



第 1 章

引论——DAMA 数据治理知识体系介绍

1.1 DMBOK 简介

1.1.1 DAMA国际和《DAMA数据管理的知识体系指南》（DMBOK）的发展历程

DAMA 国际（Data Management Association International，国际数据管理协会）是一个全球性的数据管理和专业志愿人士组成的非营利性协会，致力于数据管理的研究和实践。DAMA 国际在全球范围内拥有 50 多个分会，20000 余名数据管理专业人士会员。其宗旨是促进理解、发展和实践数据和信息的管理，以支持业务战略。DAMA 基金会是 DAMA 国际的附属研究和教育机构，致力于发展数据管理专业，并推动数据和信息作为企业资产的管理理念和实践的进步。DAMA 国际和 DAMA 基金会（统称为 DAMA）的共同使命是引领数据管理职业走向成熟。DAMA 推动管理数据、信息和知识作为企业关键资产的理解、发展和实践，独立于任何特定的供应商、技术和方法。

DAMA 国际通过多种方式致力于推动数据管理专业的成熟，其中一些举措如下。

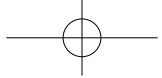
（1）与威尔希尔会议公司合作举办一年一度的 DAMA 国际研讨会（现为企业数据世界），这是世界上最大的专业数据管理会议。研讨会上的教程和会议环节为数据管理专业人士提供持续的教育。

（2）与英国信息权利管理协会（Information Rights Management UK，IRM UK）合作举办一年一度的 DAMA 国际欧洲会议，这是欧洲最大的专业数据管理会议。会议上的教程和会议环节为数据管理专业人士提供持续的教育。

（3）与计算专业人员认证协会（ICCP）合作，提供专业认证计划，授予认证数据管理专业人士（CDMP）称号。数据仓库研究所（TDWI）在认证商业智能专业人士（CBIP）计划中也使用 CDMP 认证考试。

（4）DAMA 国际教育委员会的数据管理课程框架为美国和加拿大的院校提供指导，帮助它们将数据管理作为北美高等教育模式中 IT（信息技术）和 MIS（管理信息系统）课程的一部分进行教学。

DAMA 国际自 1980 年成立以来，多年致力于数据管理的研究、实践及相关知



识体系的建设，在数据管理领域积累了极为深厚的知识沉淀和丰富经验，并先后出版了《DAMA 数据管理辞典》和 DMBOK，集业界数百位专家的经验于一体，是数据管理业界最佳实践的结晶，已成为从事数据管理工作的经典参考和指南，在全球范围内广受好评。

2009 年，DAMA 国际发布了 DMBOK1.0 版，将数据管理定义为规划、控制和提供数据资产，发挥数据资产的价值。它将数据管理划分为 10 个知识领域，分别是数据治理、数据体系管理、数据架构管理、数据开发、数据操作管理、数据安全、参考数据和主数据管理、数据仓库和商务智能管理、文件和内容管理、元数据管理和数据质量管理。其中，数据治理是高层次的、规划性的数据管理活动，其关键管理活动包括制定数据战略、完善数据政策、建立数据架构等，注重数据的使用者、使用方式、使用权限等合规性制定，强调开展数据资产全生命周期管理前的基础工作，关注数据资产管理中的相关保障措施。

《DAMA 数据管理辞典》是对 DMBOK 的有机补充，最初是作为对 DMBOK 的扩充词汇而编撰的，由于其规模和商业价值，DAMA 国际将其单独出版。该辞典中的术语定义都与它们在 DMBOK 中的用法一致。

2015 年，DAMA 国际发布 DMBOK2.0 版，将数据管理知识领域扩展为 11 个管理职能，分别是数据治理、数据架构、数据建模与设计、数据存储与操作、数据安全、数据集成与互操作、文件和内容管理、参考数据和主数据、数据仓库和商务智能、元数据、数据质量等。DMBOK2.0 版是 DAMA 国际专家对过去 30 多年数据管理领域知识和实践的总结，是一部综合了数据管理方方面面、具有权威性的基础工具书。

1.1.2 DMBOK的目标及用途

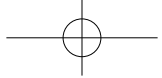
DAMA 国际组织出版的 DMBOK，是为了进一步推动数据管理专业的发展，并希望成为数据管理的权威入门读物。没有一本单独的书可以描述整个数据管理知识体系。DMBOK 并不是试图成为数据管理的百科全书，也不是涵盖所有与数据管理相关内容的完整论述。相反，DMBOK 简要介绍了数据管理的概念，并确定了数据管理的目标、职责和活动、主要交付成果、角色、原则、技术和组织 / 文化问题，描述了普遍接受的良好实践以及重要的替代方法。

作为权威入门读物，DMBOK 的目标如下。

(1) 提供通用语言：DMBOK 旨在为数据管理专业人员建立一套通用语言和术语。它定义了标准化的术语和概念，使从业人员能够有效沟通并就数据管理原则达成共同理解。

(2) 定义最佳实践：DMBOK 识别并定义了跨越各个领域和学科的数据管理最佳实践。它概述了不同数据管理领域应遵循的推荐方法、方法论和技术，为专业人员实现高效的数据管理成果提供指导。

(3) 促进一致性和标准化：DMBOK 促进数据管理实践的一致性和标准化。它



提供了一个结构化的框架，组织可以使用该框架评估其当前的数据管理能力、识别差距，并使其实践与行业标准和基准保持一致。

(4) 促进协作和知识共享：通过提供全面的知识体系，DMBOK 促进了数据管理专业人员之间的协作和知识共享。它作为一个参考指南，可以帮助从业人员提高技能、交流经验并学习行业领先的实践。

(5) 支持专业发展：DMBOK 支持数据管理人员的职业发展。它概述了数据管理领域内不同角色所需的知识、能力和技能。专业人员可以将 DMBOK 作为指导其职业发展和识别需要进一步学习和专业化的领域的路线图。

(6) 实现有效的数据治理：DMBOK 的主要目标之一是实现有效的数据治理。它提供了建立数据治理框架、定义数据政策和标准以及实施数据管理制度实践的指导。通过遵循 DMBOK，组织可以建立健全的数据治理实践，确保数据以一致、安全且符合业务目标的方式进行管理。

(7) 提高数据质量和完整性：DMBOK 强调数据质量和完整的重要性。它提供了评估、改进和维护数据在整个生命周期中的质量的方法和技术。通过实施 DMBOK 中概述的实践，组织可以提高其数据资产的准确性、完整性和可靠性。

(8) 促进数据驱动决策：DMBOK 的最终目标是促进数据驱动决策。通过推广有效的数据管理实践，组织可以利用其数据资产来获得有意义的洞察力、做出明智的决策并提高业务价值。DMBOK 作为一个指南，确保数据以支持组织目标和实现数据驱动决策的方式进行管理。

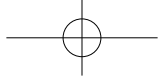
总的来说，DMBOK 的目标是建立标准、最佳实践和对数据管理原则的共同理解，以使组织能够有效地管理和利用其数据资产，实现战略和运营目标。

DMBOK 的受众包括以下几种。

- (1) 认证数据管理专业人士和有志于成为认证数据管理专业人士的人员。
- (2) 与数据管理专业人士合作的其他 IT 人士。
- (3) 所有类型的数据管家。
- (4) 对管理数据作为企业资产感兴趣的高管人员。
- (5) 认识数据作为企业资产的重要性的知识工作者。
- (6) 评估和帮助改善客户数据管理工作的顾问。
- (7) 负责开发和提供数据管理课程的教育工作者。
- (8) 数据管理领域的研究人员。

关于 DMBOK 的使用，DAMA 国际预计会有以下几种潜在用途。

- (1) 让不同受众了解数据管理的性质和重要性。
- (2) 帮助在数据管理社区内做好术语及其含义的标准化工作。
- (3) 帮助数据管家和数据管理专业人士理解他们的角色和职责。
- (4) 为评估数据管理的有效性和成熟度提供基础。
- (5) 指导实施和改进数据管理工作的举措。



- (6) 指引读者查阅更多关于数据管理的知识来源。
- (7) 指导高等教育数据管理课程内容的开发和交付。
- (8) 提出数据管理领域进一步研究的领域。
- (9) 帮助数据管理专业人士备考 CDMP 和 CBIP 考试。

1.1.3 DMBOK框架介绍

DMBOK 数据管理框架（也称为 DAMA 车轮图），即技术视角的 DAMA 车轮图，如图 1-1 所示。

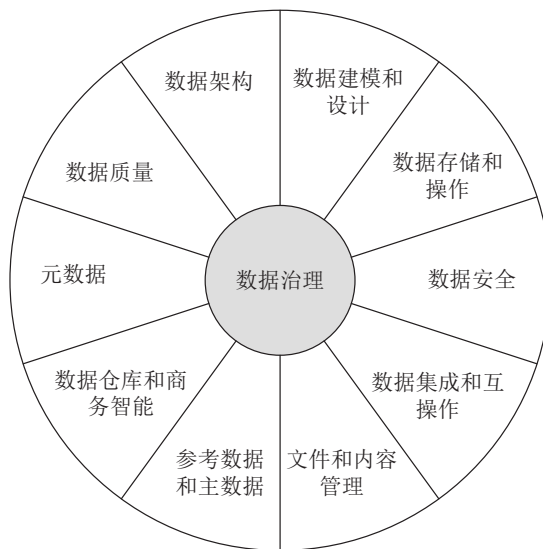


图 1-1 DAMA 车轮图

DAMA 车轮图定义了数据管理知识领域，它将数据治理放在数据管理活动的中心，因为治理是实现数据管理工作内部一致性与维持工作之间平衡所必需的。其他知识领域（数架构、数据建模等）围绕车轮平衡。它们都是成熟数据管理工作的必要组成部分，但根据各组织的需求，它们可能在不同的时间实现。

DAMA 数据管理框架也被描述为另一种形式的 DAMA 车轮图，即业务视角的 DAMA 车轮图，数据治理范围内的应用活动围绕着数据管理生命周期内的各项核心活动进行，如图 1-2 所示。

