frontiers in Blockchain

ORIGINAL RESEARCH
published: 29 October 2021
doi: 10.3389/fbloc.2021.667388

Blockchain Native Data Linkage

James Cunningham[1], Gail Davidge[2], Nigel Davies[3], Sarah Devaney[2], Søren Holm[2], Mike Harding[3], Gary Leeming[4], Victoria Neumann[3] and John Ainsworth[1]*

[1]Division of Informatics, Imaging and Data Sciences, Health EResearch Centre, Manchester Academic Health Science Centre, University of Manchester, Manchester, United Kingdom, [2]CSEP, Department of Law, School of Social Sciences, The University of Manchester, Manchester, United Kingdom, [3]School of Computing and Communications, Lancaster University, Lancaster, United Kingdom, [4]Civic Data Cooperative, Faculty of Health and Life Sciences, University of Liverpool, Liverpool, United Kingdom

Data providers holding sensitive medical data often need to exchange data pertaining to patients for whom they hold particular data. This involves requesting information from other providers to augment the data they hold. However, revealing the superset of identifiers for which a provider requires information can, in itself, leak sensitive private data. Data linkage services exist to facilitate the exchange of anonymized identifiers between data providers. Reliance on third parties to provide these services still raises issues around the trust, privacy and security of such implementations. The rise and use of blockchain and distributed ledger technologies over the last decade has, alongside innovation and disruption in the financial sphere, also brought to the fore and refined the use of associated privacy-preserving cryptographic protocols and techniques. These techniques are now being adopted and used in fields removed from the original financial use cases. In this paper we present a combination of a blockchain-native auditing and trust-enabling environment alongside a query exchange protocol. This allows the exchange of sets of patient identifiers between data providers in such a way that only identifiers lying in the intersection of sets of identifiers are revealed and shared, allowing further secure and privacy-preserving exchange of medical information to be carried out between the two parties. We present the design and implementation of a system demonstrating the effectiveness of these exchange protocols giving a reference architecture for the implementation of such a system.

Keywords: blockchain, data linkage, health informatics, distributed ledger technologies, digital health

Edited by:
*Jonathan Passerat-Palmbach,
Imperial College London,
United Kingdom*

Reviewed by:
*Alexander Vladimirovich Bogdanov,
Saint Petersburg State University,
Russia
Gisele Waters,
Nova Southeastern University,
United States*

****Correspondence:***
*James Cunningham
james.a.cunningham@
manchester.ac.uk*

Specialty section:
*This article was submitted to
Blockchain for Science,
a section of the journal
Frontiers in Blockchain*

Received: *12 February 2021*
Accepted: *29 April 2021*
Published: *29 October 2021*

Citation:
*Cunningham J, Davidge G, Davies N,
Devaney S, Holm S, Harding M,
Leeming G, Neumann V and
Ainsworth J (2021) Blockchain Native
Data Linkage.
Front. Blockchain 4:667388.
doi: 10.3389/fbloc.2021.667388*

1 INTRODUCTION

The increasing prevalence of electronic health record (EHR) data and its use for both administrative and research purposes has generated the ability for data consumers to derive novel and unexpected results from an increased breadth of available data. Data linkage is the process of combining disparate data sources into a single source, in particular unifying items of data that relate to the same entity identifier and extending the available set of information that pertain to a given individual. Through linking data sets in such a way data consumers, that is the data scientists or administrative analysts, can access richer sources of data and derive otherwise inaccessible results.

While there are clear benefits to providing the ability to link data sources in this way, there are serious issues of privacy to consider. Data Linkage systems need to ensure that privacy maintained, both in terms of ensuring pseudonymized or protected data remains hidden and in preventing information leakage through sharing data that was it not necessary to share in order to link data. Risks inherent in linking sources of medical data also include the deliberate exfiltration of private data through incorrectly authenticated requests to link data, incorrect linkage of non-related entities and the increase in potential security attack surface. Currently data linkage systems often rely on a trusted third party to link data to mitigate against the privacy risks inherent in linking data between individually sensitive sources. These trusted third parties act as a firewall between individual data sources and act to preserve the privacy of the records being linked and to provide guarantees regarding data integrity.

Blockchain is the core set of technologies derived from the underlying architecture of the Bitcoin cryptocurrency—a decentralized, trustless digitally based form of currency. A blockchain is an immutable data structure that encodes a ledger of signed transactions between cryptographically encoded addresses in as a chain of time-ordered blocks of data. The approach taken by Bitcoin as originally conceived was a means of addressing the 'Byzantine Generals' problem (that is the problem of communicating canonical information between a group of peers where certain members of that group may be untrustworthy) for particular use within the financial domain. It has also subsequently found wider adoption in fields outside of finance. While there has been a lot of hype and overpromize regarding the applicability and usefulness of blockchain in many areas, there have still been a number of innovative and successful uses of the technology. In the medical field in particular blockchain technologies have found legitimate use in a number of areas. The underlying primitives of blockchain implementations are cryptographic in nature which is of use in terms of both ensuring privacy and providing mechanisms to prevent data leakage. In the context of medical data, and particularly the protection of privacy within data being used and transmitted, these features are of particular importance.

In this paper we analyze the requirements for a system of medical data linkage that utilizes core blockchain functionality in place of a trusted third party linkage provider. We then present a reference architecture that will enable the creation and use of such a system. The remainder of this section discusses in more depth relevant aspects of blockchain technology and data linkage.

1.1 Blockchain

Bitcoin is a decentralized crypto-currency proposed and originally implemented in 2009 by the pseudonymous Satoshi Nakamoto (Nakamoto, 2009). Originally envisaged as a form of electronic cash that could be implemented without the need for reliance on a centralized or trusted third party to ensure the validity of transactions, Bitcoin has grown from a nascent network protocol with participating nodes in the network producing unvalued 'coins', to having a market cap of around 700 billion US dollars at the time of writing. Previous efforts to implement similar ideas had failed to successfully address the issue of 'double spending'—given an electronic asset that is hence easily duplicated, how can a system prevent the transmission of that asset (where it acts as a unit of value) to two separate parties simultaneously, thus undermining the perceived value that asset holds. For an electronic system having a central database (controlled and verified by a 'third party' with respect to the transaction taking place) addresses this problem, but for philosophical and political reasons there was the perceived need for a system not reliant on such a trusted third party. Such a decentralized system would need to rely on a network of nodes acting such that a canonical ordering of transactions was passed between nodes, ensuring that assets could not be double spent. However given that there is clear economic incentive for nodes to 'cheat' in certain circumstances the protocol underlying the network would need to be constructed in such a way as to ensure honesty between peers. The problem of ensuring the honest communication of canonical information between potentially untrustworthy nodes in a network is, in general, conceived of as the 'Byzantine Generals Problem' (Avizienis et al., 2004). The solution presented by the Bitcoin protocol was *via* the creation of the blockchain data structure.

