Data & Policy (2024), 6: e9
doi:10.1017/dap.2024.4

RESEARCH ARTICLE

Decentralized crowdsourcing medical data sharing platform to obtain chronological rare data

Stefan Kambiz Behfar[1,2] and Jon Crowcroft[2,3]

[1]Department of Information Systems, Geneva School of Business Administration (HES-SO Genève), Geneva, Switzerland
[2]Department of Computer Science and Technology, University of Cambridge, Cambridge, UK
[3]Alan Turing Institute, London, UK
Corresponding author: Stefan Kambiz Behfar; Email: stefan-kambiz.behfar@hesge.ch

Received: 25 July 2023; **Revised:** 20 November 2023; **Accepted:** 12 December 2023

Keywords: chronological data; crowdsourcing; decentralized app; medical data sharing; rare and elusive data

Abstract

Researchers have encountered many issues while studying rare illnesses such as lack of information, limited sample sizes, difficulty in diagnosis, and more. However, perhaps the biggest challenge is to recruit a large enough sample size for clinical studies; at the same time, obtaining chronological data for these patients is even more difficult. This has urged us to implement a decentralized crowdsourcing medical data sharing platform to obtain chronological rare data for certain diseases, providing both patients and other stakeholders an easier and more secure way of trading medical data by utilizing blockchain technology. This facilitates the obtention of the most elusive types of health data by dynamically allocating extra financial incentives depending on data scarcity. We also provide a novel framework for medical data cross-validation where the system checks the volunteer reviewer count. The review score depends on the count, and the more the reviewers, the bigger the final score. We also explain how differential privacy is used to protect the privacy of individual medical data while enabling data monetization.

Policy Significance Statement

The policy significance of medical data sharing and monetization lies in harnessing the potential of health-care data to advance medical research, improve patient outcomes, and drive innovations in health care. By establishing comprehensive and ethically sound policies, we can foster a secure and transparent ecosystem that empowers patients to control their data while facilitating responsible data sharing for research and monetization purposes. These policies will play a pivotal role in promoting collaboration between stakeholders, ensuring data privacy and security, encouraging fair compensation for data contributors, and accelerating the development of personalized medicine and evidence-based health-care interventions.

1. Introduction

There are numerous issues with conducting research on diseases, particularly rare disorders. First, clinical trial data are very expensive and almost impossible to get for non-medical researchers or external practitioners. This also causes an ethical debate on the ownership of the data. At the moment, it feels like clinics are the owners of their patients' data while, in truth, the patients actually are the owners of the said data (Koskinen et al., 2016). Thus, if a patient wants to share his data, he should be able to do it as one

https://doi.org/10.1017/dap.2024.4 Published online by Cambridge University Press

stakeholder in this environment or ecosystem (Zhang et al., 2022). Indeed, some of the main challenges regarding health data scarcity are:

- Getting data from a large spectrum of patients (ethnicity, age, etc.).
- Getting data from expensive medical tests such as magnetic resonance imaging, magnetic resonance angiography, etc.
- Getting chronological data covering a large time span (this is particularly important for diseases such as dementia, which could even span 20 years).

The last point is especially critical. Indeed, even though any type of medical data is valuable and, at the moment, very hard to get, getting chronological data is exponentially more difficult because of the different steps and places that data go through. When a patient changes his practitioners of choice or undergoes treatment in different clinics or hospitals, his data get diluted and it's near impossible to reassemble it into a cohesive whole. The adoption of blockchain data sharing platforms could significantly simplify the process of acquiring chronological data, thereby offering immense benefits to the field of medical research. Indeed, having access to chronologically filtered data is the key to predict further health complications using artificial intelligence (AI).

Blockchain in general is also a means to remove trusted authority in any information system, allowing the patients to freely use and monetize their data while stopping the hegemony of clinics and hospitals over the said data. Thus, our solution with the goal of obtaining chronological rare patient data aims to bring social impacts on many aspects such that it offers an efficient, secure, trustable, and decentralized way to share health data (Behfar, 2023). To the best of our knowledge, there has been very limited or no prior study in the literature discussing such a crowdsourcing solution considering blockchain technology for data monetization. While we have come up with an initial technology choice mostly adapted for users, we are closely inspecting all the available possibilities in terms of blockchain choice to ensure a good balance between performance and security while keeping a maximum of what makes blockchain a disruptive technology (Swan, 2015). Our biggest contribution in this study is to make the medical data sharing process optimal and efficient in terms of patients' data collection, medical data monetization, and cross-validation. Other researchers, to the best of our knowledge, didn't explore this, which can be explained because of the different constraints that this brings. Indeed, in most cases, putting data on chain can be complicated in a blockchain network, and our team has already conducted various tests on Hyperledger Fabric (Behfar et al., 2023). For the time being, there are some constraints on the number of files that could be put on-chain, but this limit is going to be decreasingly restrictive as time goes on and technology improves.

Blockchain is widely considered as a promising solution, which can build a secure and efficient environment for data sharing. Performance has become a key challenge that prevents blockchain from being used as an all-in-one data sharing platform (Crosby et al., 2016). The general scalable blockchain platforms can be divided into two categories of on-chain and off-chain. The former is through on-chain scaling solutions, adjusting the block size and interval (e.g., Bitcoin Cash) or new consensus mechanisms like Proof of Stake, Practical Byzantine Fault Tolerance. The latter is through off-chain scaling solutions, reducing the redundancy on the main blockchain using Sidechains, Multichains (e.g., Cosmos, AION), etc.

We mostly focus on the research papers that treat both the data sharing aspect and the incentive mechanisms. For the literature review, we have chosen studies that specifically treat medical data sharing and implement a solution or application. We keep only the most up-to-date research as older papers are likely to not be in line with the current state of the art and technologies. We can mention the articles by Liang et al. (2019), Zhu et al. (2019), Chen et al. (2021), Miyachi and Mackey (2021), Rajput et al. (2021), Butt et al. (2022), and Lee et al. (2022) with the focus on medical data sharing with the implication of different constraints, for example, in accordance with the General Data Protection Regulation.

Liang et al. (2019), Zhu et al. (2019), Shichang et al. (2020), Shresta et al. (2020), and Jaiman et al. (2022) have dealt with variable incentives, which is an important theme in medical data sharing. Liang

https://doi.org/10.1017/dap.2024.4 Published online by Cambridge University Press

et al. (2019) based its incentive model on the entropy of the data that are being shared. The variable incentive implemented by Xuan et al. (2020) is tied to the number of participants in the network. Indeed, they increase it when there aren't enough participants and go back to normal when the network is active enough. Zho et al. (2019) does something similar by utilizing Sharpley value to define which actor deserves what part of the incentive. Indeed, while Liang et al. (2019) does it by calculating the entropy of the data, they do not explore the data quantity bur rather focus on its quality, whereas Xuan et al. (2020) and Zhu et al. (2019) mainly focused on actual data quantity. A big quantity of data doesn't necessarily mean the said data are chronologically ordered. In fact, we can even argue that some research like Xuan et al. (2020) did not really put their users in the best position to collect the most amount of data, as they are diminishing the incentive when the network is working autonomously. As such, encouraging the data sharing process at any time on a crowdsourcing platform allows for an easier collection of chronological data, which is a lacking element in the aforementioned papers. The blockchain technology enables this possibility by offering an easier connection between all the pairs and keeping the data secret, which is also something crucial in the context of medical data.

As far as actual functionalities are concerned, Naz et al. (2019) and Liang et al. (2019) provided some kind of data verification as an important theme in medical data sharing. Naz et al. (2019) provided an automated AI-driven review system, while their implementation is limited for two reasons. At first, all the reviewing takes place on an external tool, notably to detect botting which is not adapted for medical data, as it is way more sensitive in case of hacks and/or data stealing. And second, the process doesn't take place on the chain, which would be impossible due to the constraints of smart contracts and probably the heaviness of AI models. Also, the final validation of the reviewing takes place outside of the chain as well, with some back-and-forth process for the seller to accept the review. Liang et al. (2019) also provided a solution for data verification but does not provide a real data validation scheme, rather the inner characteristics of data entropy, which correlates with the amount of information the data contain in the world of data science. Indeed, by calculating the entropy, the tool can decide whether or not a dataset is valid in the verification process.

