基于数据空间的电子病历数据融合与应用平台

包小源[1,2]，张凯[3]，金梦[1,2]，谢双莲[3]，宋锴[3]
1. 北京大学医学信息学中心，北京 100191；2. 国家医疗服务数据中心，北京 100191；
3. 北京大学医学部，北京 100191

摘要
为了建立高效可扩展且易于管理的数据融合与应用平台，利用数据空间技术，按照数据敏感性将电子病历数据按照原始数据空间、匿名数据空间、模型数据空间的框架进行集成、融合，对匿名数据进行二次分析与挖掘，并针对各数据空间设计实现了不同的存储、安全保护、数据访问机制。平台已在国家医疗服务分析以及北京大学附属医院医疗能力、质量、效率的分析中得到应用。

关键词
电子病历；数据平台；数据空间；数据质量；数据脱敏

中图分类号：C931.6　　文献标识码：A　　doi: 10.11959/j.issn.2096-0271.2019049

A data-space based platform for the integration and application of electronic health records

BAO Xiaoyuan[1,2], ZHANG Kai[3], JIN Meng[1,2], XIE Shuanglian[3], SONG Kai[3]
1. Peking University Medical Informatics Center, Beijing 100191, China
2. National Health Care Data Center, Beijing 100191, China
3. Peking University Health Science Center, Beijing 100191, China

Abstract
In order to build an efficient, scalable and easy-to-manage data integration and application platform, using data space structure, electronic medical records were integrated in original data space, anonymous data space, and model data space according to data sensitivity, and anonymous data were used for data mining and secondary analysis. Different storage, security protection and data access mechanisms were designed and implemented for each data space. The platform has been applied in the analysis of national health care performance and the evaluation of medical capabilities, quality, and efficiency of affiliated hospitals of Peking University.

Key words
electronic medical records, data platform, data space, data quality, data desensitization

1 引言

我国电子病历的应用越来越广泛，使用电子病历数据进行临床研究、医院管理以及数据共享利用的研究越来越常见。做到数据收集、数据质量控制、数据分析处理、分析模型发布的“兼容差异、深入利用”，是承担国家医疗数据中心数据平台建设任务的基本要求。其中“兼容差异”规则是指在数据输入端，可以读入目前主流应用生成的数据文件格式，可以识别语义相容的数据内容，不同版本不同标准的数据（如疾病编码标准、手术编码标准、病历编码标准）都可以向一个版本进行映射与转换等；在输出端，则可以按照需求定制输出接口与输出格式，包括变量的定制、值的自定义等。“兼容差异”的规则主要用于应对我国由于各种实际系统建设、应用差异所导致的数据差异，最大限度地兼容各个医院的数据，并使之能在一个基准线上进行分析。同时，要对差异不大的数据（如病案首页）、差异较大的数据（如电子病历文档以及病例系统数据）进行区分处理，最大限度地提高处理效率。“深入利用”规则既要求设计能够集成、融合所有数据进行各个维度、各个层面的分析建模的平台，又需要平台的结构能够保护敏感数据，同时面向特定需求发布匿名数据，进而利用各种优质资源进行数据挖掘分析、二次利用，并将分析结果、模型也作为数据进行存储、管理。

2 国家医疗数据中心系统架构

国家医疗数据中心所获得的数据主要来自医院的不同数据源，包括病历系统、影像系统（PACS）、检验系统（LIS）等，数据类型也包括了文本、图像、视频等多种形式。对于同一家医院，数据可能经历多次迭代，期间可能有错误数据的替换、缺失数据的补充等数据层面的操作。为了管理分散、异构的数据，国家医疗数据中心建立了以数据空间技术为基础的三层结构，在层次内部，针对数据模式固定的数据采用数据仓库进行管理。

数据空间是与主体相关的数据及其关系的集合，主体、数据集、服务是数据空间的3个要素[1]。在数据模型上，内部的数据不依赖严格的数据模式，可以以一种松散的数据模式来组织[2]。在构建方式上，数据空间不需要提前提出所有可能的需求以设计合适的数据库模式，而是在演化过程中，根据新增的需求建立主体、数据集和服务三者之间的关系和逻辑，同时可以根据不断改变的需求，以较低的成本重新建立新的关系。数据空间包含围绕数据集提供的服务，可以对业务过程进行很好的分层和组织。

数据仓库是一系列具有继承性、主体性和持久性的数据集合[3]，与数据空间不同，数据仓库需要有固定的数据模式，对于数据的查询效率有很好的提升，但对于数据变化的适应比较迟钝[4]，因此国家医疗数据中心仅对一些有固定数据模式的数据（如病案首页）采用基于数据仓库的管理。

目前国家医疗数据中心主要提供数据集成、匿名化处理及数据查询与分析服务。为保证敏感数据的安全，从数据存储结构和结构内部脱敏操作两个层面进行了处理。根据涉及的数据的敏感性，通过物理隔离的3层数据空间进行数据管理，即原始数据空间、匿名数据空间、模型数据空间。

原始数据空间的数据集为直接从安全

通道获取的原始数据，这部分数据未经过任何脱敏操作，因此所有数据都以加密形式存储，并且有物理隔离和严格控制的访问策略。在这一层次主要进行数据清洗以及基本的数据有效性的校验，因此在这一层次的数据迭代次数是最多的。符合数据有效性检验的数据均视为合格数据，进行脱敏处理后，下发至匿名数据空间，使得数据迭代的成本降至最低。

匿名数据空间主要进行匿名数据的管理。首先去除相应字段，再使用训练好的机器学习模型识别自由文本中的敏感信息，予以去除。将经过脱敏的匿名化数据输入匿名数据空间，建立匿名数据库；提取的敏感数据被存储在与匿名数据空间有物理隔离的模型数据空间的敏感信息数据库中。在匿名数据空间中，部分数据（如病案首页）有较固定的数据模式，还需进行部分关键信息的抽取和加载，并存入数据仓库。

模型数据空间的数据集为下发的模型数据，根据用户的需求，将所需的数据下发至用户的虚拟空间，进行模型计算。模型数据空间整合用户的需求，同时，这些需求也进一步完善了各数据空间的数据组织和管理。

各层次的数据存储均使用多级存储机制，采用Hadoop开发团队开发的开源Hadoop分布式文件系统（Hadoop distributed file system，HDFS）。在不同的物理磁盘上保存至少3份数据的备份，以保证数据的可靠性。

总体而言，由于医疗数据格式多样，国家医疗数据中心主要采用数据空间技术进行数据管理，对于其中数据模式较为固定的部分，在层次内以数据仓库的方式进行管理，提升查询效率。