核心活动位于框架中心，包括元数据管理、数据质量管理和数据架构。生命周期管理活动可以从多个方面定义，如计划的角度（风险管理、建模、数据设计、参考数据管理），实现的角度（数据仓库、主数据管理、数据存储和操作、数据集成和互操作、数据开发技术）。生命周期管理活动源于数据的使用：主数据使用、文件和内容管理、商务智能、数据科学、预测分析、数据可视化。许多情况下都会基于现有数据进行增强性的开发，获取更多洞察，产生更多的数据和信息。数据货币化的机会可以确定源于数据的使用。数据治理活动通过战略、原则、制度和数据管理，

提供监督和控制。它们通过数据分类和数据估值实现一致性。

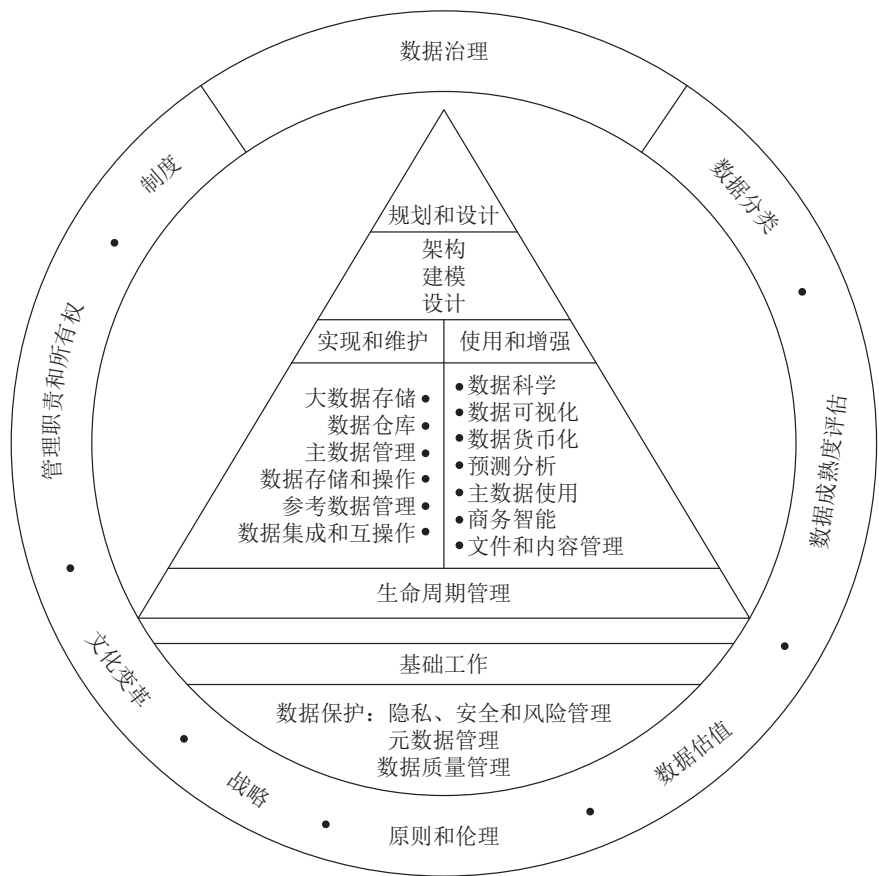
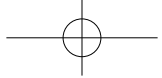


图 1-2 另一种形式的 DAMA 车轮图

DAMA 数据管理框架采用多种方式描述的原因是期望提供额外的视角，并就如何应用 DMBOK 中呈现的概念展开讨论。随着数据管理重要性的提高，这些框架成为数据管理社区内以及数据管理社区和利益相关方之间沟通的有力工具。

DMBOK2.0 是围绕 DMBOK 数据管理框架的 11 个知识领域构建的。各个知识领域章节遵循相同的结构。

- 引言
- 业务驱动因素
- 目标和原则
- 基本概念
- 活动
- 工具
- 方法
- 实施指南



- 与数据治理的关系

- 衡量指标

数据管理知识领域描述了各种数据管理活动集合的涵盖范围和工作背景，并阐释了数据管理的基本目标和原则。由于数据在组织内呈横向流动，这些知识领域涵盖的活动之间以及与其他组织职能之间难免会存在交叉和相互影响。

(1) 数据治理通过建立一个能够满足企业需求的数据决策体系，为数据管理提供指导和监督。

(2) 数据架构定义了管理数据资产的蓝图，通过与组织战略保持一致，建立战略性数据需求和设计以满足这些需求。

(3) 数据建模和设计是以一种称为数据模型的精确形式发现、分析、表示和沟通数据需求的过程。

(4) 数据存储和操作包括数据的设计、实施和支持存储以实现其价值最大化。操作在整个数据全生命周期内提供支持，从数据规划到数据处置。

(5) 数据安全确保维护数据隐私和机密性，防止数据泄露并确保适当地访问数据。

(6) 数据集成和互操作包括与数据存储、应用程序和组织内外的数据移动和整合相关的流程。

(7) 文件和内容管理包括用于管理各种非结构化媒体（尤其是支持法律和法规遵从性要求所需的文件）中的数据和信息的生命周期的规划、实施和控制活动。

(8) 参考数据和主数据包括持续核对和维护核心关键共享数据，以便在整个系统中一致使用最准确、及时和相关的核心业务实体事实版本。

(9) 数据仓库和商务智能包括计划、实施和控制过程，以管理决策支持数据，并使知识工作者通过分析和报告从数据中获取价值。

(10) 元数据包括计划、实施和控制活动，以实现高质量、集成的元数据的访问，包括定义、模型、数据流和其他关键信息，这些信息对于理解数据及其创建、维护和访问的系统至关重要。

(11) 数据质量包括规划和实施质量管理技术，以衡量、评估和改进数据在组织内的适用性。

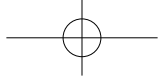
除了有关知识领域的章节，DMBOK2.0 还包含以下主题的章节。

(1) 数据处理伦理描述了关于数据及其应用过程中，数据伦理规范在促进信息透明、社会责任决策中的核心作用。数据采集、分析和使用过程中的伦理意识对所有数据管理专业人员有指导作用。

(2) 大数据和数据科学描述了随着收集和分析大型、多样化数据集的能力的提高而涌现的技术和业务流程。

(3) 数据管理成熟度评估概述了评估和改进组织数据管理能力的方法。

(4) 数据管理组织和角色期望为组织数据管理团队和实现成功的数据管理实践提供了最佳实践和注意事项。



(5) 数据管理和组织变革管理描述了如何计划和成功地实施将有效的数据管理实践嵌入组织所需的文化变革。

一家特定的组织如何管理它的数据取决于它的目标、规模、资源、复杂性以及对数据如何支持总体战略的认识程度，但大多数企业并不会实施每个知识领域中描述的所有活动。然而，了解数据管理更广泛的背景将使组织能够更好地做出在改善这些相关职能内和跨职能之间的实践时应该关注哪些方面的决策。

1.2 数据治理概述

1.2.1 数据治理概念

DMBOK 对数据治理的定义是：指对数据资产管理工作履行职权和实施管控（如规划、实施、监督和执行等）的行为。DAMA 认为数据治理是建立在数据管理基础上的一种高阶管理活动，是各类数据管理工作的核心，指导所有其他数据管理工作的执行。

我国《信息技术服务 治理 第5部分：数据治理规范》（GB/T 34960.5）对数据治理的定义是：数据资源及其应用过程中相关管控活动、绩效和风险管理的集合。

中国银行保险监督管理委员会在2018年下发的《银行业金融机构数据治理指引》（银保监发〔2018〕22号），对数据治理的定义是：指银行业金融机构通过建立组织架构，明确董事会、监事会、高级管理层及内设部门等职责要求，制定和实施系统化的制度、流程和方法，确保数据统一管理、高效运行，并在经营管理中充分发挥价值的动态过程。

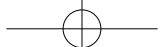
随着数据要素产业的蓬勃发展及数据要素在各行各业的应用落地，更多的主体意识到数据治理的重要性并参与其中，根据各行业、各业务场景的特点，数据治理被赋予了不同的含义和作用。考虑到数据治理的发展趋势，为更全面地体现数据治理的内涵和知识体系，本书采纳《信息技术服务 治理 第5部分：数据治理规范》（GB/T 34960.5）中关于数据治理的概念，与 DMBOK 中数据治理的概念相比，它是广义的数据治理，包含了 DAMA 车轮图中数据管理的内容。

1.2.2 数据治理原则

DMBOK 将数据治理原则划分为领导力和战略、业务驱动、共担责任、多层面、基于框架和原则导向，这些原则在2021年中国通信标准化协会发布的《数据治理标准化白皮书》（2021年）中被借鉴并予以完善，完善后的数据治理原则被概括为以下内容。

(1) **战略重视、组织保障**：规划数据治理中长期路线图、明确职责分工、建立数据治理组织架构，监督各项任务执行情况、解决组织间矛盾及冲突、及时调整规划内容。

(2) **责任共担、协调配合**：明确各部门的职责及任务，制定工作原则，明确各自任务及边界，建立配合机制，共同确保数据治理整体任务的实现和目标的达成。



(3) **业务驱动、问题导向**: 基于业务活动中发现的数据不标准、不一致、不准确、不可信、用数困难等问题,通过业务驱动开展计划、控制、开发、运营等数据治理活动,并通过数据治理考核机制来监督落实。

(4) **流程嵌入、实用落地**: 数据治理是管理、业务、技术三位一体的系统工程,将数据治理的活动、工具、输入输出物、人员角色等嵌入管理、业务、技术的关键流程中,并达成用户体验好、自动化程度高、简单适用的成效。

(5) **服务导向、量化评价**: 以服务为核心理念,为数据应用提供可用、可信的高质量数据,满足数据需求、赋能业务发展。设置量化指标评价数据治理的工作成效,反映数据治理存在的成绩和不足,提出针对性的改进和优化措施。