A blockchain is a ledger of transaction consisting of a time-ordered series of 'blocks' of data. Each block contains a list of signed transactions between addresses (in the Bitcoin protocol addresses are the hashes of public keys, where transactions can be signed by the corresponding private key) along with a hash of the previous block in the chain. Bitcoins highly novel innovation as a protocol was that, to ensure that there was an agreed upon canonical chain of blocks (thus comprising a common distributed ledger of transactions) each block would also contain a salt value such that the hashed value of the block fell below a certain number. Valid blocks could thus be produced by nodes 'mining' salt values for blocks by guessing multiple values for the salt until a valid one was found. This ensures that blocks can be produced at controlled time intervals in an agreed upon manner underpinned by the protocol. A valid block also rewards the producer of that block with additional bitcoin, ensuring an economic incentive to produce blocks in the chain.

Removed from the economic value of the currency it encodes and its application as a 'store of value,' Bitcoin, *via* the blockchain data structure mechanism, exhibits core properties of:

- Security–the public key underpinnings of blockchain technology ensure that information stored in the digital ledger that a blockchain provides is cryptographically sound.
- Decentralization–blockchains are produced and secured by decentralized networks of nodes acting as equal peers in the network. There is no single source of trust that operation of the network relies on.
- Consensus–the economic incentive to produce blocks, and the corresponding cost of producing competing chains of blocks to undermine the currently agreed upon order, ensures that there is consensus between nodes in the network as to what information has been recorded in the distributed ledger.
- Immutability–the fact that each block in a blockchain data structure must contain a hash of the previous block in the chain means that no information recorded in the ledger can be removed; data recorded on a blockchain is, by definition, immutable.

Following the launch and success of the Bitcoin currency the blockchain technology has been further utilized in an ever growing number of alternative crypto-currencies, with varying degrees of difference to the bitcoin protocol; most notably Ethereum (Buterin, 2016) which introduced the concept of smart contracts (computational rule based transactional specifications that can record additional data on the chain and conditionally control currency transactions). Beyond the crypto-currency space though blockchain technology has been utilized, with varying degrees of validity and success in a number of other fields (Miraz and Ali, 2018). To an extent some of these applications can be seen as solutions in search of a problem, often adding little of value over non blockchain-based solutions (Golosova and Romanovs, 2018). However, where one or more of the core features we have highlighted (namely security, decentralization, consensus and immutability of data) are necessary properties of a potential solution then blockchain technology can offer genuine progress in a given area (Underwood, 2016). In particular, the properties of security and consensus can be seen as particularly relevant to the healthcare domain.

1.2 Blockchain in Digital Health

In the healthcare domain in particular the properties of blockchain can find genuine application in a range of areas (Angraal et al., 2017; Hölbl et al., 2018; Agbo et al., 2019). Active research in the application of the technology to health informatics ranges from the potential use of the blockchain as a means to store and process medical records (Zhang and Ji, 2018), being used as a permissions management layer (Azaria et al., 2016), through use as an administrative tool (Cyran, 2018), to applications in supply chain management (Clauson et al., 2018). Looking in particular at applications from the perspective of the properties offered by blockchain, namely security, decentralization, consensus and immutability, we can see that these different properties can be leveraged by health informatics applications across a variety of use cases.

1.2.1 Security

Guardtime (Williams-Grut, 2016) is an organization providing 'zero trust' systems aimed at governmental and enterprise level clients utilizing the KSI blockchain stack (Nagasubramanian et al., 2020). A particular application provided by the Guardtime platform is vaccine guard, which aims to securely and efficiently provide and record Covid 19 vaccination records against patient identities, while preserving the privacy of the individual to whom the vaccination data pertains (Vazirani et al., 2019). The vaccine guard system combines information about a formally identified patient, with vaccination data and verification by registered healthcare organisations. This information is recorded on a blockchain. The use of blockchain technology here is reliant on the security that the blockchain provides in terms of the cryptographic soundness of the verification protocols used to establish authenticity of certificates and the guarantees provided by the underlying protocol that prevent the forgery, replication or deletion of validly recorded information.

1.2.2 Decentralization

The Carechain project seeks to enable patients "to offer both healthcare professionals and researchers access to their entire health history as well as to directly purchase services in a global marketplace to improve their health" (Leeming et al., 2019). This provision of a marketplace mechanism for patient data and services is enabled and mediated *via* the public Ethereum blockchain network. Marketplace mechanisms rely on efficiency of the communication of price and availability data and the establishment of peer-to-peer mechanisms for matching marketplace participants. Key to the engendering of both trust and reliability in such a system is the decentralized nature of the underlying blockchain technology. Removing any central point of failure enhances both the security and reliability of the market place and promotes fairness between provider and purchaser, where, particularly in the case of healthcare data, there is potential for patients to be taken advantage, breaking inherent ethico-legal constraints (Hoffman, 2015).

1.2.3 Consensus

Medrec is system that enables patients to control and enable access to medical records and healthcare provision information from different sources (Azaria et al., 2016). The Medrec system records metadata about medical records and enables access and use of these records by patients and authorized organisations. The implementation of the Medrec system uses blockchain technology to enable medical record providers to record and share information regarding this access. The mechanism used by Medrec enables groups of verified providers to efficiently come to consensus agreement as to the validity and meaning of the metadata being shared.

1.2.4 Immutability

The MediLedger Project from Chronicled provides a blockchain-based system for Pharmaceutical supply chain (Mattke et al., 2019). The project is a cross pharmaceutical industry initiative enabling participating organisations to meet the requirements of the US Drug Supply Chain Security Act (Brechtelsbauer et al., 2016). Using a private blockchain instance organisations can track and trace the supply and provision of pharmaceutical products (Woods and Iyengar-Emens, 2019). The immutability of blockchain data ensures that information about events recorded at specific times regarding supply chain events can be trusted to have been accurately recorded and not altered or revoked, allowing for accurate auditing and proof of compliance with relevant legislation.