3层数据空间的功能如**图1**所示。

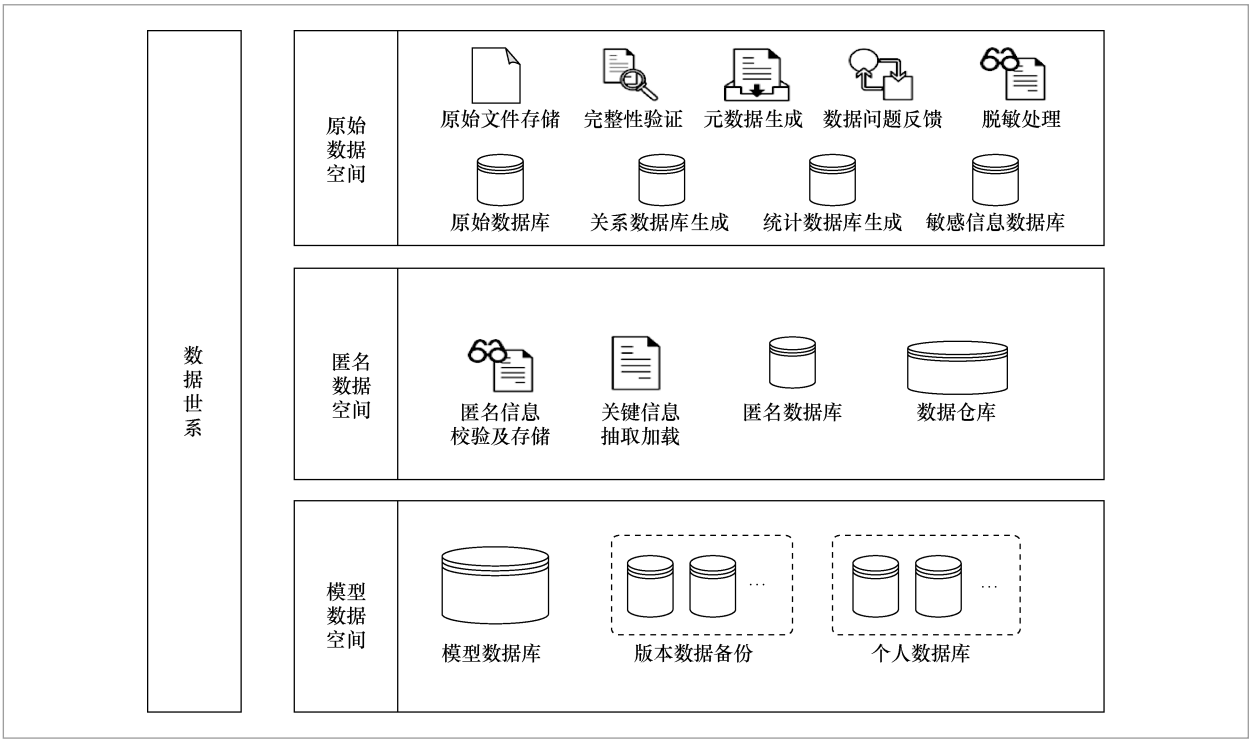


图1 3层数据空间的功能

3 基于数据空间结构和空间内功能的敏感信息保护

医疗数据涉及病人隐私，信息内容复杂，从安全通道获得的原始数据从自动清洗到数据分析与发布，涉及多个数据处理环节，每一环节所需的数据结构均不同，涉及的敏感数据也不同，需要不同的数据安全级别。因此本文提出基于数据空间的数据管理，不同数据空间存在物理隔离，数据空间之间的数据流动也有详尽的安全控制和日志记录。

对于敏感信息的保护，本文在两个层面上进行实现。一是从数据存储和管理结构上，根据数据的敏感程度，定义了3层数据空间：原始数据空间、匿名数据空间和模型数据空间；二是在数据空间中设置了多层次的脱敏处理及敏感信息的严格管理。除此之外，记录所有操作生成的数据世系也可对每步操作进行回溯。

3.1 基于3层数据空间结构的敏感信息保护

3.1.1 原始数据空间

原始数据空间处理和存储的数据集为直接从安全通道获得的原始数据经自动清洗和标准化转换后的结果数据集。这一数据空间中的输入数据包含了可识别身份的敏感数据，因此安全级别最高。在这一数据空间中，主要进行敏感信息的检测、提取，存储以及数据匿名化处理，输出匿名数据到匿名数据空间。本层数据空间存在物理隔离的数据层和应用层，这是由于在抽取敏感信息的过程中，国家医疗数据中心需要针对敏感信息进行必要的数据统计，统计结果存储于统计数据库中，并向部分经过严格安全审计的用户开放统计数据的查询功能，这一过程归入应用层的范围。

3.1.2 匿名数据空间

匿名数据空间包含匿名化处理、匿名化数据存储及匿名化数据下发过程，分为数据层及应用层。数据层主要执行匿名化数据存储和管理，应用层主要提供数据需求的审核及定制数据的下发。

3.1.3 模型数据空间

模型数据空间主要处理数据请求、下发数据至用户虚拟机以及对下发数据进行数据存储。不同数据请求单独建立数据库文件，通过安全通道下发至个人工作区，同时在数据备份存储空间备份。

3.2 数据空间的存储、安全与访问机制

3.2.1 存储机制

数据空间包含以下数据。

（1）各医院提交的原始数据

由于各医院病案室采用的文件归档系统不同（如DBase系统的DBF文件、Excel格式文件和CSV格式文件等），这部分数据经过自动清洗并生成元数据后，主要以文本文件形式进行存储。

（2）各数据层中的数据

这部分文件已经经过清洗，形成了完整的数据结构，因此主要以数据库形式进行存储，常见的格式有MySQL、SQL Server数据库文件格式。

（3）用户使用过程中生成的数据

这部分数据是用户对个人数据库操作产生的，主要以文件（如CSV）和数据库（如MySQL、SQL Server）形式存储。

在数据的存储模式上，首先根据各数据空间中数据的敏感程度进行物理隔离的数据分区，将3层数据空间的数据严格存储在不同的服务器集群中，设立不同的安全机制。在各数据空间内部，主要采用分区、分片的分布式存储方式。

在数据的分区上，对数据量大、集成度要求高而数据查询和分析又较为频繁的匿名数据空间的分区机制进行了较为详细的探索。在数据库层面，最频繁的查询有2种：第一种是按医院的多列数据查询与提取，用于DRG计算[5-6]、秩序列、TOPSIS[7]等模型的计算；第二种是按主要疾病分区的数据查询与提取，由于主要疾病频数的差异较大，因此在分区时需要考虑将频数在前10位（或100位）的疾病按照历史数据统计结果进行分区策略的动态调整。

