ICS号

中国标准文献分类号

团 体 标 准

团体标准编号T/BRA-CDCHE/ XXXX-XXXX

 代替的团体标准编号（修订填写）

胃癌专病队列数据集建设规范要求

Specifications for the Construction of the Gastric Cancer Disease-Specific Cohort Dataset

XXXX年XX月XX日 XXXX年XX月XX日实施

北京慢性病防治与健康教育研究会 发布

 **目****次**

前言……………………………………………………………………………………………Ⅱ

引言……………………………………………………………………………………………Ⅲ

1 范围…………………………………………………………………………………………1

2 规范性引用文件……………………………………………………………………………1

3 术语和定义…………………………………………………………………………………1

4 数据建模……………………………………………………………………………………2

5 数据采集……………………………………………………………………………………7

6 数据治理……………………………………………………………………………………10

7 数据存储及更新……………………………………………………………………………17

8 数据安全……………………………………………………………………………………18

附录A（资料性/规范性）胃癌数据集………………………………………………………20

参考文献………………………………………………………………………………………27

**前 言**

本文件参照GB/T1.1—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。

本文件由山东大学齐鲁医院提出。

本文件由北京慢性病防治与健康教育研究会归口。

本文件起草单位：山东大学齐鲁医院、复旦大学、北京大学肿瘤医院、复旦大学附属中山医院、北京协和医院、北京大学第三医院、山东大学临床研究中心、山东省医学会临床流行病学与循证医学分会。

本文件主要起草人：吕明、于文滨、陈兴栋、李文庆、陈世耀、杨红、丁士刚、刘联、杨孝荣、张同超、张媛、陈浩、闫治波、魏猛、殷晓霖、白浩、张雪凝、李真、崔晓晗、李紫琳。

1. **引****言**

我国是胃癌高发国家，患者数量在全球占有较大比例。然而胃癌领域仍缺乏大规模、多中心、统一规范的专病队列数据集，仅有少数医院建立了单中心胃癌专病队列数据集。全国范围内各医院使用的电子病历系统不同，导致数据格式一致性差，数据库融合共享可能性较低。此外，我国人口流动性大，随访数据难以收集，存在数据丢失的情况。因此，现有的胃癌专病队列数据集存在缺乏统一的构建、管理标准问题。如何构建规范的胃癌专病队列数据集，促进不同数据间的开放，避免数据孤岛的形成，避免信息资源的浪费，是亟需解决的问题。整合资源建立大数据平台，能提高胃癌三级预防的能力，从而降低我国胃癌严峻的疾病负担。

胃癌数据集建设标准规范了数据建模、数据采集、数据治理、数据存储及更新和数据安全五个方面。该标准的完成旨在通过全面系统整合医疗数据，提供胃肠肿瘤分类、诊断、治疗发展、临床决策、预后情况等所需信息；基于数据库的深度处理和分析，实现对胃癌的“防诊治康”全流程管理，全面提升医学科研、医疗服务的质量和效率。

1. **胃癌专病队列数据集建设规范要求**
2. 范围

本文件规定了可支撑胃癌标准数据集构建方法，包括数据建模、数据采集、数据治理、数据存储及更新和数据安全等。

本文件适用于本标准适用于医疗卫生机构和研究机构等胃癌数据集设计、研发和管理，其他相关领域可参考使用。

1. 规范性引用文件

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

ISBN 978-7-117-28858-3 胃癌标准数据集（2019版）

T/BISSC 01-2022 专科疾病标准数据集建设规范

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

NCCN 临床实践指南：胃癌指南（2018.V2）

中国临床肿瘤学会（CSCO）胃癌诊疗指南 2019

ICD-9-CM-3 世界卫生组织国际疾病分类手术码

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

GB/T 42755-2023 人工智能 面向机器学习的数据标注规程

DB3209/T 1259.1-2023 公共数据平台 数据治理规范 第一部分:总则

DB3206/T 1076-2024 医疗卫生公共信用平台建设技术规范

GB/T 36344-2018 信息技术 数据质量评价指标

GB/T22239-2019 信息安全技术 网络安全等级保护基本要求

GB/T35273-2020 信息安全技术 个人信息安全规范

1. 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

3.6

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]

3.7

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]

3.8

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。

3.9

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。

3.10

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。

1. 数据建模

4.1 临床讨论

制定数据模型之前，应召集临床医护人员（胃癌数据集使用人员、管理人员等）、临床研究方法学人员与数据建模工程师一起讨论，结合胃癌专病队列数据集的建库目标，共同确定数据范围。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。

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，如胃癌专病数据集建设需消化内科、胃肠外科、肿瘤科、检验科、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。临床研究方法学参与人员需符合以下资质要求：a) 具有统计学、流行病学或相关领域的硕士学位及以上学位，并有相关项目研究经验；b) 熟悉医学统计方法（如回归分析、生存分析等），并能熟练使用统计软件（如R、SAS、SPSS、Stata等）；c) 理解临床数据特性及数据质量控制方法，能够处理缺失值和异常值；d) 具有参与大型临床研究或医学数据分析项目的实际经验；e) 能与临床医生、数据建模工程师等相关人员有效沟通，提供统计支持和数据解释。

4.2 数据分层、分类设计

在胃癌数据集体系中，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。

结合胃癌特点，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，如表1：

表1 胃癌专病队列数据分类

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **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 | 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 | 随访随访 | 气候数据 | 蛋白质序列 |
| 科室信息 | 月经史信息 | 治疗随访 | 医保数据 | 基因组 |
| 医务人员信息 | 住院信息 | 伴随疾病随访 | …… | 蛋白质结构 |
| 医疗项目信息 | 诊断信息 | 生存状态随访 |  | …… |
| 医疗设备信息 | 门（急）诊信息 | 终末状态随访 |  |  |
| …… | 检验信息 | 转移及其他情况 |  |  |
|  | 手术信息 | …… |  |  |
|  | 护理信息 |  |  |  |
|  | 疾病进展与转归 |  |  |  |
|  | …… |  |  |  |