健康医疗领域的的数据治理也可以借鉴上述原则,将其作为参考性原则。

1.2.3 数据治理参考框架

数据治理国家标准《信息技术服务 治理 第5部分:数据治理规范》(GB/T 34960.5),将数据治理框架划分为顶层设计、数据治理环境、数据治理域和数据治理过程四大部分,如图1-3所示。

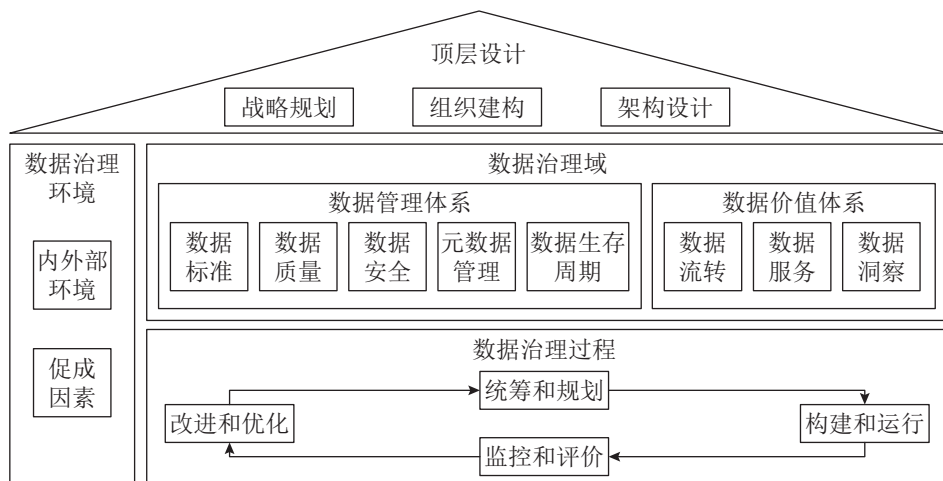


图 1-3 数据治理框架

数据治理相关国家标准、行业标准为相关行业数据治理工作提供了参考指引,推动了数据治理事业的发展,但又不像 DMBOK 那样能够提供一套完整的知识体系指导数据治理工作的开展,不利于数据治理工作的具体实施。然而,由于 DMBOK 是一个通用性的数据管理知识体系指南,未能够体现行业特点及与特定行业的结合应用,所以在特定行业应用上尚不能得到充分应用。为此,DAMA 中国(上海市静安区国际数据管理协会)于 2021 年决定基于 DMBOK 组织编写一套数据治理丛书,选取医疗健康、金融、制造业、交通、零售等行业作为试点,计划建立起与行业特点相结合的行业性数据治理知识体系指南,以期为特定行业的数据治理工作提供适

用性强且系统性的指导，赋能多个行业数据治理工作的开展，支撑不同行业数字化转型的高效落地及数据要素的价值发挥。

本书结合 DMBOK 数据管理框架和其他数据治理相关框架以及健康医疗领域数据治理的特点，形成健康医疗领域的数字治理参考框架，该参考框架包括数据战略、数据治理机制、数据治理关键领域、数据应用及服务、数据共享与安全、数据治理成熟度评估六个部分，如图 1-4 所示。

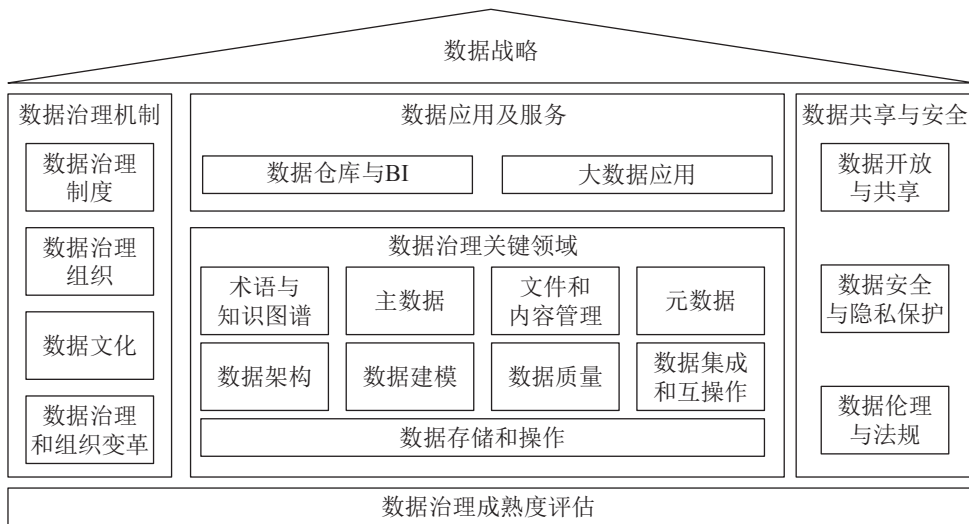


图 1-4 数据治理参考框架

(1) 数据战略：根据法律法规、行业监管要求、业务战略、未来科技发展趋势等对数据的需求，建立数据战略，明确数据治理愿景、目标、原则、任务、路径等要素，指导未来数据治理工作的开展。

(2) 数据治理机制：数据治理机制包括数据治理制度、组织及团队建设、组织变革以及数据文化，它是数据治理和数据应用各项工作有序开展的基础，为数据治理提供依据和机制保障。

(3) 数据治理关键领域：数据治理关键领域包括术语与知识图谱、主数据、文件和内容管理、元数据、数据架构、数据建模、数据质量、数据集成和互操作、数据存储和操作，它体现了数据治理的具体工作，能实现数据治理关键领域工作的落地。

(4) 数据应用及服务：数据应用及服务包括数据仓库与 BI、大数据应用，它是实现数据价值和作用的重要手段。

(5) 数据共享与安全：数据共享与安全包括数据开放与共享、数据安全与隐私保护、数据伦理与法规，它是防范数据风险、保障数据合法合规合理使用的前提。

(6) 数据治理成熟度评估：通过建立数据治理的相关评估指标，对上述数据治理相关工作进行成熟度评估和现状诊断，发现数据治理工作存在的问题及不足，明确数据治理工作的改进方向及措施，促进数据治理工作的持续改进。



第2章

医疗行业数据治理现状

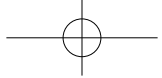
数据治理是指对数据资产管理行使权力和控制的活动集合。相比数据管理，医疗行业数据治理更强调国家级、省级卫生主管部门、医疗信息行业协会以及机构高层管理者统一协调、共同协作，建立起一套自上而下、分工科学、协作紧密、流程明晰的数据治理组织体系。目前我国在医疗行业领域的的数据治理处于起步阶段，存在着“重创造轻管理、重数量轻质量、重业务轻增值”的现象，亟须健全的治理体系和有效的解决方案。近年来，国内外相关机构在数据治理领域开展了深入的理论研究，提出国际标准数据治理框架、数据治理协会（Data Governance Institute, DGI）数据治理模型、DAMA 数据治理模型、国家标准化管理委员会数据治理框架等多种代表性的理论体系。在理论研究的基础上，我国医疗行业开启了数据治理实践的探索。

2.1 医疗卫生机构数据治理现状

医疗卫生机构包括医疗机构和公共卫生机构，是产生健康医疗数据的主要场所。随着医疗和公共卫生机构信息化建设的不断深入，尤其是近年来智慧医疗和智慧公卫的快速发展，医疗卫生机构的数据呈爆炸式的增长，但数据的质量和利用率不如人意，越来越多的医疗卫生机构认识到数据治理的重要性和紧迫性，开始了数据治理的相关工作。尽管数据治理已经提上了医疗卫生机构运营管理的日程，但就目前医疗卫生机构数据治理的现状来看，总体而言，我国医疗卫生机构数据治理仍处于起步阶段，各医疗卫生机构仍在积极探索有效可行的数据治理体系和具体实施方法。

2.1.1 医院数据治理现状

医院在推进信息化建设的过程中，系统越建越多，功能越来越复杂，因缺乏整体规划能力，各信息系统产生了众多数据，存在数据质量不高和利用率低的现状。数据是医院精细化管理的重要载体，在智慧医院建设中发挥着重要作用。面对如此现状，医院迫切需要构建数据治理体系，加快推动医疗数据的治理，提升数据质量，促进数据共享，发挥数据价值。当前，我国医院数据治理还处在起步阶段，存在以



下挑战：专业数据治理人才不足；数据中心重建设，轻管理，少应用；数据治理对人工还存在较强的依赖性，缺少自动化、实时性管理工具；重视数据安全的同时也限制了数据的共享和应用；数据治理未形成体系，难以支撑智慧医院的建设等。数据治理涉及医院所有业务和管理领域，需要明确相关部门和科室管理活动的职责范围和协作模式，方可切实保障数据治理工作的顺利推进。

2.1.1.1 医院产生的健康医疗数据类型及特征

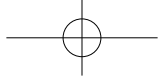
医院是产生健康医疗数据的主要场所之一，根据数据产生的源头，医院数据主要可以分为医疗服务产生的业务数据和医院运营管理数据。业务数据即临床数据，包括门急诊数据、住院数据、EMR 数据（医嘱、病历记录等文书、诊断、护理）、检验检查数据、手术麻醉数据、药品使用数据、随访数据等；医院运营管理数据包括财务数据（患者缴费、医保结算等）、医院成本数据、上报至统计信息中心和人口信息平台的数据等。

近年来，随着医院业务的快速发展，医院信息化建设不断完善，患者就诊流程中越来越多相关的数据被采集和保留下来。一方面，这些数据数量庞大，体系架构差异大，数据结构复杂，缺乏统一数据标准，即使在同一家医院，由于数据库的分割，信息孤岛的问题仍然存在，信息共享程度相对不足。另一方面，医院掌握着大多数患者的健康医疗数据，与其他健康医疗数据相比，准确性相对较高，极具开发价值。