In terms of medical data sharing implementation, Liang et al. (2017), Rajput et al. (2021), Chen et al. (2021), and Butt et al. (2022), all implemented an application based on Hyperledger, which is what we aim to perform as well but while they all bring interesting ideas to solve the specific problem, they don't explore on- and off-chain data balancing as a means to optimize this process. On the contrary, there are papers such as those from Dubovitskaya et al. (2019), Cheng et al. (2020), Lee et al. (2022), and Naga Srinivasu et al. (2021), where the analysis of blockchain data sharing is made in a more mathematical and algorithmic manner. Liang et al. (2017), Naveen and Dakshayini (2020), Chen et al. (2021), and Rajput et al. (2021), and Butt et al. (2022) use Hyperledger (Fabric and/or Composer). The way they do it is by either utilizing Hyperledger for the customization of the chaincode (smart contract) or as a way to retrieve data that are being completely stored off-chain, for example on IPFS (InterPlanetary File System). While they mostly propose finished applications that support such schema, they don't explore its feasibility of actually putting some data on-chain.

Some researchers use public blockchains to implement their data sharing ideas. Ethereum is in fact the most prominent solution for this, which has been studied by Miyachi and Mackey (2021). For the time being, it's impossible to put enough data on most public chains to implement such an idea, but this could become easier as Ethereum has changed its consensus to proof-of-stake with its "Ethereum 2.0," with lower gas fee and scalability improvements. Thus, while we won't focus our attention on public blockchains, we aim to develop our solution in an adaptive way to make it as technology agnostic as possible, in order to make it close to usable with future improvements on the blockchain field such as the aforementioned Ethereum update.

Our platform serves a primary purpose—to enable secure and efficient sharing of medical data among diverse stakeholders, including patients, healthcare providers, and researchers. It achieves this by leveraging blockchain technology to establish trust and transparency in data transactions. While in this article, we explore the implementation process of our platform; in our future study, we aim to illustrate the

https://doi.org/10.1017/dap.2024.4 Published online by Cambridge University Press

platform's functionality, showcasing instances where medical data sharing has been streamlined, which leads to improved patient care, and more efficient health-care processes.

The use of blockchain technology for medical data monetization has the potential to bring significant changes to policies related to data privacy, security, ownership, and consent. Here are some ways in which it can impact policy:

- Enhanced data privacy: Blockchain can provide a more secure and private method for storing and sharing medical data. Policies can be developed to ensure that patient data are anonymized, encrypted, and stored on the blockchain in a way that protects individual privacy rights. Additionally, access to medical data can be controlled through smart contracts, which can enforce strict rules regarding data usage and consent.
- Patient data ownership: Traditionally, patients have had limited control over their medical data. Blockchain technology can enable patients to have greater ownership and control over their health information. Policies can be implemented to ensure that patients have the right to access, manage, and share their data as they see fit. Patients can also be given the option to monetize their data by granting access to researchers, pharmaceutical companies, or other organizations, with appropriate consent mechanisms in place.
- Transparent data transactions: Blockchain's transparent and immutable nature can bring transparency to data transactions. Policies can be developed to ensure that data transactions on the blockchain are auditable and accountable, reducing the risks of unauthorized access, data breaches, or unethical use of medical data. Smart contracts can also enforce data sharing agreements, ensuring that organizations adhere to predefined terms and conditions.
- Incentivizing data contribution: Blockchain-based systems can introduce tokenization and incentivization mechanisms to encourage individuals to contribute their medical data. Policies can be formulated to define how individuals are rewarded for sharing their data, such as receiving tokens or other forms of compensation. Such policies can help address concerns regarding fairness, ethics, and equitable distribution of benefits derived from medical data monetization.
- Interoperability and data exchange: Blockchain can facilitate secure and seamless data exchange between different health-care providers, researchers, and organizations. Policies can be established to promote interoperability standards and incentivize participation in blockchain networks. This can lead to improved collaboration, data sharing, and research outcomes, ultimately benefiting patients and the health-care ecosystem as a whole.

Finally, this study aims at bringing novelty in the domain of blockchain-based data sharing policy by introducing new architecture and implementation theory for data monetization and data cross-validation. We aim for the result of our research to be implemented into a usable solution or industrial application, and our objective is to make this process efficient. In the next section, we discuss the research questions and our novel system architecture, as well as methodology and implementation theory, then we jump to making a prototype of our solution and will explain the step-by-step implementation process.

2. Research design

2.1. Research questions

We have initially realized that sharing data through blockchain is challenging, even when utilizing private blockchains which impose less restrictions compared to public ones. Thus, one of the goals of this research is to actually see if the use case of blockchain for data sharing is sensible, of course based on the current state of the blockchain technology. If this implementation is relevant, we would carefully select all the different technologies to achieve this data sharing, ranging from the actual blockchain itself, the off-chain technology supporting the blockchain, etc. If it's not relevant, we could show it by comparing the metrics (e.g., related to performance and security) between a blockchain-based data sharing scheme and a more traditional peer-to-peer system, for example, based on IPFS which is a protocol and network

https://doi.org/10.1017/dap.2024.4 Published online by Cambridge University Press

designed for storing and sharing data in a distributed file system. IPFS uses content addressing to uniquely identify each file in a global namespace, connecting all computing devices (Benet, 2014).

As previously mentioned, it is clear that utilizing blockchain as the sole technology for a data sharing platform would be difficult. Thus, we want to see how far can the blockchain go as a standalone data sharing technology before mixing it with an off-chain technology to actually handle the aforementioned limits. This is the goal of this research question:

RQ1: What are the limits of on-chain data sharing and how does the support of an off-chain technology allow for an improvement of current data sharing solutions?

The work on the previous research question will also prove useful as a starting point for the next one. Now that blockchain is a valid solution for patient's data collection and sharing, we jump to our next research question on how to customize incentives in a medical data sharing blockchain platform.

RQ2: How is it possible to customize incentives in a medical data sharing blockchain platform based on quality and scarcity of data?

Once the data are collected by patients, in order to assure data quality, we need to implement a schema for data cross-validation. This brings us to our next research question:

RQ3: How can one implement a framework for medical data cross-validation based on customizable incentives according to data anomaly?

2.2. Hybrid on-/off-chain blockchain solution

On-chain data sharing, which refers to data storage and processing directly on the blockchain, has several limitations that can hinder its efficiency and scalability. Some of these limitations include:

- Scalability: On-chain data storage can be inefficient for large-scale applications as every piece of data is recorded on the blockchain, leading to increased block size and slower transaction times.
- Cost: Storing data on the blockchain incurs transaction fees and requires more gas consumption, making it expensive, especially for large volumes of data.
- Privacy: Public blockchains have transparent and immutable data, which can compromise the privacy and security of sensitive information.
- Storage capacity: Blockchain's decentralized nature means that every node must store a copy of the entire blockchain, which can lead to storage capacity issues, especially for data-heavy applications.
- Regulatory compliance: Storing sensitive or personally identifiable information on the blockchain can create challenges regarding compliance with data protection laws and regulations.

To overcome these limitations, off-chain technologies are utilized to complement on-chain data sharing solutions. Off-chain technologies involve storing data outside the blockchain while still maintaining a connection to the blockchain for security and verification purposes. Here's how off-chain technology improves data sharing solutions:

- Scalability: By moving data off-chain, the burden on the main blockchain is reduced, improving scalability and speeding up transaction times.
- Cost efficiency: Off-chain solutions often offer lower transaction costs, making data sharing more affordable, especially for applications that require large data storage.
- Privacy: Off-chain solutions can use encryption and access controls to protect sensitive data, providing better privacy and security than on-chain storage.
- Storage capacity: Storing data off-chain allows for more flexible storage options, including utilizing cloud-based solutions, which can easily scale as needed.

https://doi.org/10.1017/dap.2024.4 Published online by Cambridge University Press

- Regulatory compliance: Off-chain solutions can incorporate data governance mechanisms and comply with data protection regulations more effectively, as data can be more easily controlled and audited.

