呼吸道传染病个案流行病学画像标签体系构建研究*

杨子良 赵自雄 张业武 马家奇

（中国疾病预防控制中心 北京 102206）

〔摘要〕 构建个案流行病学标签体系，详细阐述标签体系构建方法及示范场景应用，指出构建的标签体系可用于流行病学调查过程中提取个案流行病学特征，为进一步构建知识图谱提供流行病学调查信息分级分类的基础。

〔关键词〕 流行病学调查；用户画像；标签体系

〔中图分类号〕R－058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2022.04.008

Study on the Construction of a Case Epidemiological Portrait Labeling System for Respiratory Infectious Diseases

YANG Zilang, ZHAO Zixiong, ZHANG Yewu, MA Jiaqi, Chinese Center for Disease Control and Prevention, Beijing 102206, China

〔**Abstract**〕 A case epidemiological labeling system is constructed. The paper expounds the construction method and the application of demonstration scenarios of the labeling system in detail, and points out that the constructed labeling system can be used to extract epidemiological features of cases during epidemiological investigation, and provide a basis for further construction of knowledge graphs for hierarchical classification of epidemiological survey information.

〔**Keywords**〕 epidemiological investigation; user portrait; labeling system

1 引言

流行病学调查（以下简称流调）是用流行病学方法进行的调查研究，是传染病暴发中寻找传染源、追踪和判定密切接触者、分析传播特征与传播关系的重要手段[1]。现实工作中流调信息纷繁复杂且具有不确定性、主观性和缺少统一标准等特点。仅靠人工从海量非结构化流调信息中梳理出传播特征和传播关系费时、费力、时效性差。基于标签体系建立个案流行学画像，可在一定程度上解决流调信息无法直接通过计算机批量处理和分析的难点[2]，具有重要现实意义。用户画像是在收集用户属性及行为信息基础上抽取用户信息全貌、形成数字化标签集以表征和预测用户的行为，是对真实用户的数字化建模[3]。标签是通过人为概括或定义以唯一性语义说明其对应实体的具体含义，无需文本分析等过多预处理便可为后续信息读取、计算、分析和可视化展示提供便利[4]。本研究将用户画像和标签理论与流调相关核心业务结合，以呼吸道传染病病例个案流调相关要素为研究对象构建个案流行病学标签体系，将零散、复杂的流调信息转换为形象、结构化、易懂的标签，为提取个案流行病学特

〔修回日期〕 2022－01－18

〔作者简介〕 杨子良，硕士研究生；通讯作者：马家奇，主任医师。

〔基金项目〕 国家自然科学基金项目“面向人群健康和重大疾病的大数据集成共享平台研究及示范应用”（项目编号：91846303）。

征、建立个案流行病学画像提供模型基础。

2 研究对象与方法

2.1 研究资料

以呼吸道传染病病例个案流调相关要素为研究对象。研究资料为肺鼠疫、严重急性呼吸综合征（Severe Acute Respiratory Syndrome，SARS）以及肺炭疽以及不明原因肺炎的病例个案流行病学调查表（以下简称流调表）。

2.2 标签抽取方法

使用指标标签化与活动要素模型从流调表中抽取标签。指标标签化是从数据指标出发，在保持指标含义不变的条件下，将指标转换为以短词语或短词组形式为主的精简标签[5]。活动 3 要素模型将活动看作由一系列事件构成，将事件概括为何人（Who）、何时（When）、何地（Where）3 要素集合[6]。

2.3 标签分类方法

根据流调表业务属性，采用等级列举式分类法构建 1 级和 2 级标签。等级列举式分类法是一种以知识分类为基础，依据概念划分与概括原理，将概括文献内容与事物的各种类目组成一个层层隶属、详细列举的等级结构体系的分类法。其特点是采用等级隶属方法规定类目间相互关系，通过全面列举表达一组完整的并列类目，以树型结构在类目表上将众多类目系统排列起来组成一个体系[7]。