2.1.1.2 医院数据治理的发展阶段

医院数据的重要性逐渐凸显，而医院数据由于其复杂性等因素未能得到充分利用，越来越多的医院开始对医院数据进行数据治理，以期深入挖掘医院数据的价值。2017 年 8 月，Dimensional Insight 对美国 104 名医院的首席信息官进行调查显示，约有 44% 的医院在整个医院内部已经实现了对健康医疗数据的管理，而剩下的 56% 的医院未能实现完整的数据管理或者尚未开始实施数据管理。2017 年，中国数字医疗网 HC3i 组织开展的《医院数据管理状况调查》活动显示，其中 84% 的医院在 2017 年针对健康医疗数据治理的预算达到 20 万元以上，50% 的医院数据治理的预算金额达到 50 万~100 万元。而 2016 年同期为 84% 的医院在数据管理上的投入高于 20 万元，33% 的医院投入资金在 50 万~100 万元。全球范围内，当前大多数医院对数据治理的概念有着一定的认识，也认可其重要性。可以发现在受调查的医院中有相当一部分医院对于数据治理都安排了资金投入，并且在数据治理上的预算不断增加，显示出医院对于数据治理重视程度在加深。

尽管数据治理已经提上医院运营管理的日程，就目前的医院数据治理现状来看，总体而言，我国的医院数据治理仍处于起步阶段，各地医院仍在积极探索有效可行的医院数据治理体系和具体实施方法。数据治理可以解释为“管理数据管理”，是



对数据管理的统筹，医院在数据治理中更关注对数据管理的执行，缺乏对数据治理的统筹规划以及组织制度建设层面的考虑。在数据治理概念理解上，大多数医院在实际工作中并未将数据治理、数据管理和数据处理进行明确的区分。在数据治理组织建设中，很大一部分医院的数据治理工作属于医院信息化建设的工作内容，并未组建独立的医院数据治理组织，而是直接由医院信息化相关科室开展。在数据治理人才方面，医院并不安排专职的数据管理专员，通常由信息科室的成员兼顾数据治理的工作。不过在一些大型三甲医院中，院方已经认识到组织结构在数据治理方面的重要意义，如北京天坛医院、北京清华长庚医院等医院在信息科内设立了数据小组，专门负责数据的挖掘和利用。还有医院成立了独立于信息科之外的部门，如陆军军医大学第一附属医院成立医院大数据与人工智能中心，并安排全职工作人员负责医院数据的相关工作。目前，医院数据治理的重点主要集中在数据质量提升、数据可用性、数据互操作性、数据安全等方面。

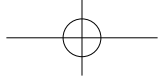
2.1.1.3 目前医院数据治理的主要方面

1) 数据质量管理

医院数据在其生命周期的各个环节均有可能出现不同程度的质量问题。数据的质量将直接影响数据的价值，也可能会对临床诊疗活动造成严重的后果。医院数据质量问题主要有数据缺失不完整、格式错误、重复记录、数据颗粒度不统一、数据录入不及时等问题。医院通常利用数据监测、数据核查等手段，对数据库中的数据进行抽取，并依据一定的标准对其进行梳理和检查，并对数据质量进行反馈。在核查和监测的数据项方面，由于医院数据的数据项众多，医院数据质量管理一般聚焦在那些最直接影响医疗诊疗活动的数据上。例如四川大学华西医院制定了数据质量控制框架，从数据质量属性和数据质量特征 2 个角度切入，分别通过 8 个和 5 个监测维度进行医院数据质量监测，覆盖了医疗服务能力、医疗服务效率、医疗收入、医疗费用控制、患者辐射范围、医疗质量终末指标、医疗保障等方面共 42 项核心管理数据。

2) 数据标准管理

医院在实施数据治理的过程中，制定统一可行的数据标准体系是数据治理的基础。对于医院来说，建立医院数据标准体系及其管理机制，有利于支撑数据的标准化建设，保证数据在数据治理过程中的一致性。各个国家和地区发布了多套医疗健康领域的标准，可供医院选择。我国经过十多年的发展也已形成了比较完善的健康医疗数据标准体系框架，现行的是基于大数据参考架构和全民健康信息标准体系框架提出的健康医疗大数据标准体系框架，包括 6 大类标准，分别为基础类、数据类、技术类、安全与隐私类、管理类和应用与服务类。同时，我国将部分健康医疗数据标准作为国家医疗健康信息医院信息互联互通标准化成熟度测评和区域全民健康信息互联互通标准化成熟度测评的依据，数据资源标准化建设情况占测评总分



的 30%，其中数据集标准化情况和共享文档标准化情况各占总分的 15%。该测评的举措在一定程度上促进了我国医院数据标准的统一，不但有利于医院自身的数据治理，也为医院间的区域性数据治理提供了重要保障。

3) 数据集成管理

医院在探索医院数据治理的过程中，通过建立临床数据仓库（Clinical Data Repository, CDR）、大数据中心等数据平台，与医院现有的 HIS、CIS、LIS、RIS/PACS 等关键系统数据库相结合，逐步搭建医院数据治理平台，形成医院数据治理框架。临床数据仓库是一个以业务支撑为主，面向电子病历，整合多个来源的数据中心，提供以患者和医护人员为中心的统一视图的数据库。与 CDR 相比，医院大数据中心强调数据的二次利用，主要起数据分析的作用，对医院管理和临床科研等方面至关重要。大数据中心包含了数据收集、数据治理、数据计算、数据挖掘分析、数据利用等方面的功能。医院利用数据平台，建立医院自身的临床样本库，并利用数据平台开展临床研究与临床决策支持，通过数据平台让数据产生临床价值。

4) 数据安全治理

医疗健康数据体现的价值不仅是经济价值，更是生命，也正因如此，数据安全一直是医院在进行数据管理活动时需要考虑的关键。病毒和人为的恶意攻击、人为误操作是影响医院数据安全的主要原因。为了加强医院数据安全，一方面，医院建立数据安全管理制度，包括业务系统密码管理、数据查询与登记、数据库后台安全监管、业务系统使用权限等相关制度，不断完善数据安全防范策略，从网络终端控制、数据存储、数据访问权限、数据交换、数据加密等方面着重保障数据的安全。另一方面，医院通过加强信息系统硬件设备，如增设备或线路、多路备份、增加异地备份等手段，降低数据安全风险。

2.1.1.4 医院数据治理的典型案例

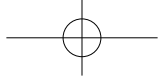
案例 1：美国堪萨斯大学医院（University of Kansas Hospital）

1) 面临的问题

堪萨斯大学医院面临医疗卫生机构几个常见的数据挑战：数据收集方面，随着多源数据的数据量日益增长，通过手工的方式收集、获取和验证数据，非常耗时耗力，医院需要寻找一个更有效的替代方案；数据质量方面，数据不准确、完全丢失或数据相互矛盾对数据的应用造成重大的影响，用户对数据不信任问题突出；数据标准化方面，存在数据术语和定义的非标准化问题等。

2) 数据治理目标

堪萨斯大学医院希望通过实施更好的数据管理流程和实践，来改善数据收集和 data 质量方面的问题，同时提高用户对数据的信任程度。



3) 数据治理实施

(1) 顶层设计

堪萨斯大学医院首先从组织构建、战略规划改进、架构设计三方面进行了数据治理的顶层设计。

在组织构建上,堪萨斯大学医院成立了一个数据治理委员会,该委员会由医院高层领导批准成立,并拥有管理所有 BI、数据和信息资产的权限。如图 2-1 所示,该委员会被划分为数据咨询组、数据治理执行组和数据治理支持团队等。数据咨询组由相关研究领域的专家组成,主要在战术策略的跟进上发挥作用,负责处理数据质量问题、工作优先级和工作组的创建等;数据治理执行组负责监督改进工作的策略和远景;其他主要委员会成员包括数据治理支持团队,由技术、流程改进和临床专家等组成。

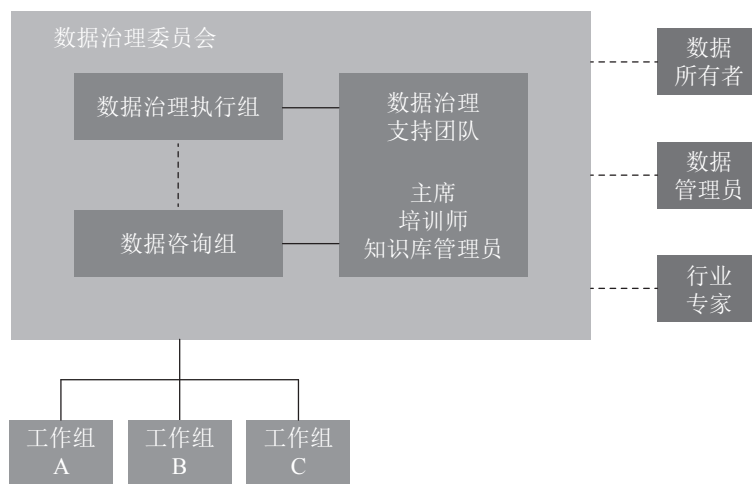


图 2-1 堪萨斯大学医院的数据治理组织架构

在战略规划改进上,堪萨斯大学医院选择以现代分析基础设施作为劳动力,替代持续增加的全职员工的工作量,从而满足不断增长的数据报告需求。通过这些现代化分析基础设施的部署实现数据收集过程的自动化,获取及时、准确的决策数据,并实现为用户访问正确的数据,从数据报告转向识别和执行等环节改进。

康萨斯大学医院也对架构设计进行了调整。从管理大型数据集和使用预测分析创建最佳结果的理想视角出发,堪萨斯大学医院开发了一个更具凝聚力的商业智能路线图(如图 2-2 所示),并根据该路线图跟踪数据治理的进展。

