ICS 35.240.80

# 团　体　标　准

T/BISSC 01—2022

# 专科疾病标准数据集建设规范

Specifications for the construction of specialized disease standard datasets

2022-12-08 发布　　2022-12-08 实施

四川生物信息学会　发布

# 目 次

# 前　言

本标准按照 GB/T　1.1—2009《标准化工作导则　第 1 部分：标准的结构和编写》给出的规则起草。

本标准由四川生物信息学会提出并归口。

本标准起草单位：四川大学华西医院、四川大学、电子科技大学、湖南大学、中国医学科学院生物医学工程研究所、四川久远银海软件股份有限公司、杭州锘崴信息科技有限公司、北京元影科技有限公司。

本标准主要起草人：张伟、殷晋、曾筱茜、陈蕾、邝俊、刘忠禹、彭玉兰、钟晓蓉、袁勇、段磊、罗婷、刘晶焰、黄娟、秦科、彭绍亮、蒲江波、李春漾、胡耀、应志野、张超、王俊人、蒋静文、朱亭西、陈一龙、邱甲军、辜永红、王国泰、樊琪琪、丁雪峰、丁雪、段贵多、王增、杨波、王爽、甄浩。

## 1 范围

本标准规范了可支撑专科疾病标准数据集构建方法，包括数据采集、治理、标注、质控及应用等。

本标准适用于医疗机构、研究机构、企业等专病数据集（或数据库）设计、研发和管理，其他相关领域可参考使用。

## 2 规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅注日期的版本适用于本文件。凡是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。

GB/T 25000.12-2017 系统与软件工程 系统与软件质量要求和评价(SQuaRE) 第 12 部分：数据质量模型

GB/T 35295-2017 信息技术 大数据 术语

GB/T 34960.5-2018 信息技术服务 治理 第 5 部分：数据治理规范

GB/T 39725-2020 信息安全技术 健康医疗数据安全指南

GB/T 5271.28-2001 信息技术 词汇 第 28 部分：人工智能 基本概念与专家系统

WS 445-2014 电子病历基本数据集

WS/T 671-2020 国家卫生与人口信息数据字典

T/CESA 1109-2020 智能医疗影像辅助诊断系统技术要求和测试评价方法

T/CMDA 001-2020 肝胆疾病标准数据规范：肝癌 CT/MRI 影像采集和处理标准

T/CMDA 002-2020 肝胆疾病标准数据规范：肝癌 CT/MRI 影像标注和质控标准

## 3 术语和定义

下列术语和定义适用于本文件。

### 3.1 数据 data

信息的可再解释的形式化表示，以适用于通信、解释或处理。

[GB/T 25000. 12-2017，定义 4. 2]

### 3.2 特征 features

能表达模式本质的功能或结构特点的度量属性，比如大小、纹理、形状、表现等。好的特征能使同类模式的数据聚集、不同类模式的数据分离。

[计算机科学技术名词 ISBN 978-7-03-059487-7，08. 0386]

### 3.3 数据质量 data quality

在指定条件下使用时，数据的特性满足明确的和隐含的要求的程度。

[GB/T 25000. 12-2017，定义 4. 3]

### 3.4 数据集 data set

数据记录汇聚的数据形式。
[GB/T 35295-2017，定义 2.1.46]

3.5 数据清洗 data cleaning

检测和修正数据集中错误数据项，以及对数据进行平滑处理等操作的数据预处理过程。
[计算机科学技术名词 ISBN 978-7-03-059487, 07.0392]

3.6 数据治理 data governance

数据资源及其应用过程中相关的管控活动、绩效和风险管理的集合。
[GB/T 34960.5-2018，定义 3.1]

3.7 数据采集 data acquisition

数据由生产装置按照数据采集规范生成，以数字化格式存储并传输到对应的目标系统的过程。

3.8 数据脱敏 data masking

对个人敏感信息通过去标识化或匿名化，实现敏感隐私数据的可靠保护。

3.9 数据标注 data annotation

对数据进行人工判断和标识，建立参考标准的过程。

3.10 标注任务 annotation task

按照数据标注规范对指定数据集进行数据标注的过程。

3.11 标注规则 annotation instruction

数据需求方用于明确标注任务和标注数据的操作规范，应包含标注对象定义、所用标注工具和标注平台、标注前的准备工作、标注后的处理工作等。

3.12 标注流程 annotation process

产生标注结果所需要遵循的步骤。

3.13 仲裁 arbitration

在标注人员对原始数据的标注结果不一致时用于决定最终结果的过程。

3.14 仲裁方式 arbitration method

在标注人员对原始数据的标注结果不一致时用于决定最终结果的方式。

### 3.15 人员考核 personnel examination

为保证标注人员/仲裁人员的能力与标注要求一致的测试过程。

### 3.16 数据建模 data modeling

对现实世界各类数据的抽象组织，确定数据库需管辖的范围、数据的组织形式等直至转化成现实的数据库

### 3.17 训练集 training set

用于模型学习和拟合的数据集。

### 3.18 调优集 tuning set

用于对模型参数进行调整并对模型性能进行初步评估的数据集，也称验证集（validation set）。

### 3.19 测试集 test set

用于最终评估模型性能以及泛化能力的数据集。

### 3.20 人工智能 artificial intelligence

表现出与人类智能（如推理和学习）相关的各种功能的功能单元的能力。
[GB/T 5271.28-2001，定义 28.01.02]

### 3.21 机器学习 machine learning

功能单元通过获取新知识或技能，或通过整理已有的知识或技能来改进其性能的过程
[GB/T 5271.28-2001，定义 28.01.21]

## 4 缩略语

AI：人工智能（Artificial Intelligence）
CT：计算机断层成像（Computed Tomography）
ROC：接受者操作特性曲线（Receiver Operating Characteristic Curve）
MRI：核磁共振成像（Magnetic Resonance Imaging）
TXT：文本格式（Text）
NLP：自然语言处理（Natural Language Processing）
ETL：数据仓库技术（Extract-Transform-Load），用来描述将数据从来源端经过抽取（extract）、转换（transform）、加载（load）至目的端的过程
SQL：结构化查询语言（Structured Query Language）
OCR：光学字符识别（Optical Character Recognition）
ODS：操作数据存储（Operational Data Store）

DW：数据仓库（Data Warehouse）
ADS：数据应用层（Application Data Service）
MPP：大规模并行处理系统（Massively Parallel Processing)
OLTP：联机事务处理过程（On-Line Transaction Processing）
CURD：处理数据的基本原子操作原子操作（Create、Update、Retrieve、Delete）
API：应用程序编程接口（Application Programming Interface）

## 5 数据集建设

数据集建设架构在设备设施支撑的基础上，通过数据建模、采集集成形成贴源数据层，通过数据治理、数据融合、数据质控、数据安全管理形成数据仓库层，向外提供 API、可视化、搜索引擎等服务，支撑 AI 智能应用的研发与评价，整个建设过程中由行业标准和数据安全作为保障，如图 1：


图 1 专病数据集建库架构图

数据集建设主要包含数据建模、数据采集、数据治理、数据存储和数据安全五个方面。

### 5.1 数据建模

数据建模简单来说就是基于专病数据集的应用需要，将各种多源异构的数据进行整合和关联，形成新的数据资源目录。

临床业务系统的数据模型是根据业务需求而设计，围绕事件活动而展开，且因厂家不同而标准不同，导致各个业务系统无法直接被整合和利用。而专病数据集的数据模型是围绕患者、围绕疾病、围绕临床应用场景而展开的，需充分考虑科研、AI 研发、医疗技术临床使用管理和资源优化配置等等实际应用

需求，结合临床医生的经验、参考诊断学、临床数据采集等相关标准进行建模设计。形成以患者主索引（EMPI）为基础的病患数据管理，增强数据的可用性和可读性，让使用方能快速地获取到自己关心的有价值的数据并且及时地做出响应。

数据集应具有充分的多样性，以提高算法模型的泛化能力和准确性，为保证数据集的多样性，在数据模型设计阶段需尽可能的纳入更多具有通用性的统计维度，同时降低数据集的覆盖偏移，需考虑不同地区、不同临床机构在人群组成、临床疾病特征、数据采集设备、操作流程规范性等方面的差异。

具体的建模步骤如图 2：


**图 2 数据建模流程图**

### 5.1.1 临床讨论

制定数据模型之前，应召集临床医护人员（专病数据集使用人员、管理人员等）与数据建模工程师一起讨论，结合专病数据集的建库目标，共同确定数据范围。如研究方向需覆盖雾霾等环境污染对专科疾病发病率的影响，则除了采集患者临床诊断相关数据，还需采集环保部门和气象局的数据；如需研发疾病 CT 智能筛查，则需采集 CT 影像数据；如需跟踪患者预后情况，则需采集患者出院后的随访数据；等等。

临床参与人员需符合以下资质要求：

a) 二甲医院的主任医师、或三甲医院的主治医师及以上级别；

b) 主导过市级以上临床科研项目；

c) 建议多学科医师团队共同参与，如肾脏病专病数据集建设需肾脏内科、泌尿外科、病理科等学科团队参与。

数据建模工程师需符合以下资质要求：

a) 熟悉医疗业务，以及各临床业务系统的数据构成；

b) 熟练掌握 Oracle、Mysql、PostgreSql 等主流关系型数据库，HBase、NebulaGraph 等非关系型数据库，以及影像、文件等存储技术；

c) 熟悉数据仓库各类建模理论、数据仓库数据层级关系；

d) 了解基本算法和至少一种数据建模工具；

e) 具备数据集构建 3 年以上工作经验。

### 5.1.2 数据分层、分类设计

在专病数据集体系中，采用自下而上划分为 3 个层级：操作数据存储层、数据仓库层、应用数据层。如图 3 所示。

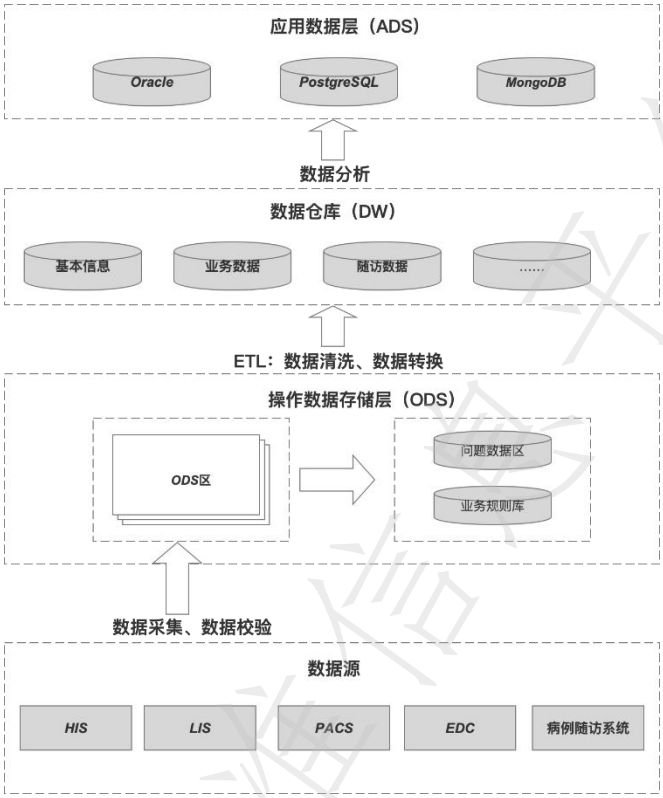


图 3 数据分层设计

结合专病特型，数据可按照基础数据、临床数据、随访数据、生物信息数据、外部数据等维度进行分类，如图 4：

| 基础数据 | 临床数据 | | 随访数据 | 外部数据 | 生物信息数据 |
|---|---|---|---|---|---|
| 患者基本信息 | 社会人口学 | 门（急）诊信息 | 随访基本信息 | 环境数据 | 基因序列 |
| 医疗机构信息 | 暴露危险因素 | 检验信息 | 随诊随访 | 气候数据 | 蛋白质序列 |
| 科室信息 | 月经生育史 | 检查信息 | 治疗随访 | 医保数据 | 基因组 |
| 医务人员信息 | 住院信息 | 手术信息 | 伴随疾病随访 | …… | 蛋白质结构 |
| 医疗项目信息 | 诊断信息 | 护理信息 | 生存状态随访 | | …… |
| 医疗设备信息 | 医嘱信息 | 疾病进展与转归 | 转移及其他情况 | | |
| …… | …… | | …… | | |