2.4 类目标签提取方法

在 2 级标签基础上采用分面组配式分类方法形成多个类目标签。分面组配式分类方法是将概括事物的主题概念分解为简单概念，组成“分面－亚面－类目”结构体系[8]。本研究应用分面组配式分类法中冒号分类法的 5 个基本范畴理论进行分面分析，从本体、空间、时间、动力、物质 5 个范畴[9]，对 2 级标签的流调具体元素进行分面分析，见表 1。

表 1 冒号分类法的 5 个基本范畴

基本范畴	意义
本体	事物本身
物质	构成事物的材料或要素
动力	事物各种活动、影响、状态和问题
空间	事物存在或发生的地点
时间	事物存在或发生的时间

2.5 标签属性定义方法

标签属性可以理解为针对标签进行的再标注，即标签的标签，主要包括标签类型、标签取值与标签组合规则等[10]。为便于理解和表达，本研究参照数据元属性相关理论定义标签属性。数据元属性包括数据名称、定义、数据类型表示格式以及数据元允许值等多个维度[11]。本研究采用多个维度混合定义标签属性，定义标签数据类型和标签表示格式。数据表示格式中，可枚举字符型参照流调表中指标选项进行定义，不可枚举字符型对标签的取值进行定义。

2.6 研究技术路线

个案流行病学画像标签体系构建研究技术路线包括标签体系构建与示范场景应用研究两部分。首先分析流调表，从中抽取关键指标并进行标签化。通过等级列举式分类法构建 1 级和 2 级标签，利用分面组配式分类法构建类目标签并定义类目标签属性，形成个案流行病学标签体系。再通过个案流行病学画像、病例间相互关系及传染病传播链分析 3 个场景进行应用示范，见图 1。

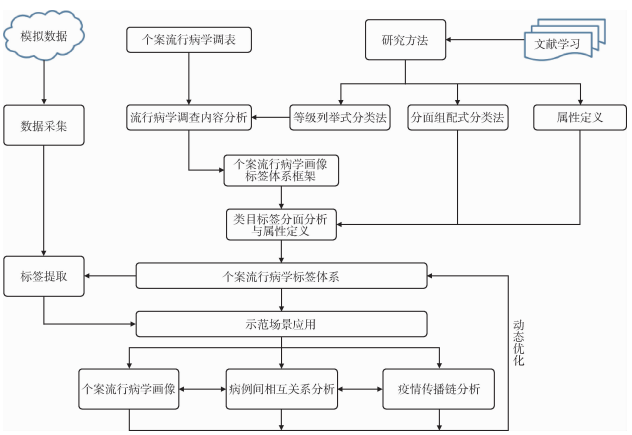

图 1 个案流行病学画像标签体系构建研究技术路线

3 标签体系构建

3.1 流行病学调查内容分析

流调表集合个案调查对象基本信息、病例发现与就诊、暴露史与接触史、实验室检验等信息，可在一定程度上反映流调数据特征。通过归纳分析 4 种呼吸道传染病流调表的内容并统计其包含的流行病学要素，见表 2。本研究重点关注呼吸道传染病早期暴发的流行病学调查，病例入院日期、出院日期、既往病史、转归及并发症等疾病临床发病过程暂不作为本次研究内容。因此将人口学特征、临床症状、临床体征、病原学检测、血清学检测、暴露史与轨迹史、疫苗接种史、病例发现 8 个流调要素纳入标签体系构建。

表 2 流调要素分析

流调要素	频次	频率（%）
人口学特征	4	100
临床症状	4	100
临床体征	4	100
病原学检测	4	100
血清学检测	4	100
暴露史与轨迹史	4	100
发病过程	3	75
疫苗接种史	2	50
病例发现	1	25

