采用基于 MedDRA 词典的《医学术语自动编码系统》进行医学术语编码的问题和分析

王燕妮，郝 颖，周健文，李 玲，孟凡强*

翰博瑞强医药科技有限公司 上海

【摘要】**目的** 总结 MedDRA 编码的实践经验以提高药物临床试验数据编码的准确性。**方法** 通过回顾性分析 20000 个词汇的编码实践，以具体示例总结应用 MedDRA 在药物临床试验编码过程中遇到的各种情况及策略，包括词汇拆分、自动编码的审核和编码质疑。**结果** 自动编码系统的使用大大提高了编码的效率，同时也有一定的局限性。一词多义、多词一义、需要拆分和澄清的词汇都需要进行人工判断和编码，目前还不能由系统来替代。**结论** 当前，虽然自动编码系统的使用显著提高了工作效率，但是人工编码和医学审核的流程仍然是不可或缺的。

【关键词】医学编码；药物临床试验；自动编码

Problems and Analysis of Medical Terms Coding Using MedDRA Dictionary-based Medical Terms Automatic Coding System

Yanni Wang, Ying Hao, Jianwen Zhou, Ling Li, Fanqiang Meng

Hanbo Ruiqiang Pharmaceutical Technology Co., Ltd. Shanghai

【**Abstract**】**Objective** To summarize the practical experience of MedDRA coding to improve the accuracy of drug clinical trial data coding. **Methods** Through a retrospective analysis of the coding practice of 20,000 words, various situations and strategies encountered during the coding of drug clinical trials using MedDRA were summarized with specific examples, including word splitting, automatic coding review and coding challenge. **Results** The use of automatic coding system greatly improved the coding efficiency, but also had some limitations. Words with multiple meanings, multiple words with one meaning, and words that need to be split and clarified all require manual judgment and coding, and cannot be replaced by the system at present. **Conclusion** Currently, although the use of automatic coding systems has significantly improved work efficiency, the process of manual coding and medical review is still indispensable.

【**Keywords**】Medical coding; Drug clinical trials; Automatic coding

前言

在药物临床试验中，医学编码不仅为数据管理和统计分析服务，通过医学编码的术语也是药品说明书撰写的基础，因此医学编码的准确性、客观性尤为重要。

国内外发布的多个指导原则中都对药物临床试验的编码工作提出要求：国际人用药品注册技术协调理事会(The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use，ICH)《M1：ICH 药事监管活动国际医学用语词典（Medical Dictionary for Regulatory Activities，MedDRA）》[1]、ICH《E2B（R3）：临床安全数据的管理：个例安全性报告传输的数据要素》[2]、国家食品药品监督管理总局药品审评中心 2018 年 06 月发布的《药物临床试验期间安全性数据快速报告标准和程序》[3]，2020 年 07 月发布的《药物临床试验数据递交指导原则（试行）》[4]和 2021 年 09 月发布的《药物临床试验数据管理与统计分析

*通讯作者：孟凡强

计划指导原则（征求意见稿）》[5]等都要求对临床试验期间受试者自发汇报的或者研究者记录的既往病史、不良事件（Adverse Events，AE）、严重不良事件（Serious Adverse Events，SAE）进行统一、规范的医学编码。

目前国内常用的医学编码工具有 MedDRA 、WHOART 和 ICD10。药物临床试验（包括中美双报临床试验）使用 MedDRA 编码。

临床试验不良反应涉及身体多个系统，患者汇报的症状和研究者的异常发现都存在一词多义、多词一义的情况。翰博瑞强数据管理团队既往 12 年使用 MedDRA 词典进行编码的超过 100 万个词条，在此基础上开发出《医学术语自动编码系统》（专利号 CN111797593A），试图部分解决医学编码遇到的各种难题。本文以真实临床试验项目的医学术语为研究对象，采用《医学术语自动编码系统》，研究临床试验中对不良事件编码过程中出现的问题，以期找出临床试验中文医学编码汇的规律，以供未来更好的工具开发积累经验。

