

证候数据库录入的方法学探讨

张志斌¹, 王永炎¹, 张启明², 王 耘³, 张庆祥²

(1. 中国中医科学院, 北京 100700; 2. 山东中医药大学, 山东 济南 250011; 3. 北京中医药大学, 北京 100029)

关键词: 证候研究; 数据库; 方法学

中图分类号: R2-05 文献标识码: C 文章编号: 1005-5304(2006)09-0106-02

古今医案及现代证候规范研究成果数据库的建立是“证候的规范及其与疾病、方剂相关的基础研究”中“证候规范与辨证方法体系的研究”课题的重要的基础性工作。在此项研究中, 将对相关资料中涉及到的四诊信息与现代理化信息, 还有来自社会和自然对证候有影响的信息全面采集, 高度开放。对于获取的证候信息依其已知可用的和未知待用的, 贡献度大的与贡献度不大或今天不大、今后可能有鉴别厘定意义的分层次加以采纳运用。数据库的可信程度如何直接关系到研究的结论。因此, 证候数据库录入的方法是一个值得充分探讨的问题。

在数据库录入方法的设计时应着重注意两点: 其一保证数据的可靠性; 其二, 便于进行质量控制。为了保证数据录入的可靠性, 我们对所有的资料都进行三重处理, 在此基础上, 建立质量控制机制。

1 建立文本数据库

选择合适可信的版本, 将资料进行扫描及分辨处理, 然后进行仔细的校对, 以保证文本数据库资料与原文本完全一致, 以备结构录入与核查之用。

2 结构数据库

2.1 编制课题专用录入软件

按照课题研究的需要, 以 Visual FoxPro 为开发平台, 编制课题专用录入软件使得证候资料所涉及的病名、症状和体征、病因病机、病位、辨证、用药都成为取值为 0 和 1 的二值变量(dichotomous variable)。

2.2 进行录入用语规范

根据录入软件所涉及的内容, 研究拟定可能需要的录入用语。由于在不同的医案著作及不同的证候规范研究成果中, 可能对同一事物的表述不一致, 比较突出的问题是症状的表述。本课题首先采用最新由科技部基础性项目“中医药学名词术语规范研究”组完成, 经全国名词委中医药名词审定委员会审定, 2003 年 12 月通过科技部验收的《中医药学名词》对症状表述进行规范。《中医药学名词》尚未涉及的症状名词采用其他术语规范成果进行对照规范及专家咨询等方法研究论定。在上述研究的基础上, 制定出数据库录入密钥, 这是每一个参与录入的人员务必要十分熟悉的内容。

2.3 制定数据录入规则、培训医案录入人员

将数据录入的操作步骤、内容及注意事项制定出录入规则, 并结合既往的经验, 将数据录入过程中可能出现的错误及避免错误的方法编辑成《数据库录入工作手册》, 下发到每一个录

入员手中。这个录入手册还需包括几个相应的附件, 如《结构数据库录入密钥》(包括录入用语)、《录入名词注释》(包括重要名词的定义)等等。例如, “腹胀”通常是指患者有腹部胀满的自我感觉, 而“单腹胀”则是指腹部膨隆胀满而躯体四肢皆消瘦的表现, 二者不可混为一谈。所以, 在每一位录入人员参与录入工作之前, 必须进行培训。

2.4 录入资料的信息分类

古今医案及现代证候规范研究成果的证候相关资料记载的内容一般包括病名、症状(或体征)、病因病机、病位、辨证、用药 6 个部分。一般来说, 如果某一资料的这 6 个部分中某些部分缺如, 软件将把缺如的部分的所有变量赋值为零。这时, 可通过在 SAS 软件中编程, 使得所有变量都被赋值为零的部分数据缺如。