3.2 指标标签化

从流调要素中归纳提炼出关键指标，将指标统一归类、合并重复。该过程共提炼出 37 个关键指标，其中有无疫苗接种史、发病前是否去过外地以及可能感染来源 3 个指标表达复杂不宜直接拿来作为标签，因此标签化为短语形式的标签，见表 3。流调表中轨迹史与暴露史部分的指标复杂且不具体，无法直接进行标签化。暴露史与轨迹史可看作随时间变化的事件集合，因此通过构建活动要素模型对其进行描述。传染病的传播依靠人 - 人或人 - 物 - 人之间的接触，因此人 - 物、人 - 人之间的关系是流调过程中重点关注的要素[12]。结合流调轨迹史与暴露史特征，本研究在 3 要素活动模型的基础上添加关系要素，提出流行病学 4 要素活动模型，用时间（When）、地点（Where）、对象（Item）、关系（Relationship）描述调查对象的暴露史与轨迹史。为使活动要素与流行病学标签体系相衔接，将流行病学 4 要素模型标签化为接触时间、接触地点、接触对象和与接触关系 4 个标签，其中接触对象包含接触的人、环境与物品。

表 3 指标标签化结果

个案调查表指标	标签化结果
有无疫苗接种史	预防接种史
发病前是否去过外地	流动状态
可能感染来源	感染来源

3.3 个案流行病学标签体系框架

在指标标签化的基础上，根据《中华人民共和国卫生行业标准》中呼吸道传染病诊断依据，即传染病的诊断需要有确切的流行病学史、相应临床表现以及特异性实验室检查结果[13-15]。采用等级列举式分类法将个案流行病学调查活动列举为基本信息、临床症状与体征、实验室检测与流行病学史 4 个基本大类，作为 1 级标签。在基本大类基础上，按照其构成列举出 8 个基本类，作为 2 级标签，形成个案流行病学画像标签体系框架，见图 2。

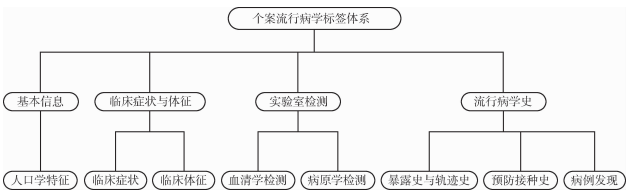

图 2 个案流行病学画像标签体系框架

3.4 类目标签的分面分析与属性定义

3.4.1 类目标签的分面分析 类目标签是在 2 级标签基础上采用分面组配式分类法分析形成的标签，表 4 为分面分析得到的 35 个类目标签，其中标 * 号的为按照 5 个基本范畴理论，结合具体业务得

到的拓展标签，其余为在流调表的基础上进行分面分析得到的基础标签。

表 4 类目标签分面分析结果

1 级标签	2 级标签	类目标签分面分析			
		本体（P）	空间（S）	时间（T）	动力（E）
基本信息	人口学特征 *	患者信息	现住地址	调查时间	临床分型
					流动状态
临床症状与体征	临床症状	症状类别	发病地点	发病日期	诊断结果
				就诊日期	
	临床体征	体征类别	就诊地点	检查日期	检查结果
				诊断日期	
实验室检测	血清学检测	标本类型	检测地点 *	采样时间	检测结果
	病原学检测	标本类型	检测地点 *	采样时间	检测结果
流行病学史	预防接种史	疫苗种类	接种地点 *	接种时间	接种剂次 *
	暴露史与轨迹史	接触对象	接触地点	接触时间	接触关系
	病例发现	发现途径 *	发现地点	发现时间	感染来源

