医疗机构
医疗大数据平台
建设指南

中国医院协会
信息专业委员会

《医疗机构医疗大数据平台建设指南》编辑委员会

主审：王才有　薛万国

主编：衡反修、王力华

编者：（按姓氏笔画排序）

王力华　计　虹　田宗梅　包国峰　朱卫国　刘会文
刘敏超　张　军　张琼瑶　袁　浩　夏洪斌　徐　浩
曹晓均　路　健　薛万国　衡反修

文字编辑：朱小兵　王立军　黄伊玮

序　言

2017 年 12 月，习近平总书记在中共中央政治局第二次集体学习时强调，大数据发展日新月异，我们应该审时度势、精心谋划、超前布局、力争主动，深入了解大数据发展现状和趋势及其对经济社会发展的影响，分析我国大数据发展取得的成绩和存在的问题，推动实施国家大数据战略，加快完善数字基础设施，推进数据资源整合和开放共享，保障数据安全，加快建设数字中国，更好服务我国经济社会发展和人民生活改善。

医学是数据密集型行业，无论是公共卫生、临床服务、医学研究都离不开数据循证的支撑：一方面是医疗健康活动中产生大量数据；另一方面是这些数据对于提升医疗质量，有效控制费用，保障医疗安全具有潜在的价值。医疗保健领域的大数据可广泛应用于疾病流行预测，临床治疗服务，改善居民健康方式，对于提升人民健康福祉，满足人民日益增长的健康需求，可发挥出重要作用。

围绕健康中国发展战略，国务院、国家卫生健康委发布了一系列政策法规，明确了医疗健康大数据发展目标及其应用功能规范，极大地推动了大数据在医疗健康行业中的应用发展。各医疗机构也根据自身发展需要，开展不同形式的大数据应用实践，特别是以北京、上海、福建等发达地区为代表，已经初见成效。

但是在医疗健康大数据应用的发展中，大家也面临着诸多挑战，包括：数据标准开发和应用、合理的数据技术、严格的数据安全、完善的数据治理、适宜的应用场景等。为了规范和指导我国医疗机构医疗大数据建设及应用工作，中国医院协会信息专业委员会（CHIMA）专门设立研究课题，组织专家撰写了《医疗大数据建设指南》（简称《建设指南》）一书。本指南的编制目的是为医院、医院管理机构、医疗大数据厂商提供医疗大数据建设、应用、运维工作的参考方案，帮助医疗信息工作者面对医疗大数据发展建设挑战、解决当前医疗大数据建设中的问题，推广医疗大数据的建设和应用。

《建设指南》共包括四个章节：第一章引言，介绍医疗大数据发展的政策背景、现状和问题，以及医疗大数据及平台的概念、意义和作用；第二章总体设计，明确医疗大数据平台的需求和建设目标，介绍医疗大数据平台的总体架构、技术路线和

功能范围；第三章建设要点，详细介绍了医疗大数据平台建设的安全体系、硬件部署、数据接入以及数据治理策略；第四章应用场景，从临床应用、科研应用、医院管理、患者管理和药物临床试验等方面介绍了医疗大数据的建设及应用。最后，本指南还设有附录，包括：专门为本指南主题征集的专家观点，回应了业界关心的医疗大数据的建设中存在的热点、难点问题，以及建设经验；医院大数据平台的建设案例，大数据平台的数据管理制度等。

针对发展中遇到的挑战，《建设指南》收集整理了医疗信息化业内专家意见，从信息安全、数据利用、平台建设、平台管理、推广应用和厂商选择等方面讨论了大数据平台建设与应用的重点与难点。此外，本指南针对性选取医疗大数据平台建设成果较为显著的代表单位，如：解放军总医院、北京大学第三医院、上海市第十医院、福建省立医院和北京大学肿瘤医院，分享其建设案例和研究成果，供读者参考。

《建设指南》力图根据当前医院信息技术应用水平、医院管理和技术能力，以及国内医疗大数据建设现状，形成适合国内医院发展、可落地的建设方案。相关医疗信息工作者可根据本单位实际加以借鉴参考。本指南的撰写也得到了医渡云（北京）技术有限公司、北大医疗信息技术有限公司、北京嘉和美康信息技术有限公司、系联软件(北京)有限公司等医院信息产品供应商的积极支持。在此，对所有参加讨论、撰写的个人及单位表示衷心感谢！由于我们的水平有限，本指南如有不当之处，恳请各位读者提出宝贵意见。

王才有
中国医院协会信息专业委员会

2019 年 5 月

目 录

第一章 引言 1
1.1 政策背景 1
1.2 编制目的 2
1.3 现状与问题 3
1.4 医疗机构的医疗大数据平台概念 7
1.5 医疗机构建设医疗大数据平台的意义及作用 9
1.6 标准与规范体系 10
1.7 《建设指南》主要内容 12
第二章 总体设计 13
2.1 需求和目标 13
2.2 总体架构 14
2.3 技术路线 15
2.4 功能范围 17
2.4.1 大数据采集汇聚 17
2.4.2 大数据治理 17
2.4.3 大数据挖掘分析 18
2.4.4 大数据利用 18
2.5 隐私保护和数据安全 18
第三章 建设要点 21
3.1 安全体系 21
3.1.1 平台部署安全 21
3.1.2 安全措施保障 24
3.2 硬件部署 28
3.3 数据接入范围 29
3.4 数据接入方式 30
3.4.1 备份恢复 32
3.4.2 数据同步 33
3.4.3 物化视图 34
3.4.4 ETL 抽取 35
3.4.5 数据增量抽取 36
3.4.6 集成平台数据提取 38
3.5 数据脱敏加密 39
3.5.1 数据脱敏 39
3.5.2 数据加密 41
3.6 数据处理 45
3.6.1 数据验收 45
3.6.2 数据生产 46
3.7 数据扩展 56
3.8 数据授权 56
3.9 数据验证 57
3.10 平台验收 58
3.10.1 速度、性能 59
3.10.2 功能验收 59
3.11 平台培训 60
3.12 数据管理 61
第四章 应用场景 62
4.1 临床应用场景 62

4.1.1 临床大数据搜索 62
4.1.2 多学科诊疗(MDT) 65
4.1.3 患者全息视图 66
4.1.4 临床决策支持 68
4.2 科研应用场景 71
4.2.1 科研思路探索与发现 71
4.2.2 基于时间模型的科研分析 73
4.2.3 专科疾病数据库 74
4.3 管理应用场景 75
4.3.1 医院精细化管理 75
4.3.2 大数据病案管理 77
4.3.3 病历评分体系 78
4.3.4 VTE 风险评估 79
4.3.5 ICD 辅助编码 80
4.4 患者服务场景 80
4.4.1 智能导诊 81
4.4.2 智能候诊 82
4.5 药物研究场景 83
4.5.1 受试者智能招募 84
4.5.2 RBM 质控核查 84
4.5.3 AE/SAE 自动报警 85
4.5.4 试验数据辅助采集 85
4.6 教学应用场景 85
4.6.1 基于真实世界数据的疾病图谱 85
4.6.2 临床数据与知识库关联应用 85
4.7 应用展望 86
附录 A 专家论点 .....
1. 医疗大数据平台如何在医院立项？ .....
2. 医院上了临床数据中心，还需要上大数据平台吗？ .....
3. 如何判断大数据平台供应商？ .....
4. 是否能让临床科室直接从大数据平台导出数据？ .....
5. 数据申请和使用的管理，您认为主要由医院哪个职能科室管理？ .....
6. 为提高数据质量，而修改业务系统，如何操作？ .....
7. 大数据平台如何在院内推广使用？ .....
8. 如何进行大数据平台效益评价，大数据平台成果不显著，可能的原因是什么？ .....
9. 大数据平台管理的难点有哪些？为什么？ .....
10. 大数据技术依靠医院自我学习还是和厂商合作？如何保证持续性和自主性？ .....
附录 B 医疗大数据管理制度示例 .....
附录 C 医疗大数据平台建设案例 .....
案例一：医疗大数据中心建设案例-解放军总医院 .....
案例二：基于大数据技术的全量数据中心建设与应用-北京大学第三医院 .....
案例三：基于人工智能的临床科研一体化平台建设-上海市第十人民医院 .....
案例四：医疗大数据应用-福建省立医院 .....
案例五：医院大数据平台建设及应用-北京大学肿瘤医院 .....
附录 D 术语定义 .....
附录 E 参考文献 .....
附录 F 编委会介绍 .....

第一章 引言

医疗大数据是提升医疗服务质量、提高医疗服务效率、降低医疗费用的数据基础，而医疗大数据平台则是管理、分析和应用医疗大数据的强有力工具。在国家宏观政策和行业需求的推动下，医疗大数据平台建设获得了长足的进步，但在建设过程中仍面临着巨大的挑战。

本章节主要介绍医疗大数据发展相关的国家和行业政策法规，分析医疗大数据发展的现状和问题，同时明确了医疗大数据平台的概念、建设意义及对行业发展的作用。

1.1 政策背景

2015 年 9 月，国务院发布了《国务院关于印发促进大数据发展行动纲要的通知（国发[2015]50 号）》（以下简称：通知）。《通知》指出，要构建以人为本、惠及全民的民生服务新体系。围绕服务型政府建设，在健康医疗等领域全面推广大数据应用。健康医疗大数据作为国家重要的基础性战略资源，它的应用发展将带来健康医疗模式的深刻变化，有利于激发深化医药卫生体制改革的动力和活力，提升健康医疗服务效率和质量，扩大资源供给，不断满足人民群众多层次、多样化的健康需求，有利于培育新的业态和经济增长点。

为顺应新兴技术发展趋势，规范和推动健康医疗大数据融合共享、开放应用，国务院办公厅于 2016 年发布了《关于促进和规范健康医疗大数据应用发展的指导意见（国办法[2016]47 号）》（以下简称：《指导意见》）。《指导意见》提出，通过“互联网+健康医疗”探索服务新模式培育发展新业态，努力建设人民满意的医疗卫生事业，为打造健康中国提供有力支撑。健康医疗大数据以居民电子健康档案、电子病历、电子处方等为核心，融合了可穿戴设备、智能健康电子产品等产生个人健康数据资源，构建人口健康信息资源库。《指导意见》要求以保障全体人民健康为出发点，消除“信息孤岛”，建设健康医疗大数据平台，推进健康医疗大数据的共享和应用。医疗大数据作为健康医疗大数据的重要组成部分，其主要目的是基于临床医疗数据和医学专业知识分析患者疾病信息，为患者提供精准医疗服务。为了推进医疗大数据的应用，有必要建设医疗大数据平台，为临床医疗、医学科研、个人健康管理奠定坚实的技术基础。

2018 年 4 月，国家卫生健康委员会规划与信息司发布了《全国医院信息化建设标准与规范（试行）》（简称《标准与规范》）。它是在 2016 年《医院信息平台应用功能指引》和 2017 年《医院信息建设应用技术指引（试行）》基础上，形成的较为完整的医院信息系统体系框架。《标准与规范》主要针对目前医院信息化建设现状，对未来 5 年～10 年全国医院信息化应用发展提出建设要求。《标准与规范》不仅指导各级医院信息化建设，要求医院实现信息共享和业务交互，数据标准化、业务规范化，同时也对大数据技术在医疗业务中的应用提出明确要求和基本功能描述。《标准与规范》明确，医疗机构需要借助医疗大数据平台来管理、分析、利用医疗大数据，以实现提升医学科研及应用效能，推动智慧医疗发展的目标。

为了从标准、服务、安全、监督等方面更好地指导医疗机构和管理部门加强健康医疗大数据服务管理，2018 年 9 月，国家卫生健康委员会出台了《国家健康医疗大数据标准、安全和服务管理办法（试行）》（简称《试行办法》）。《试行办法》明确健康医疗大数据的定义、内涵和外延，以及制定办法的目的依据、适用范围、遵循原则和总体思路等，明确各级卫生健康行政部门的边界和权责，各级各类医疗卫生机构及相应应用单位的责权利，并对三个方面进行了规范。在标准管理方面，明确开展健康医疗大数据标准管理工作的原则，以及各级卫生行政部门的工作职责，提倡多方参与标准管理工作，完善健康医疗大数据标准管理平台，并对标准管理流程、激励约束机制、应用效果评估、开发与应用等作出规定；在安全管理方面，明确健康医疗大数据安全管理的范畴，建立健全相关安全管理制度、操作规程和技术规范，提出了数据分级分类分域的存储要求，对网络安全等级保护、关键信息基础设施安全、数据安全保障措施等重点环节提出明确的要求；在服务管理方面，明确相关方职责以及实施健康医疗大数据管理服务的原则和遵循，实行“统一分级授权、分类应用管理、权责一致”的管理制度，强化对健康医疗大数据的共享和交换。同时，在管理监督方面也强调了卫生健康行政部门日常监督管理职责，并提出大数据应用的安全监测、评估、追究制度。

1.2 编制目的

医疗大数据平台的作用，是有针对性地采集、存储海量医疗数据，并且进行标

准化处理，让医疗数据在聚合、分析后，能够驱动临床医学、精准医学等实践应用。迄今为止，我国尚没有医疗大数据平台的建设指南。因此，中国医院协会信息专业委员会（CHIMA）组织专家编写《医疗大数据平台建设指南》（简称《建设指南》），旨在为医院建设医疗大数据平台提供规范和指导，推进我国医疗大数据的应用发展。

编制《建设指南》的目的是，在推进医院信息化和健康医疗大数据发展的背景下，研究国内外医疗大数据平台发展趋势，并结合目前国内医疗大数据平台建设的主要实践，为各级医疗机构建设医疗大数据平台提供行动指南，为医疗信息化服务商的产品研发工作提供重要依据，为卫生健康相关行政管理部门推进医疗大数据相关工作提供有力参考。

《建设指南》适用范围和对象包括：

（1）各级医疗机构

协助医院制定医疗大数据平台建设规划，并规范指导医疗大数据平台的建设和运营，为其提供平台架构、技术路线、功能选型、数据安全等参考指南。

（2）各级医疗管理机构

为各级医疗管理机构提供医疗大数据应用总体情况，综合分析本地区医疗大数据发展及应用信息，同时监督医疗大数据安全保障情况。

（3）医疗信息化建设服务商

面向医疗信息化企业，指导其开展医疗大数据建设方案规划、研发、实施以及运营，为医疗机构提供适用的大数据平台和服务模式。

1.3 现状与问题

我国已经出台一系列有关医疗大数据的政策和法规，以加快推进医疗大数据的发展。江苏、福建、山东、安徽、贵州等五个省率先开展了国家健康医疗大数据中心试点。

医疗大数据正在成为医疗机构的重要资产，并且是医疗行业相关企业不可忽视的战略资源。医疗大数据相关工作在我国已开展多年，但尚处于行业发展初期。各大医院的信息资源基本还是躺在数据库中“沉睡”。由于数据收集、存储、整合、管理不规范，导致数据利用率不高，加之跨部门、跨机构之间数据共享机制缺失，直接影响到大数据的有效利用。最终导致有数据的单位不愿共享，需要数据的单位

得不到数据，形成牢不可破的“信息孤岛”。对于医疗机构来说，数据“沉睡”的成本颇高。很多临床诊疗和科研项目，在数据可以共享的情况下，完全能够大幅提高效率。所以，建立衔接各方数据系统的医疗大数据平台，刻不容缓。

1.3.1 国内发展现状

政策对医疗健康大数据的推动、医疗行业对大数据应用的需求、电子病历等医疗数据的爆炸性增长、对公众健康管理数据的聚合、医疗数据分析技术和工具的进步，这些主要因素促进了国内医疗大数据的发展。下文从几个层面来介绍国内医疗大数据建设现状：

（1）政策有力促进

为推进和规范医疗健康大数据的应用发展，2016 年 10 月，国家卫计委在京召开医疗健康大数据中心与产业园建设国家试点工程启动推进电视电话会，会议围绕贯彻落实全国卫生与健康大会精神和《指导意见》，明确试点思路，确定福建省、江苏省及福州、厦门、南京、常州作为为第一批试点省市，启动第一批医疗健康大数据中心与产业园建设国家试点工程。

与此同时，根据《指导意见》要求，各地也相继出台了地方医疗健康大数据指导意见，例如，2017 年 3 月，安徽出台《关于促进和规范健康医疗大数据应用发展的实施意见》。这些政策的相继出台，从宏观层面指导医疗健康大数据应用规范、有序地向前推进。

（2）卫生行业监管推动

各地卫生监管机构在国家鼓励医疗健康大数据发展的大背景下，相继构建地方公共卫生、疾病预防、健康体检、卫生监督等数据中心，以便掌握地方整体的医疗卫生资源、疾病预防控制、妇幼卫生系统、健康体检情况及卫生监督系统情况。以北京市为例，完成 30 家三级医院电子病历信息的互联互通，共享内容包括门急诊的信息，住院病案首页、医疗机构的信息，以及患者的出院小结、用药情况，以及检查放射等所有与患者相关的信息实现跨机构调阅查询。同时在慢病管理监测过程中，开展了针对脑血管、心血管、糖尿病、高血压等疾病的监测。

（3）医疗机构内在需求

医疗机构通过大数据技术整合患者就诊数据，以患者为中心构建集成电子病历，以便医务人员能够便捷地调阅患者就诊信息。利用大数据的搜索、处理和分

析能力，对医疗机构整体运营情况进行分析和监控。机器学习结合临床决策支持系统，将临床多个维度的数据进行整合，为医生和患者提供精细化、个体化的诊疗指导。对于医疗机构来说，大数据的价值在于能够提升医疗机构管理水平、服务效率以及临床诊疗的效果。

（4）患者健康层面

通过使用医疗可穿戴设备或社区健康体检设备等便携式医疗设备，患者将医疗健康数据共享给医疗机构。医疗机构可以监控患者健康状况，并且对患者医疗健康数据进行分析，为患者提供更优质的医疗服务。

以上几个层面，也是目前医疗大数据平台的主要服务领域。针对不同领域的应用需求，医疗大数据平台可以提供不同的服务主题，以便医疗大数据能够为各领域提供合理、优质的大数据应用服务，逐步实现国家医疗健康大数据发展战略。

1.3.2 国外发展现状

2012 年 3 月 29 日，白宫宣布启动大数据研究和开发。2013 年 1 月 15 日到 17 日，美国国家标准与技术研究院（NIST）联合各行业专业人士，召开了“云和大数据论坛”会上 NIST 决定创建一个公共工作组，开发大数据互操作性框架。该框架应当定义并区分大数据技术需要满足的需求，包括互操作行、可移植性、可重用性、可扩展性、数据使用、分析及技术架构。

2013 年 6 月 19 日，NIST 大数据公共工作组成立，旨在对大数据的定义、分类、安全参考架构、安全隐私需求和技术路线图形成共识，最终形成一个中立于供应商并在技术和基础设施方面独立的框架，即《NIST 大数据互操作性框架草案》。草案明确了大数据的安全、隐私、架构、标准等内容。

2014 年 12 月，美国 ONC（美国国家卫生信息技术协调办公室）发布了《美国联邦政府医疗信息化战略规划（2015-2020）》，其总体目标是提高健康信息的安全可及性和使用率，让公众在医疗服务提供者的帮助下有能力进行健康管理，提高生命和健康质量。随后，美国大数据厂商积极开展医疗健康行业大数据建设布局。IBM 公司组织医生和研究人员汇集数千份病人的病历，近 500 份医学期刊和教科书，1500 万页的医学文献，训练出 IBM Watson 系统。IBM Watson 通过认知计算为人们创造一种全新方式，挖掘出隐藏于大量数据中的知识和模式。IBM Watson 分析医

疗记录（结构化的数据和非结构化的数据），通过分析各种医疗数据，为患者提供建议治疗方案，并给建议治疗方案排序，注明其医疗证据，医生可以根据患者病情选择合适的治疗方案。

Dignity Health 是美国最大的医疗健康系统之一，致力于开发基于云的大数据平台，带有临床数据库、社交和行为分析等功能。该平台将连接系统中 39 家医院和超过 9000 家相关机构并共享数据，通过其大数据应用方向可以看到一些机会：诸如，个人和群体医疗规划，包括预防性疾病管理；定义和应用最佳病例、减少再入院率；预测败血症或肾衰竭风险，提早进行干预减少负面结果；更好地管理医药成本；创建工具来改进患者的就医体验。

英国积极发展个性化医疗，首个综合应用大数据技术的医药科研卫生机构“李嘉诚卫生信息与发现中心”于 2013 年在英国牛津大学正式揭牌。它包括“靶标研究所”和“大数据研究所”两个机构，旨在利用大数据技术收集、存储和分析大量医疗信息，确定新药物的研究方向，减少药物开发成本，同时为发现新的治疗手段提供支持。

1.3.3 挑战与问题

虽然国内医疗机构为临床、科研、患者、药物研究等方面提供数据服务，也能够借助医疗大数据平台形成一定范围的医疗大数据生态圈，但是目前我国医疗大数据缺乏统一的标准规范。医疗机构建设医疗大数据平台的初衷也是为了满足自身医疗信息化发展的需求，未能充分考虑医疗大数据的数据共享和数据应用，无法最大限度地应用医疗大数据。为了构建完善的医疗大数据体系，构建成熟的医疗大数据平台，需要解决以下几个主要问题：

（1）数据标准要统一

健康医疗大数据是国家重要基础性战略资源。国家卫生健康委员会 2018 年发布《国家健康医疗大数据标准、安全和服务管理办法(试行)》明确指出要加快推进健康医疗大数据的标准制定工作，鼓励医疗卫生机构、科研教育单位、相关企业或行业协会、社会团体等参与健康医疗大数据标准制定工作。目前，医疗数据并不缺乏标准，但汇集之后的健康医疗大数据尚缺乏统一的国家或者行业标准，各个医疗大数据平台通常只是部分借鉴了成熟标准，大多平台建设还是采用各自的数据标准规范。不仅医疗大数据厂商的标准不统一，各个地区甚至各个医院都未使用统一数

据标准，影响了健康医疗大数据平台的数据质量和数据治理效果。目前，医疗信息化行业内尚未有指导医疗机构大数据平台建设的具体、规范、适用性强的操作文件，由于建设规范的缺乏，一定程度上制约了数据标准、技术规范和共享规范的统一。

（2）数据技术要适宜

大数据技术在各行各业都有所涉及，但是由于医疗业务不同于其他行业，构建医疗大数据平台需要根据医疗大数据特性来遴选相关技术，不能一味地照搬其他行业大数据技术。特别是在数据采集、数据挖掘、数据治理等关键环节，要能够满足医院实际情况，从作业、模型、物理资源等各方面综合评估，选择合适的大数据技术和架构。

（3）数据安全要保障

医疗信息已经从纸质文本时代迈向数据电子化时代，医疗数据和应用呈现指数级增长趋势，这也给动态数据安全监控和隐私保护带来极大的挑战。医疗机构有必要建立医疗大数据安全管理体系，保障数据存储，网络设备、基础设施等安全工作，并提供数据安全保障相应措施，做到数据流转全程留痕、数据安全监测和预警、数据泄露事故可查询可追溯等数据安全保障工作。