1.3 Data Linkage

Data Linkage can be defined as the act of combining "two or more different sources, data that relate to the same individual, family, place or event" (Holman et al., 2008). While potentially done using paper-based medical record sources, the shift of medical data into predominantly digital forms (Johnson et al., 2014) has increased the ease with which linkage between data sets can be performed. The ability to link disparate data sets into single unified sources brings benefits in both the administrative and research spheres in terms of the ability to bring about novel results (Dong and Srivastava, 2013). Linked administrative data can benefit healthcare organisations financially and enhance clinical management practices (Harron et al., 2017). Researchoriented data sets when combined, can drive novel research in areas ranging from population-level studies (Medalia et al., 2019), to results studies of eating disorders for example (Demmler et al., 2020).

Given that it is medical data that is being linked, there are a particular set of privacy and ethico-legal concerns that constrain the way in which data can be linked (Willison et al., 2008; Haddow et al., 2011). Linked data, derived from two sub-sets of information about a given entity being combined, may disclose otherwise private information about that entity (Zheng et al., 2018). Techniques exist that can preserve privacy during linkage of records (Schnell et al., 2009) or offer pooled analysis of results without revealing private information (Wolfson et al., 2010). Even where disclosure of information does not take place though there is the issue of whether explicit patient consent should be required in linking personal data (Xafis, 2015). A further point at which information leakage can occur during data linkage is if a subset of identifiers in one data set is not present in the dataset being linked to (Niedermeyer et al., 2014). If that subset of identifiers is then passed to the corresponding data provider, in the act of linking the two sets, then information not previously known by the second provider has now been revealed. This information in itself can be extremely sensitive (for example the presence of a patient identifier in a particular sexual health registry).

In order to ensure privacy protections and to enact ethico-legal constraints on the construction and distribution of linked data sets, most data linkage systems employ a Trusted Third Party (TTP) mechanism (Niedermeyer et al., 2014). A TTP is an organization or software system separate and distinct from the parties providing data, which assumes responsibility for matching identifiers across data sets, imposing privacy preserving constraints on the data such as the pseudonymization and anonymization of relevant identifiers, and the distribution linked data sets to the requestor of the data (Harron et al., 2016). The reliance on TTPs to facilitate data linkage also exposes them as a singular point of failure in terms of system security and reliability (Durham et al., 2013). While there exist approaches to removing the use of a TTP through probabilistic methods (Lazrig et al., 2018) for example, data linkage systems that require a) accurate matching between identifiers and b) guarantees around preventing disclosure of private information (both features particularly prevalent in medical linkage systems), are not suited to probabilistic approaches and must again rely on a TTP model.

In the remainder of this paper we present a reference architecture of a system that allows the role of the trusted third party in data linkage applications to be replaced by a decentralized and trust-less mechanism provided through the use of blockchain technology.

2 MATERIALS AND METHODS

Using an architecture that relies on a trusted third party for data linkage exposes a number of weaknesses: the TTP can be a single point of failure, there is potential for trust to be compromised either maliciously or accidentally and it presents a single attack surface for security adversaries. A system that could remove this reliance on a trusted third party would represent a step forward in terms of the usability and applicability of such systems. We looked to design a reference architecture that would serve as a guide for designing more complex systems based around the need for participating data providers to link datasets to provide to data consumers (such as research organisations). In order to proceed with this design we constructed two related sets of requirements for the architecture. First, from an analysis of the factors that need to be accounted for in designing such a system, we drew out a set of requirements in terms of the general principles that such a system should follow. Second, taking as our starting point the use of blockchain technology as the mechanism by which we would enable the trust-less linking of data, we derived the core technical requirements that the system should exhibit.

2.1 Generic System Requirements

The development of a system which aims to enable the sharing of patient data that is otherwise held and controlled by individual healthcare organisations needs to meet a set of requirements that are based around the value needs of the various stakeholders involved in the control of the records. The requirements of the various groups of stakeholders (management, db administrators, etc) involved in the management of patient data can generally be seen as being fairly homogenous between those groups–i.e., there are a consistent set of principles that drive localized use and management of patient records. From our analysis of stakeholder needs we drew out the following set of requirements pertaining to general features that a trust-less data linkage platform should exhibit.

- Historical–For an existing system with as rich an historical background as medicine, simple inertia and a natural conservatism play a large part in determining the extent that radical change can be tolerated. Elements of the system designed to accommodate or allay this resistance to change are classified as historical requirements. In the case of medical record provision and access meeting these historically driven requirements translate to maintaining as much of existing provisioning as possible; radical redesign impacting on existing infrastructure is to be avoided where possible.
- Political–Stemming from, but distinct to, historically driven requirements are those that can be categorized as politically motivated. The subject of electronic health records, particularly aspects relating to security, is highly politically charged and many aspects of electronic healthcare system design need to cater to the fact that often appearance trumps practical issues. Relating in particular to the adoption of a new (and potentially controversial) technology such as blockchain, such political considerations translate into a need to minimize any over reliance or over emphasis on this technology where not strictly necessary.

- Social–Even within the data-centric world of electronic patient records the use of those records is driven by human interaction. Doctors make use of the records to treat patients, with the protection of the patient's wellbeing being at the heart of medical practice. Human, or social, pressures therefore play a significant role in the potential use of electronic records. If systems supporting such records are perceived to have a negative net impact on individual needs (such as the protection of privacy and the avoidance of commercial exploitation) then, regardless of technical merit, they will tend not to be adopted.
- Commercial–Even within the UK's publicly funded monolithic NHS commercial sensibilities, driven by the existance of an internal market, impact on the desire of healthcare organisations to share data. In healthcare systems with more of an emphasis on private healthcare provision (the US healthcare system being the natural example) such commercial concerns come even more to the fore. Data linkage systems act as glue between existing medical data sets, and such data sets carry commercial value. If the design of a system that enables data linkage removes direct control of data from the data provider organisations, then it can also have the effect of damaging commercial concerns. As such the system design should not remove existing ultimate control of access to data from data providers.
- Legal–Where patient data is captured in electronic form the protection and confidentiality of that information is recognized as being of primary importance. As such there are often detailed legal frameworks put in place in order to protect such data. In the United Kingdom the primary entry point to understanding the protection of electronically stored sensitive data is the 1998 Data Protection Act (Data Protection Act, 1998) and GDPR requirements (Goddard, 2017).
- Security–The primary concern of a system dealing with the storage, access and transmission of patient data is that of security. Any system dealing with provisioning healthcare data should have security as its primary concern.
- Trust–Trust, as opposed to security, issues involve the results of the perception of a system's security on its willingness to be used. That is, even the perception that a new system will be less secure than an existing one can lead to a rejection of that new system. Stakeholders are also often unwilling to transfer trust to another party, even if that other party is a representative of the same stakeholder group. Ensuring trust exists on both a technical and perceptual level is a key requirement.

