

WS

中华人民共和国卫生行业标准

WS/T 306—2023
代替 WS/T 306-2009

卫生健康信息数据集分类与编码规则

Rules for health information dataset classifying and coding

2023-08-07 发布

2024-02-01 实施

中华人民共和国国家卫生健康委员会 发布

前 言

本标准为您推荐性标准。

本标准代替WS/T 306—2009《卫生信息数据集分类与编码规则》。与WS/T 306—2009相比，主要为编辑性修改。

本标准由国家卫生健康标准委员会卫生健康信息标准专业委员会负责技术审查和技术咨询，由国家卫生健康委统计信息中心负责协调性和格式审查，由国家卫生健康委规划发展与信息化司负责业务管理、法规司负责统筹管理。

本标准起草单位：中国人民解放军总医院、国家卫生健康委统计信息中心、中国人民解放军空军军医大学。

本标准主要起草人：刘丽华、张黎黎、冯丹、李顺飞、王珊、王霞、胡建平。

卫生健康信息数据集分类与编码规则

1 范围

本标准规定了卫生健康信息数据集分类与编码需遵循的基本原则、技术方法以及应用规则。
本标准适用于卫生健康领域各类卫生健康信息数据集分类与编码方案的制定。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本标准必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本标准；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本标准。

- GB/T 2260 中华人民共和国行政区划代码
- GB/T 2659 世界各国和地区名称代码
- GB/T 7027 信息分类和编码的基本原则与方法
- GB/T 10113 分类与编码通用术语
- GB/T 11714 全国组织机构代码编制规则
- GB/T 13745 学科分类与代码
- GB/T 14396 疾病分类与代码

3 术语和定义

GB/T 10113中确立的及下列术语和定义适用于本标准。

3.1

数据集 dataset

具有一定主题，可以标识并可以被计算机化处理的数据集合。

3.2

数据集编码 dataset coding

在分类的基础上，给数据集赋予的有一定规律性、计算机容易识别与处理的符号。

3.3

数据集标引 dataset indexing

指分析数据集的内容属性及相关外表属性，并用特定语言表达分析出的属性或特征，从而赋予数据集检索标识的过程。

4 数据集概述

4.1 数据集基本概念

数据集是具有主题的、可标识的、能被计算机处理的数据集合。

- a) 主题:围绕着某一项特定任务或活动进行数据规划和设计时,对其内容进行的系统归纳和描述。通常数据集主题应具有划分性和层级性,划分性是指主题间可通过不同的命名,将相同属性的主题归并在一起形成相同的类,将不同属性的主题区分开形成不同的类;层级性是指主题可被划分成若干子主题或子子主题。

- b) 可标识:指能通过规范的名称和标识符等对数据集进行标记,以供识别。标识与名称的取值需要通过具体的命名或编码规则来规范。
- c) 能被计算机处理:指可以通过计算机技术(软硬件、网络),对数据集内容进行发布、交换、管理和查询应用。这些数据可以由不同的物理存储格式来实现,按照数据元的定义与数据类型,在计算机系统中以数值、日期、字符、图像等不同的类型表达。
- d) 数据集合:指由按照数据元所形成的若干数据记录所构成的集合。例如,病案首页数据集由主索引、出入转、诊疗、护理、手术、费用等不同数据组成。

4.2 数据集基本属性

每个数据集所具有的属性都包括通用属性和特有属性两部分。数据集的通用属性包括数据集主题、标识、实体和数据项,又称“基本属性”。数据集特有属性包括类别、区域、专业、学科、建立时间、涉及的疾病等。

4.3 卫生健康信息数据集

卫生健康信息数据集是在卫生健康领域,为满足政府卫生决策、业务处理、科学研究、信息发布与绩效评价等需求,按照数据集概念设计、归纳、整合的主题信息集合。卫生健康领域的数据集主要可以归纳为三个方面:

