ICS 35.020
CCS L70

团 体 标 准

T/ISC XXXX—XXXX

医疗健康画像体系构建规范

Specification for construction of healthcare portrait system

（征求意见稿）

在提交反馈意见时，请将您知道的相关专利连同支持性文件一并附上。

XXXX - XX - XX 发布 XXXX - XX - XX 实施

中 国 互 联 网 协 会 发 布

目　　次

前　言

本文件按照GB/T 1.1—2020《标准化工作导则 第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由中国互联网协会提出并归口。

本文件起草单位：

本文件主要起草人：

本文件及其所代替文件的历次版本发布情况为：

——

医疗健康画像体系构建规范

1 范围

本文件规定了医疗健康画像在构建过程需满足的技术要求，标准从医疗健康画像构成要求、应用场景要求、构建质量评价要求等维度规范医疗健康画像的构建过程。

本标准适用于医疗健康画像的构建、应用、质量评估及相关技术研发活动，覆盖医疗机构、医疗健康数据服务机构、医疗科技企业、公共卫生管理部门等相关单位，可作为医疗大模型训练推理、精准诊疗支持、疾病风险评估、健康档案管理、公共卫生决策等场景中医疗健康画像规范化应用的依据。

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文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

WS/T 363-2023 卫生健康信息数据元目录

WS/T 364-2023 卫生健康信息数据元值域代码

WS/T 846-2024 医院信息平台交互标准

3 术语和定义

3.1

医疗健康画像 healthcare portrait

是指将居民个人多维度健康数据进行采集、清洗、整合、分析或推理，形成的全方位、数字化、标签化的健康模型。

3.2

事实医疗健康画像 factual healthcare portrait

是指是指将居民个人多维度健康数据进行采集、清洗、整合与分析后，基于真实医疗数据形成的健康模型。

3.3

推理医疗健康画像 reasoned healthcare portrait

是指在事实医疗健康画像基础上，通过大模型算法整合多维度信息推理后生成的健康模型。

4 符号和缩略语

下列符号和缩略语适用于本文件。

FN：假阴性（False Negative）

FP：假阳性（False Positive）

TN：真阴性（True Negative）

TP：真阳性（True Positive）

5 医疗健康画像构建总体要求

医疗健康画像按照构建对象的范围与使用的不同核心元素分为：个体医疗健康画像、家庭医疗健康画像、区域医疗健康画像。基于画像的形成逻辑，分为完全基于核心元素的事实医疗健康画像，与结合大模型算法的推理医疗健康画像。

医疗健康画像构建流程：按通用要求构建画像基础实例，由健康画像对象主体确定核心元素，按照对应核心元素及其值域要求，由各元素构建出事实画像，在此基础上按需结合大模型生成推理画像。

6 医疗健康画像构建规范

6.1 个体医疗健康画像构建

6.1.1 核心元素

6.1.1.1 基本信息类元素

个体医疗健康画像包含的基本信息类元素如表1所示，主体与维度为示例，应包括但不限于以下主题与指标：

表 1 基本信息类元素

主题重要度	主题	维度
核心	基础属性	性别、年龄、出生日期、血型...
	家族史	疾病名称、家庭关系...
	吸烟情况	吸烟状态、吸烟年龄、吸烟时长...
	戒烟情况	戒烟状态、戒烟时长...
	饮酒情况	饮酒状态、饮酒时长、饮酒频率...
	戒酒情况	戒酒状态、戒酒时长...
	过敏史	过敏原、过敏状态...
	月经史	月经周期、经期天数、末次月经时间...
	婚育史	婚姻状态、孕产次...
	…	…
其他	饮食情况	规律饮食、食物类型、盐摄入量...
	运动情况	运动方式、运动频率、运动时长...
	接触史	危险因素接触情况、接触时长...
	预防接种	疫苗名称、接种时间...
	…	…