Looking at this set of generic principles that a data linkage enabling architecture should meet we distilled the following set of design goals:

- The prevention of disclosure of data is paramount.
- The set of data shared must be minimal.
- Data retained and used must be on a 'just in time' rather than 'just in case' basis.
- Data held locally is controlled locally.
- Obligations defined within the system must be translatable to legal obligations.
- A site is the solely responsible for determining all of its actions within the system.
- A site can leave exit the system instantly, with no action or input required from any other parties.

2.2 Technical Requirements

On the technical side the primary concern of the design of the system was in the removal of reliance on a trusted third party enabling the linkage of medical data sets. We found that utilizing the core features and underlying principals of blockchain technology would meet this key requirement. As outlined in **section 1.1** the core features provided by blockchain as a technological platform and protocol are security, decentralization, consensus and immutability. From matching against these core features we derived the following set of requirements:

- Security: the system must ensure that invalid entities cannot request data *via* the system and data about participating entities and the results of requests they have made should remain private.
- Decentralization: No single organization or entity should be responsible for the provision of linkage services and access to or modification of the service itself should be on the protocol level rather than the responsibility of an individual entity within the system.
- Consensus: participants within the system should, at any given point in time, agree to what there responsibilities are with respect to the provision of data, and should have a common view as to what actions other entities are and have been performing.
- Immutability: The public record of actions that has been facilitated by the system should remain permanently auditable and a matter of public record. We designed the reference architecture presented in **section 3** around these two core sets of requirements.

2.3 Implementation

In order to demonstrate the validity of our design we implemented a working prototype of the reference architecture. This system was implemented on a public test net of the Ethereum blockchain. Smart contracts encoding the role and actions of the decentralized trusted third party were written in the Solidity programming language version 0.7.1 with surrounding web-based functionality written in the javascript programming language, in particular utilizing the Web3.js library for interacting with the Ethereum blockchain.

Deploying and running smart contract based applications on the live Ethereum main network accrues 'gas' fees–market valued payments made in terms of the number of primitive operations executed by a call made to a smart contract. Gas fees act to constrain the resources consumed by the nodes running the network in terms of the processing required to compute the outputs of smart contracts at any given time. As such running live applications is generally unfeasible in terms of costs unless the

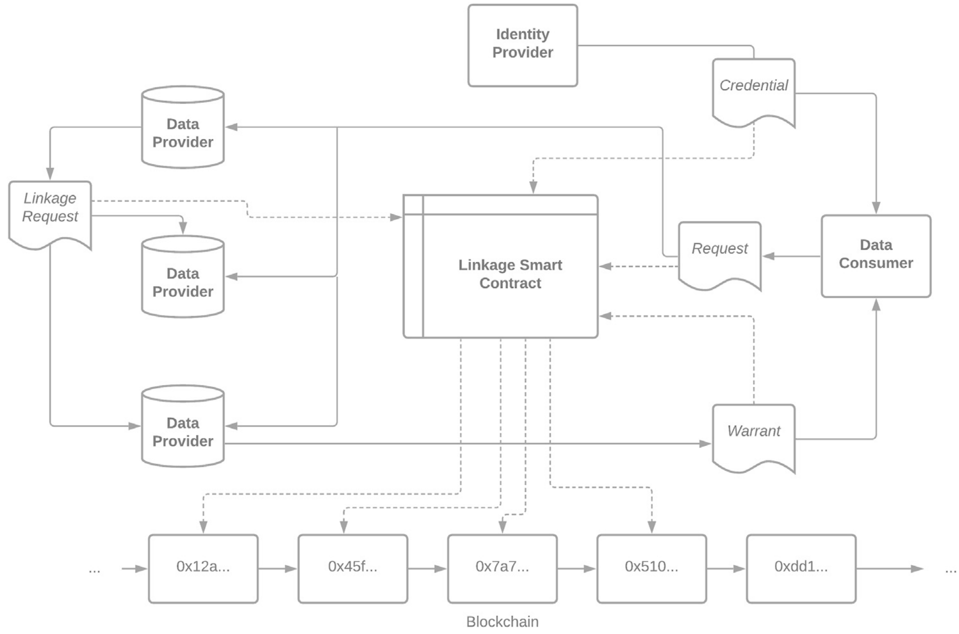

FIGURE 1 | A Reference Architecture For Blockchain Native Data Linkage.

running of the application is backed by a sustainable business model or income stream (Ranganthan et al., 2018). As an alternative, and for testing purposes, several test networks exist, identical in terms of software and operation to the main Ethereum network, but with the corresponding Ether (the coins, or units of value, backing the network and used to pay gas fees) freely or easily obtainable, as opposed to priced on the open market. We deployed, ran and tested the reference implementation against the Ropsten test network. This allowed us to deploy and run test versions of our smart contract architecture without incurring cost. The implications of running a version of the system against the main network are discussed in 4.

3 RESULTS

In this section we present the reference architecture for blockchain enabled trust-less data linkage. The architecture is derived from the requirements analysis outlined in **section 2** and a proof-of-concept implementation of this architecture was engineered and deployed on an Ethereum test network (using the tooling outlined in 2.3). This section presents the architectural design along with a description of the components of the architecture and their roles within the system. We then outline the flow of data that the system enables and describe the interaction of the various components. The results of the design in terms of its conformance to the previously derived requirements are then examined. Potential future work, consequences of the design process and issues regarding the proof-of-concept implementation are further looked at in **section 4**.

3.1 Reference Architecture

Figure 1 gives a graphical overview of the reference architecture that we have specified, designed and implemented. The components of the diagram and the actions that it enables in terms of the requesting and linking of a data query are described in the following sections.

The components of this architecture are as follows.

3.1.1 Entities

The following entities are specified by the reference architecture:

3.1.1.1 Data Consumer

A *Data Consumer* is an entity or organization which wishes to receive linkable data sets from participating *Data Providers* in the system. The role of the *Data Consumer* within the system is to receive valid credentials for participation from an Identity Provider and to publish Requests to *Data Providers* to provide data. The actual sending and receiving of the data itself happens, post verification and linkage, outside the bounds of this system.

3.1.1.2 Data Provider

A Data Provider is an entity that is responsible for holding and distributing a given healthcare dataset. The role of the Data Provider within the system is to process Requests coming from Data Consumers, to process the results of those Requests and to compute subsets of linkable identifiers through interaction with other Data Providers within the system.