基于这2种查询模式，通常以医院和主诊断来进行分区，其中医院节点数目相对小，而主诊断的节点数目较大。在分配主分区键和次分区键时，常见的方法有2种：第一种是以医院为主分区键，以主诊断为次分区键；第二种方法是以主诊断为主分区键，以医院为次分区键。从并行计算的角度考虑，越分散查询效率越高，但网络开销也会相应增大，此时要根据具体的需求平衡网络开销和查询效率，例如提取某个医院的某个疾病时，在集群中可能只会集中在一台机器上，可能会导致查询效率下降；而在模型计算时，一般的查询会分布在多家医院，因此查询会被分发到不同节点上去。2种方法在网络开销和查询效率上各有优劣，应注意其中的平衡点。主诊断数目相对节点数目庞大得多，需要专门配置映射文件，对分区进行映射转换后进入数据库。

3.2.2 安全机制

由于3层数据空间本身是根据数据的敏感性划分的，因此对于各层数据空间，本文设立了不同的安全机制，其中原始数据空间的安全级别最高，模型数据空间的安全级别最低，各层数据空间之间保持物理隔离。

在原始数据空间中，网络层面运行在与其他空间物理隔离的计算机集群上，用户认证等方面则从严格的审计机制、操作日志记录机制等多角度实现对原始数据的完全隔离。查询、处理等均局限于数据库，而文件则经过加密压缩后，密码文件独立存放，非特殊权限或特殊原因不再打开或提取。

在模型数据空间中的安全保障机制方面，本文为每个用户分配相互隔离的虚拟机，用户以虚拟桌面的方式登录，以实现个人数据的独立、安全。针对每个用户提供不同的数据，在个人虚拟机上实现不同的应用，以解决整个平台上多用户的不同需求。

3.2.3 访问机制

在访问内容上，本文只提供对数据库的访问，各医院上传的原始数据文件均不开放对外访问权限，数据库访问主要以B/S结构查询。传统关系查询可以使用Oracle BI等平台型工具，将原始数据作为后台数据模型，直接将一些可以维度化的列建立为维度，在此模型下，直接用OBIEE客户端对相关数据进行查询、展现即可。元数据查询也会提供B/S查询接口，但只开放基本的统计数据，目前包含医院上传数据的问题、反馈次数、修改问题而带入的新问题等。关键字查询的接口依然是B/S结构，但其查询结果以表关联的方式返回，在该表上可以查询对应的数据条目。

模型数据空间中的访问接口与其他两个数据空间没有很大的区别，只是在

用户的数据权限(列、行、导出、计算、数据总量)方面,需要在大数据虚拟语言环境模型(model in virtual language environment of big data, MVLB)中进行监控,并记录实际操作序列等数据。由于访问方式在接口方面区别不大,本文在MVLB环境中的入口访问集群框架设计方面,采用了相同架构、面向不同需求的定制化配置部署方式。

3.3 数据空间多结构数据集成与敏感信息保护

3.3.1 多结构数据集成

国家医疗数据中心获取的数据类型多样,囊括了关系数据、半结构化数据以及非结构化数据(基于openEHR修正模型的集成逻辑框架),而在原始数据空间中,最重要的技术是对多结构的数据进行集成。

数据集成的方法主要有2种:全局视图方式和局部视图方式。考虑到病案首页的格式是中华人民共和国卫生部规定的标准格式,虽然各地区对具体内容会有所调整,但其数据结构具有相对稳定性。本文采用了全局视图的方式(即各医院病案首页数据模式向全局数据模式映射的方式),其步骤包括目标模式确定、数据收集、源包装器构造、并行集成执行及结果数据集的合并等[8]。

数据空间具有数据组织松散的特征,使用索引和映射查询数据较为低效。多数据集成针对数据空间中结构化较好、查询频繁的数据建立数据仓库,利用数据仓库查询速度快的特点,提升数据查询效率,实现高效、准确的数据查询。对于数据空间中存储的电子病历文本数据,本文采用关系数据库(SQL server)存储并建立全文索引,以实现对病历文本的检索。

3.3.2 敏感数据提取和匿名化处理

首先参照敏感信息条目和国家电子病历数据接口标准,提取原始数据中涉及个人信息的数据,将这一部分数据定义为敏感数据,用于后续的操作。敏感信息条目的制定参考了美国HIPAA法案[9]、国家标准GB/T 35273-2017《信息安全技术个人信息安全规范》以及相关文献提及的敏感数据条目,并人工复核了医院上报的数据,最终确定了包含个人信息(如姓名、年龄、联系电话、详细地址等)、病历识别号(如医保卡号、病历号、影像号等)、就诊详细日期(如入院日期、手术日期、出院日期)、就诊过程隐私数据(如床号、主治医师姓名、手术医师姓名等)在内的多项数据。然后对上交的包含自由文本的字段进行脱敏处理,在这一步,本文使用已有的机器学习方法,对数据进行两遍扫描,第一遍进行元素值的特征计算,第二遍将数据分为敏感信息和非敏感信息,并去除敏感信息。

匿名化数据还要进行重新识别风险的评估。每次有新的数据源加入后,都进行一次全面的评估。在日常使用时,定期随机抽取数据,以评估重新识别的风险,根据重新识别患者所需要结合的字段数来评估数据的安全性。

3.3.3 敏感数据关联机制

将匿名数据空间中提取的敏感信息存入敏感信息数据库后,会返回与存入信息对应的唯一ID,将此ID作为识别码与提取的敏感信息一并存入匿名数据库,建立匿名数据库与敏感数据库的关联。识别码不作为可下发字段,仅在有特殊需求时,作为与敏感信息数据库关联的方式。在评估特殊需求时,要根据

计算结果是否返回敏感信息进行严格的评估和审核。

3.4 数据世系的生成与查询

在数据世系信息的生成、查询及管理方面，目前比较关心的是每一个处理步骤都抽取了哪些数据、有多少量以及结果存储在哪里，因此针对每个中间结果集，都要记录其查询语句并进行反向计算，以便追踪到起点或其前驱处理节点的信息。目前采用查询语句与查询结果一一关联映射的方式实现数据世系的管理。为实现数据世系的自动生成，需要在Perl或其他高级语言的基础上加一层命令解析器，这样，每一次查询及其结果都会被写到日志中，之后的数据世系信息均以专门的解析器抽取日志文件的方式形成。每个处理模块完成任务处理后，都需要运行自动的日志信息处理语句，其目的是识别原始程序中的查询语句、查询输出目标、查询输入、当时运行该数据处理的程序本身等，然后在原始程序的特定位置，增加输出到日志文件的语句，这样做的优点是数据处理本身会专注于业务处理，而日志输出等常规、普遍性要求都会通过系统来自动完成。

在原始数据空间中，只提取匿名数据进入匿名数据空间的过程也需要将查询处理和处理结果的对应关系记录下来，整个过程参照数据世系模型、数据集成指令（包括选项）的类型，进行业务数据世系的内容生成。

在模型数据空间中，通过基于环境支撑层对处理工具中嵌入处理日志的强制记录方式来实现个人空间的数据世系信息生成。另外，模型空间的处理定制化需求非常明显，而处理方式非常复杂，因此目前在MVLB中，将数据世系的记录方式简化为输入数据、处理程序源码（或指令序列）、输出数据。

