文 献 与 数 据 学报
第 2 卷 第 4 期 2020 年 12 月

应用区块链构建健康医疗大数据溯源机制 *

郭少友 胡斐然
（郑州大学信息管理学院，郑州 450001）

摘 要：[**目的 / 意义**] 区块链具有去中心化、安全性高、可追溯性强等特征，将其引入到健康医疗大数据溯源中，有助于增强溯源过程的安全性和溯源结果的可信性。[**方法 / 过程**] 采用 PROV-O 本体、FHIR Provenance 溯源模型、健康医疗区块链数据模型等来描述数据，采用非对称加密、数字摘要、数字签名来保障溯源过程中的数据安全与隐私。[**结果 / 结论**] 所提出的溯源机制涉及溯源体系架构、数据描述、数据安全与隐私控制、溯源流程等几个方面，可实现对健康医疗大数据的演变历史和操作历史的多级可信溯源。

关键词： 区块链 健康医疗大数据 数据溯源
分类号： G203 TP391.1
DOI： 10.31193/SSAP.J.ISSN.2096-6695.2020.04.02

0 引 言

健康医疗大数据是国家重要的基础性战略资源，其中包含大量的个人隐私信息和国家机密，相关责任单位应确保数据的采集、加工、存储、访问、删除等各种操作都在可控的范围内进行，所有数据的来源和去向都可查询、可追溯[1]。为此，卫生健康主管部门应联合医疗机构、医保机构等共同建立一套健康医疗大数据溯源体系。区块链是一种按照时间顺序将不同机构的数据组织成区块（Block）、并将区块以链表形式组织起来的技术，且利用密码学原理保证数据的隐私安全和不可篡改。如何在上述溯源体系中引入区块链技术，利用其安全性高、可追溯性强等特征来实现健康医疗大数据的可信溯源，是一个值得深入研究的问题。

* 本文系国家社会科学基金项目“基于区块链的个人医疗信息语义组织与多粒度可信溯源方法研究”（项目编号：20BTQ063）的研究成果之一。
[作者简介] 郭少友（ORCID：0000-0001-7735-4542），男，教授，研究方向为网络信息资源管理，Email：guoshy@zzu.edu.cn；胡斐然（ORCID：0000-0003-0076-656X），男，硕士研究生，研究方向为网络信息资源管理，Email：hufeiran20@163.com。

1 相关研究

1.1 数据溯源

国内外学者对数据溯源这一研究主题的关注基本上始于 2000 年，涉及数据溯源的概念、描述模型、方法、应用领域等诸多方面。

（1）概念。目前有两类观点：①数据溯源是对数据在整个生存周期内的演变信息和演变处理内容的记录[2]；②数据溯源是对数据谱系以及与再现数据谱系相关的输入、实体、系统、过程等额外数据的记录[3]。前者强调数据的演变历史，后者同时兼顾数据的演变历史及相关额外数据。

（2）描述模型。较为成熟的溯源数据描述模型主要有三个：PROV 模型[4]、OPM 模型[5]、ProVO C 模型[2]。从本质上看，上述三个模型是相似的，都包括三类核心概念：实体，如网页、图表、数据集等；代理，如人、机构等；活动，如生成、翻译、删除等。

（3）方法。现有研究所提出的溯源方法可归纳为两类：基于标注的方法和基于非标注的方法，前者需要通过标注来记录原始数据的一些重要上下文信息[6]，后者只在需要数据溯源时才进行计算，通过构造逆置函数进行反向查询来获得数据的溯源信息[7]。

（4）应用。数据溯源可应用于对数据真实性要求比较高的各个领域，如生物、历史、考古、天文、健康医疗等[8]。戴超凡等进一步将其归纳为三个方面：在数据库中的应用、在工作流中的应用、在其它方面的应用[9]。

此外，还有一些学者对大数据溯源面临的挑战、机遇等问题进行了分析[10-11]。

1.2 基于区块链的健康医疗数据溯源

已有一些研究成果针对将区块链技术应用于健康医疗数据溯源进行了探讨。

（1）适用范围。借助于区块链的不可篡改特性，可实现医疗设备的跟踪、药品追踪、临床试验数据和医疗保险数据的溯源[12]。通过文献检索发现，在利用区块链实现健康医疗数据溯源方面，关于药品溯源的相关成果占大多数。

（2）溯源系统设计与实现。现有的第二代区块链系统（如以太坊）已支持各种应用开发，一些研究者已在此基础上设计、实现了溯源系统，如禹忠等设计了一种基于区块链的医药防伪溯源系统，消费者可查看包括药品生产信息、物流信息及使用信息在内的全部溯源信息[13]；Marbouh D 等设计了一个基于区块链的 COVID-19 疫情数据跟踪系统，可以对人员、捐款、疫情爆发情况、防疫用品物流等数据进行跟踪[14]。

（3）溯源模型构建。可以将区块链技术与现有的数据溯源模型（如 PROV 模型）、现有的健康医疗数据标准（如 HL7 CDA）相结合，构建面向健康医疗领域的溯源模型。Masi M 等提出一个基于区块链和 PROV 模型的健康数据溯源模型[15]；Margheri A 等将分布式电子病历系统与 W3C PROV、HL7 CDA 等标准结合起来，提出一种基于区块链的去中心化医疗数据溯源模型，可以对 PDF、CDA、CCDA 格式的文档进行溯源[16]。

（4）实现途径。从现有文献来看，查阅分布式日志、调用智能合约是实现数据溯源的两种主要途径。Tanaka K 等提出一种跨机构电子病历数据溯源架构，通过记录分布式日志数据的索引信

息来实现电子病历数据的溯源查询和分析[17]；Maslove D 等对临床试验各个阶段的元数据生成和使用情况进行分析，在此基础上构建一个可用于临床试验数据溯源的区块链 BlockTrial，病人和研究者可以通过智能合约与其进行交互[18]。

综合来看，关于传统数据溯源的相关研究较多，将区块链技术引入大数据溯源的相关研究较少；后一类研究中，讨论健康医疗供应链数据溯源的居多，综合考虑各类健康医疗数据溯源的较少，讨论将 FHIR（Fast Healthcare Interoperability Resources）和区块链相结合来实现健康医疗大数据演变历史和操作历史溯源的更少。本文探讨一种将 PROV 模型、FHIR 及其 Provenance 溯源模型、区块链相结合来实现多级可信溯源的健康医疗大数据溯源机制。