堪萨斯大学BI路线图	2015财政年		2016财政年		2017财政年			
	零散分布		企业视角		高级分析		大数据	
BI架构	没有或少量关键解决方案		1 实施中央基础设施	2	核心&自助服务设施	1	优化的基础设施（如：数据集市、ODS）	
数据源/数据时效	来自一个系统或来自内部系统有限数量的特定BI系统的交易申请		1 主要数据源构建的ETL	2	辅助数据源构建的ETL	1	网页、患者、基因组学&其他外部来源（如：堪萨斯大学、eMeasure、对比分析）	
分析类型/分析的利用	自动内部报告（部分）		1 企业KPI&自动外部报告	2	预测性&规范性分析和基于证据的分析	1	结合多个复杂数据源的分析	
数据模型	部门的		1 通用词汇、起始模式、维度	2	多种数据模型	1	无	
数据治理	独立的&部门的		1 通用政策/标准、集中管理的KPI&安全管理	1	商定议程&优先事项、数据规范化&系统变化的初始来源	1	内部&外部数据的所有者&管理者、复杂分析的审查和交付	
工具	冗余信息、桌面工具集		1 统一数据管理工具	1	拓展的分析能力	1	专门的、针对性的能力	
技能	SQL、Excel、Access轻量化的数据模型、轻量化的可视化		1 对人体有深度理解、逻辑数据模型、轻量化统计和术语标准	2	对统计有深度理解、操作分析、程序设计	1	NLP、基因组学、规则引擎	
文化/企业数据素养	数据价值被低估、足够好的决策		1 越来越强调基于事实的决策	2	数据素养培训识别BI机会、改变	1	深刻理解BI能力的局限性	

图 2-2 数据治理长期路线图

（2）数据治理重点

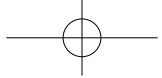
堪萨斯大学医院数据治理委员会主要从信息治理、数据质量、数据的可用性和数据的有效性四方面入手改进其数据质量和数据可信问题。

信息治理方面，通过将临床、管理和技术合作伙伴联合起来，根据组织的战略和目标共同设计和优化信息资产，包括对事实来源的声明、记录系统、角色和职责、信息交付标准和经过认证的企业报告 / 仪表板，以及数据安全和访问决策的治理。

数据质量方面，以数据的完整性、及时性、准确性和一致性为依据，通过标准体系和数据质量评价评估体系的创建和数据工作流程的监控来确保数据质量。

数据的可用性方面，开发易于使用的应用程序，并以数据字典、数据模型和元数据存储库作为工具，尽可能实时地访问可信的数据集以促进数据的交互和知情决策，创建一种通用的数据语言来理解组织绩效。

数据的有效性方面，实施适当的分析基础设施，使临床和业务用户需要时能够获得各自的临床和业务需要的数据。



（3）数据治理流程的改进

在统筹和规划层面，堪萨斯大学医院着重对统筹层面的领导力进行优化。最有效的领导者应该具备必要的技术技能和客户服务意识，以便与临床和行政领导建立伙伴关系，同时重视确保高级领导人的承诺，在实施先进的分析基础设施和有效的改进策略时需获得高级领导人的积极支持。由医院数据治理委员会带动建立数据驱动的核心文化，并将数据所有权的责任落至临床医生和操作决策者，由终端用户承担数据的所有权，并承担管理数据质量的责任。医院数据治理在构建和运行层面，通过构建高级商业智能体系结构，确保管理团队具备该项目所需的必要技术、操作和临床专业知识，并提供足够的资源和专用资源推进数据治理；以透明度为目标，向终端用户提供用于日常工作的高质量数据；从多个源系统中提取质量数据，并将其提取为一个单一的真理源（Enterprise Data Warehouse, EDW）；尽可能自动化数据提取和报告；基于通用词汇表和术语进行高级数据建模；教育培训用户，向他们提供性能优化的高级分析和改进技能，允许用户能够有效地拥有和管理他们的数据需求。

（4）监控和评价

堪萨斯大学医院还计划构建一个数据治理绩效评估表以跟踪和管理治理指标，这些指标包括：确定的数据所有者数量、记录的流程数量、节省的成本 / 避免的费用和企业数据仓库的使用情况（例如获得的效率、减少非企业数据的使用）。

案例 2：首都医科大学宣武医院

1) 面临的问题

首都医科大学宣武医院在数据应用过程中面临院内数据质量冲突问题、复杂数据的获取困难问题及数据治理实施过程中的数据安全保障问题。

2) 数据治理目标

通过数据治理实现不同就诊场所、不同时间的患者及诊疗信息贯通；实现全院级别的影像数据的统一管理；在保证信息安全的前提下增强了各类临床数据的可及性。

3) 数据治理实施

（1）顶层设计

在组织架构上，宣武医院专门成立了数据治理委员会，数据治理委员会是信息管理委员会的分支机构，涉及职能部门、主要数据采集部门和应用部门、信息中心等多类人员。

（2）数据治理实施

从四个方面进行数据治理，分别是数据内容管理、数据安全与隐私保护及数据获取便捷性、数据整合与数据质量、主数据管理。

在数据内容管理中，持续不断地扩展医院级别统一管理的数据内容，例如整合床旁设备数据，逐渐将监护仪、呼吸机、输液泵、床旁血气分析等设备输出数据整

合到信息系统；通过流程和程序改造，将专科系统内存储的影像（如眼科影像、内窥镜影像等）加入患者标识，逐步联通整合到影像平台，进行全院级别的统一管理；增加院内外数据的联动性，通过提供患者报告数据的入口，构建医疗数据生态圈；通过拓展 APP 功能，提升随访、医患互动的便捷性以及对患者报告数据的采集可行性；

数据安全与隐私保护及数据获取便捷性方面，制定面向院内和院外，涉及和未涉及患者隐私的不同数据提取流程。数据提供的内容只能按照需求方、迫切度、持续时间、需求频度等具体要求去分析，综合决策。将数据治理和信息安全委员会合并，从而促进关于数据利用的长期决策合理化，保证数据安全和隐私保护。

数据整合与数据质量方面，主要从数据完整性、数据的真实可信性和数据时间线三个维度考察数据质量，关注数据形成的闭环，对数据质量进行多重验证。例如患者出病房、入手术室、开始麻醉、开始手术等一系列时间的先后顺序，系统要进行验证，不合理的数据不能存入系统。

主数据管理方面，应用主数据管理工具，保证数出一源，口径一致，如对诊断字典进行规范统一，尽量减少诊断字典的种类，并通过集成平台将最终统一的字典向各个业务系统分发。统一不同系统中同一指标的统计口径。并由数据治理委员承担起主数据管理中定义、鼓励使用和解决数据冲突管理者的角色。除了指定编码数据标准，数据治理委员会还需参与制定数据计算、分析的算法，将数据绑定到分析算法中，并在整个机构内统一、持续使用。业务主数据治理范围包括机构数据标准（材料编码、人员编码、设备编码、部门编码等），以及区域和行业标准（ICD、SNOMED、LOINC）等。患者主数据管理除了使用患者主索引（EMPI）这样通用的患者主数据管理工具，还基于医院实际门诊、急诊、住院患者身份标识（ID）不统一的实际情况，对患者各个场景下住院的信息进行贯通，并设计多个验证环节以弥补 EMPI 功能无法覆盖的情况。

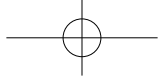
2.1.2 公共卫生机构数据治理现状

广义的公共卫生机构指一切能够促进健康、预防疾病、保护健康的机构。本书中的公共卫生机构聚焦在疾病控制机构，这些机构的数据治理的主要目的是应对公共卫生事件，通过对公共卫生机构收集的相关数据进行治理，并基于数据，对公共卫生事件及时地进行干预和处理。疾控机构收集的数据主要是人群的公共卫生监测数据。

当前公共卫生机构的数据治理重点主要集中在数据质量管理、数据标准管理和公共卫生监测网络。

1) 数据质量管理

公共卫生监测数据的质量问题大部分集中在数据采集录入环节。公共卫生机构在设计数据收集报表时，就已将数据质量因素考虑在内，利用规范、逻辑清晰的报表来确保录入数据的准确性和完整性。电子化的报表还可通过限制录入字段的格式



控制输入的数据质量。利用双录的形式输入数据也是公共卫生机构保证数据质量的常用手段。在收集数据之后,工作人员还需要对数据进行审核、抽检,以提高数据质量。另外,我国公共卫生机构还建立了反馈机制,上级主管部门在收到公共卫生机构的相关报告时,会对数据质量进行反馈,从而提高机构的数据质量管理能力。

2) 数据标准管理

我国针对疾控系统的数据标准不够完善,各地疾控机构在建立数据标准体系时,一般参考《中国公共卫生信息分类与基本数据集》《中国公共卫生信息分类与基本数据集》等数据标准,结合机构自身的业务需求,对疾病预防控业务进行分类,并利用该业务分类构建疾病预防控制数据标准体系框架。疾控机构的数据标准体系覆盖疾病监测管理、健康影响因素监测与干预以及公共卫生服务等方面。

3) 公共卫生监测网络

各国部署了公共卫生监测网络来收集疾病监测数据,并对数据进行分析。我国自2004年启动国家级疾病监测信息报告管理系统,实现传染病监测数据从县级到中央的实时报告。系统通过自动对收集到的相关疾病监测数据进行分析,还可以起到疾病预警的作用,实现对传染病的事前监测,对及时控制传染源和切断传播途径起到重要作用,有效阻止传染病的进一步蔓延和扩散。例如,中国疾控于2008年4月运行了国家传染病自动预警系统,系统利用数据模型,自动分析各地上报的疾病监测数据,实现对超过30种传染病的自动监测,并将疾病异常发展的预警信号发送给相关区县的疾控机构,同时收集对应机构对该预警信号的处理反馈。2016年,该预警系统共发出325208条预警信号,涉及2996个区县,在一定程度上控制了传染病扩散的规模。

2.2 政府行业管理机构数据治理现状

2.2.1 卫生健康管理机构数据治理现状

国务院和国家卫健委等政府管理机构高度重视健康医疗大数据、“互联网+医疗健康”工作,为推动医疗大数据的快速发展,国家相继出台了一系列相关政策:如2015年9月国务院发布了《促进大数据发展行动纲要》;2016年6月国务院印发《关于促进和规范健康医疗大数据应用发展的指导意见》;2017年2月发布的《“十三五”全国人口健康信息化发展规划》提出要夯实健康医疗大数据基础、深化健康医疗大数据应用、创新健康医疗大数据发展三大任务;2018年9月,为加强健康医疗大数据服务管理,促进“互联网+医疗健康”发展,充分发挥健康医疗大数据作为国家重要基础性战略资源的作用,国家卫生健康委员会正式发布了《关于印发国家健康医疗大数据标准、安全和服务管理办法(试行)》;2020年4月,《中共中央、国务院关于构建更加完善的要素市场化配置体制机制的意见》正式公布,数据首次被正式纳入生产要素范围,与土地、劳动力、资本、技术等传统要素并列

