临床医学研究数据库的创建和质量控制要点

魏 然

湖北科技学院 湖北省咸宁市 437000

【摘 要】建立高质量的标准化数据库是临床医学研究中的一个关键步骤，它是今后数据的统计和分析、结果的展示，以及以后的高质量的学术论文的出版平台。本文概述了临床医学研究中常见的数据库类型（Excel 数据库、EpiData 数据库、SPSS 数据库、EDC 数据库）以及建立数据库的方法，并着重阐述了数据库变量设置、输入规则、数据质量控制等方面的知识，为临床科研工作者提供了科学的依据。

【关键词】临床研究；数据库创建；质量控制；数据录入

1 临床研究数据库的分类与特点

目前临床上使用最多的数据库有 Excel 数据库、EpiData 数据库、EpiInfo 数据库、SPSS 数据库、EDC 数据库等。

（1）使用 Office 办公软件建立的数据库。Excel 数据库易于上手，易于创建，数据录入方便，适合小型临床试验。

（2）利用统计分析软件建立的数据库，包含 EpiInfo 和 SPSS 数据库. 要想建立这样的资料库，必须具备运用统计分析软体及资料库结构与设定的基本知识。如果是在使用 SPSS 软件建立数据库时，用户需要在 “Variable View（变量窗口）中设置一个变量名称，在数据视窗中输入数据。

（3）以 EpiData 数据库为代表，使用专门用于临床研究的数据库软件建立的数据库。EpiData 软件具有创建数据库、设置质量控制、输入数据、合并数据库、一致性检验、数据输出等功能，因其功能强大、使用方便、开放资源自由等优点，逐渐得到了临床和流行病学的广泛应用。

（4）一个由 EDC 数据库代表的电子数据采集系统所建立的数据库。该系统利用国际临床数据交换标准（CDISC）的标准建立了一个数据库，可以将 EDC 和医院的医院信息系统、实验室信息管理系统连接起来，可以方便地获得临床研究需要的数据。

2 临床医学研究数据库系统的发展演变

这些类型的临床医学研究数据库都具有各自的特色和重点，但是在系统的功能方面存在着一些共同的要求，如：数据的输入、采集、质量控制、电子病历的提取。随着医学信息技术的不断发展，电子病案的应用越来越广泛，现实数据的研究也随之兴起，医学科学的数据库体系也随之发生了变化。

2.1 临床医学研究数据库系统的平台化

由于临床研究的类型、目的以及科研问题的差异，各研究课题或病种数据库所需要的数据内容、录入表单、校验规则、存储结构等各方面的差异，使得临床研究数据与数据管理系统的个性化开发模式存在工作量大、变更维护不便等问题。平台化是目前临床研究数据库的一个发展方向，它可以为临床研究库的通用平台提供通用的平台，用户可以根据自己的研究项目要求，自定义数据的内容和功能，从而达到通用化和个性化的结合。这些平台具有以下几个主要的作用：

（1）数据项定义。数据项目的格式和标准规格由研究人员自行确定。

（2）病例表单定义。案例表格中的数据项、表现形式和数据检验规则都是由研究人员和工程师来决定的。根据回顾性的研究或前瞻研究的需求，将资料输入表格。在数据来自于电子医疗记录的情况下，可以对数据进行自动提取和映射。

（3）病例录入。根据病例表格的定义，为用户提供了数据的输入接口，并对其进行了验证和质量的审查。

（4）病例检索。为科研工作者在数据库中的病例数据提供了一种灵活的查询方法。

（5）病例数据统计。为研究人员提供了基础数据，并对各变量进行了相关分析。

近年来在国际上广泛应用的 REDCap 临床数据收集系统和国内部分医院新开发的临床研究数据库都是这种平台。

2.2 体系结构的演变

在早期，临床研究的数据库系统大多是比较独立的，并不能与医院的信息系统进行整合，这时的系统架构如下（图 1A），还有很多科室的数据库系统。随着电子病历的广泛使用，临床医学领域的研究日益增多，对电子病历进行提取和产生科学数据的需求也日益增加，因此，临床研究数据库必须与电子病历数据进行整合，此时的体系结构如下（图 1b）所示，仅有少数研究数据库已具备电子病历数据采集功能。

但是，各临床研究数据库的数据内容不同，其数据结构也不尽相同，由病例表格与不同类型的病历数据进行交互，造成界面开发工作量大，不同专业数据库之间的界面可复用性较差，这就使得临床研究数据库系统在体系结构上进一步增加了病例原始数据库构件。