4.3 数据库概念模型设计

根据胃癌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，确定胃癌数据集的数据范围。

数据库概念设计应符合以下要求：

a) 不受数据来源限制，充分考虑胃癌 AI 应用研发需求；

b) 内容满足胃癌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。

4.4 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，形成胃癌数据集最终的数据

结构目录，包含以下步骤：

a) 数据库选型，一般数据量的胃癌数据集存储可选择 Mysql、PG 等关系型数据库，TB 级大型胃癌数据集可选择 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。

4.5 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。

4.6 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。

4.7 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2；

表2 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（GB/T2261.1-2003） |
| 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/T2659-2000） |
| 4 | 国家标准 | 中华人民共和国行政区划代码表（GB/T2260-2002） |
| 5 | 国家标准 | 职业分类与代码表（GB/T6565-1999） |
| 6 | 国家标准 | 专业技术职务代码表（GB/T8561-2001） |
| 7 | 国家标准 | 政治面貌代码表（GB/T4762-1984） |
| 8 | 国家标准 | 婚姻状况代码（GB/T2261.2-2003） |
| 9 | 国家标准 | 文化程度代码表（GB/T4658-1984） |
| 10 | 国家标准 | 学位代码表（GB/T6864-2003） |
| 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/T4761-1984） |
| 13 | 国家标准 | 健康状况代码表（GB/T2261.3-2003） |
| 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/T14396-2001） |

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。

胃癌数据集部分示例数据字段如表3：

表3 胃癌专病队列数据集的部分数据字典

|  |  |  |
| --- | --- | --- |
| 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） |
| 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 | 1-机关医师 2-机关护士 21-对讲师 31-技术药剂师 32-执业药师 33-其他药剂人员 41-检验人员 42-影像人员 43-医务管理人员 44-其他医务人员 60-其他技术人员 70-管理人员（1~70） |
| 10 | 管理服务 | 1-党委（副）书记 2-院（所、局）长 3-副院长（所、局）长 4-本科室主任 5-科室副主任（1~5） |
| 11 | 专业技术职务 | 专业技术职务代码（GB/T 8561-2001） |
| 12 | 医师资管 类别 | 1-临床 2-口腔 3-公共卫生 4-中医（1~4） |
| 13 | 所调动 类别 | 1-注册 11-考试注册 12-军转人员 13-其他注册 19-其他调动 20-调岗 21-出入境 22-辞职 23-继续教育 24-调配 29-其他 |
| 14 | 婚姻状况 | 婚姻状况代码（GB/T2261.2—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-学生/本科 4-大专 5-中专 6-高中 7-初中及以下（1~7） |
| 18 | 所学专业 | 1-基础医学 12-临床医学 13-内科学 14-口腔医学 15-公共卫生 16-护理 17-药学 18-预防医学 19-公共管理 20-人力资源管理 33-信息管理 41-经济学 42-会计学 43-统计学 51-法律 61-信息技术/计算机 62-工程管理（1~99） |
| 19 | 民族 | 民族代码（GB/T3304-1991） |
| 20 | 健康状况 | 健康状况代码（GB/T2261.3-2003） |
| 21 | 胃癌病理组织学类型 | 1-良性肿瘤 2-癌前 3-鳞癌 4-腺癌 5-腺鳞癌 6-印戒细胞癌 7-非上皮性癌 8-小细胞未分化癌 9-其他 |
| 22 | 医院等级 | 1-一级甲等 2-二级甲等 3-二级乙等 4-三级甲等 5-三级乙等 6-其他医院/专科医院（1~6） |
| 23 | 医院类别 | 1-综合医院 2-中医院 3-传染病医院 4-精神病医院 5-牙科医院 6-专科医院 7-护理院/康复中心 8-其他医院（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 | 有效区域代码（GB/T2260-2002） |
| 26 | 病情 | 1-危 2-急 3-一般（1~3） |
| 27 | 病原学现 状 | 1-标识 2-体检（1~2） |
| 28 | 疾病分类（ICD-10） | 疾病分类代码（GB/T 14396-2001） |
| 29 | 诊断依据 | 1-病理（包括骨髓） 2-脱落细胞（包括涂片） 3-X光 4-CT 5-超声波 6-核磁共振 7-超声波 8-实验室 9-其他 |
| 30 | 肿瘤分期 | 1-CT 2-N M\_ 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 7-其他（1~9） |
| 32 | 治疗结果 | 1-治愈 2-好转 3-未愈 4-死亡 5-其他（1~5） |
| 33 | 病案质量等级 | 1-甲 2-乙 3-丙（1~3） |
| 34 | 费用类别 | 1-CT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） |

1. 数据采集

5.1 数据源调研

根据胃癌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。是否明确分级数据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。

5.1.1 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。

数据 一般产生自临床相关业务系统，包括但不限于：

a) CDR/RDR:部分机构已经成立CDR（临床数据中心）和/或RDR（科研数据中心），可直接从CDR/RDR中获取患者的临床治疗相关数据。

b) HIS: 胃癌患者的门诊及住院的病情诊治、诊断、医嘱、检验、检查、手术等信息。

c) EMR: 胃癌患者的门诊及住院病历、病程记录、术前讨论、术后情况、手术小结、出院小结、会诊记录等全部文档。

d) 护理: 胃癌护理评估、护理单、护理记录、护理措施、危重记录、体征、PICC、置管等。

e) 手术麻醉: 胃癌患者的术前麻醉检查、术中麻醉、手术麻醉总结、麻醉效果评价、术后麻醉记录。

f) RIS: 胃癌患者的放射影像检查信息，检验项目、检查编号、细项项目及正常值范围。

g) LIS: 胃癌患者的检验项目及结果，包括CT/MRI/PET等检查及其报告原始文件。

h) 其他系统: 其他与胃癌相关的检验检查信息、检查记录、超声图像信息。

i) 医院信息系统: 可能包括各种院内会议纪要、会诊意见、专家意见等。

j) 其他: 影像信息库中的临床治疗数据。

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。

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。

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) 调研数据集成服务和系统调研方式。

d) 调研接口调研计划和更新周期。

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。

5.1.2 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。

5.1.3 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 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，如胃癌数据采集需“出院诊断”包含<胃恶性肿瘤>或<胃癌>，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。