- a) 信息发布类统计数据集:如中国卫生健康统计年鉴中卫生机构设置及规模、卫生人员资源的地区分布、卫生经费的筹集及分配等数据集,各类卫生机构的统计月报、年报,以及满足某一专项统计需求通过统计收集、归纳、整理、报告形成的数据集。
- b) 业务系统建设类的基本数据集:包括医疗服务、公共卫生、药品供应、医疗保障、卫生管理等领域为了满足业务信息系统规范化建设和领域内部以及领域间数据交换与共享需求,设计归纳的各个子系统(或者功能模块)所包含的最小数据元素的集合。如:儿童出生登记、食品卫生许可、个人健康档案、住院病人出入转、居民死亡登记报告等基本数据集。
- c) 为满足特定目的收集整理制作的数据集:包括通过调查、观察、监测、检测、试验、实验等方式获取的满足科学研究、业务咨询或卫生服务决策等需求的数据集。如:近年来国家投入建设的国家人口健康科学数据集、卫生服务调查数据集、疾病及危险因素调查等内容的数据集。

5 数据集分类

5.1 分类基本原则

- a) 系统性原则:综合考虑数据集主题一致性,按其内在联系进行系统化排列,确保类目唯一、结构合理、层次清晰,减少冗余。
- b) 实用性原则:满足数据集分类编目的简便性、可操作和通用性需求和数据集查询的一致性理解。
- c) 可扩展性原则:保证分类体系框架适应数据集不断丰富的内容和日益增长的种类与数量。
- d) 兼顾科学性原则:自顶向下,优先选择最能代表卫生健康信息数据集主题的语言、词条定义类目名称,编制受控分类体系表。
- e) 稳定性原则:使用稳定的因素作为分类依据,同时提高分类体系的可延展性或兼容性,促进稳定性。

5.2 分类方法

5.2.1 主题优先法

卫生健康信息数据集分类体系首先依据主题构建主分类框架,分类框架各层级类目可表示为一级、二级、三级类目等,见表1。

表 1 卫生健康信息数据集主分类框架

一级类目	二级类目	三级类目	n 级类目	数据集系列
一级类目 1	二级类目 1	三级类目 1	n 级类目 1	数据集 1 数据集 2
			n 级类目 2
		
		三级类目 2	n 级类目 1 n 级类目 2
		三级类目 3	n 级类目 1 n 级类目 2
		n 级类目 1 n 级类目 2
	二级类目 2	三级类目 1	n 级类目 1