病案原始资料库与不同类型的病案资料进行对应，提取出完整的病案资料。在建立病历表格时，可以更加方便地从基础数据库中提取和处理数据。

在这种架构下，系统需要将患者的原始数据模型和外部的数据模型进行一次映射，将专家数据库与患者的数据进行关联，从而简化了数据库的开发和定制。

该系统允许医师获取个别病人的资料，而不仅仅是检查病人的一般特性。CDR 的资料主要有：临床化验结果、病人人口统计、药物信息、放射报告和影像、病理报告、入院&出院日期、出院小结、记录等。

医务工作者不仅可以把所搜集的资料应用于病人的治疗，而且可以应用于特定的手术、部门或治疗计划的管理。对于研究者来说，这是一种非常有价值的数据，能够为他们提供更多的临床信息。由于一些法律上的理由，获得此类资料有一定的限制，因为不适当的利用会给病人的资料带来泄露的危险。

因此，建立一个临床研究数据库，必须要有一个强有力的电子医疗记录系统，把所有的历史资料都扫描或输入到该系统。通过实验室检测、医学影像研究、医学检测等过程中所获得的新的信息，将被记录在电子文件中，最后将其存入到临床研究数据库中，从而可以及时更新和提供更多的信息。医务工作者也可以在病人许可的情况下与别人分享病历。

2.3 临床研究数据库的应用

2.3.1 患者护理

医生能够看到病人的全部病历，并且在任何时候都能发现异常。并且可以轻松地从各个位置调取记录。首先是搜寻，其次是分析。这不但有助于医师的诊断，而且对治疗慢性或后期的病症也有一定的帮助。同时，对数据进行集中管理还能降低出现差错的几率。

2.3.2 疾病研究

临床医学研究数据库能够为患者、病人的医疗状况和疗效提供大量

的信息。以往的流行病学研究也许不如临床资料丰富，从而使资料或结果缺乏说服力。因此，对于科研人员来说，这是一笔非常珍贵的资源。研究者们可以通过对比病人的健康档案来评价不同治疗方法的疗效，这对长期从事多种人群的研究很有帮助。比如，研究人员可能会对护理方面的差别感兴趣，从而能够通过现有的临床资料来判断谁得到了照顾，护理的质量和效果。

2.3.3 医院管理

通过临床资料仓库中的资料，医院管理人员和其它相关人员可以发现相关的问题，了解各种方法的疗效，并讨论其它相关的问题。这将有助于管理人员制订新的医疗计划和医疗标准。通过统计和监控资料，提高病人的医疗质量。

2.3.4 监测疾病及药物使用

比如，医院里的抗生素监控，还有传染病的监控。

3 临床研究数据库的创建方法

临床试验数据库的建立主要分为 3 个阶段：变量名设置、变量属性定义、变量之间逻辑关联设置。四种类型的临床试验数据库的内部结构、特点和核心需求不同，其建立的方式和操作程序也不同。

3.1 Excel 数据库

本研究以调查问卷为基础，建立 Excel 资料库，假定调查问卷的主要内容有：①调查对象的年龄、性别、学历、调查日期；②吸烟，包括吸烟、每日吸烟量、吸烟年数；③是否患有肿瘤、高血压、糖尿病等疾病。首先，研究人员按照问卷的内容设定变量名称。变数名称可以用英文的名字或者英文的字母加上数字来命名。比如，可以把一个变量的名字叫做“age”或者“A1”，可以把它的英文翻译成“age”。如果把变数名称设为“A1”，需要在以后的数据输入中作注释。在 Excel 数据库的第 1 行中列出了所有已设定的变量。其次，对每一个变数进行定义。例如，年龄、每日吸烟量、吸烟年数等量化变数，必须对这些变数进行明确标注；性别、文化程度、是否吸烟、是否患有肿瘤、高血压、糖尿病等定性变项，应分别标注不同的变量。最后，将变量之间的逻辑关联统一起来。比如，如果“smoking”是“no”，则应该自动忽略后面的“number_smoke”以及“year_smoke”。三个步骤都做完了，Excel 数据库就可以可以输入数据了。

3.2 SPSS 数据库

本文以以上问卷的主要内容为例，运用 SPSS16.0 开发 SPSS 数据库，按以下几个步骤进行：第一，按照问卷的内容设定变量名称。需要在 SPSS“Variable View”中进行变量名称的设定（图 2 的上半部分），可以直接用英文或英文字母加上数字来命名变量名。比如，在性别方面，可以把“gender”或者“A2”作为一个变量，而建议将“gender”作为英文的翻译。如果把变数名称设为“A2”，需要在以后的数据输入中作注释。其次，定义了年龄、每日吸烟量等量化变量，如“Variable View”中“Label”，定义了性别、教育程度、是否吸烟、是否患有癌症等定性变量，如“Variable View”中“Values”。最后，将变量之间的逻辑关联统一起来。完成上述 3 个步骤，就可以建立 SPSS 数据库。在 SPSS“Data View”中需要输入数据。

