MediBank: A Novel Approach to Blockchain Incentivization and Data Integrity

Ying-ke Chin-Lee
Harvard University
Cambridge, USA
ychinlee@college.harvard.edu

Allison Lee
Harvard University
Cambridge, USA
allison_lee@college.harvard.edu

Vincent Viego
Harvard University
Cambridge, USA
veviego@college.harvard.edu

Abstract—

To advance medical care and research, individuals must obtain agency over their medical records. No centralized system exists which securely manages and distributes electronic medical records (EMRs). We propose an Ethereum-based solution: MediBank. Our model, which has a fully functional prototype, differs from other blockchain-based solutions in that its incentivization scheme and bootstrapping methods support adoption and data integrity. The proposed solution (1) provides patients with full data agency while incentivizing them to share anonymized data for research purposes, (2) is fully interoperable with current database systems, and (3) disincentivizes fraudulent data generation through an adaptive data pricing model.

I. Introduction

In our data driven society, information access has become invaluable, particularly with regards to healthcare and medical research. With access to individual and aggregate records, medical professionals can better diagnose and treat patients. Despite this data's importance, there is no standardized system for EMR exchange between healthcare providers, researchers, and patients. The Office of the National Coordinator for Health Information Technology has identified such a degree of "health information blocking"[2]--the systematic interference of authorized personnel accessing electronic medical records--that Congress has been called upon to find a solution.

Recently, interest in applying blockchain technology to this problem has expanded. However, no proposed solution has gained widespread support or adoption. Such a system must fulfill five parameters of success. First and foremost, the system's incentivization scheme must ensure data integrity for researchers. Second, the system must allow all parties to access data efficiently. Third, it should maintain patient agency and privacy when providing aggregate information to researchers. Fourth, it must be cost-efficient for all parties. Finally, the system must integrate seamlessly with existing medical record storage platforms to facilitate adoption.

Our work focuses on effectively integrating healthcare data storage systems with blockchain technology. There have been previous attempts at this; Azaria et al. [1] showed through a fully functional Ethereum prototype, MedRec, that the integration of healthcare data systems and blockchain technology is feasible. However, MedRec's protocol lacks provisions for ensuring data integrity as it does not effectively disincentivize patients, medical professionals, or healthcare stakeholders (insurance companies, public health authorities, etc.) from manipulating the technological infrastructures (e.g., contributing fake data) for monetary gain.

Other attempts at blockchain-based EMR platforms have been implemented. To our knowledge, these attempts have all failed to address one or more of the aforementioned requisite features of an EMR sharing system. For example, Medicalchain is one solution that also incorporates researchers, patients, and doctors [5]. However, their solution fails with regards to interoperability with existing storage solutions and data security, as it stores sensitive medical data on the blockchain, in violation of HIPAA's regulations [6].

Simultaneously, technologies have been developed that are applicable to blockchain technology to progress towards a more robust system. Zyskind et al. [4] have shown that it is possible to integrate blockchain technology with existing local databases while ensuring data ownership, transparency, and access control. Their system "combines blockchain and off-blockchain storage to construct a personal data management platform focused on privacy"[4] with private keys granting access to pointers to where the actual data is stored.

Our concept and prototype hope not only to improve the robustness of systems such as MedRec and Medicalchain, but also to modify Zyskind's methods to enable integration with local data storage solutions used by healthcare providers (e.g., SQL databases). Ultimately, MediBank's goals are to fulfill all of the previously described success criteria of a viable blockchain-based EMR sharing system.

II. Proposed Approach

A. *Intellectual Points*

Our solution focuses on the following intellectual points: patient agency, data integrity, and incentivization. In developing a system that facilitates patient management of medical records, preserving anonymity and establishing patient agency over data is of the utmost importance. Therefore we consider it a critical failure that proposed solutions violate this fundamental right to privacy by rewarding miners with access to medical records. A consequence of monetizing data is the threat that malicious doctors and patients, in the absence of proper disincentives, will collude to post fake data to amass large sums of tokens. Beyond economic problems, failure to address this flaw compromises the value of this entire system to medical researchers. MediBank's incentivization scheme prevents this.

With healthcare providers already entrenched in their current recordkeeping systems, any new system must provide significant incentives, not only integrating easily with existing storage solutions, but also convincing potential adopters that any barriers to entry are worth the expense. MediBank stands to contribute a new system design that specifically addresses each of these unanswered problems.

We now describe the specifics of our approach for managing EMRs. The fundamental structure of the dapp will be similar to other works, but our prototype differs significantly from other models in several ways:

1. Patients have full agency over their personal data.
2. While many systems such as Medicalchain attempt to store data directly on the blockchain, MediBank will instead implement the methods proposed by Zyskind et al., modifying them to apply to transactions of *pointers to data* stored in a SQL database.
3. MediBank's *incentivization scheme* is novel in incorporating insurance companies, which has many positive implications for the overall infrastructure. It also discourages the generation of fake data by imposing fees and adaptively restricting rewards to be zero-sum (see *Figure 1* and *Experiments* for more detail).