为要素之一。

在国家层面的积极倡导下，各地政府、医疗机构和相关企业等开始从不同环节切入，进行医疗大数据建设，持续推动大数据、互联网、人工智能等新兴技术在医疗健康领域落地应用。通过建立一套完整的数据治理体系来加强全民健康信息化建设，推动国家级全民信息平台和省统筹区域全民健康信息平台的建立，还在规范医院信息化建设方面制定了《全国医院信息化建设标准与规范（试行）》及其相关的数据标准化指导文档，组织全国公共卫生信息化标准与规范文档的编制工作，为医疗健康行业的数据治理奠定了坚实的基础，并且在国家和地方层面都开启了健康医疗大数据的治理工程。

2.2.1.1 国家健康医疗大数据治理现状

1) 国家全民健康信息平台数据治理

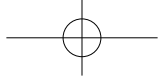
国家卫生健康委统计信息中心在实施全民健康保障信息化工程一期项目，以及推进国家全民健康信息平台建设过程中，高度重视数据治理工作，包括推进数据统一采集、资源目录统一注册、数据汇聚集中、数据质量多重管控、使用的规范化管理、数据综合分析利用等。国家全民健康信息平台数据治理内容如下。

(1) 组织制度保障

为有效保障全民健康保障信息化工程一期项目建设，推进数据治理工作，在项目建设之初，国家卫生健康委就制定、印发了相关规章制度，明确了工程实施组织架构，相关司局承担数据资源的组织、协调等职责，各实施机构负责相应业务系统的建设。国家卫生健康委统计信息中心牵头各实施机构，共同建设统一应用支撑平台，以及部分业务应用系统。为加强项目实施、系统开发和数据管理规范，项目组织制定了包含管理类、技术类、数据类和安全类等一系列规范和办法，结合部门相关工作规章制度，规定了各部门、单位工作机制和流程，基本奠定了数据治理的组织架构和制度基础。

(2) 资源中心构建

通过统一应用支撑平台，实现数据的统一采集交换与汇聚。业务应用子系统提供数据采集需求，由业务系统负责制定、发布数据集、数据元等标准，由平台进行统一注册管理业务系统、系统表单、数据字段和术语字典。国家级平台以及部分业务应用系统采用平台对接或系统对接方式实现联通，数据进入平台采集交换库，转发给业务应用子系统生产库。国家全民健康信息平台数据资源中心最终整合省级平台、相关业务应用系统、业务应用分平台数据资源，支撑业务应用子系统之间的业务协同，并可进行各项数据的管理、分析与利用，从逻辑上可以分为支撑库、全员人口库、健康档案库、电子病历库、生产库和分析库等。数据资源中心按照业务划分对数据分类存储，并组织成基础库、主题库、索引库、字典库等数据库。根据数据挖掘和分析需求将数据组织成主题类数据，在数据仓库内分类存放。通过服务管



理支撑库、数据管理支撑库和共享库向外提供数据服务。

（3）数据质量控制

数据质量控制体现在业务系统建设、数据仓库建设。各重要阶段设置数据检查监控点，并能实现跨监控点、数据源的比较分析。全民健康保障信息化工程通过建设应用支撑平台数据质量控制系统，实现对数据生命周期的数据质量控制，涵盖数据源到应用端的全过程分级负责，重点是业务系统端和数据资源中心端的数据质量控制。应用支撑平台采集交换库经过技术质控，将质控的数据转发给业务应用子系统的生产库，生产库可根据业务需求自行对数据进行质控。应用支撑平台提供统一数据监控能力，为平台数据采集交换各组件，提供完整的数据采集日志和数据交换监控报告信息功能，对业务层、服务层数据质量提供质量评估、质量评价、质量报告、质量监控等功能，为信息资源中心建设提供质量保证支撑。

（4）数据安全

在项目整体信息安全体系基础上，加强平台数据安全。主要措施包括提供数据加解密服务，实现数据安全，对数据进行秘密性保护；对平台业务数据进行存储备份；对接入实现权限控制，要求提供口令验证、IP 地址验证、时间控制等；加强安全策略管理、数据安全检测、数据权限管理和数据安全审计；加强采集与共享数据传输的安全性，兼顾平台网络传输效率；设计数据脱敏，确保国家全民健康数据中心数据高效、可靠及业务应用安全。通过数据治理安全设计实施，进一步加强平台在数据采集、上传录入、数据利用、跨区域共享等环节的管理，保证数据的完整性，加强对敏感数据的防护。

（5）数据资源管理

基于数据资源池构建数据开放目录、数据共享系统，建立数据集与数据表之间的映射关系。在此基础上，实现数据之间的影响分析和血缘管理分析。基于集中管理的元数据，建立数据监控系统。在数据生命周期管理过程中，数据监控负责监控数据生成、传入、存储、处理、应用、销毁等环节，主要包含数据访问监控、数据状态监控、数据质量监控和数据安全监控等。可提供对数据的可视化访问监控、状态监测、统计分析、告警、报表生成功能，提供完整的日志和监控报告信息功能，确保数据使用安全、系统运行稳定。

（6）数据价值利用

通过安全审批、统一监控、权限管理等功能，设计保证数据安全、高效、便捷使用，为分析利用提供支撑。利用数据开放目录创建、管理的能力，提供较为通用的预测、预警、对比、关联等模型。未来逐步面向国家部委、地方省市相关机构、行业、科研团体等不同的用户，进行权限设计、开放共享。通过融合各类结构化、非结构化及半结构化的数据资源，实现对海量的医疗健康数据的有效利用，聚焦在分级诊疗、现代医院管理、医疗卫生服务资源等方面作出分析示范。加强了平台数据的可视化能力，提供可视化公共服务，力求数据价值不断提高，推进数据管理科学化。

2) 国家健康医疗大数据中心（北方中心）数据治理

2016年6月国务院办公厅印发《关于促进和规范健康医疗大数据应用发展的指导意见》，将健康医疗大数据应用发展纳入国家大数据战略布局，随后原国家卫生计生委按照“1+5+X”的总体规划，确定福建（南部中心）、江苏（东部中心）、山东（北方中心）、安徽（中部中心）和贵州（西部中心）为5大区域数据中心试点省份，承接国家健康医疗大数据中心和产业园建设。2018年北方中心在山东济南落地，率先完成了国家卫生健康委试点评估。北方中心汇聚了山东省近5年报送系统数据，包括省人口库、省健康档案库、省电子病历库和省卫生资源库，以及济南市近百家医院的全量数据。为打通健康医疗数据汇聚、传输、存储、处理、交换、销毁全生命周期链路，有效支撑惠民、惠医、惠企、惠研、惠政等行业应用，北方中心实施了数据治理工程，具体内容如下。

针对数据质量无法满足业务需求的数据资源，按照统一标准对数据资源进行标准化处理，形成数据质量较高的可以支撑上层应用的DWD数据层。数据治理主要针对原始数据进行数据元解析、数据映射、数据清洗、数据归一、值码标准转换等方面的处理，理清医院的库表结构与表间关系，并对整个数据处理过程形成可跟踪、可追溯的处理流程记录、脚本，最终形成完整的数据处理报告。待治理数据来源包括省市级全民健康信息平台、各级医院单位系统等。

（1）数据注释 / 解析

根据健康医疗大数据北方中心的实际状况，对原始数据提供数据元注释 / 解析服务。数据来源：解析来自省级医疗数据平台和市级医疗数据平台的全量数据；解析来自三级医院、二级医院以及二级以下医院的全量医疗数据。

（2）数据映射

依据医院表结构、表数据及数据标准，提供数据映射服务，对医院原始数据进行数据归一化映射（包含表与表、字段与字段级的关联关系）、值域标准化、值码标准化，确保归一后的数据与医院生产的数据一致，满足支持上层业务应用的需求。

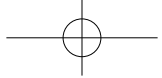
（3）数据质量管理

完善、规范数据标准与数据质量管理规则，包括数据质量类别管理、数据质量检核入库、问题数据展现、问题数据趋势分析、数据质量检核监控、检核日志管理、数据质量报告、专项数据质控规则处理模块等。根据数据标准与数据质量管理规则，针对省市级健康医疗大数据平台与各级医疗机构系统汇聚上来的原始数据进行数据汇聚质量评估标准服务。数据质量管理规则将最终落实到系统工具中，输出可供追溯、分析报告。

2.2.1.2 地方健康医疗大数据治理现状

1) 甘肃省健康医疗大数据治理

甘肃省卫生健康委于2016年5月启动了省级全民健康信息平台建设，建设内容



主要包括：完善全员人口数据库、电子健康档案数据库和电子病历数据库，依托省级交换平台建成医疗协同服务、公共卫生服务、人口计生服务、医疗保障管理、药品采购配送、综合卫生管理等6大应用系统，实现面向社会公众的服务、跨区域业务应用协同和数据交换共享、面向各级卫生计生机构的综合信息统计分析、大数据应用和决策支持功能。经过三年多的建设和发展，甘肃省的卫生信息化取得了明显的成效，并且积累了大量、优质的健康医疗数据。但是，由于医疗机构信息系统的建设时间、建设主体和承建单位的不同，采用的数据库、平台、网络结构等不同，造成了大量形态不同的数据。为进一步提高甘肃省健康医疗大数据的质量，使甘肃省的健康医疗大数据发挥真正的作用，省卫生健康统计信息中心开始探索对大数据进行治理。本次数据治理在保证不对原有系统做任何改动的前提下，数据保留在原系统不做任何迁移，逐项对甘肃省的全员人口数据库、电子病历数据库、健康档案数据库、健康扶贫数据库和卫生资源数据库进行了治理。数据治理的过程如下。

