临床医学数据库的建立，采用纵向还是横向数据结构？

——理论兼解读CDISC研究数据模型

刘岳鹏

徐州市中心医院

在建立临床医学数据库过程中，确定了数据库相关软件和需要收集的变量之后，如何有序地组织数目众多试验数据是紧接着要考虑的一个问题。目前的解决方案是采用关系型数据库，将整个试验数据分类并存储在多个二维表格中，这些二维表格之间通过共同且唯一的“ID”变量而发生关联，以便利表格间的数据查询整合。在一个二维表格中，一“列”代表一个测试（特征、变量），一“行”代表一个受试者（观察单位），而两者交叉点则代表该受试者某种测试值。这些二维表格根据数据的组织形式大概分为纵向数据结构和横向数据结构两大类。在横向数据结构中，所有变量横向展开，一列代表一个变量，一行代表一个受试者，因为这样的二维表格通常较宽，也称“宽表”；而在纵向数据结构中，多个变量整合成一类，统一在一个变量名下（或可称作“类变量”），因为横向分布的变量的数目减少并且相同受试者的不同测试会录入在一个变量下而使表格纵向延展，这样的二维表格会比较长，也称“长表”。

纵向和横向数据结构各具优缺点，采用何种数据结构主要考虑是否便利于后续的操作——数据的采集还是数据的分析。在数据采集阶段，数据不断被添加到数据库中。纵向数据结构更适合进行数据的录入，只需要在表格的末尾进行数据的添加即可，相反，在横向数据结构中则需要定位相关的行和列；另外，如果有新的测试名目需要添加，纵向数据结构只需要在相关“类变量”下进行录入即可，而横向数据结构则需要添加一个新的列而涉及了更多的操作。所以，在数据采集阶段采用纵向数据结构更加方便。而对于另外一个主要的操作——数据分析，纵向数据结构对其支持有限，相反，横向数据结构对数据分析的支持是全面的，尤其是当我们需要分

DOI: 10.14218/MRP.2021.004
通讯作者：刘岳鹏 Email: liuyp2080@163.com

析两个变量之间的相关关系，或者进行大数据相关的分析时，只有横向结构的数据能够满足要求。这时，横向数据结构是最佳的数据组织形式。

说到医学数据的组织，不得不提 CDISC（临床数据交换标准协会）制定的数据模型，这个在医学研究数据组织方面被广泛采用的一个规范。在 CDISC 数据模型的整个设计中，首先将数据模型的类型分为采集型（研究数据列表模型，SDTM）和分析型（分析数据模型，ADaM）。采集型数据模型建立的目的是为了全面且便利地采集相关数据，数据组织形式是采用纵向和横向数据结构相结合的形式：首先，数据被分为多个域（大类），每个域存储在一个二维表格中，比如在 CDISC SDTM 3.1.2 版本中，所有的临床试验数据分为若干大类，例如，心电图检查、既往伴随用药、暴露、生命体征、嗜好品使用、不良事件、受试者访视、问卷、病史、体格检查、实验室检查、一般情况、人口学特征等。 每个域（二维表格）中有且只有一个主题变量，而其它的变量类型，识别变量、时间变量、修饰变量等都为围绕主题变量而存在，这里主题变量的组织就是采用纵向数据结构，多个种类的测试都归到一个主题变量之下，其它的变量的组织是采用横向数据结构；而分析型数据模型的数据组织形式推荐为纯粹的横向数据结构，其目的是为了后续的数据分析提供支持。实际工作中，分析型数据模型的数据库是由采集型数据库衍生出来，一个采集型数据库可以衍生出多个不同分析目的的分析型数据库。

总之，纵向数据结构更灵活，便于录入数据和增加新的测试名目；而横向数据结构更便于分析，两者可以结合使用，从而避免表格过长也避免表格过宽，不失灵活而又便于分析。在了解了纵向和横向数据机构的特点，特别是有 CDISC 的研究数据模型作为借鉴，相信大家更加能够有序地组织自己的临床研究数据，为后续的数据采集和数据分析奠定基础。