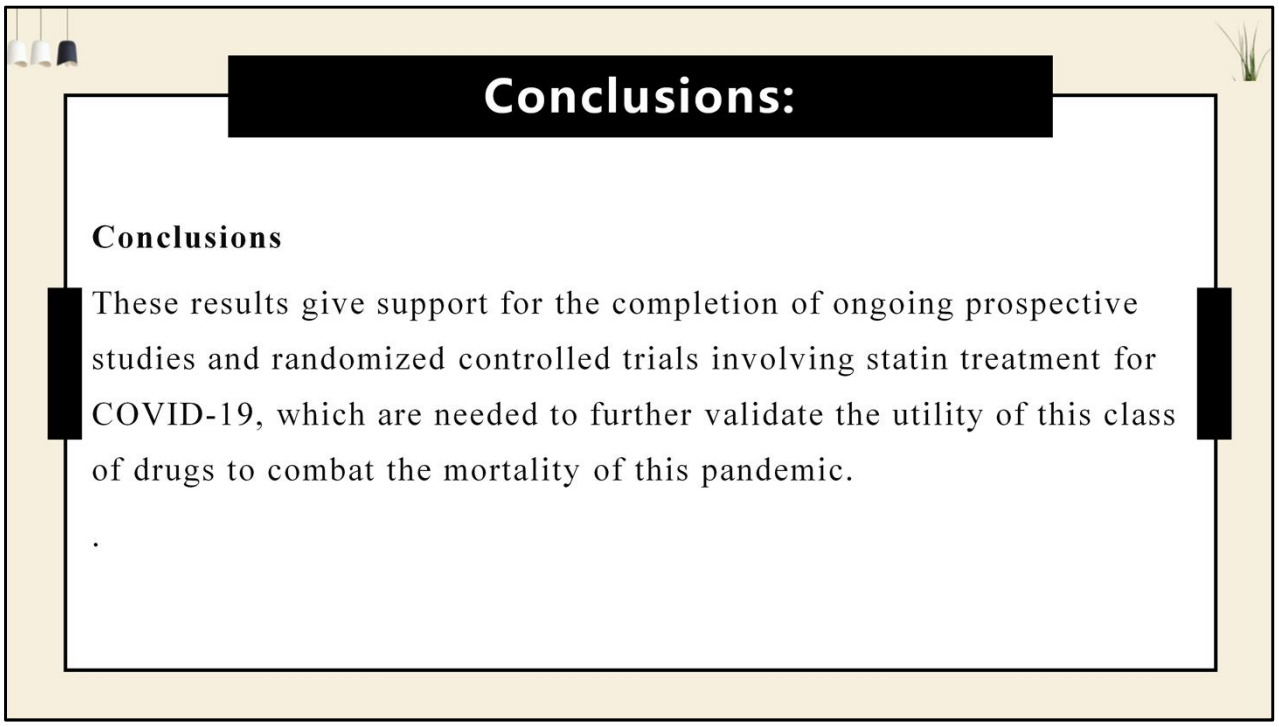
”