By combining on-chain and off-chain technologies, developers can create hybrid solutions that take advantage of the strengths of both technologies. This approach is known as layer 2 scaling solutions, where most transactions occur off-chain, but their validity is anchored to the main blockchain, providing a balance between efficiency, security, and data management for decentralized applications. Layer 2 scaling solutions are in fact technologies that aim to address the performance limitations of blockchain systems. These solutions operate on top of the base layer of the blockchain and provide mechanisms for increasing transaction throughput and reducing transaction fees without modifying the underlying blockchain structure (Hafid et al., 2020). Examples of layer 2 scaling solutions include state channels, sidechains, and off-chain computation networks like Plasma and Rollups, ZK-rollups, optimistic rollups, computing oracles, and privacy-preserving blockchain solutions (Hafid et al., 2020).

There are multiple on-chain metrics that allow for a benchmarking of blockchain. Indeed, tools such as Santiment, Covalent, or Etherscan provide many metrics that all serve a certain purpose. For the first research question, we focus on the ones that actually allow for bigger security in the broad sense, as security in a blockchain network can signify different things and is not a clearly defined concept. By finding and optimizing those metrics, the solution will undeniably suffer in other categories, such as throughput of data.

As a matter of fact, we aim to investigate these metrics, parameters, and overall functionalities by such a hybrid on/off-chain blockchain solution. We look at customization as a dynamic or adaptive strategy to determine when/how/what data should be stored on-chain or off-chain to face the trade-off between performance and security. We explore the relevant research questions and the metrics/parameters by implementing a proof-of-concept solution using Hyperledger-fabric and IPFS (see Behfar et al., 2023). This leads to a balance customization problem depending on the data amount required to be on-chain, the infrastructure, the performance requirement, and some other metrics. Finally, it offers an optimal solution for data sharing; by passing the data type, volume, and chain metrics manually as input parameters, one could provide the best configuration in terms of performance and security.

2.3. Customizable incentive

Medical data sharing even using blockchain technology is not a new topic of research, but our solution provides innovation on how it performs it. Indeed, there are some projects that have studied and/or implemented a blockchain-based data sharing solution. The current state of the art is either focused on how the technology is used to implement the solution, on the tweaking and testing of certain setups to provide better metrics on that application, or on the implementation of incentives to ensure activity on the network. Thus, to the best of our knowledge, what we are trying to achieve has not been yet explored by any of the previously mentioned work. In our solution, the uniqueness comes from the way we tackle the incentive. The way this would work is by adapting the incentive for a data seller depending on the characteristics of his data. In a use case, we focus on patients with cognitive disorders, Alzheimer's disease in particular, because it is an almost 20-year-long disease where obtaining chronological data is quite important. In addition, not only data obtention in different stages of this disease is critical, but also age, ethnicity, genetics, and other factors play a role. For instance, data from a European individual suffering from Alzheimer's disease does not provide the exact same information as data from an Asian affected by the same disease. Also, while having a few years of data from those patients is useful, large chronological data are an invaluable asset for research purposes.

Thus, we need to encourage the owners of that kind of specific and rare data to share it, and that's precisely what our customizable incentive would do. For every transaction made on the network, a small fee would be added to fill up an ETH pool (see Figure 1). The initial pool reserve would be covered by some of the research budgets. Our user structure would contain multiple attributes such as age, ethnicity, the actual disease, and the time range of the data. Our platform would track the most sought-after data and,

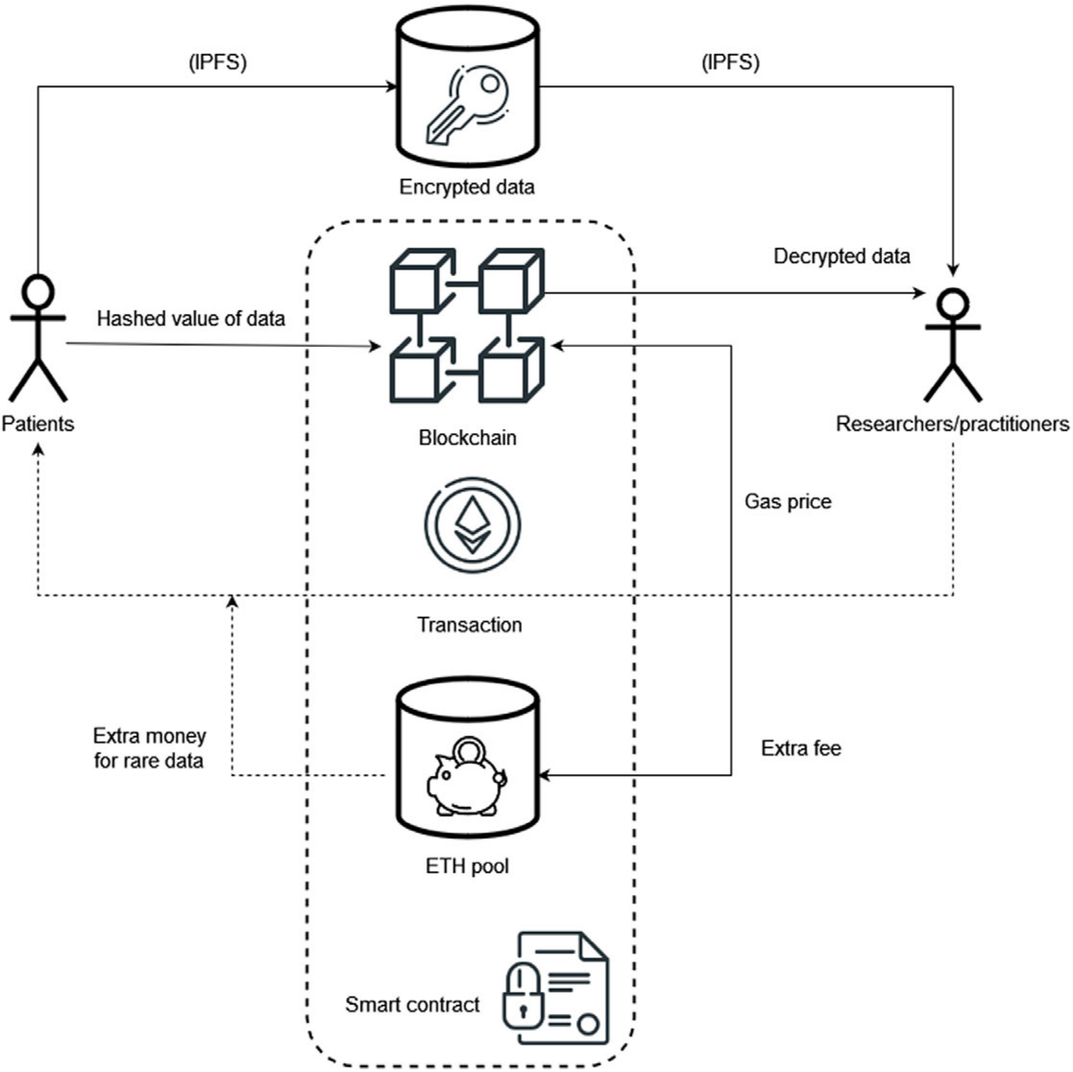

Figure 1. *Architecture of customizable incentive.*

when a user provides corresponding data to sell on the network, the base smart contract and the specific contract made with the buyer would automatically adjust the money received by gathering from the pool. Not to forget that, although test data need to be completely anonymous, we request the sellers to digitally sign some ethics formulaire that they are fully aware of how these data will be used.