3.1.1.3 Blockchain

The underlying Blockchain acts as a canonical source of truth, hosts the linkage smart contract and is used to record auditable information regarding actions that have been performed within the system. In our reference implementation of this design the particular blockchain used was an Ethereum test network.

3.1.1.4 Linkage Smart Contract

The Linkage Smart Contract is the piece of decentralized code that is relied upon to coordinate and process requests passing through the system. It publishes relevant information to the underlying blockchain supporting the system, and provides a canonical mechanism for sending data objects between participants in the system.

3.1.1.5 Identity Provider

An Identity Provider is an entity whose role it is to verify the legitimacy of participating actors within the system. There can be one or more Identity Providers acting within the system. An Identity Provider grounds the trust that Data Providers have that Data Consumers are legitimate entities, and *vice versa.* A Data Consumer or Data Provider does not need to rely on the verification provided by a given Identity Provider, but must rely on at least one such provider acting within the system in order to be able to interact with other participating entities.

3.1.2 Data Objects

The following data objects are specified by the reference architecture:

3.1.2.1 Credential Object

A Credential object is a signed message provided by an Identity Provider attesting to the legitimacy of a participating Data Consumer. The content of this object is published *via* the Linkage Smart Contract.

3.1.2.2 Request Object

The purpose of a Request Object is to specify a request that is being sent by a Data Consumer to a Data Provider. The object consists of:

- A query, being a specification of the query that the originating data consumer wishes to run.
- A list of credential references, specifying the signed credential objects that can be used to confirm the validity of the originator of the request.
- A list of data provider identifiers that specify the data providers that the data consumer wishes to run the query against.
- A unique identifier for that object, which can be used to index request objects, in particular to allow the identification of previous requests from data consumers that may impact the validity of a current request (for example to protect against statistical disclosure of identifiable information from pseudonymized data sets obtained from multiple sources).

3.1.2.3 Linkage Request Object

The role of a Linkage Request object is to initiate interaction between two Data Providers such that they can compute a subset of identifiers that are common to both Data Providers and to constrain the future results returned through enacting the query contained in a given Request. A Linkage Request object consists of:

- Request identifier, a link to the request object against which the linkage is taking place.
- An originating provider id specifying the data provider sending this linkage request.
- A receiving provider id specifying the data provider that the originator wishes to compute a linked identifier set against.

3.1.2.4 Warrant Object

A Warrant is the final specification, given to a Data Consumer by a Data Provider enabling them to query the data held by the Data Provider. It contains a signed message originating from the Data Provider and a Request identifier.

3.2 Data Linkage Procedure

The general process for linking data from data providers following a request from a data consumer proceeds as follows:

1. Prior to the initialization of a data linkage request the data consumer obtains a valid credential object from an identity provider. the details of the credential are recorded on chain.
2. The data consumer constructs a request object pertaining to a query for a particular set of identifiers and containing a signed reference to a valid set of credentials and a set of identifiers of data providers.
3. The data consumer publishes the request object *via* the linkage smart contract.
4. Following the publication of a Request object, each Data Provider for which that Request pertains does the following:

a. Consumes the request.
b. Independently verifies that the published credentials within the request are valid.
c. Internally computes the set of identifiers that satisfy the query portion of the request.
d. Publishes a linkage request object *via* the linkage smart contract.

5 Following the registration of a linkage request object, each data provider which has also internally computed a set of identifiers pertaining to the query id of the linkage request does the following:

a. Consumes the linkage request object.
b. Sends an encrypted set of identifiers to the originator of the linkage request.
c. Receives back a subset of those identifiers that are the union of those identifiers and the set computed by the originator of the linkage request.
d. If the request object to which the linkage request relates contains any data provider identifiers against which the data provider has not yet computed a linked set of identifiers then the data provider publishes a corresponding linkage request object.

e. Once the data provider has computed subsets against all other data providers referenced in the query, then a warrant object is published *via* the linkage smart contract that contains signed permission for the data consumer to send a request for data to that data provider.

6. The originating data consumer then:

a. Consumes any published warrant objects.
b. Sends a data query request to the relevant data consumer in an out of band channel to the data provider of the warrant.
c. Receives the relevant query results set.

The mechanisms behind the production and matching of patient identifiers within a given linkage query may vary depending on the type of data being linked against and the general system requirements pertaining to holding that data (for example any legal constraints on the retention and representation of such identifiers. In our proof-of-concept reference implementation we assumed that identifiers were represented in a universal format across the platform.

3.3 Adherence to Requirements

In **section 2** we outlined two sets of requirements that our reference architecture design needed to exhibit to, falling into categories of the generic and the technical. Here we outline the ways in which the architecture adheres to these requirements.

3.3.1 Generic Requirements

- The prevention of disclosure of data is paramount. Within the operation of the system data at all times remains under the control of the Data Provider entity that supplies that data. Additionally the operation of the system is based around the computation of subsets of identifiers that are only those linkable with the data sets held by other Data Providers. It is thus a core principal of the system that it acts to prevent the disclosure of unnecessary data.
- The set of data shared must be minimal. The role of the Linkage Request object within the system flow is to allow Data Provider entities to compute the minimum releasable subset of identifiers that they should provide to a Warranted Data Consumer. As such this requirement is met.
- Data retained and used must be on a 'just in time' rather than 'just in case' basis. While broadly the retention of data by the Data Consumer falls outside the remit of the system as specified here, the fact that there are immutably recorded and formally specified Warrants specifying the right to request data, it would be possible to extend the system such that the Warrant also specified the obligations of the Data Consumer with respect to the retention of data and to potentially translate this into a legal obligation (see the point below).
- Data held locally is controlled locally. At every point in the operation of the system it is under the control of the Data Provider as to whether they ultimately release data to the requesting Data Consumer. The Data Provider is responsible and has control over the decision to trust the Identity Provider, the computation of releasable sets of identifiers through interaction with other Data Providers, the issuing of the signed Warrant that is supplied by the Data Consumer in requesting data, and ultimately the final act of releasing data against that Warrant itself. Hence, at all points, the data that is held by the provider is controlled by the provider.
- Obligations defined within the system must be translatable to legal obligations. Although no component of the system explicitly exhibits the principle of this requirement, the fact that all actions performed within the system, particularly the action of publishing a specific Warrant pertaining to the granting of a Data Consumer the right to request data from the issuing Data Provider, lends itself to future work in mapping the specification of this Warrant to a legally sound contractual obligation specified outside the bounds of the system.
- A site is the solely responsible for determining all of its actions within the system. No action within the operation of the system is automated, and all actions performed by Data Providers are initiated and processed by themselves. As such this requirement is met.
- A site can leave exit the system instantly, with no action or input required from any other parties. Given that it is under the control of the Data Providers within the system to accept and process a request for data, and there is no system specific obligation to continue to act after any given previous action then leaving the system is simply a case of no longer participating in it.