图 4 专病数据分类

### 5.1.3 数据库概念模型设计

根据专病数据集的实际应用目标，参考《电子病历基本数据集》等标准与规范，确定专病数据集的数据范围。

数据库概念设计应符合以下要求：

a) 不受数据来源限制，充分考虑专病 AI 应用研发需求；

b) 内容满足专病专业医护的研发方向；

c) 数据概念的颗粒度不宜过细，可参考以一种医治/就诊行为事件为一个概念，如门诊、住院、手术、护理、检查、检验等；

d) 将概念进行细化，抽象出具体的实体和实体属性，以及实体与实体之间的关系。

如图 5：

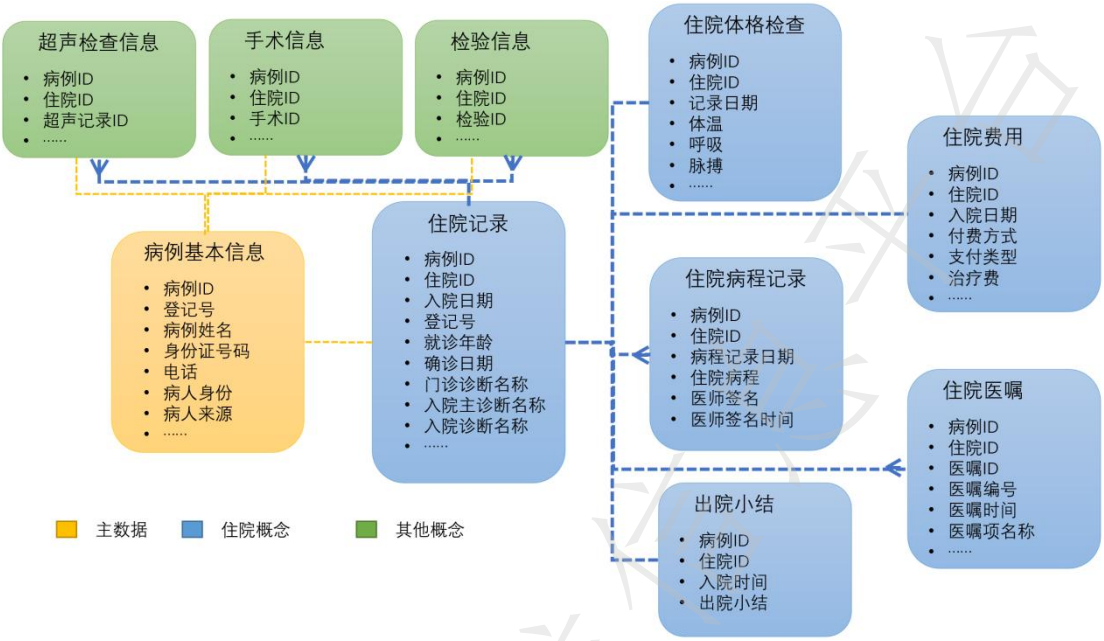


图 5 专病集部分概念模型设计图

### 5.1.4 数据库逻辑模型设计

数据逻辑模型设计是将概念模型在数据库中以表结构的方式呈现出来，形成专病数据集最终的数据结构目录，包含以下步骤：

a) 数据库选型，一般数据量的专病数据集存储可选择 Mysql、PG 等关系型数据库，TB 级大型专病数据集可选择 HBase、Redis 等非关系型数据库；

b) 生成表、字段、主键、外键及其他数据对象，包括视图、序列、索引、约束以及函数、触发器、存储过程等；

c) 详细定义字段的数据类型、长度、是否必须，以及标准编码等。

专病数据逻辑模型如图 6（以乳腺癌为例）：

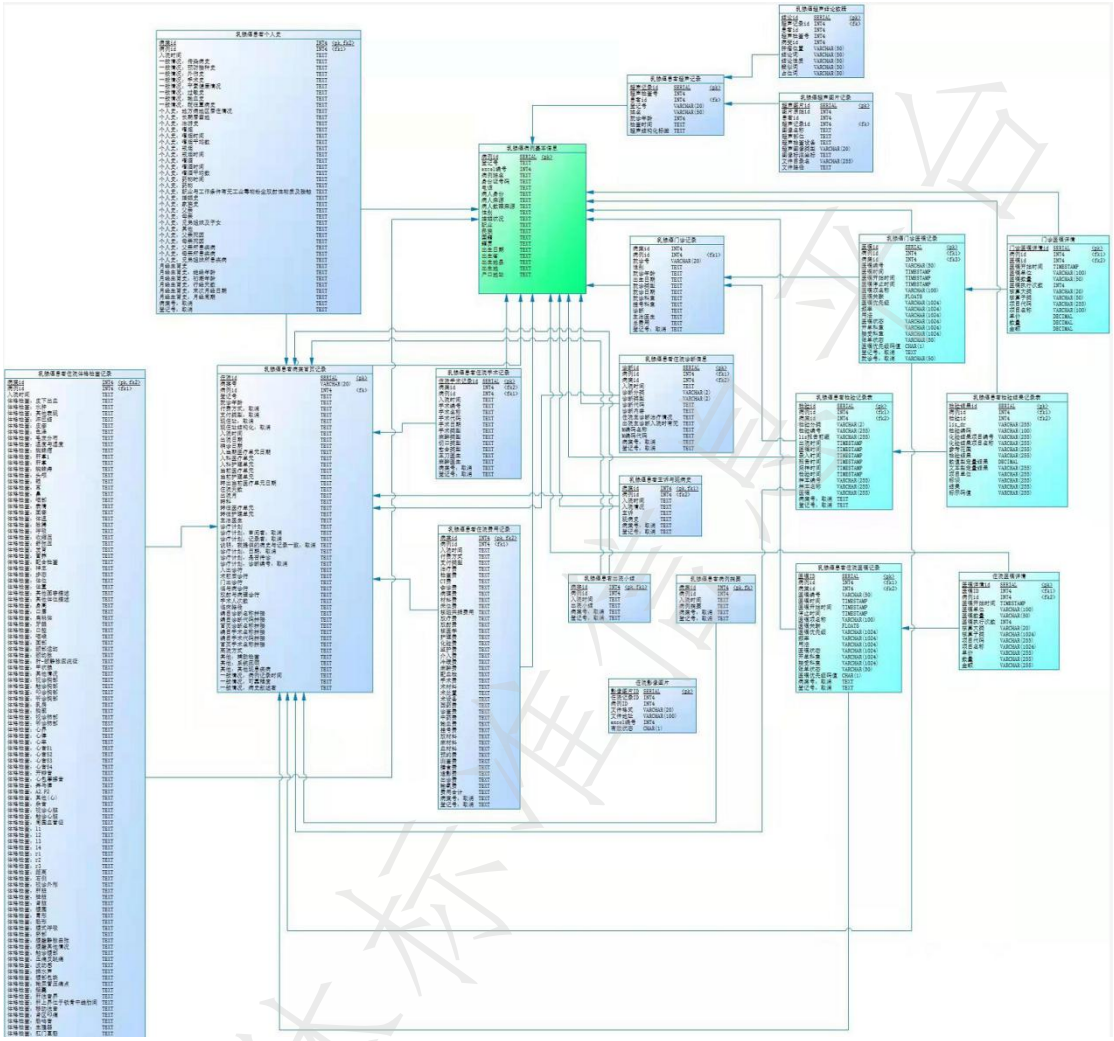

图 6 乳腺癌专病集数据逻辑模型计图

### 5.1.5 数据库命名规范

数据表和字段的命名规范应满足以下要求：

a) 表和字段的命名均禁止使用数据库关键词和保留词。

b) 表和字段的标识符由英文字母、下划线、数字构成，首字符应为英文字母。

c) 表名称长度原则上不超过 40，字段名称长度原则上不超过 30。

d) 表和字段的标识符是中文名称关键词的英文翻译，可采用英文译名的缩写命名（ODS 层例外）。

e) 按照中文名称提取的关键词顺序排列关键词的英文翻译，关键词之间用下划线分隔；缩写关键词一般不超过四个，后续关键词应取首字母。

f) 表和字段的标识符采用英文译名缩写命名时，单词缩写主要遵循以下规则：

1) 英文关键词有标准缩写或行业通用缩写的应直接采用。如中国 CHINA 可缩写为 CHN。

2) 没有标准缩写的，取单词的第一个音节，并自辅音之后省略。

3) 若出现中文同义词或英文译名缩写相同时，参考压缩字母法或取中文拼音首字母等常见缩写方法以区分不同关键词。

4) 若关键词本身翻译简洁，则可以不缩写，如名称或姓名使用 NAME，但关键词进行组合时

需要缩写，如单位名称，则使用 ORGAN_NM 表示。

g) 相同的实体和实体特征在要素类表、关系类表、属性类表中应采用一致的标识。

### 5.1.6 元数据设计

元数据是描述数据的数据，其使用目的在于识别数据、评价数据、追踪数据在使用过程中的变化，是数据资源管理的重要手段，元数据的内容项应包含数据项的名称、编码、类型、长度、业务含义、数据来源、质量规则、安全级别、域值范围等，以及数据项之间的关联关系。

### 5.1.7 数据标准规范

数据标准的建立将充分借鉴行业标准，结合医疗健康的行业规范及专病对数据的实际应用要求，数据标准规范建议参考如下：

a) 遵守国家、行业标准代码规范，部分规范见表 1；

**表 1 部分国家/行业代码标准**

| 表号 | 标准分类 | 代码表名称 |
|---|---|---|
| 1 | 卫生部标准 | 年龄（段）代码表 |
| 2 | 国家标准 | 人的性别代码(GB/T2261.1-2003 ) |
| 3 | 国家标准 | 世界各国和地区名称代码表(GB/T 2659-2000 ) |
| 4 | 国家标准 | 中华人民共和国行政区划代码表(GB/T 2260-2002 ) |
| 5 | 国家标准 | 职业分类与代码表(GB/T 6565-1999) |
| 6 | 国家标准 | 专业技术职务代码表（GB/T 8561-2001） |
| 7 | 国家标准 | 政治面貌代码表（GB/T 4762-1984） |
| 8 | 国家标准 | 婚姻状况代码（GB/T 2261.2－2003） |
| 9 | 国家标准 | 文化程度代码表（GB/T 4658-1984） |
| 10 | 国家标准 | 学位代码表 (GB/T 6864-2003) |
| 11 | 国家标准 | 民族代码表（GB/T3304-1991） |
| 12 | 国家标准 | 家庭关系代码分类（GB/T 4761-1984） |
| 13 | 国家标准 | 健康状况代码表（GB/T 2261.3-2003） |
| 14 | 卫生行业标准 | 卫生机构（组织）分类代码表(WS218-2002) |
| 15 | 国家标准 | 疾病分类代码 ICD-10（(GB/T 14396-2001)） |

b) 遵守卫生行业相关标准规范，如《国家卫生与人口信息数据字典》；

c) 遵守院内信息化相关标准规范。

专病集的部分示例数据字段如表 2（以食管癌为例）：

**表 2 食管癌专病集的部分数据字典**

| 序号 | 元素名称 | 代码值 |
|---|---|---|
| 1 | 性别 | 0-未知的性别 1-男性 2-女性 9-未说明的性别（0~9） |
| 2 | 年龄（段） | 年龄（段）代码 |
| 3 | 通讯联系方式类别 | 1-地址 2-邮政编码 3-电话号码(总机/查询台) 4-单位电子邮箱(E-mail) 5-单位网站域名（1~5） |
| 4 | 身份证件类别 | 1-居民身份证 2-军官（文职干部）证 3-护照（1~3） |
| 5 | 医疗档案 | 1-住院病例 2-门诊病历 3-居民健康档案（1~3） |

| | 类别 | |
|---|---|---|
| 6 | ABO 血型 | 1-A 2-B 3-AB 4-O 5-其它（1~5） |
| 7 | Rh 血型 | 1-Rh 阳性 2-Rh 阴性 3-Rh 血型不详（1~3） |
| 8 | 职业 | 职业分类与代码(GB/T 6565-1999) |
| 9 | 从事专业 | 11-执业医师 12-执业助理医师 21-注册护士 31-执业药师 32-执业中药师 33-其他药剂员 41-检验人员 42-影像人员 43-卫生监督员 44-其他卫技人员 60-其他技术人员 70-管理人员（11~70） |
| 10 | 管理职务 | 1-党委(副)书记 2-院(所.站)长 3-副院(所.站)长 4-科室主任 5-科室副主任（1~5） |
| 11 | 专业技术职务（评） | 专业技术职务代码（GB/T 8561-2001） |
| 12 | 医师资格类别 | 1-临床 2-口腔 3-公共卫生 4-中医（1~4） |
| 13 | 工作调动类别 | 调入:11-考试录用 12-军转人员 13-卫生机构调入 19-其他机构调入；调出:21-退休 22-辞职(辞退) 23-自然减员 24-调往卫生机构 29-其他 |
| 14 | 婚姻状况 | 婚姻状况代码（GB/T 2261.2－2003） |
| 15 | 政治面貌 | 政治面貌代码（GB/T 4762-1984） |
| 16 | 文化程度 | 文化程度代码 (GB/T 4658-1984) |
| 17 | 学历 | 1-博士 2-硕士 3-学士/本科 4-大专 5-中专 6-高中 7-初中及以下（1~7） |
| 18 | 所学专业 | 11-基础医学 12-临床医学 13-中医学 14-口腔医学 15-公共卫生 16-护理学 17-药学 18-中药学 31-公共管理 32-人力资源管理 33-信息管理 41-经济学 42-会计学 43-统计学 51-法学 61-信息技术/计算机 62-工程 99-其他（11~99） |
| 19 | 民族 | 民族代码(GB/T3304-1991 ) |
| 20 | 健康状况 | 健康状况代码（GB/T 2261.3－2003） |
| 21 | 恶性肿瘤种类 | 1-胃肿瘤 2-肝肿瘤 3-肺肿瘤 4-食管肿瘤 5-结直肠肛门肿瘤 6-白血病 7-子宫颈肿瘤 8-鼻咽肿瘤 9-女性乳房肿瘤 10-膀胱肿瘤 11-其他（1~11） |
| 22 | 医院等级 | 0-三级特等 1-三级甲等 2-三级乙等 3-三级丙等 4-二级甲等 5-二级乙等 6-二级丙等 7-一级甲等 8-一级乙等 9-一级丙等（0~9） |
| 23 | 医院类型 | 1-综合医院 2-专科医院 3-门诊部 4-其他医院 5-疗养院 6-专科防治所（站） 7-卫生防疫站 8-其他卫生（1~8） |
| 24 | 家庭成员关系 | 家庭关系代码分类（GB/T 4761-1984） |
| 25 | 出生地 | 行政区划代码(GB/T 2260-2002 ) |
| 26 | 病情 | 1-危 2-急 3-一般（1~3） |
| 27 | 疾病发现方式代码 | 1-就诊 2 -体检（1~2） |
| 28 | 疾病分类(ICD-10) | 疾病分类与代码(GB/T 14396-2001) |
| 29 | 诊断依据 | 1-病理（包括骨髓片） 2-脱落细胞（包括血片） 3-手术 4-内窥镜 5- X 光（包括造影） 6-CT 7-核磁共振 8-超声波 9-同位素扫描 10-免疫 11- |

| | | |
|---|---|---|
| | | 生化 12-临床（1~12） |
| 30 | 肿瘤分期 | 1-CT__N__M__ 2-ST__N__M__ 3-PT__N__M__ 4-T__N__M__ 5-aT__N__M__ 99-不详 （1~99） |
| 31 | 费用支付方式 | 1-社会基本医疗保险 2-商业保险 3-自费医疗 4-公费医疗 5-大病统筹 6-新型农村合作医疗 9-其它（1~9） |
| 32 | 经费支出 | 1-医疗支出 2-药品支出 3-财政专项支出 4-其他支出（1~4） |
| 33 | 治疗结果 | 1-治愈 2-好转 3-未愈 4-死亡 5-其他（1~5） |
| 34 | 病案质量等级 | 1-甲 2-乙 3-丙 (1~3) |
| 35 | 费用类别 | 1-床费 2-护理 3-西药 4-中药 5-化验 6-诊察治疗 7-手术 8-检查 9-其他费用(1~99) |

## 5.2 数据采集

### 5.2.1 数据源调研

根据专病数据集模型规划的数据范围，调研各字段所属数据源，明确需要采集的数据源类型及现状。

应首先明确区分数据集模型字段中的回顾性数据和前瞻性数据，不同的数据类型有不同的调研内容和采集方式。

#### 5.2.1.1 回顾性数据调研

回顾性数据指已经发生的医疗行为所产生的数据。

（1）数据一般产生于临床相关业务系统，包括但不限于：

a) CDR/RDR：部分机构已形成 CDR（临床数据中心）和/或 RDR（科研数据中心），可直接从 CDR/RDR 中获取专病患者的临床诊疗相关数据。

b) HIS：专病患者门诊及住院的就诊情况、诊断、医嘱、检查、检验、手术等信息。

c) EMR：专病患者的门诊病历、入院病历、病程、术前讨论、术后情况、手术小结、出院小结、会诊记录等全部文书。

d) 护理：护理首页、护理评估、护理记录、护理措施、危重记录、体征、PICC、置管等。

e) 手术麻醉：麻醉记录单、手术记录单、监护仪器数据等。

f) LIS：专病患者检验相关基本信息、检验项目、检验细项、细项结果及正常值范围。

g) RIS：专病患者影像检查相关基本信息、检查报告、CT/MRI/PET 等各类文字报告原始文件。

h) 病理：专病患者病理检查相关基本信息、检查报告、涂片图像原始文件。

i) 超声：专病患者超声检查相关基本信息、检查报告、超声图像原始文件。

j) 专病患者医疗或者疾病相关的其他系统。

（2）数据也可能来源于其他第三方系统或组织，包括但不限于：

a) 第三方穿戴设备

b) 日常体征监测设备

c) 气候情况

d) 其他医疗机构的临床诊疗数据

e) ……

以上数据的数据获取方式一般包括业务系统数据库对接、服务接口和文件。如果被采集数据源支持数据库直接访问，应优先考虑数据库对接。如果不支持直连数据库，考虑服务接口方式。除此之外，还