As we already stated, patients are the actual owners of the said data; this could be a debatable concept. If the patient signs a contract turning over the ownership of the data to the research facility, that‘s perfectly legal, and to argue about ethics is misguided and irrelevant. If one assumes, for the sake of argument, that patient has irrevocable and unassignable rights to their health data, the problem increases in complexity for the purpose of data sharing, as some patients may deny such sharing. Other studies have also discussed medical data ownership concept (see Mirchev et al., 2020). Zhang et al. (2022) have mentioned that if a patient wants to share his data, he should be able to do it as one stakeholder in this environment or ecosystem.

https://doi.org/10.1017/dap.2024.4 Published online by Cambridge University Press

2.4. Cross-validation

Similar to our customizable incentive model for patients who provide rare data, we would take advantage of the same kind of incentive for data validators. The way it would work is with a reviewing system, where a set of network validators can review the data that are being uploaded on the network to verify if the formatting is good. Validators would get a score associated with their reviews. For example, if someone rates some data well while most of the others treat it as unusable, this particular validator will lower his personal score and, on the contrary, good reviews will increase his score. The reviews would also provide a bonus score for good review streaks. They would also not be consultable before a certain time, and would all be released at the same time, making them reviewable by the other members.

This implementation solves two problems at once: firstly, it's obviously a way to automize cross-validation without needing to employ people for doing so, which also provides extra decentralization to our solution. This also prevents the usage of botting, as bots wouldn't be able to consult other reviews to base theirs on. Also, the inner complexity of such health-care data would necessitate very sophisticated bots to provide good reviews consistently, and thus, they wouldn't be able to benefit from the streak bonus and would most likely have their review score decreased at some point.

The only downside to this implementation is another reliance on the ETH pool, and we would need to make sure that it is sufficiently filled up to pay both the data sharers and the data validators. A solution to this problem is an alternative incentive for data validators by giving them an equivalent status to the premium members (see Figure 2). They would get, instead of money from the pool, a priority on the data they are most interested in, which would alleviate some of the pressure on the pool.

We need to define the best blockchain to use for our implementation process. For that, we have chosen to implement Hyperledger Fabric for its ease of use and its inner characteristics. Hyperledger Fabric is a private/permissioned ledger technology from the Linux Foundation which is commonly used to build enterprise blockchain solutions. It is in fact an open-source blockchain platform that provides a foundation for developing enterprise-grade blockchain solutions (Ullah et al., 2021). It is designed to be a permissioned blockchain, meaning that participants must be granted access and follow certain rules to join the network (Makhdoom et al., 2020). Hyperledger Fabric distinguishes itself from other blockchain technologies by having a modular architecture that separates the blockchain ledger and the world state, which is a database that keeps track of the ledger states (Makhdoom et al., 2020). We also look at more options such as Multichain, IBM blockchain, and Proprietary blockchain for "hybrid" solutions. In our future works, we are also going to explore the possibility of utilizing public blockchain to increase the security and the decentralization aspect of our solution. In our current case, Hyperledger Fabric is very adapted to the platform we build for multiple reasons:

- It is highly configurable, which means we can easily adapt the technical specifications of the blockchain according to the results of the study. One example of that is what they call "pluggable consensus protocol," which basically allows for a fully customized consensus depending on our needs. For instance, the framework comes with a pBFT (Practical Byzantine Fault Tolerance) consensus implementation, which might be a good fit to our application. Indeed, it allows for high throughput and good security at the expense of high scalability which might not be necessary in our case.
- It is permissioned, which allows for the users of the application to have a certain level of trust in the system without having to use some more expensive/energy-unfriendly blockchains like Ethereum, which is also commonly used for application-focused blockchain solutions.
- It uses a "channel"-based architecture where we can ensure that only the participants of this particular channel can see the data that are being transferred, which is not only crucial but very adapted to our core idea in medical data sharing scenario.

For the configuration, we need to define a few criteria such as the creation of the previously mentioned channels, but also the configuration of the base network to start the customization.

https://doi.org/10.1017/dap.2024.4 Published online by Cambridge University Press

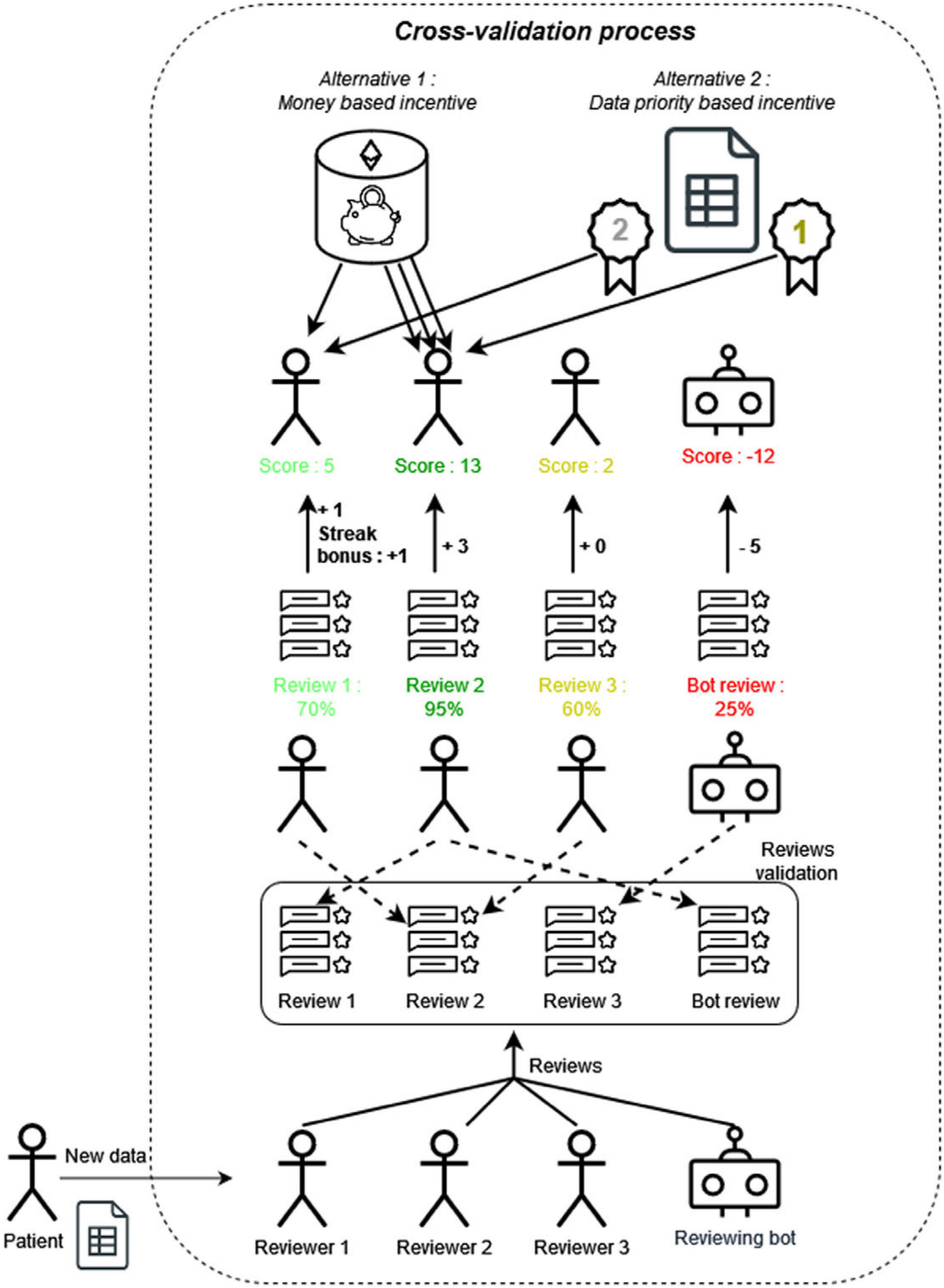

Figure 2. *Illustration of the data validation process.*

https://doi.org/10.1017/dap.2024.4 Published online by Cambridge University Press

As previously mentioned, Hyperledger is an open-source distributed ledger framework enabling the development of smart contracts and decentralized applications. Unlike Bitcoin or Ethereum networks, where anyone can freely create and own a node to join the network in order to secure it and be rewarded for this work, Hyperledger is a private blockchain and is only accessible by the users, recognized parties, and the members of a channel; see the online resource in the reference "Hyperledger Fabric, A Blockchain Platform for the Enterprise." The operation is based on a majority consortium that takes place during the validation of transactions. Here, the players are all considered "in good faith"; there is no consensus based on proof of work or proof of stake. The advantage of Hyperledger among its competitors lies in the many customization possibilities of the system. We can cite the choice of the consensus algorithm, the management of access to data, or the channels, acting as a compartment for communications between the different organizations. Indeed, the modularity as well as the flexibility of Hyperledger has a price to pay, which is the difficulty of the implementation. In order to understand better the Hyperledger transaction process, Nijssen and Bollen (2019) made it quite illustrative.