3.5 数据流动过程及处理流程

通过安全通道获得的原始数据在原始数据空间中进行数据清洗、入库，形成关系数据，并下发至匿名数据空间，在匿名数据空间中进行匿名化处理，提取敏感信息，并保存匿名化数据。经过审计的用户提出数据需求后，被提取的匿名数据下发至模型数据空间。如果用户获得了随访数据查询的许可，必要的敏感数据也将从敏感数据库下发至模型数据空间。

数据在数据空间中的所有操作日志都被记录在以数据空间为主键的日志数据库中，便于生成直观的数据世系信息。整体系统框架及处理流程如**图2**所示。

3.5.1 原始数据空间框架及处理流程

在原始数据空间中，通过安全渠道获取的数据经过定制的数据包装器框架，将文本、电子表格、数据库文件、XML等格式的文件转化为可识别和导入的数据格式，以文本形式插入输入数据库。这一步需要验证数据的完整性，对于缺失必填项的文件，则只存入元数据存储空间备份，而不做导入操作，待相关医院重新上传补充缺失项的文件后，再导入数据库。完整的数据文件导入输入数据库后，原始文件经过强密码加密，存入元数据存储空间。

进入输入数据库的数据将经过进一步的数据清洗，首先根据国家医疗数据中心发布的数据接口标准对数据列定义进行数据类型的验证和转换，对于不符合定义数据格式的数据，必要时要求相关医院进行自查和重新上传。经过数据格式转换的

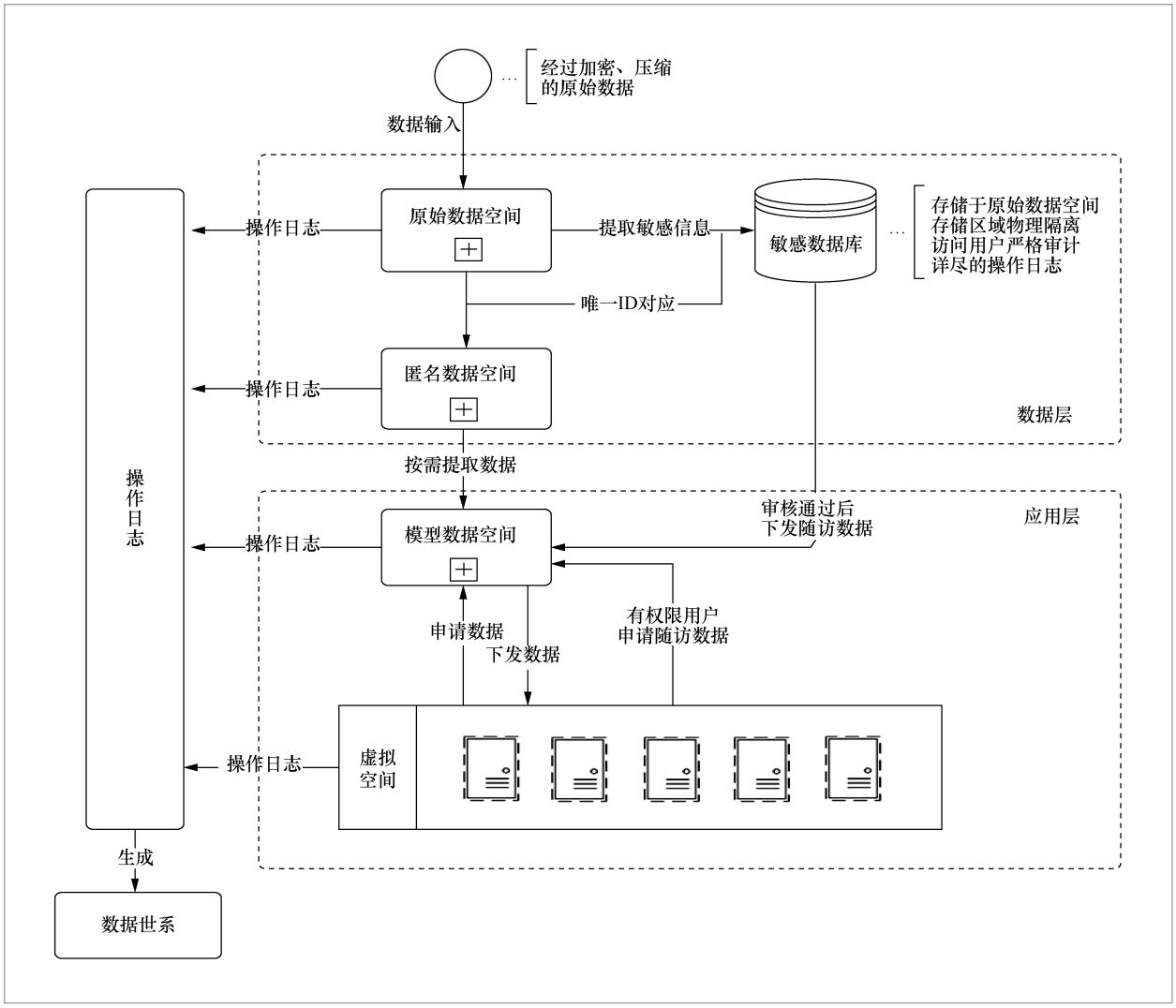


图 2 整体系统框架及处理流程

数据，将根据数据接口中对各部分数据的定义，建立关系数据表，形成多维度的数据，保存于原始数据关系数据库，并进一步进行数据匿名化处理。除根据数据列定义去除涉及个人信息的数据列外，还对包含自然语言的文本使用深度学习识别姓名、地名等信息，并进行脱敏处理。将敏感信息存入敏感信息数据库，生成唯一对应的ID，并将此ID与非敏感信息下发至匿名数据空间。

有关原始数据的一些必要的统计信息被存入统计数据库，供有权限的用户通过查询系统进行查询。原始数据空间框架及处理流程如图3所示。

3.5.2 匿名数据空间框架及处理流程

匿名数据空间主要进行匿名数据的存储与管理，将原始数据空间下发的脱敏数据存入匿名数据库，并在此层进行模式固定的数据的集成。同时，可以通过敏感数据ID在模型数据空间中查询