1 方法

1 1 医学编码工具

本文中编码所采用的词典为 MedDRA。MedDRA 是 ICH 针对临床试验开发的标准医学术语集，有预先界定的 5 个术语层级归属：包括低位术语（Lowest Level Term，LLT）、首选术 语（Preferred Term，PT）、高位语（High Level Term，HLT)、高位组术语（High Level Group Term，HLGT)和系统器官分类（System Organ Class，SOC)。使用时用户将药物临床试验中收集的原始医学术语（Verbatim Term)采用 MedDRA 词典逐一归属到相应层级的“标准医学术语”的过程称为“编码”。ICH MedDRA 维护和支持服务组织（MSSO）对于编码词典的使用提供了详细的视频和培训，但是对于某个具体的词汇应该如何编码，需要用户进行判断，MSSO 并不负责提供答案。本文所有编码由经过专业培训的编码人员完成。

1.2 材料来源

翰博瑞强既往承接的 1135 项 I-IV 期、BE 临床试验，治疗领域涉及肿瘤、心血管、CNS、消化、呼吸等系统的创新药、仿制药。鉴于 MedDRA 版本的更新换代，本文研究只采用使用 MedDRA 24.0 完成编码的临床试验，所编码词汇共计 20000（2%）个。

1.3 编码方法

先由数据管理经理采用翰博瑞强自主开发的《医学术语自动编码系统》（专利号：****）进行初筛：将所有临床试验收集的“原始医学术语”进行查重以去除重复的术语；再将剩余的术语与 MedDRA 的地位术语 LLT 和首选术语（PT）进行自动匹配，筛除完全一致的医学术语。将完全匹配的术语和剩余的无法与 MedDRA-PT 匹配的“原始医学术语”提交医学专业人员进行医学编码。

1.4 医学编码

由具有临床医学工作经验的“医学顾问”对临床试验中收集的与 MedDRA-PT 直接匹配和无法匹配的“原始医学术语”在基于 MedDRA 的《医学术语自动编码系统》上进行逐一核查并进行 PT/HLT/HLGT/SOC 的归属判定。对于概念模糊、描述不清的术语则向研究者发出质疑（Query）以便进一步编码。对于疑难的编码，由两位医学顾问和研究者沟通后解决。

2 结果

2.1 编码的词汇

采用 MedDRA 24.0 完成编码的词汇共计 20000（2%）个。来自 5 个临床试验；研究的适应症包括抗肿瘤、治疗糖尿病和抗凝药物；涉及的躯体器官、系统 27 个。

使用基于 MedDRA 的《医学术语自动编码系统》初步筛选，去重后的词条共计 3738 个，80.3%的“原始医学术语”是重复性的术语，符合临床试验不良事件出现的规律。由于存在“一词多义”的情况，例如同为“疼痛”，可能有“头痛”、“腹痛”、“伤口疼痛”等区别。所以，对于重复的原始术语词条，依然需要医学顾问逐条核对。

2.2 编码结果

采用《医学术语自动编码系统》，15643（78%）个词汇可以直接与 MedDRA 的 PT 术语相匹配，其过程仅仅需要 1-2 分钟的时间（如果人工逐项与 MedDRA 匹配，需要大约累计 98 个工作日）。将这些“自动编码”的词汇由医学顾问进行 100%人工核查，发现自动编码的准确率为：80.7%；表明即使是“原始医学术语”与 MedDRA 的 PT 完全匹配，依

然可能在 HLT/HLPT/SOC 的归属上发生错误。当前阶段，医学人工核查环节必不可少。

采用《医学术语自动编码系统》初步筛选不能直接与 MedDRA 的 PT 匹配的词汇总计 4357（22%）个。由医学顾问对这些词汇进行人工编码，发现 503/20000 个词汇（2.5%）编码时需要进行拆分，例如头晕伴耳鸣需要拆分为“头晕”和“耳鸣”；263/4357（1.3%）个词汇需要进行质疑，例如研究者由于转氨酶升高判定的原始术语“肝功能不全”，可能最终被编码为“肝功能衰竭”。