2.5. Methodology and implementation theory

Data monetization in a blockchain platform involves various mathematical formulas and calculations to determine data pricing and economic incentives. Below are some key math formulas used in data monetization on a blockchain platform:

- *Data pricing formula*

Data pricing can be calculated based on factors such as data quality, uniqueness, demand, and supply. A basic formula for data pricing can be

$$Data\_Price = Base\_Price + Quality\_Factor + Uniqueness\_Factor + Demand\_Supply\_Factor,$$

where:
Base_Price: A fixed base price set by the data provider (based on certain medical test data).
Quality_Factor: A factor representing the quality or accuracy of the medical data.
Uniqueness_Factor: A factor representing uniqueness or exclusivity of the data (how rare medical data is).
Demand_Supply_Factor: A factor representing the demand and supply dynamics of the medical data in the marketplace.

- *Data usage metrics formula:*

Data usage metrics can be used to measure the utilization of data by consumers. A simple formula for data usage metrics can be

$$Data\_Usage\_Metrics = Usage\_Time + Usage\_Frequency + Insight\_Derived,$$

where:
Usage_Time: The total time for which the data were accessed by consumers.
Usage_Frequency: The frequency of access to the data by consumers.
Insight_Derived: A measure of the specific insights or value derived from the data.

- *Data privacy formula:*

Data privacy is a critical consideration in medical data monetization. Blockchain-based techniques, such as homomorphic encryption or zero-knowledge proofs, can be used to ensure secure data sharing while

https://doi.org/10.1017/dap.2024.4 Published online by Cambridge University Press

preserving patient privacy. The specific formulas depend on the encryption and privacy-preserving techniques used.

- *Homomorphic encryption formula:*

Here, we assume there is a medical data value x that needs to be encrypted for sharing, and the encryption function is denoted as $E(x)$. The formula for homomorphic encryption can be represented as

$$Encrypted\_Data = E(x).$$

With homomorphic encryption, the blockchain platform or data consumers can perform specific computations on the encrypted data without knowing the actual value of x. The encrypted data can be stored on the blockchain or shared among authorized parties, and operations can be performed on it in its encrypted form. For example, consider a basic arithmetic operation on encrypted data:

Addition: Suppose we have two encrypted medical data values, $E(x)$ and $E(y)$. The blockchain platform can perform an addition operation on the encrypted values without decrypting them:

$$Encrypted\_Result = E(x) + E(y).$$

The result of the addition, *Encrypted_Result*, will still be in encrypted form, ensuring that the actual values of x and y remain private. By employing homomorphic encryption or other privacy-preserving techniques, blockchain-based medical data sharing platforms can enable secure and confidential data sharing while maintaining patient privacy. It allows authorized parties to perform computations on encrypted data without compromising the sensitive information contained within it.

Here, we explain this system with detailed steps to better comprehend how the cross-validation process works:

1. A patient wants to sell his data. He inputs it on the platform and some nodes that flagged themselves as volunteer validators get access to some bits of data.
   This last point is very important: we cannot give access to the entire data as this would imply the reviewers could basically steal the data instead of reviewing it. Getting snippets of the data is sufficient to evaluate whether or not the data are formatted nicely, but it also diminishes its size which is critical as it would most likely be inserted in a smart contract transaction. This also means that the final data uploaded on IPFS would not be changed, but it needs to be validated as good before being actually uploaded.
2. The system checks the volunteer reviewer count. The review score mentioned in the former step would depend on the count: the more reviewers there are, the bigger (exponentially) the final score, and thus, the final incentive. The reasoning behind this is to avoid the scenario where a single reviewer or a very small number of them that could actually be collaborating together provide a botched review, for example, by people not actually knowledgeable in the field or misleading, for example, flagging good data as bad purposely. That also prevents botting to some extent, as bot-made reviews would be flagged as bad by other reviewers.
3. The reviewers analyze the data with the process being made outside the platform and review it. When they are done, they upload a review tied to the data and the second review process begins. The reviewers would now go through the other reviews that have been done and give them a pertinence score (on the visual workflow, this is the percentage). Note that reviewers would purposely not be placed in the same review process multiple times before a certain amount of time has passed to prevent self-serving team motives.
4. There are now two possibilities:

https://doi.org/10.1017/dap.2024.4 Published online by Cambridge University Press

a. Everybody agrees that the data are bad, and the reviews all mention it. It would, thus, not be uploaded on the platform, and the reviewers would all share a base remuneration for their work coming from the ETH pool (with the amount depending on their score).
b. The data are good, and some reviews explain the reason based on the knowledge of the reviewer. Each reviewer, by evaluating colleagues' reviews, gives a score to the said reviews. Depending on the quality of the review, every reviewer would get a personal score, adding up with each review they make on the platform and further increasing with good review streaks. The higher the score, the more money they get from reviewing. As such, it encourages activity on the network but also prevents bad reviews with the possibility of diminishing one's score.

There's also a sub-possibility that involves giving priority to data purchase instead of money to the reviewers. The reason is to prevent the ETH pool running out of money for sellers' incentives, which is the main goal of this project. Indeed, with the pool's money coming from transaction fees, it is possible that, depending on the activity on the network, it doesn't contain a lot of tokens. This is also something that the dynamic incentive would take into consideration when assigning the extra gain on a data sale.

We use the following parameter(s) and get the following output(s) (Table 1):

By inputting information such as the disease, one is affected by, and the time range of the said disease and providing the data about it, the patient can enter the network. There would be an initial gas fee for doing so, but that could be either covered by the ETH pool or be added as a fee for the buyer to cover the expense. Different parameters would be mapped to certain filters, which would allow potential buyers to better look for them on the platform.

2.6. Differential privacy

Applying differential privacy to blockchain-based medical data monetization involves implementing privacy-preserving techniques that protect individual data privacy while allowing the use of aggregated data for monetization purposes. Here's how one can achieve this:

- Data aggregation: Instead of sharing raw individual medical data, blockchain-based systems can perform data aggregation on the blockchain. Aggregated data hides specific details of individuals while retaining useful insights for monetization and analysis.
- Noise addition: One of the core principles of differential privacy is adding controlled random noise to the aggregated data. This noise obscures individual values, ensuring that the presence or absence of a single individual's data does not significantly impact the aggregated results.
- Smart contracts and consent management: Implement smart contracts on the blockchain to manage data sharing and consent. Patients should provide informed consent for their data to be used for monetization purposes, and smart contracts can enforce the terms of consent.
- Data budget: Set a privacy budget or threshold to limit the amount of privacy loss that can occur during data aggregation and monetization. This ensures that the cumulative privacy loss remains within an acceptable limit over time.