2 基本概念与相关数据模型

2.1 健康医疗大数据

现有文献对健康医疗大数据的定义较多，归纳起来可分为 3 类：（1）从个体视角进行的概念界定，如刘瑛等提出的观点：健康医疗（大）数据是反映特定个人的生理及心理状态的信息[19]；（2）从管理视角进行的概念界定，典型代表是文献[1]所提出的观点：健康医疗大数据是指人们在疾病防治、健康管理等过程中产生的与健康医疗相关的数据，国内很多学者的观点与其基本一致[20-21]；（3）从大数据特性视角进行的概念界定，认为健康医疗大数据是与健康医疗相关且满足大数据基本特征的数据集合[22-23]。

以上述后两类观点为基础，本文从内涵和外延两个方面对健康医疗大数据概念进行界定。（1）内涵方面，健康医疗大数据是指人们在疾病防治、健康管理等过程中产生的与健康医疗相关的数据，除了具备 Volume、Velocity、Variety、Value 等常规的大数据特征之外，还具备隐私性、可追溯性、规范性、长期保存性等特性。隐私性是指健康医疗大数据涉及个体的健康隐私，非授权用户不能查看个体的隐私数据；可追溯性是指健康医疗大数据保存着个体的诊断、治疗历史数据，承担着为医疗纠纷、医保纠纷提供证据的职责，应能为个体和机构提供数据溯源服务；规范性是指健康医疗大数据采用健康医疗领域的数据标准来描述数据，规范性较强，易于实现不同机构之间的共享；长期保存性则是指为了更好地保证数据的可追溯性，应尽可能长久地保存个体的诊疗历史等具有长期保存价值的数据。（2）外延方面。健康医疗大数据涉及人们从出生到死亡等一系列生命过程所产生的个体数据，以及与之相关的各种公共数据，本文在借鉴现有文献[24-26]相关内容的基础上，认为如下 8 种类型的数据都属于健康医疗大数据：计生、人口等基础数据，电子病历、医学检查与检验结果等临床数据，各类医疗保险数据，疾控、卫生应急等公共卫生数据，个体基因序列等生物医学数据，药物研发、临床试验及健康医疗科研数据，医政管理、药政管理数据，线上健康医疗数据、移动健康医疗数据等其它类型数据。

2.2 健康医疗大数据溯源

按照文献[2]的定义，数据溯源是对数据在整个生存周期内的演变信息和演变处理内容的记录。文献[7]认为，这种记录既可能是一种从数据当前状态向源头数据追溯的逆向回溯过程，也

可能是从源头数据运动之始就对数据变动信息进行捕获和记录的正向跟踪过程。在此基础上，可将健康医疗大数据溯源定义为：对健康医疗大数据在整个生存周期内的演变信息和演变处理内容的记录。

根据生命周期内数据是否发生变化，可将健康医疗大数据溯源分为两种类型：一是数据流转溯源，即对数据在不同存储位置之间的“漂流”过程进行记录，例如某个公众人物的某次体检报告从体检医院被泄漏到某个网站，可对整个泄露过程进行数据层面的溯源；二是数据演变过程溯源，即对数据的原始状态、促使数据发生改变的各种条件和动作及改变后的状态进行记录，例如糖尿病患者可通过溯源系统查询到自第一次血糖检验结果超标以来的所有诊疗数据。

根据被溯源对象的粒度，健康医疗大数据溯源可分为三级：数据集级、文档级、数据元级。数据集级溯源是指用户需要对数据流转全过程或数据演变全过程进行溯源，如某个患者对其在多家医院诊疗史的数据溯源，溯源结果可以是一个包含文本、影像、音频等多种媒体类型和多种数据格式的数据集，以及数据集中各个文档之间的演变关系。文档级溯源是指用户对指定文档的追溯，如某个患者对其某个时间在某个医院所做的核磁共振影像文件的追溯，溯源结果可以是CDA格式的结构化临床文档、PDF格式的非结构化诊断报告或DICOM格式的医学影像文档，以及与该文档相关的各种操作记录。数据元级溯源是指用户对某个特定数据项目的追溯，如某个患者对其某个时间在某个医院进行体检时甘油三酯值的追溯，溯源结果可以是XML格式或RDF格式的三元组。

根据被溯源对象的性质，健康医疗大数据溯源还可分为面向数据的溯源和面向过程的溯源，前者追溯的是数据，后者追溯的是对数据的处理过程及各种相关因素。

2.3 相关数据模型

（1）W3C PROV数据模型。W3C发布的PROV模型是一种面向数据溯源的数据模型，包括PROV-DM（PROV Data Model）、PROV-O（PROV Ontology）等多个组件。PROV-DM定义了实体（如网页、图表、数据集）、代理（如人、机构）、活动（如生成、翻译、删除）等概念及其相互关系，可用其描述数据的起源与演变。PROV-O以PROV-DM为基础，定义了Entity、Activity、Agent等3个基本类及其9个属性，7个扩展类及其16个属性，20个资格类及其25个属性，可用其描述溯源数据，形成RDF格式文档。

（2）FHIR数据模型。FHIR是Health Level Seven组织（简称HL7）在其现有标准v2、v3、RIM、CDA等的基础上研制的新一代健康医疗信息交换标准，其数据模型的基本组件是资源，分为Foundation、Base、Clinical、Financial、Specialized五个大类，共146个资源[27]。应用系统可以通过各种资源的组合来解决资源中涉及到的疾病、药品、观测指标等语义互操作问题。相较于HL7的其它标准，FHIR具有更强的灵活性和可扩展性，更易于快速部署和实现，适合用于健康医疗大数据及其溯源数据的描述与共享。FHIR包括一个名为Provenance的资源，可视为FHIR的数据溯源模型。任何一个资源都可内嵌一个Provenance资源，用于对资源的When、Where、Why、Who、How、What等多种类型的上下文信息进行描述。FHIR Provenance虽然基于W3C PROV，但只包含target、occurred、recorded、policy、location、reason、activity、agent、entity、signature等10个属性，对溯源数据的描述能力远远低于后者，属于轻量级的溯源数据描述模型。

此外，HL7 还在 FHIR 数据模型的基础上构建了 FHIR 本体，该本体包含 822 个类、6022 个属性[28]，可用于实现健康医疗数据的语义化描述。

3　溯源机制

从溯源体系架构、数据描述、数据安全与隐私控制、溯源流程 4 个方面来讨论基于区块链的健康医疗大数据溯源机制。

3.1　溯源体系架构

本文提出的健康医疗大数据溯源机制将现有文献所采用的标注方法和非标注方法进行结合：医疗机构、医保机构等数据控制者采用通用的数据溯源本体 PROV-O 对其所控制健康医疗大数据的演变情况进行标注，形成语义化溯源数据；溯源消费者通过应用层的数据溯源模块对保存在区块链中的健康医疗大数据查询情况、增删改情况、数据传递情况等操作史进行遍历，形成结构化溯源数据。两类数据融合，形成完整的溯源数据，可较为全面地反映目标数据的演变史和操作史。