6.1.1.2 健康状态类元素

个体医疗健康画像包含的健康状态类元素如表2所示，主体与维度为示例，应包括但不限于以下主题与指标：

表 2 健康状态类元素

主题重要度	主题	维度
核心	一般情况	睡眠情况、精神情况...
	身体特征	特征指标、特征结果...
	体征	生命体征项、生命体征结果、查体体征项...
	症状	症状名称、症状时间、症状时长...
	疾病	疾病名称、疾病时长、疾病状态...
	检查	检查名称、检查结果、检查时间...
	检验	检验明细、检验结果、检验时间...
	药品	药品名称、用药频次、用药开始时间...
	手术	手术名称、术后并发症、手术时间...
	…	…
其他	评估	评估类型、评估名称、评估结果...
	操作	操作名称、操作时间、操作并发症...
	诊疗建议	复查条件、饮食指导、健康评价...
	…	…

6.1.1.3 诊疗服务类元素

个体医疗健康画像包含的诊疗服务信息类包含所有健康状态类元素，同时额外包含如表3所示主题与维度，主题与维度为示例，应包括但不限于以下主题与指标：

表 3　诊疗服务信息类额外元素

主题重要度	主题	维度
核心	就诊信息	就诊类型、就诊机构、就诊时间...
	…	…
其他	输血	输血成分、输血日期、输血反应...
	辅助康复治疗	辅助康复治疗名称、辅助康复治疗并发症...
	会诊信息	会诊意见、会诊时间...
	…	…

6.1.2　构建方法

个体医疗健康画像构建以单一个体的全维度健康数据为核心，通过自然语言处理直接提取信息或通过基于规则的数据预处理流水线完成数据分析并汇总为事实画像，后续通过推理，形成精准的个体医疗健康画像：

a) 应支持采集个体的人口学特征、临床诊疗记录、电子健康档案、体检报告、可穿戴设备监测数据等多源数据；
b) 应支持结构化数据与非结构化数据（如病历文本、医生手写医嘱、检验报告描述性等）的兼容接入；
c) 应支持采用命名实体识别技术，从病历、医嘱等文本中提取疾病名称、用药类型、检查部位、症状描述等核心医疗实体；
d) 宜支持借助关系抽取技术，建立医疗实体间的关联关系，例如构建“症状 - 疾病 - 用药”的关联链条；
e) 应支持通过文本归一化技术，统一非标准化表述，例如统一对齐为符合 ICD-10 术语，实现不同医疗机构诊断术语的对齐；
f) 应支持剔除重复诊疗记录，修正可穿戴设备监测数据的异常值，填补人口学特征的缺失值；例如剔除同一时间重复上传的血压数据，采用均值法填补缺失的血糖监测值；
g) 应支持对结构化数据进行标准化编码，对性别、血型等分类变量做标签编码，对血压、血糖等数值变量做归一化处理；系统宜支持将处理后的结构化数据，按照“分类 - 主题 - 维度 - 值域”四层结构进行映射归类；
h) 应支持整合全量标准化数据生成基础画像，涵盖个体全生命周期健康特征；系统宜支持过滤、提炼最新健康数据生成个体医疗健康画像，精准反映个体当下健康状态；
i) 应支持在静态画像基础上引入医学推理模型，结合遗传史、基因数据等信息，计算疾病发生的风险比值（odds ratio）等指标。

6.2　家庭医疗健康画像构建

6.2.1　核心元素

除个体医疗健康画像包含的核心元素外，家庭医疗健康画像应额外包含表4所示主题与维度，主体与维度为示例，应包括但不限于以下主题与指标：

表 4　家庭医疗健康画像额外主题与维度

主题重要度	主题	维度
核心	基础属性	家庭特殊人群、家庭居住环境...
	家庭属性	家庭生命阶段、健康风险类型、家庭健康等级...
	生活属性	家庭饮食结构、家庭运动习惯...
	就诊信息	家庭就医频度、家庭就医成员、家庭医疗支出...
	药品	家庭药品储备、药品储备方式...
	…	…