（4）数据治理要全面

医疗大数据平台给医疗机构带来的价值取决于医疗数据质量，医疗机构往往需要通过全面完善的大数据治理来保障数据质量。在传统数据平台阶段，数据治理的目标主要是做数据管控，为数据部门建立一个的治理工作环境，包括标准、质量等。在大数据平台阶段，用户对数据的需求持续增长，用户范围从数据部门扩展到整个医疗机构，数据治理不仅要面向数据部门，还要面向所有医疗业务场景，使得数据治理能够覆盖元数据、隐私、数据质量、业务流程整合、主数据整合和数据生命周期管理等环境。

1.4 医疗机构的医疗大数据平台概念

医疗大数据是健康医疗大数据的重要组成部分，原卫计委将健康医疗大数据解读为：健康医疗大数据涵盖人的全生命周期，既包括个人健康，又涉及医药服务、疾病防控、健康保障和食品安全、养生保健等多方面数据的汇聚和聚合。但是，医疗卫生相关部门还没有对医疗大数据做出明确的解释和定义，目前关于什么是医疗

大数据，业内理解有所不同，不过普遍认为医疗大数据应具备“医疗”和“大数据”双重特性。

本《建设指南》认为，医疗大数据主要是指医生对患者诊疗和治疗过程中产生的数据，包括患者的基本数据、电子病历、诊疗数据、医学影像报告数据、医学管理、经济数据、医疗设备和仪器数据等。即以患者为中心，构成医疗大数据的主要来源。医疗大数据不仅具有大数据的“4V”特点外，还包括时序性、隐私性、不完整性等医疗领域固有的主要特征：

a) 时序性：患者就诊、疾病发病过程在时间上有一个进度；医学检测的波形、图像均为时间函数。

b) 隐私性：患者的医疗数据具有高度的隐私性，泄露信息将造成严重后果。

c) 不完整性：大量来源于人工记录，导致数据记录的残缺和偏差；医疗数据的不完整搜集和处理使医疗数据库无法全面反映疾病信息。

基于医疗大数据平台，医疗机构可以有效地聚合、分析、管理、利用医疗大数据，实现医疗大数据的有效管理和应用。关于医疗大数据平台，原卫计委也给出医疗大数据平台的范围和定义。

《医院信息化建设应用技术指引（试行）》中医疗大数据平台的标准解释为：①数据交换汇集。多种数据采集接入技术(包括 ETL、爬虫等)，实现医疗机构内部数据、医疗相关科室数据、健康数据和互联网数据等多源异构数据的解析、汇集和共享。②数据存储。基于列数据库、文件数据库、分布式数据库、集群等多种文件存储技术，支持结构化、半结构化、非结构化文件的分布式存储。③分布式计算。基于分布式计算框架，利用集群资源，实现计算任务的分布式并行执行，提高多源异构海量数据的计算效率。④数据可视化。基于统一时空框架，利用可交互的可视化界面方式，实现医疗卫生大数据综合展现。

医疗大数据平台管理的范围为：①平台运维。支持可视化开发界面、计算任务调度、智能部署、资源监控等能力。②大数据平台加固。支持大数据平台组件的统一配置，对其安全管理措施进行统一配置，规范平台安全配置管理。③大数据平台日志管理。通过 syslog 等方式，记录各组件的操作和运行日志，记录身份验证信息，统一存储，并支持检索和分析。④安全告警。支持自动化的异常行为分析、告警及分析规则自定义。⑤访问控制。利用身份认证技术对组件操作权限进行管控。

⑥数据权限管理。通过字段级别的访问控制措施，进行结构化数据访问权限、非结构化数据访问权限的管理。

医疗大数据平台提供的大数据服务为：①数据挖掘和建模。提供基于大数据架构下海量数据读取、数据处理、数据计算服务，通过可视化的数据探索工具、数据挖掘模型、简易模型训练支持数据挖掘与分析服务。②数据应用服务。支持快速数据集成、在线数据检索、多人协同等工具，提供大数据的检索、归并等应用服务。③数据治理。通过规范流程和规则库，基于流程引擎构建统一的、可配置的数据转换、清洗、比对、关联、融合等加工处理过程，对异构异源海量离散的数据资源加工生产，生成易于分析利用的、可共享的数据。

1.5 医疗机构建设医疗大数据平台的意义及作用

1.5.1 医疗大数据平台的意义

建设医疗大数据平台，运用大数据的分析和挖掘技术，可以在一定程度上帮助医疗机构提高生产力，改进护理水平，增强竞争力。基于医疗大数据平台，实现历史医疗资源的再利用，并借助大数据的思维和方法进行研究，完成过去传统思维、方法、技术无法完成的任务，解决过去无法解决的问题，使得数据加以利用，形成从量变到质变的过程，同时通过多维度的分析研究，实现对医疗数据的高效检索、后结构化、分析计算。

同时，建设一个的高效、稳定运行医疗大数据平台，实现现有各种医疗数据库的数据共享与交换，可让大数据处理更加便捷、快速、贴近用户，有效实现数据的流通及使用价值的增值，为患者、医务人员、科研人员及管理人员提供服务和协助，成为未来信息化工作的重要方向。

1.5.2 医疗大数据平台的作用

（1）面向医务人员

建设医疗大数据平台，可以为医务人员提供基于大数据技术的医疗服务，可深入洞察病症诊疗手段与成果，为相关病症研究提供数据支撑。利用医疗大数据平台，可以为医务人员提供辅助诊疗服务，借助过往各类病例和各类数据源，深入分析相关病症并需找、推荐最优治疗方案，为个性化诊疗提供基础。通过大数据技术完成病历文档后结构化处理，在保证医生书写的原始病历数据可溯源的基础上，实现对

既往病历的结构化处理，满足科研数据采集需求。

同时，在临床科研领域，面对海量数据，科研人员只需从中直接查找或挖掘所需信息、知识和智慧，甚至无需直接接触需研究的对象。在科研过程中，大数据的利用、开发和整理，可以颠覆以往很多研究结果，带来意想不到的发现。

（2）面向患者

建设医疗大数据平台，可以使患者主动参与医疗过程，结合患者的健康数据、既往病史，更有利于医生做出正确的疾病诊断。基于医疗大数据的医疗服务，可以创新医疗模式，减少医患矛盾。因为有效的数据整合模式，大数据医疗满足了以患者为中心的个性化医疗，提升现有医疗技术平台的服务能力。医疗大数据的运用，从医疗研究、临床决策、疾病管理、患者参与以及医疗卫生决策等方面，推动了医疗模式的转变，尊重患者的价值观、个性化特征和需求，协调并整合不同专业的医疗服务，保持医疗服务的连续性和可及性，提高医疗质量。

（3）面向管理人员

建设医疗大数据平台，提供统一可视化分析展示平台，为医院管理运营相关决策提供数据依据，实现医院精细化管理。医院精细化管理是以规范化为前提，系统化为保证，数据化为标准，信息化为手段，把服务者的焦点专注到满足被服务者需求上，以获得更高效率、更多效益和更强竞争力。通过大数据分析平台对医院门诊量、手术量、入/出院病人数、床位使用率、床位周转率、设备使用率、 疾病图谱、患者分布区域、费用支出等数据分析。将当前数据与同期数据、前期数据进行对比分析。同时，可以对比本地区医院运营情况，找出不断提高医院经济运行质量的成因和差距，抓住自身工作的薄弱环节，切实采取改进措施。

1.6 标准与规范体系

医疗信息标准和规范体系医疗大数据采集、治理、共享和应用基石，脱离标准规范体系的医疗大数据无法实现有效地融合、分析、应用和共享，也就无法将医疗大数据价值最大化。但是，医疗大数据的融合、分析、应用和共享并非几个数据标准就能完成，需要借助术语标准、内容标准、交换标准、技术标准、应用标准、安全标准等各种标准规范构建的医疗大数据标准体系来体现医疗大数据价值。

大数据标准体系由五类标准构成：①语义标准，包括术语、标识等标准；②语

法标准，用于规范数据、信息的描述格式；③传输标准，用于支持跨系统数据信息传输；④安全标准，用于规范数据信息的安全访问和传输；⑤服务标准，用于规范数据共享、处理、分析服务。

医疗大数据具有多源、异构、非统一等数据特性，这使得医疗大数据标准体系有别于以往国家和行业发布的数据标准，需要根据不同的医疗大数据服务需求，基于用例驱动方法，利用更为宽泛、灵活的数据标准来构建大数据标准体系。因此，医疗大数据标准体系要依据数据相关方利益、数据流通环节、平台服务协议等因素来构建符合医疗大数据价值利用的标准体系，而非一个简单、统一、永恒不变的标准体系，它需要在实践中不断升级和完善。

医疗大数据标准规范体系的升级和完善包含以下工作：

（1）医疗大数据相关利益方梳理

梳理医疗大数据平台中服务的提供者和消费者，以及数据流通环节其他的利益方、责任方。提供者包括数据提供者、大数据应用提供者、大数据框架提供者，相关利益方从数据、应用、技术不同层面对数据利用提供支持。通过对这些数据相关利益方的梳理，就可以得到数据流通各个环节的业务需求和相关技术需求，这些需求共同组成大数据平台需求，作为医疗大数据用例，用以评估和验证标准规范体系是否合理、适用。

（2）医疗大数据环节服务协议

为保障大数据环节服务有效实施，需要从数据流维度和 IT 服务维度来构建标准规范体系。从数据流维度来看，数据价值是通过数据采集、集成、分析、使用结果来实现的。从 IT 维度上分析，数据价值是通过平台 IT 服务来实现的。IT 服务既包括基础设施、存储设施、网络设施等硬件服务，也包括 IT 架构（云计算、分布式等）、IT 工具（Hadoop、Spark 等）等软件服务。通过 IT 服务来保证医疗大数据平台数据的一致性、安全性、准确性，实现医疗大数据平台数据价值的有效利用。

（3）医疗大数据标准体系升级完善

现有医疗信息标准规范体系已经在医疗机构得到检验，建医疗大数据标准体系升级完善首先就需要与现有标准体系作差异分析，通过大数据用例来驱动新标准体系的构建。新标准规范体系构建的宗旨是满足医疗大数据的独特需求。

1.7《建设指南》主要内容

本《建设指南》从深化医院信息化改革和推广医疗大数据应用入手,结合大数据时代下的国内外医疗现状,给出了医疗大数据平台的建设内容和建设要点，包括总体架构、技术路线和功能范围；其次，针对不同医疗业务场景,介绍了基于医疗大数据平台的临床、科研和公共卫生等各种大数据应用，并解析了医疗大数据为医疗机构提供何种数据服务的问题，同时，给出了一些详细介绍医疗大数据平台实际建设和应用情况的案例；最后，行业专家围绕建设医疗大数据平台给出了一些具有建设性的意见和建议，并展望了医疗大数据平台的后续发展。

第二章 总体设计

《建设指南》第一章明确了医疗大数据平台概念、意义和作用。本章将进一步明确医疗大数据平台建设所需要满足的行业需求、平台的发展目标。同时，为了从技术层面明确界定医疗大数据平台，本章还详细介绍了医疗大数据平台总体架构、技术路线、功能范围以及安全与性能要求。

2.1 需求和目标

医疗大数据平台是医院信息化建设重要组成部分。在促进和规范健康医疗大数据应用发展的形势下，为了有效地聚合、分析、管理、利用医疗大数据资源，有必要构建医疗大数据平台，为医院的管理、诊疗、科研和教学提供高效的服务。

（1）建设医疗大数据平台的主要需求

- 稳定的大数据平台：基于大数据相关技术和框架，提供稳定、高效的数据采集、数据融合、数据计算、数据挖掘、数据分析、数据治理的医疗大数据平台。
- 多样的大数据采集：支持增量抽取、全表采集等各种数据采集方式；支持日志管理和异常监控。
- 有效的大数据治理：支持结构化和非结构化数据、集中式和分布式数据的统一建模；支持大数据清洗、脱敏的数据治理；以统一的数据标准对多源异构数据进行归一化处理。
- 丰富的大数据应用：利用数据中心的大数据资源，对医疗服务、科研管理、医院治理等的辅助决策支持应用。
- 灵活的大数据展示：提供大数据数据模型可视化配置，提供大数据分析结果的可视化展示。
- 安全的大数据服务：支持大数据存储、传输、访问等服务的安全保障，对数据进行安全评估和数据流转监控，防止隐私数据泄露。

（2）建设医疗大数据平台的主要目标

- 利用大数据平台，实现医疗数据共享开放，并且能够提高医疗数据利用率，充分挖掘医疗数据潜藏的价值，使其最大限度地服务医疗业务。
- 以患者为中心，借助大数据技术整合患者医疗数据，为医生提供患者全生

命周期的数据服务。

- 为患者提供个性化医疗数据服务，保持医疗服务的连续性和可及性，提高医疗质量。
- 为医院临床辅助决策和管理辅助决策、临床科研提供基础数据。

2.2 总体架构

根据《医院信息化建设应用技术指引（试行）》中有关医疗大数据平台建设的标准，医疗大数据平台总体架构见图 2.1 所示。

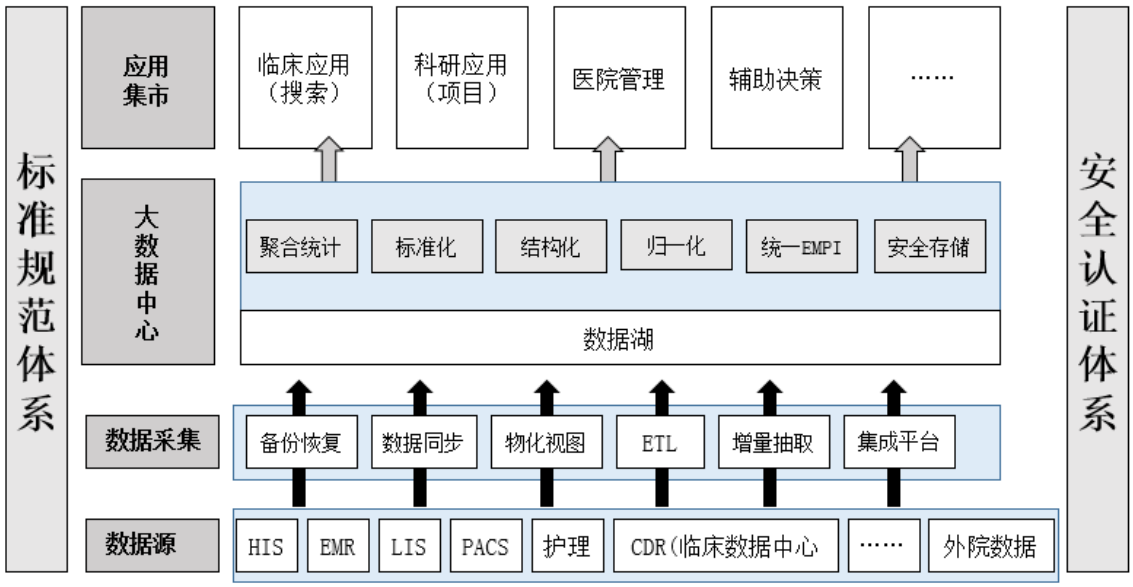


图 2.1 医疗大数据平台总体架构

医疗大数据平台总体架构包括四层架构，即数据源层、数据采集层、大数据中心、应用集市。

数据源层：本层主要指医疗大数据平台涉及的数据范围。根据医疗大数据平台的功能和应用，本层数据范围一般包括医院涉及临床医疗工作的与患者有关业务系统数据。常见业务系统包括：医院信息系统（HIS）、电子病历系统（EMR）、检验系统（LIS）、放射系统（PACS）、病理系统、超声系统、心电系统、手术麻醉系统、其它检验检查类系统等。医疗大数据平台数据源层数据一般为结果性数据，针对过程性数据建设单位可以根据医院实际情况进行处理。另外数据源层一般不包括医院财务数据、患者基因检测数据等。针对医联体医院，建议有条件的核心医院可以将医联体单位数据纳入医疗大数据平台管理，实现医联体内数据综合应用和分享。

数据采集层：本层实现将业务系统源数据抽取到医疗大数据平台功能。结合目

前医疗行业普遍使用的数据抽取方式总结如下：数据库备份恢复、集成平台、物化视图、数据同步工具（比如 OGG）、ETL（抽取工具）等，建设单位可根据本单位信息系统建设情况、信息人员技术掌握情况等选择数据采集方案。

大数据中心：本层为医疗大数据平台核心技术层，主要用来处理医疗数据，实现非结构化数据的结构化。通过数据采集层采集数据汇集到大数据中心形成业务数据湖，大数据中心的分析加工功能基于数据湖中数据进行处理，基本处理功能包括：①数据整合，将从各个业务系统获取的医疗数以患者主索引 EMPI 为中心进行整合，实现数据综合利用。②数据自动化处理，处理过程可以实现自动处理，避免人工实时干预。③数据安全性，数据处理过程要保证数据安全性，做到数据保密性、数据完整性、数据可利用等。

应用集市：本层以医疗大数据平台数据中心数据为基础，建设各种医疗大数据平台基础应用。包括但不限于：临床数据搜索、患者全景诊疗视图、患者数据服务、临床科研应用、临床知识库、科室运营、临床辅助决策等。

标准规范体系：本平台搭建主要参照卫计委数据标准、HL7CDA 文档、相关术语标准和国家相关数据标准，具体标准内容请见下表：

表 标准规范系统参考标准清单

卫计委数据标准	中国卫生信息数据元值域代码 WS364.X-2011（X 从 1 到 17,共 17 部分）
	电子病历基本数据集 WS445.X-2014（X 从 1 到 17,共 17 部分）
HL7CDA 文档	HL7ChinaCDA 规范试行 2013 版(共 5 个）
相关术语标准	国际疾病分类第九版临床修订第三卷：手术与操作 ICD-9-CM-3
	国际疾病分类第 10 版 ICD-10
	国际疾病分类肿瘤学专辑第三版 ICD-O-3
	观测指标标识符逻辑命名与编码系统 LOINC V2.42
国家相关数据标准	GB/T 2261.1-2003 个人基本信息分类与代码 第 1 部分 人的性别分类
	GB/T 4671-2008 家庭关系代码

2.3 技术路线

大数据的特征是海量的数据规模、动态的数据体系、多样的数据类型和巨大的数据价值。面对数据量的指数级增长，传统的存储和运算模式已经不足以应对当前的数据量和数据复杂程度，尤其传统的分析模式无法深入挖掘数据的潜在价值。以 Hadoop 为代表的分布式存储与计算框架是当前主流的大数据技术架构，是一种具体的实现技术。目前在建的医疗大数据平台基本采用 Hadoop 技术来实现海量数据

存储与处理，本《建设指南》不对 Hadoop 技术做深层次说明，只简单介绍下相关技术特点。

Hadoop 具备高拓展性、高可靠性和低成本的优点，为海量数据的存储和计算提供了技术支持。

（1） 高扩展性：分布式系统支持不停机线性扩容，同步提高存储能力和计算能力，保证未来可扩展性。

（2） 高可靠性：分布式存储和冗余提高数据可靠性，保证数据长期可靠保存。

（3） 低成本：基于廉价的 x86 开放架构，软硬件成本可控。

Hadoop 的框架最核心的设计就是：HDFS（Hadoop Distributed File System）和 MapReduce。HDFS 为海量的数据提供了存储能力，MapReduce 为海量的数据提供了计算能力。

（1）HDFS 主要特点

- 处理超大文件：可以存储几百万个大型文件，文件系统的容量可达数 PB 级别。
- 支持动态扩展：利用横向扩展模式（scale-out），节点动态加入集群，可以数百数千个。
- 流式数据读写：HDFS 的设计思想“一次写入，多次读写”，一个数据集一旦由数据源生成，就会被复制分发到不同的存储节点中。
- 数据可靠性：多副本机制。
- 高容错性：可容忍机器某些部件故障和磁盘失效。
- 可运行廉价的商用机器：HDFS 设计时充分考虑了可靠性、安全性及高可用性，因此 Hadoop 对硬件要求比较低，可以运行于廉价的商用机器集群，无需昂贵的高可用性机器。

（2）MapReduce 特点

- 面向大数据并行处理的计算模型、框架和平台。
- 大任务切割为小任务。
- 多机多线程并发控制处理。
- 处理阶段：

Map 阶段：接收计算任务，将接收到的任务打散为小任务，分配到不同的机器上。

Reduce 阶段：集群机器接收打散后的小任务进行运算，并将结果整理汇总，反馈给客户端。

基于 HDFS 分布式存储特点，所以在建设大数据平台时是要求多硬件、硬件配置要求不高的特点。

2.4 功能范围

医疗大数据平台建设遵循《全国医院信息化建设标准与规范（试行）》等相关标准与规范。主要包括以下几方面功能：

2.4.1 大数据采集汇聚

基于平台数据集成，以服务器作为基础硬件平台，采用集群技术、分布式存储技术、分布式计算技术、ETL 技术，制定数据采集标准及处理流程，对结构化数据抽取入库，对非结构化数据进行结构化改造，主要包括病人的基本信息、病历信息、病程信息、医嘱信息、检验信息、影像信息、护理信息等内容。实现数据存储与共享，针对不同的需求提供更精细化、精准化的支持。

2.4.2 大数据治理

对采集汇聚的数据进行清洗加工处理，并做标准化整理。主要包括制定数据清洗流程、清洗流程控制、清洗质量控制、清洗过程管理等。通过规范流程和规则库，基于流程引擎构建统一的、可配置的数据转换、清洗、比对、关联、融合等加工处理过程，对异构异源海量离散的数据资源加工生产，生成易于分析利用的、可共享的数据。

通过部署大数据计算框架，基于多种算法库，实现大数据存储访问及分布式计算任务调度、多维索引数据的深度搜索和全文检索等功能。建立基于分布式并行计算架构，部署服务器集群，具备横向扩展能力，可以动态增加或减少计算资源和存储资源，支持 PB 量级离线计算和在线计算。部署非关系型数据库 HBase、数据仓

库Hive、数据处理工具Sqoop、机器学习算法库Mahout、一致性服务软件ZooKeeper、管理工具Ambari等，或者其他大数据计算框架如MapReduce、Spark、Tez等，部署搜索引擎Elasticsearch用于全文检索、结构化检索和分析。

2.4.3 大数据挖掘分析

从大量的医疗数据中通过算法搜索隐藏于其中信息的过程。一般通过包括审计、在线分析处理、机器学习、专家系统（依靠过去的经验法则）和模式识别等方法来实现。提取出的信息合成知识库、诊疗规则，服务于大数据应用。

2.4.4 大数据利用

通过医疗大数据平台开展各类应用，提高医务人员诊疗水平、辅助医院管理人员决策、加速科研成果落地、为患者提供精准化的医疗服务。包括临床辅助决策、单病种大宗病例统计分析、治疗方法与疗效比较、精准诊疗与个性化治疗、不良反应与差错分析提醒、健康预测与预警、精细化管理决策支持、科研结果验证、辅助用药分析与药物研发等。

2.5 隐私保护和数据安全

医疗大数据平台建设应在满足安全需求的前提下，结合实际情况，基于现代信息安全理论，遵循国家和行业标准，采用目前国内外先进的信息安全技术，采取有效的安全策略和技术手段，建立覆盖硬件网络、操作系统、数据库、应用软件和管理等各个方面的统一、安全、稳定、高效的信息安全保障体系，确保平台承载业务信息的安全可靠及业务服务的连续且稳定运行，并可随着未来业务及管理所需的不断发展而动态性调整，最终实现“政策合规、资源可控、数据可信、持续发展”的生存管理与安全运维目的。