5.2.2 多属性复分法

卫生健康信息数据集以相同主题构建，同时在若干相同主题数据集中，又根据其个性，如时间、地域、疾病、学科、机构等多属性，构成辅助分类体系，即复分类体系。

卫生健康信息数据集复分类见图 1：

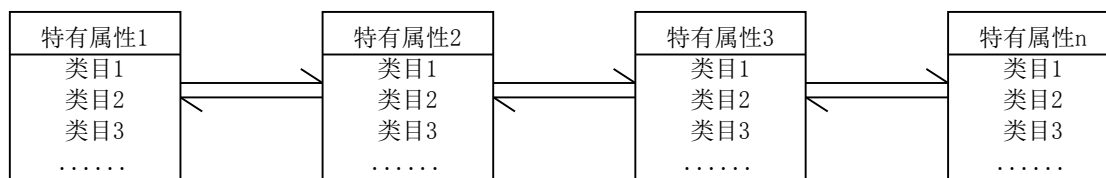


图 1 卫生健康信息数据集-复分类

5.2.3 组配分类法

组配分类法是按照多维度、交叉性方式，将数据集主分类框架与复分类表组合使用，对数据集进行分类的规则。在某个主分类下的卫生健康信息数据集的组配分类示意图 2：

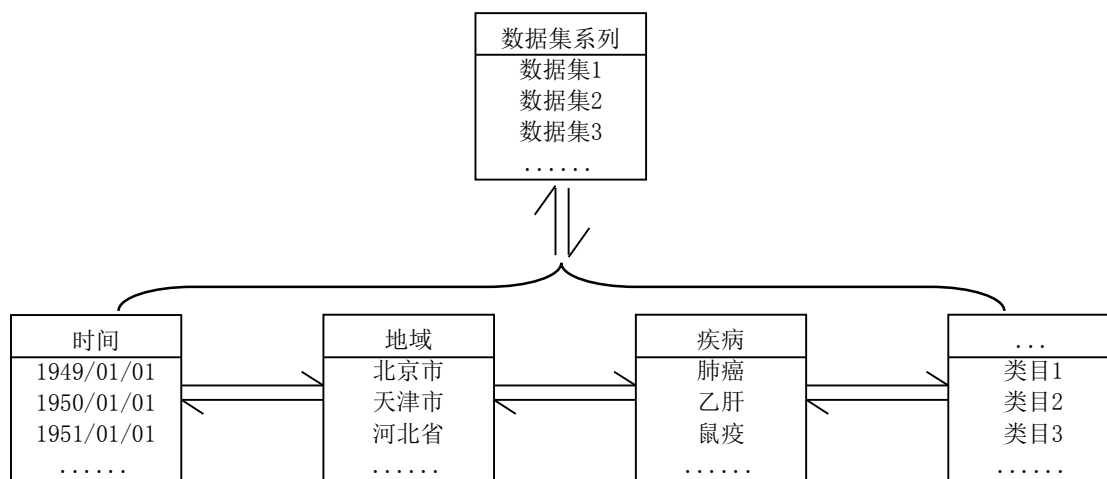


图2 卫生健康信息数据集组配分类

5.3 类目设置规则

5.3.1 主分类表类目设置

5.3.1.1 类目界定规则

卫生健康信息数据集分类与编码中的每一个类目都围绕特定的主题,表达一定主题知识的内涵和外延。遵照本分类与编码规则,科学分类体系所列的类目,能够容纳大量卫生健康信息的最小单元—数据集,从而成为管理数据集的工具。

- a) 类目名称与主题对应: 类目反映主题内容,一一对应。
- b) 类目名称规范: 名称应简洁、规范、通用、无歧义,且确切表达主题内容、范围,必要时应给出类目注释。
- c) 同位类互斥: 同位类之间应界限分明,非此即彼。
- d) 数据集存在性: 每个类目应确保存在一定数量的卫生健康信息数据集。
- e) 类目均衡: 类目均衡展开,允许突出或合并同类。
- f) 层次逐级隶属: 类目从总到分的结构须逐级隶属,每层划分应有单一、明确依据,连续展开。

5.3.1.2 类目层级约束规则

卫生健康信息数据集分类体系的展开层次决定着分类导航系统的导航程度,层次越多越深,知识被组织得越细密,用户查找信息耗费的时间和精力就越多。

卫生健康信息数据集主分类层次应控制在3级-6级之间。

5.3.2 复分类表类目设置

5.3.2.1 类目界定规则

类目界定规则如下:

- a) 复分类表的类目根据特有属性分区段设置,每个区段对应数据集的一个特有属性,构成一个复分类的类目。
- b) 复分类的类目间无层级、无关联,无顺序约束,属于分区段的松耦合约束。
- c) 复分类目区段个数可根据实际需要进行扩展。

5.3.2.2 属性有限性控制

复分类表属性宽度的控制应该充分考虑卫生健康信息数据集内容的特有属性进行设置,详细描述、全面覆盖数据集内容信息,多属性组配,增加数据集管理维度、提高数据组织和生产的效率和质量,增强数据集聚类的灵活性、增加检索入口,适应计算机检索的需要。

卫生健康信息数据集分类中的复分类表可以包括多种属性分类，例如，地区表（国内地区按GB/T 2260、国际地区按GB/T 2659）、组织机构代码表按GB/T 11714、疾病分类表按GB/T 14396、学科分类表按GB/T 13745的规定。