2.4.1 关于病名 虽然本课题研究的最终目标——建立辨证方法新体系在设计中并没有涉及病名, 但本项目的研究目标却是非常强调病证相关的。所以, 结构数据库的病名项也是我们充分注意的内容。一般来说, 在结构数据库中, 我们使用经过规范的病名, 如《中医药学名词》、《中华人民共和国国家标准·中医病证分类与代码》、《中华人民共和国国家标准·中医证候名称与分类代码》、《中华人民共和国中医药行业标准·中医病证诊断疗效标准》、《中华人民共和国国家标准·中医临床诊疗术语疾病部分》及中医高等院校教材等。如遇到古医案中无法进行规范归位的原始病名, 则按原样录入。

2.4.2 关于症状与体征 这是本项研究中十分重要的内容。本课题以证候规范为研究的总体目标, 辨证方法新体系必须建立在规范语言的基础上。因此, 结构数据库中的症状(及体征)表述将尽可能使用规范用语。如“半身不遂、半身瘫痪、半身肢体痿废不用”等表示的其实是一样的意思, 都是指“左侧或右侧肢体不能随意运动的表现”, 所以均用“半身不遂”来表述。“呕恶”实际上是“恶心”(感觉胃中有物上拱, 急迫欲吐的表现)与“呕吐”(胃内容物, 甚至胆汁、肠液通过食道反流到口腔, 并吐出的反射性动作)两个名词的缩写, 输入时将仍然用“恶心”与“呕吐”来表述。但是, 鉴于中医诊断特色, 在相似症状的描述中, 有时存在着程度的差别, 这种程度的差别可能具有中医诊断学意义。比如, “完谷不化”与“食谷不化”都是指“粪便中夹有未消化食物的腹泻表现”的意思, 但“完谷不化”可能是指吃什么泻什么, 即完全没有被消化; 而“食谷不化”则可能是指粪便中夹有不消化食物, 即消化不良。对于中医诊断来说, 这两种情况所反映的脾胃虚弱的程度差别则具有意义。因此, 即便是进行症状名词规范, 也应该保留这种

基金项目: 国家重点基础研究发展计划(973)2003CB 517101

不能被忽略不计的差别。在录入工作开展之前,课题组需要对症状规范表述做较为细致的研究工作。不仅要规范症状名称,还要对各个症状名称下一个尽可能准确的定义。当然,这项工作在此前也已经有了较好的基础,有许多其他课题组已经完成的工作可以供我们参考使用。

2.4.3 关于病因病机与病位 本课题研究的目的是为了提取证候要素。证候要素的提取有两个原则:其一,同一层面的证候要素必须是同类概念;其二,证候要素必须是不可分解的最低单元,即单要素。证候研究结构数据库是为提取证候要素而建立,根据课题研究分为“病机层面”与“病位层面”的二元设计,在数据录入的时候要对复合病因病机、复合病位均将进行适当的尽可能准确的分解。如“肾阴虚”在录入过程中,将被分解为病位“肾”和病机“阴虚”;“脾肾阳虚”在录入过程中,将被分解为病位“脾”、“肾”和病机“阳虚”。当然,复合病因及病位的分解是有条件的,即被分解的证候要素还能够合成原来的复合状态。在复合病因或复合证候表述很不规范的特殊情况下,也要尽可能多地保留原有的信息,使得将被分割的证候要素组合时,能更接近原本的复合状态。分解时,应参照课题前期工作中专家建议的29个证候要素^[1],但不受此限制,如果不符合这29个证候要素,则按原始文献提到的证候要素进行分解。

2.4.4 关于辨证 这一部分的处理对于数据库完成之后的数据分析是至关重要的。鉴于本课题研究的重点在于对证候规范提出了与此前的证候规范研究工作不同的思路,为了尽可能地尊重原始文献,以保证被分析的数据是真实的、可靠的,这一部分采用的方法是将原始文献中的证候名称忠实录入,以供分析证候要素应证及其组合规律时所需。如原文献缺乏证候名称的诊断,则保持数据的缺如状态。

2.4.5 关于用药 本课题的研究内容是“证候规范与辨证方法体系的研究”,原本不涉及用药的问题。只是为了今后的进一步研究,在古今医案数据库中,也注意到了用药的录入。在结构数据中,使用规范药名,即2005版《中国药典》所用的中药名,将别名、地方名等其他用药名称均统一到规范药名。