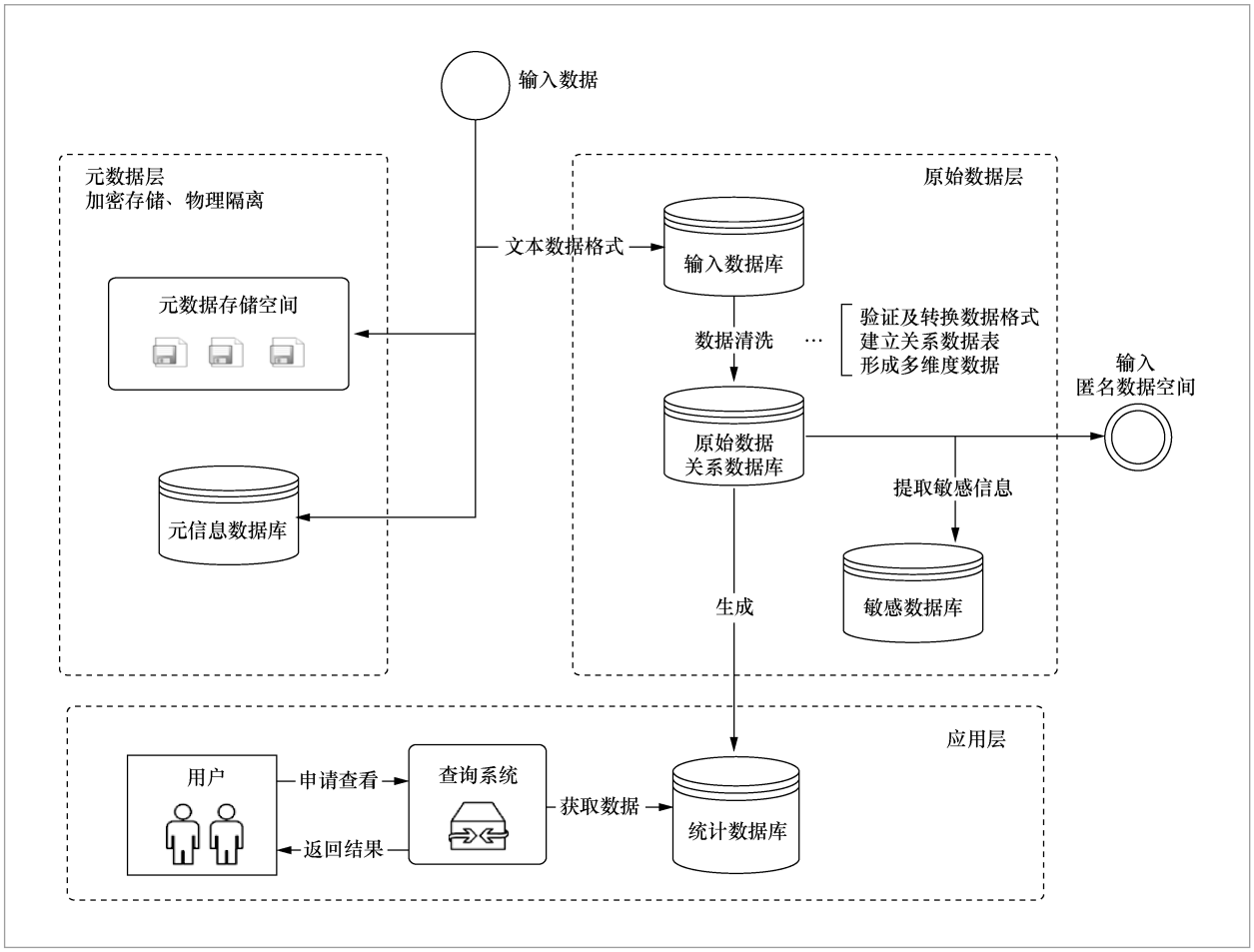


图 3 原始数据空间框架及处理流程

原始数据。

用户通过模型数据空间向匿名数据空间发出的数据下发请求，此请求在应用层得到处理。在查询需求通过审核后，按照申请的新数据字段，生成需要下发字段名和数据列列表，根据此列表，从匿名数据库中提取相应的数据，记录日志并生成新版本号，将以版本号命名的数据作为模型数据空间的输入数据。匿名数据空间框架及处理流程如**图4**所示。

3.5.3 模型数据空间框架及处理流程

在模型数据空间中，用户个人提出数据申请后，会在初步审核后生成包含所需字段名的请求，并提交给匿名数据空间处理。在模型数据空间进行的初步审核主要审核用户是否具有获取该字段的权限。当匿名数据空间通过审核，确定可以提供相关数据列，并下发数据后，数据首先存入模型数据库，并备份至数据备份存储空间，随后下发到用户的虚拟机上。

用户可以在虚拟机上从请求的数据库中提取需要的数据，并存入虚拟机的个人数据库进行处理。其中，提取的数据也记录操作日志，以实现数据世系的追踪。模型数据空间框架及处理流程如**图5**所示。

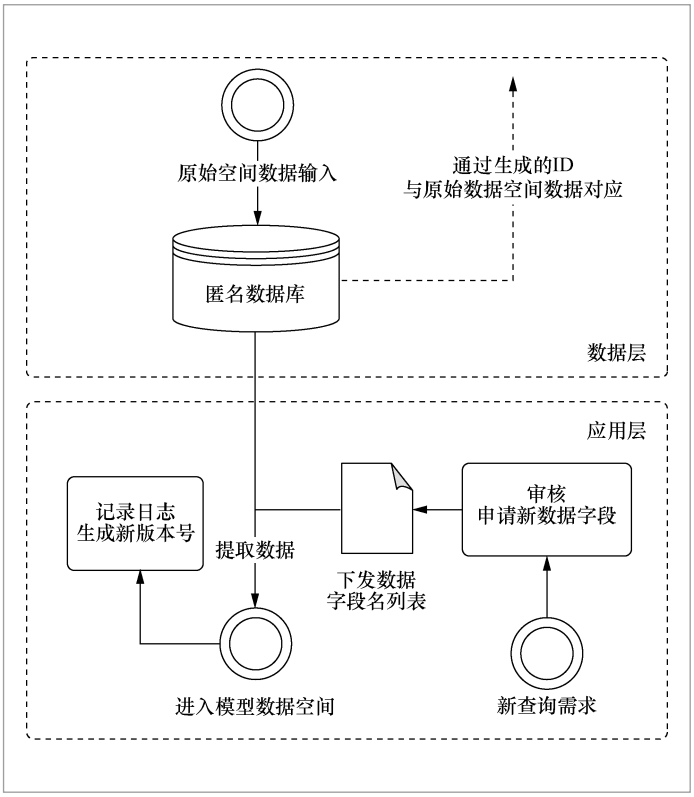


图 4 匿名数据空间框架及处理流程

4 系统运行情况

国家医疗数据中心利用3层数据结构已经平稳运行6年，3层结构业务及产出如图6所示。原始数据层已经拥有成熟的数据接口工具，而对于未标注使用接口标准的数据，也已有了用于判断数据接口标准的模型，国家医疗数据中心共收集并整合了全国总计500余家医院的数据。在匿名数据空间脱敏的过程中，形成了用于数据脱敏的匿名语料库和匿名知识库。