上述目标的实现需要在本地系统、云存储系统、区块链系统的基础上形成溯源系统。该系统包括应用层、区块链存证层、云服务层、本地层、保障层，如图 1 所示。

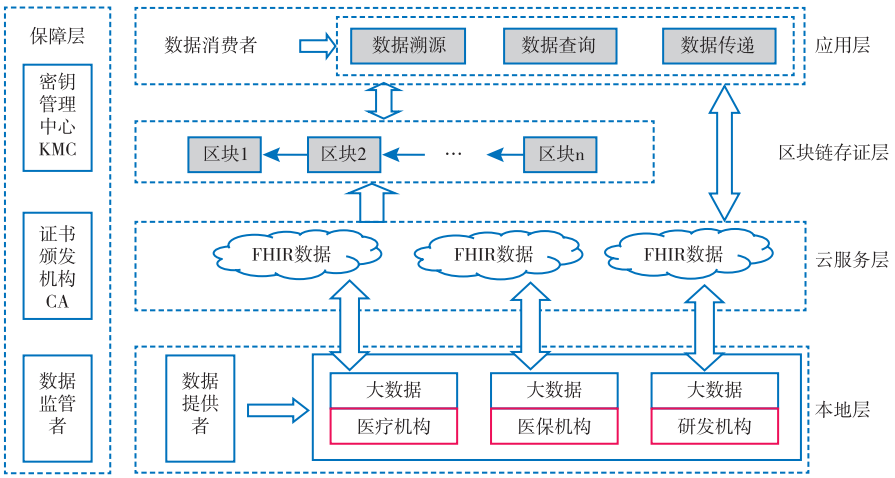

图 1　溯源体系架构

应用层包括数据溯源、数据查询、数据传递等模块。数据溯源模块可以实现药品供应链溯源、疾控数据溯源、患者诊疗数据溯源、保险数据溯源、临床试验数据溯源等，该模块可以同时查看区块链中的存证数据、云服务器中的 FHIR 数据、本地层的 PROV-O 溯源数据，并将操作记录保存在区块链中，以便发生隐私泄露时进行调查取证、定位泄露源。数据查询模块允许患者、医护人员等数据消费者经过授权后对保存在云服务器中的各种数据进行查询，查询操作记录也保存在区块链中。数据传递模块实现药品数据、诊疗数据、医保数据等在不同数据控制者之间的传

送、共享，源数据和目标数据都以 FHIR 数据的形式保存在云服务器中，传递操作记录保存在区块链中。

区块链存证层可采用联盟式区块链（以下简称为联盟链）来保存健康医疗大数据的操作历史，作为证据以供溯源。联盟链是一种带有准入限制的区块链，实体只有获得入链许可，才能作为节点加入。与公有链相比，联盟链可以更好地控制节点规模；与私有链相比，联盟链可以更好地实现不同类型健康医疗实体之间的数据共享。医疗机构、医保机构、药品研发机构等相关实体可在卫生健康主管部门的监督与管理下加入健康医疗区块链，在规定的权限范围内操作区块链并共享区块链中的信息。应用层可以只根据区块链中保存的操作历史进行数据溯源，本文称之为 I 级溯源。

云服务层由医疗机构、医保机构、药品研发机构等数据控制者的云服务器组成，用于提供 FHIR 格式健康医疗大数据当前版本的查询共享服务。数据消费者可以在权限范围内查询这些数据，法院、公安机关等部门在调查取证时可以通过卫生健康主管部门获取这些数据的查看权限。经过数据提供者的授权，数据控制者可以在云服务器中进行数据的增加、删除、修改操作，并将操作记录保存到区块链中。应用层可根据区块链中保存的操作历史和云服务器中的 FHIR 数据进行溯源，本文称之为 II 级溯源。

本地层由各个数据控制者节点组成，每个节点可以有自己的健康医疗大数据平台，用于提供原始的本地格式健康医疗大数据、原始的 PROV-O 格式溯源数据。节点从患者、医护人员、研发人员等数据提供者处采集数据，根据国际通用的 FHIR 本体将本地健康医疗大数据映射为 FHIR 数据，并将其发布到云服务器中。应用层可根据区块链中保存的操作历史、云服务器中的 FHIR 数据、本地大数据平台中的 PROV-O 溯源数据进行溯源，本文称之为 III 级溯源。应用层可在取得数据控制者的授权后，在 III 级溯源的基础上进一步查找保存在本地的原始数据如医学影像文件、化验单等，本文称之为 IV 级溯源。

保障层由数据监管者、密钥管理中心（Key Management Center，KMC）、证书颁发机构（Certificate Authority，CA）等组成，通过安全与隐私控制机制来保障数据溯源的安全性和可信性。卫生健康主管部门承担数据监管者的角色，通过用户管理来控制用户对溯源系统的使用权限，具体可通过部署在区块链上的智能合约来实现。用户权限操作过程也应该能溯源，需要保存到区块链中，可通过 I 级溯源，与数据操作历史一起呈现给数据消费者。

3.2 数据描述

按照存储位置的不同，可将溯源体系中的数据分为区块链数据、FHIR 数据、本地数据三类，每类数据可分别采用不同的描述方法和存储方法。

3.2.1 本地数据的描述

医疗机构有多种类型的信息系统，如医院信息系统（Hospital Information System，HIS）、临床信息系统（Clinical Information System，CIS）、医学影像归档和通信系统（Picture Archiving and Communication Systems，PACS）、实验室检验信息系统（Laboratory Information System，LIS）、电子病历系统（Electronic Medical Record，EMR）等，医保机构、研发机构等其它数据控制者也都有多种相应的系统。这些系统在描述健康医疗大数据时可以采用两种思路：一是采用我国国家卫

生健康委员会《电子病历基本架构与数据标准》等国内现有标准或数据控制者自行制定的本地数据模型来描述数据，同时提供可将数据转换为 HL7 FHIR 格式数据的接口；二是直接采用 HL7 FHIR 标准作为本地数据模型来描述数据。