题目

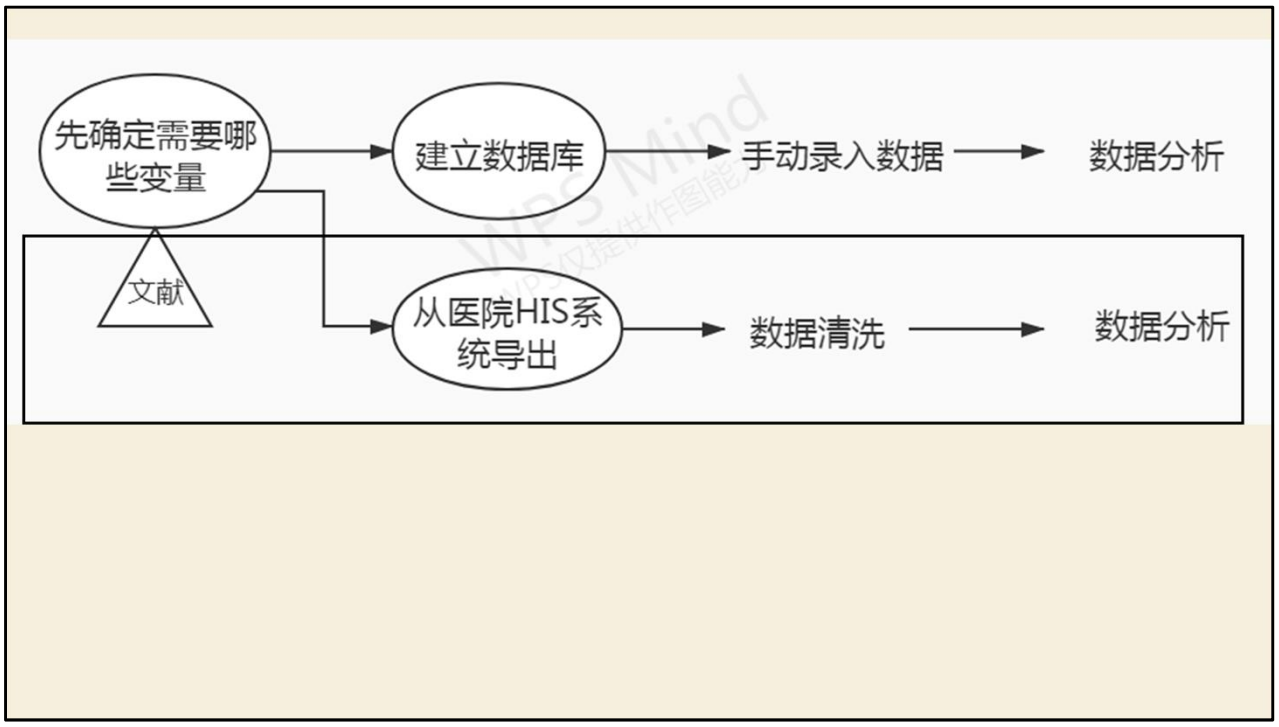
In-Hospital Use of Statins Is Associated with a Reduced Risk of Mortality among Individuals with COVID-19

Cell Metab 2020 Aug 4;32(2):176-187.e4. IF: 22.4150

譬如这篇例文，X是他汀，结局是新冠患者的死亡风险。



这篇文章的来源即为此前发表的RCT。因此，需要对此前发表的RCT文献进行阅读，即可获取需要哪些Z（协变量）



策略2，则为基于医院his系统导出数据，先进行清洗，再通过文献中提到的调整变量，进行数据分析。

HIS系统的导出需要一系列的 行政流程。不同的医院各有不同。因此，需要区别对待，无可一概而论。

HIS系统导出的数据主要包含

1 人口学数据 ■ 多不随时间变化，一次性获取

2 实验室数据 ■ 随时间变化，获取较麻烦，手动输入太累

3 治疗相关数据 ■ 如果针对真实世界研究，随时间变化
仅仅只是作为协变量调整，不随时间变化， 单次获取

3 影像超声数据 ■ 多为文字，先根据具体需求，将需要的变量列出，
再行处理

普通数据库存储形式1（横向数据）

住院号	性别	年龄	Cr1	Cr2	Cr3	BUN1	BUN2	BUN3	抗生素
1001	1	39	22	24	33	21	23	23	0
1002	2	55	35	67	45	44	19	45	1

普通数据库存储形式2（纵向数据）

住院号	性别	年龄	Cr	BUN	抗生素
1001	1	39	22	21	0
1001	1	39	24	23	0
1001	1	39	23	23	0

这是我们此前提到的纵向与横向数据库

住院号	性别	年龄	变量名	值	抗生素
1001	1	39	Cr	21	0
1001	1	39	Cr	23	0
1001	1	39	Cr	23	0
1001	1	39	BUN	34	0
1001	1	39	BUN	45	0
1001	1	39	BUN	45	0

HIS系统导出的数据

然而，HIS系统导出的数据，其“纵向”的程度更高。因此，需要用易侕的数据清理工具中的“非结构化数据工具”中的“医院化验室电子记录信息处理”及“医院医嘱电子记录信息处理”进行提取和清洗。

“

因此，在清洗HIS系统数据时，
采取的策略是各个击破

”

谢谢聆听

谢谢