5.3 回顾性数据采集

5.3.1 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。

5.3.2 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。

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胃癌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。

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。

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。

5.4 前瞻性数据采集

5.4.1 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胃癌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。

5.4.2 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。

1. 数据治理

6.1 数据预处理

6.1.1 结构化数据清洗

结构化数据是指可以使用二维表结构表示和存储的数据, 具有易于输入、存储、查询和分析的特点。数据清洗需要对数据进行准确性、完整性、一致性、唯一性、实时性、有效性的检验，过滤不符合规则要求的数据。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。包括以下工作：

1）空值处理：可捕获字段空值，进行加载或替换为其他含义数据；

2）数据标准：统一元数据、统一标准字段、统一字段类型定义；

3）数据拆分：依据业务需求做数据拆分，如身份证号，拆分区划、出生日期、性别等

4）数据验证：时间规则、业务规则、自定义规则；

5）数据替换：业务发生调整时，可实现无效数据、缺失数据的替换：

6）数据关联：明确数据关联规则，建立数据映射关系，确保不同数据源之间的数据一致性。

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。

数据清洗规则包括但不限于：

1）重复数据检验。数据库中属性值相同的记录被认定是重复数据，通过判断记录的属性值来检测记录是否相等，相等的数据合并为一条数据（合并/清除），非重复数据进入下一步数据处理环节；

2）缺失数据检验。对于信息目录要求必填项目进行检验，对检查中数据项缺失的数据，直接反馈给数据提供单位，经确认后重新提交：

3）不一致数据检验。以代码提供的数据，需要根据数据字典表进行比对，发现数据的不一致或者数据字典不存在的数据，直接反馈给数据提供单位，经确认后重新提交。

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。

胃癌专病结构化数据清洗流程如图所示：

图1 胃癌专病结构化数据清洗流程图

6.1.2 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。

6.1.2.1 标注任务确定

标注内容由标注需求方在标注任务说明中提供，标注任务说明一经确认，不可修改，如需修改则进入需求变更环节，标注任务应包括但不限于：

a）版本信息：明确当前版本编号、发布日期、发布人、发布说明（发布原因或选代原因）；

b）历史选代信息（历代版本编号、发布日期、发布人、发布说明等）；

c）项目背景：明确数据标注需求产生的原因，以及数据标注结果的应用场景：

d）任务描述：明确数据标注任务，包括数据形式、数据规模、标注规则、相关术语、标注样例、质量要求、指标计算方式、验收流程、交付时间等；

e) 主客观描述：明确说明数据标签是根据个人专业领域知识进行标注，还是客观认识进行的,标注；

f）标注人员资质：约定标注任务参与人员的资质要求；

g）标注结果：明确数据标注结果的交付形式；

h）知识产权：明确数据的知识产权归属。
6.1.2.2 待标注数据分析

数据标注前，数据需求方应对待标注数据进行分析，核对标注任务，包括：

a）数据核查：检查待标注数据是否与标注任务说明书中的数据定义相符，核查结果及时同步给数据需求方；

b）数据整理：建立完善的数据追踪机制，实现数据整理，以及最小粒度的数据追踪；

c）数据处理：根据标注任务以及标注数据的特性，通过数据聚类、组合排列、数据杂质去除等方法提高标注质量

6.1.2.3 标注任务评估

数据标注前，标注管理方应对标注任务进行评估，包括：

a）根据标注任务说明，评估标注任务可行性、标注规则合理性；

b）在数据需求方提供的小规模样本上进行预标注，将标注结果提交给数据需求方验收。在获得数据需求方确认后，再正式启动数据标注任务。

注：及时记录数据预标注流程中标注规则与数据相悖、覆盖不全或规则之间相悖的情况，并向数据需求方反馈完善标注规则。

6.1.2.4 标注需求变更

标注需求方需求变更时，应在标注管理方评审同意后更新标注任务说明，重新进入标注任务评估阶段。

6.1.2.5 标注人员

a）标注人员范围：数据标注方应根据标注任务内容中规定的标注人员资质需求，确定符合要求的人员进入标注人员培训环节。

b）标注人员培训：数据标注方应根据标注任务说明，对标注人员进行岗前能力培训。标注能力考试合格者，方能参与标注任务。

c）标注人员能力档案：数据标注方应建立标注人员能力档案，记录标注人员承担标注任务的相关内容，用于进行标注人员能力评估与标注质量追踪。

6.1.2.6 标注环境

a）选择标注工具：数据标注方应根据标注任务难度、数据处理规模及数据属性特征、数据安全控制层级与方式，合理选择标注工具，完成数据标注任务。

b）选择标注场地：数据标注方应根据标注任务中必要的数据安全要求，搭建数据标注场地。

c）标注软件：使用稳定的专业医学标注软件进行标注工作，可灵活配置标注参数，以不断优化标注过程。

6.1.2.7 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。

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。具体数据标注需求和条件请专家结合领域知识设计。

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。

6.1.2.8 标注任务质量检查

在标注过程中，应采用多种检查方法对标注任务质量进行检测，对不满足标注任务要求的，及时预警反馈，并查明问题原因。根据项目特性，标注任务质量检查方法可归纳为以下三种。

a）机器验证：在任务进行期间，安排超过一名人员做同一个子任务，选择出最优、最正确的标注结果。结果选择可通过下列方式。

1）标注工具自动选择：通过与标注工具匹配的模型推理，或拟合若干个标注结果，选择其中置信度最高的标注结果，作为最终结果；

2）人工辅助选择：人工对多个标注结果进行对比，从而挑选出置信度最高的标注结果作为最终结果。对于需要特定专业知识标注的领域，进行人工辅助标注时应以多个专家的共同商议结果作为最终结果。

b）埋题验证：在任务进行期间，除了常规标注子任务外，在任务中混进若干已知结果的测试题，以此验证标注质量。在此操作的过程中注意以下事项。

1）针对数据特征专题专用：对于理题验证，应保障测试题在真题中间处于混淆的状态。因此，在出题的过程中，应针对数据的自身特征（数据的类别、场景、内容等），准备相应的题目，避免题目暴露于操作者，失去验证的效果。