为了提高本地数据的可追溯性，数据控制者需要为本地数据提供相应的溯源数据，具体可用现有的溯源本体 W3C PROV-O 对这些溯源数据进行描述。例如，针对患者 XXX 的电子病历（xls 格式，手掌骨手术病历，具体数据略），根据 PROV-O 本体，可形成如下 RDF 溯源数据（Turtle 格式）：

```
@prefix emr: <http://localhost/users/1410317/>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix dcterms: <http://purl.org/dc/terms/>.
@prefix prov: <http://www.w3.org/ns/prov#>.
emr:202004030095
    a prov:Entity;
    dcterms:title “患者 XXX 电子病历” ^xsd:string;
    prov:generatedAtTime “2020-04-03T13:35:23Z” ^xsd:dateTime;
    prov:wasDerivedFrom emr: 手掌骨 X 光 .dicom;
    prov:wasDerivedFrom emr: 血常规 .xls;
    prov:wasGeneratedBy  :activity;
    prov:wasAttributedTo “X 大夫” ;
    prov:wasAttributedTo :agent.
:activity
    a prov:Activity;
    prov:wasAssociatedWith :agent;
    rdfs:label “数据抽取” ^xsd:string;.
:agent
    a prov:Agent，prov:SoftwareAgent;
    rdfs:label “dcm4che” ^xsd:string.
```

其中 emr:202004030095 是患者 XXX 电子病历 URI，该电子病历由 X 大夫于 2020 年 4 月 3 日生成，病历中的数据由软件 dcm4che 从“手掌骨 X 光 .dicom”、“血常规 .xls”两个文件中抽取得到。

3.2.2　FHIR 数据的描述

可将 W3C PROV、FHIR 和 FHIR Provenance 结合起来用于描述 FHIR 数据，即：在 FHIR 数据中，通过 FHIR Provenance 的 10 个属性来描述资源的 10 项直接上下文信息，其它更多的直接上下文信息以及所有间接上下文信息（指上下文的上下文）可在本地数据的 PROV-O 溯源数据中描述。当数据消费者想查看某个资源的简单溯源数据时，只需从云服务器中获取该资源内嵌的 Provenance 资源即可；如果想查看详细的溯源数据，可进一步从本地数据中获取 PROV-O 溯源

数据。

医疗机构等数据控制者可以利用 FHIR 本体对本地数据进行 FHIR 化，从而实现共享数据的语义化描述。基本思路如下：第一，构建本地数据模型与 FHIR 资源之间的映射表，同时构建 PROV–O 本体与 FHIR Provenance 之间的映射表；第二，根据第一个映射表，将本地数据中的实体映射为 FHIR 资源、实体之间的关系映射为 FHIR 资源的属性；第三，根据第二个映射表，将本地数据中的 PROV–O 格式溯源数据映射为 FHIR Provenance 格式溯源数据；第四，根据 FHIR 本体，将上述映射结果整合为 FHIR 数据。

以上述患者 XXX 电子病历及其 PROV–O 溯源数据为例，可按照上述 FHIR 化思路，形成如下 FHIR 数据（Turtle 格式）：

```
@prefix fhir: <http://hl7.org/fhir/>.
@prefix fhirdata: <http://fah.zzu.edu.cn/fhir/procedure/>.
fhirdata:202004230047
  a fhir:Procedure;
  fhir:Procedure.identifier “http://fah.zzu.edu.cn/fhir/procedure/202004230047”;
  fhir:Procedure.status “完成”;
  fhir:Procedure.category “外科手术”;
  fhir:Procedure.subject  http://fah.zzu.edu.cn/fhir/patient/202004220036;
  fhir:Procedure.performer.actor  http://fah.zzu.edu.cn/fhir/practitioner/doctor0015;
  fhir:Procedure.bodySite “手掌骨”;
  fhir:Procedure.outcome “成功”;
  fhir:Procedure.report http://fah.zzu.edu.cn/fhir/DiagnosticReport/2020042200082;
  fhir:Procedure.note :provenance
  :provenance          // 内嵌的 provenance 资源
    a fhir:Provenance;
    fhir:Provenance.target  http://fah.zzu.edu.cn/fhir/procedure/202004230047;
    fhir:Provenance.occurredDateTime “2020–04–03T13:35:23Z”;
    fhir:Provenance.activity “数据抽取”;
    fhir:Provenance.agent.who “dcm4che”;
    fhir:Provenance.entity.what http://localhost/users/1410317/202004030095;
    fhir:Provenance.entity.what “8f...46a9”.
```

其中 fhirdata:202004230047 是 FHIR 数据在云服务器中的 URI，数据的前一部分是电子病历数据，由前文提及的 xls 格式手掌骨手术病历数据映射得到；后一部分是溯源数据，由前文所列的 PROV–O 溯源数据映射得到，其中最后一行为 HASH 函数值，表示 PROV–O 溯源数据的数字摘要。

3.2.3　区块链数据的描述

从本质上看，区块链是一种数据只进不出的链表，由一系列的区块及附加数据组成。由于患

者的每一次就诊数据都会形成一个交易，并最终打包到区块链的新增区块中，且这种交易的数量非常多，因此，从顶层设计的角度看，区块的数据结构应尽量简单，只保留必备的区块链管理数据以及与溯源密切相关的数据即可，大量的结构化健康医疗数据可保存在各个数据控制者的云服务器中，医学影像数据等一些超大量的非结构化数据可保存在数据控制者的本地数据库中。

区块需要借助于健康医疗区块链数据模型进行描述后才能通过联盟链事先规定的共识机制追加到区块链的尾部，形成一个新的区块。在借鉴比特币区块链数据模型、以太坊区块链数据模型的基础上，本文设计了一种健康医疗区块链数据模型，该模型规定了区块头、区块体、交易等实体各自所含字段的名称、类型、长度及取值范围。如图 2 所示，区块头包含 5 个字段；区块体包含 3 个字段，其中交易列表是一个数组，由 n 个交易 HASH 组成，n 是一个常数，其取值可在进行联盟链顶层设计时确定；每个交易均包含 6 个字段，其中交易内容字段是复合型的可变长字段，其结构由交易类型字段决定：

（1）交易类型 =“增加”时，交易内容字段由 FHIR 数据类型、FHIR 数据提供者公钥、FHIR 数据 HASH、FHIR 数据 URI 四个子字段组成。FHIR 数据类型子字段的取值分为“open”、“non-open”两种情况，前者适用于 FHIR 数据为开放数据的情况，FHIR 数据 URI 子字段的值不加密，所有的数据消费者都可以通过不加密的 URI 访问云服务器中的 FHIR 数据；后者适用于 FHIR 数据包含隐私信息的情况，FHIR 数据 URI 子字段的值为加密后的值，数据提供者或经过授权的其它数据消费者可通过 FHIR 数据提供者的私钥进行解密方可获取云服务器中的 FHIR 数据。

