建立临床数据库前的准备

刘岳鹏

徐州市中心医院

【摘要】 临床数据库是临床科研必不可少的基础，其建立的准备过程可以归纳为三步：定调、确定内容和形式标准化。通过以上三个步骤的实施形成一个临床研究病例报告表和一个可靠、易用的数据载体，最终为临床数据库的建立做好准备。

【关键词】 临床数据库

循证医学时代，临床数据库通过收集有研究价值的临床诊疗数据，为临床经验的积累提供数据支撑，并孕育出一篇篇临床研究论文，是科室发展和个人成长过程中不可或缺的助力。现阶段，各临床科室也逐渐认识到建立临床数据库的重要性和必要性。那么，在建立临床数据库之前，哪些准备是必要的呢？这里总结了三点：

一、定调

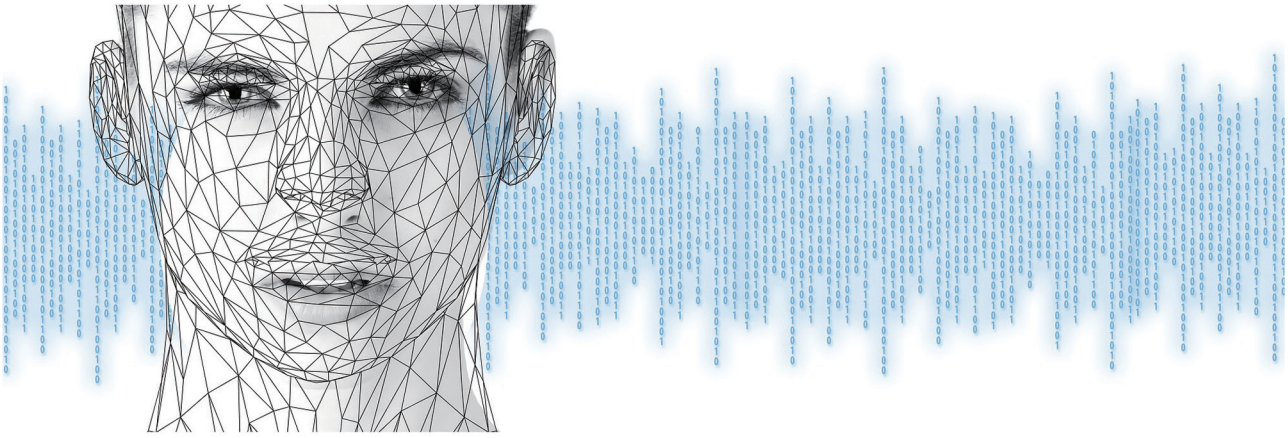
首先，根据个人或者科室目前的能力和需求，确定需要哪种类型的数据库。根据复杂程度，数据库分为三类：单任务型，其内容是针对一个试验项目，包含一个核心结局变量集，数据量少，收集工作量不大，但是数据不能反复利用；多任务型，其内容针对多个可能的试验项目，包含多个核心结局变量集，数据量大，收集工作量大，数据可以反复利用[1]；多中心型，是建立在多个机构之间的多任务型数据库，除了兼具多任务型的特点外，在数据管理、数据保密等方面具有自己的特点（表1）。表格中提到的“核心结局变量集”是临床科研标准化的概念之一[2]，是指通过最少且必需的一系列变量对某个结局事件进行评价，其目的是避免变量过多，浪费人力物力，又避免变量过少，不能全面评价结局事件。

二、确定内容

1. **确定临床数据库的主题和特色。**确定数据库的主题，即确定个人或科室感兴趣的研究领域，比如乳腺癌、甲状腺疾病等。数据库主题可以有其独特之处，比如特殊的疾病，也可以和其它的数据库一样。唯一注意的是，对于感兴趣的疾病是否能够获得足够的病例数量，没有足够的病例数量就无法得出确实的研究结果。另外，数据库建设一定要有特色，一方面是考虑到撰写有新意的论文，另一方面也是为了做出创造性的工作。这里是“创造性”，不是“创新性”。“创新性”在医学基础研究中比较重要，然而，在临床研究中不单纯强调创新性，因

表1. 临床数据库的类型及其特点

类型	内容	特点
单任务型	针对一个试验项目,包含一个核心结局变量集	数据量少,数据不能反复利用
多任务型	针对多个可能的试验项目,包含多个核心结局变量集	数据量大,数据可以反复利用
多中心型	多个研究机构	兼具多任务型数据库特点



为许多临床研究的目的是为了为临床诊疗提供证据支撑，而不是为了产生新的医学理论和方法，相似设计的多个临床研究为后续系统性综述的实施提供了便利。

数据库的特色从以下几个方面考虑：1）特色的预测变量（也称作自变量）。结局变量是评价某个疾病转归的一系列变量，通常是标准化的，而预测变量是影响结局的因素的变量，其随着新治疗技术、新的诊断方法、新的药物的出现而不断变化，是一个数据库特色的、可定制的方面；2）足够多的病例数。大样本的观察可以为临床实践提供最确实的证据，增加论文结论的可信度。一个临床数据库在各方面都普通，却包含大量的样本，也是非常有价值的数据库；3）系统地保留血液、组织等样本。物以稀为贵，组织、血液等样本在临床上是珍贵、不易得的，包含这些样本的临床数据库自然就是珍贵的。即时或者未来对样本的检测都可以为我们了解疾病提供重要的信息；4）特色的研究人群。不同的人群可能是不同的民族，有不同的生活习惯或对某疾病的具有不同的易感性的人群，不同人群的临床研究为了解疾病提供了多样的信息。