将 4357 个术语进行逐个编码，共使用 28 个累计工作日完成编码。医学顾问平均 3 分钟完成一个编码。在实际工作中往往所用时间更长。由此可见，医学顾问编码的时间主要花费在这些“不规范”的原始术语上。

2.3 医学编码

（1）一词多义的情况分析

MedDRA 的术语具有“多轴性”，即一个术语出现在多个 SOC 中，并根据不同的类别分组（如：按病因或发病部位），这些术语可以在不同的数据集进行检索和表达[6]。

多轴性表示一个 PT 可能存在于多个 SOC 中。这样可以使术语从医学角度按照不同方式进行分组（例如，按病因或按器官系统）。每个 PT 被分配了一个主 SOC；该 PT 被分配的所有其他 SOC 称为“次 SOC”。一个 PT 只有一个主 SOC 是为了避免从所有 SOC 输出数据时，重复计算事件[7]。

MedDRA 是一个标准术语集，有预先界定的术语层级结构，不应更改。用户不得对 MedDRA 进行临时的结构改动，包括变更主 SOC 分配；这样做将有损该标准的完整性。如果发现术语的 MedDRA 层级结构不正确，应向 MSSO 提交变更申请[7]。如“肺部感染”的主 SOC 是“感染及侵染类疾病”，次 SOC 是“呼吸系统、胸及纵隔疾病”，编码时只可以选择主 SOC “感染及侵染类疾病”。MedDRA 对于“感染”的主 SOC 规则为“所有感染性疾病以 SOC 感染及侵染类疾病为主 SOC，这类术语的次 SOC 分配原则是发病部位”[7]。

（2）多词一义的情况分析

在编码时常常会遇到一些词典里没有完全匹配的 LLT 术语的情况，对于多词一义的术语可通过同义词替换来找到合适的 LLT 术语进行编码。

①形容实验室检查值异常的同义词替换

形容实验室检查值异常的减少、减低、下降、降低均为同义。如在 MedDRA 24.0 中没有白细胞计数减少、白细胞计数减低、白细胞计数下降，但有白细胞计数降低，采用同义词替换的方法均可用白细胞计数降低进行编码。类似的还有增多、增加、偏高、升高等等均为同义，加上具体的检查项后会出现多词一义的情况，可采用同义词替换的方式进行编码。

②不良事件名称的同义词替换

由于汉语表达的多样性，相同的不良事件可能记录不同的名称，如不良事件“胃部疼”、“胃疼”，在 MedDRA 24.0 中没有完全匹配“胃部疼”的 LLT 术语，但是可以使用它的同义词“胃疼”进行编码。“鼻腔出血”、“鼻部出血”在 MedDRA 24.0 中没有完全匹配的 LLT 术语，可以使用它的同义词“鼻出血”进行编码。“室性早博”、“室早”在 MedDRA 24.0 中没有完全匹配的 LLT 术语，可以使用它的同义词“室性期前收缩”进行编码。“癌痛”可采用同义词“癌症疼痛”进行编码。“乙肝病毒肝炎”可采用同义词“病毒性乙型肝炎”进行编码。

③扩大范围编码

有些不良事件或病史名称记录的很具体，词典没有特别具体的 LLT 与之对应，可采用扩大范围进行编码，如“左侧睾丸术”可采用“睾丸手术”进行编码。“双肾多发囊肿”、“双肾囊肿”可采用“肾囊肿”进行编码。

④关键词编码

MedDRA 词典中绝大多数的术语不带有严重程度，如轻度、中度、重度，当遇到记录了严重程度而词典中无带有严重程度的术语时，可忽略严重程度，编码关键词。如“轻度贫血”、“贫血”，可均使用“贫血”进行编码。“轻度脂肪肝”、“脂肪肝”，可均使用“脂肪肝”进行编码。“轻度颈部疼痛”、“中度颈部疼痛”、“颈部疼痛”，可均使用“颈部疼痛”进行编码，等等。