（2）交易类型 =“删除”时，交易内容字段由交易 HASH、FHIR 数据提供者公钥、被删除 FHIR 数据三个子字段组成，其中第一个子字段指明被删除 FHIR 数据在区块链中的交易 HASH，第三个子字段指明经 FHIR 数据提供者公钥加密的被删除 FHIR 数据。

（3）交易类型 =“修改”时，交易内容字段由交易 HASH、修改内容、FHIR 数据类型、FHIR 数据提供者公钥、FHIR 数据 HASH、FHIR 数据 URI 六个子字段组成，其中第一个子字段指明被修改 FHIR 数据在区块链中的交易 HASH，第二个子字段指明修改的内容，其余四个子字段的含义同“增加”交易类型。

（4）交易类型 =“查询”时，交易内容字段由交易 HASH、FHIR 数据提供者公钥两个子字段组成，其中第一个子字段指明被查询 FHIR 数据在区块链中的交易 HASH。

（5）交易类型 =“传递”时，交易内容字段由 FHIR 数据提供者公钥、FHIR 数据接受者公钥、发送内容三个子字段组成，其中发送内容子字段经过 FHIR 数据接受者公钥加密。

以上述 FHIR 数据 fhirdata:202004230047 为例，将其发布到云服务器中属于“增加”操作，从云服务器中查询该数据属于“查询”操作。可将两个操作的相关信息以交易的形式永久保存到区块链中，以备数据消费者追溯该数据的操作历史。交易数据内容如下：

[66...7ef31] // 交易 HASH，256bit
[1594820322] // 交易时间戳，2020-07-15 21:38:42
[MI...LwIDAQAB] // 交易生成者公钥，2048bit
[增加] // 交易类型

［［non-open］［MI...mQIDAQAB］［9d...45d9］［msKg...Cw==］］　　// 交易内容
［306c...6711］　　// 交易数字签名，1024bit。以上 6 行为“增加”交易

［a8...02c4］　　// 交易 HASH，256bit。以下 6 行为“查询”交易
［1594902245］　　// 交易时间，2020-07-16 20:24:05
［MI...LwIDAQAB］　　// 交易生成者公钥，2048bit
［查询］　　// 交易类型
［［66...7ef31］［MI...mQIDAQAB］］　　// 交易内容
［885d...c622］　　// 交易数字签名，1024bit

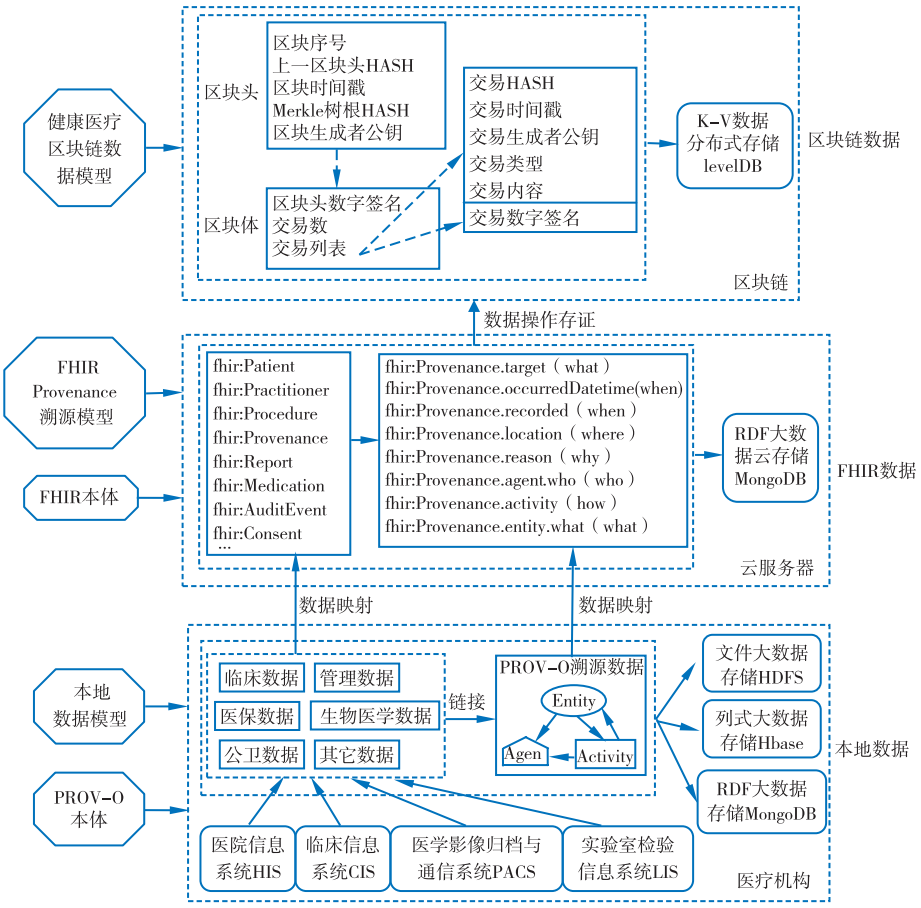

图 2　溯源数据的描述（以医疗机构为例）

3.3　数据安全与隐私控制

健康医疗大数据在安全与隐私方面的要求较高，涉及数据的保密性、完整性和有效性。保密性是指数据在传输、存储与使用过程中，只有授权者才能看到；它是健康医疗大数据溯源的基本要求，用于保证数据消费者的溯源行为只发生在拥有查看权限的数据范围内。完整性是指数据在

传输、存储与使用过程中须确保数据来自于已认证的数据源，数据内容不被篡改；它是确保健康医疗大数据溯源结果可信的保证。有效性是指数据一旦产生便不能被否认；它是确保健康医疗大数据溯源结果不被相关数据操作者抵赖的保证。

为了满足上述要求，所有参与到溯源系统的个人和机构均需向政府认可的证书颁发机构和密钥管理中心申请数字证书和密钥对，以便据此对具有保密性要求的数据进行加密、对具有完整性要求的数据进行数字签名。此外，区块链存证层作为一个联盟链，还需要一个中心化的用户权限管理机构来增、删用户并进行权限管理，可由作为数据监管者的卫生健康主管部门来承担这一角色。访问溯源系统的用户可分为三类：监管机构、普通机构、个人。监管机构进一步分为中央、省、市、县四级，上级机构决定哪些下级机构可以加入联盟链并为其授予权限，每个机构都可自行决定辖区内的哪些医疗、医保或研发等普通机构可以加入联盟链并为其授予权限。普通机构经过数据提供者的授权，可将数据发布到云服务器；作为联盟链的节点，还具有向联盟链提交交易申请的权利；同时还可为个人开设账户并进行授权。个人用户只能通过应用层的数据溯源模块和数据查询模块对系统中的开放数据和本人具有查看权限的数据进行查看和溯源，无权查看其它数据，也无权向联盟链提交交易申请。