可考虑文件导入的方式。

#### 5.2.1.1.1 数据库对接方式相关调研

a) 调研源数据库的运行环境、性能状况及网络状况；

b) 调研源数据库的数据库基本信息和参数配置，例如：数据库软件的版本信息、补丁集，数据库软件的安装情况、数据库的系统日志等等；

c) 调研源数据库的存储空间划分情况，例如：各个逻辑设备的大小、划分情况和存储操作特性，数据库和各个逻辑设备之间的对应关系，各个数据库的配置选项等等；

d) 调研源数据库的库表结构；

e) 调研源数据库的数据指标；

f) 调研源数据库的数据体量，以及每年的数据增长情况；

g) 调研源数据库的备份策略；

h) 调研是否需要前置机。

#### 5.2.1.1.2 服务接口调研

a) 调研接口协议；

b) 根据不同接口协议调研编码格式、提交方法、传参要求等；

c) 调研数据接口传输量；

d) 调研增量数据识别方式和更新方式。

#### 5.2.1.1.3 文件导入方式相关调研

a) 调研数据提供方式，如分布式文件系统、网络文件、线下拷贝等；

b) 调研需对接文件格式，如 excel、txt、csv、图像文件、影像文件、音频文件等；

c) 调研数据结构情况，是否可直接导入数据库，是否需要数据处理程序；

d) 调研数据体量；

e) 调研增量数据识别方式和更新周期。

### 5.2.1.2 前瞻性数据调研

前瞻性数据指以现在为起点追踪到将来情况所记录的数据，一般需要通过随访、临床研究、设备检测等方式人工或自动采集。建议调研内容包括但不限于：

a) 调研前瞻性数据采集表的内容及格式要求；

b) 调研前瞻性数据采集人员要求及相关设备、系统情况；

c) 调研采集流程要求，如被采集对象应符合的条件、采集频率、采集次数、采集过程。

### 5.2.1.3 数据类型调研

除了区分前瞻性和回顾性数据，还应根据数据类型进行调研。

a) 结构化数据通常是由二维表结构来逻辑表达和实现的数据，其严格地遵循数据格式与取值规范，主要通过关系型数据库进行存储和管理。

b) 非结构化数据为无法定义结构的数据。常见的非结构化数据为文本信息，图像信息，视频信息

以及声音信息等等，他们的内容不能用一个固定的结构来描述。针对非结构化数据一般有标注需求，以将其内含的医疗信息提取表达出来。

c) 除了结构化和非结构化数据之外，还需要对半结构化数据进行采集。半结构化数据和前面介绍的两种类型的数据都不一样，它是结构化的数据，但是不遵循关系型数据库或其他数据表相关数据模型的层次结构。在半结构化数据中，同一类的不同实体数据的结构可能会有一定程度的不同，即不同实体所具有的属性会有一定程度的不同，而同时，对于这些实体来说，不同的属性之间的顺序是并不重要的。

### 5.2.2 规定采集数据范围

根据数据集应用目标和数据源库表情况，定义被采集对象要求，如乳腺癌专病数据采集需“出院诊断”包含<乳腺恶性肿瘤>或<乳腺癌>，年龄大于等于 18 岁且小于等于 70 岁等。

如果数据集存在数据有效期要求或限制，也需要定义采集数据的时间范围，如入院日期为 2012-2020 年间的住院数据等。

### 5.2.3 回顾性数据采集

#### 5.2.3.1 采集人员要求

a) 人员选拔

数据采集人员的资质建议要求具备数据库知识和相应开发技能，能通过工具或编程手段汇集数据并管理数据。

b) 人员培训

根据数据采集要求对参与数据采集的人员进行培训。主要包括：相关数据采集流程、采集设备操作培训、操作规范培训、数据安全培训等。

[来源：T/CMDA 001-2020，有修改]

#### 5.2.3.2 采集过程要求

回顾性数据采集是使用服务器作为基础硬件平台，搭建软件系统平台，采用服务总线技术、集群技术、分布式存储技术、分布式计算技术、ETL 技术等制定数据采集标准和处理流程，对回顾性数据实现统一的采集、存储和管理。

数据采集流程图如图 7 所示：

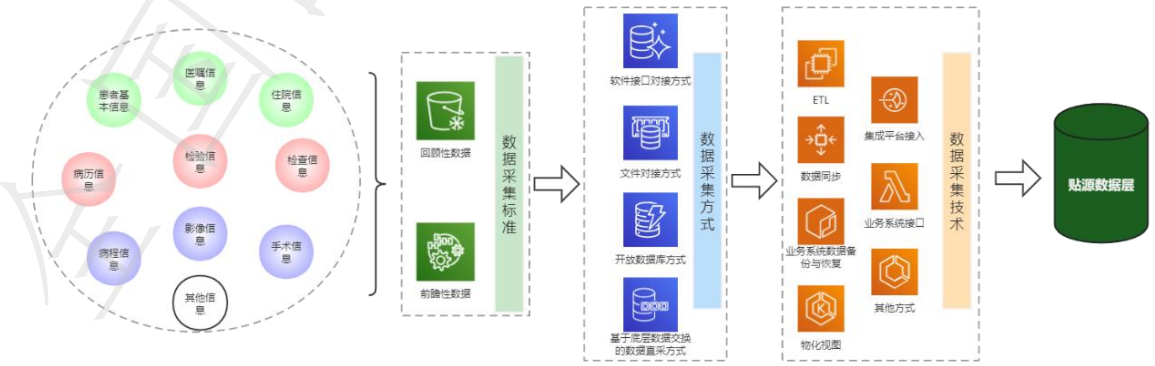


图 7 数据采集流程图

回顾性数据采集流程包括采集对象、数据采集方式、数据采集技术、数据采集结果四个部分。要求

如下：

a) 数据采集的对象数据需要根据专病数据集需求范围进行采集，主要包括患者的基本信息、病例信息、病程信息、医嘱信息、检验信息、检查信息、影像信息、护理信息等。

b) 数据采集方式需要根据具体情况进行选型。数据采集方式包括软件接口对接方式、文件对接方式、开放数据库方式、基于数据库交换的数据直采方式等，如表 3。

**表 3 数据采集方式应用场景介绍**

| 数据采集方式 | 简要介绍 |
| --- | --- |
| 服务接口对接方式 | 各个业务系统直接提供数据接口，可以直接通过变成调用数据接口采集数据。 |
| 文件对接方式 | 专病数据的来源有一部分是 excel 文件、txt 文件、影像文件、图片文件、压缩文件等，针对不同文件类型会采用不同的采集技术。 |
| 开放数据库方式 | 业务系统直接开放数据提供数据采集方采集数据，一般情况下，业务系统开发方是不会直接提供这种方式的。 |
| 基于前置数据库交换的数据直采方式 | 通过获取软件系统的底层数据交换、软件客户端和数据库之间的网络流量包，进行包流量分析采集到应用数据，同时还可以利用仿真技术模拟客户端请求，实现数据的自动写入。 |

c) 数据采集技术也需要根据数据采集方式和实际情况选择一种或多种技术的组合来完成数据采集。数据采集可选择的技术也比较多，例如 ETL、数据同步、业务系统备份与恢复、物化视图、业务系统接口等技术，数据采集技术简介如表 4 所示：

**表 4 数据采集技术介绍**

| 数据采集技术 | 技术简介 |
| --- | --- |
| ETL | ETL 将数据从数据源经过抽取、清洗转换、加载到目的地数据仓库的过程，目的是将分散、凌乱、标准不统一的数据整合到一起。<br>ETL 数据源可以是业务系统数据库，也可以是文件或其他来源。<br>ETL 的实现有多种方法，常用的有三种。一种是借助 ETL 工具实现，一种是 SQL 方式实现，另外一种是 ETL 工具和 SQL 相结合。前两种方法各有各的优缺点，第三种是综合了前面二种的优点，会极大地提高 ETL 的开发速度和效率。建议采用第三种方式。<br>ETL 清洗的主要对象为残缺数据、异常数据、重复数据，需要事先定义详细的清洗规则并按照清洗规则完成数据清洗。 |
| 数据同步 | 数据同步是指数据从业务数据库到贴源数据集的技术，可以通过程序编码实现，也可以在数据库层面实现。一般都是使用数据库厂商提供的发布订阅工具实现，该技术在实现过程中也必须考虑到数据安全问题，需要在数据完成脱敏加密后再同步到贴源数据集。 |
| 业务系统数据备份恢复 | 业务系统可以直接将脱敏和加密后的数据库备份文件提供给数据采集工具，数据采集工具只需要将提供的数据库备份文件进行还原即可完成数据采集工作。 |
| 物化视图 | 物化视图是一个查询结果的数据库对象，它是远程数据的的本地副本，物化视图存储基于远程表的数据，也可以称为快照。 |
| 业务系统接口调用 | 如果业务系统可以提供数据接口，数据采集可以调用业务系统的数据接口进行数据采集 |

d) 数据采集的结果是形成一个贴源数据集，贴源数据集指从数据源导入到数据仓库的第一步。

e) 数据采集可以选择全量数据采集和增量数据采集。全量数据采集是指每次从数据源采集全部数据（包括之前采集过的数据），而增量数据采集则是指只抽取从上次抽取之后数据库中的新增或修改的数据。对于全量数据采集，每次采集之后都需要重新进行数据治理。增量数据采集只需要对增量数据完成数据治理工作。同样需要根据具体情况选择使用全量数据采集或增量数据采集。全量和增量数据采集都必须保证数据的准确性和系统的性能稳定。

#### 5.2.4 前瞻性数据采集

##### 5.2.4.1 采集人员要求

a) 人员选拔

数据采集技师的资质建议要求在三甲医院从事专科疾病相关诊疗工作 5 年以上。

b) 人员培训

根据数据采集要求对参与数据采集的人员进行培训。主要包括：相关数据采集流程、采集设备操作培训，数据安全培训等。

c) 人员考核

采集人员考核标准要求熟悉专病数据采集相关技术要点，能根据不同数据类型及采集需求，获得质量最佳的数据。建议从以下方面进行综合评价：

（1）数据采集规范熟悉程度：可采用书面答题形式，通过设置数据采集规范相关问题，对采集人员回答进行打分和评估；

（2）采集设备操作熟练程度：通过采集人员操作设备过程中的操作合规程度及完成时的操作时间进行综合考量，将过程中的不合规操作作为评估时的罚项；

（3）采集过程数据安全程度：对采集人员进行数据采集过程中发生数据遗失、泄露等安全风险进行评估，按照人员操作的数据安全风险程度对综合评估进行扣分；

（4）采集结果质量合规程度：通过最终操作人员采集到的数据质量（如在采集过程中产生数据失真、噪声、畸变等问题）合乎后续数据使用规范的程度进行评价。

[来源：T/CMDA 001-2020，有修改]

##### 5.2.4.2 采集过程要求

被采集对象应对采集内容知情并同意，采集员应对被采集对象进行采集过程的介绍，如设备检查时应如何配合，被采集对象的口述内容以及检查姿势、状态等要求进行告知以便被采集对象能更好配合完成采集过程，确保所采集数据的真实性、准确性和有效性。

### 5.3 数据治理

[来源：T/CMDA 001-2020，有修改]

#### 5.3.1 数据预处理

##### 5.3.1.1 结构化数据清洗

###### 5.3.1.1.1 结构化数据清洗原则

结构化数据清洗应遵循一定的原则，从而达到较高的数据清洗质量，清洗原则包括：

a) 方法一致性原则：

数据资源清洗加工工作应统一决策，同一数据库范围内工作方法、技术指标均应当统一，从而达成数据产品的一致性。

b) 数据可信性原则

数据可信性包括精确性、完整性、一致性、有效性、唯一性。

c) 数据可用性原则

数据可用性包括时间性、稳定性等。

#### 5.3.1.1.2 结构化数据清洗流程

专病结构化数据清洗流程如图 8 所示：

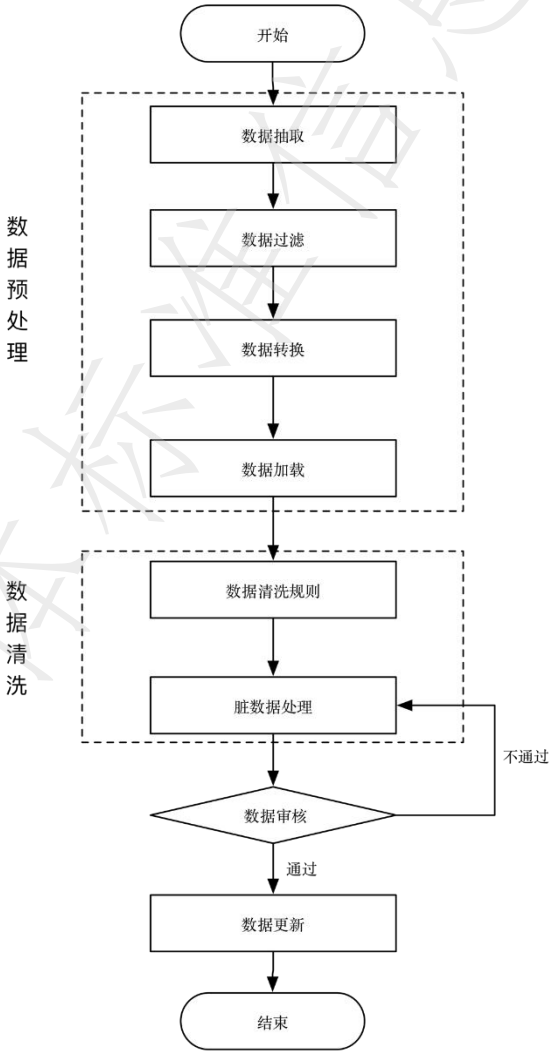


**图 8 结构化数据清洗流程图**

结构化数据清洗流程包括：

a) 数据抽取

从数据库中抽取数据包括全量抽取和增量抽取两种方式，根据具体情况进行选择。

b) 数据过滤

数据过滤要初步实现对业务数据中不符合应用规则或者无效的数据进行过滤操作，确保数据标准统一。

c) 数据转换

数据转换要实现对数据的格式、信息代码等进行转换。

d) 数据加载

数据加载过程进行的主要操作是插入操作和修改操作。将干净数据及脏数据分别插入到不同的数据表中。对于数据加载工作，一般会搭建数据库环境，如果数据量大（千万级以上），可以使用文本文件存储结合脚本程序处理进行操作。

e) 数据清洗

数据清洗过程是指根据现有的数据清理规则对“脏数据”进行数据处理转换，将“脏数据”转化为满足数据质量要求或应用要求的数据的过程。

通用数据清洗规则包括但不限于：

（1）数据格式校验：通过检查表中属性值的格式是否正确来衡量其准确性，如身份证格式、手机号格式、邮箱格式、姓名格式、时间格式、币种格式、乱码检测等。

（2）非空校验：要求字段为非空的情况下，需要对该字段数据进行校验。

（3）主键重复：多个业务系统中同类数据经过清洗后，在统一保存时，为保证主键唯一性，需进行校验工作。

（4）非法代码值清洗：非法代码问题包括非法代码、代码与数据标准不一致等，非法值问题包括取值错误、格式错误、多余字符、乱码等,需根据具体情况进行校核及修正。

（5）记录数校验：指各个系统相关数据之间的数据总数校验或者数据表中每日数据量的波动校验。

f) 问题数据处理

问题数据处理需要处理缺失值数据、错误数据和错误关联数据三种问题数据。

g) 错误数据处理

错误数据处理包含格式内容问题数据和逻辑问题数据两类问题数据的处理。

#### 5.3.1.2 非结构化数据标注

常见的非结构化数据为文本信息、图像信息、视频信息以及声音信息等，结构差异大。针对这类非结构化数据需要进行数据标注，数据标注的质量、全面性、体系统一及标注过程的质量控制体系都将决定数据集的临床可靠性和使用价值。