（1）基本数据平台安全标准

目前，国家卫健委已经给出了构建数据平台的安全保障体系的方法和规范。

数据平台应当按照计算环境、区域边界、通信网络三个环节进行分等级的安全防护建设，同时，在此基础上还需要建设集中的安全管理中心，对部署在计算环境、区域边界、通信网络上的安全策略与安全机制实现集中管理。

①安全计算环境主要针对主机安全性保障提出，对于数据平台，应实现二级增强的计算环境所要求的身份鉴别、访问控制、安全审计以及数据保密性和完整性等内容，此外，应根据实际情况，建立数据的备份及存储恢复措施，如条件具备，可构建集中的数据和系统灾备中心，保证在发生安全事件时能够尽快恢复数据、系统，快速恢复业务等；②安全区域边界针对隔离与访问控制而提出，对于数据平台，应实现二级增强的区域边界所要求的防火墙隔离、安全审计、入侵防护以及恶意代码监测与过滤等内容；③安全通信网络则针对网络及通讯的安全保障而提出，对于数据平台，应当实现二级增强的安全通信网络所要求的通信机密性、完整性保护、网络设备安全性保护、网络设备冗余等内容

数据平台应该建有安全管理中心，其是对上述三个层面所采取的安全措施的集中管理，包括系统管理、安全管理等相关内容。其次，物理安全方面，要根据实际情况建立相应的安全防护机制；需要加强计算机房的安全建设，机房必须具备防水、防潮、抗震、防雷击、防盗窃、防静电、防电磁辐射的措施。安全管理方面，要考虑政策、法规、制度、安全培训等，制定切实有效的管理制度和运行维护机制。

（2）大数据平台安全要求

- 可扩展性：支持动态增加和删除系统节点，集群搭建方式灵活可控；
- 高效性：以分布式文件系统进行存储数据，支持海量数据的快速读/写、查询操作；采用分布式计算进行数据分析与业务操作，各业务节点独立计算互不干，节点数量越多运算速度越快；
- 可靠性：系统自动容灾(HA)；采用主-从机制(Master-Slave)进行集群搭建，系统内节点之间数据互相实时备份，当节点宕机时直接切换至备份节点，运算单元宕机时直接切换至备份运算节点；
- 低成本：对系统中各节点设备硬件要求不高，而且开发技术可跨平台。

（3）大数据隐私保护要求

医疗大数据平台除了要实现《电子病历基本规范（试行）》要求：“对操作人员的权限试行分级管理，保护患者隐私”，还要提供数据脱敏和个人信息去标识化功能，提供满足国内密码算法的用户数据加密服务。

在数据加解密方面，能通过高效的加解密方案，实现高性能、低延迟的端到端和存储层加解密（非敏感数据可不加密，不影响性能）。同时，加密的有效使用需

要安全灵活的密钥管理，这方面开源方案还比较薄弱，需要借助商业化的密钥管理产品。此外，加解密对上层业务透明，上层业务只需指定敏感数据，加解密过程业务完全不感知。

除此以外，可以引入主动隐私保护技术，构建隐私模型，对具有隐私泄露风险的数据，进行风险监测、风险评估、主动保护、责任追溯。其中，风险监测与评估达到事前预警的目的，主动保护提供事中整体防御措施，责任追溯提供事后溯源和追踪，做到主动防御隐私泄露。

第三章 建设要点

本章节主要介绍医疗大数据平台建设过程，从平台搭建、数据接入、数据处理、平台应用产出等各个环节介绍医疗大数据平台建设内容。

3.1 安全体系

在医疗大数据平台建设工作中，平台的安全性至关重要。目前，各医疗大数据平台公司在数据平台建设方面，对于安全保障有不同的解决方案。最普遍方式是侧重于使用杀毒软件、防火墙、IPS 等系统，建立网络安全边界，有效隔离外界访问。

借着大数据如火如荼的发展大势，网络安全问题越来越成为人们关注的焦点。据统计，在 2016 年 6 月 21 日国务院下发的《关于促进和规范健康医疗大数据应用发展的指导意见》中，“安全”一词出现了 33 次，可见国家对于健康医疗大数据安全的重视程度。

黑客入侵、系统漏洞、勒索病毒等问题助推了数据安全、隐私保护的热度，这让医疗行业不得不清醒地意识到，发展健康医疗大数据的前提是保护数据安全和患者隐私。只有在安全的前提下，才能真正实现数据融合共享、开放应用。

然而，由于网络安全人才与经费保障缺口大、缺乏具备行业特色的网络安全指导框架、网络安全标准等，我国健康医疗数据安全形势依然严峻。习近平总书记在 2016 年 4 月 19 日的全国网络安全和信息化工作会谈会上曾表示，网络安全和信息化发展相辅相成，安全是发展的前提，发展是安全的保障，安全和发展要同步推进。因此如何保障医疗数据的安全，将成为医疗大数据平台建设的首要内容。

3.1.1 平台部署安全

医疗大数据平台建设，在安全方面主要涉及的医院业务系统数据，以及大数据平台的算法优化更新。医疗大数据平台结构化数据精准性在于平台算法的持续性更新、优化，目前各大医疗大数据平台技术服务商在建设大数据平台时都需要不定期对平台进行算法更新，同时由于算法团队人员一般不在项目现场工作，而在服务商驻场，就牵扯到如何保证技术服务商时刻可以连接大数据平台进行算法优化和平台更新。所以既要保证平台数据安全，避免数据丢失。还要考虑公司对算法不定期的优化与更新。目前医疗大数据平台部署、数据存储模式有多种，但具体选择何种模

式需医院根据自身情况来进行选择。

在当前医疗环境下，大型医疗机构在院内一般采用业务内网和办公网两网运行模式。业务内网主要运行医院信息系统，如 HIS、EMR、LIS、PACS 等，属于保证医疗行为正常开展的内部网络，简称院内网。办公内网主要运行系统办公、邮件、图书馆等非医疗核心支撑系统，简称办公网。上述两网一般采用逻辑隔离或物理隔离模式，院内网与互联网完全隔离，办公网可以访问互联网。结合医院网络环境和医疗新技术、医疗大数据平台数据以及所需计算资源和应用资源，提供以下医疗大数据平台数据存储部署模式：1、院内网模式（逻辑隔离或物理隔离的业务内网）。2、公有云模式（互联网云平台，比如医疗机构租用的电信、移动、阿里、华为等公有云机房资源）。3、医院内网+VPN 模式（是上述二者结合的混合云模式，既数据存储在内网，技术服务商通过安全出口对平台进行日常维护）。三种方式的优缺点如下：

院内网模式

优点：

- ✧ 纯内网环境部署，医院无需顾虑来自互联网的安全威胁。
- ✧ 患者隐私等信息原则上无需做脱敏或者加密处理。
- ✧ 原生库到大数据平台数据交互，不受网络、安全设备等限制。

缺点：

- ✧ 由于完全内网部署，平台算法无法及时被公司更新，产品更新迭代时效受限。不像院内生产系统（如：HIS、EMR、LIS 等）相对成熟，大数据平台产品更需要快速、频繁的产品迭代。
- ✧ 维护成本增加。医疗大数据平台业内引用先进的 Hadoop 分布式存储与计算框架，目前院内信息化工作人员大多未能有效掌握这门技术，后续依靠信息部门自身力量进行维护有一定困难。
- ✧ 医疗大数据大数据工程师（科学家）属于稀缺资源，服务成本较高，不可能像传统医疗信息化合作企业为医院提供相对低廉的驻场运维维护服务。

安全：如医院选择该模式进行平台部署，建议服务器放置在医院业务核心区进行常规网络和数据保护，参照国家信息系统安全等级保护三级备案与评测标准，要求细节不再赘述。

公有云模式

优点:

- ✧ 降低院内大数据服务器资源投入。
- ✧ 降低医院维护成本。
- ✧ 便于未来医院更多“互联网+”业务拓展。

缺点:

- ✧ 网络安全方面，医院无法掌控。
- ✧ 数据安全、数据使用医院无法掌控。
- ✧ 院内临床科室访问大数据平台受限。
- ✧ 院内生产系统数据到大数据平台之间存在数据传输安全风险。
- ✧ 大数据平台数据自动更新很难实现，这方面需要长期投入人力资源去实现，人工干预增加了故障率和数据泄漏的风险。

安全：多中心、疾病联盟等科研数据可进行数据加密处理后适度存放。建议医院主体和服务商签订数据安全协议，以明确安全责任。

院内网+VPN 模式

优点:

- ✧ 大数据安全架构，可以被医院直接审计监控。
- ✧ 患者隐私、医生姓名等数据加密，医院完全受控。
- ✧ 院内数据到医疗大数据平台之间的数据安全、数据交互医院可控。
- ✧ 医疗大数据平台软件版本可快速更新迭代。
- ✧ 减少院内医疗大数据平台的运维成本。
- ✧ 技术服务商人员操作医疗大数据平台完全可控，操作行为完全被监控。
- ✧ 产品的更新迭代，更多的用机器代替人工，减少了人为操作带来的错误率和安全风险，更易被管理。
- ✧ 可以较为快速响应用户客户化需求。

缺点:

- ✧ 医疗数据存在对外出口，技术服务商服务存在不确定性因素。
- ✧ 增加了医院信息化管理工作内容（对外 VPN 通道）。

✧ 数据安全方面，提出更加高的要求（数据脱敏加密方面）。

安全：建议数据去隐私和加密处理后独立存储，服务器存储架构可置于二级安全区。有条件的可以设置为三级安全，同时建议和核心业务区适当隔离保护。

综上三种方式的优缺点分析，建议建设单位可依据本单位实际情况进行方案选择部署。第二种公有云方案是将来的发展趋势，特别有助于开展面向患者的互联网服务，以及多中心的临床研究需要。但目前情况下，公有云还不为医疗机构管理者以及信息部门负责人接受。医疗大数据平台建设初期，考虑到数据安全和产品更新，建议采用院内网+VPN 模式为医疗大数据平台建设方案。在平台建设和应用成熟后，医疗机构可适时管理（打开或关闭）VPN 通道，从网络技术层面保障数据安全可控。本《建设指南》以下内容主要针对第三种建设模式进行指导说明。

方案部署图如图 3.1 所示。

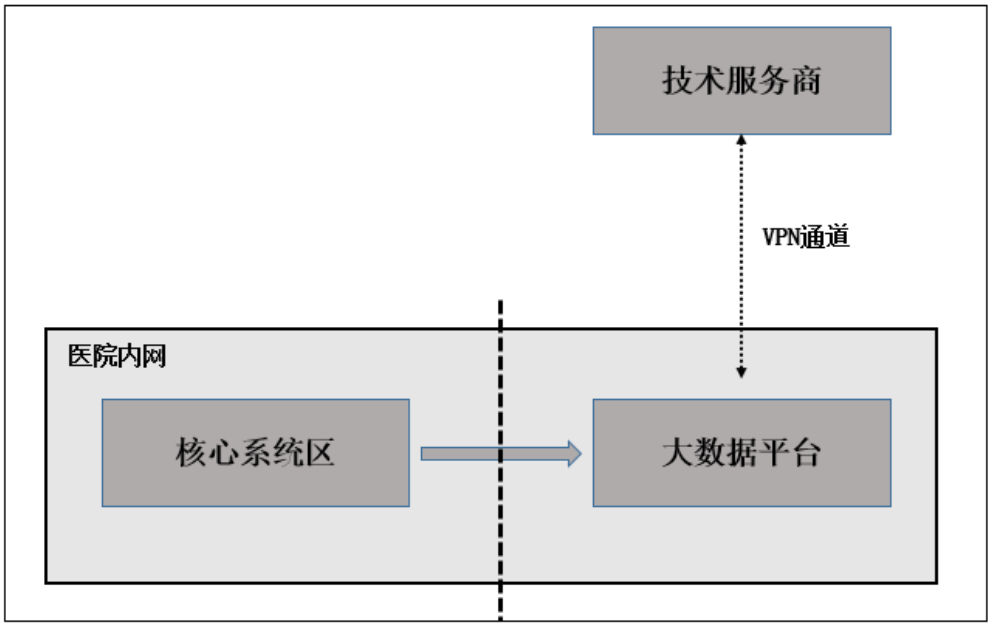


图 3.1 院内网+VPN 架构

3.1.2 安全措施保障

医院作为建设方，对技术服务商要从软件、硬件、数据、链路等方面提出安全体系要求。首先技术服务商提供的医疗大数据平台应达到信息安全等级保护三级要求，针对未评审的医疗大数据平台，医院也可将医疗大数据平台纳入医院等级保护三级管理中，要求技术服务商配合医院以医院等级保护三级为基础，实现平台等级

保护要求。

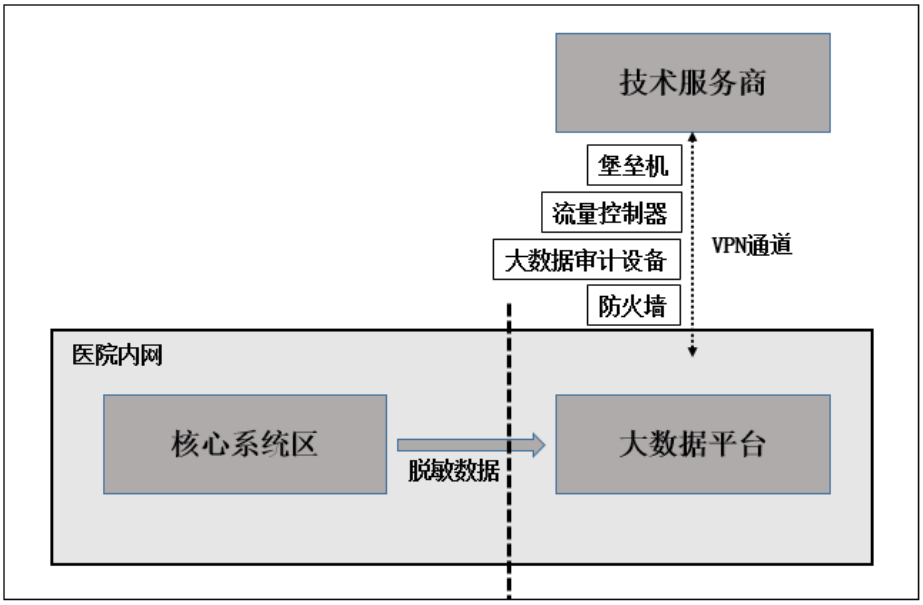


图 3.2 院内网+VPN 安全体系架构

（1）硬件安全：

为确保医疗大数据数据安全，所有平台相关的硬件设备应部署在医院中心机房，统一纳入医院固定资产管理，包括硬件日常巡检、维护等。同时医院应对硬件具有访问、维护等权限。

（2）链路安全：VPN 主要提供服务商到大数据平台访问链路，为整个链路的私密性和安全性提供保证。服务商提供的对医院的访问入口应采用 https 安全访问链接，保证数据浏览过程的网络安全。同时医院要对 VPN 进行管理，可以在服务商进行日常维护时开通 VPN，在平台不维护稳定运行情况下断开 VPN，以保证链路安全。

- 防火墙

医院大数据存储的防火墙部署在医院内网，主要负责服务商 VPN 到医院数据平台之间安全通信和访问控制； 防火墙具备访问控制功能、IPS、AV 等功能，除访问控制功能以外，IPS、AV 等功能根据实际情况开启。

- 流量控制器

准确记录 VPN 通道流量的上下行流量，在预设的时间窗口内，上行流量累计超过阈值时，会自动中断技术服务商数据平台的 VPN 网络链路，防止超额异常流量流出医院 。

流量控制器的流量控制配置和管理员权限，由医院管理。医院可以完全控制服

务商 VPN 服务中产生的流量总量，并且可以在医院的设置下自动或手动关闭或开通服务商间的网络链路。

- 堡垒机

堡垒机可记录所有账号在服务器的操作行为，核心是记录服务商 VPN 流入流出的所有流量内容，供医院或服务商进行实时的行为监控和事后的留档审计。对应监控配置管理、审计功能等管理员权限，均向医院开放，保证医院可以实时监控和审计技术服务商数据平台的所有行为和流量内容，确保数据平台的数据安全。

- 大数据审计设备

主要审计服务商通过 VPN 链路对大数据平台所做的任意操作，包括增、删、改、查、拷贝、黏贴等，对数据库的操作进行全面的审计。（目前市面上还没有相对成熟的产品，需要更多验证）。

为了保障医院对技术服务商员工的安全审计可查，防火墙、流量控制系统、堡垒机等安全设备的最高管理权限应上交给医院相关负责人。同时，所有安全设备的 syslog 按照医院要求，输出并记录到医院内部日志服务器。

（3）数据安全管理

- 数据脱敏

在数据处理前，参考国内相关法律法规或者国际较为认可的美国 HIPAA 法案。该法案要求对 18 类敏感数据脱敏，比如姓名、身份证号码、电话号码以及能够唯一标识某一个病人的信息。在数据库中，通过 MD5 加密或者其它加密算法进行处理，前端展现用“*”来代替。

同时对医生姓名信息在数据库中通过 DES、AES、RSA、MD5 加密或者其它加密算法进行处理，前端展现用“*”来代替，防止有人想用数据平台进行医疗数据统方。

针对敏感数据，数据库的每一字段都有脱敏规则。脱敏规则的设置和更改，由医院负责人员操作。建议要求技术服务商人员在处理数据时，通过保密 PC 和虚拟桌面来保障医院数据安全。所有数据加工处理都在保密 PC 和虚拟桌面上完成，以确保数据不会流出到服务商员工的个人笔记本等设备上。

所有脱敏的操作均在医院内部专用服务器上，由统一的工具完成，保证整个脱敏过程的安全。

- 数据加密

医疗大数据平台应使用加密存储和传输。

存储加密，即医疗数据平台处理后的数据先加密再存储，防止重要数据被非法窃取或窥探。加密过程选择了加密强度较高的加密算法，如 AES 等国际通用算法或 SCB2 等我国规定的国有商密算法。同时采用集中化的数据加密密钥管理与分发机制，实现对数据加密密钥的安全管理。

传输加密，在用使用浏览器搜索病历查看病历全文时，即数据传输过程中，也对传输的数据进行加密。用户首先需要获得由医院信息中心保管并分发的私钥。用户查看病历时，使用信息中心提供的私钥才能正确地解密病历，如同银行用户使用网银前，先向银行索取 UKey，或者下载安全证书一样。医院如果每名医生都有 UKey，即可采用 Ukey 的管理模式，实名管理每名用户，做到安全控制。

数据存储加密会在一定程度上影响计算及读取速度，技术服务商可以对医院进行按需调整。

数据加密机制保证数据不会被第三方轻易侦听，在数据应用服务的过程中难以被解读和察觉数据内容，以及部分数据因非预期的不可控因素遗失后，也会因不可解密而不会产生恶劣影响。

（4）应用安全

平台还应对数据的访问权限做全面控制。不同用户、不同业务系统运维人员或系统管理员的权限不同，登录平台看到的业务系统的数据亦不同。除此之外，还需经过安全审计与管理手段，包括但不限于以下内容：

① 所有权限由医院控制和分发；

② 单独用户群组、角色及权限管理；

③ 权限细化至每个人每个字段；

④ 用户无法自行注册，必须由医院管理员开通；

⑤ 平台必须先登录再使用；

⑥ 密码强度必须为大小写字母加数字的组合；

⑦ 支持用户证书登录；

⑧ 要求在医院内网使用，外网使用必须使用 VPN。

同时，平台会记录所有用户的数据访问及操作记录，方便事后审计。

4）主机系统安全

技术服务商除了具备针对医疗数据平台的特点而制定的安全防护策略外，在整个服务层面也需要采用第三方系统安全服务来建立全方位的系统安全机制，来守护整个系统的安全性。服务包括：

- 实时入侵监控

实时监控服务器和网络上的异常活动和数据流向，发现异常实时报警。

- 漏洞追踪和修补服务

系统漏洞没有及时修复，也是大部分系统被入侵的主要原因，应尽量全面收集公开和未公开系统漏洞，和服务系统比对，第一时间升级系统补丁，保证系统补漏洞，不给恶意攻击人员可乘之机。

- 渗透测试服务

在医院平台稳定和安全使用允许的前提下，院内 数据平台应定期模拟第三方进行入侵，检验整个系统在安全措施是否全面到位。

- 系统操作审计

记录 Linux / windows 系统上的所有操作行为，为安全审计系统信息记录。

- Web 安全防御系统

通过最新最全的扫描工具，对技术服务商提供的 Web 服务进行扫描，防止黑客通过 Web 的漏洞进行入侵,可以防范 SQL 注入,XSS,上传漏 洞等致命问题。系统会进行实时更新,可以阻断最新的攻击行为,还可以有效防止各种攻击变形。

- 登录安全系统

登录安全防御系统主要用于防止暴力破解和异常登录。可以有效识别和阻止针对 SSH, RDP,VPN,MAIL 系统等多种系统的暴力破解攻击。还可以在用户密码泄露后发现和锁定异常登录行为。

3.2 硬件部署

大数据平台所需的服务器、交换机、防火墙等硬件根据医院生成系统的数据量、医院病人总数、医院病例数及平台可持续性年数进行评估规划。但从数据安全、应用等角度考虑，一个基础的医疗大数据平台应至少包含以下硬件：

硬件类型	数量	用途

应用数据服务器	根据医院数据情况确定	ETL 服务 X 台、ES & Mongo X 台
数据处理服务器	根据医院数据情况确定	主要用于离线的大数据计算
交换机	根据医院数据情况确定	作为互备，提供内部网络连接
流量控制器	建议一台	监控流量、保证数据安全
堡垒机	建议一台	拦截非法访问和恶意攻击，内部操作的审计监控
防火墙	建议一台	有效监控和保证内部网络安全
审计服务器	建议一台	负责 VPN 专线操作审计

针对部分医院提到大数据平台能否采用虚拟化平台实现，建议不采用虚拟化平台实现。主要原因为：大数据计算在调度的时候会尽量把任务部署到离数据近的地方，比如同机器、同机架，进而极大减少网络传输。但底层存储如果是虚拟化的，在虚拟机上看起来是本机磁盘，实际上不知道存在哪里。网络的延时和带宽可能都会成为瓶颈。整体表现就是平台反应慢很多。除非虚拟化和大数据做联合的深度定制优化，否则效果一般都会很差。

所以一般大数据集群都建议采用物理机独立部署，虚拟化用来运行一些线上的 Web 产品。

3.3 数据接入范围

按照前述医疗大数据平台定义范围，即接入与患者相关就诊及治疗数据，及部分费用经济数据，包含但不限于：

- HIS：患者（含门诊、住院）的基本信息、就诊情况、病历、诊断、医嘱、用药、耗材、手术、输血、检查、检验等信息。
- EMR：门诊患者的门诊病历，住院患者的入院病历、病程、术前讨论、术后情况、出院小结、会诊记录等全部文书。
- 首页：包括临床首页和编目首页，以及临床随访和病案随访数据、经济数据、部分院外诊疗数据。
- 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。
- 手术麻醉：麻醉记录单、手术记录单、监控仪器数据。