6 数据集编码

6.1 编码基本原则

编码基本规则如下：

- a) 唯一性：在卫生健康信息数据集编码体系中，每一个类目仅有一个代码，一个代码只唯一表示一个主题概念。
- b) 匹配性：代码结构应与分类体系相匹配。
- c) 可扩充性：代码应留有适当的后备容量，以便适应不断扩充的需要。
- d) 简洁性：代码结构应尽量简单，以便减少代码的差错率，便于管理。
- e) 区段性：在卫生健康信息数据集编码体系设置若干区段，每个区段表达不同的类目属性，便于实现不同维度检索。

6.2 编码方法

按GB/T 7027规定的方法。

6.3 编码设计规则

卫生健康信息分类与编码体系设计为两类系统，一类是主分类编码系统，一类是复分类编码系统。给予数据集主分类码和复分类码，进行分段编码标识。

6.3.1 主分类表编码

6.3.1.1 码位设计

本标准规定主分类框架中每一个类目设定唯一的编码，码位结构对应各级类目分段组合而成，见表2，此示例中，一级类目为1位英文字母，二级、三级等类目分别为2位数字，从01到99，后3位给出数字流水码，从001-999，各层级按从属关系逐级顺序编码。

表2 卫生健康信息数据集主分类码位结构

码位	第1位	第2位	第3位	第4位	第5位	第6位	第7位	...	后3位
名称	一级类目码	二级类目码		三级类目码		四级类目码		...	流水码
类型	字母码	数字码		数字码		数字码		数字码	数字码
取值	A—Z	01—99		01—99		01—99		01—99	001—999

6.3.1.2 编码规则

编码规则如下：

- 整体编码是数字字母混合码，同位类目位数、类型一致；
- 各级类目中的“其他”类编码定为“99”。

6.3.2 复分类表编码

6.3.2.1 码位设计

复分类表编码需依据卫生健康信息数据集的特有属性采用分段设计，其基本内容可以由诸如卫生机构、疾病分类和学科分类等多个区段组成，表述格式为××××××××××（GB/T 11714）—×××.×××（GB/T 14396）—×××.×××××（GB/T 13745）—……。

对扩增的复分属性，其代码按照同一规则设计。

6.3.2.2 编码规则

编码规则如下：

- 标准优先性原则。复分类表编码须参照其相应特有属性的标准化代码表，各类代码表的采用中，依次优先采用国家、地区、行业标准代码或积极采用适宜的国际标准代码，没有参考标准时可自定义代码。
- 数据集内容不体现某一复分类目属性时，相应编码位须用等位 0 表述。

7 分类与编码标准文档结构

在卫生健康领域层面对卫生健康信息资源进行统筹规划、规范描述和系统关联时，会涉及大量的数据集分类与编码标准的制定工作。为便于起草这类标准，给出了卫生健康领域各专业《分类与编码标准文档》编写格式与规范，参见本标准附录A。

附录 A

(资料性)

卫生健康信息数据集分类与编码标准文档

A.1 分类与编码标准文档结构

卫生健康信息数据集分类与编码标准的内容和层次划分应符合GB/T 1.1的规定。一般来说，它包含以下内容：

封面
目次
前言
名称
1. 范围
2. 规范性引用文件
3. 术语和定义
4. 分类框架
5. 编码方法
6. 分类体系与代码表
7. 分类与编码的应用
附录（资料性附录）
参考文献

A.2 分类与编码标准结构说明

A.2.1 封面

标准封面上端要注明标准的级别，统一写为：中华人民共和国卫生行业标准；封面中间部分要求写明标准名称的全称及其英文名称；封面的下端则应标明发布日期、实施日期及发布者。

需要特别说明的是：标准名称应简短，并应明确地反映卫生健康信息数据集分类与编码标准化对象或标准的主题。卫生健康信息数据集分类与编码标准名称的形式一般为《×××分类》、《×××代码》或《×××分类与代码》。