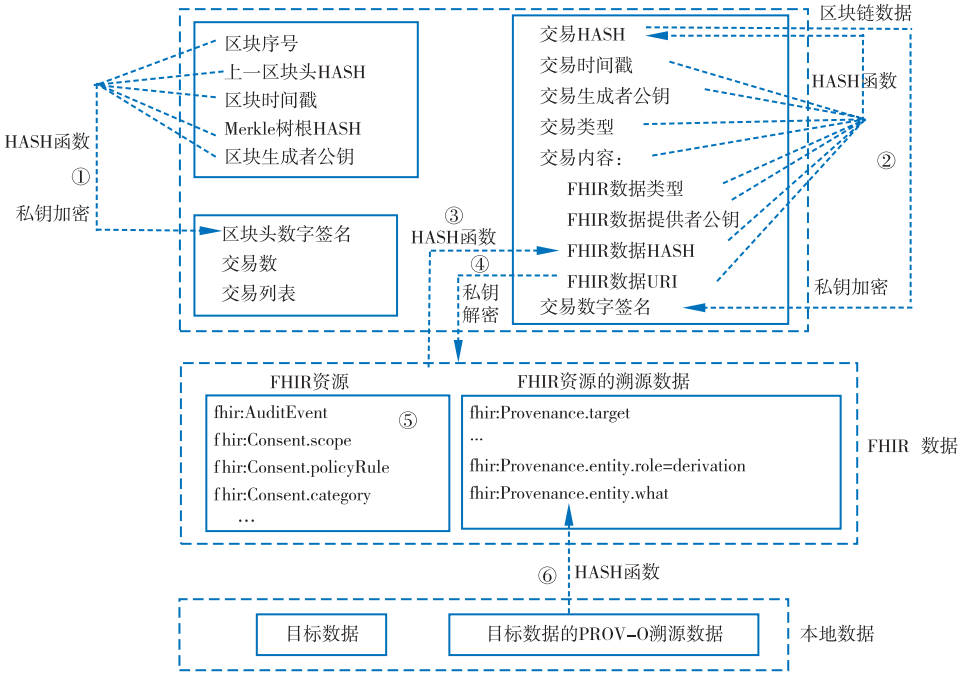

图 3　数据安全与隐私控制

卫生健康主管部门以及经其授权的机构可通过智能合约来实现上述的用户权限管理。该智能合约是一段可通过 ABI JSON 接口和合约地址进行调用并通过区块头中的 Merkle 树根 HASH 进行完整性验证的代码，具体实现时，该代码应至少包括如下两个函数：（1）用户权限设置函数，当卫生健康主管部门通过管理系统增、删、改用户权限时触发该函数，产生并向区块链提交一个

新的交易，其中交易类型为“用户管理”，交易内容包括用户名、用户类型、用户权限等子字段；（2）结果返回函数，当上述交易通过联盟链共识机制获得认可并被打包到新的区块中时，触发该函数返回权限设置结果，管理系统根据该结果来判断设置是否成功。当权限设置成功后，新的用户权限信息将保存在管理系统中，同时，对用户权限的增、删、改等操作信息将保存在区块链中，以防止卫生健康主管部门以及经其授权的机构抵赖。

如图 3 所示，可以非对称加密、数字签名、数字摘要等技术为基础，通过以下具体措施实现数据的安全与隐私控制：（1）在生成区块数据时，通过 HASH 函数生成区块序号、上一区块头 HASH 等 5 个字段的数字摘要，并利用区块生成者私钥对其进行加密，生成区块头数字签名。数据消费者在溯源过程中获取某个区块后，可通过区块头中的区块生成者公钥对数字签名进行逆运算，验证区块头数据的完整性。（2）在生成交易数据时，通过 HASH 函数生成交易时间戳、交易生成者公钥、交易类型、交易内容 4 个字段的数字摘要（即交易 HASH），并利用交易生成者私钥对其进行加密，生成交易数字签名。数据消费者在溯源过程中获取某个交易后，可通过交易生成者公钥对交易数字签名进行逆运算，验证交易数据的完整性。（3）将 FHIR 数据发布到云服务器中时，需要在区块链交易内容字段的 FHIR 数据 HASH 子字段中保存该数据的数字摘要（HASH 函数值），用于判断该数据是否被修改过。（4）将 FHIR 数据发布到云服务器中时，需要在区块链交易内容的 FHIR 数据 URI 子字段中保存该数据的 URI；如果 FHIR 数据是 non-open 类型，则需要利用数据提供者公钥对该 URI 进行加密，数据提供者或其委托的其它数据消费者可以利用数据提供者私钥对 URI 进行解密并从云服务器中访问该 FHIR 数据。（5）在 FHIR 数据的 fhir:AuditEvent 属性中记录与隐私、安全相关的事件，在 fhir:Consent 属性中记录数据提供者的同意信息，可根据这两项数据判断数据提供者隐私信息被侵犯的情况。（6）将本地目标数据的 PROV-O 溯源数据转换为 FHIR 溯源数据时，需要将前者的数字摘要作为属性 fhir:Provenance.entity.what 的值保存到云服务器中，用于判断 PROV-O 溯源数据是否被修改过。

3.4 溯源流程

溯源是由患者、医护人员、管理人员等数据消费者提出的，既可以对数据的演变历史进行溯源，也可以对数据的操作历史进行溯源，还可以同时对两种历史数据进行溯源。对于普通数据消费者来说，溯源只能在一定范围内的数据中进行，这些数据包括未加密的开放数据、数据消费者有权访问或被授权访问的数据。法院、公安等部门需要数字证据时，可以通过卫生健康主管部门获得溯源系统的使用权，通过 KMC 获得特定的数据提供者密钥对，在此基础上实施溯源。

如图 4 所示，溯源流程包括如下三个阶段：

（1）区块链数据溯源阶段。数据消费者提出溯源请求后，数据溯源模块以其公钥为关键词对区块链进行遍历，检索出所有与其相关的区块，对每个区块的区块头进行数字签名验证，对每个区块所包含的所有交易的数字签名进行验证，在确定区块数据完整无误后，根据每个交易的交易类型和交易内容，对 FHIR 数据的操作历史和用户权限设置历史进行还原。该阶段的数据溯源属于 I 级溯源，当数据提供者的信息发生泄露时，数据提供者或公检法等执法部门可通过 I 级溯源排查泄露源；当数据提供者想了解其数据操作历史、用户想了解其权限变迁史时，可通过 I 级溯源获取相应的数据。