6.2.2　构建方法

家庭医疗健康画像构建以家庭成员个体画像为基础，通过自然语言处理直接提取信息或通过基于规则的数据预处理流水线完成数据转化并汇总为静态画像，通过提取家庭共性特征、联合推理健康风险，形成家庭医疗健康画像：

a) 应支持采集并汇总家庭所有成员的个体医疗健康画像，明确各成员的健康状态、诊疗记录、生活习惯等核心信息；
b) 宜支持不同医疗机构诊疗数据进行入参协议转换；
c) 应支持采用文本分类技术，对家庭成员的饮食、运动、作息等非结构化描述文本进行分类；
d) 宜支持运用实体链接技术，对齐不同成员的共病信息；
e) 宜支持通过聚类算法对家庭成员的饮食偏好、运动频率等数据进行特征聚合，生成“家庭饮食结构”“家庭运动习惯”等聚合维度；
f) 应支持基于标准化的成员个体数据和家庭共性特征，开展联合推理分析（如结合多人相似的消化道症状，推断食物中毒或传染病风险；依据遗传病史，评估家族性疾病的发病概率）；
g) 应支持输出涵盖家族健康风险、共患病管理建议、家庭医生服务需求等内容的医疗健康画像；
h) 宜支持家庭下一代遗传病风险预测家庭画像。

6.3 区域医疗健康画像构建

6.3.1 核心元素

除个体医疗健康画像与家庭医疗健康画像包含的核心元素外，区域医疗健康画像应额外包含表5所示主题与维度，主体与维度为示例，应包括但不限于以下主题与指标：

表 5 区域医疗健康画像额外主题与维度

主题重要度	主题	维度
核心	自然环境	地理地貌、气候条件…
	区域风险	风险类型、人群占比、风险趋势…
	医疗机构分布	医院、诊所等数量、级别、分布位置…
	医疗人员配备	医生、护士等数量、专业结构、职称水平…
	医疗卫生政策	医保覆盖范围、报销比例、公共卫生政策…
	传染病流行情况	染病种类、发病率、流行季节…
	慢性非传染病发病率	高血压、糖尿病等发病率、患病率…
	地方病情况	地方病种类、病因、分布范围…
	…	…

6.3.2 构建方法

区域医疗健康画像构建以区域内个体及家庭医疗健康画像为数据源，应通过自然语言处理与数据预处理技术完成大规模数据治理，聚焦公共卫生需求开展统计分析与关联推理，形成服务于区域健康决策的医疗健康画像：

a) 应支持整合区域内所有个体和家庭的医疗健康画像数据，建立区域健康数据池，数据类型应覆盖个体医疗健康画像、家庭医疗健康画像、流行病学调查报告、区域医疗资源数据；
b) 应支持采用批量命名实体识别技术，对区域内的流行病学调查报告、各层级医疗健康画像进行批量处理，提取发病时间、地点、人群、病因等核心实体；
c) 宜支持借助自然语言处理技术或数据处理流水线，从各级医疗健康画像中识别聚集性发病事件；
d) 应支持处理跨机构、跨区域数据的格式差异；
e) 应支持结合个体与家庭的接触史等数据，开展流行病学关联分析，梳理疾病传播链，判断流行病的发展趋势；
f) 应支持生成涵盖区域疾病防控重点、医疗资源配置建议、公共卫生干预策略等内容的医疗健康画像。

7 健康画像构建质量控制要求

7.1 完整性质控

7.1.1 质控指标

完整性反应目标画像维度填充的全面性，使用召回率进行评估，指目标画像所有应被填充的维度中，被正确填充的比例，计算公式如下：

$$Recall = \frac{TP}{TP+FN}$$

式中：

Recall——召回率；

TP——真阳性数量，画像中所有应被填充且实际被填充的维度数；

FN——假阴性数量，画像中所有应填充但未填充的维度数。

7.1.2 指标阈值要求

健康画像元素召回率应大于85%。

7.1.3 质控方法及流程

健康画像构建的完整性可采用分层抽样法进行评估，按照个体、家庭、区域画像类型在真实世界的占比，根据比例抽取对应的样本，保障评测数据集符合实际分布，评估流程整体如下：