A.2.2 目次

卫生健康信息数据集分类与编码标准的目次一般采用二级目次（也可采用三级目次）。若标准的篇幅较少，也可以不要目次。

A.2.3 名称

标准的名称是识别标准的标志之一，标准名称应简练、主题明确、与其他标准易区分。

A.2.4 前言

卫生健康信息数据集分类与编码标准的前言应阐明本标准参考了哪些标准，最后部分应另起一段，按照如下格式来写：

本标准由××××××提出。

本标准由××××××归口。

本标准的附录×是规范性附录（或资料性附录）。

本标准主要起草单位：

本标准主要起草人：

A. 2.5 范围

卫生健康信息数据集分类与编码标准的范围应简明扼要：

本标准规定了××××××××××××的分类代码。

本标准适用于××××××××××××。

A. 2.6 规范性引用文件

卫生健康信息数据集分类与编码标准的规范性引用文件的编写格式是固定的：

下列文件中的内容通过文中的规范性引用而构成本标准必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本标准；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本标准。

GB/T ×××××—××××× （标准名称）

GB ×××××—××××× （标准名称）

ISO ×××××:××××× （标准名称）

注：若没有引用文件，注明：本标准无规范性引用文件。（此节为可选）

A. 2.7 术语和定义

卫生健康信息数据集分类与编码标准的术语和定义，应给出为理解标准中某些术语所必需的定义。在编写时应符合GB/T 1.1中6.3.1的规定。

注：若没有术语和定义，注明：本标准无术语和定义。（此节为可选）

A. 2.8 分类框架

卫生健康信息数据集分类与编码标准的分类原则遵循本标准的规定，但在分类标准文本中，需在此处阐明标准的分类依据和框架结构。

A. 2.9 编码方法

卫生健康信息数据集分类与编码标准的编码原则和方法遵循本标准的规定，但在分类标准文本中需在此处阐明标准所采用的代码种类和代码结构。

当代码结构较复杂时，可用示例说明。若代码较长需校验时，应阐明代码的校验方法（参照GB/T 17710）。

A. 2.10 分类体系与代码表

A. 2.10.1 主分类框架与代码表

卫生健康信息数据集分类与编码标准的主分类框架与代码表（或代码表）可以用表格的形式给出，表格从左到右一般由类目名称栏、代码栏、数据集名称栏、数据集流水号栏组成，并可根据实际需要适当增减栏目，如下所示：

一级类目名称	一级代码	二级类目名称	二级代码	n级类目名称	n级代码	数据集名称	流水码
一级类目 1	A	二级类目 1	01	n级类目 1	01	数据集 1 数据集 2	01 02
				n级类目 2	02
			
		二级类目 2	02	n级类目 1	01	数据集 1
			
			
.....	

A. 2. 10. 2 复分类代码表

卫生健康信息数据集分类与编码标准的复分类属性与代码表需给出数据集复分类类目和引用（自定义）的代码标准。内容可以用表格形式列出，表格从左到右一般由属性段标识位次栏（简称“属性段标识位”）、属性名称栏、引用的编码标准栏组成，如下所示：

属性段标识位	属性名称	引用的编码标准
1	疾病属性	GB/T 14396 疾病分类与代码
2	组织机构属性	GB/T 11714 全国组织机构代码编制规则
3	学科属性	GB/T 学科分类与代码
.....

A. 2. 11 分类与编码的应用

卫生健康信息数据集分类与编码标准的应用需要阐述以下几方面问题的处理：

- 关于交叉主题的处理方法；
- 关于集中与分散问题的处理；
- 关于扩展某一类目的问题的处理；
- 需要特殊强调的类目注释；
- 其他问题。

A. 2. 12 资料性附录

A. 2. 12. 1 卫生健康信息数据集分类与编码标准的资料性附录是为了对标准的正文内容进行补充说明。一般来讲，是针对标准中涉及的主要问题进行完整的例证说明。

A. 2. 12. 2 当分类编码对象较多，可以编制代码索引。代码索引应位于附录之后。索引可采用下述形式编排：

- a) 按编码对象名称的汉语拼音字母顺序；
- b) 按编码对象名称的英文对应词字母顺序；
- c) 按有关的其他关系排序。

A. 2. 13 参考文献

参考文献置于最后一个附录之后，参考文献的起草应遵照 GB/T 7714 的规定。