3.3 EpiData 数据库

就上面所说的问题来说，利用 EpiData 软件进行 EpiData 数据库的制作过程如下：第一，在 EpiData 软件中，可以直接用英文翻译名或者英文字母加上数字来命名变量名。比如，你可以把一个变量的名字设定成“age”或者“A1”。由于 EpiData 数据库在输入数据时会出现一些问题，因此建议将变量名称以英文和数字来命名。但是，在使用这个名字的时候，必须要用空白来分隔变量名和问题提示，比如 A1 的年龄 15 岁。其次，将资料输入格式设定为变量的类型。“#”是数字类型的变量，1 个“#”是一个；“_”表示字符类型，“_”为一个汉字；使用“yyyy/mm/dd”或者“mm/dd/yyyy”来表示日期类型的变量。在 QES 文件存盘上设定好了变量名称和输入格式，并按照 QES 文件产生 REC 文件，存储磁盘，此时就能进行数据输入。为了确保输入数据的准确度和效率，需要为数据库创建 CHK 档案，并设定数据录入质量的有关要求，主要有“Range/Legal”、“Jump”（Jump）、“Must enter”（必需输入）和“重复”四个部分。

3.4 EDC 数据库

与以上三种类型的建立方式相似，EDC 的建立也包含了变量名称、属性定义以及变量之间的逻辑关联。例如，上海申康医疗发展中心所建立的 EDC 数据库“CRIP 数据库”。为了便于多中心临床试验数据的拼接以及随后的合并分析，变量名称必须具备国际通用性。其次，与创建 EpiData 数据库时的 CHK 文件一样，将数据输入到“DEV”的各个变量中，并进行了仿真试验。最终，通过仿真验证，将数据库锁定，并将其发送到 PROD 上，然后开始正式运行。EDC 的建立要求有专门的技术和知识，通常都是由专业的企业来完成。

4 临床研究数据库的质量控制要点

为了确保建立的临床研究数据库的质量，研究者必须根据一定的原则和要求建立数据库，规范数据的输入和数据的质量。

首先，在小型临床试验中，EpiData 数据库作为首选，建立 CHK 档案以保证数据输入的准确度和效率，同时进行两次输入数据的一致性检验，以保证数据的正确性。在数据量少的情况下，可以采用 Excel 数据库或者 SPSS 数据库来缩短系统的建库时间。需要注意的是，由于 Excel 和 SPSS 数据库在输入数据时没有进行逻辑校对和品质的监控，所以在输入数据时一定要认真仔细，避免出现错误。在规模大，变量多，尤其是多中心的临床试验时，如果资金允许，推荐专业机构进行 EDC 数据库的开发。

其次，不同的临床试验数据库需要输入阿拉伯数字，而不能输入汉字（如性别、性别、是、否等），否则就不能进行数据的统计分析。在输入资料之前，研究者必须对问卷内容进行全面审查，以确保问卷的内容清楚、逻辑性。在数据量大的临床试验中，为了确保输入的准确性和一致性，可以在数据输入之前对输入人员进行统一的训练。

第三，对已经录入数据的数据库，在进行数据统计分析之前，必须进行数据质量控制。数据质量管理的重点是：①资料的完整性；研究人员应该对数据库中的变量数据进行彻底的核实，尽量不遗漏主要结局变量。如果主要终点和核心变量的数据有缺失，则可以通过数据填补（均值填补、k 近邻填补、回归填补、随机森林填补、多重插补、热卡填补）等方法来填补。②资料的逻辑。主要检查资料之间的逻辑关系，如身高、体质量、年龄等一般人口特征，有无不合逻辑的异常，时间资料之间的逻辑性等。

最后，为了确保数据的真实，需要从数据库中提取出一份（通常为 5%~10%）的调查表，并将其与数据库中的数据进行对比，以评估整个数据库的输入质量。若查对结果显示，输入准确率小于 80%，则视为资料库资料输入品质不佳，一般推荐重新输入资料。只有当以上各项工作都已完成，资料输入质量评估为优良时，数据库将被锁定，并根据已锁定的数据库进行数据统计分析。

结论：

本文综述了目前临床医学研究中常用的几种数据库类型和特征，并对其从孤立的、面向平台的发展和技术结构的演进进行了综述，分析了临床医学研究数据库的创建和质量控制要点，希望可以为相关工作人员提供一定的帮助。

参考文献：

[1]王瑞平，李斌.临床医学研究数据分类浅谈[J].上海医药，2022，43（1）：3-6.

[2]王瑞平，李斌.临床医学研究数据统计分析思路概述[J].上海医药，2022，43（1）：7-9.

[3]王瑞平.临床研究规范设计 PICO 原则[J].上海医药，2022，43（3）：67-72.