- LIS：检查患者基本信息、身份信息、检查项目、检查细项、细项结果及正常值范围。
- RIS：检查患者基本信息、身份信息、检查报告、CT/MRI/PET 等各类文字报告原始文件。
- 病理：检查患者基本信息、身份信息、检查报告、涂片图像原始文件。
- 心电图：检查患者基本信息、身份信息、检查报告、心电图原始文件或 pdf 文件。
- 超声：检查患者基本信息、身份信息、检查报告、超声图像原始文件。
- 体检：患者基本信息、单位基本信息、体检项目清单、各项检查结果及正常值范围、各科室检查结论、终检结论、相关影像原始文件等。
- 患者医疗或者疾病相关的其他系统。

医联体单位在建设医疗大数据平台开展数据接入时，有条件情况下建议将医联体内所有医院涉及上述数据纳入平台管理。

3.4 数据接入方式

医疗大数据平台数据接入方式选择要考虑数据接入源、源数据库及接数模式。数据接入源包括原业务数据库、数据中心、集成平台三种。源数据库一般为 Oracle、MySQL、SQL Server 等。数据接入方式包括业务系统数据库备份恢复、数据同步（如 OGG）、物化视图、ETL（抽取工具）、集成平台等。

具体选择哪个方案？可以根据以下条件满足情况以及医院意愿进行选择。

需求	备份恢复	数据同步 (如 OGG)	物化视图	ETL 抽取工具
医院对修改生产服务器/库参数比较敏感	√	×	×	×
前置机数据脱敏 / 加密 / 自定义处理	√	√	×	×
特定类型数据库(MySQL/Caché 等)需要获取增量数据以支持增量生产	√	×	×	×

医院业务网与数据平台之间带宽不足（不能在需求所需的时间内完成备份传输）	×	√	√	√
秒/分钟级时效性（指生产库到前置机，不包括数据生产时间）	×	√	√	√
医院只能提供部分表，不能提供整个数据库	×	√	√	√
医院存在非数据库类型数据	√	×	×	×
数据库结构变更频繁	√	×	×	×
表无主键	√	×	×	×
计算机名与 SQL Server 内部服务器名无法一致	√	×	×	×

3.4.1 备份恢复

- 总体方案

此方案原则上不需要生产系统做大调整，只需要将生产系统日常备份数据库提供给平台即可。备份数据库在提交给平台前要做好数据脱敏、加密处理，考虑到无法在生产系统脱敏、加密后导出备份数据，所以先备份生产库，然后对生成库进行异地恢复，执行脱敏、加密算法，二次备份后提交给平台。

- 优缺点

✧ 优点

- 生产库只需进行备份操作即可，很少需要对数据库参数进行修改

Oracle 数据库需要开启归档，其它数据库没有参数需要特别设置。

- 只占用生产服务器网络带宽，对于其它资源配置没有特别要求
- 灵活配置备份、恢复的执行时间

可以配置成在业务低峰时段进行，减少对生产系统的影响。

- 最大程度的利用医院已有备份方案

如果院内的备份策略满足要求，可以直接使用已有的备份。

- 统一逻辑处理各种不同数据库类型

Oracle/SQL Server/MySQL/Caché/Sybase/DB2 文件形式存储的电子病历。

- 对于特殊数据类型、DLL(表结构变更)无需特殊处理，可以原生支持
- 支持对数据的自动脱敏/加密以及其它自定义处理（比如删除表或数据）
- 通过数据比对，可以获取差异数据

✧ 缺点

- 异构数据库支持不足

对于 AIX/Solaris 平台的数据库，无法使用普通的备份恢复，需要采用逻辑备份或数据库提供的异构平台迁移工具，执行时间比较长。

- 跨版本支持不足

Oracle 不支持跨版本备份恢复。

- 无法提供秒/分钟级的时效性（指生产库到平台，不包括数据生产时间）

只能提供小时/天/周/月级的时效性。

- 对于网络带宽要求较高

由于备份文件比较大，传输时间集中，所以要求医院生产库与平台之间的网络带宽不低于千兆。

- 只能对整个数据库级别进行数据备份，无法对表/字段级别进行处理

- 适用场景

  医院生产库与平台之间的网络带宽>=1024M

  不要求秒/分钟级的时效性（指生产库到前置机，不包括数据生产时间）。

3.4.2 数据同步

- 总体方案

利用数据库或第三方提供的数据同步工具实现生产库到平台的实时同步模式。此方案同样考虑到数据安全，需要在生产库和平台之前增加前置机，实现数据先从生产库到前置机，前置机完成数据脱敏、加密之后再同步到平台。

一般采用 OGG 方案。Oracle Golden Gate（简称 OGG）软件是一种基于日志的结构化数据复制软件。它能够实现大量交易数据的实时捕捉、变换和投递，实现源数据库与目标数据库的数据同步，保持亚秒级的数据延迟。

Golden Gate 能够支持多种拓扑结构，包括一对一，一对多，多对一，层叠和双向复制等等。

- 优缺点

  ✧ 优点

  - 支持异构数据库、跨版本数据库

  异构数据库、不同版本数据库之间可以同步数据。

  - 提供秒/分钟级的时效性（指生产库到前置机，不包括数据生产时间）
  - 对网络带宽要求相对较低

  数据同步实时传输，相对于备份的集中传输，对带宽要求较低。

  - 对前置机磁盘空间要求相对较低

  因为只保留同步的数据库，所以磁盘空间只需大于生产库实际大小即可。

- 可以对数据库/表/字段级别进行同步处理

✧ 缺点

- 需要在生产库安装/配置数据同步工具，并对数据库相关参数进行修改
- 需要占用生产库磁盘空间、内存等资源
- 不同数据库需要使用特定数据同步技术/工具，没有统一的处理逻辑
- 对于特殊数据类型、DLL(表结构变更)等，不同的同步工具支持程度不同
- 不支持对数据的脱敏/加密以及其它自定义处理（比如删除表或数据）
- 不同的同步工具，对于差异数据的获取支持程度不同
- 对数据库结构有特定要求，比如要求表必须存在主键

● 适用场景

✧ 医院生产库与前置机之间的网络带宽有限

✧ 要求秒/分钟级的时效性（指生产库到前置机，不包括数据生产时间）

3.4.3 物化视图

物化视图是包括一个查询结果的数据库对像，它是远程数据的本地副本。物化视图存储基于远程表的数据，也可以称为快照。

允许数据中心通过 DBLINK 链接到业务系统,增加一些之前是全量采集的大数据表的物化视图，方便数据中心增量采集这些表数据。

物化视图刷新方式

Fast 增量刷新：假设前一次刷新的时间为 t1,那么使用 fast 模式刷新物化视图时,只向视图中添加 t1 到当前时间段内,主表变化过的数据.为了记录这种变化，建立增量刷新物化视图还需要一个物化视图日志表，日志表要定期清理无用日志。

✧ 优点：

- 减轻对远程表即基表的访问压力，对业务库影响小
- 根据基表本身数据库时间戳增量刷新变化的数据；
- 将基表数据原始的粗糙的直接抽取到业务环境外的集结区,清洗,转换，集成等操作在一个专门的环境中完成，因为数据抽取对 CPU 和内存的要

求比较高，所以更加专门纯粹的环境更容易管理控制。

✧ 缺点：

- 数据存储占用磁盘资源可能是基表的几倍，需要足够大的磁盘空间支持；
- 由于作业调度等原因，占用 CPU 的性能也比较大，因此对数据分发 DTC 服务压力较大。

3.4.4 ETL 抽取

ETL 的过程一般可以这样描述：从操作型的数据源，经过数据中转区，最后到达数据集市的数据处理过程。底层是整个 ETL 过程中都涉及到的数据存储层。

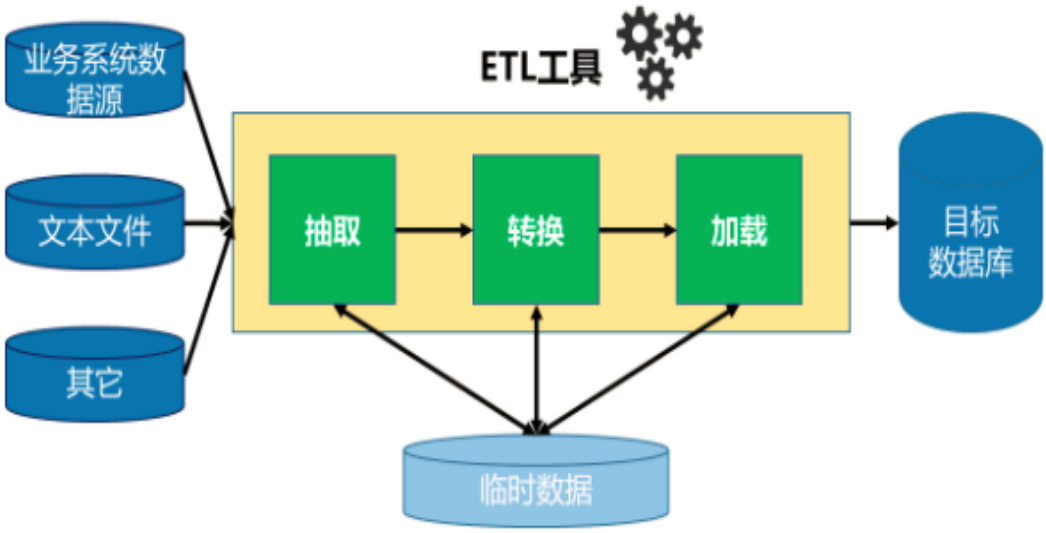


图 3.4.3 ETL 过程逻辑架构

- 数据源

上图的左边是数据源的提供者，业务系统是业务信息的来源。源系统数据存储类型由源系统规定，如一般的关系型数据库，文本文件、影像等。了解源系统的本质对于创建数据集市结构、ETL 过程结构等非常关键。

- 数据中转区（ETL 工具）

数据中转区是数据准备的工作台，其主要主要包括：

a. 可快速接受数据采集系统传过来的大量数据，缩短数据采集时间，减少数据采集对应用系统的冲击。

b. 实现对多个数据源的统一数据采集，提高了采集数据的可靠性、一致性。

c.暂时保存了要加载的数据，避免了数据转换系统对数据源的直接操作，减少了对数据源的影响。

d.对数据进行转换清洗的操作。

- 数据集市

数据集市包括维度表和事实表的存储，数据集市的数据结构是根据用户分析的主题需要来组织的，将所有数据组合为对组织的单一而又有相关性的视图。上图展示了 ETL 过程中三个数据的存储层。ETL 过程的数据移动过程如下：

数据源-数据中转区-数据集市。

每一层数据存在一定的依赖关系。最上层描述的是 ETL 的活动过程。通过抽取程序，把所选的数据源的当前快照或者捕获的数据源的数据变化，按需求追踪和充实新数据，抽取并挑选出数据。首先传到数据中转区，根据目标的要求，按照一定的规则把数据进行转换和清洗。合并后，把结果通过加载程序刷新或更新到数据集市的事实表与维表中。ETL 过程保证来自不同系统、不同格式的数据和信息模型具有一致性和完整性，并按要求装入 CDSS 的数据集市。ETL 的过程就是数据流动的过程，从不同异构数据源流向统一的目标数据。其间，数据的抽取、清洗、转换和加载形成串行或并行的过程。ETL 的核心还是转换过程，而抽取和加载一般可以作为转换的输入和输出，或者将抽取和加载作为一个单独部件，其复杂度没有转换部件高。

本系统主要关注如何解决在适当的时间通过适当的转换，将数据从一个位置正确地转移到另一个位置。了解如何将数据存储联系起来，帮助确定 ETL 活动过程中需要包括哪些内容。实际需要的数据存储依赖于业务需要以及提取和转换处理的复杂性。

3.4.5 数据增量抽取

增量抽取，只抽取自上次抽取以来数据库中要抽取的表中新增或修改的数据。在 ETL 使用过程中，增量抽取较全量抽取应用更广。如何捕获变化的数据时增量抽取的关键。

对捕获方法一般有两点要求：第一，准确性。能够将业务系统中的变化数据

按一定的频率准确地捕获到；第二，性能。不能对业务系统造成太大的压力，影响现有业务。

目前增量数据抽取中常用的捕获变化数据的方法有：

- 触发器方式

触发器方式又称快照式，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，临时表中抽取过的数据被标记或删除。此方式的优点：数据抽取的性能高，ETL 加载规则简单，速度快，不需要修改业务系统表结构，可以实现数据的递增加载。触发器方式也存在缺点：要求业务表建立触发器，对业务系统有一定的影响。

- 时间戳方式

时间戳方式是一种基于快照比较的变化数据捕获方式，在源表数上增加一个时间戳字段，系统在更新修改表数据时，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戳字段的值，来决定抽取哪些数据。有的数据库时间戳支持自动更新，即表的其他字段的数据发生改变时，自动更新时间戳字段的值。有的数据库不支持时间戳的自动更新，这就要求业务系统在更新业务数据时，手工更新时间戳字段。

该方式的优点：同触发器方式一样，时间戳方式的性能也比较好，ETL 系统设计清晰，源数据抽取相对清楚简单，可以实现数据的递增加载。缺点：时间戳维护需要由业务系统完成，对业务系统也有很大的倾入性（加入额外的时间戳字段），特别是对不支持时间戳的自动更新的数据库，还要求业务系统进行额外的更新时间戳操作，工作量大、改动大、风险大。另外，无法捕获对时间戳以前数据的删除和修改操作，在数据准确性商受到了一定的限制。

- 全表删除插入方式

每次 ETL 操作均删除目标表数据，由 ETL 全新加载数据。该方式的优点：ETL 加载规则简单，速度快。缺点：对于维度表代理键不适应，当业务系统产生删除数据操作时，综合数据库将不会记录到所删除的历史数据，不可以实现数据的递增加载；同时对于目标表所建立的关联关系，需要重新进行创建。

- 全表比对方式

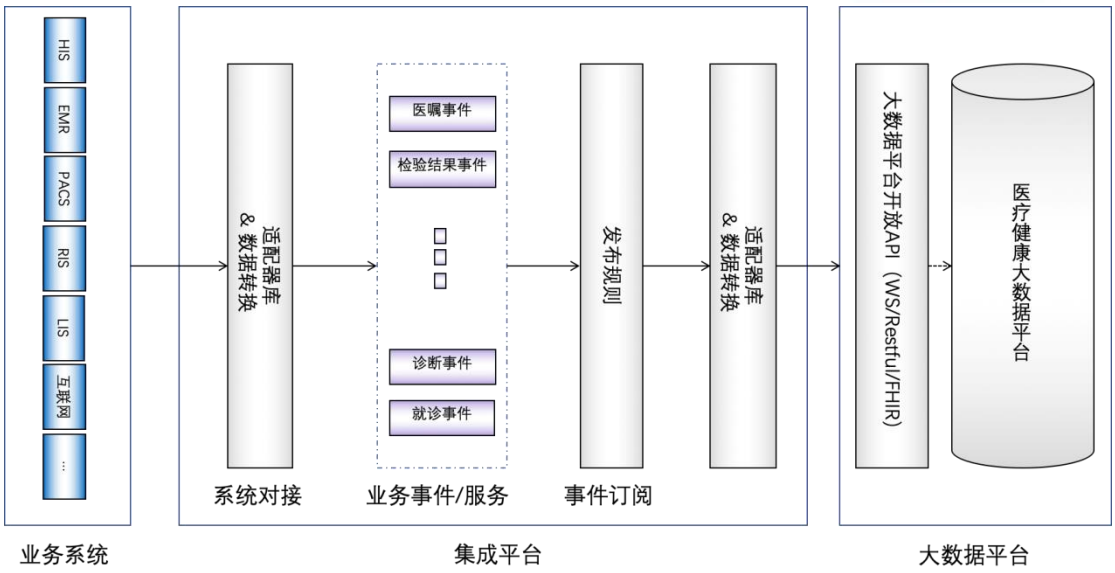
全表比对的方式是采用 MD5 校验码。ETL 工具事先为要抽取的表建立一个结构类似的 MD5 临时表，该临时表记录源表主键以及根据所有字段的数据计算出来的 MD5 校验码，每次进行数据抽取时，对源表和 MD5 临时表进行 MD5 校验码比对。如有不同，进行修改操作。如目标表没有存在该主键值，表示该记录还没有，即进行插入操作。该方式的优点：对已有系统表结构不产生影响，不需要修改业务操作程序，所有抽取规则由 ETL 完成，管理维护统一，可以实现数据的递增加载，没有风险。缺点是：ETL 比对较复杂，设计较为复杂，速度较慢。与触发器和时间戳方式中的主动通知不同，全表比对方式是被动进行全表数据的比对，性能较差。当表中没有主键或唯一列且含有重复记录时，全表比对方式的准确性较差。

- 日志表方式

在业务系统中添加系统日志表，当业务数据发生变化时，更新维护日志表内容，当做 ETL 加载时，通过读日志表数据决定加载哪些数据及如何加载。优点：不需要修改业务系统表结构，源数据抽取清楚，速度较快。可以实现数据的递增加载。缺点：日志表维护需要由业务系统完成，需要对业务系统业务操作程序作修改，记录日志信息。日志表维护较为麻烦，对原有系统有较大影响。工作量较大，改动较大，有一定的风险。

3.4.6 集成平台数据提取

针对部分已经上线集成平台的医疗机构可以采用医疗大数据平台从集成平台订阅消息，获取实时数据的解决方案，详细结构图如下：



通过集成平台实时获取数据架构图

使用集成平台获取数据主要经过一下过程：

适配器：由于各个业务系统从设计到实现技术的差异性，集成平台需要具备连接不同技术接口的能力，通过适配器可以适配不同业务系统的连接口，从而将这些业务系统对外接口封装成相应的业务事件。

数据转换：平台通过适配器获取数据后，需要对数据进行标准化处理，包括：数据过滤、数据格式标准化、语义（术语）标准化、以及数据有效性校验四个步骤，通过这四个步骤的处理，保证进入到大数据平台的数据的完整性和有效性。

消息订阅/事件驱动：一旦数据接入完成，并进行了标准化处理后，这些服务就形成了标准化的事件。通过消息订阅机制，针对大数据平台关注的所有业务事件相对应的数据进行订阅，一旦业务系统产生了一条新的数据，集成平台就能捕获到该事件，进而实时将事件对应的数据发送给订阅方（大数据平台），最后通过大数据平台提供的接口发送到大数据平台。

3.5 数据脱敏加密

3.5.1 数据脱敏

对于医疗大数据平台的数据应用而言，由于平台集成了医院所有医疗信息系统的患者信息，数据体量大；同时，临床基于医疗大数据平台的应用在使用

时一般可以直接主动访问平台数据，导致数据泄露风险增大。所以，原则上要求在建立大数据平台时要有数据脱敏、数据加密机制。

具体来讲，数据脱敏（Data Masking）是一种通过替换数据中的敏感信息，或者对数据中的敏感信息进行变形等处理的技术，又称数据漂白、数据去隐私化或数据变形。处理后的数据看似真实，但是不会暴露任何敏感信息。对于想要滥用此数据的人而言没有任何使用价值，保证在开发环境、测试环境和其它非生产环境，以及数据共享环境中安全地使用脱敏后的真实数据集。

基于大数据的数据脱敏机制，大数据平台在接收数据时，医院要对医疗数据进行脱敏处理。推荐的数据脱敏方法包括：替代、混洗、数值变换、加密、遮挡、空置插入或删除等。

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。

但无论选择何种数据脱敏机制，考虑到部分数据在临床医疗研究中的价值，平台应该还具备反脱敏功能（将脱敏数据还原为原始数据）。反脱敏机制应建立在一整套平台授权机制上，包括依据医院管理规定，那些数据允许临床可以看到，那些数据必须执行脱敏，或经过一定授权后可以给临床患者所有明文数据等。

大数据平台的脱敏和加密是否是必选环节，各医疗机构可以根据自身的使用目标和范围进行选择。一般来说如果提供全院全员的数据查询，科研发现，以及对外多中心的科研课题，一般应选择对数据脱敏处理；如果前期未规划和实施脱敏，后期再进行处理，对平台架构设计和数据处理流程的稳定性都会有挑战。所以建议先期统筹考虑，同时加密算法应掌握在信息部门手中。脱敏后的数据应具有可逆性，既通过严格授权机制可逆，以满足不同业务需求（比如

临床随访等）。

3.5.2 数据加密

医疗大数据平台的数据除了执行脱敏操作，还应执行加密部分信息的操作。具体到加密方案，可以借鉴美国的 HIPAA 法案的相关思路。

HIPAA（Health Insurance Portability and Accountability Act）是美国 1996 年制定的联邦健康保险便携与问责法案。这项法律的主要目的是让人们更容易保持他们的医疗保险，保障医疗信息的保密性和安全性，并帮助医疗行业控制行政成本。

HIPAA 分为五个部分（Title），每部分解决医疗保险改革的一个特定问题。该法案解决的第二个问题是简化管理，重点在医疗相关的信息安全问题。HIPAA 建立了一整套用于接收，传送和维护医疗信息，并确保隐私和个人身份信息的安全标准。这部分的核心是对 PHI（Protected Health Information）等信息的隐私和安全保护。