对外发布的数据包括根据匿名数据空间及原始数据空间计算的数据质量报告以及模型数据空间用户训练的模型。自2013年以来，已经完成1 600余份质量报告的发布。通过模型数据空间提取和处理的数据，已经提供给了DRG模型、临床分层评价模型进行计算。

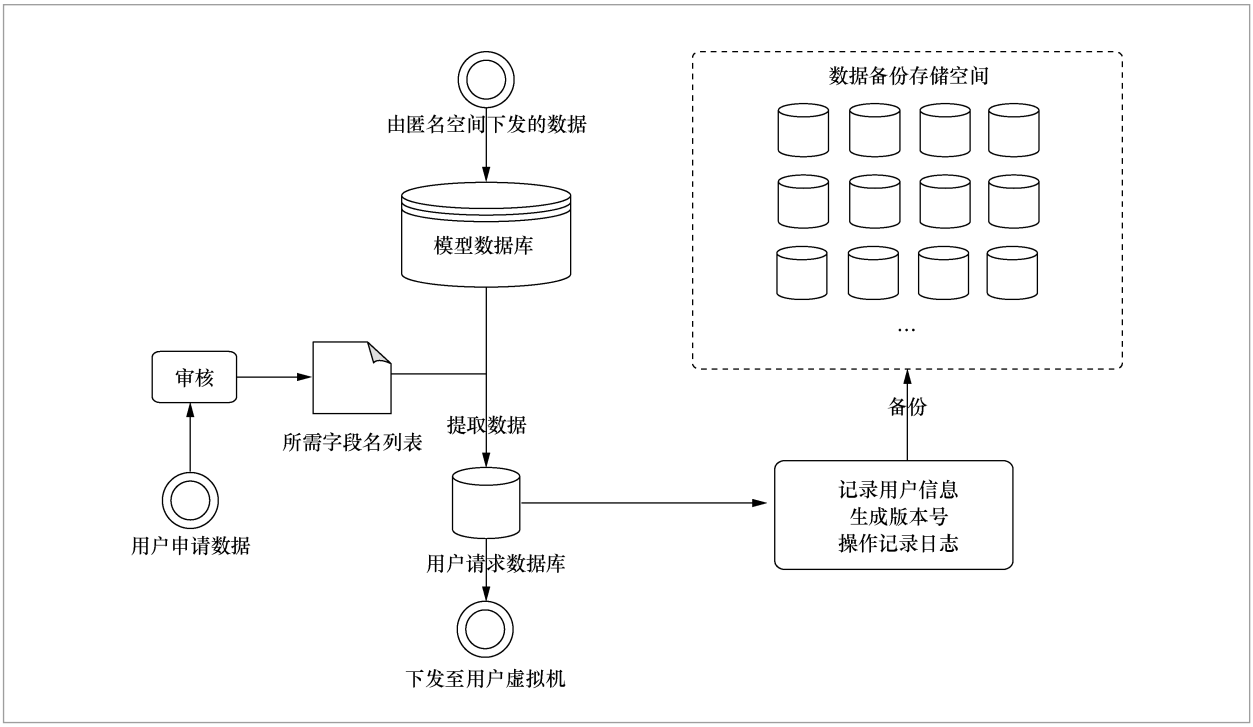


图 5 模型数据空间框架及处理流程

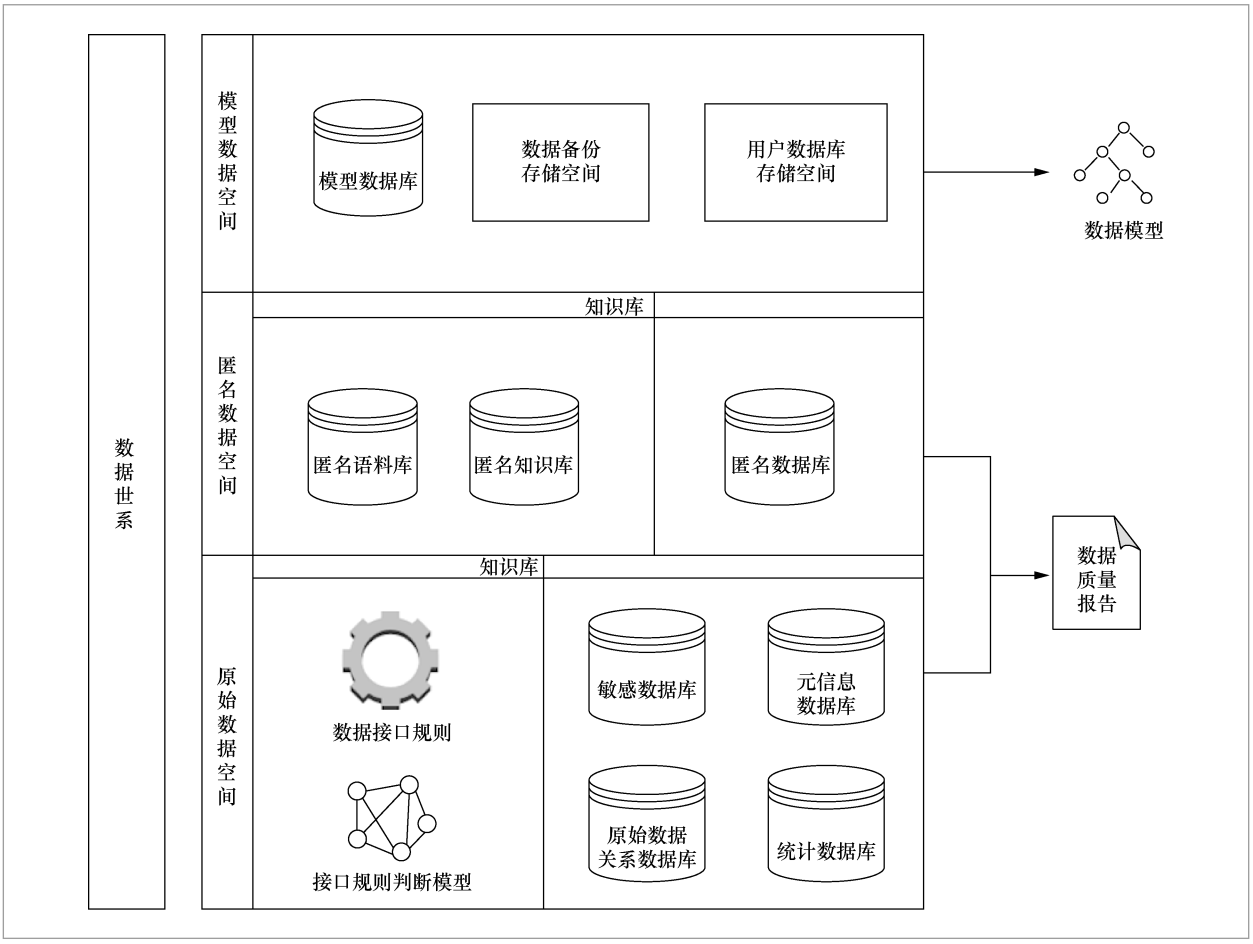


图 6 3 层结构业务及产出

本文使用数据世系结构[10]来表示数据产生和数据演变的过程，追溯模型数据空间计算结果的原始数据。数据世系作为表示数据演变的技术，被广泛应用于互联网及物联网大量的数据管理中[11-13]，用于追踪数据的演变过程[14-15]。根据用户使用数据库的版本号，首先可以在数据备份存储空间中找到原始数据，如果发现数据有问题，可以从数据世系中找到匿名数据库下发数据的时间、内容及版本号，判断在下发时间点后有无数据更新，还可以进一步通过匿名数据版本号，对应到原始数据空间中的数据。如果确认数据有误，可反馈给相应医院进行数据的迭代更新。