3.4.2 类目标签的属性定义 通过整理提炼各个类目标签的属性特征，对标签属性进行归类，得到类目标签属性定义的结果，见表 5。

表 5 个案流行病学画像类目标签属性定义

1 级标签	2 级标签	类目标签	属性定义
基本信息	人口学特征	患者信息	身份证号、姓名、年龄、职业
		现住地址	S1（病例现住地区的省、市、县（区）、街道等名称）
		调查时间	D（D8）
		临床分型	S3（皮肤型、肠型、肺型、其他）
		流动状态	S3（境外、跨省、跨市、跨县（区）、本区县）
临床症状与体征	临床症状	症状类别	S3（呼吸道症状、消化道症状、全身症状）
		发病地点	S1（发病时居住地的街道/社区名称）
		发病日期	D（D8）
		就诊日期	D（D8）
		诊断结果	S3（肺炭疽、肺鼠疫、SARS、不明原因肺炎、其他）
	临床体征	体征类别	S3（血常规、血生化、肺部影像学）
		就诊地点	S1（所就诊的医疗机构名称）
		检查日期	D（D8）
		诊断日期	D（D8）
		检查结果	S2（正常、异常）
实验室检测	血清学检测	标本类型	S3（血清、咽拭子、血液、粪便、痰）
		检测地点	S1（检验检测机构名称）
		采样时间	D（D8）
		检测结果	S2（阳性、阴性）

续表 5

	病原学检测	标本类型	S3（血清、咽拭子、血液、粪便、痰）
		检测地点	S1（检验检测机构名称）
		采样时间	D（D8）
		检测结果	S2（阳性、阴性）
流行病学史	暴露史与轨迹史	接触对象	S3（所接触人、物品、环境）
		接触地点	S1（发生接触的具体语义地点）
		接触时间	D（DT）
		接触关系	S3（接触、家人、邻居、同事、朋友）
	预防接种史	疫苗种类	S1（呼吸道传染病疫苗）
		接种地点	S1（疫苗接种单位）
		接种时间	D（D8）
		接种剂次	N（对应疫苗剂次）
	病例发现	发现途径	S2（主动、被动）
		发现时间	D（D8）
		发现地点	S1（发现病例的省、市、县（区）名称）
		感染来源	S3（环境、物品、人）

3.5 个案流行病学画像标签体系

标签体系是通过对多种标签进行归类并对标签属性加以定义形成的。通过等级列举式分类法构建 1 级与 2 级标签，利用冒号分类法进行分面分析构建类目标签，并对标签体系每一层添加标识符定位标签位置（如患者信息可表示为 1aP），形成个案流行病学画像标签体系，见图 3。

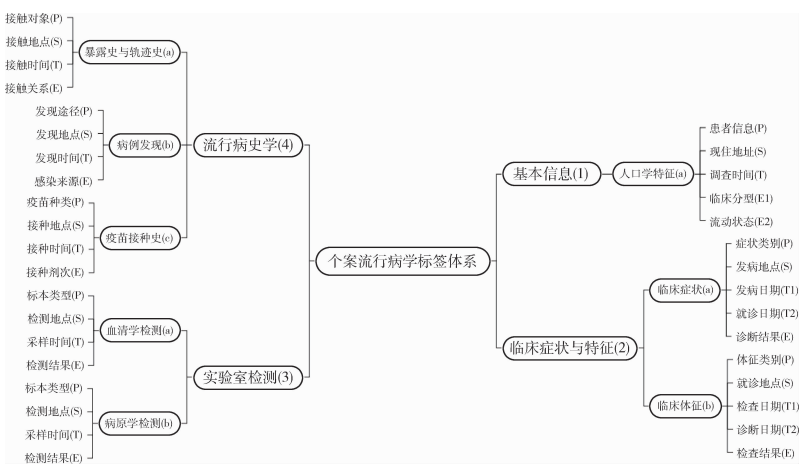

图 3 个案流行病学画像标签体系

4 示范场景应用

4.1 描述个案流行病学画像

通过自然语言处理等技术，从复杂流调信息中提取标签体系所需要信息要素形成数字化标签，刻画个案流行病学信息，通过画像可视化方式展示。此方式有别于传统意义上的文本描述，能够让研究者更加生动、直观、全面地了解该调查对象的流行病学特征。利用本研究构建的标签体系可对病例进行流行病学画像，见图 4。