In conclusions all of the above generic requirements were either met in the design and implementation of the system, or are outside the bounds of the present specification, not prevented by it and could be enabled through extension of the current specification.

3.3.2 Technical Requirements

The following technical requirements were derived for the system:

- Security: the system must ensure that invalid entities cannot request data *via* the system and data about participating entities and the results of requests they have made should remain private. The act of preventing invalid participants interacting with the system is assumed by the role of the Identity Provider which signs Credentials for participants. The results of requests instantiated through the system are encoded and transported outside the bounds of the system, and the communication of intersecting sets of identifiers is encrypted by the originating Data Provider entities.
- Decentralization: No single organization or entity should be responsible for the provision of linkage services and access to or modification of the service itself should be on the protocol level rather than the responsibility of an individual entity within the system. Actions that take place within the bounds of the system are coordinated by the Linkage Smart Contract. The fact that this smart contract is hosted on and interacted with *via* the underlying Ethereum blockchain

ensures that the coordinating actions of the system remain decentralized and outside the control of a third party actor.

- Consensus: participants within the system should, at any given point in time, agree to what there responsibilities are with respect to the provision of data, and should have a common view as to what actions other entities are and have been performing. The actions of the system are specified as a protocol, with the required actions of entities within the system clearly specified at a given time. The common view of the current and previous actions of other entities participating in the system is coordinated through the Linkage Smart Contract and results of interaction written to the underlying Blockchain instance. These components underpin the element of consensus within the system.
- 1Immutability: The public record of actions that has been facilitated by the system should remain permanently auditable and a matter of public record. Again the role of the Linkage Smart Contract is to coordinate actions of system components and to record the results on the underlying Blockchain instance. Where that blockchain is public then the requirement of immutability is met.

It is a natural outcome of designing the system based on an underlying blockchain implementation that this set of technical requirements is met by the architectural specifications.

4 DISCUSSION

We have presented a reference architecture for a data linkage model that replaces the need for a centralized trusted third party with a blockchain mediated trust-less system. Core features of blockchain, particularly decentralization, have allowed us to replace the trusted third party component present in most data linkage platforms. The use of blockchain technology to replace a central trusted authority with a decentralized system mirrored the primary driver for the original development of blockchain technology through Bitcoin. In constructing the requirements for the design of this system we mapped our derived requirements against the core underlying features provided by blockchain technology (security, decentralization, consensus and immutability). This mapping between high-level systemic features begins to illustrate a conceptual framework that allows for both assessing the genuine worth of blockchain-oriented projects and systems, and for informing the design of such systems. Given a propensity of some projects to buy into the hype of blockchain and to base system designs around blockchain technology where it does not perhaps offer any particular benefit, then formulating an analytic framework around the way in such features are matched against a potential use case may be interesting.

While replacing a trusted third party in a system such as medical data linkage may increase trust in the system in a technical sense, by removing a potentially insecure or manipulatable component of the system, there is still the issue of perceived trust in the system as a whole. Given that the role of medical data linkage systems is in essence handling highly sensitive personal data, any data linkage system must engender trust not just on that technical level but also on the public level. Potentially the use of blockchain technology could have a negative impact on perceived public trust in a system on two fronts. Firstly, as a new technology in general, despite any purely technical merits, there can be more inherent distrust in the use of that technology (Mittelstadt, 2017). With blockchain being a relatively new and groundbreaking technology the may be natural tendency for critical systems handling sensitive data to shy away from the use of such technologies for purely perceptual issues. Additionally a theme of negative media coverage of blockchain and associated cryptocurrencies could negatively impact willingness to adopt blockchain technology in general, apart from it being a generically new technology. Future research examining both of these factors will be crucial in the integration of blockchain technologies as components of existing systems, particularly in the case of health informatics.

The architecture we have described revolves around the roles of data provider (an entity acting as the custodian and distributor of data) and data consumer (an entity wishing to process portions of data) as the key actors in the exchange of medical data. An additional consideration is the role of the individual patient in such a system. We have implicitly assumed that these data consumers would at times be acting as direct intermediaries between patients and data providers (for example as vetted apps allowing patients access to portions of their data) and serving a filtering and management role between individuals and their data. . Where the patient assumes the role of data consumer more directly (i.e., can request data directly from a provider for their own use), the specific implementation of the architecture may need to be changed to accommodate both the increased load on the system brought about by there being a large number of consumer entities (i.e., the individual patients) and the potential of additional ethico-legal requirements placed on the system though the direct involvement of patients, which would be dependant on the jurisdictions in which the system was being deployed.

The reference implementation we built of this trust-less data linkage architecture was developed using the Ethereum blockchain and associated tooling. The use of the public Ethereum network for such a system raises a number of points regarding costs, alternatives, business model implications and performance issues. While we deployed our reference architecture against one of the Ethereum test networks, meaning that there was no incurred costs in terms of deploying and running the smart-contract code base, such networks offer no guarantees of availability or persistence of data, nor the economically grounded security guarantees offered by the main network. As such deploying the architecture as it stands for use in a real world scenario would incur the costs associated with running applications against the main Ethereum network. In practice then the long term deployment and use of the system would require a sustainable business model, either in terms of support from a research or healthcare organization or potentially a pay-per-use or pay-per-volume pricing model. However the market based nature of the fee structure of the Ethereum network is such that the cost of performing smart contract operations at a given time is in essence determined by the current level of use of the network. At times where the network is under heavy load fees can increase significantly. This potentially hinders the consistent use of the network as the foundation of a research tool. Going hand-in-hand with this issue is the fact performance (measured in terms of the time it takes to process a transaction on the network) is not guaranteed, and is a factor of both market-based pricing and the fact that the scalability of the network as a whole is limited and there are ongoing development efforts to address it (Park et al., 2020). Both of these factors suggest that use of the Ethereum network as an application layer may be better suited to applications that are immediately transactional in terms of cost (such as distributed market places, for example) as opposed to service-oriented applications were costs are assumed to be more constant. Alternative blockchain networks and implementations exist, potentially offering better solutions in terms of cost and efficiency. In particular we have looked at the relatively new Avalanche protocol, and its associated public blockchain deployment and novel consensus mechanism (Rocket, 2018), which could potentially address these issues. Further explorations of the trade-offs between blockchain implementations and the implications that their use would have on the design and running of data linkage systems is an avenue for future research.