2）限制题目的使用次数：为避免题目多次出现，引起被测者的注意，从而失去验证效果，应限制题目的使用次数。尤其是拥有容易记忆的特征点的题目（如特定脸部特征、特定文字、特定场景等），应严格限制出现的次数。

c）标注人员状态验证：通过对标注人员的操作规范性、实时注意力状态、标注准确率等方面进行检查与监测，及时发现操作违规问题，保证数据质量；在发现操作违规问题、数据质量有下降时，应根据时间段等特征，对标注人员在这一状态内操作的标注数据进行检查或者返工等操作。

d）标注任务一致性检查：在标注任务进行期间应使用统计规则或模型验证等方法，得到标注任务一致性水平，一旦发现离群点或明显的降低趋势，及时对标注人员预警和警告。

6.2 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

6.3 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。

6.3.1 基本要求

数据质量应贯穿数据的设计、产生、汇聚、应用阶段。数据质量应明确组织层面的数据质量目标，统一数据质量需求相关模板和管理机制。

6.3.2 数据质量内容

数据质量包含以下内容：

a）数据质量需求：

1）制定完善的数据质量管理方法，包括质量目标制定、质量检查、质量分析、质量评估,质量提升等环节；

2）建立数据质量管理机构和机制，明确质量管理的角色和职责；

3）设计组织统一的数据质量评价体系以及相应的规则库。

b）数据质量检查：

1）制定数据质量检查计划，明确相关人员责任，对数据进行定期的质量检查和评估；

2）明确数据质量检查方法，一般有人工对照校核、程序自动校核和人机交互等；

3）识别数据生命周期各个阶段的数据质量关键因素，构建数据质量评估框架（包含但不,限于数据的准确性、完整性、一致性、可访问性、及时性、相关性和可信度）。

c）数据质量分析：

1）制定组织层面的数据质量问题评估分析方法，制定统一的数据质量报告模板，明确数据质量问题分析的要求；

2）制定数据质量问题分析计划，对关键数据质量问题的根本原因、影响范围进行分析，定期输出数据质量分析报告。

d）数据质量提升：

1）制定组织层面的数据质量提升制度，明确数据质量提升方案的构成；

2）定期开展数据质量提升工作，针对重点质量问题进行汇总分析，制定数据质量提升方,案，从业务流程优化、系统改进、制度和标准完善等层面进行提升。

6.3.3 数据质量评价过程

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。

数据质量评价过程如图：



图2 数据质量评价流程图

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。

6.3.4 数据质量评价指标框架

数据质量评价指标应符合GB/T 36344-2018中数据质量评价指标的定义。具体框架见图。



图3 数据质量评价指标框架

1. 数据存储及更新

7.1 数据存储

实现不同类型数据归档存储的技术支撑包括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。

7.2 数据更新

7.2.1 更新方式

对于数据更新方式，主要有全量更新和增量更新两种。对于首次报送平台的数据，应采用全量更新的方式，一次性报送所有历史数据。对于更新的数据，应采用增量更新方式，按照数据时间戳分批次报送新数据到平台。

7.2.1.1 全量更新

全量更新宜采用直接覆盖目标数据集的方式。全量更新数据处理应采用抽取方法直接处理，通过ETL工具，直接抽取数据覆盖到目标数据集。

7.2.1.2 增量更新

增量更新数据处理应抽取自上次抽取以来数据库中要抽取的表中新增、修改、删除的数据。一般增量数据抽取方法有以下两种：

a）触发器。在要抽取的表上建立需要的触发器，一般要建立插入、修改、删除三个触发器，每当源表中的数据发生变化，就被相应的触发器将变化的数据写入一个临时表，抽取线程从临时表中抽取数据。触发器方式的优点是数据抽取的性能较高，缺点是要求在业务数据库中建立触发器，对业务系统有一定的性能影响；

b）时间截。它是一种基于递增数据比较的增量数据捕获方式，在源表上增加一个时间戳字段，更新修改表数据的时候，同时修改时间戳字段的值。当进行数据抽取时，通过比较系统时间与时间截字段的值来决定抽取哪些数据。

7.2.2 异议数据处理

当监管单位、被监管机构对平台所提供的数据结果存在异议时，可以通过平台或维护渠道对数据提出异议。根据异议情况，平台运维团队将对数据问题进行排查，确认异常发生原因以及解决方案，反馈给到异议提出方。如果出现对敏感数据有重大异议时，应将对应用进行下线处理，待异议确认和修复后重新上线。

1. 数据安全

8.1 数据安全基本原则

a) 数据最小化：应明确数据的使用目的及所需范围，提供适当的管理和技术措施保证只采集和处理满足目的所需的最小数据;

b) 责任不随数据转移：当前控制数据的机构应对数据负责，当数据转移给其他机构时，责任不随数据转移而转移;

c) 最小授权：应控制数据活动中的数据访问权限，保证在满足业务需求的基础上最小化权限，并及时回收过期的数据访问权限;

d) 可审计：应实现对数据使用和业务各环节的数据审计，记录数据活动中各项操作的相关信息，且保证记录不可伪造和篡改，数据活动的所有操作可追溯。

8.2 数据加密

专病数据集在存储前要加密以防止重要数据丢失或盗用，传输过程中对数据进行加密也同样重要。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校验功能，对用户输入信息进行安全检查，降低数据库管理员权限被攥改、滥用等数据安全风险。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。

8.3 数据脱敏

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，敏感数据或身份数据信息需要隐藏，避免信息泄漏。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。