a) 根据分层抽样法，按要求完成评测数据集构建；
b) 按照健康画像构建规范，组织医生专家，对评测数据集的目标画像维度进行全面填充构建；
c) 按照不同评测对象的调用方式，对评测数据集进行预处理，通过 API 或 Web 等形式获取待评测对象的回复结果；
d) 通过对比医生专家的标注结果和待评测对象的回复结果，得到健康画像构建的召回率；
e) 数据域不完整时(如专科医院、社区医院且数据流通不佳场景等特殊情况)健康画像元素召回率阈值要求宜通过相关领域专家意见重新确定。

7.2 准确性质控

7.2.1 质控指标

准确性反应画像维度填充的准确率，使用准确率进行评估，指目标画像实际填充的维度中，正确填充维度数所占比例，计算公式如下：

$$P = \frac{N_c}{N}$$

式中：

P——准确率；

N_c——画像中实际填充且准确的维度数；

N——画像实际填充的维度总数。

7.2.2 指标阈值要求

健康画像元素准确率应符合以下要求：

——核心元素准确率≥95%；

——其他元素准确率≥90%。

7.2.3 质控方法及流程

健康画像构建的准确性可采用分层抽样法进行评估，按照个体、家庭、区域画像类型在真实世界的占比，根据比例抽取对应的样本，保障评测数据集符合实际分布，估流程整体如下：

a) 根据分层抽样法，按要求完成评测数据集构建；
b) 按照健康画像构建规范，组织医生专家，对评测数据集的目标画像维度进行全面填充构建；
c) 按照不同评测对象的调用方式，对评测数据集进行预处理，通过 API 或 Web 等形式获取待评测对象的回复结果；
d) 通过对比医生专家的标注结果和待评测对象的回复结果，得到健康画像构建的准确性。

7.3 一致性质控

7.3.1 质控指标

一致性反应画像使用术语是否统一规范，避免因术语差异导致的误解，使用一致率进行评估，指画像填充中使用术语项数中，使用规范术语的项数占比，计算公式如下：

$$C = \frac{N_{consistent}}{N}$$

式中：

C——一致率；

$N_{consistent}$——画像中实际填充且使用规范术语的维度数；

N——画像实际填充的维度总数。

7.3.2 指标阈值要求

健康画像整体元素一致率应大于等于90%。

7.3.3 质控方法及流程

健康画像构建的一致性可采用包括但不限于如下方法，评估流程整体如下

a) 医生专家通过 ICD-10、资源梳理扩充等形式，总结形成术语映射表；
b) 针对评测对象输出的画像名称，在术语映射表中进行检查，确认是否属于规范术语，统计得到健康画像构建的一致率；
c) 针对不一致的画像名称描述，应当由医生专家定期进行审视，评估是否需要融合至术语映射表，确保术语映射表的全面性与时效性。

7.4 合理性质控

7.4.1 质控指标

合理性反应健康画像推理内容的结果是否合理，使用合理率进行评估，指画像中涉及推理内容的结果合理的占比，计算公式如下：

$$P_{reasonable} = \frac{N_{reasonable}}{N_{reasoned}}$$

式中：

$P_{reasonable}$——合理率；

$N_{reasonable}$——画像中推理内容中结果合理的数量；

$N_{reasoned}$——画像中存在推理内容结果的总数。

7.4.2 指标阈值要求

健康画像推理元素合理率应大于等于90%。

7.4.3 质控方法及流程

健康画像构建的合理性可采用分层抽样法进行评估，按照个体、家庭、区域画像类型在真实世界的占比，根据比例抽取对应的样本，最终保障评测数据集符合实际分布，估流程整体如下：

a) 根据分层抽样法，按要求完成评测数据集构建；
b) 按照不同评测对象的调用方式，对评测数据集进行预处理，通过 API 或 Web 等形式获取待评测对象的回复结果；
c) 针对涉及推理内容的画像结果，提交医生专家按照健康画像构建规范进行合理性审核，得到推理性健康画像构建的合理率。