（3）医学顾问人工核查的结果

当低位术语（LLT）与药物临床试验中录入的词汇不完全匹配时，需要医学顾问人工复核。表 1 举例罗列当 LLT 与原始术语不一致时各层级的归属。

以下示例为程序与人工编码的差异，可见如果任由程序自动匹配，可能出现南辕北辙的现象。

（4）医学顾问词汇拆分的结果

根据法规《MedDRA®术语选择：考虑要点》要求：医学顾问对能提供更多临床信息的不良反应/不良事件报告进行“拆分”，选择多个 MedDRA 术语[7]。表 2 为拆分编码示例。

（5）医学编码的质疑

本研究发现，医学编码过程中发出的质疑按照要澄清的内容分为 8 种类型：①确认是否为不良事件。②确认不良事件名称是否为死亡或是否有明确的死亡原因。③确认疾病名称和严重程度。④确认不良事件名称为疾病诊断还是实验室检查异常。⑤确认病因。⑥确认病变部位。⑦确认是否为感染性疾病。⑧确认诊断或实验室检查名称（详见表 3）

（6）医学编码的质疑与澄清举例

例 1 “血压升高”

健康人的血压也可能因为情绪、运动等原因一过性升高。MedDRA 词典中对于“高血压”的 SOC 为“血管与淋巴管类疾病”，“血压升高”的 SOC 为“各类检查”。

例 2 “阻塞性肺炎”

“阻塞性非感染性肺炎”及“阻塞性感染性肺炎”对应的 PT term 分别为“肺部炎症”及“感染性肺炎”，因此需要质疑疾病的病因是否为感染。类似的情况还有：牙龈炎、腹膜炎、膀胱炎等。

例 3 “肝功能不全”

受试者实验室检查 ALT、AST 升高，可能报告为 ALT 升高、肝功升高、肝功异常、肝功能不全等。实际上，单项 ALT、AST 轻度升高可能不具有临床意义。严重 ALT、AST 升高一定伴有其他指标的异常，包括血/尿胆红素异常、肝脏肿大、黄疸等。有时一个简单的 ALT 升高被判定为肝功能异常（PT code：10019670），而 SOC 为肝功能衰竭。完全脱离了实际情况。

表 1 当低位术语（LLT）与药物临床试验中录入的词汇不完全匹配时，需要医学顾问人工复核

药物临床试验中录入词汇（Verbatim Term）	编码方式	LLT	PT	HLT	HLGT
周围神经损伤	自动	神经损伤	神经损伤	各种神经损伤（不另分类） 神经系统异常（不另分类）	各种损伤（不另分类） 神经类疾病（不另分类）
	人工修正	外周神经损伤	外周神经损伤	各种周围神经损伤 各种周围神经疾病（不另分类）	各种损伤（不另分类） 外周神经类疾病
精神差	自动	精神恶化	精神损害	认识和注意力紊乱和障碍（不另分类） 精神损害（痴呆和记忆力丧失除外）	认识及注意力紊乱和障碍 精神损害类疾病
	人工修正	精神萎靡	倦怠	心境障碍性疾病（不另分类） 虚弱状态	心境障碍和混乱（不另分类） 全身系统疾病（不另分类）
饮食欠佳	自动	食欲下降（未特指）	食欲减退	各种食欲障碍 各种感受与感觉（不另分类）	食欲及一般性营养障碍 全身系统疾病（不另分类）
	人工修正	摄食量减少	摄食量减少	各种食欲障碍 各种摄食障碍（不另分类） 各种胃肠道症状和体征（不另分类）	食欲及一般性营养障碍 饮食障碍和混乱 胃肠道系统症状和体征
无症状菌尿	自动	菌尿症	菌尿症	各种细菌感染（不另分类） 各种泌尿系统障碍	细菌感染性疾病 各种泌尿道体征和症状
	人工修正	无症状性菌尿	无症状性菌尿	各种细菌感染（不另分类） 各种泌尿系统障碍	细菌感染性疾病 各种泌尿道体征和症状
冠心病	自动	冠状动脉疾病	冠状动脉疾病	各种冠状动脉疾病（不另分类） 冠状动脉坏死和血管功能障碍	冠状动脉类疾病 动脉硬化、狭窄、血管功能不全和坏死
	人工修正	冠状动脉粥样硬化性心脏病	冠状动脉硬化	各种冠状动脉疾病（不另分类） 冠状动脉坏死和血管功能障碍	冠状动脉类疾病 动脉硬化、狭窄、血管功能不全和坏死