。

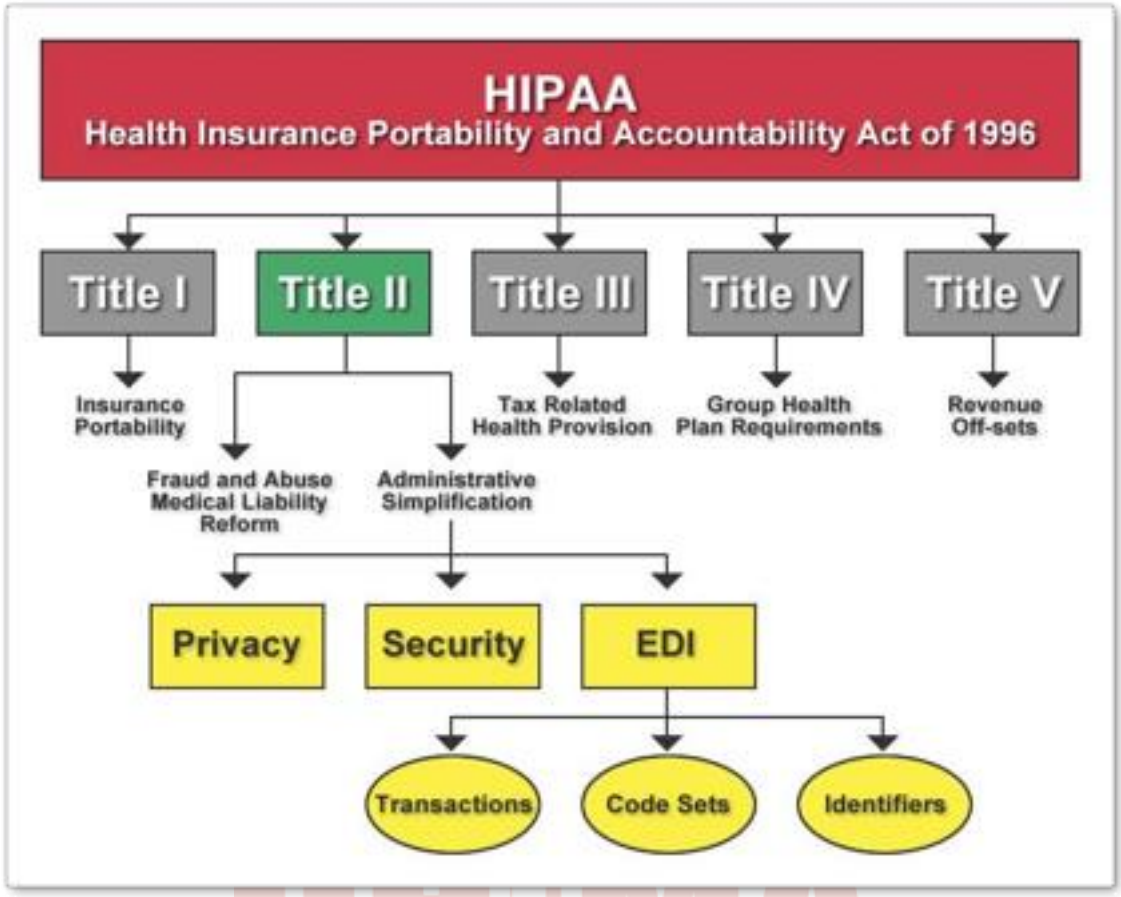


图3.5.1 HIPAA法案

如上图所示，第二部分的标黄部分的内容与 PHI 信息的保护及脱敏处理直接相关。

（1）加密字段范围

为确保患者医疗隐私数据不被泄露，需首先完成患者相关隐私数据的脱敏处理（属静态脱敏方式）。

结合现有字段定义、实际使用情况以及防统方需求，并参照 HIPAA 对 PHI 定义的 18 项内容，逐项整理了字段对应分析和相应脱敏建议。

具体如下表所示：

	HIPAA PHI	字段	脱敏建议	备注
1	姓名 Names	P_N	滤除（*号代替）	
		C_NAME	滤除（*号代替）	
2	地址类信息	B_P	保留	大于 20000 人范围
		S_D_NAME	仅保留城市区县	大于 20000 人范围
		D_NAME	保留	大于 20000 人范围
		DISTRICT_CODE	保留	大于 20000 人范围
		D_CODE	滤除（*号代替）	患者家属信息，与科研无关
		D_NAME	滤除（*号代替）	患者家属信息，与科研无关
		S_D_NAME	滤除（*号代替）	患者家属信息，与科研无关
		C_ADDRESS	滤除（*号代替）	患者家属信息，与科研无关
		P_CODE	滤除（*号代替）	患者家属信息，与科研无关
		C_E	滤除（*号代替）	患者家属信息，与科研无关
3	日期类信息	B_DATE	保留年	同 HIPAA 要求
		P_AGE	保留，>90 岁写年龄段	同 HIPAA 要求
		DIAGNOSIS_DATE	保留	
4	电话号码	M	滤除（*号代替）	
		T	滤除（*号代替）	
		C_TELEPHONE	滤除（*号代替）	
		C_MOBILE	滤除（*号代替）	
5	传真号码		滤除（*号代替）	
6	电邮地址	EMAIL	滤除（*号代替）	
7	社保编号	INSURANCE_CARD_NO	滤除（*号代替）	
8	病历号	BED_NO	滤除（*号代替）	
		V_SN	保留	不直接体现用户信息，用来关联各个数据表

		P_SN	保留	不直接体现用户信息，用来关联各个数据表
9	医疗保险受益人 ID 号		滤除（*号代替）	
10	账号	MEDICAL_CARD_NO	滤除（*号代替）	
11	身份证号	C_NO	滤除（*号代替）	
		C_TYPE_NAME	滤除（*号代替）	
12	驾照号码/汽车牌照		如后期增加，则滤除	
13	其他设备的序列号		如后期增加，则滤除	
14	URL		如后期增加，则滤除	
15	IP 地址		如后期增加，则滤除	
16	指纹、虹膜、声纹		如后期增加，则滤除	
17	全脸照片及任何可识别出患者的影像		如后期增加，则滤除	
18	任何其他独特的识别号码	WORK_PLACE	滤除（*号代替）	
		N_NAME	滤除（*号代替）	
		NATION_CODE	滤除（*号代替）	
19	医生姓名	Doctor_name	滤除（*号代替）	反统方

（2）加密方法

加密方法有两类，可逆加密算法和不可逆加密算法。

可逆加密算法，就是通过加密秘钥给数据加密，当然数据可以通过解密秘钥

进行数据解密，在特定的业务场景或者高级账号权限的基础上，可以通过调用解密秘钥算法进行数据机密。

不可逆机密算法，行业内我们比较熟悉的就是 MD5，是一种不可逆的哈希值算法，将数据加密为一组 32 位无规则字符串，且过程不可逆，也就是说加密后数据不可能通过反向算法进行解密。

如果贵单位在医疗大数据平台使用中，有需要使用患者隐私信息的可能性（比如随访业务场景的需要），建议采用可逆加密算法进行相关数据加密处理。反之建议采用“MD5 加密”不可逆加密算法。

（3）数据加密环节

无论是数据加密采用可逆加密算法和不可逆加密算法，建议加密在数据离开院内网环境前，进行数据加密处理。

3.6 数据处理

数据处理属于医疗大数据平台的核心功能，对从业务系统接入数据进行处理，实现数据应用，为临床、管理、科研提供数据支持。数据处理过程包括两个环节：数据验收、数据生产。数据验收主要是核查接入平台的原始数据是否跟业务系统数据匹配，避免在前期的同步、脱敏、加密等过程中丢失原始数据。原则上该过程在数据生产之前必须完成，以保证进入数据生产过程的数据无问题。

3.6.1 数据验收

本阶段的数据验收，指大数据平台从业务系统接收的经过脱敏、加密后的数据的完整性、一致性、正确性等进行验收。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。如果从业务系统接收过来的数据经过脱敏、加密仍存在问题，未通过验收，这部分数据即使经过后期处理，也是不准确的。

数据验收包括：验证方法、全量验证、逻辑验证，验证部分要明确公司和医院两方分别采取何种验证方式。

全量验证：主要比对平台接收到数据与业务系统数据在同一逻辑条件下数据总量是否一致。比如某一个出院时间范围内，出院病人总数一致。

逻辑验证：主要验证部分数据是否合规。逻辑验证内容一般比较多，比如可以验证某一数据表中某一字段情况与原业务数据一致性等。

3.6.2 数据生产

数据生产属于医疗大数据平台的核心功能。一般大数据平台厂商系统都具备该功能，但不同厂商数据生产模式、步骤不一样。本《建设指南》在此举例介绍一种常见数据处理、生产模式，不作为本《建设指南》的指导性内容。

将原始的数据库结构映射到统一的基础数据模型上，进行统一的清洗、对照、EMPI、EMOI 等处理，最后将数据按照统一的数据交换标准（如 HL7）的业务单元规范进行整合处理。

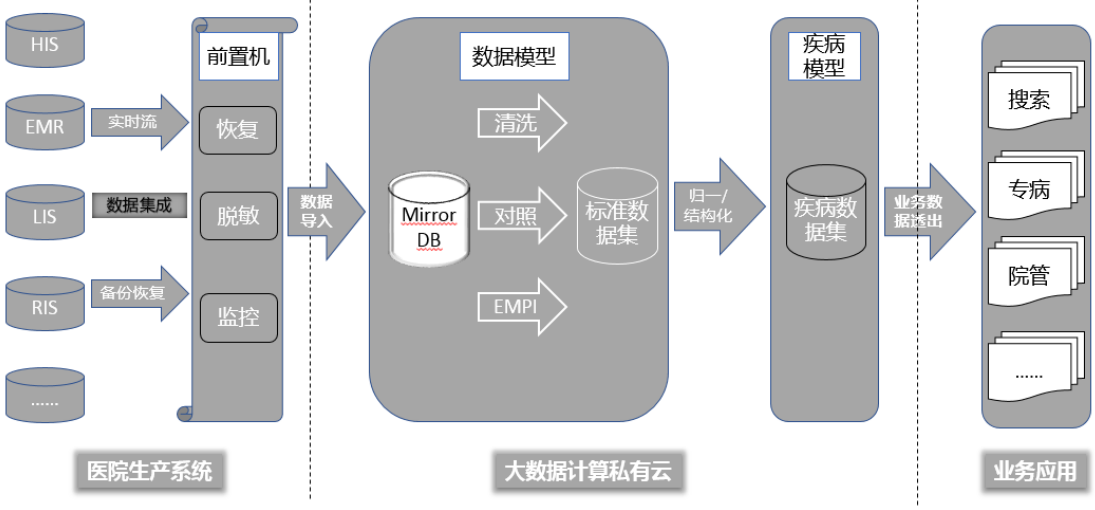


图 3.6.4 数据处理流程图（待更换）

医疗大数据平台对数据的处理，要从数据的重构开始，就是将各个业务系统数据打散之后，患者诊疗模型重构。

总体数据模型分为以下业务单元：

图 3.6.5 总体模型

3.6.2.1 ETL 映射

患者诊疗模型经过重构之后，需要采用相关数据标准进行数据抽取。

临床各业务系统之间数据格式和数据标准不一致，需要大数据中心制定相关数据存储标准，对历史数据和实时数据进行统一转换和清洗。

按照互联互通标准化成熟度、电子病历系统应用水平分级评价等相关规范，制定符合医院后期数据利用要求的数据存储规范、数据转换对照（目前市场尚未规范对照标准），也可选择成熟案例较多的大数据厂商，对实例进行确认和修正。

数据转换，主要是根据大数据中心标准，与原始业务数据进行对照。按照数据元的方式，对数据类型、限定条件和值域等进行匹配映射，把原始业务数据转换为符合大数据中心标准的数据格式。在数据转换过程中，校验每一个数据项的匹配情况。当发现必需项、日期类型等与数据元定义不匹配时，及时记录不规则数据日志，后期进行数据修正。

利用大数据技术的分布式计算功能，对实时接收业务系统所推送的消息，进行数据逻辑关系映射，有利于数据的快速查询和展现。

将原始医院 HIS、LIS、EMR、RIS、PACS 等系统的数据映射到统一的数据模型，成熟的大数据厂商一般可以对国内市场主流 HIT 厂商的数据模型进行归纳，并进行算法特征提取，形成一整套 ETL 的知识库，提升系统映射的效率和准确性。

3.6.2.2 数据清洗

数据清洗是针对单行单字段为粒度的数据格式转换。医院数据来源于多个系统，不同医院同一系统数据格式也不尽相同，且医院原始系统数据中存在各种异常的数据格式。为统一化处理，数据在清洗时需对必要字段做格式的统一化处理。

清洗举例：

字段	清洗前	清洗后
住院日期-出生日期	3 岁	3YOMOD
患者姓名	&张 三 ！	张三

（1）数据清洗规则

数据清洗规则包括：非空检核、主键重复、非法代码清洗、非 法值清洗、数据格式检核、记录数检核。

非空检核：要求字段为非空的情况下，需要对该字段数据进行 检核。

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，需进行检核工作。

非法代码、非法值清洗：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。

数据格式检核：通过检查表中属性值的格式是否正确来衡量其 准确性，如时间格式、币种格式、多余字符、乱码。

记录数检核：指各个系统相关数据之间的数据总数检核或者数 据表中每日数据量的波动检核。

业务约束核验，应在平台建设商实施处理数据过程中，与医院业务人员共同确定。业务人员要从业务的正确性、一致性、有效性等角度，考虑数据的核验规则，如：建档日期、入学日期、民族信息等的有效性核验。

（2）脏数据处理

对于常见的空缺值、离群值和不一致等脏数据，可以采用人工检测、统计学方法、聚类、分类、基于距离、关联规则等方法来实现数据清洗。

根据缺陷类型分类，可以将脏数据分为缺失值数据、错误数据和错误关联数据这三种典型问题数据，分别进行数据清洗。

3.6.2.3 对照

将医院的数据字典映射到统一的数据字典，对照的范围包括：科室对照，检查类型对照、诊断类型对照等。

以诊断类型为例， 统一的字段包括如下：

No	编码	诊断类型
1	DN000001	门诊诊断
2	DN000002	入院诊断
3	DN000007	死亡诊断
4	DN000008	出院诊断
5	DN000010	术前诊断
6	DN000011	术中诊断
7	DN000012	术后诊断
8	DN000015	转入诊断
9	DN000016	转出诊断
10	DN000019	尸检诊断
11	DN000025	会诊诊断
12	DN000031	损伤中毒

图 3.6.7

对照分为两个步骤：

✧ 第一步，通过特征模型进行机器对照；第二步，机器无法准确识别的，采用人工对照。

✧ （1）EMPI

✧ 患者主索引处理， 将相同的患者信息进行合并，合并规则如下：

✧ a.姓名、性别、出生年月都合法，姓名大于一个字，性别为男或女，出生年月非空；

✧ b.身份证号、姓名、性别、出生年月 4 个字段完全一致；

✧ c,身份证号不合法但非空时仍然按 4 个字段完全一致。

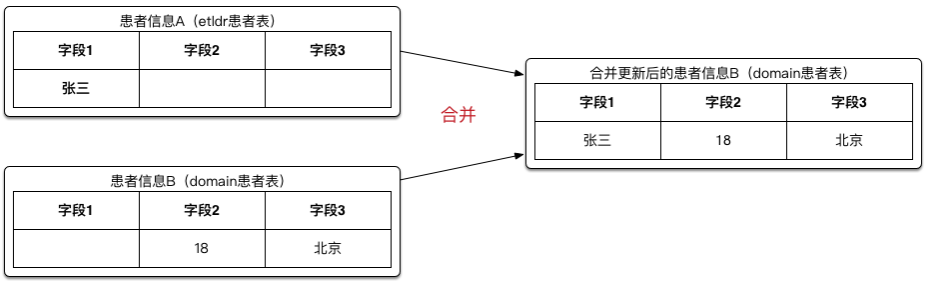


图 3.6.8

（2）EMOI

很多医院的医技系统、电子病历系统等和 HIS 分离，无法互联互通，系统之间回传的就诊号无法相关联，可能导致检查、检验、电子病历数据无法和就诊及患者相关联，会对数据完整性造成较大影响。

针对以上情况，根据数据在医院业务系统的产生状况，通过现有系统之间的信息，建立匹配规则，将满足规则的检查、检验、电子病历数据与就诊及患者相关联。

EMOI 逻辑示例：

患者此次就诊能够对应到患者信息，就诊表中的就诊时间或入院时间早于或等于检查业务时间，检查业务时间从以下字段中选取：申请时间、登记时间、检查时间。

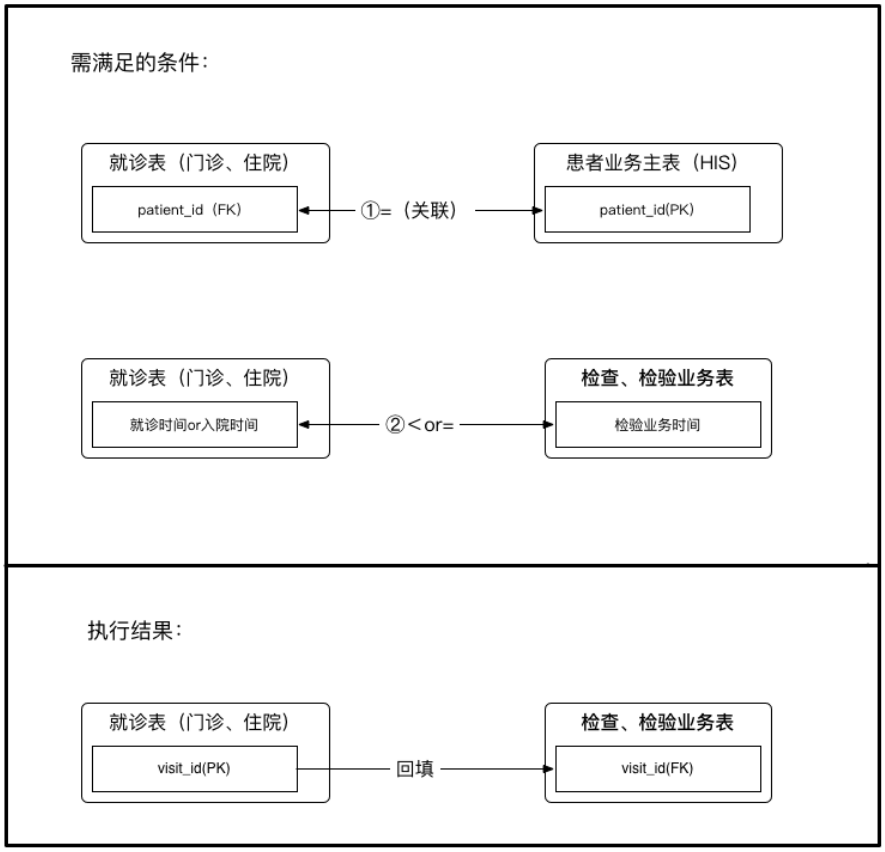


图 3.6.9

3.6.2.4 主数据标准化处理

由于临床数据的不规范性、随意性等特点，在进行医疗大数据利用前，需要将临床主数据进行标准化处理。

主数据：临床系统中的字典库，例如：科室字典、诊断字典、药品字典、检验检查字典等。

可参考国内外相关标准，对数据进行标准化处理。参考的规范有：ICD 10、ICD11、MESH（医学主题词表）、ICD-9-CM-3、LOINC、CFDA、ATC 分类、国家卫计委 – 医疗机构诊疗科目名录等。

序号	分类	标准名称
1	卫计委数据标准	■ 中国卫生信息数据元值域代码WS364.X-2011（X从1到17,共17部分）
2		■ 电子病历基本数据集WS445.X-2014（X从1到17,共17部分）
3	HL7CDA文档	■ HL7ChinaCDA规范试行2013版(共5个)
4	国际性肿瘤数据库结构	■ 美国国家癌症研究所（NCI）SEER计划编码和分期手册2015版
5		■ 美国外科医生学院机构肿瘤注册数据标准FORDS2015版
6		■ 英国国家癌症智能网络国家癌症数据储存库数据定义NCDRR_V5.2
7		■ 美国肿瘤临床协会（ASCO）治疗计划最终版（Online）
8		■ 英国国家临床分析与特定应用小组(NCASAT)放疗数据集RTDS数据手册及实施指导V4.0.8
9	肿瘤学国际诊治指南	■ AJCC/UICC临床分期手册（第7版）
10		■ 美国国家癌症研究所常见不良事件评价标准第四版CTCAE V4
11		■ 美国国家癌症研究所肿瘤放疗小组远期放疗反应评估表
12		■ 实体肿瘤疗效评估标准(RECIST) V1.1（美国，英国，加拿大，欧洲等）
13		■ ……
14	相关术语标准	■ 国际疾病分类第九版临床修订第三卷：手术与操作 ICD-9-CM-3
15		■ 国际疾病分类第10版 ICD-10
16		■ 国际疾病分类肿瘤学专辑第三版 ICD-O-3
17		■ 观测指标标识符逻辑命名与编码系统LOINC V2.42
18		■ Karnofsky功能状态评分标准（Online）
19	国家相关数据标准	■ GB/T 2261.1-2003 个人基本信息分类与代码 第1部分 人的性别分类
20		■ ……
21		■ GB/T 4671-2008 家庭关系代码

图 3.6.10 医学数据行业标准

以诊断为例，需要建立疾病相关同义词图谱，对临床书写不规范的诊断进行标准化处理。

图 3.6.11 疾病同义词图谱

以诊断“贲门恶性肿瘤”为例，ICD 10 诊断编码 C16.001，同一家医院对应的临床诊断原词可能有上百种，常见的有：贲门癌、贲门癌复查、贲门 Ca 等等，比如某 肿瘤医院的数据，具体数据情况参见下图：

图 3.6.12

数据经过疾病诊断数据模型标准化处理后，才可以得到充分利用。无论医生书写的是标准词汇还是非标词汇，都可以通过平台的转化，从而实现数据的充分利用。

3.6.2.5 结构化处理

医疗数据主要包含患者的基本信息、病历、医嘱、护理文书、检查所见、检查结论等。这些数据反映了患者的基本信息、临床诊断、治疗过程和结果。随着医疗系统信息化建立和完善，越来越多的医疗数据由人工记录的方式转为电子化录入，对于病历、医嘱、护理文书、检查报告等临床信息主要由医疗人员通过自然语言的方式书写而成，信息结构较为复杂，如何使得计算机能够理解这些医疗信息中所

包含的语义，能够高效对这些数据进行存储、检索、统计、分析和挖掘将是医疗信息化建设的一个重要问题。

结构化具体目标是基于医疗信息学的角度，将以自然语言方式录入的医疗数据根据医学语境转化为可用于存储、查询、统计、分析和挖掘的数据结构。依据目前实际业务需求，目前结构化对象数据包括但不限于如下：

- 病历中一诉五史：主诉、现病史、既往史、月经史、社会史、家族史；
- 手术记录：手术过程描述、术中出血量等；
- 物理检查报告：检查所见和检查结果；
- 病理报告：检查所见和检查结果；
- 诊断结果：术前诊断、术后诊断和出院诊断。

结构化主要从若干个独立维度来进行，对数据依据主题字段进行划分，主要主题字段有：症状、体征、烟酒情况、病理诊断、病理表现、过敏情况、婚育状况等。根据病理或报告中不同字段的语义复杂程度和实际需求，目前结构化框架主要由正则抽取和通用框架组成：

正则抽取主要针对语义比较单一或规则性较强的字段，可以采用正则表达式直接对文档中的关键信息进行抽取。例如，对于句子“月经情况 ： 初潮 15 岁，5/26 天 ， 末次月经 2016-2-14”，可通过正则表达式对日期的模式进行提取末次月经的时间，结果为{‘末次月经’:’ 2016-2-14’}。

对于比较复杂的文本，例如患者症状，通用框架基于一些基本语法规则、医疗常识基础上，通过自然语言处理的技术方法来分析文本的隐含语义和上下文结构关系，主要包括：分词、消歧、模版匹配、语义分析、规则过滤等技术方法。

基于结构化框架需求，目前主要通过专业医疗人员标记和数据挖掘技术两个方面结合来提供解决方案。一方面，医疗文本数据不同于主流语料，包含很多专有医疗相关名词、词性和语义，只通过传统的自然语言处理技术很难完成；另一方面，又需要借助一些自然语言处理算法，在已有标记数据基础上，发现更多的医疗语义规则，降低医疗人员的人工标记成本。解决方案主要表现为，根据不同的技术方法，发现和挖掘结构化所需要的知识模块，主要包括：