8.4 数据备份

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。

8.5 其他

数据集应符合GB/T22239—2019中等级保护三级的要求，个人信息安全应符合GB/T35273—2020要求。

1. **附 录 A 胃癌数据集**

（资料性/规范性）

附录名称

建成胃癌专病数据库，收集人口统计学信息、就诊记录、病史信息、体格检查、专科检查、诊断、检验、检查、手术治疗、术后恢复与并发症、病理、内科治疗、放射治疗、随访、不良反应等信息，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，为后续胃癌诊疗与预防，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。胃癌数据集主要包括但不限于以下内容：

1. 患者人口学信息

病案号、性别、姓名、性别、民族、职业类别、出生日期、发病年龄、ABO血型、RH血型、籍贯市、籍贯省、文化程度、婚姻、身份证件号码、本人电话、常住地址、户口地址、是否手术、手术单位、手术前治疗标志、手术前治疗方式、手术日期。

2. 就诊记录

入院日期、入院科室、主要诊断、出院日期、出院科室、是否转科、转科科别、主治医生、就诊日期、就诊科室、主要诊断、主治医生。

3. 病史信息

症状、症状病程、症状初发到确诊时间、病情进展及加重时间、伴发完全梗阻、伴发不完全梗阻、伴发出血、伴发穿孔、体重变化、体重变化数值、体重流失程度、体重变化时长、院前初检、院前手术、院前用药、院前辅助检查、院前治疗史、腹腔内手术具体部位、腹腔内手术具体器官、腹腔外手术史、腹腔外手术具体部位、腹腔外手术具体器官、ESD手术史、ESD手术日期、糖尿病、高血压、肝炎、结核、呼吸系统慢性病、心血管系统慢性病、消化系统慢性病、肾脏系统慢性病、脑血管系统疾病、免疫系统慢性病、其他疾病、其他手术史、其他肿瘤史、肿瘤名称、肿瘤发生时间、过敏史、过敏原名称、输血史、输血反应、输血成分、非肿瘤家族史、疾病名称、关系类型、肿瘤家族史、肿瘤名称、关系类型、饮酒史、饮酒种类、戒酒时长、饮酒时长、吸烟史、吸烟种类、戒烟时长、吸烟时长、高盐食物、盐渍食物、烟熏食物、其他饮食偏好、饮食偏好时长。

4. 体格检查

舒张压、心率、脉搏、体表面积、ECOG评分、Karnofsky评分、WHO体力状态评分、ASA评分、NRS 2002营养风险筛查评分。

5. 专科检查

检查日期、检查内容、腹部压痛、腹部压痛部位、腹部包块、腹部包块位置、腹部包块大小、腹部包块移动度、腹肌抵抗、腹肌抵抗部位、胃型、直肠指诊、盆腔积液、结节活动度、冰冻盆腔、指示密集、振水音、左侧肋下脾包块大小、移动性浊音、其他阳性体征。

6. 诊断

病灶数量、大体分型 - 早期胃癌、大体分型 - 进展期胃癌（Borrmann分型）、病灶位置 - 纵轴、病区组合、病灶位置 - 横轴、病理确诊、T分期、N分期、M分期、M分期的具体部位、TNM分期、诊断名称、诊断日期、诊断类型、分期类型、诊断名称、诊断日期、诊断来源、诊断医师。