2.5 资料中潜在信息的发掘

由于中医学的人文含量比较大,医学著作中也会使用较多的修饰方法,使原本要表述的意思显得较为含蓄而且间接。因此,有些资料中的相关信息可能需要进行从潜在到显现的转化,如“水亏”一般可以按“肾阴虚”来处理。另外,根据“益气养阴”的治则,可以推测出“气虚”与“阴虚”的病机。潜在信息的发掘尤其要注意的是,推出的结论必须是唯一而准确的,不能随意推测。如原文献中说病位在“中”,不可推断为病位在“中焦”,因为也有可能是说病位在“里”而已。遇到此类情况,只能舍弃病位的推断,保留资料的缺如状态。

3 建立用语对照附件

由于原始文献在录入过程中被人为地进行了以上所讨论的用语规范及要素分解的处理,不同的录入人员可能有不同的具体掌握标准而做出不同的录入处理。为了便于进行质量控制与审核检查,我们采取的补助措施是:凡是原始文献表述用语与录入语言不同,必须建立用语对照附件。附件格式见表1。

表1 《XXXX》(原始文件名)录入用语对照附件

录入号	对照内容	原始文献同语	录入用语	备注
XXXXXXX	病因病机	痰湿内停	①内湿,②痰	
	症状	食谷减少,纳呆呕恶	①恶心,②呕吐	

4 数据录入的质量控制

4.1 选择合适的著作与版本

为了保证资料的可靠性,著作与版本的选择是十分重要的。因为本次研究强调的是海量数据,所以只要相关的著作将做尽可能多的选择。当然,就古今医案来说,著作实在是太多,想要竭泽而渔是不可能的。著作的选择上将在时代、地区、科别(以内、妇、儿为主)方面做一些平衡,强调有代表性的著作不能空缺。现代规范研究成果相关著作,则将做尽可能多的搜罗,力求将有一定影响的著作均做收入。版本问题上,首先要强调的是选择第一手资料,不从第二手资料中引用。第二,同一种著作如果有不同的版本,要在出版者与印刷质量方面做一些比较,并请相关专家提出好的建议。

4.2 确立录入号

凡是古今医案著作中的经过按一定标准选择的一个医案或证候规范研究成果中的一个独立的证候均可视为一条资料。在录入任何一条资料的时候,进入录入软件的界面,首先要为这条资料确立一个录入号,据估计,整个数据库资料完全录入之后,会超过10万条以上,所以录入号可能以7位数较为合适。这个录入号确立有两项要求:①必须在包括各个小组分别完成的整个关联数据库中保证是独一无二的,即肯定不会与其他任何一条不同的资料的录入号重复;②在文本数据库、结构数据库与录入语言对照附件中,同一条资料的录入号相同,以确保每一条资料都可以随时被准确地调出进行相应的核查或修改。如在审查或使用过程中对录入的资料有疑问,可以同时调出文本数据资料、结构数据资料、录入语言对照附件来进行核实及必要的修正。这样,既方便于课题组对录入人员的工作质量进行抽查检测,也方便于审查专家对课题组的工作质量进行审核把关。

4.3 录入的资料必须保持相对完整的信息

每一条资料在录入之前,首先添加资料的详细来源。就古今医案来说,包括:医家姓名、著作名称、著作者、出版社、出版时间、页码等;就现代证候规范研究成果来说,包括:著作(完成)者、著作(成果)名称、出版(公布)者、出版(完成)时间、页码等。

4.4 录入质量的控制与审核

要求录入人员严格地尊重原始文献,不容许任何主观意志的介入,缺如的信息一定不能人为补充,保持其缺如状态,并在自己录入的每一条资料的备注中署名,以备核查。经过培训的录入人员初步录入的资料经由核查人员的两次随机审核合格后方被正式录用。进入工作状态之后,课题组仍安排质量把关人员随时抽查。

参考文献:

- [1] 张志斌,王永炎.证候名称及分类研究的回顾与假设的提出[J].北京中医药大学学报,2002,(2):1-5.

(收稿日期:2005-07-07,编辑:华强)