A consideration to be had in a practical deployment of the architecture would be whether to deploy it on a fully public blockchain instance or a private or hybridized one. There is explicitly nothing in the architecture that precludes a public deployment (i.e., the recording, on-chain, of identifiable personal information), but ethico-legal considerations and public perception may point toward the use of a private instance (where participants in the network are invited in and blockchain data is not publicly accessible) may be more appealing. The choice of Ethereum as an implementation and deployment platform may mean that there is an implicit choice made with regards to the underlying consensus mechanism that such a practical deployment of the architecture would use. Ethereum currently uses proof-of-work consensus where ‘miners’ spend computational resource competing to produce canonical blocks in the Ethereum blockchain. While this is justifiable in the case of a public blockchain where the use of the computational resource is used as a means to underpin the security of the system, if the architecture were to be deployed on a private instance of a blockchain, the use of such a proof-of-work consensus mechanism to should be considered overly complex and perhaps wasteful. In this case alternative implementations utilizing more efficient consensus mechanisms would be more applicable, for example the Hyperledger technology stack, which is more suited to such deployments (Androulaki et al., 2018).

The research we have presented has demonstrated the feasibility of applying blockchain technologies to problem spaces which benefit from the removal of entities or components that rely on an inherent degree of trust to maintain the integrity of the system as a whole. In general the underlying features provided by blockchain technology, namely security, decentralization, consensus and immutability can provide a conceptual framework for leveraging blockchain as a tool for addressing outstanding system design problems. In the particular instance of data linkage, the application of blockchain technology can provide security, consensus and immutibility to an area that is crucial for pushing forward research while maintaining patient privacy.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

JC led the writing of the paper, and lead on the design and writing of the technical architecture with design and input from JA and GL. GD, ND, SD, SH, and MH contributed to the design, analysis and testing of the reference architecture. All authors drafted, edited, and contributed to the design of the paper.

FUNDING

This work was supported by the Engineering and Physical Sciences Research council as part of the investigator led Digital Economy Standard Grant Ref: EP/S02770X/1.

REFERENCES

Agbo, C., Mahmoud, Q., and Eklund, J. (2019). Blockchain Technology in Healthcare: a Systematic Review. *Healthcare* 7, 56. doi:10.3390/healthcare7020056

Androulaki, E., Barger, A., Bortnikov, V., Cachin, C., Christidis, K., De Caro, A., et al. (2018). Hyperledger Fabric: a Distributed Operating System for Permissioned Blockchains. In Proceedings of the thirteenth EuroSys conference. 1–15.

Angraal, S., Krumholz, H. M., and Schulz, W. L. (2017). Blockchain Technology: Applications in Health Care. *Circulation: Cardiovascular quality and outcomes* 10, e003800. doi:10.1161/circoutcomes.117.003800

Avizienis, A., Laprie, J.-C., Randell, B., and Landwehr, C. (2004). Basic Concepts and Taxonomy of Dependable and Secure Computing. *IEEE Trans. Dependable Secure Comput.* 1, 11–33. doi:10.1109/tdsc.2004.2

Azaria, A., Ekblaw, A., Vieira, T., and Lippman, A. (2016). Medrec: Using Blockchain for Medical Data Access and Permission Management. 2016 2nd International Conference on Open and Big Data (OBD) IEEE, 25–30.

Brechtelsbauer, E. D., Pennell, B., Durham, M., Hertig, J. B., and Weber, R. J. (2016). Review of the 2015 Drug Supply Chain Security Act. *Hosp. Pharm.* 51, 493–500. doi:10.1310/hpj5106-493

Buterin, V. (2016). What Is Ethereum? *Ethereum Official Webpage*. Available at:http://www.ethdocs.org/en/latest/introduction/what-is-ethereum.html (Accessed May 14, 2021).

Clauson, K. A., Breeden, E. A., Davidson, C., and Mackey, T. K. (2018). Leveraging Blockchain Technology to Enhance Supply Chain Management in Healthcare: an Exploration of Challenges and Opportunities in the Health Supply Chain. *Blockchain in healthcare today* 1, 1–12. dx.doi.org/10.30953/bhty.v1.20

Cyran, M. A. (2018). Blockchain as a Foundation for Sharing Healthcare Data. *Blockchain in Healthcare Today* 1, 1–6. doi:10.30953/bhty.v1.13

Data Protection Act (1998). Available at: https://www.legislation.gov.uk/ukpga/1998/29 (Accessed May 14, 2021).

Demmler, J. C., Brophy, S. T., Marchant, A., John, A., and Tan, J. O. A. (2020). Shining the Light on Eating Disorders, Incidence, Prognosis and Profiling of Patients in Primary and Secondary Care: National Data Linkage Study. *Br. J. Psychiatry* 216, 105–112. doi:10.1192/bjp.2019.153

Dong, X. L., and Srivastava, D. (2013). Big Data Integration. 2013 IEEE 29th international conference on data engineering (ICDE). IEEE, 1245–1248.

Durham, E. A., Kantarcioglu, M., Xue, Y., Toth, C., Kuzu, M., and Malin, B. (2013). Composite Bloom Filters for Secure Record Linkage. *IEEE Trans. Knowl. Data Eng.* 26, 2956–2968. doi:10.1109/TKDE.2013.91

Goddard, M. (2017). The Eu General Data Protection Regulation (Gdpr): European Regulation that Has a Global Impact. *Int. J. Mark. Res.* 59, 703–705. doi:10.2501/ijmr-2017-050

Golosova, J., and Romanovs, A. (2018). The Advantages and Disadvantages of the Blockchain Technology. 2018 IEEE 6th workshop on advances in information, electronic and electrical engineering (AIEEE). IEEE, 1–6.

Haddow, G., Bruce, A., Sathanandam, S., and Wyatt, J. C. (2011). 'Nothing Is Really Safe': a Focus Group Study on the Processes of Anonymizing and Sharing of Health Data for Research Purposes. *J. Eval. Clin. Pract.* 17, 1140–1146. doi:10.1111/j.1365-2753.2010.01488.x

Harron, K., Dibben, C., Boyd, J., Hjern, A., Azimaee, M., Barreto, M. L., et al. (2017). Challenges in Administrative Data Linkage for Research. *Big data & society* 4, 2053951717745678. doi:10.1177/2053951717745678

Harron, K., Mackay, E., and Elliot, M. (2016). *An Introduction to Data Linkage*. Available at: http://eprints.ncrm.ac.uk/4282/ (Accessed May 14, 2021).