7. 检验

检验日期、检验定性结果、检验定量结果、检验定量结果单位、正常参考值范围、白细胞（WBC） - 静脉血、白细胞计数（WBC） - 静脉血、红细胞（RBC） - 静脉血、红细胞计数（RBC） - 静脉血、血红蛋白（Hb）测定 - 静脉血、血小板（PLT） - 静脉血、中性粒细胞绝对值（NEUT#） - 静脉血、中性粒细胞百分比（NEUT%） - 静脉血、淋巴细胞绝对值（LY#） - 静脉血、淋巴细胞百分比（LY%） - 静脉血、单核细胞绝对值（MONO#） - 静脉血、单核细胞百分比（MONO%） - 静脉血、平均红细胞体积（MCV） - 静脉血、平均红细胞血红蛋白（MCH） - 静脉血、平均红细胞血红蛋白浓度（MCHC） - 静脉血、嗜酸性粒细胞绝对值（EO#） - 静脉血、嗜酸性粒细胞百分比（EOS%） - 静脉血、嗜碱性粒细胞绝对值（BASO#） - 静脉血、嗜碱性粒细胞百分比（BASO%） - 静脉血、红细胞体积分布宽度（RDW-CV） - 静脉血、红细胞分布宽度（RDW-SD） - 静脉血、血小板分布宽度（PDW） - 静脉血、平均血小板体积（MPV） - 静脉血、大血小板比率（P-LCR） - 静脉血、红细胞比容（Hct） - 静脉血、红细胞压积（PCV） - 静脉血、血小板压积（PCT） - 静脉血、酸碱度（pH）测定 - 尿液、比重测定SG - 尿液、白细胞WBC（尿流式）- 尿液、尿糖GLU - 尿液、尿酮体KET - 尿液、尿胆原URO - 尿液、胆红素BIL - 尿液、隐血（HB）- 尿液、红细胞RBC - 尿液、白细胞LEU（干化学）- 尿液、尿蛋白定性试验PRO - 尿液、尿蛋白定量 - 尿液、上皮细胞 - 尿液、隐血试验 - 便检、便常规、临床试验 - 便检、葡萄糖（Glu）- 静脉血、乳酸脱氢酶（LD）测定 - 静脉血、天门冬氨酸氨基转移酶（AST）测定 - 静脉血、丙氨酸氨基转移酶（ALT）测定 - 静脉血、总胆红素（TBIL）测定 - 静脉血、结合胆红素（DBIL）测定 - 静脉血、非结合胆红素（IBIL）测定 - 静脉血、碱性磷酸酶（ALP）测定 - 静脉血、γ-谷氨酰基转移酶（GGT）测定 - 静脉血、α-羟丁酸脱氢酶（HBDH）测定 - 静脉血、总蛋白（TP）测定 - 静脉血、白蛋白（Alb）测定 - 静脉血、钾（K+）- 静脉血、钠（Na+）- 静脉血、钙（Ca2+）- 静脉血、无机磷（P）- 静脉血、镁（Mg2+）- 静脉血、肌酐（Cr）测定 - 静脉血、尿素（Urea）- 静脉血、尿酸（UA）测定 - 静脉血、尿素氮肌酐比（BUN/Cr）- 静脉血、总胆固醇（TC）测定 - 静脉血、高密度脂蛋白胆固醇（HDL-C）测定 - 静脉血、低密度脂蛋白胆固醇（LDL-C）测定 - 静脉血、极低密度脂蛋白胆固醇（VLDL-C）测定 - 静脉血、非高密度脂蛋白胆固醇（nHDL-C）测定 - 静脉血、甘油三酯（TG）测定 - 静脉血、尿素氮（Bun）测定 - 静脉血、球蛋白（Glb）测定 - 静脉血、白蛋白/球蛋白（ALB/GLO）比值 - 静脉血、前白蛋白（PA）测定 - 静脉血、谷氨酰胺氨（GDN）测定 - 静脉血、C反应蛋白（CRP）测定 - 静脉血、胃蛋白酶原（PG）测定 - 静脉血、促甲状腺素（T-G7-17）测定 - 静脉血、铁蛋白（SF）测定 - 静脉血、维生素B12（VitB12）测定 - 静脉血、血浆D-二聚体（D-Dimer）测定 - 静脉血、凝血酶原时间（PT）测定 - 静脉血、活化部分凝血活酶时间（APTT）测定 - 静脉血、凝血酶时间（TT）测定 - 静脉血、纤维蛋白原测定（FIB）- 静脉血、血浆抗凝血酶III（AT-III）测定 - 静脉血、纤维蛋白原降解产物（FDP）测定 - 静脉血、凝血酶原活动度（PTA）测定 - 静脉血、凝血酶原国际标准化比率（INR）- 静脉血、凝血酶原时间比率（PTR）测定 - 静脉血、活化部分凝血活酶时间比率（APITTR）测定 - 静脉血、糖类抗原CA12-5测定 - 静脉血、糖类抗原CA19-9测定 - 静脉血、糖类抗原CA72-4测定 - 静脉血、糖类抗原CA15-3测定 - 静脉血、癌胚抗原（CEA）测定 - 静脉血、甲胎蛋白（AFP）测定 - 静脉血、神经元特异性烯醇化酶（NSE）测定 - 静脉血、VEGF测定 - 静脉血、乙型肝炎表面抗原定性（HBsAg）测定 - 静脉血、乙型肝炎表面抗原定量（HBsAg）测定 - 静脉血、乙型肝炎e抗体定性（抗HBe）测定 - 静脉血、乙型肝炎e抗体定量（抗HBe）测定 - 静脉血、EB病毒VCA-IgA抗体（VCA-IgA）- 全血、梅毒螺旋体抗体（FTA-ABS）测定 - 静脉血。

8. 检查

心电图：检查方法名称、检查日期、检查所见、检查结论、是否异常、期前收缩、期前收缩类型、传导阻滞、阻滞类型、劳损肥厚、劳损肥厚部位、心肌缺血、ST-T改变、陈旧性心肌梗死。

X线：检查日期、检查项目名称、检查部位、检查所见、检查结论。

上消化道造影：检查日期、检查项目名称、检查部位、病灶位置 - 纵轴、病灶位置 - 横轴。

CT：检查日期、检查项目名称、检查部位、检查部位分类、检查所见、检查结论、原发灶位置 - 纵轴、原发灶位置 - 横轴、原发灶大小、可见肿块器官部位、淋巴结大小、淋巴结评价情况、腹水、浆膜面情况、浸润脏器、术后改变、复发转移、转移脏器、肝转移单发/多发、转移部位、转移灶个数、转移灶大小、浸润脏器、MRI：检查日期、检查部位、检查部位分类、检查项目名称、检查所见、检查结论、原发灶位置 - 纵轴、原发灶位置 - 横轴、原发灶大小、可见肿块器官部位、淋巴结大小、淋巴结评价情况、腹水、浸润脏器、术后改变、复发转移、转移脏器、肝转移单发/多发、转移部位、转移灶个数、转移灶大小、PET-CT：检查日期、检查部位、检查部位分类、检查项目名称、检查所见、检查结论、原发灶位置 - 纵轴、原发灶位置 - 横轴、原发灶大小、可见肿块器官部位、淋巴结大小、淋巴结评价情况、腹水、浸润脏器、术后改变、复发转移、转移脏器、肝转移单发/多发、转移部位、转移灶个数、转移灶大小、转移灶位置、骨扫描：检查日期、检查项目名称、检查所见、检查结论、转移部位、心脏超声：检查日期、检查项目名称、心功能 - 左室舒张末容积、心功能 - 左室收缩末容积、心功能 - 射血分数、心功能 - 心排血量、心功能 - 心指数、心功能 - 每搏量、心功能 - 每搏输出量、检查所见、检查结论、颈部超声：检查日期、检查项目名称、检查部位、颈部淋巴结大小、颈部淋巴结提示转移、其他超声：检查日期、检查项目名称、检查部位、检查所见、检查结论。

胃镜：检查日期、检查项目名称、检查所见、检查结论、病灶大小、大体分型 - 早期胃癌、大体分型 - 进展期胃癌（Borrmann分型）、齿状线距门齿距离、病灶上缘距齿状线距离、病灶位置 - 纵轴、病灶位置 - 横轴、Hp试验、超声胃镜：检查日期、检查项目名称、检查所见、检查结论、齿状线距门齿距离、病灶位置 - 纵轴、病灶位置 - 横轴、病灶所侵厚度层级、病灶所在层面标志、病灶所在层面标志线、胃周肿大淋巴结、腹水。