Table 1. *List of input parameters and the outputs*

Parameters/input	Result/output
Type of profile (patient, practitioner, …)	Chronological medical tests
Disease/medical data type	Customizable incentive
Time range	Medical data cross-validation
Raw medical data	

https://doi.org/10.1017/dap.2024.4 Published online by Cambridge University Press

- Access control: Enforce access control mechanisms through smart contracts to restrict data access only to authorized parties who have the required consent from patients. Access can be granted based on predefined terms and conditions.
- Regular auditing and monitoring: Regularly audit and monitor the blockchain-based system to ensure that privacy measures are correctly implemented and privacy guarantees are maintained. Any potential breaches or privacy risks should be promptly addressed.
- Collaboration with privacy experts: Designing and implementing a differential privacy mechanism requires expertise in privacy-preserving techniques. Collaborating with privacy experts ensures that the system meets privacy standards and best practices.
- Transparency and communication: Transparently communicate with patients about how their data will be used, aggregated, and protected. Maintain open communication about the benefits and risks of data monetization and how privacy is being safeguarded.

By incorporating these measures, blockchain-based medical data monetization can achieve a balance between preserving individual privacy and enabling valuable data monetization opportunities. Differential privacy techniques provide a strong foundation for ensuring that patient data remain protected while still allowing researchers, pharmaceutical companies, or other stakeholders to derive valuable insights from aggregated data.

The mathematical formula for differential privacy in blockchain-based medical data monetization can be expressed as follows. A mechanism satisfies ε-differential privacy if, for all possible datasets D1 and D2 that differ in only one data point, and for all possible outcomes S in the output space of the mechanism:

$$P(M(D1) \in S) \leq e^{\varepsilon} * P(M(D2) \in S) + \delta,$$

where:

$M(D1)$ represents the output of the mechanism when the dataset D1 is used.

$M(D2)$ represents the output of the mechanism when the dataset D2 is used.

The probability on the left side represents the probability of the mechanism outputting result S using dataset D1, and the probability on the right side represents the probability of the mechanism outputting result S using dataset D2, with the additional privacy breach probability δ.

The ε parameter controls the privacy budget or the maximum allowable privacy loss, while the δ parameter represents the probability of the mechanism failing to provide ε-differential privacy. Smaller values of ε and δ indicate stronger privacy protection and lower risk of privacy violations, respectively.

In the context of blockchain-based medical data monetization, differential privacy techniques are applied to ensure that the presence or absence of individual data points does not significantly impact the monetization results while preserving the privacy of the participants' medical data. By controlling the values of ε and δ, stakeholders can strike a balance between privacy and the utility of aggregated data for monetization purposes.

3. Prototype

3.1. Creation of the Ethereum-based platform

This is basically where all the configuration is going to be done to start working on the most innovative part of our solution. As previously mentioned, the blockchain we use for this project is going to be Ethereum. Thus, there are a few tools we will need to first install and configure in order to create our decentralized platform such as wallets for the transactions and tools for the development. The tools that we chose are considered standard for Ethereum-based DApps (decentralized applications):

- Truffle: framework for DApp development contains:
- Solidity: the language that we use to develop our smart contracts.

https://doi.org/10.1017/dap.2024.4 Published online by Cambridge University Press

- Ganache: simulator that comes with a local blockchain and fake Ethereum addresses loaded with ETH. It allows for testing without having to wait nor pay gas fees.
- Metamask: a reliable Ethereum wallet for most usages.

The goal is to create a working implementation allowing for Ethereum transactions on a set of nodes. We also create an initial front end at this point to test things more easily and prepare the mapping with the blockchain through Metamask.

3.2. Blockchain incentive customization model on the platform

Here is where a big part of the research is done. Indeed, we need to create the final model for our customizable incentive, which involves medical research, thorough calculation, tests, and calibration. Indeed, the flexibility of the model depends on many factors, so a lot of thinking still needs to be done to have a balanced yet appealing solution for end users. We need to assure that the model is working properly, which means we need to develop all the features and components of the model and map everything with IPFS to answer the following research questions:

- How is it possible to store chronological medical tests via patients onto a hybrid on-/off-chain blockchain platform?
- How is it possible to customize incentives in a medical data sharing blockchain platform based on quality and scarcity of data?

3.2.1. Development of the model and the smart contract

According to the results of our research, we actually implement the different inputs necessary for our model, such as:

- The type of profile:
  - Patient
  - Buyer/reviewer
    - Practitioner
    - Researcher
- Diseases/medical data type
- Time range
- Raw data.

Here are the steps for the creation of the smart contract:

1. *Creation of the base smart contract*: This is made in Solidity, which is the de facto language for Ethereum smart contract coding. The contract has its own Ethereum address (20 bytes) which serve as the pool. This is defined in the constructor with the msg.sender instruction.
2. *Creation of the data mapping:* To keep privacy in place, we don't associate the data with a user profile per say. As such, we could use a mapping (which is a dictionary equivalent in Solidity) with, as key, the Ethereum address of a user, and as value, either a hash (CID) to an off-chain (IPFS) encrypted collection of data for a patient or null for practitioners. Note that the data seller would still need to use a little bit of ETH of his own to accommodate for fees that could be reimbursed when sold as previously mentioned.
3. *Creation of the buy/sell function*: This kind of function uses a special Solidity keyword (payable) to specify its money transfer capacity.

   Choosing from a list of available data shown by looping through the values from the mapping, a user can apply to buy the data. By doing so, the funds (price + fee) would be checked by the

https://doi.org/10.1017/dap.2024.4 Published online by Cambridge University Press

contracts to verify they are sufficient and, if so, the patient would be linked with it through a private smart contract transaction including the money and the CID. We might use a specific private Ethereum transaction solution such as TornadoCash for this. We should still figure out details about how to prevent user scams (e.g., reclaiming data with CID when we work on the incentive model).

4. *Implementation of the incentive model*: This is made with special Solidity functions called "modifiers," which allows for automatic checking of a condition. When a sell call is made, the system would check the properties and dynamically increase the incentive before showing the gain to the seller and before putting it on the platform.

   For the data inputs, we test them with some data, which means we will use some medical data (publicly available in ADNI website) which include the different metrics we are looking for. Alternatively, we also create a dataset ourselves based on our prior work and experience in the medical field.
5. *Calculate and calibrate:* The incentive being dynamic in many ways, we need to properly calculate all the possible outcomes to make sure the model is both appealing for the user and balanced because of its reliance on the ETH pool.

Thus, there is a lot of calculation, testing, and calibration done. The IPFS to data hash mapping is done at this stage, and we need to test the model properly.

3.3. Medical data cross-validation and mainnet implementation

Some of the implementation part to create our cross-validation framework has already been done in the prior section. Here, we also do calculation, test, and calibration for the most part, but in relation to the review system for the cross-validation instead of the base customized incentive. When the cross-validation model is finished, we add it to the base solution and run some tests. Here, we work on the research question RQ3: How do we implement a framework for medical data cross-validation based on customizable incentives according to data anomaly?

Here are the steps for the creation of the cross-validation framework:

3.3.1. Update of the incentive in the smart contract

We need to update the model and the related smart contract made in the last section to implement the reviewing system. We also add the support for the premium version at the same time, as it is also related to the users' status on the platform.

- *Data slicing and integration into transactions:* The data field is a staple of all Ethereum transactions. This is where the data for reviewing are going to be inserted. When inputting a file with the data using the front end, the platform will slice the data to save up on gas and ensure there's no data theft (see the cross-validation process details) and format it in accordance with the specifications of the ABI format (application binary interface)
- *Reviewers' validation and data snippet sending:* The smart contract is responsible for keeping a list of reviewer addresses when subscribing as such on the platform and for making sure they are still active (as if they interacted with the platform during the last month) as well as keeping their score updated (see the cross-validation process details). With that list, it would redistribute some transactions with the data snippet to the aforementioned reviewers.
- *Review upload and cross-validation:* This is basically a combination of the two previous steps into one. Instead of snippets of data being included in a transaction data, the review is redistributed to other reviewer nodes for cross-validation. When that second process is over, the scores associated with the review (see the cross-validation process details) are then aggregated and upgraded in the smart contract data structure.
- *Incentive calculation:* This is the most crucial part of the smart contract, which is splitted in two parts:

https://doi.org/10.1017/dap.2024.4 Published online by Cambridge University Press

- ○ If the data are not good enough according to the reviewers, the smart contract would allocate some compensation incentive to pay reviewers for their work. That calculation will be based on an equation involving the current state of the ETH pool, the score of the reviewers, and the number of reviewers who participated in that review.
- ○ Based on the score of the reviewers, the smart contract would loop through them in order, following a kind of round robin system. Here, two options would be available for them: the first is to have priority on buying the data. If the currently looped reviewer is indeed interested in buying it, the remaining ones with a decent score would get money from the ETH pool with the smart contract automatically calculating the amount based on what's left in the pool and on the respective scores of reviewers that need to be paid. If he's not interested in the data per se, the same calculation would account for it, he would get the most amount of money from the pool and the priority would be offered to the next one to be looped, repeating the process until the list of remunerable reviewers is over.