##### 5.3.1.2.1 标注人员要求

a) 选拔

为保证标注医师的代表性，建议面向全国公开考试选拔标注医师。医师的资质建议要求在三甲医院从事专科疾病相关领域工作 5 年以上，职称为住院医师及以上。

b) 培训

入选的标注医师应当接受培训，统一对标注规则的认识，熟悉标注软件操作。

c) 考核

由临床领域和数据技术领域专家进行审核和把关，从包括但不限于以下方面进行综合考核标注人员素质以及技术水平：

（1）数据标注规则的熟悉程度：通过设置数据标注范围、标注方式及数据安全等规则相关书面问答题的形式，对标注人员对于规则的熟悉程度进行评估；

（2）标注软件的操作熟练程度：通过标注人员操作标注软件进行数据标注的时间以及发生误操作

等事故的频次及错误等级进行综合评估；

（3）操作过程中数据安全程度：通过标注人员标注数据过程中由于操作导致数据发生数据遗失、泄露等数据安全风险的频次和等级进行综合评估；

（4）完成标注的数据合规程度：通过标注人员标注完成后的标注数据质量、标注正确率及与金标准标注数据（一般由临床专家完成标注）一致性等方面进行综合评估。

[来源：T/CMDA 002-2020，有修改]

#### 5.3.1.2.2 标注场所要求

为保证数据安全，标注应于院内非联网或仅连接内部网络的计算机设备上进行。对于超声图像、MRI 图像等影像类数据，应在符合阅片条件的光照环境下进行标注操作。

#### 5.3.1.2.3 标注软件要求

标注工具、平台等软件系统应满足以下要求：

（1）稳定性：保证软件与院内设备系统的兼容性及运行过程中不易发生崩溃、损坏等事件；

（2）易操作性：标注工具应降低标注人员的操作难度，提供交互方式的自有标注；

（3）规范性：标注工具的数据导出格式，应满足或可转换到要求的格式；

（4）高效性：标注工具应保证标注任务的完成效率；

（5）数据安全性：保证数据安全保密，确保软件运行无数据遗失、泄露等安全事故发生；

（6）合规合法性：满足医学领域相关法规要求，具备资质/资格证书、许可证等，如：当涉及医学伦理标注时，标注软件系统应通过相应机构伦理委员会的论证程序。

#### 5.3.1.2.4 标注过程要求

a) 标注人员

为提高标注的准确性和敏感度，降低假阳性率，避免记忆偏倚，标注流程建议多轮次分组交叉进行，优化人力资源。由于在不同环节上的工作量和人员资质存在差异，标注工作需要标注医师、标注组长和仲裁专家 3 种级别的医师参加。标注组长由工作经验 10 年以上的副主任医师担任，每一批标注任务由标注组长带领两名标注医师承担。

b) 标注流程

在数据标注过程中需要确认标注的目标以及具体的标注条件。以下以乳腺癌病历报告中的乳腺肿瘤标注作为样例（图 9），具体数据标注需求和条件请专家结合领域知识设计。考虑到本标准涉及较多的数据模态，在必要情况下需要针对模态设计对应的标注要求和标注流程规范。

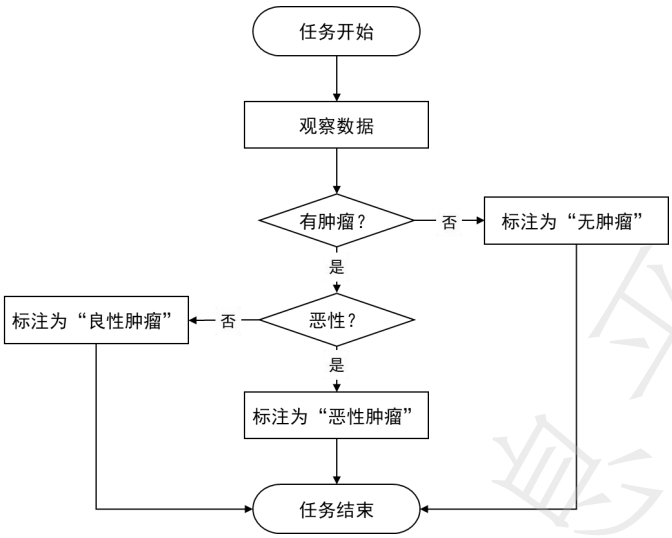


**图 9 乳腺肿瘤标注流程**

c) 标注任务

按照不同数据类型及需求，具体标注任务包括如下方面：

（1）分类标注：即将数据按照人工判读得到的数据类别进行标签标记，通常可分为二分类标记和多分类标记，例如通过解读病历文本进行“无肿瘤”、“有肿瘤”（二分类）或“无肿瘤”、“恶性肿瘤”、“良性肿瘤”、“无法确定”（多分类）标记；

适用：病历文本、医疗影像数据。

（2）区域标注：对数据拟标注目标的范围进行标注，如对病历文本中的实体起始字符位置进行标记或对医疗图像中目标区域范围进行标记；

适用：病历文本、医疗影像数据。

（3）标框标注：框选标注需要检测的目标对象，如在超声图像中框选出肿瘤位置；

适用：医疗影像数据。

（4）描点标注：对于细致特征的要求中需要将拟检测目标进行描点标记，如描点标记出超声图像中的肿瘤具体形状；

适用：医疗影像数据。

（5）OCR 转写标注：识别图片格式中的手写文字，转写为计算机可识别的文本格式，如识别图片扫描格式的病历文本；

适用：图片扫描的病历文本。

d) 标注细则

标注过程中应尽量做到无错标、漏标，对无法确定具体类别的数据样本纳入集中管理并呈交专家组进行标记和复审。对于图像标记应做到以不同颜色区分主要征象和次要征象，以标记主要征象为主、尽量多标次要征象；在标注病灶轮廓时，对内部细节辅以文字进行描述。

e) 自动标注

允许利用标注系统自动进行标注，但其结果仍需医疗专家审核和评价。

[来源：T/CMDA 002-2020，有修改]

#### 5.3.1.2.5 标注质量评估

a) 评估人员

仲裁专家由工作经验 15 年以上的专科疾病相关领域副主任医师或主任医师担任。

b) 评估方法

为提高标注的准确性，建议将标注流程进行多轮次、分组、交叉设计，整个标注工作由多级别医师分组（标注医师、标注组长、仲裁专家）参与。标注流程（图 10）建议如下：

（1）数据标注工作由标注组长带领至少 3 名标注医师独立完成，可在标注数据中设置一定数量的重叠样本以检测标注一致性；

（2）标注完成后，先由计算机对标注结果进行自动比较和归集，对于标注一致结果进行合并，对于标注不一致结果以特殊颜色提醒仲裁专家进行标注；

（3）由标注组长对标注结果进行复审、仲裁专家进行终审从而形成最终标注结果。

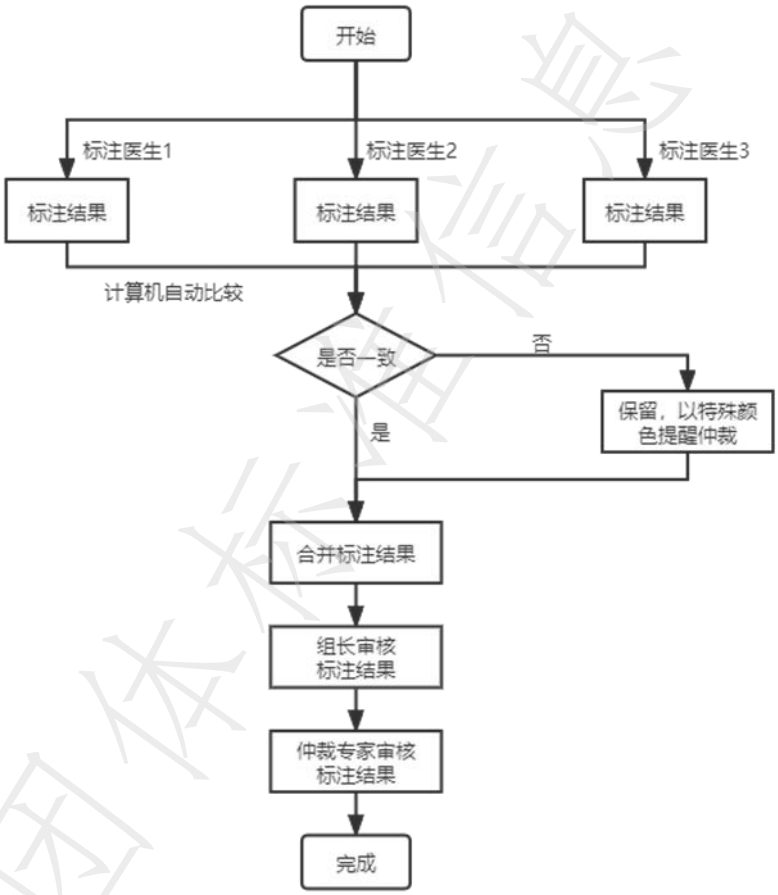


图 10 标注工作流程

c) 评估指标

（1）如无金标准标注数据集，可用标注人员间的标注一致性指标评估：

$$\text{标注一致率} = \frac{\text{重叠样本中标注一致的样本数量}}{\text{重叠样本数量}} \times 100\%$$

（2）如有金标准标注数据集，可用以下指标评估：

$$\text{标注准确率} = \frac{\text{重叠样本中标注一致的样本数量}}{\text{已标注数据集与金标准数据集重叠样本数量}} \times 100\%$$

d) 通过准则

（1）标注环节：由 3 名标注医师背靠背独立标注，然后用计算机自动对标注结果进行评估指标计算，以所有标注人员结果的并集作为结果；

（2）审核环节：由其他标注组长和仲裁专家各自独立对标注结果进行审核与修改，纠正漏诊、误诊和误判。

#### 5.3.3 主数据管理

主数据管理要做的就是从各部门的多个业务系统中整合最核心的、最需要共享的数据（主数据），集中进行数据的清洗和丰富，并且以服务的方式把统一的、完整的、准确的、具有权威性的主数据提供给需要使用这些数据的操作型应用系统和分析型应用系统。

主数据管理流程及要求如下：

a) 识别主数据，建立主数据模型。

b) 识别主数据域的业务职责。

c) 制定主数据采集标准，确定主数据存储模型，采集分散在各个业务系统的主数据集中存储到统一存储库。

d) 制定主数据清洗标准，根据业务规则和数据治理标准对采集到的主数据进行加工清洗，从而形成符合专病数据集标准的主数据。

e) 制定主数据变更标准，例如变更版本管理等（当主数据变化时能记录主数据的变更内容），从而保证主数据修改的一致性和稳定性。

f) 制定主数据历史版本管理标准，需要对主数据进行分层并记录不同的版本值。

g) 确保在多个业务场景中主数据共享的一致性。

#### 5.3.4 数据质控

数据质控，即数据验收，指对进行了脱敏、加密、转换等数据处理流程后的数据的完整性、一致性、正确性等进行验收。如果数据未通过验收，在使用中是不准确的。只有通过验收的数据，在执行后期数据生产、处理时，才是可以使用的数据。

##### 5.3.4.1 数据质量管理流程

数据集应由专人对数据的合规性、质量、容量、多样性和临床依从性等方面建立评价指标体系，评价结果形成报告储存。定期进行数据稽查，由不直接参与研究的人员对数据的一致性、合规性等方面进行检查。

数据质控总体流程如图 11：

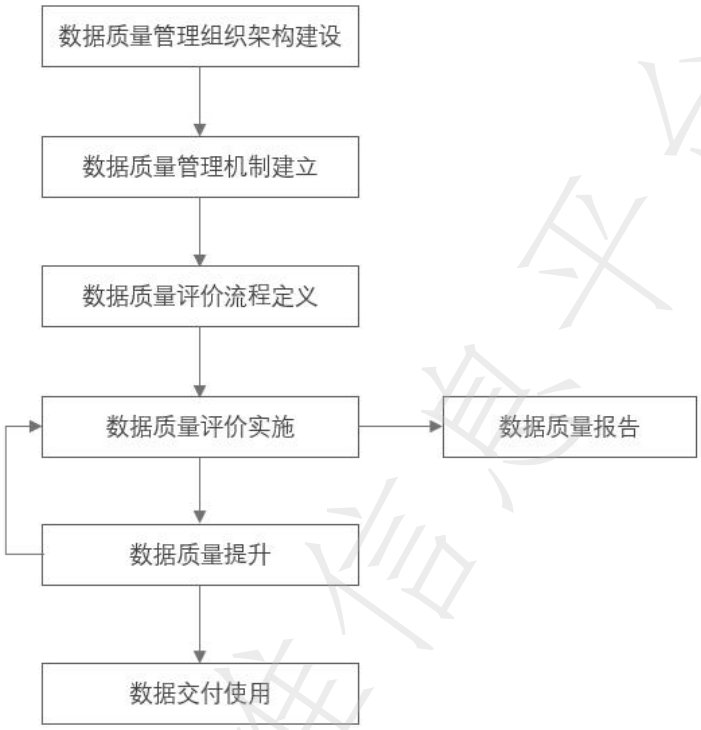


图 11 数据质控总体流程

a) 数据质量管理组织架构建设

数据质量管理是一个体系化的工作，需要多部门人员参与，建立较为完善的人员组织架构，其中较为关键的岗位包括：

数据质量管理岗：牵头数据质量标准、数据质量检查规则的订立和维护，数据质量评估模型的定制和维护、数据质量相关办法的编制、修订、解释、推广和落地，以及专项数据质量整改工作。