（1）明确数据的标准与规范

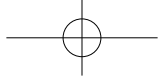
甘肃省全民健康信息平台是根据国家信息化建设的要求，以全员人口、电子病历、健康档案等五大基础数据库为基础，实现了医疗保障、免疫规划和慢病管理等多项业务系统的互联互通，新建了远程医疗、双向转诊等新业务系统，通过居民电子健康卡的创新应用，为甘肃省健康医疗大数据的可持续发展奠定了坚实的基础。该平台以国家制定的疾病诊断、手术操作、药品和医疗服务项目编码为标准，以全省统一的接口规范完成了医疗机构的接口改造工作，基本实现了省、市、县、乡、村医疗机构的全覆盖。数据标准和规范的制定，在一定程度上解决了医疗机构信息化建设的历史遗留问题，通过数据标准和规范的明确使医院内部多个系统之间、不同医疗机构之间的数据实现互联互通，凸显数据深度挖掘的价值。

（2）最大限度整合历史数据

过去卫生行政部门和医疗机构也积累了大量的历史数据，由于历史原因，数据的利用率和开发程度处于非常低的水平。由于数据标准和规范的不同，历史数据的使用必须经过数据的转化，将其转化为满足现行标准和规范的数据。在历史数据的转化中，需要建立较多的对应关系和对应规则，需要充分考虑当时的政策背景和业务场景。数据整合结果的验证是难度比较大的工作，主要通过抽样的方法从医疗机构的HIS系统中进行验证，其次就是采用循证医学的方法在相关的研究文献中提取验证规则。

（3）有计划地进行数据清洗

数据清洗是指在对数据进行审核时，将脏数据清洗为干净数据的过程。而所谓的脏数据则是指在数据审计的过程中发现的有问题的数据，即低质数据。目前，对于数据的清洗研究，主要集中体现在缺失数据的清洗、冗余数据的清洗、异常数据的清洗。在数据的清洗过程中，科学地分析脏数据的类型，根据不同类型的脏数据定义不同的清洗策略。数据清洗的粒度必须是从最小粒度的“字段”开始，以“字段”



为单位制定数据的转化规则。选择合适的清洗工具，将计算机决策和人工清洗有效结合，并对清洗的结果进行抽样验证。

（4）科学选择数据清洗技术

随着信息化技术的不断发展，数据的清洗技术也随之发生了较大的进步。数据清洗技术是为了提高数据质量而剔除数据中错误记录的一种技术手段，在实际应用中通常与数据挖掘技术、数据仓库技术、数据整合技术结合应用。由于历史的原因，目前采集到的数据，种类多样，结构复杂，所以有大量的数据需要进行清洗和处理以后，才可以进行利用。传统的人工处理的成本会非常大，而且效率和准确率都有明显的差异，所以自动化应用程序的清洗工具也受到了各方的关注。本项目对数据的清洗以函数清洗为主，结合人工清洗，函数清洗是目前大数据清洗过程中比较常用的一种手段，对于结构性较好的数据有明显的作用，可以同时纠正数据的异常、重复、错误和缺失等多个方面的问题。数据的清洗工作主要是在数据的预处理环节进行，力求从数据源头减少噪声数据，有效提高了数据的清洗效率和分析结果的准确性。人工数据清洗主要是定义清洗规则和规则的优先级别，清洗过程中对数据采用了“分类分级”的清洗思路，针对不同优先级的数据制定不同的清洗规则。

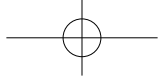
通过数据治理取得如下成果。

（1）**人口基础数据的质量明显提高。**治理后，甘肃省的全员人口数据库总记录条数 4.8 亿条，占用存储 124.42GB，分属于 12 个数据库表。甘肃省常住人口为 2600 多万人，录入基本实现全覆盖。治理后的全员人口数据库可以实现到乡镇级的数据查询，并支持个人查询。

（2）**电子病历数据的价值密度大幅提升。**经过治理后，甘肃省健康医疗大数据平台电子病历数据库目前实现了门诊、住院、手术、麻醉、孕产和护理等 16 项电子病历的查询、统计等功能。

（3）**盘活了居民的电子健康档案。**通过治理后，甘肃省健康档案数据库集 15 个，数据表 52 个，累计获取居民健康信息 16506 万条，占用存储约 92GB。累计获得个人基本信息 2733.78 万条，个人疾病史 306.01 万条，个人家族史 280.23 条，个人手术史 110.81 万条，个人输血史累计 108.22 万条，个人外伤史累计 168.94 万条。目前，甘肃省的健康档案数据库涵盖了个人档案、家庭档案、健康体检、孕产妇健康、儿童健康等诸多领域，对高血压、糖尿病和严重精神障碍患者实现了全方位监测。

（4）**健康扶贫工作的监管更加精准。**经过数据治理后，甘肃省的健康扶贫数据库实现了对贫困人口精准筛选，实现了对贫困人口就医环节的全程监管，对健康扶贫工作进行实时监管。目前主要分析的指标包含全省的建档立卡贫困户的总人数、参加社会保险人数、患病人数、患病人数中的已治愈人数和未治愈人数；健康状况分布情况（健康人群数、患病人群数、死亡人数）；婚姻状况分布；兜底保障户的统计情况；患病的疾病分类情况、因病负债情况等。健康扶贫数据库包含贫困人口档案、入户调查记录和“一人一策记录”三大模块，实现了 188.28 万条入户调查的



记录,核对了 163 万条贫困人口的健康状态,对 3.18 万人的治愈情况进行核实。

(5) 卫生资源大数据的治理。经过治理后的甘肃省卫生资源数据库,涵盖了 2.89 万家医疗机构,总数据 1.99 亿条,建立 57 个数据库表。实现了对 200 余家二级及二级以上医院的卫生资源查询;覆盖了全省的疾病预防控制中心,包含了 1300 多家卫生院以及 1.70 万家村卫生室。

2) 成都市成华区健康医疗大数据治理

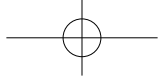
成都市成华区卫生健康局对目前辖区内 13 家小区卫生服务中心及 5 家公立医院进行了联网,目前已经具备了全区卫生专网,为实现区域医疗健康大数据发展进行医疗健康大数据治理与应用开发的工作,核心工作在于完成医院和基层医疗健康信息系统数据的融合治理和可视化建设,达到数据共享和统一监管的初步愿景。成华区在对医疗健康大数据治理过程中在元数据的生命周期管理和变更管理的基础上来实现功能层的建设和应用。基于元数据的健康医疗大数据治理过程如下。

(1) 元数据解析和融合

元数据的采集首先会从区域内各医疗机构内原始数据进行 ETL 抽取与解析。在不同的机构中,不同的厂商对相同的业务所定义的元数据命名不同、含义不统一。为便于统一认识和管理。需要将不同的元数据翻译成统一的语言以及描述,在此需要建立元数据语料库,并对所有元数据基于语料库进行翻译。在平台中需要对所有元数据翻译结果进行维护。在传统的业务系统建库时,针对一个业务所涉及的表,表之间是有相互业务关联关系的,但在建库时并非所有的业务关联关系都会建立在数据库中,无法通过元数据之间关联关系完整地映射出一个业务场景,在对元数据管理过程中,需要通过业务分析,将各种表间关联关系维护到元数据库中,形成一个完整的业务流程。基础元数据语料库主要包含数据元及数据集,通过元数据语料的审核和知识库来管理元数据。在知识库中,用一套可维护的,统一的语料表达数据元和数据集,实现对具备相同含义的不同数据元和数据集给予统一的描述。利用基础元数据语料库,通过自然语言处理(Natural Language Processing, NLP)算法对数据进行去重、归一、梳理、消歧然后将元数据同基础元数据语料库进行对比识别,完成不同系统中的不同元数据的统一理解过程,将处理后的业务数据模型添加进知识库;再将数据按照对应业务数据模型抽取、清洗到目标数据库中。当某一厂商的数据源通过智能算法及人工核检统一其元数据表述后,将会以厂商为单位,按分类分别保存和展示每个版本的元数据信息,包含原始元数据与通过算法识别和人工校验后统一的元数据信息。

(2) 元数据生命周期管理

元数据生命周期管理核心步骤包括对元数据采集,元数据中识别,值域映射,结构标准化 ETL 和数据质量的管理。一个系统库的处理流程代表一个任务,每个任务是一条完整的监控链,通过监控链可查看每个任务执行过的步骤、跳过或未执行的步骤。在处理流程中,可以通过查看元数据生命周期中每个节点的进度进行管理。



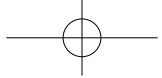
每个版本的元数据通过一条完整的元数据生命周期管理，保证元数据的一致性和连续性，避免元数据错乱。利用数据的生命周期管理完成对数据流向过程中的监控，把控数据流向的每一个过程，完成对数据处理过程的追溯。通过数据流，以数据源为基准，分析每张表里，每个元数据与其他数据源中元数据之间的联系，可能存在一对多的关系，推断出该元数据的影响力，即该元数据可能影响到的其他数据，从而为保证元数据的一致性和正确性提供保障。经过了数据溯源和数据血缘分析后，记录每个元数据在数据流通网上与其他元数据的关系。在进行元数据生命周期管理过程中所有对元数据的操作过程，包括元数据的导入和导出都将信息转换为使用标准元数据格式的 XML 文件。

（3）元数据变更管理

随着业务场景的变化，系统的升级改造，元数据会随着业务的改变而发生变化，通过对元数据进行变更管理，尤其在监控上层数据流中的元数据，以规避对后续计算统计结果的影响。保留业务元数据历史变更记录，以便后续可对此进行版本管理追溯，统一对元数据版本进行管理追溯。对元数据的变更进行监控，查看明细信息，随时监察，消除问题隐患，掌握数据资产的变化。通过定时检测生产库中元数据的变化，并通过影响分析查看该元数据变化后对上层应用的影响。再进行元数据变更后对比，包含元数据及其所有下游数据的比对，监控元数据的改变；对比相同厂商的元数据结构，完成元数据差异的对比，对于其本身元数据结构的更改，同步更改到元数据标准库中；对比不同数据来源的单位，完成同库下元数据版本的对比。