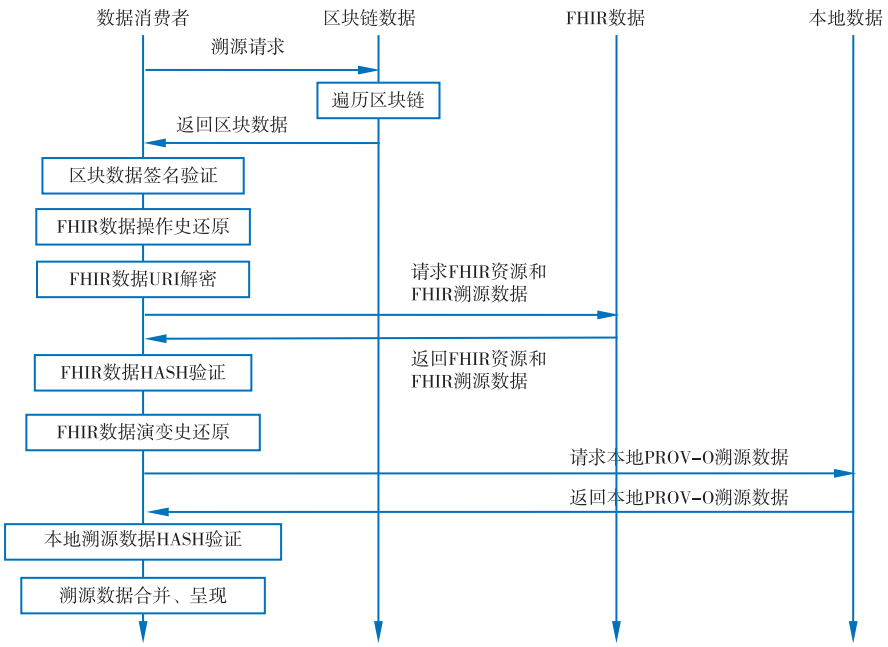

图 4　溯源流程

（2）FHIR 数据溯源阶段。通过每个区块交易内容字段中的 FHIR 数据 URI 子字段和 FHIR 数据 HASH 子字段，获取 FHIR 数据的 URI（当 FHIR 数据类型为 non-open 时，需要通过数据提供者私钥对其解密）和 HASH，根据 URI 从云服务器中获取对应的 FHIR 资源和 FHIR 溯源数据，根据 HASH 对 FHIR 资源和 FHIR 溯源数据进行完整性验证，在此基础上对 FHIR 数据的演变历史进行还原。上述两个阶段的溯源构成 II 级溯源，适用于数据提供者通过互联网获取其 FHIR 数据及其简单的溯源数据、同一患者通过互联网追溯其在不同医疗机构的所有 FHIR 格式诊疗数据等场合。

（3）本地数据溯源阶段。通过 FHIR 数据中的 fhir:Provenance.entity.what 属性，获得本地 PROV-O 溯源数据的访问路径和数字摘要，获取本地 PROV-O 溯源数据并进行 HASH 验证，进一步得到数据在本地的演变历史，并与上述 FHIR 数据操作历史和 FHIR 数据演变历史进行合并，形成完整的数据溯源结果。上述三个阶段的溯源构成 III 级溯源，适用于数据提供者或患者查看分布在区块链、云服务器、各个医疗机构的完整溯源数据等情况。如果患者本人或其主治大夫想进一步查看保存在医疗机构本地数据库中的医学影像文件、化验单等原始数据，或者发生医疗事故、出现医保纠纷时需要查看这些原始数据，则可进一步进行 IV 级溯源。

按照图 4 所述流程，对 3.2 节所述患者 XXX 的电子病历进行溯源，结果如图 5 所示，其中上半部分为区块链数据溯源阶段的溯源结果，右下部分为 FHIR 数据溯源阶段的溯源结果，左下部分为本地数据溯源阶段的溯源结果。

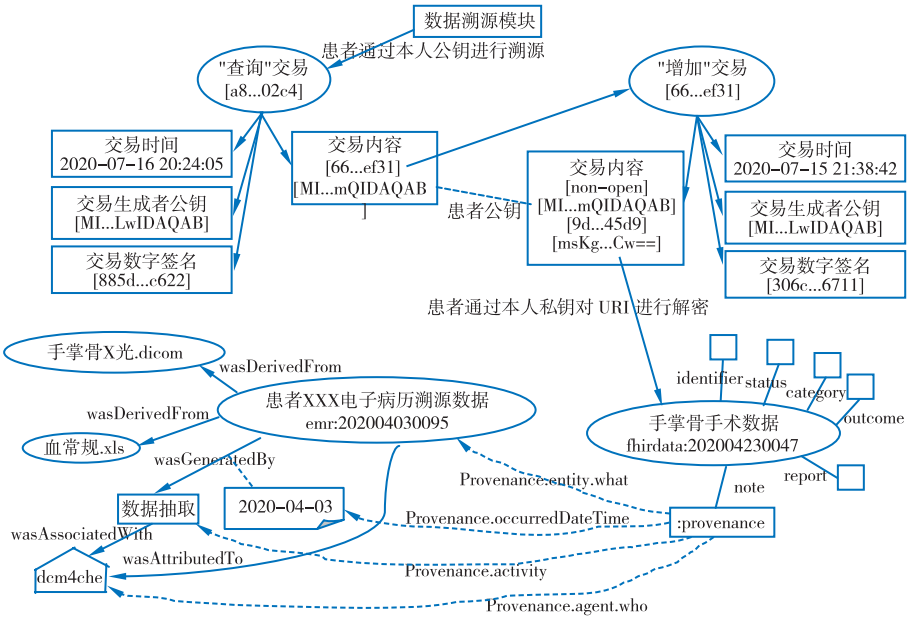

图 5　溯源结果示例

4　结　语

将区块链技术引入到健康医疗大数据溯源中，有助于增强溯源过程的安全性和溯源结果的可信性。本文对基于区块链的健康医疗大数据溯源机制所涉及的溯源体系架构、数据描述、数据安全与隐私控制、溯源流程等几个关键问题进行了分析，采用 PROV-O 本体、FHIR Provenance 溯源模型、健康医疗区块链数据模型等来描述数据，采用非对称加密、数字摘要、数字签名来保障溯源过程中的数据安全与隐私，可达到对健康医疗大数据的演变历史和操作历史进行多级可信溯源的目的。首次提出了四级溯源概念以及采用 PROV-O 本体、FHIR 本体、FHIR Provenance 溯源模型、健康医疗区块链数据模型对健康医疗大数据进行多级描述的方法。

本文只是从理论上对基于区块链的健康医疗大数据溯源机制进行了探讨，没有利用实际的数据集进行实验验证；溯源流程中，区块链遍历、FHIR 数据操作史还原、FHIR 数据演变史还原等环节没有给出具体的算法；区块链中 K-V 数据的存储、RDF 大数据云存储、文件大数据和列式大数据的本地存储等健康医疗大数据存储方式只在图 2 中提及，并没有展开论述。未来将针对这些问题做进一步的研究。

【参考文献】

［1］国家卫生健康委员会 . 国家健康医疗大数据标准、安全和服务管理办法［EB/OL］.［2019-12-23］. http://www.cac.gov.cn/2018-09/15/c_1123432498.htm.