数据协调员：数据协调员来自于涉及数据治理的相关部门，职责在于代表本部门参与数据质量相关的评审、决策，配合、协调、推动数据质量管理在本部门的的执行。

b) 数据质量管理机制建立

建立数据质量管理各个参与部门的沟通机制。

建立数据持续校验、周期校验机制。

建立质量问题数据反馈、处理机制。

c) 数据质量评价

数据质量评价通过设计数据质量模型、订立数据质量规则，对数据进行校验，并根据校验结果对数据质量进行评价，定位问题数据，指导数据问题的解决，从而提升数据质量，数据质量评价是数据质量管理的核心。

d) 数据质量提升

根据数据质量评价中暴露的数据质量问题和问题数据反馈机制，对问题数据进行继续跟踪处理，从而提升数据质量。

e) 数据交付使用

对于达到数据质量要求进行交付使用，常见的使用方式包括库表、文件、接口等。

#### 5.3.4.2 数据质量评价

数据质量评价，首先需要基于数据标准确定需要进行校验的数据范围，为每个需要校验的数据创建数据质量模型，模型中包括主校验表、规则评分卡等要素，数据质量模型确定之后，可以借助数据治理平台或开发 SQL 脚本的方式来对数据质量模型进行落地上线，并定期运行数据质量模型，调度监控数据质量任务，确保数据质量任务定期按时完成，最后根据数据质量模型运行结果，形成数据质量报告、评分、问题数据集。

数据质量评价流程见图 12。

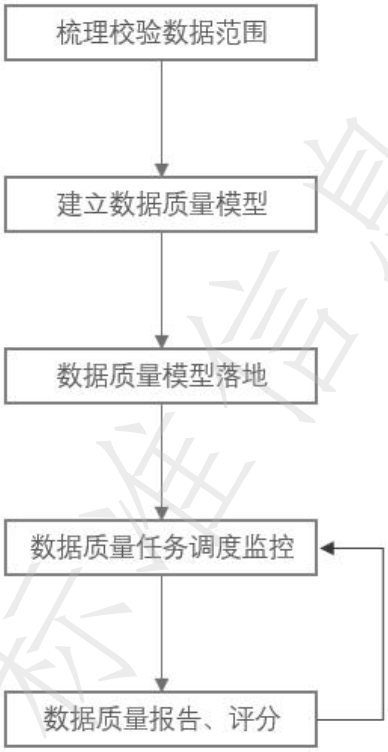


**图 12 数据质量评价流程图**

其中，建立数据质量模型是核心步骤，数据质量模型定义为：标识和评价一个数据集的数据质量情况的形式结构。一个数据质量模型主要包括：主校验表、关联校验表（可选）、校验规则、评分卡（必选）、质量分析维度（可选）、校验结果、数据质量报告等内容。

a) 主校验表

主校验表是当前数据质量模型校验的目标数据集，整个数据质量模型的配置都围绕主校验表进行。

b) 关联校验表

关联校验表是对主校验表进行跨数据集的数据一致性校验时，需要关联的其他数据集。

c) 校验规则

校验规则也叫数据质量规则，是数据质量模型核心内容，它通过定义被校验数据应该满足的质量标准来指导质量校验任务的运行，并根据运行结果计算被校验数据在该规则上的得分。质量规则应该包括规则代码、规则描述、所属规则维度、规则权重等信息。

具体维度定义见表 5。

**表 5 质量评价维度定义**

| | |
|---|---|
| 完整性 | 衡量所必须的数据的完整程度，主要包括数据记录条数完整和字段值内容完整。 |
| 一致性 | 同一数据元素的类型和含义在不同的系统和不同的数据处理环节上必须保持一致和清晰，主要包括元数据一致和数据内容的一致。 |
| 准确性 | 确保数据必须反映真实的业务内容。不仅仅是与原始文本或单据比较准确性，也可以是数据的源头与目标作比较。 |
| 规范性 | 对于数据的值、格式要求符合数据定义或业务定义的要求，如某些电话、邮箱的格式。 |
| 唯一性 | 针对某个数据项或某组数据，没有重复的数据值，值必须是唯一的，如 ID 类数据。 |
| 及时性 | 对于数据更新频率的满足程度，针对用户对信息获取的时间及时性要求，确保数据及时更新。 |

d) 模型评分卡

多个数据质量规则和其在数据质量模型中的权重组合形成该模型的评分卡，对模型校验的数据进行综合评价打分。（实例见图 13）


**图 13 数据质量模型评分卡实例**

其中，规则权重是用于标识数据质量模型中各个规则的重要性（以正整数表示），此重要性会体现

在该模型总体的质量评分计算中。数据质量模型根据各个规则的校验得分和规则权重，进行加权平均值计算，最终计算出整个数据的数据质量评分。

单个规则的数据质量评分为：

$$规则得分 = \frac{校验数据总量 - 不合格数据量}{校验数据总量} \times 100$$

数据质量模型总体评分为模型内所有规则得分的加权平均值：

$$模型总体评分 = \frac{规则\ 1\ 得分\times规则\ 1\ 权重 + 规则\ 2\ 得分\times规则\ 2\ 权重 + \ldots + 规则\ N\ 得分\times规则\ N\ 权重}{规则\ 1\ 权重 + 规则\ 2\ 权重 + \ldots + 规则\ N\ 权重}$$

e) 分析维度

数据质量模型中可以预先自定义分析维度，如日期、地区等，在后续的数据质量报告中，可以按预定义的维度进行分析。

f) 质量任务调度

数据质量模型上线后，可以定期运行，运维人员可监控质量任务的运行情况，处理任务调度中出现的问题，以确保每期数据质量任务全部运行成功。

g) 校验结果

数据质量任务运行完成后，可以查质量模型中查询校验结果，包括总体评分和各个规则的得分以及问题数据明细。

#### 5.3.4.3 数据质量报告

每个数据质量模型可单独形成数据质量报告，数据治理中心也可形成总体数据质量报告，报告中可按不同的维度分析数据质量模型的情况，包括默认评价维度和自定义的数据质量分析维度，其中数据质量规则分类维度为默认评价维度，按 6 个数据质量评价分类对数据质量进行分析。（参考图 14）

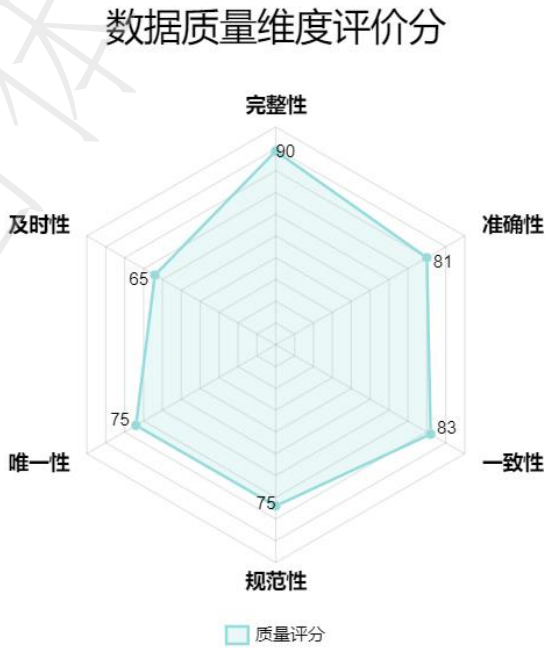


图 14 数据质量维度评价分参考

各个维度的评分规则为该维度内所有规则得分的加权平均值：

$$维度质量评分=\frac{维度内规则\ 1\ 评分\times维度内规则\ 1\ 权重+...+维度内规则\ N\ 评分\times维度内规则\ N\ 权重}{维度内规则\ 1\ 权重+...+维度内规则\ N\ 权重}$$

同时，报告中可进行必要的数据质量趋势分析，用于反映整个数据中心或某个数据质量模型的数据质量变化情况。

#### 5.3.4.4 总体数据质量评分

数据治理中心的总体数据质量评分由所有数据质量模型的加权平均值确定。每个数据质量模型需要拥有自己的权重等级，权重等级定义了该模型在整个数据治理中心内的重要性，预设级有：普通模型、重要模型、核心模型，其在大数据治理平台中默认权重比值为：1:3:5。

数据治理中心总体质量评分为各个数据质量模型评分的加权平均值：

$$总体质量评分=\frac{模型\ 1\ 评分\times模型\ 1\ 权重+模型\ 2\ 评分\times模型\ 2\ 权重+...+模型\ N\ 评分\times模型\ N\ 权重}{模型\ 1\ 权重+模型\ 2\ 权重+...+模型\ N\ 权重}$$

数据治理中心各个质量维度评分为：各个模型内维度评分的加权平均值：

$$各质量维度评分=\frac{模型\ 1\ 中该维度评分\times模型\ 1\ 权重+...+模型\ N\ 中该维度评分\times模型\ N\ 权重}{模型\ 1\ 权重+...+模型\ N\ 权重}$$

### 5.4 数据存储与计算

数据存储与计算包含两部分内容：数据存储技术、数据计算引擎。针对不同产品的设计，存储与计算不是绝对分离，有些是存储与计算分离的设计，比如 Hadoop 技术体系多数采用存算分离的技术，另外一部分采用存算一体的设计，如 MPP 数据库、传统关系数据库等。

#### 5.4.1 数据存储管理技术

数据存储管理是实现不同类型数据归档存储的技术支撑；其主要有文件管理系统和数据库管理系统两大类，其中，数据库管理系统根据不同的应用场景，包括分析型数据库、关系型数据库、图数据库、NoSQL 数据库等四大类。

a) 分布式文件系统

分布式文件系统是基于多存储节点，实现海量数据存储访问的文件管理系统，主要提供非结构化数据存储管理，存储容量支持弹性扩展。用于存储专病数据集的影像、出院小结、现病史、检查结论等非结构化数据。

b) 分析型数据库

分析型数据库是指基于 MPP 架构，实现分布式存储和分布式计算的数据库，面向专病数据集诊断、医嘱等业务场景，存储数据量大且有复杂统计分析计算的结构化数据，可用于存储决策支持、专病专题分析等业务数据，以满足大数据量 AD-HOC 查询的应用需求。

c) 关系型数据库

关系型数据库是指传统业务系统所使用的事务型数据库，面向专病数据集 OLTP 业务场景，存储有高并发 CURD 功能需求的结构化数据，可用于存储专病数据集病例查询、门诊记录、住院记录等业务数据。

d) 图数据库

图数据库是基于图论实现的一种新型数据库，其数据存储结构和数据查询方式都是以图论为基础，存储具有图数据结构且需进行复杂关联关系查询与分析的结构化数据。可用于专病数据集中病例关联图谱、诊断关联图谱等图结构数据。

c) NoSQL 数据库

NoSQL 数据库即非关系型数据库，主要类型有文档型，键值型、时序型、列存储型等。其中，文档型 NoSQL 数据库用于存储专病数据集的各类自定义表单和文档等非结构化数据，方便随时变更数据指标。键值型 NoSQL 数据库可以用于存储专病数据集的分析类应用中间计算结果数据或业务系统字典缓存数据。时序型 NoSQL 数据库可以用于存储基于时间序列的事件数据，如专病数据集病例就诊信息更变日志数据。列存储型 NoSQL 数据库可用于存储专病数据集业务系统产生的海量日志数据；在实际应用中，根据数据类型及数据应用的业务场景，选择合适类型的 NoSQL 数据库。

### 5.4.2 数据计算引擎

数据计算引擎是数据处理技术框架的关键，数据计算引擎为各类数据加工、数据分析任务提供高性能的算力支撑；计算引擎主要有批量计算引擎、流式计算引擎、科学计算引擎三大类。

a) 批量计算引擎

主要用于大数据量、时间不敏感的数据处理场景，其特点是数据吞吐量大、数据处理过程复杂、延时高；支撑专病数据集大批量的数据加工、大规模数据的清洗转换以及挖掘分析等。如对批量数据清洗与转换、决策支持系统中的大数据量统计分析等业务场景。

b) 流式计算引擎

主要用于对数据加工计算和应用有较强的时效性要求场景，其特点是数据处理时效高、响应速度快；用于支撑数据中心大屏的实时动态展现、风险预警等场景。

c) 科学计算引擎

是一种分布式场景下大规模数据科学计算的新计算库（引擎），提升在图像处理、机器学习、深度学习等领域的计算性能，专病数据集数据挖掘模型、深度学习模型的开发等应用场景需要用到科学计算引擎。

## 5.5 数据安全

### 5.5.1 安全设计标准

- 全国医院信息化建设标准与规范（试行）
- 通用数据安全体系
- 信息安全管理标准
- 信息技术开放系统互连开放系统安全框架
- GB/T 9361-2011 计算机场地安全要求
- GB 50174-2008 电子计算机房设计规范
- GB/T 18336-2008 信息技术安全技术信息技术安全性评估准则
- GB/T 18018-2007 路由器安全技术要求
- GB/T 9387.2-1995 信息处理系统开放系统互连基本参考模型第 2 部分：安全体系结构
- GB/T 18237-2000 信息技术开放系统互连通用高层安全
- GB/T 18231-2000 信息技术低层安全
- GB/T 2887-2000 计算机场地通用规范

- GA243-2000 计算机病毒防治产品评级准则
- GB/T 17963-2000 信息技术开放系统互连网络层安全协议
- GB/T 17900-1999 网络代理服务器的安全技术要求
- GB/T 18019-1999 信息技术包过滤防火墙安全技术要求
- GB/T 18020-1999 信息技术应用级防火墙安全技术要求
- GB 17859-1999 计算机信息系统安全保护等级划分准则
- GB/T 17143.7-1997 信息技术开放系统互连系统管理安全报警报告功能
- GB/T 17143.8-1997 信息技术开放系统互连系统管理安全审计跟踪功能

### 5.5.2 数据访问权限控制

专病数据集采用多级授权管理模式，对用户权限进行管理，可针对用户、数据等进行权限配置。主要包括授权权限管理和业务授权管理：

授权权限管理，中心系统管理员给业务管理员授权，授权业务管理员对其业务范围内的操作授权。

业务授权管理，各业务管理员根据数据集管理员所授业务范围的权限，对业务范围内的具体业务操作权限授权。

### 5.5.3 数据加密

数据加密分为存储加密和传输加密。

存储加密，即专病数据集的数据先加密再存储，防止重要数据丢失，防止数据文件被复制、攥改或盗用；加密过程中选择加密强度较高的加密算法，提供数据逻辑性和有效性的自动交验功能，对用户输入信息进行安全检查，降低 SQL 注入，数据库管理员权限被攥改、滥用等数据安全风险。

传输加密，在使用专病数据集过程中，有着大量的数据通过网络进行传输，因此，数据传输的安全显得格外重要。采用加密技术实现用户身份和鉴权口令、用户资源等关键信息的加密传输或加密存储，防止信息在网络传输或存储中被窃取、破坏及篡改。数据加密可支持多种加密形式及数据加密提醒。

数据存储加密一定程度上影响算及读取速度，可按需调整。

### 5.5.4 数据脱敏

需提供数据脱敏功能，将需要导出的敏感信息自动替换为随机乱码或特殊符号，保障所输出的信息中，敏感数据或身份数据信息需要隐藏，避免信息泄漏。对于安全级别较高的数据，进行重点监控，此部分数据的修改、访问、拷贝等操作，都需记入操作日志。

