零成本建立中小型临床数据库

刘岳鹏

徐州市中心医院

临床学术界有句话—“要想“富”，先建库”，是说要想多出研究成果，首先就要有一个有质量的临床数据库作为基础。我们往往羡慕别人有优质的数据库可以利用，有时也欣喜国外的哪个数据库又对外开放了，其实，通过一定努力和学习，我们也可以建立自己的数据库，并且随着数据的不断积累和建库经验的不断丰富，我们的数据库也可以达到较高的质量。

建立数据的紧迫性和必要性

我们处在循证医学时代，医学经验的形成和传播，都需要统计数据的支持；我们又处在一个大数据时代，研究采用的数据量越大，研究结果越精确，研究结论越可信，研究成果用途越广泛。与一些使用 Excel 收集的、数据量较小的、用完即弃的临床研究数据不同，用数据库存储的数据具有延续性强和数据量大的优点。这带来了两点好处，首先，数据的延续性可以为持续性研究提供支持。一时一地的研究往往具有局限性，结论可能是有偏差的，这需要持续不断的研究来弥补，这样我们才能了解疾病的真相。毕竟我们做研究，不仅仅是为了发表论文，更是为了总结和传播真实的临床经验。其次，数据库大量的数据，为获得更可靠的研究结果和更广泛的研究用途奠定了基础。临床研究根据数据量的大小分为三个层次或阶段：第一层次：假设检验，包括参数检验和非参数检验；指标：*P* 值；能力：确认两个指标之间有没有关系；数据量：小；第二层次：构建统计模型，比如线性回归、逻辑回归和 Cox 回归；指标：OR、RR、HR 及其置信区间以及 *P* 值；能力：在说明两个指标之间是否有关系的同时，说明是什么样的关系，比如直线关系，或者与指标的二次方之间呈直线关系等等；数据量：中等；第三层次，人工智能阶段，比如，机器学习和深度学习；指标：算法评分；能力：预测未来发生的医学事件；数据量：大。

正是认识到数据库在医学创新中的重要性，众多的国家级的医学数据库正在建设中。目前，我们国家已经建成了“国家人口健康科学数据中心（https://www.ncmi.cn/）”、“公共卫生科学数据中心（http://www.phsciencedata.cn/）”、“中医药学科学数据中心(http://dbcenter.cintcm.com/)”、“药学数据中心（http://www.pharmdata.ac.cn/）”和“临床医学科学数据中心（http://101.201.55.39/#/）”等，另外，国内某些单位，比如，南京医科大学牵头各附属医院正在建立“专病队列”。与此同时，以机器学习和深度学习为代表的大数据分析的方法在医学领域的普及也为大型医学数据的利用提供了可能[1]。可以预见的是，几年以后，国家和省市重要的医学课题和成果普遍都会建立在大、中型的数据库研究分析的基础上；重大医学创新的背后也一定有大型数据库的支持。

从临床科研的各个关键环节看，无论试验设计、

DOI: 10.14218/MRP.2021.009
通讯作者：刘岳鹏 Email: liuyp2080@163.com

数据收集、数据分析，数据发表，都是围绕数据展开的，而贯穿各个环节的一个基础的工具就是临床数据库。目前临床上使用的数据库主要来自商业化的数据库，普遍价格不菲，使一些科室和个人望而却步，我们在实践中通过摸索形成了零成本地建立一套相对成熟的、标准化地临床数据库的步骤，在这里和大家分享。

数据库的特点介绍

我们构建的数据库可以满足以下需求：方便的数据的录入、处理、输出到统计软件，具有一定的安全性。然而与成熟的商业数据库相比，也存在一定的不足，比如不支持移动设备录入等。

数据库建立的核心内容，叙述如下：

1. 采用 MySQL 这种关系型数据库作为数据库的基础，关系性数据库是目前大型数据存储流行的形式，各种商业数据库都普遍采用。搭配 phpMyAdmin 可视化界面（有条件的单位或个人可以采用收费的 Navicat），并借助数据库操作语言 SQL（推荐掌握），可以方便地对数据库中多张表进行查询、增加、修改和删除等操作，以满足管理数据的需求。MySQL 和 phpMyAdmin 的搭配在安全性方面，提供了用户管理系统，数据备份等，并且 MySQL 可以设置进行本地访问，也可设置进行远程访问。如果是个人的数据库，可以在个人电脑上进行安装，默认本地访问即可；如果是科室的数据库，可以有一台电脑作为服务器，其它的电脑通过用户名和密码进行远程访问。MySQL 可以存储中等数据规模，较 Excel 的数据存储能力大大提高，完全满足目前数据的需求。