5 系统结构的设计思路

在架构选择方面，国家医疗数据中心主要采用了基于数据空间的数据结构。当前有一些开源的医疗数据存储解决方案，如应用比较广泛的架构openEHR，该架构是由国际openEHR组织于1999年提出的开放式电子健康档案规范，它采用由参考模型和原型模型组成的两层结构，以实现医疗领域知识和实际临床信息的分离，使信息模型具有高可扩展性[16]。一些研究

者已经论证了国家医疗数据中心采用基于openEHR架构的系统设计的可行性[16-17]，证实了openEHR是一种部署起来较为便利的有更强语义互操作性的系统[18-19]。

国家医疗数据中心收集的数据是各医院提交的临床数据，其关键不在于建立内容的逻辑关系，而在于如何存储管理已有数据，进行进一步处理、分析及发布。由于openEHR更关注内容逻辑，对于数据的内容敏感度没有严格的划分，使得数据匿名化和发布面临较大困难，因此，本文并没有选择以openEHR架构为基本框架，而采用了能够更好地体现数据敏感度的基于数据空间的3层结构，以较好地区别管理原始数据和匿名数据。

从传统数据集成的角度考虑，传统的数据库管理模式一般需要在整体设计、全面标准化的基础上，从数据源到目标平台进行完整的设计，包括数据抽取、清洗、加载，并存放于标准的数据仓库中。而数据空间管理与传统的数据管理有以下4个区别：一是数据空间需要支持所有类型的数据；二是数据空间提供数据更新的能力，因此不像传统数据库对数据有完全的控制能力；三是对于数据查询的需求，数据空间只能根据数据的情况返回最好的结果，而不一定都能返回准确的结果；四是数据空间需要有数据集成的能力[20]，数据空间还可以将用户反馈加入数据管理的过程中，使得数据空间可以不断演化，满足更多的需求[21]。

在业务相对成熟的行业，使用传统数据管理模式是非常有效的。但是，就医疗行业本身而言，其收集的数据不仅包含大量的数据类型，已收集的数据也可能有部分数据列缺失的情况（但此时非缺失的数据已经可以用于分析）[22]，而且随着学科发展而新出现的诊疗会呈现出新的数据内容、数据格式等（譬如近年来兴起的基于基因技术的精准医疗就产生了大量的基因数据），加之对数据的需求也更加具体和复杂，在建设大数据平台时需要遵从pay-as-you-go的方式进行，即边建设、边应用、边改进、边融合，进行渐进的、螺旋式的数据平台建设。因此，在医疗行业使用数据空间管理，是更加符合实际情况的。

在设计系统结构时，本文主要考虑数据敏感性。由于个人的医疗数据具有独特性，在匿名化过程中不仅需要考虑去除明确的涉及患者隐私的数据列，还要考虑重新识别的风险，即使用者通过结合多个数据列识别出患者的风险。例如根据患者在既往史和现病史中披露的就诊医院、时间和所做手术就能较准确地识别出患者。因此在系统设计上，应该考虑控制匿名化数据重新识别的风险。参考文献[23]讨论了评估系统重新识别风险的3个方面：数据接收方的数据安全性、数据泄露对病人隐私侵犯的程度以及数据使用方重新识别患者的收益。而对于医疗数据，显然数据泄露对病人隐私侵犯程度是极高的，因此设计系统架构时需要严格控制接收方数据的安全性，通过提高重新识别的成本来降低重新识别的收益。本文使用3层数据空间的结构，针对接收方数据的安全性，使用模型数据空间来管理用户及用户数据，以实现对用户数据安全性的完全掌控；针对提高重新识别的成本，则采用对匿名数据空间进行匿名化和按需下发数据来解决。

使用数据空间来管理数据也呈现出了一些问题。由于在数据检索和计算时不一定能返回准确的结果，数据空间具有一定的不确定性，同时查询效率也不如传统数据管理模式高。针对这一问题，本文将部分数据模式固定的数据集成在匿名数据空间的数据仓库中，解决了部分常用数据的查询效率问题。另外，由于数据空间具有数据优先、淡化模式的特点，数据质量也

有所下降。本文在模型数据空间进行了基本的数据质量控制，但是有些数据问题在模型计算时才显现，笔者仍然将这部分数据视为合格数据，将反馈后更新的数据视为这些数据的新版本进行管理。总之，使用数据空间作为医疗数据管理的主要技术是符合实际情况的，因为数据空间在保护了敏感数据的前提下，提供了更多二次利用的可能。它提供的pay-as-you-go的模式，可以容纳由于学科进步、信息化水平提高而产生的新的数据。对于部分成熟的数据模式，还可以在数据空间内用数据仓库进行优化，能够最大化地从数据中获取信息。

6 结束语

从国际、国内大数据应用的趋势考察，笔者发现大集成和大融合是临床数据管理的基本模式，而专项、细分的定制化分析与挖掘则是数据利用的基本方式。本文基于数据空间所构建的数据平台正是顺应了这一基本趋势。大集成和大融合在原始数据空间、匿名数据空间完成，而定制化分析则在模型数据空间中实现个性化支撑。

下一步将对智能数据管理方法做进一步探索，实现平台对数据质量控制、数据集成融合、数据脱敏、基本数据分析的智能赋能，建立基于分类自治的索引框架，支持高效查询，进一步提高平台管理的效率，实现个人数据空间的易用性。

参考文献:

[1] 李玉坤, 孟小峰, 张相於. 数据空间技术研究[J]. 软件学报, 2008, 19(8): 2018-2031.
LI Y K, MENG X F, ZHANG X Y. Research on dataspace [J]. Journal of Software, 2008, 19(8): 2018-2031.

[2] MIRZA H T, CHEN L, CHEN G. Practicability of dataspace systems[J]. International Journal of Digital Content Technology, 2010, 4(3): 233-243.