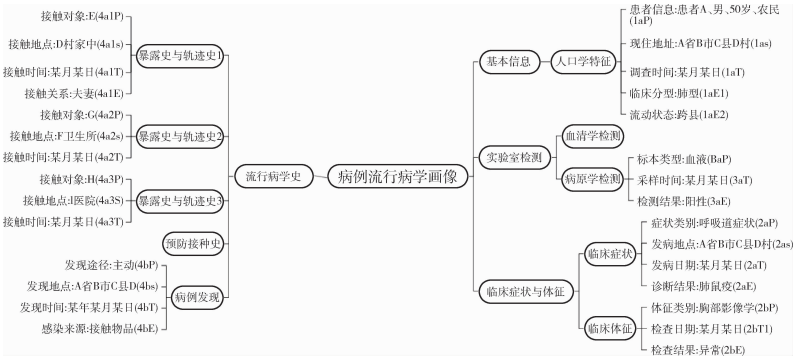

图 4 病例流行病学画像应用示范（已隐藏未取值标签）

4.2 发掘病例间的关系

在构建个案画像的基础上，通过对两个病例暴露史及接触史中有联系的标签进行组合碰撞，可发掘其关联关系，提升分析效率。利用本研究构建的标签体系，对病例 A 和病例 B 进行流行病学画像，通过标签碰撞发掘出病例间关系，虚线框内为两病例共同标签交汇信息，见图 5。

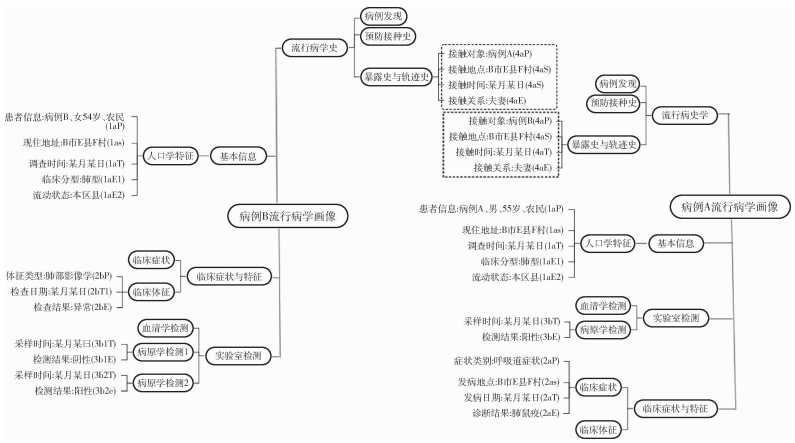

图 5 病例间相互关系分析应用示范（已隐藏未取值标签）

4.3 传播链分析

在发掘两病例间关系的基础上，可进一步拓展多个病例间的传播关系，辅助绘制传播链，为防控传染病和排查密切接触人群提供指导。对模拟的一起病例进行流行病学画像，通过寻找共同轨迹史中的类目标签建立关联关系，形成的传播链示意图，见图 6。