- *Scoring criteria:* the criteria used to score the data according to the reviewers are based on the number of reviewers, also factors like relevance and completeness in the validation process.
  - ○ Reviewer count: The primary criterion for the review score is the count of volunteer reviewers. The system designed in the study checks the number of reviewers participating in the evaluation of a data submission.
  - ○ Scoring mechanism: The review score is directly dependent on the reviewer count. Specifically, the more reviewers there are, the larger (exponentially) the final score becomes. This approach is taken to ensure a broad consensus among reviewers, which adds robustness to the review process.
  - ○ Purpose of the mechanism: This scoring system aims to mitigate the risk of biased or uninformed reviews. By requiring a larger number of reviewers, the system dilutes the impact of any single reviewer who may not be knowledgeable or who might intentionally provide misleading feedback. This method is particularly effective against collaborative efforts to provide botched reviews or manipulation by a small group of reviewers.
  - ○ Botting prevention: Another significant aspect of this scoring mechanism is its capacity to prevent or minimize the influence of automated, bot-generated reviews. Reviews suspected to be generated by bots are likely to be flagged as bad by human reviewers, further ensuring the integrity of the review process.
  - ○ Incentive correlation: The final incentive for data submission is correlated with the review score. A higher score, influenced by a higher reviewer count, results in a larger incentive. This linkage encourages data submitters to seek reviews from a larger group, promoting transparency and accuracy in the data validation process.

3.3.2. Implementation of the platform on Ethereum mainnet

At this point, all the work on the project is done on Ganache to avoid gas fees. https://github.com/stefankam/Ethereum-medical-data-sharing. Then, we implement our smart contract on the Ethereum mainnet to do the final testing of the platform. Most of the testing is done on Ganache, and when we are confident that it is working as intended, we move on Ethereum mainnet to properly test the solution while keeping gas fees as low as possible.

- Application effects of the proposed scheme and platform

In assessing the application effects of the proposed decentralized crowdsourcing medical data platform, a multifaceted approach is essential. This includes evaluating the efficiency of data transactions, cost implications, and enhancements in data security.

The evaluation of data transaction efficiency should focus on how the platform optimizes the speed and accuracy of data exchanges. This is crucial in the context of medical data, where promptness and precision are paramount. By utilizing blockchain technology, the platform is expected to streamline these processes, thereby offering a more effective solution compared to traditional methods.

Cost reduction is another critical aspect. The platform's impact on reducing expenditures related to IT, operations, support functions, and personnel needs to be carefully quantified. The unique efficiency of blockchain could potentially result in significant savings for the health-care industry, and this should be clearly demonstrated through comparative cost analysis before and after the platform's implementation.

Enhancing data security is a key goal of the platform. This involves not just safeguarding sensitive health information but also ensuring compliance with data protection laws. The immutable nature of blockchain's ledger provides a robust framework for this purpose. Evaluating the platform's effectiveness in this regard would require analyzing metrics related to data breaches, unauthorized access attempts, and regulatory compliance.

To complement these technical assessments, real-world case studies and user testimonials are indispensable. They provide tangible evidence of the platform's practical application and effectiveness. Case studies should cover diverse scenarios, highlighting the benefits realized by different stakeholders like patients, health-care providers, and researchers. User testimonials, on the other hand, offer direct feedback from those who have interacted with the platform, providing valuable insights into user satisfaction and the platform's ability to meet varied needs.

- Experimental comparisons with relevant studies

A comprehensive comparison with relevant studies in the field is necessary to establish the platform's superiority in key aspects like data sharing efficiency, security, and accessibility.

The methodology for these comparisons should be thorough and systematic. It involves selecting benchmark studies that serve as a reference point and defining clear performance metrics and criteria for evaluation. These comparisons are not just about highlighting the strengths of the proposed platform but also understanding areas for potential improvement.

A detailed analysis of data sharing efficiency is crucial. This should involve a comparison of quantitative data on the speed, reliability, and overall effectiveness of data transactions facilitated by the platform versus alternative solutions. Such an analysis will provide clear evidence of the platform's advancements in improving data sharing processes.

Security measures are another vital area of comparison. The platform's approach to ensuring the safety and privacy of medical data needs to be assessed against existing solutions. This includes an evaluation of data encryption methods, access controls, and compliance with established security standards.

Finally, assessing accessibility and user experience is essential to understand the platform's practicality and user-friendliness. This encompasses a review of the platform's interface design, ease of use, and the process of adoption compared to other systems in the field. It's important to ensure that while offering advanced features, the platform remains accessible and intuitive for its users.

4. Conclusion

Our solution is innovative and offers concrete social value creation, providing both patients and other stakeholders an easier and more secure way of trading medical data by utilizing blockchain technology, but also facilitates the obtention of the most elusive types of health data by dynamically allocating extra financial incentive depending on the scarcity of the data. It also provides a way to ease the obtention of long-term chronological data, which is not only extremely important but very difficult at the moment. For example, if somebody suffers from Alzheimer's disease or dementia, which is a long-term disease, probably possesses some test data from 10 to 15 previous years, this constitutes an invaluable resource for the medical and scientific field. Thus, our solution would automatically increase the money received by that patient when he sells his test data on the platform by using tokens gathered in a pool that gets filled by some fees coming from all the transactions made on the platform. This could also serve as compensation if the said patient just wants to make his data public without necessarily selling it to a particular individual.

A lot of research has already been done on the subject of data sharing, but we mostly position ourselves as innovators in terms of how the data are incentivized and verified with cross-validation. In this context,

https://doi.org/10.1017/dap.2024.4 Published online by Cambridge University Press

there are some relevant industrial projects, the closest one would be Medibloc (https://medibloc.com/en); they also provide a data sharing platform for health-care data with incentive mechanisms and a token pool. We could also cite dHealth (https://www.dhealth.com); they also build a blockchain-based solution for sharing health-care data, with an even bigger emphasis on tokenomics than Medibloc. Their whitepaper goes in-depth in terms of how they're handling their token ($DHP), but their solution looks quite centralized due to their consensus choice. We also cite Medicalchain (https://medicalchain.com); they incentivize users to monetize their health data and utilize a dual blockchain structure for doing so (Hyperledger and Ethereum); however, they don't implement any sort of customizable incentive for specific data either.

This project would consist of a web-based application to interact with the blockchain and the other participants and would aim at reaching individuals who are affected by the diseases and the stakeholders who could be interested in buying the collected data (research institutes, pharmaceutical companies, insurance firms, etc.). In the implementation section, we explained all our future-intended research work in terms of 1) application effects of the platform and 2) experimental comparisons with the relevant studies. Finally, the use of blockchain technology for medical data monetization has the potential to bring significant changes to policies related to data privacy, security, ownership, and consent, which were explained throughout the paper.

Data availability statement. The platform data will be available at https://github.com/stefankam/Ethereum-medical-data-sharing following a primary implementation of the platform.

Acknowledgments. The authors acknowledge the help of Mr. Cyril Schranz in writing relevant funding proposals, where we also partly benefitted from in writing this article.