基于大数据的数据脱敏机制，专病数据集在接收数据时，要对专病数据集的数据进行脱敏处理。推荐的数据脱敏方法包括：

替代：使用伪装数据替换源数据中的敏感数据以保证安全；

混洗：对敏感数据进行随机变换打破原有的关联关系；

数值变换：通过随机函数对数值型数据进行可控的调整，是常用的脱敏方法；

加密：加密处理待脱敏数据，外部用户只能看到无意义的加密数据；

遮挡：指对敏感数据的部分内容用掩饰字符如“术”、“#”等进行统一替换，从而使得敏感数据保持部分内容公开；

空值插入：将敏感数据删除或置为 NULL 值；

反脱敏：支持反脱敏机制，将已脱敏的数据进行恢复 。

### 5.5.5 安全审计

专病数据集使用应该采取以下审计措施：

a) 审计内容宜包括人员审计、管理审计、技术审计（系统、网络、操作、日志审计等）；

b) 任何操作，包括登录、创建、修改和删除记录的行为，都宜自动生成带有时间标记的审计记录，包括但不限于修改时间、修改原因、修改内容、修改人及签名等信息，并可供审计；

c) 应制定和部署专病数据集活动审计政策，重点对于专病数据集的访问及操作的合规性进行审计，确定必要的审计控制范围和需要审计的数据；

d) 应制定适当的标准操作流程，确定异常报告所需的审计跟踪数据和监视程序的类型；

e) 审计记录应安全存储并实施访问控制，只允许授权人员能够查看相关记录，保存的内容需反映临床医学研究整个过程。

### 5.5.6 数据操作日志

专病数据集应该建立数据访问操作日志记录，通过高易用性的日志功能，对所有应用系统的操作过程轨迹进行记录，以便为以后对操作痕迹进行追踪，为应用系统操作漏洞分析提供原始证据。访问日志记录，包括了操作轨迹的记录、应用系统错误日志记录、访问日志数据存储和分析等。以访问日志的形式，确保应用操作的不可抵赖性。

### 5.5.7 数据备份

本地备份是数据库容灾的重要组成部分，需要建立健全的备份和恢复机制，对数据库进行保护。通过备份恢复设备，对重要数据进行备份保护。根据业务特点及数据性质，制定相应的备份策略，并安排运维人员，对备份进行监控，如遇备份失败，第一时间进行故障定位并重启备份任务。

建议备份策略：

a) 根据数据库厂商官方的备份建议，以周为备份周期；

b) 周日全备份；

c) 周一至周六增量备份；

d) 每天每 12 小时执行一次备份（依据实际日志产生数量）；

e) 备份保留周期，依据实际要求确定。

## 6 数据应用标准

数据应用主要包含对专病标准数据集进行采样和处理，形成适用于支撑医疗 AI 器械研发和评测医疗 AI 器械性能的数据集，并进行相关数据集或评测数据报告开放的过程，整体应用过程如图 15 所示。

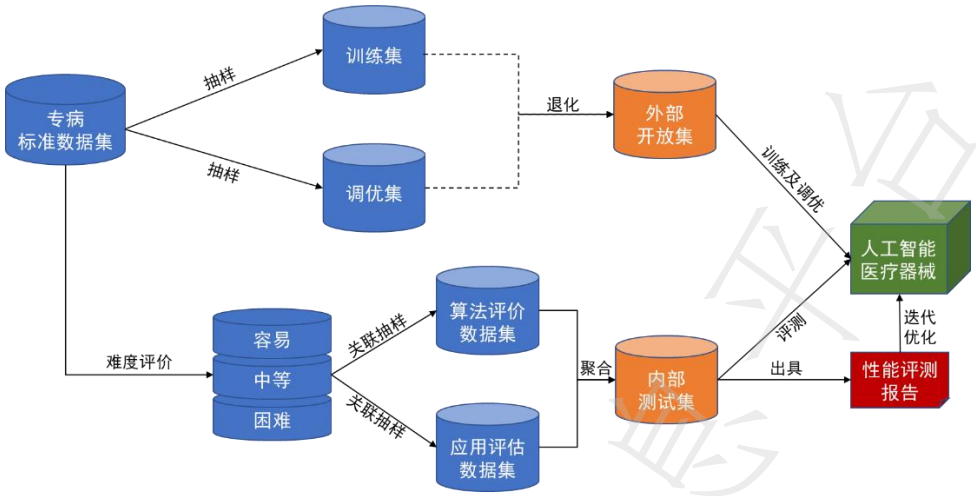


图 15 专病标准数据集应用流程示意图

### 6.1 数据采样及处理

通过对专病标准数据集进行采样和处理，形成外部开放集和内部测试集两部分数据子集。外部开放集用于支撑医疗 AI 器械研发，内部测试集用于评测医疗 AI 器械性能，外部开放集和内部测试集无数据交集。

#### 6.1.1 外部开放集建设要求

通过对专病标准数据集进行随机采样或关联采样形成训练集和调优集，训练集和调优集无数据交集。训练集数据量一般大于调优集并成一定比例（如 4:1，9:1 等）。训练集和调优集经过严格脱敏和数据退化聚合形成外部开放集，用于支撑医疗 AI 器械研发。

#### 6.1.2 内部测试集建设要求

通过对专病标准数据集进行数据难度测算形成不同难度（如容易、中等和困难）的数据子集，从数据子集中分别进行两次关联抽样形成 AI 算法评价数据集和 AI 应用评估数据集。算法评价数据集和应用评估数据集聚合形成内部测试集，用于进行医疗 AI 器械算法性能和真实应用场景适配程度两方面的评估，形成评测报告。

### 6.2 数据开放

外部开放数据集在经过严格的数据脱敏、退化处理之后，可采用受控公开共享方式。如提供数据 API 接口，由实名注册并签订数据使用协议的特定用户发起数据请求，系统实时单点返回用户所需的开放集数据，由用户自行存储并用于数据处理、医疗 AI 器械研发等。

内部测试数据集采用领地公开共享的方式。即提供医疗 AI 器械测试平台，实名注册用户签订数据使用协议后，可在平台上自行部署医疗 AI 器械模型。通过后台自动在内部测试集上进行模型评估并返回测试报告，供用户下载。内部测试集原始数据不得导出，可提供模拟数据样例对测试集数据结构进行展示。

## 6.3 医疗器械评价方法

### 6.3.1 测试评价流程

利用内部测试数据集，针对专病真实医疗应用场景，开展医疗 AI 器械的测试评价。测试流程见图16。

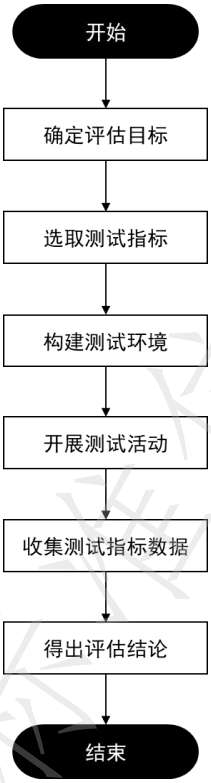


图 16 医疗 AI 器械测试评价流程

[来源：T/CESA 1109-2020，有修改]

### 6.3.2 确定评估目标

根据测试评价活动开展的不同目的以及医疗 AI 器械应用的不同场景和领域，由第三方测试机构根据实际情况确定本次测试评估的目标。

### 6.3.3 选取测试指标

根据具体专病医疗场景以及医疗 AI 应用范围进行测试指标设置，可从以下测试指标选取也可根据实际情况自行进行具体指标设计。

a) 准确率

在给定测试环境下，AI 器械对测试集中 n 个数据样本进行测算，其中测算正确的样本数占 n 个样本总数的比率。

$$准确率 = \frac{预测正确的样本数}{选取的\ n\ 个数据样本} \times 100\%$$

b) 敏感度

在给定测试环境下，AI 器械对测试集中 n 个数据样本进行测算，其中测算结果为正例的样本占总数据样本中正例样本的比率。

$$敏感度 = \frac{测算结果为正例的样本数}{总数据集中的正例样本数} \times 100\%$$

c) 特异度

在给定测试环境下，AI 器械对测试集中 n 个数据样本进行测算，其中测算结果为负例的样本占总数据样本中负例样本的比率。

$$特异度 = \frac{测算结果为负例的样本数}{总数据集中的负例样本数} \times 100\%$$

d) ROC 曲线下面积

在给定测试环境下，ROC 曲线下的面积，其中 ROC（receiver operating characteristic curve）表示接收者操作特征曲线

e) 鲁棒性

在给定测试环境下，AI 器械对于输入的 n 个非正常数据样本，能保持其准确率不发生较大偏移的能力。

$$鲁棒性 = \frac{非正常数据样本中能够保持其结果正常输出的样本个数}{测试集中的非正常数据样本个数} \times 100\%$$

f) 响应时间

在给定测试环境下，AI 器械对给定 n 个数据样本进行测算并获得结果所需要的平均时间。其中$T_{bi}$表示第 i 个数据样本输入的时间，$T_{ei}$表示第 i 个数据样本的测算结果输出的时间。

$$响应时间 = \frac{1}{n}\sum_{i=1}^{n}(T_{ei} - T_{bi})$$

### 6.3.4 构建测试环境

根据 AI 器械运行需要的软硬件要求，构建一致的测试环境。如无法构建出与专病实际使用场景相同的测试环境，则需要进一步分析由于测试环境与实际使用环境不一致所带来对测试结果的影响。

### 6.3.5 开展测试活动

在构建的测试环境下，利用测试数据集对 AI 医疗器械进行测试。即按照被测试器械的正确使用方法，将测试数据集中的数据样本全部输入被测器械。

### 6.3.6 收集测试指标数据

根据所选择测试指标的计算公式，收集被测试器械的输出数据并进行相应测试指标计算。

### 6.3.7 得出评估结论

将测试指标结果与给定的指标要求进行对比，并综合得出被测试器械的评估结论，并出具对应的性能测试报告。

# 附录 A 专病数据集（示例）

## A.1 总体数据架构设计

采用统一数据标准，全面汇聚医院内部、外部数据，形成专病数据集，为专病数据的应用和服务提供数据支撑。在数据来源上，汇集 HIS、LIS、PACS、EDC、随访等内部数据，以及医保、环境等外部部门数据。通过离线批量处理、实时同步处理、文件传输、数据接口等方式实现不同来源数据的入库。数据入库后，先保持原格式进入操作数据层（ODS），其中格式化数据进入汇聚数据库、非格式化数据进入文件服务器。对于格式化数据，通过数据治理后，进行标准化处理后，进入数据仓库层（DW），形成数据模型，并摆渡至应用数据层（ADS）支撑应用使用。对于非结构化数据，进入各自文件服务器后，待后续结构化处理或应用。

专病数据集采用分层的数据架构如下图 17 所示：

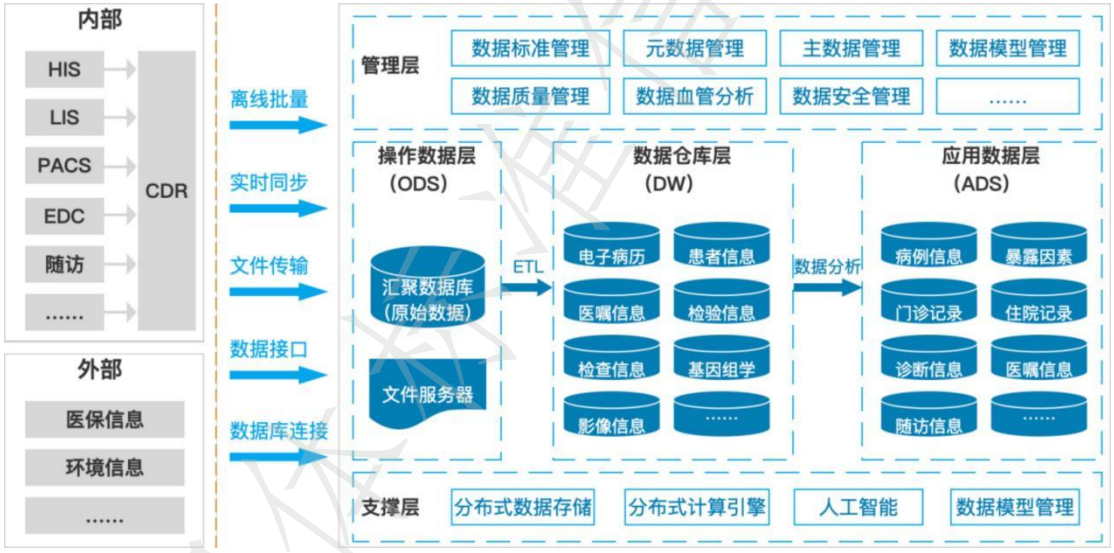


图 17 专病数据集数据架构

## A.2 专病数据模型

a) 乳腺癌数据

乳腺癌数据集数据包括但不限于四川大学华西医院 2008-2019 年间乳腺癌患者的门诊记录、住院记录、诊断、医嘱、用药、手术、检验、超声、病理、随访数据。

数据源包括但不限于 HIS、LIS、超声系统、病理系统、科研随访项目。

数据集由实体/表、属性名称、英文名称、类型、值域、参考标准等组成如图 18 所示。


图 18 乳腺癌数据集样例

b) 肾脏病数据集

数据集数据包括但不限于四川大学华西医院 2010-2020 年间肾脏病患者的门诊记录、住院记录、诊断、医嘱、用药、手术、检验、病理记录。

数据源包括但不限于 HIS、LIS、病理系统、科研随访项目。

数据集由实体/表、属性名称、英文名称、类型、值域、参考标准等组成如下图 19 所示。


图 19 肾脏病数据集样例

c) 食管癌数据集

食管癌数据集数据包括但不限于四川大学华西医院 2010-2020 年间食管癌患者的门诊记录、住院记录、诊断、医嘱、用药、手术、检验、影像、病理数据。

数据源包括但不限于 HIS、LIS、影像系统、病理系统、科研随访项目。

数据集由实体/表、属性名称、英文名称、类型、值域、参考标准等组成如下图 20 所示。


图 20 食管癌数据集样例

d) 抑郁症数据集

抑郁症数据集数据包括但不限于四川大学华西医院 2009-2019 年间抑郁症患者的门诊记录、住院记录、诊断、医嘱、用药、手术、检验、脑电数据。

数据源包括但不限于 HIS、LIS、心情量表系统、病理系统、科研随访项目。

数据集由实体/表、属性名称、英文名称、类型、值域、参考标准等组成如图 21 所示。


图 21 抑郁症数据集样例

其中，如病例基本信息包含信息如下表 6 所示：

表 6 病例基本信息样例表

| 实体/表 | 属性名称 | 英文名称 | 类型 |
|---|---|---|---|
| 病例基本信息 | 国籍 | country | 文本 |
| 病例基本信息 | 出生日期 | birth_date | 日期 |
| 病例基本信息 | 籍贯 | native_place | 文本 |
| 病例基本信息 | 职业 | career | 文本 |
| 病例基本信息 | 患者 id | patient_id | 整型 |
| 病例基本信息 | 婚姻状态 | marriage | 分类(单个) |
| 病例基本信息 | 性别 | sex | 分类(单个) |
| 病例基本信息 | 民族 | nation | 分类(单个) |
| 病例基本信息 | 费用类型 | patient_identity | 文本 |
| 病例基本信息 | 病人数据来源 | patient_src | 文本 |