表 2 词汇拆分编码示例

Verbatim Term	拆分并编码				
	LLT	PT	HLT	HLGT	SOC
疼痛（左侧腰部至腿部）	腰部疼痛	背痛	肌肉骨骼和结缔组织的各种疼痛与不适	肌肉骨骼和结缔组织疾病（不另分类）	各种肌肉骨骼及结缔组织疾病
	腿部疼痛	肢体疼痛	肌肉骨骼和结缔组织的各种疼痛与不适	肌肉骨骼和结缔组织疾病（不另分类）	各种肌肉骨骼及结缔组织疾病
头晕伴耳鸣	头晕	头晕	神经学症状和体征（不另分类）	神经类疾病（不另分类）	各类神经系统疾病
	耳鸣	耳鸣	各种内耳症状和体征	内耳及第八颅神经类疾病	耳及迷路类疾病
头颈部疼痛	头痛	头痛	各种头痛（不另分类）	各类头痛	各类神经系统疾病
	颈部疼痛	颈痛	肌肉骨骼和结缔组织的各种疼痛与不适	肌肉骨骼和结缔组织疾病（不另分类）	各种肌肉骨骼及结缔组织疾病
乙肝小三阳	乙型肝炎表面抗原阳性	乙型肝炎表面抗原阳性	病毒鉴定和血清学	微生物和血清检查	各类检查
	乙型肝炎核心抗体阳性	乙型肝炎核心抗体阳性	病毒鉴定和血清学	微生物和血清检查	各类检查
	乙型肝炎 e 抗体阳性	乙型肝炎 e 抗体阳性	病毒鉴定和血清学	微生物和血清检查	各类检查
头面部麻木	头部麻木	感觉减退	感觉异常和感觉迟钝	神经类疾病（不另分类）	各类神经系统疾病
	面部麻木	感觉减退	感觉异常和感觉迟钝	神经类疾病（不另分类）	各类神经系统疾病

表 3 医学顾问编码过程中发出的质疑类型

质疑原因	频数（百分比）
确认病因	8（5.6%）
确认部位	19（13.4%）
确认疾病名称和严重程度	1（0.7%）
确认是否为感染性疾病	11（7.7%）
确认是否为不良事件	1（0.7%）
确认死亡或是有明确的死亡原因	1（0.7%）
确认诊断或实验室检查名称	96（67.6%）
确认疾病诊断还是检查异常	5（3.5%）
总计	142（100%）

例 4 发生的“身体部位”确认

结扎术以医学编码而言，除了“输精管结扎术”和“输卵管结扎”的 LLT，还有“动脉结扎”、“静脉结扎”、“痔疮结扎术”等 LLT，单纯以“结扎术”无法进行编码。因此需要中心澄清结扎的部位，经澄清结扎部位为输卵管，则以 LLT：输卵管结扎，PT：女性绝育术（PT code：10056199）进行编码。

在 MedDRA 24.0 中，包含疖的 LLT 众多 有“鼻部的疖”、“股部的疖”、“脚趾的疖”等，PT 均是“疖”，还有 LLT“生殖器疖”，PT 为“生殖器脓肿”，LLT“外耳道疖”，PT 为“外耳道脓肿”。为了编码的准确性，需确认“疖”的部位。