- 分词词典
- 同义词词库

- 主题和属性词关系图谱
- 上下文匹配逻辑
- 规则库
- 正则模版

通过这些知识模块会将病历等本信息结构化成一个便于存储、查找、分析的数据结构。

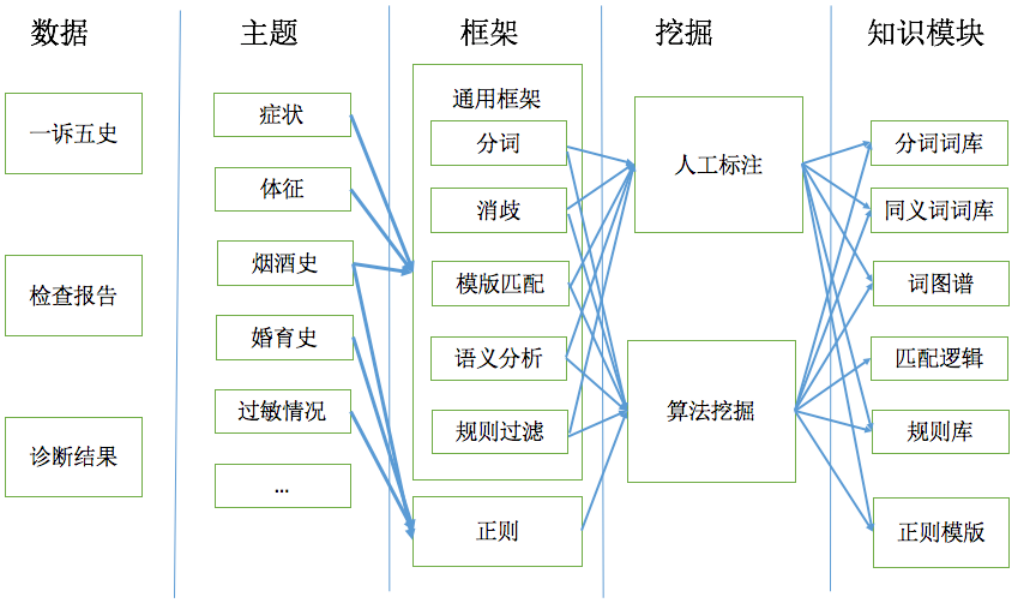


图 3.6.13 结构化整体流程示意图

最终,还需要对最后结果进行评估,评估的主要依据两个指标,准确率和召回率:准确率指结构化的结果满足文本语义信息的在所有结构化结果中的占比,召回率指成功结构化字段在本文中所有关键语义中的占比。在评估后,根据评估的结果和需求目标分析并改进在结构化中存在的不足。通过不断迭代评估和完善知识模块,使结构化具有一定的泛化性,在不同的医疗语料中都能有良好的表现。

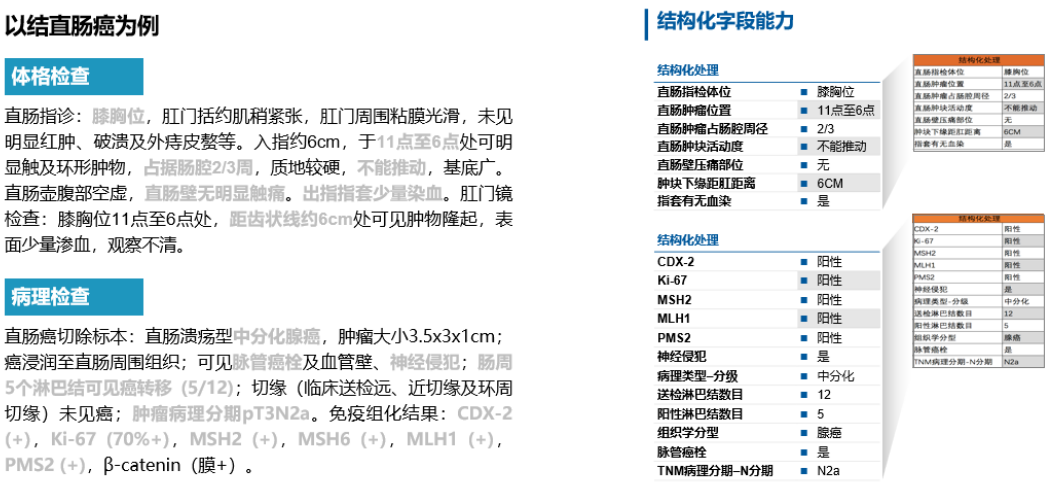


图 3.6.14 结构化示例

3.7 数据扩展

医疗大数据平台具有以下特点：（1）非实时，面向医疗运营、科研和临床的院内外数据源；（2）无特定主题，数据丰富程度极高，与业务系统松耦合；（3）面向医疗机构，对数据进行二次利用。大数据平台除了可以接入院内数据源外，还可以扩展对接院外的患者可穿戴设备数据、气象学数据、环境学数据、基因数据等，进而把各种数据融合在一起，让医疗大数据的发展不仅仅局限于院内数据，当然是保证网络安全和数据安全的前提下。

在当今医疗大数据领域，院外数据已经有成熟案例接入医疗大数据平台，例如诊疗指南数据、文献库、基因数据、院外诊疗、随访数据等。外部关联数据的接入，可进一步完善和丰富医疗大数据平台数据资源，为使用者提供更多应用支持。

3.8 数据授权

大数据平台需要对数据的访问做全面的控制，不同用户、不同业务系统运维人员或系统管理员的权限不同，登录平台看到的业务系统的数据亦不同。除此之外，建议制定一系列的安全审计与管理手段，包括但不限于以下内容：

- 所有权限由医院信息部门（或科研处、大数据中心等医院授权的管理部门）控制和分发；

- 单独用户群组、角色及权限管理；
- 权限细化至每个人、甚至一些特殊字段单独授权；
- 用户无法自行注册，必须由医院管理员开通；
- 平台必须先登录再使用；
- 密码强度必须为大小写字母加数字的组合；
- 支持用户证书登录；
- 要求在医院内网使用，外网使用必须使用 VPN。

同时，大数据平台需要记录所有用户的数据访问及操作记录，方便事后审计。

3.9 数据验证

数据验证指验证医疗大数据平台处理、生产后的数据与原业务系统数据的一致性、完整性、正确性。源数据通过大数据平台处理后，如何保证数据跟源数据无异常、无异差是必须通过数据验证环节来实现。只要验证通过后数据才可以被临床所使用。数据验证环节建议医院和承建单位双方进行共同验证。

数据质量验收建议从以下 8 个质量维度进行：

- 完整性（completeness）：数据项内容的完整情况，主要考察字段数据是否存在，NULL 等同于空。
- 规范性(normalization)：数据项的数据格式，数据长度，时间范围，数值范围是否正确。
- 标准性（standardization）：数据项的数据是否与字典一致，如性别的取值应该是男，女，未知。
- 正确性 (accuracy，correctness)：数据项内容是否正确，如入院记录中月经史不为空时，患者性别不应该为男。
- 一致性 (Integrity, concordance)：数据项间内容是否一致，如考察病案

首页上的患者、手术、诊断、费用等信息是否同 HIS 或其他系统中的信息一致。

- 时序性(currency)：时间逻辑合理性(流程时间分布)，如出院时间需要晚于入院时间
- 整合性(integration)：相关系统对应数据项目可对照或关联，如医疗费用明细表里的医嘱标识应该在医嘱标示在中西医处方表、药品类医嘱表、草药处方表或者非药品医嘱表里。
- 唯一性(uniqueness)：记录的唯一性，如 patientid 在 patient 表不能重复，相同的医嘱不应该重复。

数据验证方法：

1、初期的数据验收应要求平台建设方提供数据报告，包括接入数据范围（系统数量）、数据量（需要分类，比如患者、医嘱、处方、检验检查数量等）、数据纵深（起始到终止日期）。医院信息部门工程师可进行宏观数据核验。

2、数据质量抽样检查：比照 HIS、EMR 等业务系统的单个患者进行数据核验（完整性、一致性、整合性等）；可以由技术服务商工程师和医院工程师共同进行；

3、病案数据抽样检查：随机抽取一定数量的病例数据，交付技术服务商工程师在大数据平台前台界面对数据进行核对。

4、隐私数据检查：查看是否按照要求进行去隐私化处理，比如患者姓名、地址、身份证号等，可有意识的搜索隐私数据看可否能检索到。

检出问题后及时汇总和分析出问题原因。要么接入对接偏差，要么工程师理解偏差，或是后期后结构化归一的次生原因，作为后期调整技术方案的依据。

3.10 平台验收

平台验收指大数据平台的整体产品交付验收。前文提出的数据验收是基础，输入业务系统原始数据，大数据平台进行整合和技术加工，产出应该是符合项目要求的功能和数据。平台的验收包括基础验收、速度和性能验收，以及功能验收。基础验收功能多偏重于对技术服务商的要求，因为只有了解数据的

验收过程细节，验收才能确保数据产出的价值提升。所以系统对医院和技术服务商都有借鉴。而性能和功能的验收，更多偏重于医医院的应用体验，是以信息部门为代表的收货确认，是平台交付临床和科研、管理科室客户使用的前提。

3.10.1 速度、性能

医院信息部门可根据技术服务商交付的测试应用，选择部分患者诊疗数据进行多组合检索，测试平台响应速度。建议如下：

响应时间：

用户进行在线实时查询业务操作，实现秒级搜索（查询时间低于 5 秒）。

系统容量：

可存储的科研数据（包括 EMR 数据、生物样本库数据、实验室数据等）应满足实际工作需求，并根据发展规划，设计系统冗余。

系统在容量、性能、用户数量等方面具有扩展性，能够满足未来 5 年的业务需求（HDFS 需要 3 倍空余为佳，由于服务器架构具有扩展性，不是验收重点）。

系统并发用户数：

用户同时在线数量，须满足系统中心未来 5-10 年发展需求（硬件平台的扩展性）。

用户同时并发数量满足协同工作实际需要（用户并发数的影响因素和对平台性能的影响）。

安全性验收：1、VPN 断开测试：如果采取院内存储加 VPN 模式，需要进行断开 VPN 后平台可用性测试，以确保系统的独立可靠性。2、数据监控和审计测试，主要针对平台的堡垒机、数据库审计、防火墙等安全设备的线上运行情况的测试。看是否正常授权、行为记录、屏幕录像、流量监控和预警等功能。

3.10.2 功能验收

医疗大数据平台产品功能上至少具备如下功能：

- 支持院内临床相关数据的高效搜索引擎：可以根据检索词秒速应答并反馈结

果；

- 多条件逻辑组合的，相似病历、疑难病历搜索：高级检索功能，能根据平台归一或结构化前后的准确信息点进行精准检索；
- 搜索符合患者、病历信息的统计分析功能：平台应提供基础的可视化的检索信息统计，包括表格、图示，可点击挖掘；
- 患者全景视图：能以患者为核心展示诊疗信息，包括相关符合诊疗常规的分类；
- 患者全量数据时间轴：能按照患者就医行为时间或关键医疗事件时间，按照时间轴的方式展现；
- 患者隐私信息保护：按照管理要求，在保障诊疗数据安全可回溯的前提下，能够实现数据的去隐私化；
- 科研患者队列（人群）智能纳排（操作简易、可视化、多维度）；
- 数据导出功能：用户检索数据可实现导出功能；
- 数据使用权限管理：有数据分科室、分人员、分属性的权限管理，以及数据隐私回退权限管理。

3.11 平台培训

数据平台基本验收后，可适时交付临床和科研科室试用或正式使用。并进行系统化的培训。一般分为：

1、用户培训：面向大数据平台的使用者，比如临床大夫、护士、或科研、管理人员。培训方式可分为：

集中培训：

大数据平台操作知识等前期培训。对各科的计算机骨干进行重点的培训，加强对系统的熟悉和对维护及一般故障处理等知识的培训。

分科培训：

如场地或时间有限制时对人员采用分科培训的方式进行前期培训，和在集中培训后对部分人员再进行巩固式培训。

2、平台管理员培训：一般指信息部门或者大数据其他主管部门的培训，主要包

含用户授权、软件安装、数据加密和反加密工具使用、数据质量监控、数据接入监控、数据统计工具使用、数据导出管理等

培训一般由信息部门或者大数据平台主管部门组织，由平台建设商进行，培训前应准备培训环境和培训资料。

3.12 数据管理

医疗大数据平台的建设的核心目的是服务于最终用户，包括一线医生、护士、科研人员、管理人员。但由于平台汇集了大量包含患者敏感信息的医疗数据。如何保障科学、合理、安全的使用，需要医疗机构以及其数据管理部门制定相应的管理流程、管理制度，以便在平台推广实践中落实。

- 数据安全管理制度：数据安全管理从系统安全和审批管理两方面进行考虑。系统安全包括信息部门数据备份策略、用户权限、账户弱密码、账户有效期和数据导出管理等方面进行管理；审批管理主要指账户权限审批流程和数据申请审批流程。
- 数据使用管理规范：数据使用人需严格遵守国家有关法律法规，申请数据时需注明申请数据范围、用途、使用事件等重要信息，对使用数据有保密责任，数据使用结束要立即向管理部门报备，不可更改数据用途或进行其他违规操作。
- 知识产权管理制度：多家单位数据共享使用应遵循平等互利、诚实守信、成功共享的原则，参与方需提前拟定合作协议，所获利益按参与方贡献大小分配。
- 使用违规惩罚制度：对于擅自传播、转让、更改数据用途、未经许可使用数据或其他违规行为，医院管理部门有权利追究法律责任。

本指南收集整理了部分医疗机构数据管理制度示例，可参考附录 B。

第四章 应用场景

通过医疗大数据平台的建设，可更便捷地对医院内积累的大量数据进行深度的分析、挖掘，建立专项科研课题，进行回顾性或前瞻性科研分析；找寻体征、诊断、用药、治疗方式等的相关性，分析医生的诊疗路径，优化指南，形成更加科学的诊疗知识库，作为分级诊疗的基础。

应用大数据、人工智能的分析和优化手段，使医生、护士更专注疾病本身，患者更信赖医生、护士的医疗行为，能更好地配合治疗，优化整个诊疗流程；对于临床大量的非结构化的数据，进行结构化处理，深度挖掘出有意义的价值。同时，医疗大数据平台支持传统医学研究模式研究，也支持传统技术所不能实现的真实世界研究。

4.1 临床应用场景

4.1.1 临床大数据搜索

4.1.1.1传统搜索模式

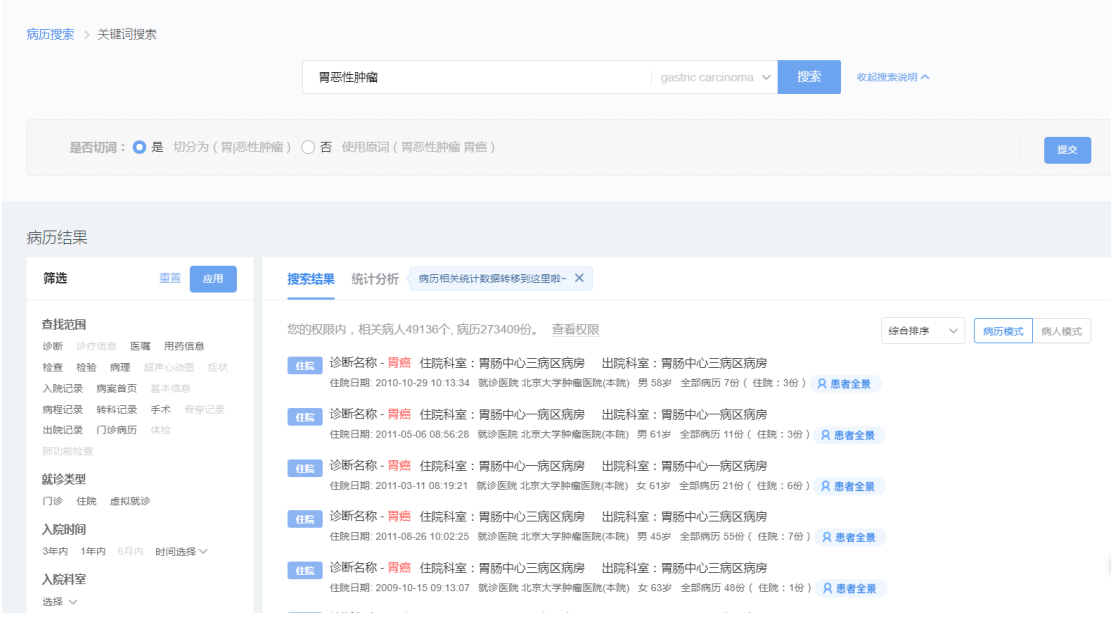
搜索场景是临床和科研中最基本也是最重要的应用。

传统的数据存储模式大部分是基于关系型数据库设计，性能效率低下，能支持的搜索方式也很有限，专业技术要求较强。通常需要医生提交申请表，信息部门检索符合医生需求的患者信息，然后将患者信息（如患者 ID）汇总成 Excel 表形式提交给检索申请者。，申请者通过翻阅扫描病历或既往电子病历收集所需要的科研数据，最终汇总成 Excel 表格供科研使用。传统检索模式有以下弊端：(1）数据体量小：传统数据查询都是基于各个业务系统进行的独立查询，关联性差，查询的数据量也小；(2）查询速度慢：传统基于关系型数据库查询在做大批量查询时速度比较慢；(3）工作效率低：传统数据查询都是手工查询完成之后再人工处理，费时费力。传统检索模式很难满足医生日益增长的科研需求。

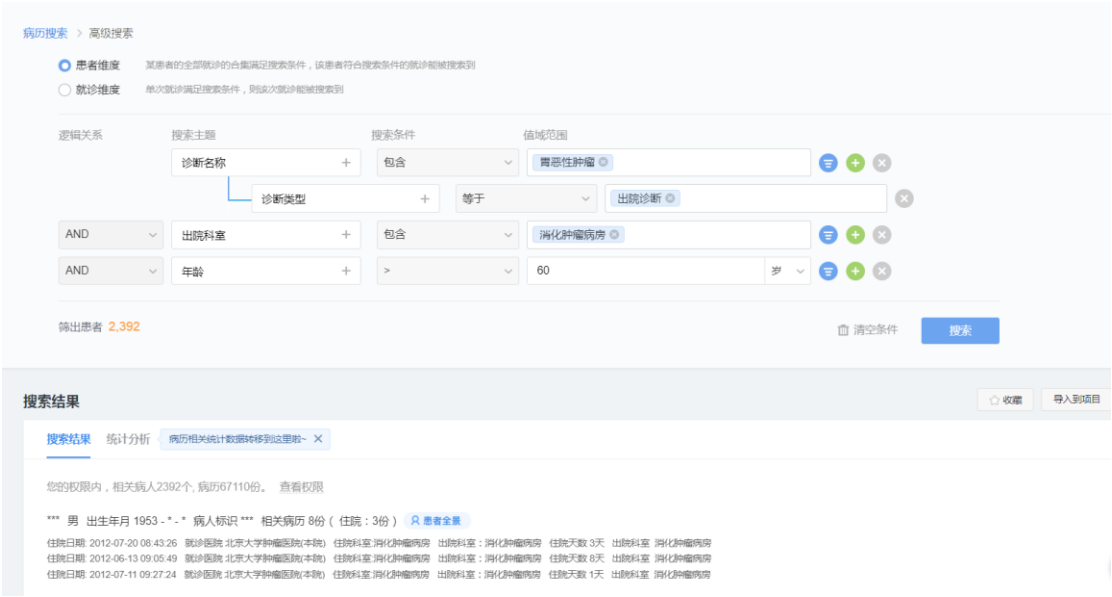
4.1.1.2大数据搜索模式

基于大数据架构的设计可有效优化传统检索问题，可以用于全文搜索、结构化搜索、分词搜索、模糊搜索以及复合搜索等多种模式。在医疗搜索的垂直领域有模糊搜素、关键词搜索、条件搜索三种具体的应用模式，且在各种复杂场景下的搜索性能表现也较为强健，均能够实现秒级别的搜索。

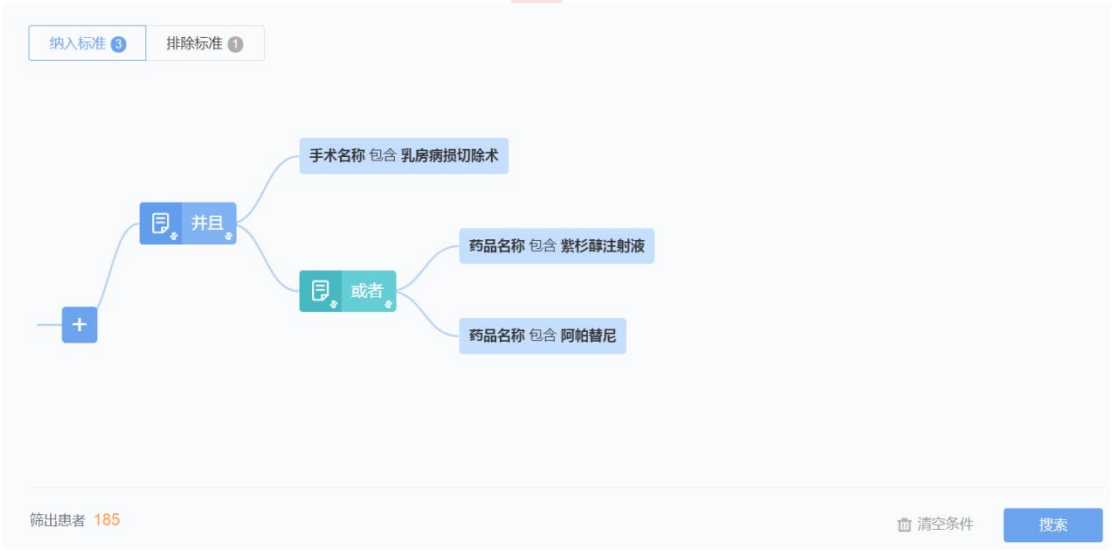
(1)关键词搜索：提供便捷的快速关键词搜索入口，通过医疗专业字典分析、切词等技术处理，检索符合请求条件的病历结果，并提供了各种灵活的筛选方式、排序方式和搜索结构的专业统计，如下所示：



(2）高级搜索：高级搜索用于描述多个复杂的检索逻辑和条件，可以精确召回需要的病历或患者。高级搜索包括了逻辑关系/搜索主题/搜索条件/值域范围四个建立高级搜索条件的变量，以及患者维度/病历维度的搜索展示。具体如下所示：



（3）条件树搜索：条件树搜索相对于高级搜索更加灵活，能够将并且、或者和排除三种逻辑关系按照需求任何进行组合。具体如下所示：



4.1.1.3大数据搜索的难题与应用

大数据搜索能为医生带来比较流畅的搜索体验，但是目前仍面临着诸多难题。由于科室间利益、医疗工作本身的个人属性（医生手术等医疗行为）等，使得数据的所有权具有争议。医生调阅不方便，调阅时间、方式需要优化，提高使用粘度。检索得到的数据如何导出，去隐私数据如何还原等问题都需要进一步解决和优化。

目前，已有的解决方案包括：（1）将大数据平台入口嵌入住院工作站、门诊工作站，方便临床调用，提高应用粘度；（2）依托工作站的账号体系和科室权限，解决账户授权、科室查询数据权限的问题；（3）建立数据导出管理制度（如审批流程），规范数据导出。

未来也需要不断的探索和总结，利用更完善的方案来解决大数据搜索应用过程中的新难题。

4.1.2 多学科诊疗(MDT)

4.1.2.1 传统 MDT 模式

在临床实际工作中，医院经常会基于某些疑难杂症、典型病例发起会诊讨论或回顾性分析讨论。在病例讨论之前相关负责人需要整理病例资料，通过制作传统 PPT 和 Word 资料，手工翻看病历总结及详细信息。参会各专业人员讨论病例并给出意见，手工记录。在这个过程中传统的收集病例信息的方式比较繁琐，讨论现场病情展示也很局限，手写的会诊记录也无法回顾和追溯。存在诸多的局限，影响临床工作效率。