2. **确定需要收集的变量。**需要收集变量也分为三类：1）结局变量（集），以多个变量从多个角度来全面评价一个临床事件结局，这些变量被称作“核心结局变量集”，其内容是相对固定的、标准化的；2）预测变量，这是有特色的，可以由研究者根据研究目的定制的部分；3）其它变量，包括与预测变量有关的变量，和与结局变量有关的变量。收集这类变量的目的是为了从中筛选出混杂变量，从而在多因素分析过程中排除混杂变量的影响而得出预测变量对结局变量的相对“独立”作用。

确定这三类变量的方法一般是通过专家共识、文献查询、参考模板和依照标准四个途径。前三个途径照字面的含义，而“依照标准”的含义是参考“临床数据交换标准协会（CDISC）”制定的标准来设计需要收集的变量的种类[3]。原则上，收集的变量不能过多，会消耗大量的人力和物力；变量不能过少，会遗漏重要数据，要在这两者之间取得平衡。变量的收集的种类最终归纳在临床研究病例报告表（CRF）中。

3. **确定纳入的人群。**建立数据库之初，要在一定程度上明确今后要进行的临床研究形式（诊断研究、病因研究或预后研究）进而确定纳入标准和排除标准，因为这关系到对照组人群的纳入。例如，预后研究是研究治疗方案的有效性，可以只纳入患者，预后好的与预后不好的患者互相作为对照；在诊断研究中要研究与诊断相关的因素，则要纳入病种或综合征相似的、需要鉴别诊断的病例作为对照；病因研究中还要纳入没有患病的人群作为对照。纳排标准的确定根据以下原则：1）纳入标准有四个根据，分别是：临床特征、时间特征、地理特征和人口学特征[4]；2）排除标准：原则上为了增加外推性，尽可能不设排除标准，但是容易失访，有潜在不良反应，无法提供数据，可以列为排除标准[4]。

三、形式上标准化

1. **变量名的标准化：**需要遵循的几个原则：尽量长使其具有自明性，尽量短而方便输入；避免使用空格和特殊字符。例如，SubjectID、FName、ExamDate、WghtKg、HghtCm、LabID[4]。个人手

工建立的，小型的数据库可以借鉴举例进行变量命名。此外，目前数据库建设领域有个“公共数据元”的概念，是“临床数据交换标准协会（CDISC）”提出的，已经被业界普遍接受的概念，其主张同领域的数据库用统一的变量类别、名称、格式、单位等来建立数据库，其目的是方便临床数据的交换和重复利用[5]。CDISC 建议的命名更系统，但是其自明性略差，一般是电子数据采集系统在采用。

2. **数据库软件：**软件是数据的载体，同时起着对数据进行管理、查询、甚至统计分析等作用。根据其特点分为三类：1）本地平台：Microsoft EXCEL 和 ACCESS 等，其优点是可在本地运行，容易上手，定制程度高；2）云数据平台：临床研究平台 2.0（MedSci）、中国临床试验注册中心 (ResMan) 等，该类型对临床数据的存储和采集有不同程度的优化，通过网路进行数据存储，数据安全性更高；3）商用 EDC：linklab，易侑 EDC 等，其特点是收费，针对临床数据的存储和采集有相当程度的优化，有专业团队进行指导数据库的建设。

3. **数据表格的标准化：**计算机数据库包含一个或多个数据表格，其中“行（Row）”对应个体记录，“列（Column）”对应变量。标准化的数据库是“多表格关系型数据库”[6]，其特点为受试者特征、既往病史、伴随用药、实验室检查结果等为单独的表，受试者具有唯一的研究对象识别编号进行标识，可以通过“查询”功能进行不同表格之间的数据连接。

总之，建立临床数据库的准备工作，一是形成一个临床研究病例报告表，二是形成一个可靠、易用的数据载体，这需要多种背景的人员共同参与。值得一提的是，数据库的建立不是一蹴而就的。首先建立的是“采集型数据库”，其目的是为了全面记录试验的信息，其特点是采用“纵向数据结构”，多用文字描述来记录信息，没有进行“数字化”，不能直接进行统计分析；接着建立的是“分析型数据库”，其目的是为了统计分析，其特点是数据结构有“纵向数据”和“水平数据”两种，其中变量直接或衍生自“采集型数据库”且经过“数字化”而归纳成“二分类变量”，“多分类变量”，“有序分类变量”，“连续变量”等可分析的形式。

参考文献

[1] 赵一鸣, 曾琳, 李楠, 等. 临床注册研究可持续发展的科学基础:多目标多任务研究方案. 中华医学杂志 2013;093(046):3649–3651.
[2] 邱瑞瑾, 李敏, 韩松洁, 等. Interpretation of the COMET handbook (version 1.0) and its insight for developing core outcome sets in clinical trials of traditional Chinese medicine. 中国循证医学杂志 2017;017(012)1482–1488.
[3] 王雅倩, 杨悦. CDISC标准与临床试验数据标准化. 中国医药指南, 2016;14(12):296–297.
[4] Stephen B. Hully, Steven R. Cummings, Warren S. Browner 等. 临床研究设计（第4版）. 北京大学医学出版社, 2017.
[5] 林玲.中医临床护理信息数据元标准体系构建. 湖北中医药大学, 2014.
[6] 张永亮, 侯俊. 关系型数据库的规范化方法研究. 通化师范学院学报 2013;034(006):31–32.