Author contribution. Conceptualization: S.K.B.; Methodology: all authors; Supervision: all authors; Validation: all authors; Writing—original draft: S.K.B.; Writing—review and editing: all authors. All authors approved the final submitted draft.

Funding statement. This work received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interest. The authors declare that they have no competing interests.

References

Behfar SK (2023) Decentralized intelligence and big data analytics reciprocal relationship. In *IEEE International Conference on Blockchain Computing and Applications (BCCA)*, Kuwait, pp 643–651.

Behfar SK, **Théodoloz F**, **Schranz C and Hosseinpour M** (2023) Blockchain-based data sharing platform customization with on/off-chain data balancing. In *IEEE International Conference on Blockchain Computing and Applications (BCCA)*, Kuwait.

Benet J (2014) IPFS - Content addressed, versioned, P2P file system. available at: https://arxiv.org/pdf/1407.3561.pdf

Butt G, **Sayed T**, **Riaz R**, **Rizvi SS and Paul A** (2022) Secure healthcare record sharing mechanism with Blockchain. *Applied Sciences 12*(2307). https://doi.org/10.3390/app12052307.

Chen Y, **Meng L**, **Zhou H and Xue G** (2021) A blockchain-based medical data sharing mechanism with attribute-based access control and privacy protection. *Wireless Communications and Mobile Computing*, *5*, 1–12. https://doi.org/10.1155/2021/6685762.

Cheng X, **Chen F**, **Xie D**, **Sun H and Huang C** (2020) Design of a secure medical data sharing scheme based on blockchain. *Journal of Medical Systems 44*, 52. https://doi.org/10.1007/s10916-019-1468-1.

Crosby M, **Pattanayak P**, **Verma S and Kalyanaraman V** (2016) Blockchain technology: Beyond bitcoin. *Applied Innovation 2*, 6–10.

Dubovitskaya A, **Novotny P**, **Xu Z and Wang F** (2019) Applications of blockchain technology for data-sharing in oncology: Results from a systematic literature review in oncology. *Oncology and Informatics – Review 98*, 1–9. https://doi.org/10.1159/000504325.

Hafid A, **Hafid AS and Smith MA** (2020) Scaling technologies; a comprehensive survey. *IEEE Access 8*, 125244–125262. https://doi.org/10.1109/ACCESS.2020.3007251.

Jaiman V, **Pernice L and Urovi V** (2022) User incentives for blockchain-based data sharing platforms. *PLoS One 17*(4), e0266624. https://doi.org/10.1371/journal.pone.0266624.

Koskinen J, **Kainu V and Kimppa K** (2016) The concept of Datenherrschaft of patient information from a Lockean perspective. *Journal of Information, Communication and Ethics in Society 14*(1), 70–86. https://doi.org/10.1108/JICES-06-2014-0029.

Lee JS, **Chew CJ**, **Liu JY**, **Chen YC and Tsai KY** (2022) Medical blockchain: Data sharing and privacy preserving of EHR based on smart contract. *Journal of Information Security and Applications 65*, 103117. https://doi.org/10.1016/j.jisa.2022.103117.

Liang J, **Zhao S**, **Shetty J and Li D** (2017) Integrating blockchain for data sharing and collaboration in mobile healthcare applications. In *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*. https://doi.org/10.1109/pimrc.2017.8292361.

Liang X, **Chen W**, **Li J**, **Mu Y and Tian Z** (2019) Incentive mechanism of medical data sharing based on information entropy in blockchain environment. *Journal of Physics: Conference Series 1302*, 022056. https://doi.org/10.1088/1742-6596/1302/2/022056.

Makhdoom I, **Zhou I**, **Abolhasan M**, **Lipman J and Ni W** (2020) PrivySharing: A blockchain-based framework for privacy-preserving and secure data sharing in smart cities. *Computers & Security 88*, 101653. https://doi.org/10.1016/j.cose.2019.101653.

Mirchev M, **Mircheva I and Kerekovska A** (2020) The academic viewpoint on patient data ownership in the context of big data: Scoping review. *Journal of Medical Internet Research 22*(8): e22214. https://doi.org/10.2196/22214.

Miyachi K and Mackey TK (2021) hOCBS: A privacy-preserving blockchain framework for healthcare data leveraging an on-chain and off-chain system design. *Information Processing & Management 58*, 102535. https://doi.org/10.1016/j.ipm.2021.102535.

Naga Srinivasu P, **Bhoi AK**, **Nayak SR**, **Bhutta R and Woźniak M** (2021) Blockchain technology for secured healthcare data communication among the non-terminal nodes in IoT architecture in 5G network in electronics. *Electronics 10*(12), 1437. https://doi.org/10.3390/electronics10121437.

Naveen S and Dakshayini M (2020) Secure sharing of health data using Hyperledger Fabric based on blockchain technology. *International Conference on Mainstreaming Block Chain Implementation (ICOMBI)*, 1–5. https://doi.org/10.23919/ICOMBI48604.2020.9203442.

Naz M, **Al-Zahrani FA**, **Khalid R**, **Nadeem J**, **Qamar AM**, **Afzal MK and Shafiq M** (2019) A secure data sharing platform using blockchain and interplanetary file system. *Sustainability 11*, 24, 7054. https://doi.org/10.3390/su11247054.

Nijssen S and Bollen P (2019) The lifecycle of a user transaction in a hyperledger fabric blockchain network. Part 1. https://link.springer.com/chapter/10.1007/978-3-030-11683-5/11.

Rajput AR, **Li Q and Ahvanooe MT** (2021) A blockchain-based secret-data sharing framework for personal health records in emergency condition. *Healthcare 9*(2), 206. https://doi.org/10.3390/healthcare9020206.

Shichang Xuan, **Li Zheng**, **Ilyong Chung**, **Wei Wang**, **Dapeng Man**, **Xiaojiang Du**, **Wu Yang**, **Mohsen Guizani** (2020) An incentive mechanism for data sharing based on blockchain with smart contracts. *Computers & Electrical Engineering 83*, 106587.

Shresta AK, **Vassileva J and Deters R** (2020) A blockchain platform for user data sharing ensuring user control and incentives. *Frontiers in Blockchain 3*, 497985. https://doi.org/10.3389/fbloc.2020.497985.

Swan M (2015) *Blockchain: Blueprint for a New Economy*. Sebastopol, CA: O'Reilly Media, Inc.

Ullah N, **Mugahed Al-Rahmi W**, **Alzahrani AI**, **Alfarraj O and Alblehai FM** (2021) Blockchain technology adoption in smart learning environments. *Sustainability 13*, 1801. https://doi.org/10.3390/su13041801.

Xuan S, **Zheng L**, **Chung I**, **Wang W**, **Man D**, **Du X**, **Yang W and Guizani M** (2020) An incentive mechanism for data sharing based on blockchain with smart contracts. *Computers & Electrical Engineering 83*, 106587. http://doi.org/10.1016/j.compeleceng.2020.106587.

Zhang D, **Wang S**, **Zhang Y**, **Zhang Q and Zhang Y** (2022) A secure and privacy-preserving medical data sharing via consortium blockchain. *Security and Communication Networks 2022*, 2759787. https://doi.org/10.1155/2022/2759787.

Zhu L, **Dong H**, **Shen M and Gai K** (2019) An incentive mechanism using Shapley value for blockchain-based medical data sharing. In *IEEE International Conference on Big Data Security on Cloud (BigDataSecurity), High Performance and Smart Computing (HPSC) and Intelligent Data and Security (IDS)*, 113–118. http://doi.org/10.1109/BigDataSecurity-HPSC-IDS.2019.00030.

Cite this article: Behfar SK and Crowcroft J (2024). Decentralized crowdsourcing medical data sharing platform to obtain chronological rare data. *Data & Policy*, 6: e9. doi:10.1017/dap.2024.4

https://doi.org/10.1017/dap.2024.4 Published online by Cambridge University Press