9. 手术治疗

手术基本信息：手术开始时间、手术结束时间、手术时间、手术名称、手术类别、手术者、第一助手、麻醉方式、术中诊断、失血量、有无输血、输血量、输入成分、术中T分期、术中N分期、术中M分期、M的具体部位、如T4b，具体累及器官/组织。

手术入路：手术入路、腹腔镜观察孔、气腹压力范围、中转开腹、中转开腹原因。

手术性质：手术性质、姑息手术原因、姑息手术、减瘤手术、探查结论原因、标本来源、取材部位。

腹腔检查情况：腹腔检查情况、腹水量、腹水颜色、粘连部位、受粘连累及器官、其他。

胃切除重建：切除方式、主要重建方式、重建具体方式、淋巴结清扫范围、淋巴结清扫范围+D2+范围、根治切除、切口长度、切口位置、联合切除、切除器官、切除器官-原因、十二指肠残端加固、食管空肠吻合口加固。

术中并发症：术中并发症、有无出血、损伤修复、其他并发症。

术中治疗情况：术中治疗、术中输血、热灌注药物名称、热灌注药物剂量、热灌注药物灌注时间、热灌注药物灌注温度、热灌注药物灌注速度。

标本解剖信息：病灶数量、病灶标本名称、病灶标本脏器、病灶标本脏器数目、病灶大小、标本解剖信息、病灶标本位置-纵轴、病灶标本位置-横轴、Siewert分型。

术中可疑转移淋巴结：判断依据、可疑淋巴结转移数、1#贲门右侧淋巴结、2#贲门左侧淋巴结、3#胃小弯淋巴结、4sa#胃左动脉旁淋巴结、4sb#胃短血管旁淋巴结、4d#胃网膜左血管旁淋巴结、5#幽门上区淋巴结、6#幽门下区淋巴结、7#胃左动脉干旁淋巴结、8a#肝总动脉干旁上区淋巴结、8p#肝总动脉干旁下区淋巴结、9#腹腔动脉干旁淋巴结、10#脾门淋巴结、11p#脾动脉旁近端淋巴结、11d#脾动脉旁远端淋巴结、12a#肝十二指肠韧带内肝总动脉旁淋巴结、12b#肝十二指肠韧带内肝固有动脉旁淋巴结、12p#肝十二指肠韧带内门静脉旁淋巴结、13#胰头后淋巴结、14v#肠系膜上静脉旁淋巴结、14a#肠系膜上动脉旁淋巴结、15#结肠中动脉旁淋巴结、16a1#主动脉裂孔上区淋巴结、16a2#主动脉裂孔下区淋巴结、16b1#腹主动脉旁中区淋巴结、16b2#腹主动脉旁下区淋巴结、17#胰下淋巴结、18#结肠旁淋巴结、19#膈下淋巴结、20#食管裂孔旁淋巴结、110#肝门淋巴结、111#肝后上区淋巴结、112#后胰头上区淋巴结、113#后胰头下区淋巴结、114#十二指肠旁淋巴结、115#直肠旁淋巴结、116#骶前淋巴结、其他可疑淋巴结。

10.术后恢复及并发症

术后恢复过程：术后入住ICU、术后入住ICU原因、转入ICU时间、转出ICU时间、ICU治疗时间、特殊诊疗措施、总住院天数、术后住院天数、术后体重、首次离床日期、肠功能恢复（肠鸣音/排气）日期、恢复至饮食日期、恢复至饮水日期、是否留置胃管、胃管留置日期、胃管引流量、胃管引流物、拔出胃管置留胃管、重置胃管时间、是否留置引流管、腹腔引流日期、腹腔引流量、拔除引流管日期、是否输血、输血量、输入成分、术后疼痛评分 - POD#1、术后疼痛评分 - POD#2、术后疼痛评分 - POD#3、术后疼痛评分 - 出院前一天、是否早期并发症（术后30天）。

术后早期并发症：切口问题、腹腔感染/脓肿、腹腔内出血、肠梗阻、吻合口漏、吻合口狭窄、十二指肠残端瘘、十二指肠残端出血、消化道出血、胆漏、胰瘘、腹腔炎症、肺部感染、泌尿系感染、甲状腺、肝衰竭、心脑血管意外、深静脉血栓、其他并发症、并发症发生日期、干预措施、并发症严重程度分级。

出血：是否出血、发生时间、出血部位、出血部位 - 其他、处理方法、输血、输血量、输血成分。

消化道瘘：是否消化道漏、发生时间、漏病灶、处理方法 - 手术、处理方法 - 保守、食管空肠漏、食管吻合漏、术后梗阻：是否术后梗阻、发生时间、梗阻部位、梗阻原因、处理方法 - 手术、处理方法 - 保守。