2. 在形式上，如果从零开始建立关系型数据库，要考虑到数据库设计，以便使用数据的时候更方便且不会出现矛盾冲突。在这里为了减少数据库建设的难度，直接借鉴临床数据交换协会 (CDISC) 制定的一整套的临床数据收集规范 [2]，该规范将临床资料分成若干个领域（比如，临床干预、实验室检查、病理等），并制定了统一的变量的名称、关键字的名称、变量的种类（标识变量、日期变量、标签变量）等，通过遵循这些规范我们就可以从形式上保证我们的数据库是符合关系型数据库设计规范的，并且将来收集的变量的种类是全面的。

3. 在内容上，虽说要收集哪些变量是由研究者根据试验的内容所决定的，但是也有一定的原则可循，即结局指标核心数据集（COS）。结局指标是反映临床结局的一系列指标，一个研究往往会用多个结局指标从不同方面进行观察。用多个指标就可能造成研究成本增加，而用一个指标又可能不能充分说明问题。COS 的概念由世界卫生组织最初在肿瘤研究领域提出，目前由 COMET（core outcome measures in effectiveness trials）工作组运作并制定相关标准 [3]。COS 可以减少同类临床研究由于不同结局指标选择导致异质性而无法纳入系统评价的情况，同时也能更容易识别出临床研究中潜在的选择性报告偏倚。根据 COMET 组织发布的指南 [4]，各个研究领域具体的 COS 通过专家共识和文献综述相结合的方式来确定，已经建立了 COS 数据库，更多的 COS 研究正在不断地纳入到该数据库中。国内学术界，特别是中医药领域，对其应用有广泛地探讨。

数据库建设相关的想法

1. 临床数据库不是一蹴而就的。我们建议分两步走，首先建立“采集性数据库”，这种数据库的目的在于全面收集数据，采用纵向数据结构，可以方便地录入数据。这种类型的数据库多是文字描述，变量没有被整理成分析时可用的“二分类”、“多分类”、“连续变量”等数据，会有缺失值、重复值、错误值等情况，也有的数据在数据分析时可能会用不到，例如一些注释的信息。“采集型数据库”可以根据不同的分析目的而选取一部分数据进行“清洗”（去重、处理缺失值、格式化）等操作形成多个分析目的不同的“分析型数据库”，用于数据分析。分析型数据库多采用横向数据结构，数据经过清洗

和整理，可直接用于分析。

2. 建立数据库要花很长时间才能看到成果? 不完全是。临床观察性研究类型分回顾性研究和前瞻性研究两种主要形式。如果采用回顾性研究的形式，收集临床上已经存在但没有被收集利用的信息来在进行研究，短时间内就会产生研究成果，可以为将来的研究提供研究线索和方向。

总之，认识到医学数据库建设对于医院诊疗经验总结和研究的必要性和迫切性，我们在实践中摸索出一套零成本建立中小型数据库的方法。这套方法在形式和内容结合了目前流行的 CDISC 和 COS 标准，采用 MySQL 交互式数据库作为基础，是适合大多数临床科室的一种临床数据库建设解决办法，同时也期望临床医生因此积累建库的经验，为将来建立更大型的数据库做准备。

参考文献

[1] 谭向龙, 赵之明.机器学习在医学中的应用现状中华腔镜外科杂志(电子版) 2020;13(01):61–64.

[2] 李庚, 李晓彦, 温泽淮. 国际临床数据交换标准协会标准在电子数据采集系统中的应用研究. 世界科学技术-中医药现代化 2017;19(2):338.

[3] 邱瑞瑾, 李敏, 韩松洁, 何天麦, 黄涯, 陈静, 商洪才. 《COMET手册》1.0版解读及其对构建中医临床研究核心指标集的启示. 中国循证医学杂志 2017;17(12):1482.

[4] 史纪元, 高亚, 马新萍, 田金薇. COMET数据库核心指标集研究现状剖析. 中国医药导刊 2020;022(001):53–58.