### A.3 结构化数据清洗规则示例

数据治理团队对原始数据进行数据结构、完整度、格式、质量等维度进行评估后，输出结构化数据清洗工作计划。数据清洗规则包括但不限于（见表 7）：

表 7 结构化数据清洗规则示例

| 清洗规则分类 | 描述 | 举例 |
| --- | --- | --- |
| 数据字典编码 | 国家、婚姻状态、病人身份、付费方式、病人来源等 | 婚姻状态：未婚、已婚、初婚、再婚、复婚、丧偶 |
| 关键信息脱敏 | 姓名、地址、电话、身份证等 | 张三姓名脱敏为张* |
| 文本类字段脏数据清洗 | 婚姻、民族、职业、籍贯等脏数据处理，几乎 80%的文本字段都需要进行脏数据清洗 | 婚姻状态字段中除了婚姻状态，可能还有人名和其他字符 |
| 文本类字段不可见字符清洗 | 部分文本类型字段存在非法不可见字符 | 例如文本“否认外伤史（不可见字符）”改为“否认外伤史” |
| 日期类字段脏数据清洗 | 各种日期类字符的脏数据清洗 | 确诊日期字段中存在文字信息 |
| 文本类型字段统一格式 | 部分文本类型的字段内容需要统一格式 | 例如“***无类似记载”改为“无类似记载”，“3--4”统一为“3-4” |
| 数值类型字段格式统一 | 部分数值类型的字段内容需要统一格式 | 例如年龄“３９”（全角）统一为“39”（半角） |
| 日期类型字段内容拆分 | 部分日期类型的字段需统一格式 | 例如统一为 yyyy-mm-dd |
| 文本类型字段内容拆分 | 对部分比较规范文本内容按规则进行拆分 | 例如月经生育史的“妊娠 4 次，顺产 1 胎，流产 3 胎，早产 0 胎”按数据集模型拆到对应字段 |
| 格式转换 | 文本类型字段转其他格式 | 文本转日期、文本转数字 |
| 空值内容格式统一 | 对一些明显是空值的字段统一格式 | 例如内容为“NUL”和字符串“NULL”的全部转为 null |

### A.4 非结构化数据清洗流程样例

a) 由专业医师团队针对现病史/出院小结的文章、超声报告文章制定实体及关系规则。例如：

现病史/出院小结文章的实体有“日期”、“临床表现”、“检查内容”、“检查结果”、“治疗”等，关系有“检查内容-检查结果”、“治疗-日期”等。

乳腺超声报告文章的实体有“部位”、“病变”、“回声”、“血流信号”、“诊断”等，关系有“部位-病变”、“部位-回声”、“诊断-诊断描述”等。

b) 组织标注员团队，由专业医师团队对标注员团队进行标注培训，目的是熟悉标注软件的操作、统一实体及关系的标注规则，培训后抽选 50-100 篇真实文章，以专家团队的标注为金标准，对各个标注员进行标注质量的评估和问题分析，如存在明显问题需再次统一认识，反复该流程直到专家团队认可当前标注质量。

c) 根据文章总数按比例向标注员团队分配任务，专家团队进行抽查复核，复核方式主要包括 1.人工检查；2.专家团队标注 1/4 的文章量，由系统程序计算 F1 值和显示差异实体及关系。如果存在不一致的部分，由专家团队进行复审和确定最终版本。

d) 形成一定量文章数后，训练自动标注算法模型。当模型的 F1 值达到认可水平后，由模型对剩下的文章进行反标注，再由标注人员—专家团队进行复核与错误纠正。

## A.5 数据安全

a) 数据信息的采集、标注、使用、以及共享等获得医疗机构的伦理审批。以知情同意书形式获得个人信息主体的知情权与授权，或伦理机构同意免除知情通知；

b) 医院向开发者传递的数据，在医院会进行第一次脱敏；存入开发者的原始数据库前，会使用开发者自行研发的脱敏工具再次脱敏，确保开发者使用的影像数据已完全脱敏；

c) 系统严格按照 GB25000.51-2016 标准的要求保障数据的保密性、完整性、可用性。

（1）应采取数据分类、重要数据备份、加密认证等措施保证数据安全；

（2）系统分为管理员和一般用户两级授权机制进行不同等级用户的数据接入和使用权限控制，功过用户名、密码登录的方式，防止非授权人员进入；

d) 系统具备日志功能，对用户认证、对数据访问进行控制、规范数据接入、使用和销毁过程进行痕迹管理，确保对数据访问行为可管、可控；对服务管理的全流程留痕，对安全、隐私、风险及事故可查询、可追溯。

为了达到数据安全，专病数据集网络拓扑图见图 22 所示：

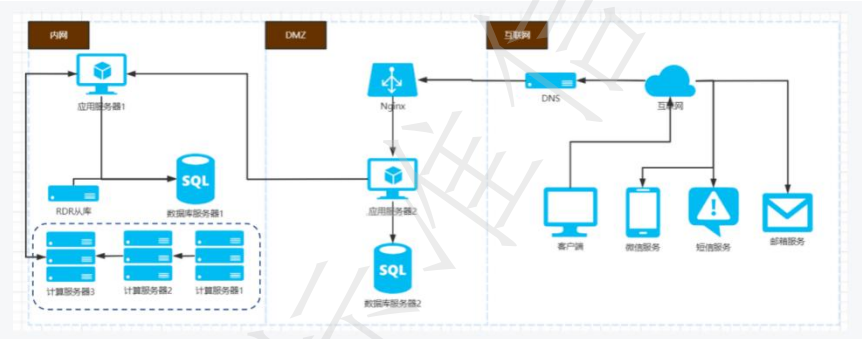


图 22 专病数据集网络拓扑图

## A.6 专病数据集视图系统展示

### A.6.1 乳腺癌

#### A.6.1.1 概述

建成乳腺癌专病数据库，包含 10281 余名患者约 3480 万条人口统计学信息、暴露/危险因素临床症状、患者月经生育史、患者住院记录、住院诊断、住院体格检查、住院影像、住院检验、住院诊疗计划、门诊/住院医嘱、现病史、出院小结以及随访等信息，全面挖掘数据价值应用，搭建具有专病特征的数据资源目录，细化研究方向，为后续乳腺癌诊疗与预防，以及科学研究及医疗 AI 器械研发等应用场景提供数据支撑。

结合大数据治理技术，对大量数据清洗标化包括：数据抽取、数据转换、数据装载。深入结合医生团队，整合专业医学知识。使用 NLP 挖掘文本价值，将文本化病历形成结构化数据。数据清洗标化是不断重复的周期性的过程，为科学研究及医疗 AI 研发提供重要依据。乳腺癌专病数据集首页如下图 23 所示。


图 23 乳腺癌专病数据集视图系统首页

### A.6.1.2 乳腺专病数据概览

支持乳腺癌专病数据库集成病历信息数据的总体视图和患者视图。基于专病数据库用户可以通过临床统一视图实现病历信息的综合概览，全方位、全周期查看病例诊疗数据。并提供临床统一视图建立患者诊疗时序模型，包括患者在不同时间节点的诊断、用药、体征、检验、检查、手术、治疗等。支持以时间轴的方式展示病例全生命周期的诊疗概览并支持详细记录的查看。

### A.6.1.3 专病数据资源目录

展示专病数据集的数据资源目录信息，以及各个目录中变量的内容、属性、有效填充率、变量相关的统计、说明、所属目录、相关变量等。用户可通过预览目录数据样例了解数据资源情况，乳腺癌专病数据资源目录如下图 24 所示：

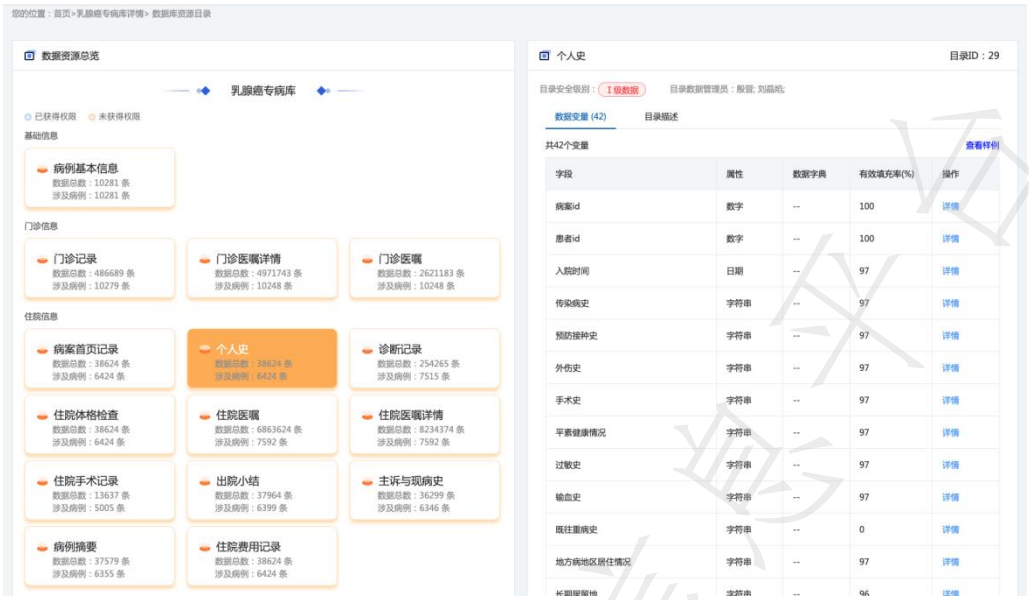


图 24 乳腺癌数据资源目录

### A.6.1.4 专病特征统计

提取乳腺癌专病数据集的关键性特征指标，通过可视化的图标展示统计分析结果，可便于用户提前了解数据情况，乳腺癌专病特征统计如下图 25 所示。

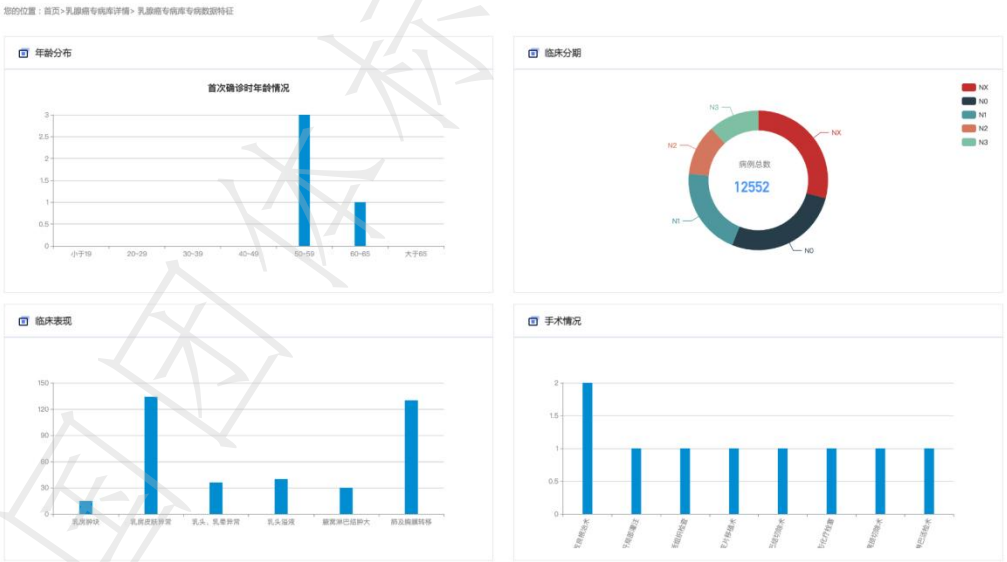


图 25 乳腺癌专病特征统计

## A.6.2 肾脏病

### A.6.2.1 概述

汇集华西医院历年来各信息系统的肾病病例信息，包括患者从门诊、住院以来所有的肾病所需信息。建成肾脏病专病数据集，目前包含 25000 余名患者约 8155 万条覆盖人口学信息、暴露/危险因素、疾病

诊断、现病史、既往史、生育史、用药情况、检查情况、防治策略、慢病管理等多维度信息。围绕大数据驱动的肾脏病防治模式创新、肾脏病数据登记和信息化标准、肾脏病真实世界数据的共享机制，实现肾脏疾病相关数据的互联融合，开放共享。以帮助研究者获得大量宝贵的病例资料，为高水平的临床科研提供坚实的基础。肾脏病专病数据集首页如下图 26 所示：


图 26 肾脏病专病数据集视图系统首页

### A. 6. 2. 2 肾脏病专病数据概览

支持肾脏病专病数据库集成病历信息数据的总体视图和患者视图。基于专病数据库用户可以通过临床统一视图实现病历信息的综合概览，全方位、全周期查看病例诊疗数据。并提供临床统一视图建立患者诊疗时序模型，包括患者在不同时间节点的诊断、用药、体征、检验、检查、手术、治疗等。支持以时间轴的方式展示病例全生命周期的诊疗概览并支持详细记录的查看。肾脏病专病数据集患者病例详情信息如下图 27 所示


图 27 肾脏病专病数据集患者病例详情信息

### A.6.2.3 肾脏病专病数据资源目录

展示专病数据集的数据资源目录信息，以及各个目录中变量的内容、属性、有效填充率、变量相关的统计、说明、所属目录、相关变量等。用户可通过预览目录数据样例了解数据资源情况，肾脏专病数据资源目录如下图 28 所示：

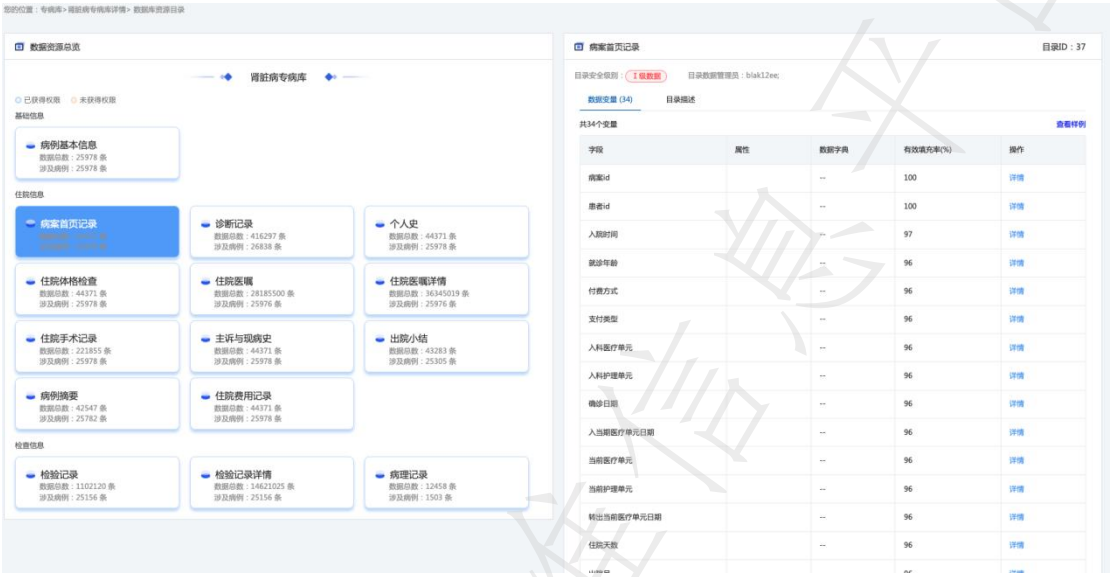


图 28 肾脏专病数据资源目录

### A.6.2.4 肾脏病专病特征统计

提取肾脏病专病数据集的关键性特征指标，包括年龄分布、住院时长、职业分布、性别统计、手术情况、愈后情况等指标，通过可视化的图表展示统计分析结果，可便于用户提前了解数据情况，肾脏病专病特征统计如下图 29 所示：