［2］信息技术数据溯源描述模型：GB/T34945-2017［S/OL］.［2019-10-15］. http://openstd.samr.gov.cn/bzgk/gb/newGbInfo?hcno=B05FB0694A5C2E9995C50F8F55764530.

[3]Data lineage VS data provenance [EB/OL]. [2020-06-05]. http://toppertips.com/data-lineage-vs-data-provenance/.

[4]BELHAJJAME K, GARIJO D, MISSIER P, et al. PROV model primer [EB/OL]. [2020-06-23]. https://www.w3.org/TR/prov-primer/.

[5]MOREAU L,FREIRE J,FUTRELLE J, et al. The open provenance model [J].Future Generation Computer Systems,2011,27(6):743-756.

[6] 倪静 , 孟宪学 . 关联数据环境下数据溯源描述语言的比较研究 [J]. 现代图书情报技术 , 2013(2):18-23.

[7] 王芳 , 赵洪 , 马嘉悦等 . 数据科学视角下数据溯源研究与实践进展 [J]. 中国图书馆学报 ,2019,45(5):79-100.

[8] 明华 , 张勇 , 符小辉 . 数据溯源技术综述 [J]. 小型微型计算机系统 ,2012,33(9):1917-1923.

[9] 戴超凡 , 王涛 , 张鹏程 . 数据起源技术发展研究综述 [J]. 计算机应用研究 ,2010,27(9):3215-3221.

[10]MCCLATCHEY R, SHAMDASANI J, BRANSON A, et al. Traceability and provenance in big data medical systems [EB/OL]. [2019-6-13]. https://ieeexplore.ieee.org/document/7167491.

[11]GLAVIC B. Big data provenance: challenges and implications for benchmarking [J]. Lecture Notes in Computer Science,2014(1):81-93.

[12]BELL L, BUCHANAN W, CAMERON J, et al. Applications of blockchain within healthcare[J/OL].[2020-12-9]. Blockchain in Healthcare Today,2018. https://www.researchgate.net/publication/325417373.

[13] 禹忠 , 郭畅 , 谢永斌等 . 基于区块链的医药防伪溯源系统研究 [J]. 计算机工程与应用 ,2020,56(3):35-41.

[14]MARBOUH D,ABBASI T,MAASMI F. et al. Blockchain for COVID-19: Review, opportunities, and a trusted tracking system [J/OL]. Arabian Journal for Science and Engineering,2020.

[15]MASI M, MILADI A, MARGHERI A, et al. Using PROV and blockchain to achieve health data provenance [EB/OL]. [2019-06-15]. https://eprints.soton.ac.uk/421292/.

[16]MARGHERI A, MASI M, MILADI A, et al. Decentralised provenance for healthcare data [J]. International Journal of Medical Informatics,2020,141:104197.

[17]TANAKA K, YAMAMOTO R. Assessment of traceability implementation of a cross-institutional secure data collection system based on distributed standardized EMR storage [J]. Studies in Health Technology & Informatics, 2019, 264(10):1373-1377.

[18]MASLOVE D, KLEIN J, BROHMAN K, et al. Using blockchain technology to manage clinical trials data: A proof-of-concept study [J/OL]. JMIR Medical Informatics, 2018,6(4):e11949.

[19] 刘瑛 , 高逸 . 健康医疗数据法律规制研究 [J]. 天津师范大学学报 (社会科学版),2020(2):59-63.

[20] 张世红 , 李磊 , 史森 . 区域健康医疗大数据中心体制机制研究 [J]. 医学信息学杂志 ,2020,41(5):43-48.

[21] 陈圣杭 , 胡晶 , 包耿珲等 . 健康医疗大数据资源开放对策研究 [J]. 福建电脑 ,2020,36(4):51-54.

[22] 金磊 . 大数据医疗的发展现状及未来展望 [J]. 安徽科技 ,2017(11):42-43.

[23] 敖勇平 . 健康医疗大数据的现状及应用场景探索 [J]. 电脑知识与技术 ,2018,14(6):1-2.

[24] 许培海 , 黄匡时 . 我国健康医疗大数据的现状、问题及对策 [J]. 中国数字医学 ,2017,12(5):24-26.

[25] 李岳峰 , 胡建平 , 张学高 . 中国健康医疗大数据资源目录体系与技术架构研究 [J]. 中国卫生信息管理杂志 ,2019,16(3):249-256.

[26] 汪冬 , 秦利 , 魏洪河等 . 健康医疗大数据发展现状与应用 [J]. 电子技术与软件工程 ,2018(11):209-210.

[27]HL7 FHIR release 5 preview [EB/OL]. [2020-12-03]. http://build.fhir.org/.

[28]FHIR Ontology [EB/OL]. [2020-12-03]. https://www.hl7.org/fhir/fhir.rdf.ttl.zip.

郭少友，胡斐然 . 应用区块链构建健康医疗大数据溯源机制［J］. 文献与数据学报，2020，2（4）：024-039.

Construction of Health Care Big Data Provenance Mechanism Based on Blockchain

GUO Shaoyou　HU Feiran

(School of Information Management, Zhengzhou University, Zhengzhou 450001, China)

Abstract:［**Purpose/significance**］Blockchain has the characteristics of decentralization, high security, and strong traceability. It can help to enhance the security of provenance process and the credibility of results by introducing blockchain into the construction of health care big data provenance mechanism.［**Method/process**］Using PROV-O ontology, FHIR provenance model and health care blockchain data model to describe the data, and asymmetric encryption, digital digest and digital signature to ensure data security and privacy in the provenance process.［**Results/conclusion**］The proposed provenance mechanism involves provenance architecture, data description, data security and privacy control, provenance process. It can achieve multi-level credible provenance of the evolution history and operation history of health care big data.

Keywords: Blockchain; Health care big data; Data provenance

（本文责编：王秀玲）