4.1.2.2 大数据驱动的 MDT

基于大数据的应用平台，可以优化整个病例讨论的流程，并且为病历讨论建立比较完整的管理流程。这将，在一定程度上提高医生工作效率，服务于临床科内、跨科以及跨院的病例讨论场景，可以实现线上会议、快速在线编辑、基于大数据平台的患者会诊资料的智能汇总，实现关联患者既往病例资料、影像资料，以及相关疑难病例、相似病例或相似患者、相关指南和文献等需求。MDT 会议期间，可实现在线的会诊病例资料投屏、语音录入和记录会诊过程专家意见、专家会诊意见记录；会后可基于平台跟进患者诊疗方案和效果评估，实现真正数据可追溯的 MDT 平台。MDT 管理部门也可进行统计分析，以及疑难案例汇总成册提供更多医生学习。

下图为 MDT 数据库病历列表：

图 4.5

4.1.3 患者全息视图

4.1.3.1 传统模式

医生在实际诊疗过程中，除了关注患者本次入院（就诊）情况之外，还会关注患者既往诊疗方案、患者既往检查检验结果、患者是否具有其他的特殊关注点，比如以前用过什么特殊的药物，是否做过什么重要手术等。传统模式下，医生需要登录院内多个系统查看患者信息，且很难查询到既往全部的诊疗行为。此场景在会给医生的实际临床以及科研工作带来极大的不便。

4.1.3.2 患者全息视图

在大数据模式下，通过整合不同系统间（HIS、LIS、RIS、EMR、护理、手麻、ICU、PACS 等）的数据通道，能够以“患者全息视图”的方式展示患者的全治疗周期，记录患者在每一个时间节点的诊断、用药、体征数据、检查、检验、治疗、手术等数据。

“患者全息视图”是基于临床数据的临床应用，临床医生可以通过清晰、友

好的统一视图对患者的就诊信息进行查阅，从而优化医生的操作流程，使临床医生在短时间内对患者就诊情况有整体了解。“患者全息视图”收集了全量的临床数据内容，可以实现临床数据的 EMR 数据、检验报告、检查报告同时查看。“患者全息视图”是临床场景中最基础的应用，相对于传统的患者全息视图，大数据应用能赋予患者全息视图更友好的诊疗支持体验：

（1）“时间轴”作为整体患者全景应用中总领全局的模块，能够在大数据平台的基础上建立时序模型，完成整体功能的建设。开发如下所示的功能应用界面：

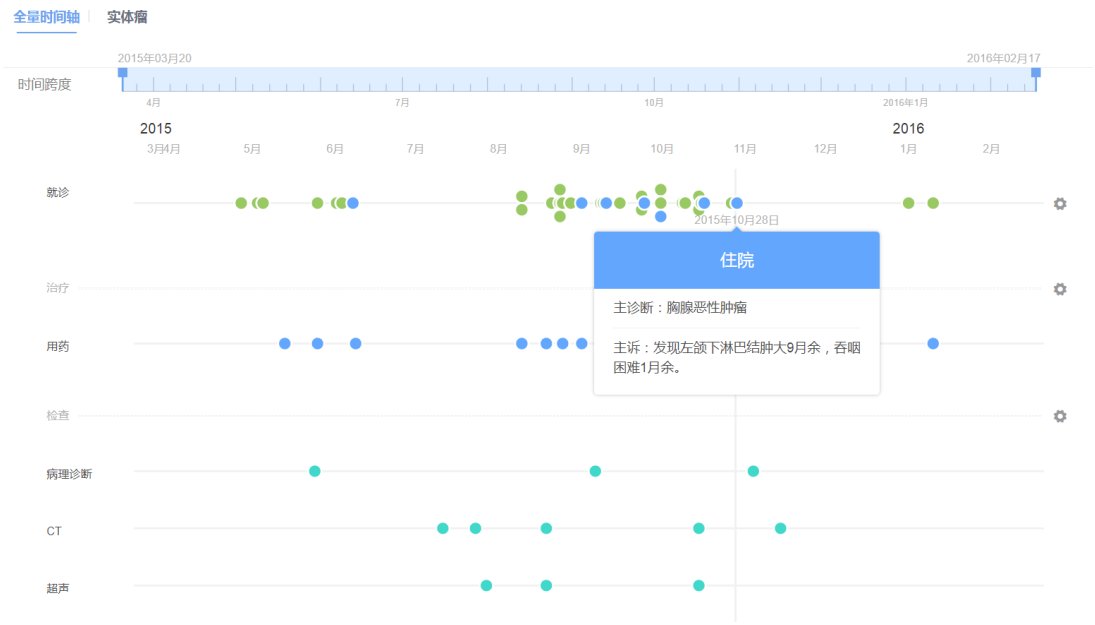


图 4.1 全景视图

（2）患者全息视图体验：因为对医院所有系统数据进行了数据和业务的重构，故在数据加工和数据增强方面有优势，数据加工和数据增强能够帮助医生更好的应用该平台，针对该患者所有就诊信息可以在一个页面综合信息的浏览，能更加快速定位所需要查看的病历；将患者的历次检查数据通过时间的顺序，按照不同类型快速解读其变化情况；通过患者全息视图的检验模块将某个检验的历史全量数据按照时间先后顺序展示其趋势图，并且可以添加更多的指标进行趋势对比等。

4.1.4 临床决策支持

4.1.4.1 CDSS 发展矛盾与瓶颈

临床决策支持系统（Clinical Decision Support System, CDSS）一直是医疗信息化领域的一个重要发展分支。基于医疗信息化的发展，CDSS 在中国医疗市场的发展和探索也已经经历了二十年左右的时间，逐渐演化出应用于临床不同场景和服务于不同层级医生的 CDSS 类型产品，并在医疗市场内进行了广泛推广和使用。

从产品应用形态上，基本可以分为：基于知识库的查询类 CDSS、基于知识规则的推荐审核类 CDSS。

无论是基于知识库的查询类产品，还是基于知识规则的推荐审核类产品，都与真实临床场景存在一道很难逾越的屏障：知识库查询产品在真实应用中偏碎片化，缺乏与医院信息化系统的深度拟合，只能对临床场景起到补充和解决特殊问题的作用，对提升医疗效率和质量的作用非常微小。而基于知识规则的推荐审核类产品，本质上是以有限的数据规则去覆盖无限的临床个性化情况，本身就存在很大的瓶颈。在该类产品的研发上，即使投入巨大也可能收效甚微，并且在实际应用中往往出现两种情况：不是机器推演出的推荐审核结果与实际情况下医生的认知存在很大差异（医生一般仍然根据实际情况该怎么做就怎么做），就是系统认知问题的维度和推演逻辑太简单，起不到根本上的帮助作用（医生不用系统推荐也会这么做）。因此，研发并演化出更加友好贴近临床逻辑的、更加符合真实世界数据的新型的临床智能辅助系统成为一个必然趋势。

4.1.4.2 大数据时代下的 CDSS

自 2018 年《电子病历系统功能应用水平分级评价方法及标准（试行）》发布以来，CDSS 相关功能成为相关细则明确的医院信息系统建设的核心要求。2018 年，国家卫健委明确，要求到 2019 年，所有三级医院要达到电子病历系统应用水平分级评价 3 级以上，即实现医院内不同部门间数据交换；到 2020 年，所有三级医院要达到电子病历系统应用水平分级评价 4 级以上，即医院内实现全院信

息共享，并具备医疗决策支持功能。

基于大数据应用的智能临床决策支持系统，是在医疗大数据应用平台的标准数据流生产的基础上，利用大量的病历作为分析样本，并结合临床知识，抽象出疾病特征字段进行建模，结合专业的临床知识库，为临床医生提供全方位的辅助决策服务。内容设计思路如下：

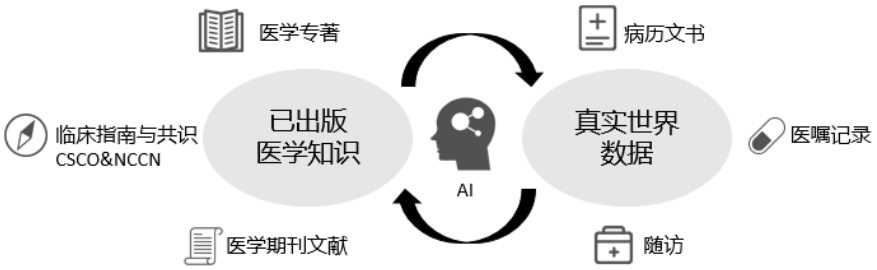


图 4.2

因此基于真实世界数据加上循证医学知识的思路，可以有两个应用方向：基于不同复杂病种的单病种智能辅助诊疗，以及应用于工作站的适用于通用疾病的临床智能助手。

（1）单病种辅助诊疗：主要是将出版的知识或者顶级临床专家的经验方案，结合真实病历数据进行分析，将这些内容机器规则化形成决策引擎，支持不同病种不同特征信息的录入后，根据该引擎的规则输出可用于临床参考的诊疗方案。同时，基于该系统，进一步实现方案采用情况的跟踪分析，从而最终对临床决策过程进行管理和优化，提升该病种的诊疗效率及质量。业务流程设计示意如下图：

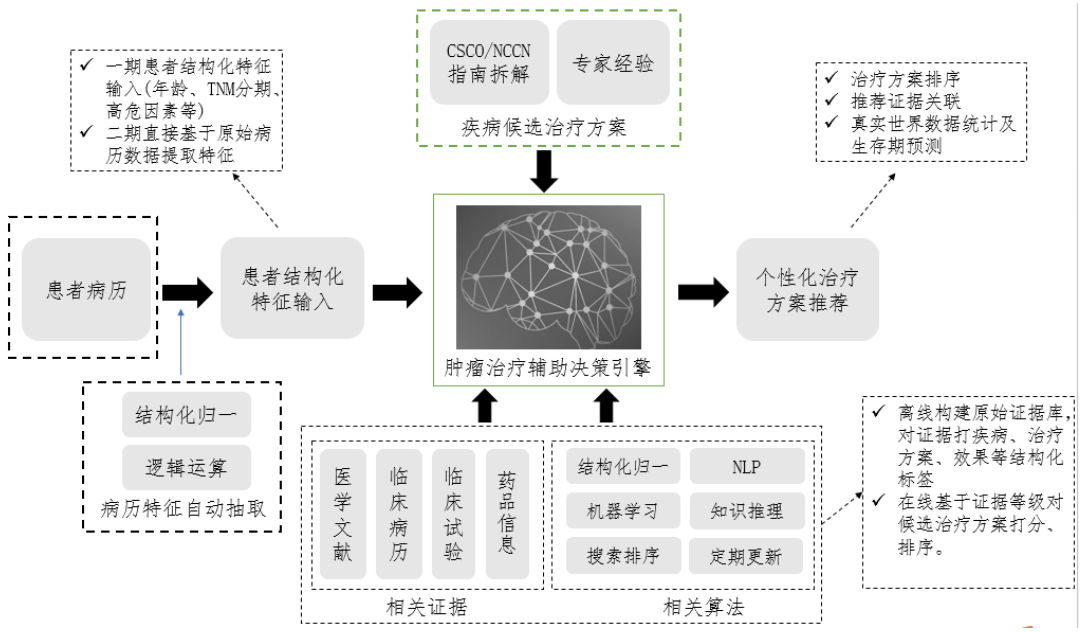


图 4.3

（2）医生工作站临床助手：主要是基于权威循证医学知识库和真实世界临床数据，构建更加精准且契合临床思维的知识图谱。在医生诊疗过程中进行预测、推荐、预警，提供知识库查询、疑似疾病预测、检查方案推荐、治疗方案推荐、异常提醒、智能审核等覆盖诊疗全流程的辅助决策支持，降低漏诊和误诊，提升高医疗效率和质量。该应用主要涵盖业务示意如下：

图 4.4

（3）相似病历和相似患者推荐：通过医疗大数据平台，可建立病历画像和患

者画像的模型。以疾病为维度，可以建立基于可参数调整的相似病历和相似患者的模型。对于整体的诊疗行为推荐、会诊讨论等诸多场景进行广泛的应用。

4.2 科研应用场景

对于科研场景来说，医生面临的难点主要有如下情况：科研思路发现困难；诸多病历的非结构化字段处理需要大量的人力；处理完的数据需要时间精力的转换和处理才能进行分析；整个科研过程想进行延续而搭建专业疾病数据库需要专业的平台。而依托于医疗大数据平台，对于上述的诸多困难均能得到一定程度的解决。大数据平台从临床科研链路的全流程（科研灵感的发现、初步调研验证、科研立项、圈定目标人群、观测指标的建立、数据如何收集以及最后的统计分析和文章撰写等）予以逐步支持，帮助临床医生进行科研工作。

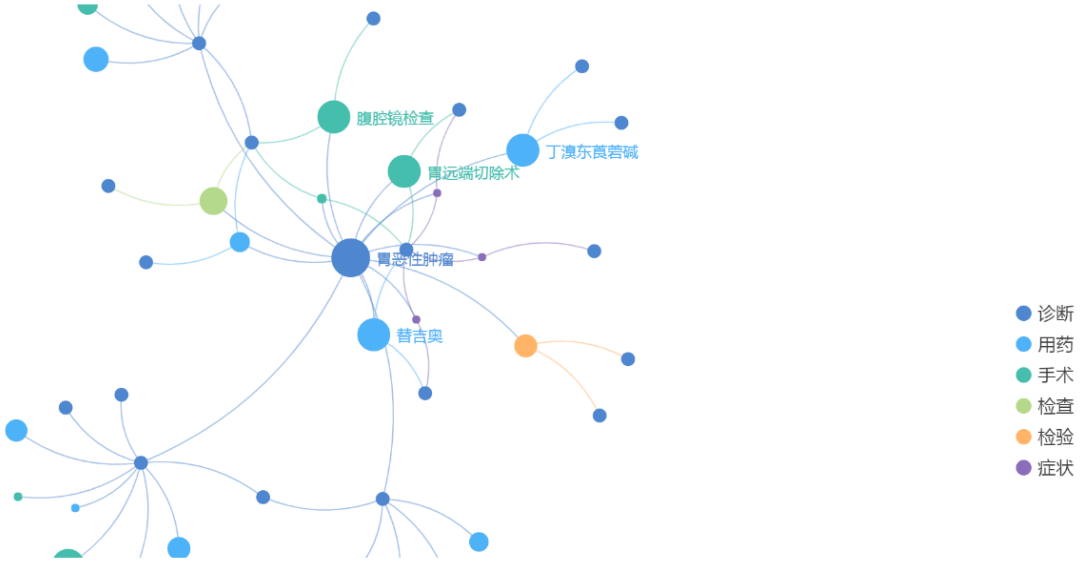
4.2.1 科研思路探索与发现

在整个科研流程中，对于医生来说，科研的思路探寻和科研场景的发现是打开科研的第一把钥匙。而在现有的无论是电子病历数据或者 CDR 的集成数据，要想挖掘科研场景探索科研思路特别的困难，“翻病历”往往成为了科研灵感发现的源头。除了上述章节所讨论的“大数据检索模式”之外，还能够通过数据挖掘和有监督以及无监督的机器学习，打开科研新模式，助力临床医生更加轻松高效的完成科研第一步。

4.2.1.1 疾病图谱

疾病图谱是大数据平台基于真实诊疗大数据，根据知识相关性展示的疾病可视化功能，展现与疾病主题关键词强相关的诊疗关键词，以及各个诊疗关键词互相关联的多层级关系网络。

构建图谱的方法，除了利用真实病历数据进行结构化、标准化处理从而形成疾病真实诊疗画像外，还需要充分参考疾病相关的循证医学知识，利用知识库进一步参与构建图谱，才能使得数据的分析和展示权威有效。



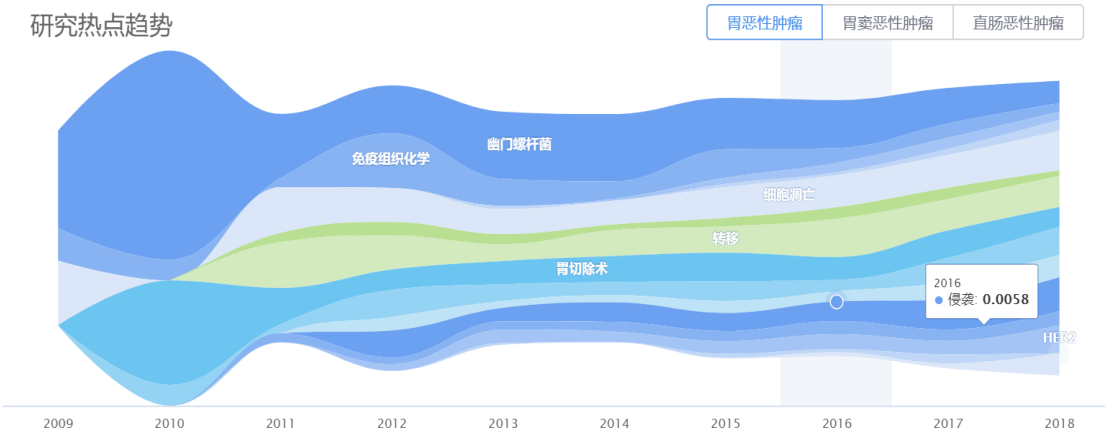
构建疾病图谱的主要目的是，基于大数据挖掘及数据可视化技术，帮助临床医生更好地从既往真实病历数据中发现临床价值和科研价值。目前常见展示的是一些基础疾病指标，譬如一些常规的诊疗行为的关键词数据，未来可以基于更智能的算法挖掘更多疾病指标，以及相关指标数据的智能相关性分析，实现临床数据价值的自动挖掘。

医学知识图谱是实现智慧医疗的基石，有望带来更优质的医疗服务。然而，现有知识图谱构建技术在医学领域中普遍存在效率低、限制多、拓展性差等问题。医疗大数据应用平台针对医疗数据跨语种、专业性强、结构复杂等特点，从医学知识获取、抽取、表达、融合和应用五个方面进行知识图谱自下而上的构建，从而能够在信息检索、知识问答、智能诊断等多个场景有深层次的应用。

4.2.1.2研究热点趋势

挖掘热点研究领域信息，是一个繁琐且耗费时间精力的过程。基于大数据处理的热点趋势展现能改善科研现状。根据 PUBMED 等文献数据库中的研究信息，智能提取 TOP 10 关键词进行计算，并绘制成一个研究热点趋势图。可以直观展示与疾病相关的研究热点以及热度变化趋势。在研究热点趋势中可集合具体的文献信息，可以查看文献描述、关键词等信息，如对具体文献内容感兴趣或有需求，可以查看或下载原文内容。这种基于大数据的研究热点汇总，可以为用户提供热点走势了解专业领域发展情况，为科研提供思路启发，便于医生快速精准地找到

科研入手点。



4.2.2 基于时间模型的科研分析

在实际科研过程中，对于采集后的数据处理对于医生来说是特别耗时耗力的事情，从最原始的字段数据到能够进入 SAS、SPSS 等统计软件进行分析，有漫长的数据中间环节需要处理。而大数据应用平台，可以基于患者重点事件建立时序模型，依托该模型进行中间数据处理过程，并且能够对采集的数据、处理的过程进行全面保障，排除了传统处理数据中的人为错误等因数。

另外，基于大数据平台的数据处理方式，会带来整体科研节奏的变化。以前需要数月才能进行数据处理和得到的数据分析结论，通过大数据处理方式，可以在小时级别内得到相应的结果，能对科研本身产生积极的影响。

以下将着重介绍基于时间模型的数据处理方案和示例：该功能主要是以医疗事件为基底，在此事件的上设计相对时间的条件进行过滤，来满足数据的阶段或者特定的数据需求的功能。其中“医疗事件”包括了 就诊、诊断、手术、用药、检验、检查、病理报告、医嘱等重要的医疗事件，而对此事件建立的时间阶段即为“相对时间”；比如 “手术”前 10 天到“手术”后 10 天之间的“白细胞计数”的值。

在此基础上，可以延伸添加到以下场景：

（1）“第一次”手术， “手术名称” 包含 “全胃切除术” 的术前 10 天到术后 10 天的， “白细胞计数” 的 “最大值”。

（2)对于部分更加复杂的需求，需要用到两个相对时间进行嵌套才能完成，具体如下需求“：

“第一次”手术，“手术名称”包含“全胃切除术”的术后 3 个月内，在这段时间内首次用药“注射头孢孟多酯钠”后，当次就诊中“丙氨酸氨基转移酶”的变化情况，此功能就需要用到两个医疗事件的相对时间嵌套完成。

基于以上时间模型的条件设置，能够将患者诸多原始数据通过需求转化成真正需要的可导入到 SPSS 等统计软件的数据，此格式数据能够保证每一个患者只有一行数据，满足任何方式的数据统计和分析。

4.2.3 专科疾病数据库

搭建专科疾病数据库，一直是科室、医院乃至国家层面重要的需求。但是，目前患者的病案或者各个业务系统的信息，在绝大部分医院都以非结构化或半结构化的方式进行存储，比如手术麻醉信息、手术过程信息、会诊讨论信息等。医生难以直接利用，在需要使用时只能采用人工誊写或摘抄的方式整理数据，这种高人工低智能的方式极大地增加了科研数据采集的时间成本。

科研数据根据课题研究目的不同，研究方向的迥异，对于数据本身的需求也会有不同的种类，而不同的数据又来源于多个业务系统。根据治疗方式和阶段的不同，数据的类别也具有多样性。比如患者的特殊检查检验数据，如患者的随访、CRF 表单以及生物样本、组学分析等数据。多维度、多系统的状况，造成医生在实际数据收集和整理过程中，需要来回在多个系统间切换，并选择合适的方法来留存数据，将多系统来源的数据人工进行关联核对，因此成本显著提升。对于从科研设计阶段、数据收集阶段、历史科研成果的延续以及跨科室、跨医院的专业疾病研究，均要形成整体的平台平台。

基于上述困境和对专业疾病数据库的需求，在大数据科研平台的基础上，搭建专业疾病数据库，成为较为重要的方案和手段。这也是近年来，以大数据为基础的应用和新热点与趋势。

图 4.6

4.3 管理应用场景

4.3.1 医院精细化管理

4.3.1.1 医院管理的实际问题与难题

医院管理是一个多学科、多部门、多方法管理学科。虽然医院针对现状采取不同的措施、不同的方案进行管理，但有一点是相同的，医院需要通过数据来生产出可以预警的医院运行指标及报告。目前医院管理仍需沉淀指标及逻辑规则，也面临如下诸多实际问题：

数据统计口径不统一

各个科室统计口径方式及上报给管理者统计口径不统一；

多业务系统分类方式不能有效集中；

无联盟医院质量对比；

无有效手段对重点学科建设及病种发展；

数据质量问题；

医疗整体质控方案未完善；

未建立病人统一索引；

主数据管理未统一；

病历数据未能有效利用；

没有数据定制化集中显示；

异常变异原因无法深究。

除了在以上实际问题，医院管理还面临如下诸多新环境下的挑战：

医疗资源配置系统调整	■ 2017年1月9日，国务院发布“十三五”深化医药卫生体制改革规划，国务院“十三五”期间医改重点任务：建立科学合理的分级诊疗制度。形成基层首诊、双向转诊、急慢分治、上下联动的就医新秩序。
医保控费科学化精细化发展	■ 2017年6月28日，国务院办公厅印发《关于进一步深化基本医疗保险支付方式改革的指导意见》全面推行以按病种付费为主的多元复合式医保支付方式，意味着医保基金科学与精细化的控费将逐步推进。2018年广州市作为按DRG付费的试点城市开展DRGs医保控费。
医疗服务价格结构改革	■ 2017年7月15日，继北京全面启动医改后，广州地区公立医院综合改革工作启动，全面取消药品加成，执行新的医疗服务价格政策，全面推进公立医院管理体制、补偿机制、价格机制、人事编制、收入分配、行业监管等方面的综合改革。
管办分离，提升自主性灵活性	■ 2017年7月25日，国务院办公厅发布了《国务院办公厅关于建立现代医院管理制度的指导意见》，（以下简称《意见》），对现代医院管理制度进行了一系列规定和调整，深化“放管服”，大大加强了公立医院管理的自主性和灵活性。