例 5 “死亡”编码

对本文中 20000 个词汇中死亡或有明确死亡原因的事件进行汇总，共有 29 例转归为死亡。其中，初始以死亡为名称的 15 例，以死亡原因为名称的 14 例，后经质疑澄清确认，部分不良事件名称修改为死亡原因“冠心病急性前壁心肌梗死”、“右肺腺癌”、“肺腺癌伴多发转移”、“肺腺癌疾病进展”等。

4 讨论

（1）在药物临床试验中，需要进行编码的词汇要几万条。词汇编码中 Verbatim term 与词典中词汇一一对应的过程耗费大量时间。自动编码系统的应用将重复单调的工作效率明显提高。但是，由于编码词汇的复杂性，对于一词多义、多词一义的词汇以及需要拆分的词汇进行医学判断，目前阶段还只能依赖人工编码完成。

（2）有些有歧义或信息不完整的词汇，无论是自动编码还是人工编码都不能直接完成，还需要质疑澄清。只有录入的数据准确无误，才有可能得到正确的编码结果。因此，有时“编码”过程不仅仅是在后台对词汇的处理，还需要与临床团队进行交互澄清才能完成。

随着自然语言处理技术和人工智能的应用推广，词汇编码工作的自动化程度有望进一步提高。但人工智能算法准确度的提升依赖于前期大量人工词汇标注和对模型的训练，需要投入大量的人力和时间。目前编码的词汇都是由临床团队进行录入完成，在编码词汇澄清时，临床团队往往需要查阅原始病历的资料，核对后进行判断和回复。若使这一环节自动化，还有待研究中心的信息化水平提高，病历的结构化处理以及研究中心与编码系统的信息交互。所以，编码工作全面自动化的实现脱离不了医疗行业整体信息化水平的提高。

参考文献

[1] 国家食品药品监督管理局药品审评中心.M1:监管活动医学词典(MedDRA)（中文版)[EB/OL].(2018). https://www.cde.org.cn/ichWeb/guideIch/toGuideIch/4/0

[2] 国家食品药品监督管理局药品审评中心.E2B（R3）:临床安全数据的管理：个例安全性报告传输的数据要素(中文版)[EB/OL].(2016-11-10).https://www.cde.org.cn/ichWeb/guideIch/toGuideIch/3/1

[3] 国家食品药品监督管理局药品审评中心.药物临床试验期间安全性数据快速报告标准和程序[EB/OL].(2018-06-03).https://www.cde.org.cn/main/news/viewInfoCommon/1293f7c3d511225fadbab12a209d152c

[4] 国家食品药品监督管理局药品审评中心.药物临床试验数据递交指导原则（试行）[EB/OL].(2020-07-20).https://www.cde.org.cn/main/news/viewInfoCommon/f649995d3a9ade8dcd67f6a2ced36f0b

[5] 国家食品药品监督管理局药品审评中心.药物临床试验数据管理与统计分析计划指导原则（征求意见）[EB/OL].(2021-09-03).https://www.cde.org.cn/main/news/viewInfoCommon/550b58058f4654812cecdc4e573e7773

[6] MedDRA 入门指南 24.0 版.[EB/OL].(2021-03). https://www.meddra.org/how-to-use/support-documentation/chinese/welcome

[7] 国家食品药品监督管理局药品审评中心.MedDRA®术语选择:考虑要点(中文版)[EB/OL].(2018-09-01).https://www.cde.org.cn/ichWeb/guideIch/toGuideIch/4/0

收稿日期：2022 年 3 月 13 日
出刊日期：2022 年 4 月 27 日
引用本文：王燕妮，郝颖，周健文，李玲，孟凡强，采用基于 MedDRA 词典的《医学术语自动编码系统》进行医学术语编码的问题和分析[J]. 国际医学与数据杂志, 2022, 6(1) : 121-126.
DOI: 10.12208/ j.ijmd.20220031

检索信息：RCCSE 权威核心学术期刊数据库、中国知网（CNKI Scholar）、万方数据 （WANFANG DATA）、Google Scholar 等数据库收录期刊