[3] 屈景辉, 廖琪梅, 许卫中, 等. 医学信息数据库的建立与数据挖掘[J]. 医学争鸣, 2001, 22(1): 88-89.
QU J H, LIAO Q M, XU W Z, et al. Technology on establishing and mining the data warehouse of medical information[J]. Negative, 2001, 22(1): 88-89.

[4] 王珊, 王会举, 覃雄派, 等. 架构大数据：挑战、现状与展望[J]. 计算机学报, 2011, 34(10): 1741-1752.
WANG S, WANG H J, QIN X P, et al. Architecting big data: challenges, studies and forecasts[J]. Chinese Journal of Computers, 2011, 34(10): 1741-1752.

[5] 白玲. 基于DRGs指标的综合医院绩效评价TOPSIS模型分析与验证[J]. 中国卫生统计, 2016, 33(5): 839-841.
BAI L. Analysis and verification of TOPSIS model of general hospital performance evaluation based on DRGs indicators[J]. Chinese Journal of Health Statistics, 2016, 33(5): 839-841.

[6] 郑小虹. 北京DRGs系统的研究与应用[M]. 北京: 北京大学医学出版社, 2015.
ZHENG X H. Research and application of Beijing DRGs system[M]. Beijing: Peking University Medical Press, 2015.

[7] 毛瑛, 王雪, 何荣鑫. 基于TOPSIS模型的公立医院医疗服务能力评价研究[J]. 中国卫生质量管理, 2016, 23(6): 99-103.
MAO Y, WANG X, HE R X. The TOPSIS model based evaluation of the medical service ability of public hospitals[J]. Chinese Health Quality Management, 2016, 23(6): 99-103.

[8] 包小源, 俞国培, 李岩, 等. 病案首页数据分布式集成管理平台的设计与应用[J]. 中国医院管理, 2014, 34(5): 30-32.
BAO X Y, YU G P, LI Y, et al. Design and application of distributed integration management of dada in the home page of medical records[J]. Chinese Hospital Management, 2014, 34(5): 30-32.

[9] BENITEZ K, MALIN B. Evaluating re-identification risks with respect to the HIPAA privacy rule[J]. Journal of the American Medical Informatics Association, 2010, 17(2): 169-177.

[10] 岳昆, 刘惟一, 朱运磊, 等. 一种基于概率图模型的不确定性数据世系表示方法[J]. 计算机学报, 2011, 34(10): 1897-1906.
YUE K, LIU W Y, ZHU Y L, et al. A probabilistic-graphical-model based approach for representing lineages in uncertain data[J]. Chinese Journal of Computers, 2011, 34(10): 1897-1906.

[11] 窦婷, 卢菁. 混合云环境下利用世系保障数据一致性的研究[J]. 计算机应用研究, 2015, 32(1): 108-111.
DOU T, LU J. Data provenance based consistency service for hybrid cloud[J]. Application Research of Computers, 2015, 32(1): 108-111.

[12] 聂娟, 孙瑞志, 邓雪峰, 等. 基于数据世系管理的精准农业不确定性复杂事件处理[J]. 农业机械学报, 2016, 47(5): 245-253.
NIE J, SUN R Z, DENG X F, et al. Uncertain complex event processing in precision agriculture based on data provenance management[J]. Transactions of the Chinese Society for Agricultural Machinery, 2016, 47(5): 245-253.

[13] 黄庆宇, 卢珞先. 基于数据世系的微博信息管理与检索算法研究[J]. 计算机科学, 2015, 42(10): 198-201.
HUANG Q Y, LU L X. Provenance based information management method for microblog messages[J]. Computer Science, 2015, 42(10): 198-201.

[14] 徐飞, 高济. 一种通用的数据追踪系统的实现[J]. 计算机工程与应用, 2003, 39(34): 191-193.
XU F, GAO J. Implemention of a common data tracing system[J]. Computer Engineering and Applications, 2003, 39(34): 191-193.

[15] 王黎维, 黄泽谦, 罗敏, 等. 集成对象代理数据库的科学工作流服务框架中的数据跟踪[J]. 计算机学报, 2008, 31(5): 721-732.
WANG L W, HUANG Z Q, LUO M, et al. Data provenance in a scientific workflow service framework integrated with object deputy database[J]. Chinese Journal of Computers, 2008, 31(5): 721-732.

[16] 刘骏健. 基于openEHR的临床数据中心设计与实现[D]. 浙江大学, 2016.
LIU J J. Design and implementation of clinical data repository based on OpenEHR[D]. Hangzhou: Zhejiang University, 2016.

[17] DEMKSI H, GARDE S, HILDEBRAND C. Open data models for smart health interconnected applications: the example of open EHR[J]. BMC Medical Informatics and Decision Making, 2016, 16(1): 137.

[18] GONZALEZFERRER A, PELEG M, MARCOS M, et al. Analysis of the process of representing clinical statements for decision-support applications: a comparison of open EHR archetypes and HL7 virtual medical record[J]. Journal of Medical Systems, 2016, 40(7): 1-10.

[19] MIN L T, TIAN Q, LU X D, et al. An openEHR based approach to improve the semantic interoperability of clinical data registry[J]. BMC Medical Informatics and Decision Making, 2018, 18(1): 15.

[20] HALEVY A, FRANKLIN M, MAIER D. Principles of dataspace systems[C]// The 25th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, June 26 - 28, 2006, Chicago, USA. New York: ACM Press, 2006: 1-9.

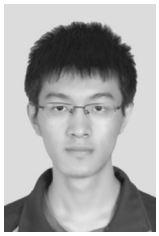
[21] JEFFERY S R, FRANKLIN M J, HALEVY A Y. Pay-as-you-go user feedback for dataspace systems[C]// International Conference on Management of Data, June 10-12,2008, Vancouver, Canada. [S.l.:s.n.], 2008: 847-860.

[22] HOFFMAN S, PODGURSKI A. Big bad data: law, public health, and biomedical databases[J]. The Journal of Law, Medicine & Ethics, 2013, 41: 56-60.

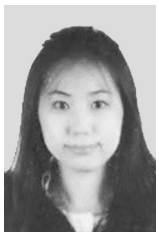
[23] El E K. Methods for the de-identification of electronic health records for genomic research[J]. Genome Medicine, 2011, 3(4): 25.

作者简介

包小源（1971- ），男，博士，北京大学医学信息学中心、国家医疗服务数据中心总工程师，主要研究方向为临床文本数据挖掘。



张凯（1996- ），男，北京大学医学部博士生，主要研究方向为临床医学、临床数据管理。



金梦（1986- ），女，北京大学医学信息学中心、国家医疗服务数据中心工程师，主要研究方向为医学信息学。

谢双莲（1996- ），女，北京大学第五临床医学院本科生，主要研究方向为临床医学、临床数据管理。

宋锴（1997- ），男，北京大学中日友好临床医学院本科生，主要研究方向为临床医学、临床数据管理。

收稿日期： 2019-08-10