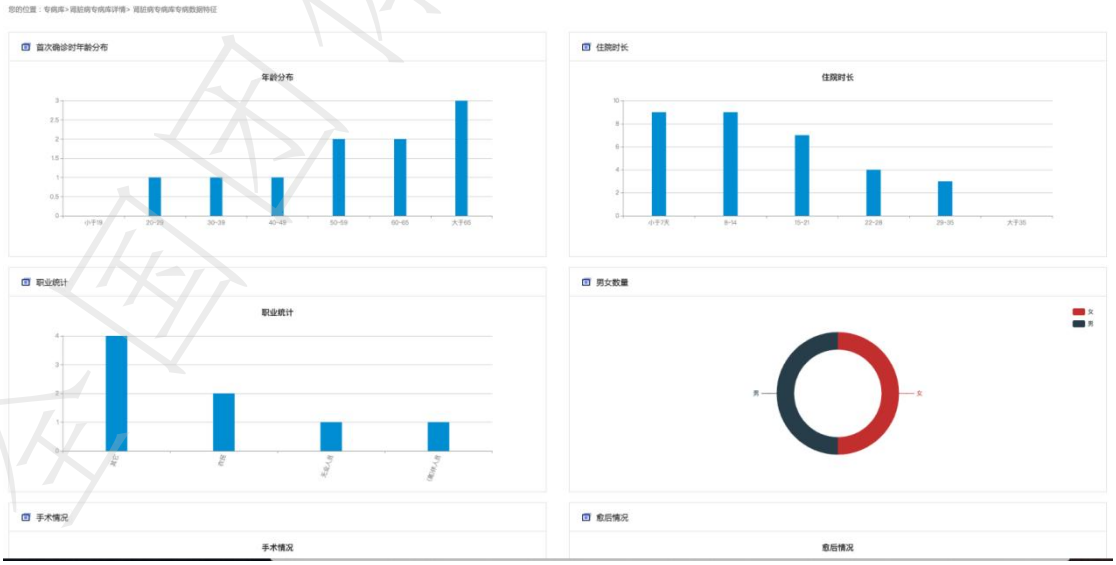


图 29 肾脏病专病特征统计

### A.6.3 食管癌

#### A.6.3.1 概述

建成食管癌专病数据库，包含 7000 余名患者约 1700 万余条人口统计学信息、暴露/危险因素临床症状、患者月经生育史、患者住院记录、住院诊断、住院体格检查、住院影像、住院检验、住院诊疗计划、门诊/住院医嘱、现病史、出院小结以及随访等信息，全面挖掘数据价值应用，搭建具有专病特征的数据资源目录，细化研究方向，为后续食管癌诊疗与预防，以及科研提供数据支撑。

结合大数据治理技术，对大量数据清洗标化包括：数据抽取、数据转换、数据装载。深入结合医生团队，整合专业医学知识。数据清洗标化是不断重复的周期性的过程，为科研决策分析提供重要依据。食管癌专病数据集首页如下图 30 所示。


图 30 食管癌专病数据集首页

#### A.6.3.2 食管癌专病数据概览

支持食管癌专病数据库集成病历信息数据的总体视图和患者视图。基于专病数据库用户可以通过临床统一视图实现病历信息的综合概览，全方位、全周期查看病例诊疗数据。并提供临床统一视图建立患者诊疗时序模型，包括患者在不同时间节点的诊断、用药、体征、检验、检查、手术、治疗等。支持以时间轴的方式展示病例全生命周期的诊疗概览并支持详细记录的查看。

#### A.6.3.3 食管癌专病数据资源目录

展示专病数据集的数据资源目录信息，以及各个目录中变量的内容、属性、有效填充率、变量相关的统计、说明、所属目录、相关变量等。用户可通过预览目录数据样例了解数据资源情况，食管癌专病数据资源目录如下图 31 所示：

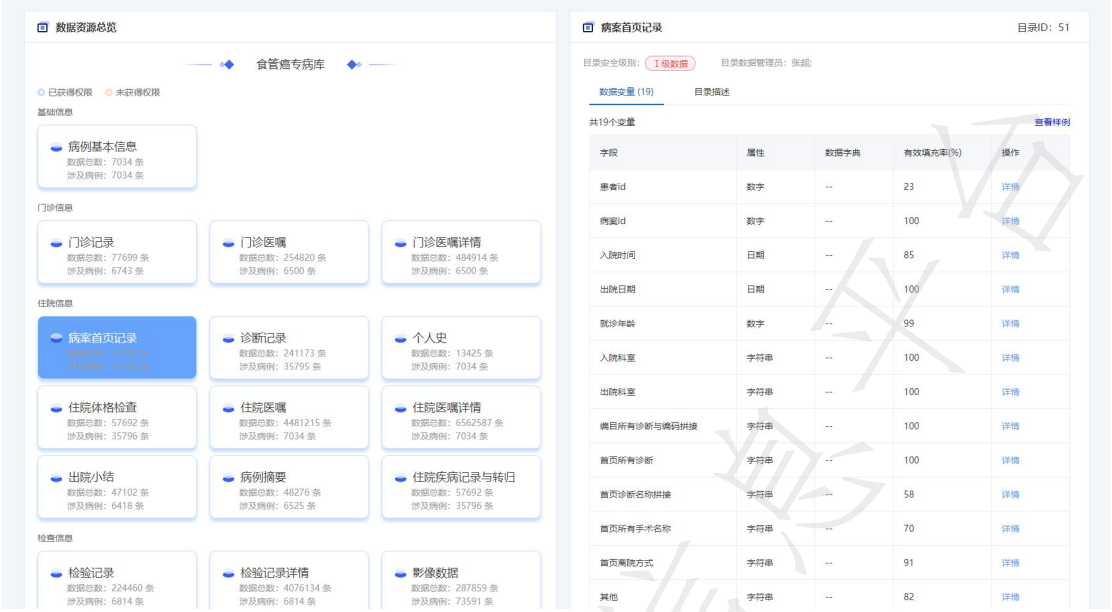


图 31 食管癌专病数据资源目录

### A. 6. 3. 4 食管癌专病特征统计

提取食管癌专病数据集的关键性特征指标，通过可视化的图标展示统计分析结果，可便于用户提前了解数据情况，食管癌专病特征统计如下图 32 所示：

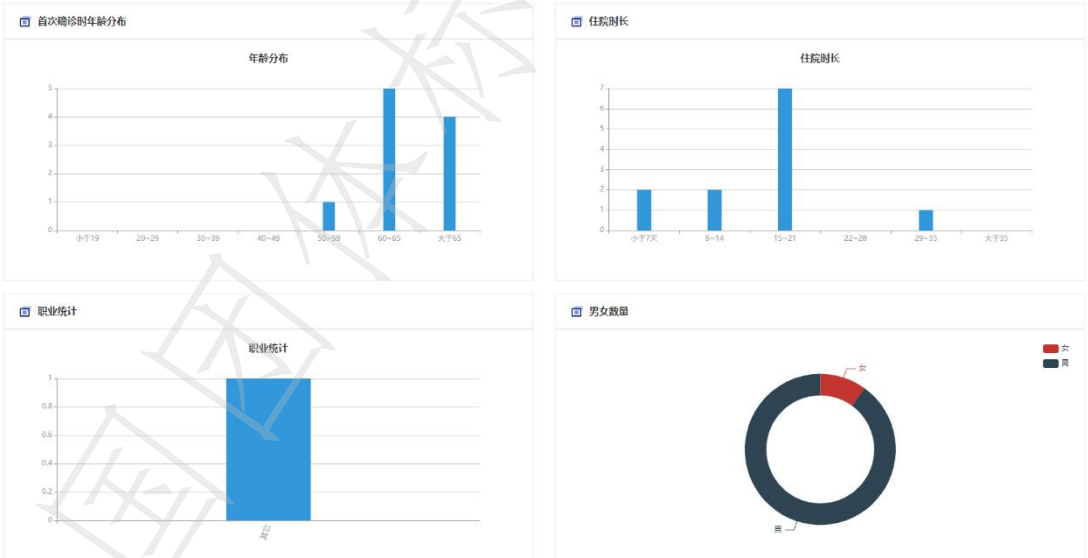


图 32 食管癌专病特征统计

## A. 6. 4 抑郁症

### A. 6. 4. 1 概述

汇聚多源异构数据，包括家族史数据、暴露/危险因子数据、门诊和住院的诊疗数据、患者检验数据、患者脑电数据、气象环境数据等。建成抑郁症专病数据库，包含 16000 余名患者约 3880 万条人口

统计学信息、暴露/危险因素临床症状、患者住院记录、出院诊断、住院体格检查、住院检验、住院诊疗计划、门诊医嘱等信息，为后续抑郁症诊疗与预防的科研提供数据支撑。

将多源异构的数据，进行数据清洗包括：数据抽取、数据转换、数据装载，过程繁琐，数据清洗标化是不断重复的周期性的过程，为科研决策分析提供重要依据。形成可视化的数据分析，其中包括年龄分布、职业分布、地区分布、住院时长、愈后情况等，同时提供数据挖掘分析，有助于更好的开展科研工作。抑郁症专病数据集首页如下图 33 所示。

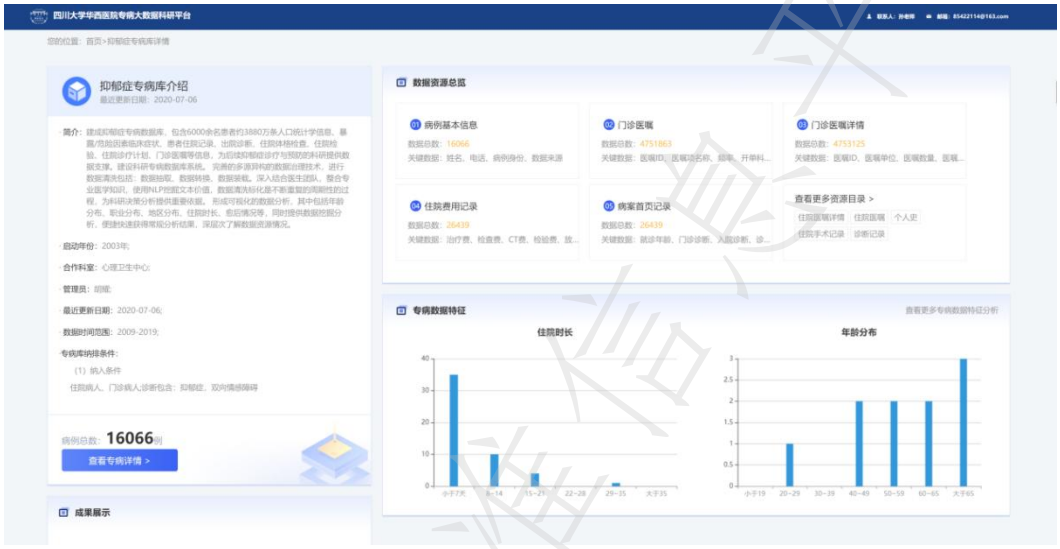


图 33 抑郁症专病数据集视图系统首页

### A. 6. 4. 2 抑郁症专病数据概览

支持抑郁症专病数据库集成病历信息数据的总体视图和患者视图。基于专病数据库用户可以通过临床统一视图实现病历信息的综合概览，全方位、全周期查看病例诊疗数据。并提供临床统一视图建立患者诊疗时序模型，包括患者在不同时间节点的诊断、用药、体征、检验、检查、手术、治疗等。支持以时间轴的方式展示病例全生命周期的诊疗概览并支持详细记录的查看。

### A. 6. 4. 3 专病数据资源目录

展示抑郁症专病数据集的数据资源目录信息，以及各个目录中变量的内容、属性、有效填充率、变量相关的统计、说明、所属目录、相关变量等。用户可通过预览目录数据样例了解数据资源情况，抑郁症专病数据资源目录如下图 34 所示：

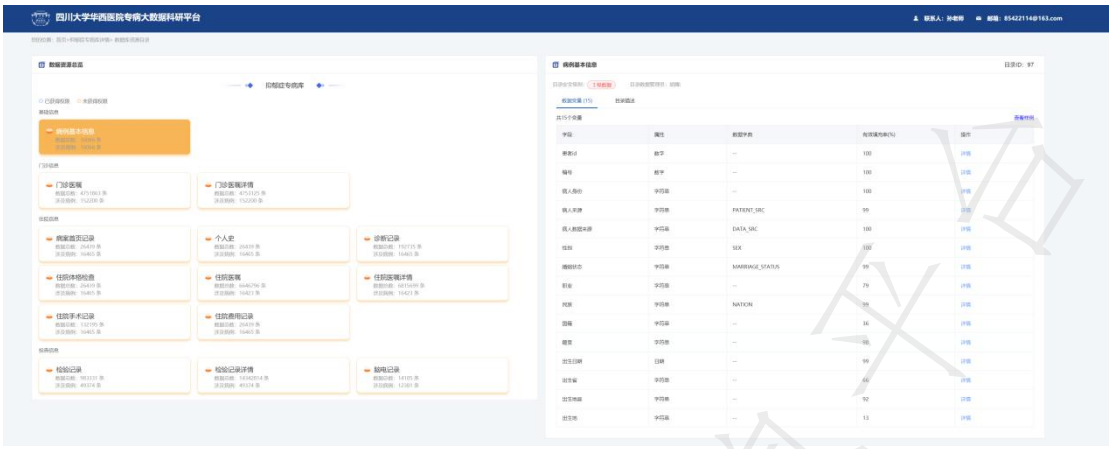


图 34 抑郁症数据资源目录

### A.6.4.4 专病特征统计

提取抑郁症专病数据集的关键性特征指标，包括年龄分布、住院时长、职业分布、性别统计、手术情况、愈后情况等指标，通过可视化的图表展示统计分析结果，可便于用户提前了解数据情况，抑郁症专病特征统计如下图 35 所示。


图 35 抑郁症专病特征统计