基于元数据管理的数据治理工作的完成使得成华区健康医疗大数据的治理达到了预期成果。首先通过对元数据的解析融合，梳理出了数据资产目录。对业务系统及计算引擎中存在的大量数据表、API 等各类数据资产进行统一管控，并提供对应数据资产管理规范。再通过数据资产管理，资产使用者可快速对资产进行查找、申请和使用等操作。编制出的数据资产目录可以给出业务场景和数据资源的关联关系，降低理解系统数据的门槛；同时考虑到随着对业务数据的抽取转换和分析次数的递增，系统中存储的元数据数量越来越大，为了避免出现混乱以及便于后续数据的利用，相应地开发了元数据检索功能，辅助用户定位和查看元数据；其次，通过元数据的治理可以进行血缘分析，指出某一类数据是从哪张表采集而来，后续又转储到哪张表中，构建出完整的数据流通网，分析打通数据在产生、处理、流转到消亡的全过程，数据能被串联起来，便于对数据进行快速查询、维护。

统一元数据标准后的主数据，能够在跨平台间实现数据采集、融合、监管等工作，并且在 HIE 融合等过程中，实现元数据向标准数据的转换映射、数据清洗过滤等数据治理流程；最后将整合后的医疗数据通过按疾病、药品、公卫监管等主题分类放入数据集中，为上层应用提供更高效的数据支撑。最终成华健康医疗大数据平台项目建设成功并持续投入使用，实现了精准到人的面向百姓的健康服务，包括慢性病管理服务，强化对家庭医生的管理，使家庭医生能够真正服务群众，对慢病做到



早预防，早治疗；实现了面向卫健局的监管服务，对各门诊机构的就诊人次、人均医疗费用等指标进行实时监管。

2.2.2 医疗保障局数据治理现状

医保大数据以人的健康为核心，以数据标准为基础，通过数据治理整合涵盖医保链条、全周期、全人群、全方位的数据资源，通过数据分析、人工智能等技术，实现医保数据信息的规范化、可视化、自动化、智能化，满足政府科学决策、智慧治理，医疗资源优化配置，方便群众就医等需要。

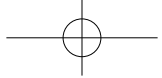
虽然医保数据蕴含着巨大价值，但是目前我国城乡二元制度造成了人员、信息及资源的长期割裂局面，医保统筹信息系统应用主要集中在区县级，且多数相对独立，各地信息系统的软硬件不一，这使得相关医保数据亟待通过数据治理后整合形成大数据。数据治理是实现医保大数据的必然路径，其可以通过分析病人收入水平、家庭负担等情况，寻找解决因病致贫的有效办法；可以通过分析社会保障基金的运行数据，结合医疗资源分布，掌握社会保障的整体效率，提出社保、医疗资源合理配置的建议；还可以通过分析病人就诊情况，研究包括临床医学、流行病学、临床路径等在内的医疗服务行为，改进治疗方式、卫生经济学评价等。值得注意的是，数据治理是医保数据整合中的前提，也是实现医保大数据面临的最大难题。例如，各地医保系统用的疾病分类代码，有用ICD-10，也有用ICD-9，还有地方自定义的一套。此外，药品分类编码、病人基本信息编码等也存在不一致。这些问题都亟待解决。

2.2.2.1 医保关键数据问题

《“十四五”全民医疗保障规划》中明确指出医疗保障是减轻群众负担、增进民生福祉、维护社会和谐稳定的重大制度安排。建成完善的基本医疗保障网，推进医疗保障高质量发展，满足人民群众多元化的医疗保障需求是进行医疗大数据治理的核心目的。目前国家在推动医疗大数据的发展和应用方面出台了一系列政策与有力的指导，但想在充分发挥大数据的效能的同时确保各类医疗健康数据的准确性、安全性、完整性，还需要更多的资源投入和数据治理相关技术与工具的研发跟进。

一是存在数据质量问题。各信息平台中存在着大量基础数据不全、不准确、标准不统一等问题。造成原因主要有：来源于各方应用的参保人数据在录入时没有相应的校验规则或同一参保人多次录入不同的个人信息；各地区医保信息系统间差异巨大，在数据采集时也没有统一标准；新旧数据之间没有有效的更替机制。

二是各信息系统间互联互通程度不高。医保、医疗机构以及监管部门之间互联互通程度低、信息系统林立各自有参考标准、数据化获取渠道单一、数据利用率低等问题，都不利于数据的获取与流动，也不利于医保监管，易出现人员重复参保、骗保以及报销困难这些民生难题。



三是医保数据治理机制与应用体系不完善。没有做好底层的数据治理工作以及没有建成医保数据应用体系，都难以进行医保数据的共享与商业化价值变现。

2.2.2.2 医保数据治理途径

1) 打破信息孤岛，加强部门间数据连通

通过加强部门间的数据共享，实现数据来源的多渠道化。不仅医保和医疗等部门要建立信息数据的共享合作机制，同时要加强与税务、公安、民政等其他相关部门的合作，建立多渠道的数据归集共享机制。通过横向联通、纵向贯通方式实现医保业务信息共享与业务协同，为医疗保障信息平台提供数据支撑。

2) 完善数据标准，提升数据质量

国家局信息平台通过制定医保数据元规范，实现数据的标准化和数据规范管理，不断开展技术标准、业务标准、总体标准等标准规范的工作，最终建立形成以医疗保障信息平台为依托的医保数据综合标准库。统一药品编码和医疗服务及医疗器械编码，统一参保人数据采集标准，逐步实现全国医保数据信息的标准化，有效提升医保数据质量，为大数据应用打下坚实的基础。

3) 结合大数据技术，建设主题仓库形式的大数据平台

常州市医疗保障局依托常州市医疗保障信息系统平台，按照医保管理决策和业务应用要求，通过对海量大数据的收集、整合等，围绕医保核心管理决策和“三医”联动的目标和标准规范，建立按主题仓库形式存储的大数据平台，涵盖基础信息管理、医保费用管理、医疗服务质量、费用监管路径、医药价格、医药交易采购分析、信用评价管理等多领域，为规范医保业务服务流程、费用审核稽查方向和信息安全共享协同提供数据基础，有助于提高精算分析和管理决策的水平。

4) 建立医保数据安全应用管理机制

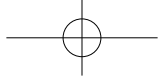
医保数据的安全是医保数据使用的前提，关乎每一个参保人的权益。医保管理部门要逐步探索建立医保数据的安全管理机制，在数据的收集、管理、使用方面出台具体的操作规程，厘清数据管理者和数据使用者之间的责任和义务。既要充分发掘医保数据在医保监管领域的价值，又要确保数据的安全。

2.2.2.3 山西省医保数据治理案例

1) 现状分析

目前医保行业已经沉淀海量的存量数据，且每年还将在持续的增长；数据分散在各地市，和人社其他系统数据在一起，没有统一割裂集中；建设厂商多，虽然参考核心平台三版标准，但在实施过程中各厂家差异标准不统一；从政策角度出发，打通各地市数据、建立省级集中、统一标准的数据中台，为挖掘医保数据的价值提供先决条件，为支撑医疗保障信息化建设打下坚实的基础。

山西省医保局数据分布在各个地市和省直，医保核心系统由7个核心供应商建



设,涉及 17 个库,标准不一,数据量大,数据治理工作任务繁重。现场的数据治理工具不稳定,性能与易用性欠缺。主要存在以下问题。

体量大。医疗数据体量巨大,一个省的医疗健康数据要以 TB、PB 量级存储和管理。

多态性。数据源是各种各样的,并且涵盖了多种类型数据,多种结构的数据对数据治理能力是一项重大考验。

不完整性。就医数据的收集和处理常常被分割,导致医疗资料库难以完整体现出所有相关信息。很多数据都来自手动记录,导致数据记录的偏差和残缺。科室、诊断信息等关键信息缺失量大,空值率高。

冗余性。医保每日会产生大批量数据,同一个人可以在不同的医院药店创建相同的信息;整个医学数据库包含许多重复的和无关的信息。无效信息缺少标识,造成数据冗余量大。300 万的参保人,库里却有 600 多万参保人信息。

时效性。数据生成速度非常快,更新速度非常高,许多数据的获取时间每周、每天、每分钟甚至每秒都在不断更新,因此,要求更高的响应速度和治理速度。

隐私性。数据隐私是医疗大数据的重要特征。医疗和卫生数据如疾病、诊断和基因数据等泄露将对个体造成负面影响,也会导致公民权利受到侵犯。

2) 数据治理平台

随着大数据时代的来临,各行各业开始认识到数据的价值。把数据视作宝贵的财富,已经成为业界的一种共识,企业也在快速探索应用场景和商业模式,并开始构建数据平台。如果在大数据“拼图”中遗忘了数据治理,那么技术投资就可能毫无意义。因为没有数据治理这一环节,其带来的后果往往是随处可见的数据不统一、难以提升的数据质量、难以完成的模型梳理、难以保障的数据安全等。源源不断的基础性数据问题会进一步产生,进而导致数据建设难以真正发挥其商业价值。因此,消除数据的不一致性,采用标准化的数据标准,提高数据管理能力,实现数据资源共享,并将数据作为公司在企业、管理和战略决策的宝贵资产使用,发挥数据资产价值变得迫在眉睫,数据治理呼之欲出。企业通过制定战略方向,建立组织架构,明确分工责任,控制数据风险,满足安全标准,提高业绩和增加价值,并提供创新的大数据服务。

数据治理平台是企业用户打造的一站式数据建仓和治理平台,由数据地图、数据血缘、数据标准、数据模型、数据质量、数据集成、数据安全、元数据管理等功能模块组成。目标是解决数据治理过程中数据不集中、标准不统一、质量不可控、交付周期长等一系列问题,通过标准化工具和智能化方法,将杂乱无章的基础数据治理成有价值的数据资产,实现数据的价值化和业务化。

数据治理平台依托 PostgreSQL 数据库加 Hadoop 混搭架构的大数据平台,提供海量数据的存储和大规模并行计算能力,支持 PB 级以上的数据存储和海量数据加工处理,实现海量存储和高效的数据处理。采取数据分层架构的方式,治理了 28 万个