Hoffman, S. (2015). Citizen Science: the Law and Ethics of Public Access to Medical Big Data. *Berkeley TechLJ* 30, 1741. Available at: https://www.jstor.org/stable/26377581 (Accessed May 14, 2021).

Hölbl, M., Kompara, M., Kamišalić, A., and Nemec Zlatolas, L. (2018). A Systematic Review of the Use of Blockchain in Healthcare. *Symmetry* 10, 470. doi:10.3390/sym10100470

Holman, C. D. A. J., Bass, J. A., Rosman, D. L., Smith, M. B., Semmens, J. B., Glasson, E. J., et al. (2008). A Decade of Data Linkage in Western australia: Strategic Design, Applications and Benefits of the Wa Data Linkage System. *Aust. Health Rev.* 32, 766–777. doi:10.1071/ah080766

Johnson, O. A., Fraser, H. S. F., Wyatt, J. C., and Walley, J. D. (2014). Electronic Health Records in the uk and usa. *Lancet* 384, 954. doi:10.1016/s0140-6736(14)61626-3

Lazrig, I., Ong, T. C., Ray, I., Ray, I., Jiang, X., and Vaidya, J. (2018). Privacy Preserving Probabilistic Record Linkage without Trusted Third Party. 2018 16th Annual Conference on Privacy, Security and Trust (PST) . IEEE, 1–10.

Leeming, G., Cunningham, J., and Ainsworth, J. (2019). A Ledger of Me: Personalizing Healthcare Using Blockchain Technology. *Front. Med.* 6, 171. doi:10.3389/fmed.2019.00171

Mattke, J., Hund, A., Maier, C., and Weitzel, T. (2019). How an Enterprise Blockchain Application in the Us Pharmaceuticals Supply Chain Is Saving Lives. *MIS Q. Exec.* 18. doi:10.17705/2msqe.00019

Medalia, C., Meyer, B. D., O'Hara, A. B., and Wu, D. (2019). Linking Survey and Administrative Data to Measure Income, Inequality, and Mobility. *International Journal of Population Data Science* 4. doi:10.23889/ijpds.v4i1.939

Miraz, M. H., and Ali, M. (2018). Applications of Blockchain Technology Beyond Cryptocurrency. AETiC 2, 1–6.

Mittelstadt, B. (2017). Ethics of the Health-Related Internet of Things: a Narrative Review. *Ethics Inf. Technol.* 19, 157–175. doi:10.1007/s10676-017-9426-4

Nagasubramanian, G., Sakthivel, R. K., Patan, R., Gandomi, A. H., Sankayya, M., and Balusamy, B. (2020). Securing E-Health Records Using Keyless Signature Infrastructure Blockchain Technology in the Cloud. *Neural Comput. Appl.* 32, 639–647. doi:10.1007/s00521-018-3915-1

Nakamoto, S. (2009). *Bitcoin: A Peer-To-Peer Electronic Cash System*. Manubot: Tech. rep.

Niedermeyer, F., Steinmetzer, S., Kroll, M., and Schnell, R. (2014). *Cryptanalysis of Basic Bloom Filters Used for Privacy Preserving Record linkage*Working Paper Series, No. WP-GRLC-2014-04. Regensburger, Nürnberg: German Record Linkage Center.

Park, D., Zhang, Y., and Rosu, G. (2020). End-to-end Formal Verification of Ethereum 2.0 Deposit Smart Contract. International Conference on Computer Aided Verification . Springer, 151–164. doi:10.1007/978-3-030-53288-8_8

Ranganthan, V. P., Dantu, R., Paul, A., Mears, P., and Morozov, K. (2018). A Decentralized Marketplace Application on the Ethereum Blockchain. 2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC). IEEE, 90–97. doi:10.1109/cic.2018.00023

Rocket, T. (2018). Snowflake to Avalanche: A Novel Metastable Consensus Protocol Family for Cryptocurrencies. Available [online].(Accessed: : *4-12-2018*)

Schnell, R., Bachteler, T., and Reiher, J. (2009). Privacy-preserving Record Linkage Using Bloom Filters. *BMC Medical Informatics and Decision Making* 9, 1–11. doi:10.1186/1472-6947-9-41

Underwood, S. (2016). Blockchain beyond Bitcoin. *Commun. ACM* 59, 15–17. doi:10.1145/2994581

Vazirani, A. A., O'Donoghue, O., Brindley, D., and Meinert, E. (2019). Implementing Blockchains for Efficient Health Care: Systematic Review. *J. Med. Internet Res.* 21, e12439. doi:10.2196/12439

Williams-Grut, O. (2016). Estonia Is Using the Technology behind Bitcoin to Secure 1 Million Health Records. *Bus Insid*. Available at: http://www.businessinsider.com/guardtime-estonian-health-recordsindustrial-blockchain-bitcoin-2016-3?r=UK=T (Accessed May 14, 2021).

Willison, D. J., Swinton, M., Schwartz, L., Abelson, J., Charles, C., Northrup, D., et al. (2008). Alternatives to Project-specific Consent for Access to Personal Information for Health Research: Insights from a Public Dialogue. *BMC Medical Ethics* 9, 1–13. doi:10.1186/1472-6939-9-18

Wolfson, M., Wallace, S. E., Masca, N., Rowe, G., Sheehan, N. A., Ferretti, V., et al. (2010). DataSHIELD: Resolving a Conflict in Contemporary Bioscience-Pperforming a Pooled Analysis of Individual-Level Data without Sharing the Data. *Int. J. Epidemiol.* 39, 1372–1382. doi:10.1093/ije/dyq111

Woods, J., and Iyengar-Emens, R. (2019). Blockchain to Secure a More Personalized Pharma. *Genetic Engineering & Biotechnology News* 39, 27–29. doi:10.1089/gen.39.01.08

Xafis, V. (2015). The Acceptability of Conducting Data Linkage Research without Obtaining Consent: Lay People's Views and Justifications. *BMC Medical Ethics* 16, 1–16. doi:10.1186/s12910-015-0070-4

Zhang, M., and Ji, Y. (2018). Blockchain for Healthcare Records: A Data Perspective. *PeerJ Preprints* 6, e26942v1. 10.7287/peerj.preprints.26942v1.

Zheng, X., Cai, Z., and Li, Y. (2018). Data Linkage in Smart Internet of Things Systems: a Consideration from a Privacy Perspective. *IEEE Commun. Mag.* 56, 55–61. doi:10.1109/mcom.2018.1701245

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.