图 4.7

4.3.1.2大数据模式下的医院管理

基于大数据平台的医院管理，可通过科学的分组方法、多维度透视分析、行业基线对比、诊疗一体化的 ICD 对接能力以及 DRGs 指标，评估全院效率能力，了解全院科室能力分布，聚焦重点科室，通过科室效能分析，确定科室整体情况与需要关注的主诊组。围绕 DRGs 分组，评估疾病分组的治疗效率，了解不同主诊组在同一分组上的治疗表现差异。以下以全院科室的能力分布，作为举例示意：

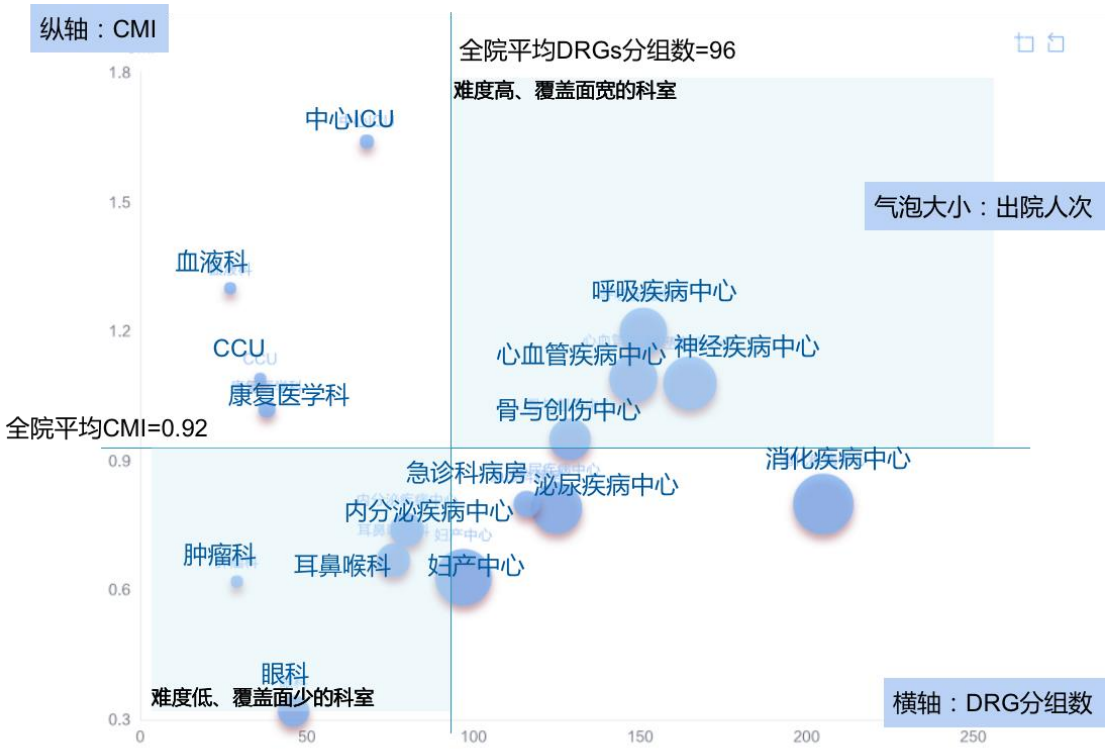


图 4.8

4.3.2 大数据病案管理

我国已经开始逐步推行 DRG 系统，进行医院的 DRG 绩效评价，这也是患者按照诊断分类定额支付的基础。ICD 编码是疾病分类的核心，编码的准确性、主要诊断选择的准确性、诊断顺序的准确性均十分重要。原国家卫计委发布的《三级综合医院评审标准（2011 年版）》明确规定：ICD10 与 ICD9-CM-3 作为病案管理的正规分类体系及基础配置，并将人员资质纳入评审。

较为传统的病案编码步骤是：病案室专业编码员对照病程病历内容，对照各地政府部门发布的 ICD 诊断目录进行文字筛选查找和理解，将临床诊断术语与 ICD 编码进行关联。这一过程普遍存在如下主要的问题：

图 4.9

通过大数据平台，可以引入 ICD 编码的规则。通过大数据平台中既往病历编码的累积和沉淀，将编码数据或者历史数据作为正确的样本输入，通过机器学习方式，结合大数据推荐和智能分组应用，逐步解决医院在病历编码时遇到的困难。

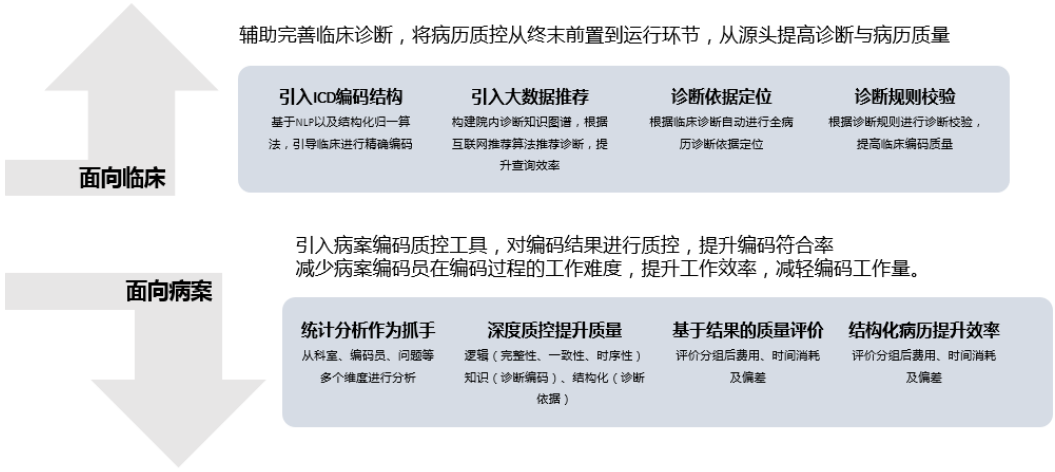


图 4.10

4.3.3 病历评分体系

病历评分是针对病历质量管理的重要手段。原始的评分手段，通过对医生完

成的病历按照《卫生部病历质量评价标准》进行人工评分，不仅工程浩大，而且人为因素过多，导致该功能在大部分医院无法真正使用。

而依托大数据平台，可以将评分的标准拆解成具体的规制。通过深度学习的方法，学习病历评分的专业人员的病历和评分数据，将评分功能从低效的人工方法变成智能自动评分。并且，可以对医院的数据质量、数据完整性丰富性、数据和疾病的专业度相关性进行评定，搭建完整的评分体系。

被评分的病历在临床、科研和病历管理等许多场景，均具有深入的应用空间。比如，通过评分来规范医生的病历编写过程，并且科研将病历的质量作为考核的一部分，以此来提高医院的整体病历质量。这使得在医生进行科研时，系统能够按照病历评分的质量维度进行筛选，排除低质量的病历，加速整个科研的进程。

如下表所示，该表是按照《卫生部病历质量评价标准》进行规制拆解后，作为规则数据的示例：

序号	项目名称	描述
1	不规范的内容	夹杂特殊符号：空格，换行（影像学表现包含\r\n），xml tag，乱码，"
		姓名中包含数字
		标准字段的格式不合逻辑：身份证号码长度，邮编长度
		相应内容理应归一化，但医生并没有执行，如：“无”，“-”，“没有”，“--”等
2	默认数据值的问题	使用默认日期，如：1970-01-01 00:00:00 或 1000-00-00 00:00:00 等
3	医生输入错误或特殊含义	体温（376℃） 心率（800次） 身高（1.75cm） 血压（11787）
		时间输入错误造成的各种不合逻辑
		检查方法、检查类别与检查结论的不匹配
		存在字典规范输入的字段，医生输入文本而不是字典
		字段输入不规范：手术护士未输入护士姓名，而写为“手术护士”
		血压：未知 检查结果：已做
		生命体征不包含单位
4	医疗系统设计问题或医生执行错误	检查部位和检查方法写在一起
		多个诊断写在同一个字段
		检查结果和检查所见写在一起
		检验套餐名称和检验标本名称用同一个字段
		用药频次、用药方法、给药途径、单次剂量等字段中的两个或多个组合写到一起
		月经史和婚育史放在一起叫月经婚育史/月经生育史
		与陈述者关系却记录患者姓名
		缺少医嘱类别
		单一表存在生产和测试数据
		电子病历状态不符合预期
		日常病程中，医生输入上级查房记录
		系统缺少对默认标签含义的理解：离异，育有1子0女 -> 离异10

图 4.11

4.3.4 VTE 风险评估

静脉血栓栓塞症（VTE）是继缺血性心脏病和卒中之后，位列第三的最常见心血管疾病。全球范围内每年的 VTE 例数可达近 1000 万例，且多达 60%的 VTE 事件发生在住院期间或出院后，已成为主要的可预防的院内死亡原因。我国 VTE 患病率在 ICU 患者中为 27%，脑卒中患者中为 12.4%-21.7%，骨科甚至达到 40%，

构成了重大的公共卫生问题。

VTE 院内防治是目前有效防治 VTE 的一种手段。目前我国的 VTE 预防做得不好，原因在于随着患者人群的增多，病人数量越来越庞大，患者风险也越来越大，需要临床给予关注和预防。传统的 VTE 防治流程是临床大夫对患者进行国际通用的 Khorana、Ay、Caprini 评分，筛选出高危人群，进行物理预防或者药物预防措施，以降低静脉血栓的发生。但 VTE 评分需要评估项目较多，操作不便捷，需要占用大夫大量的时间，更存在人为偏差，而且治疗方案难以全流程追踪预后情况。

依托医疗大数据平台，使用后结构化和分词技术，解析相关患者诊疗数据，按照 Caprini 评分规则，可实现系统的自动评分，并可嵌入临床 HIS、EMR、ICU 重症监护等系统，适时为临床大夫推送高危患者提示，并推送治疗方案。减轻临床负担，降低 VTE 发生率。

4.3.5 ICD 辅助编码

我国医院的 ICD 诊断编码工作，主要由病案编码人员参考医生书写的诊断描述来完成。不仅耗费时间，而且滞后、易出错。而 ICD 编码的正确性影响着科室治疗评价、医保付费等。随着医院电子病历系统的应用深入，一些医院将 ICD 编码工作推至前台，希望交给临床医生进行书写选择，但临床大夫书写的更多是临床诊断，而不是 ICD 诊断，实际临床应用效果并不理想。随着国家卫健委 ICD11 的下发，如何快速落实疾病诊断编码成为当前技术前提下需要克服的瓶颈。

基于大数据技术（机器学习和历史编码特征归类等）的 ICD 辅助编码，预期可以应在临床。计算机基于后台的学习模型，在医生书写病例的过程中智能学习，适时推送可参考的疾病编码，医生的选择和病例编目后的诊断校准，可以提高推送的准确性，进一步做好医生助手。

4.4 患者服务场景

医疗大数据平台除了为医院医、教、研、管提供数据和应用支持外，在患者

服务方面，目前越来越多的大数据公司也在探索研究。

4.4.1 智能导诊

门诊初诊患者自行将主诉与科室诊疗范围准确对应的难度较大，且不少疾病可能需要经历跨专业、跨科室的诊治过程，患者对疾病诊治的认知，很多与医生的专业设计相左。患者挂错号、退换号行为以及相关纠纷屡见不鲜，不仅耽误了患者就诊时间，也浪费了宝贵的医疗资源。因此，对患者进行及时科学的导诊分诊十分必要。

4.4.1.1 传统导诊服务

传统的导诊服务主要有两种：

1.就医目录指南。就医目录指南是最普遍的导诊服务形式。该类服务通过设置疾病目录／症状目录，为患者提供疾病信息解读和就医科室、医学专家推荐服务。目前各医疗机构的官方网站和专业的第三方医疗卫生信息服务网站都提供了这项服务，并与在线挂号系统集成。就医目录指南在一定程度上实现导诊功能，但是该服务起效的前提是患者需要明确所患有的疾病，而该前提是不合逻辑也不现实的，尤其对于初次患病的患者。

2.人工在线咨询。人工在线咨询也是常见的一种导医服务形式。该服务通过在线论坛、聊天工具、电话交流等形式接受患者咨询，并向患者提供就诊建议。该类服务的人性化、专业性良好，但服务质量受咨询服务人员专业水平影响较大，尤其在监管不力时，会出现欺诈风险。很多就医人工咨询服务普遍存在诱导就医的问题，以 2016 年百度贴吧事件最为典型。目前该类服务是政府重点关注和治理对象。除此以外，人工咨询服务高额的运营成本也降低了服务的持续性、可行性。

3.传统线下导诊。传统线下导诊对导诊人员的专业知识要求较高，而且难以及时满足线上挂号的导诊需求。简单的科室和专业说明也无法完全解决患者对挂号科室的疑虑。在实施初诊全预约后，患者错挂号后退换号的情况更为常见，造成诊疗效率降低、患者满意度下降、医院运营成本增加等不良影响。

4.4.1.2 大数据智能导诊

智能导诊服务是一种高科技服务形式，主要采用人工智能技术识别患者可能患有的疾病并给出就医指导。通过智能终端（APP、微信等）与医院信息系统连接，结合大数据平台疾病知识图谱、数据挖掘方面的导诊算法、语义分析等工具，让患者在线完成导诊分诊服务。通过建立相关专科疾病、症状、治疗方式知识库，利用问题生成器和病历分类器，建立知识库内容与提问内容、推荐科室之间的联系，实现分诊科室的智能推荐。通过对医学教科书和患者病例的学习，匹配最佳就诊科室；实现了通俗语言高度识别，提高易用性；减少了患者手动输入内容，给患者流畅的产品体验；进行分诊结果校验，不断提高结果准确度。

该类服务的显著优点是效率较高，不受人为因素的影响。缺点是受技术水平和数据资料质量影响很大。比如对患者自然主诉的理解水平就直接影响了后续的疾病判断，推理算法的准确率和效率也影响了咨询质量。大数据处理技术的发展为该类服务的升级与普及提供了巨大支持，是目前互联网医疗服务领域的热点之一，基于大数据技术的智能导诊服务有望成为今后主流的就医咨询服务场景。

4.4.2 智能候诊

4.4.2.1 传统候诊

目前，我国的就诊方案依旧采用排队叫号进行患者的排队分流，无论是分诊就医还是付款取药等环节，都需要患者进行排队。现有的病患排队叫号系统，主要通过发挥独立的排队管理功能来进行医疗信息的整合。单纯通过排队的功能，难以实现有序且高效的 医疗秩序管理。一旦到就诊高峰期，患者往往为了缩短就诊时间，患者或患者的家属不停在就诊室和候诊区之间走来走去，这样把本来就纷乱的医院环境变得更加纷乱，不仅影响了医生的工作，也给医院的导诊护士增加了工作量，给患者造成了不必要的麻烦。

4.4.2.2 智能候诊

利用医疗大数据平台实时获取门诊预约系统、排队叫号系统、分诊导诊系统、

HIS 系统等数据，通过数据模型精准评估对门诊患者候诊时间的量化统计和分析，发现影响候诊时间的关键因素以优化门诊就诊流程，优化每位患者等候时间。通过与医院多种方式对接，如刷卡挂号,预约挂号、刷卡入队、排队叫号等，并推送到线上预约系统及相关信息系统，实现移动端和线下显示屏同步，让患者实时掌握自己的就诊时间，避免了患者长时间的排队等候，合理有效的引导患者来院就医的时间，改善患者的就医感受，从而大大提高了患者就医满意度。

4.5 药物研究场景

中国现行的药品临床试验管理规范（GCP）来源于欧、美、日共同发起的国际标准 ICH-GCP，在基本原则和大多数的实施细则上都一致。2017 年中共中央办公厅、国务院办公厅印发了《关于深化审评审批制度改革鼓励药品医疗器械创新的意见》，并发出通知，要求各地区各部门结合实际认真贯彻落实。这份《意见》特别指出，临床试验机构资格认定实行备案管理，大力支持支持临床试验机构和人员开展临床试验，与此同时，须对临床试验数据可靠性和真实性进行更大程度的监管。

最新公布的《药物临床试验质量管理规范》中第一章总则第一条提出：为保证药物临床试验过程规范，数据和所报告结果的科学、真实、可靠，保护受试者的权益和安全，根据《中华人民共和国药品管理法》、《中华人民共和国药品管理法实施条例》，参照国际公认原则，制定了本规范。规范中针对临床试验的源数据管理、必备文件管理、质量管理、CRF、药物管理、AE/SAE、QA/QC、计算机系统（数据管理系统）、电子数据稽查轨迹等均提出明确的要求。

在新的时代背景下，临床试验的质量管理直接关系到药品研发进程及疾病防治策略，如何高质高效的开展临床试验及临床研究，促进学科和疾病领域的进展成为新的课题和攻坚方向。依托医疗大数据平台和《药物临床试验质量管理规范》的各项规定，参考专业的标准数据集和中心化统一编码—CDSIC，探索如下应用场景：

4.5.1 受试者智能招募

传统招募方式：医院或社区张贴招募广告；发放宣传单、宣传册、便捷联络卡、教育资料；社区义诊(适用于罕见病或入排标准较为复杂的试验)；医师推荐；宣传会；运用网络媒体，如电子邮件、受试者 QQ 群、微信群、定期群发招募信息、微信公众号推送等；以及研究者自行招募的方式。但是，这些方式无法确保招募宣传送达最合适的潜在患者人群，而且由于患者已经离院，需要打电话召回患者，消耗大量人力成本。

依托医疗大数据平台，基于项目的纳入排除条件，实时在临床场景发现疑似符合入组条件的受试者，通过医生工作站提醒医生，加速临床试验入组。

4.5.2 RBM 质控核查

临床药物试验中，以风险为基础的监测(Risk-Based Monitoring ，RBM)至关重要。传统质控方式：传统的质控流程主要依赖机构质控员、CRA、第三方稽查公司在院内，基于 SD 进行数据评估和数据抽检。有如下问题：人员专业水平参差不齐：目前国内 SMO、CRO 机构的专业化管理水平和执业诚信有待提高。质控员专业水平参差不齐，即使单一临床试验也经常出现质控跟踪有始无终的现象，基本上把临床试验的安全监测当作与 PI 的常规交流和工作拜访。

非全量核查：受限于临床试验受试者数据量与质控人员时间比例，无法做到全量核查，只能进行抽检再基于发现的问题重点排查，有遗漏问题的风险。

时效性差：目前核查频率基于试验进程按阶段进行，只能后置发现问题进行整改，无法做到过程中监控，实时报警。

依托医疗大数据平台，历史项目核查报告和质控规则，智能预测临床试验可能的风险，并利用大数据技术对 SD 进行质量评估，实时发现方案不依从、数据不一致、AE/SAE 漏报等质量问题，提醒研究团队进行整改。同时在 CRA、质控员、稽查人员现场质控前，基于数据模型运算，优先产出质控分析报告，告知质控人员风险点，方便质控人员做有针对性核查。

4.5.3 AE/SAE 自动报警

临床试验中的安全性通报，在很大程度上是基于国家药政的要求。这些临床症状看似与临床试验或试验用药无关联，或这些症状在该疾病中本身就比较常见，因此研究者有可能认为其不是不良事件，极易造成漏报。

依托医疗大数据平台和 CTCAE 规则，以及结构化归一等数据处理技术，系统能够自动发现疑似 AE/SAE，提示医生进行判读，方便医生进行 AE/SAE 生命周期管理，避免漏报。

4.5.4 试验数据辅助采集

临床研究人员在参与科研的同时，有大量工作要投入到临床诊疗过程中去，将研究所需数据归纳整理及填写如 CRF 中是研究必要且十分重要的环节。在精力有限、人员不足的情况下，容易出现差错，从而影响研究进度与质量。依托医疗大数据平台，进行方案规则拆解，遵循 CDSIC 标准，可以将医生的 SD 数据，通过算法模型计算，自动导入到 EDC 完成填表，减少 CRC 工作量的同时，也避免了人工录入错误。

4.6 教学应用场景

4.6.1 基于真实世界数据的疾病图谱

基于大数据的疾病图谱是临床医疗教学除了教科书、文献之外的第三个工具。对院内某一疾病的海量真实数据进行统计分析，可呈现真实的疾病特征分布，包括年龄分布、常见症状、性别比例、常用检查检验方式等。隐藏在大数据各个节点数据之间逻辑关系的透出，帮助深度解读疾病信息。做到理论和临床实际案例相结合，辅助临床教学。

4.6.2 临床数据与知识库关联应用

可以根据用户的特征信息进行智能知识推荐，能够学习用户对于推荐内容的喜好程度进行深度学习，将用户更加需要的知识推荐给用户，例如由大数据智能

技术对文献分析产生的文献热点趋势图、文献热点关键词、对应的文献作者图谱等。支持中英文文献检索，提供多种文献检索方式，进行个性化的文献推荐，有效提高文献检索效率。此外，还可提供临床指南、药品说明书、临床指南、临床试验查看，多种知识类型集中呈现，提供一站式知识查询体验。辅助青年医生学习成长，提升临床和科研能力。

4.7 应用展望

尽管大数据技术和医疗大数据平台在医疗行业还是新生事物，但是从国家和行业政策支持，以及自身应用发展角度讲，必然会在医疗行业得到更加深入的应用，提高临床效率，辅助诊疗决策，便捷患者就医，科学管理决策。如下是未来几年可期的应用场景：

1、 大数据平台将和医院临床数据中心相互融合，功能互补；

2、 大数据平台将在临床决策支持方面获得临床突破，产生更多适用于临床诊疗的知识产品；

3、 大数据平台将一步推动医疗机构科研数量和水平的提升；

4、 大数据平台将推动国家药物临床试验提速，提质，逐渐和发达国家并轨，及时引入新药，惠及民生；

5、 大数据平台将和国家开放共享政策相结合，推出更加患者参与和使用的医疗服务应用。

6、 大数据平台将成为医院，特别是研究型医院必选的数据决策信息平台。

7、 大数据平台将成为未来新型电子病历的智能引擎，实现临床科研一体化电子病历的落地；

8、 区域医疗大数据平台（专科、专病中心）的建设成为连接各级医疗机构的数据平台；

9、 医联体核心医院大数据平台建设，将推动落实医联体间分级诊疗，并促进临床科研联合体的有效建立。

10、 基于大数据的应用将趋于移动化，让院领导、临床、科研、患者更可及，更便利。

附录 A-F 征求意见稿中省略，征求意见后指南全文将在 7 月厦门 CHIMA 大会正式发布。

反馈意见可联系：

CHIMA 秘书处 杨永燕

电话：010-84279271

邮箱：yangyy@chima.org.cn

北京大学肿瘤医院信息部 王立军

电话：010-88196039

邮箱：www_539@163.com