胃瘫：是否胃瘫、动力恢复时间、切口情况：是否脂肪液化、是否感染、是否有积液、是否切口肥大、是否切口裂开、裂开程度、裂开部位、切口感染、切口感染情况、切口情况。

腹腔感染：是否腹腔感染、感染部位、处理方法、处理方法 - 手术、吻合口瘘病灶、吻合口瘘病灶 - 养料来源。

胶质感染：是否胶质感染、是否骨髓炎、处理方法、处理方法 - 手术、呼吸系统：是否肺炎、是否肺不张、是否呼吸衰竭、是否呼吸功能恢复、处理措施。

心脑血管系统：是否心衰、是否心房纤维颤动。

静脉性并发症：是否周围血管性栓塞、是否肠系膜血管性栓塞、是否静脉性栓塞、是否下肢静脉栓塞、其他静脉性并发症。

其他：处理方法、其他并发症。

11. 病理

基本信息：检查日期、送检标本类型、病理号、标本离体时间。

基本信息：标本固定时间、标本取材时间、大体检查、大体检查 - 早期胃癌、大体分型 - 进展期胃癌 (Borrmann 分型)。

标本描述：标本来源、胃大弯长度、小弯长度、近切缘、远切缘、胃腔内容、肿瘤距近切缘、肿瘤距远切缘。

肿瘤位置：纵轴、横轴。

病理分型：组织学分型、Lauren 分型、组织学分级。

其他病理所见：其他病理所见、具体内容。

放化疗反应：放化疗反应。

胃壁侵犯层次：胃壁侵犯层次 (pT 分期)。

胃壁侵犯层次：侵犯邻近脏器组织。

淋巴结：1群小弯右淋巴结、2群小弯左淋巴结、3群小弯淋巴结、4a群胃短血管淋巴结、4b群胃左动脉沿胃大弯走行淋巴结、4d群小弯左侧淋巴结、5群胃右淋巴结、6群胃左淋巴结、7群胃总动脉旁淋巴结、8a群肝总动脉上游淋巴结、9群脾门淋巴结、10群脾下淋巴结、11群脾周围淋巴结、12a群肝十二指肠韧带淋巴结、12b群肝十二指肠韧带静脉后方右肾门淋巴结、12p群肝十二指肠韧带门门静脉旁淋巴结、13群胃后壁左侧淋巴结、14a群肠系膜上动脉旁淋巴结、14v群肠系膜上静脉旁淋巴结、15群小弯左侧及远端脾门淋巴结、16a1群主动脉旁淋巴结、16a2群主动脉下淋巴结。

远处转移：pM 分期、pM 肝转移、pP 腹膜转移。

其他浸润：CY 分期 - 血管、CY 分期 - 淋巴管、CY 分期 - 神经。

免疫组化：Her-2、FISH、VEGF、EGFR、CD31、CD34、D2-40、S100、Ki-67、MMR、MSI、PDL1、CDH1。

脱落细胞学：脱落细胞学。

分子分型：TCGA 分型。

病理分期：pT 分期、pN 分期、pM 分期、M的具体部位、总分期。

12. 内科治疗。

化疗方法：化疗目的、开始时间、结束时间、化疗药物。

化疗方法：化疗剂量、剂量单位、给药途径。

化疗效果：化疗效果评估结果、化疗效果评估手段、化疗效果评估时间、不良反应。

化疗方案：化疗方案。

免疫治疗方法：开始时间、结束时间、免疫治疗药物、免疫治疗剂量、剂量单位、给药途径。

免疫治疗效果：免疫治疗效果评估结果、免疫治疗效果评估手段、免疫治疗效果评估时间、不良反应。

靶向治疗方法：开始时间、结束时间、靶向治疗药物、靶向药物剂量、剂量单位、给药途径。

靶向治疗效果：靶向治疗效果评估结果、靶向治疗效果评估手段、靶向治疗效果评估时间、不良反应。

肠外营养：开始时间、结束时间、营养药物、药物剂量、剂量单位、给药途径。

营养支持：营养支持类型、不良反应。

抗感染治疗：开始时间、结束时间、用药或吸入用名、用药或吸入商品名、每次用量、剂量单位、给药途径、下医嘱科室、针对感染部位。

抑酸护胃治疗：开始时间、结束时间、用药或吸入用名、用药或吸入商品名、每次用量、剂量单位、给药途径。

13. 放射治疗

放疗记录：是否放疗、放疗目的、开始时间、结束时间。

放疗记录：放疗技术、放疗部位处方剂量、放疗部位处方照射次数、放疗部位实际照射剂量、放疗部位实际照射次数、疗效评估。

14. 随访

患者随访详情：随访日期、随访次数、随访类型、受访人、生活情况、是否肿瘤复发、复发部位、复发时间。

患者随访详情：是否远处转移、远处转移时间、转移部位、主要症候、处理情况、死亡原因、死亡时间。

15. 不良反应

胃肠道症状：腹痛（最大程度）、腹泻（发生日期）、便秘（最大程度）、便秘（发生日期）、口腔黏膜溃烂（最大程度）、口腔黏膜溃烂（发生日期）。

肝损害：肝损害（最大程度）、肝损害（发生日期）。

泌尿系统损伤：肾损害（最大程度）、肾损害（发生日期）、血尿（最大程度）、血尿（发生日期）、蛋白尿（最大程度）、蛋白尿（发生日期）、非传染性膀胱炎（最大程度）、非传染性膀胱炎（发生日期）。

代谢与营养：高钙血症（最大程度）、高钙血症（发生日期）、低血钙（最大程度）、低血钙（发生日期）、低钠血症（最大程度）、低钠血症（发生日期）、低钾血症（最大程度）、低钾血症（发生日期）。

神经毒性：感觉异常（最大程度）、感觉异常（发生日期）、运动异常（最大程度）、运动异常（发生日期）、皮质症（最大程度）、皮质症（发生日期）、小脑损伤（最大程度）、小脑损伤（发生日期）、情绪异常（最大程度）、情绪异常（发生日期）、头痛（最大程度）、头痛（发生日期）、听力异常（最大程度）、听力异常（发生日期）、视力异常（最大程度）、视力异常（发生日期）。

皮肤：脱发（最大程度）、脱发（发生日期）、瘙痒症（最大程度）、瘙痒症（发生日期）、光敏感（最大程度）、光敏感（发生日期）、红皮病（最大程度）、红皮病（发生日期）。

其他：过敏（最大程度）、过敏（发生日期）、输液相关反应（最大程度）、输液相关反应（发生日期）、发热（最大程度）、发热（发生日期）、寒战（最大程度）、寒战（发生日期）、水肿（最大程度）、水肿（发生日期）、乏力（最大程度）、乏力（发生日期）、其他（最大程度）、其他（发生日期）。

1. **参 考 文 献**

[1] WS 445-2014 电子病历基本数据集

[2]WS/T 671-2020 国家卫生与人口信息数据字典

[3]ISBN 978-7-117-28858-3 胃癌标准数据集（2019版）

[4]T/BISSC 01-2022 专科疾病标准数据集建设规范

[5]NCCN 临床实践指南：胃癌指南（2018.V2）

[6]中国临床肿瘤学会（CSCO）胃癌诊疗指南 2019

[7]ICD-9-CM-3 世界卫生组织国际疾病分类手术码