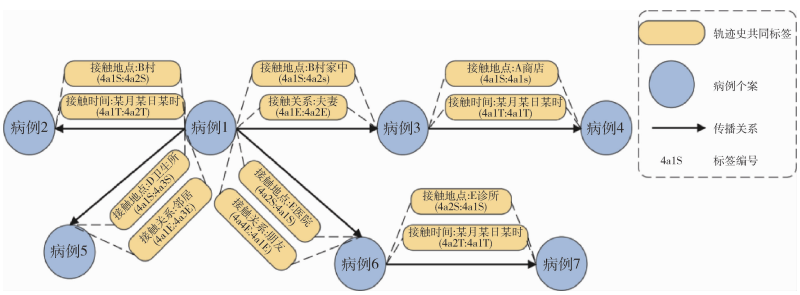

图 6 传播链分析应用示范（仅显示共同轨迹史标签信息）

5 结论

本研究从流行病学调查信息的结构化处理出发，以呼吸道传染病病例个案调查表为研究对象，按照基于个案调查表中流调指标构建个案流行病学标签体系的技术路线，应用等级列举式分类法和分面组配式分类法构建形成了包括 4 个 1 级标签、8 个 2 级标签和 35 个类目标签共 47 个标签的个案流行病学画像标签体系，其中通过流调表标签化 37 个，通过活动要素模型得到 4 个，通过分面分析拓展得到 6 个。3 个示范场景应用表明，本研究构建的标签体系兼具稳定性和可扩展性，能够为非结构化流行病学调查信息的结构化处理提供动态的分级分类规范化标引，可用于个案流行病学画像、病例间传播及相互关系的信息关联匹配，为进一步构建流行病学知识图谱，实现基于大数据和人工智能技术的智能流调提供信息分级分类的基础。本研究构建的标签体系还存在一定不足。首先，由于无法获得完整、真实的流调数据，仅以个案调查表为研究对象，构建的标签体系还有待在实践中进一步验证。其次，该标签体系主要关注传染病感染及发现过程的关键指标，并未涉及既往史、并发症、毒株分型等现实流调过程中关注较少的标签，有待根据实际工作需要进一步扩展和丰富标签体系。

参考文献

1 周朋辉，高璐，刘静，等．天津市新型冠状病毒肺炎流行病学调查报告质量分析［J］．实用预防医学，2021，28（4）：441－445.

2 赵君珂，张振宇，蔡开裕．基于自然语言处理的医学实体识别与标签提取［J］．计算机技术与发展，2019，29（9）：18－23.

3 刘彪，刘金长．基于用户画像分析预测电费敏感型客户的建模实践［J］．电力大数据，2017，20（8）：20－24.

4 朱白．图书馆读者用户“脸谱”绘制研究［J］．商洛学院学报，2017，31（5）：87－90.

5 李望月，刘瑾，陈娜．大数据技术在乡村画像中的应用研究［J］．大数据，2020，6（1）：99－118.

6 Peuquet D J. It's About Time：A Conceptual Framework for the Representation of Temporal Dynamics in Geographic Information Systems［J］. Annals of the Association of American Geographers，2015，84（3）：441－461.

7 史梦洁，王庆娟，涂莹，等．电力营销客户标签体系分类方法研究［J］．电力需求侧管理，2018，20（2）：51－53.

8 费志勇，赵新力．基于本体驱动的高校网站招生信息分面组配揭示［J］．图书情报工作，2008，52（12）：81－84.

9 熊爱民．《冒号分类法》与《中图法》整体结构比较［J］．贵州教育学院学报（社会科学），2005（1）：85－88.

10 张粲．基于客户画像的 A 银行个贷业务精准营销研究［D］．秦皇岛：燕山大学，2021.

11 原国家卫生和计划生育委员会．卫生信息数据元标准化规则：WST303－2009［EB/OL］．［2009－01－22］．http：//www.gdhealth.net.cn/uploadfile/2016/0809/20160809124454969.pdf.

12 黄亚男．多重网络中的传播动力学研究［D］．上海：华东师范大学，2018.

13 原卫生部．传染性非典型肺炎诊断标准：WS 286—2008［EB/OL］．［2008－02－28］．https：//www.chinacdc.cn/did/jszl/zxwj/bzygf/201508/W020150812366903594032.pdf.

14 原卫生部．鼠疫诊断标准：WS 279－2008［EB/OL］．［2008－02－28］．http：//www.nhc.gov.cn/wjw/s9491/200802/38803/files/d02d5b312e734f189d378fa086a9eb5f.pdf.

15 国家卫生健康委员会．炭疽诊断：WS283—2020［EB/OL］．［2020－04－21］．http：//www.nhc.gov.cn/wjw/s9491/202005/7ab2722f726541a6aae5bf427406764b.shtml.