The incentivization scheme works in the following manner:

a. *Transaction fees* will be universal (i.e., single network-wide cost) and based on global network variables such as number of users and tokens in circulation. The overall trend will be for transaction fees to decrease over time; however, miners' profits will remain constant due to increased network usage.

b. **Patients** will *sell anonymized medical data* to medical researchers, who, beforehand, set a maximum price they are willing to pay. When a patient attempts to sell data, the smart contract calculates the patient's expected annual costs (i.e., mining fees paid for downloading data, verifying PPR modifications, and sharing data with other doctors), and then using a logistic model calculates the price that each researcher must pay to obtain a year's subscription to the patient's anonymized data. The transaction is completed if the researcher's established max-price is high enough. Note that the logistic model is a function whose domain is the number of researchers to which a patient is attempting to sell their data and whose codomain is the percent of the patient's expected annual cost for which said number of researchers will collectively pay in order to acquire the patient's data. Furthermore, the system prevents additional transactions if the revenue would allow a patient to exceed their expected annual costs.

c. **Doctors** *receive economic kickbacks* from insurance companies in return for utilizing the system (as insurance companies directly benefit from this system), but this is balanced out exactly by the cost of uploading the data.

d. **Researchers** do not gain anything monetarily from this system, but can gain access to a large source of anonymized data.

e. **Insurance companies** act as the *mining party* in this system. Together, they verify all transactions by checking a status variable in the smart contracts. They are rewarded with MediTokens.

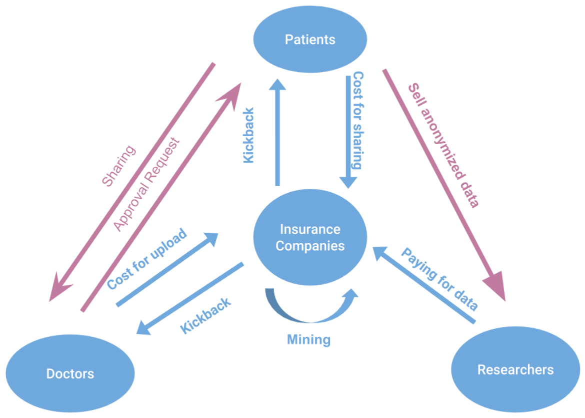

Fig. 1. MediBank's incentivization scheme. Purple represents the flow of information, while blue represents the flow of MediTokens.

B. *Success Criteria*

We will assess our system's success based on the extent to which it meets the five requirements established in *§I*. This can be demonstrated heuristically and supported through mathematical simulation of various system usage scenarios. Our model will explore the feasibility of patient and researcher costs in a zero-sum system to determine whether the proposed solution is sustainable and beneficial for all parties (though not strictly economically beneficial). Together, reasoning and data can determine our success in creating a robust incentivization scheme and EMR sharing platform.

III. Experiments

A. Description of system and theoretical framework

Our protocol draws heavily from related work [1]. In this section we define unique aspects of MediBank's protocol: incentivization and protection against fraudulent data.

Any transaction made on the MediBank network must be paid in MediTokens. Thus, anytime an authorized user wishes to manage their data (sharing or retrieving records) they will be charged a fee, which serves as a reward for the miners who validate the transaction. In MediBank, only insurance companies function as miners.

Since patients have full agency over their data, they may sell their data to medical researchers using methods proposed in [3] to protect anonymity via batch level decryption. When they do so, their insurance provider receives the token-based reward provided by the researchers. The insurance provider then rewards the patient by either distributing MediTokens as rewards when patients sell data or when doctors post data or possibly providing off-the-chain rewards (enforceable via smart contracts) such as reduced premiums for patients and increased billings for healthcare providers.

B. Description of Experiments

All experiments conducted analyze the effectiveness of the MediBank protocol. Certain aspects of the protocol lend themselves more readily to heuristic analysis, while others are more easily supported by numerical methods. First, we justify that the MediBank protocol, by construction, fulfills requirements two, three, and five of an effective EMR sharing platform as defined in *§I.* Second, we identify the following parameters within our model: number of patients and researchers using the network, growth rate, median, and limit of the logistic model, number of MediBank transactions per patient-doctor visit, and size of transaction fees. Then, assigning default values to each of these parameters, we conduct a thousand trials of an experiment for each of these parameters in which we observe the effect of random variations in the specified parameter's value on the following metrics: expected annual patient costs, costs for researchers to acquire an individual's data based on the number of researchers to whom said individual is offering their data, and the number of patients whose data a researcher could afford to purchase for a year based on the standard signup reward granted to each researcher. In performing these simulations we account for the disparity in the use of healthcare resources by patients in varying degrees of health, by breaking down our analyses into three cases (as done by the Agency for Healthcare Research and Quality): "adults with three or more chronic illnesses and no functional limitations, adults with three or more chronic illnesses and any functional limitation,"[8] and all remaining, healthier adults.

IV. Results

A. Analytical and Experimental Results

The MediBank protocol satisfies properties two, three, and five of an effective EMR sharing system defined in *§I* for the following reasons. First, by design, MediBank allows patients, doctors, and researchers with the appropriate permissions to easily access data by storing SQL queries within smart contracts, which point to the appropriate records on the healthcare provider's local database (properties two and five). Second, in the Medibank protocol patients are never obligated to contribute anonymized data to researchers; additionally, they can manage their data sharing settings on an annual basis (property three).

In the experiments where the number of patients is the variable, we observe that this parameter is independent from the metrics we are analyzing. Intuitively this makes sense since the reward system is built around the individual, independent of other patients on the network.

When the number of researchers is the variable, it is observed that the cost for a researcher to acquire a patient's data experiences an exponential decay in the number of researchers on the network. Consequently, the number of patients' whose data an individual researcher can purchase for a year with their initial tokens increases exponentially as the number of researchers on the network increases.

Letting the growth rate of the logistic model vary does not produce any significant results, other than showing that the system maintains the greatest stability (as measured by how metrics behave as functions of the parameters relative to expectation) with the growth rate at its default value of 1.

For experiments in which the median of the logistic model varies, the average costs for researchers to acquire patient data decreases as the median increases; however, this effect is only observed when patients sell their data to fewer than 100 percent of the researchers. Additionally, as expected based on this result, it is also observed that as the median increases, patients are able to cover less of their expected annual costs by selling data to researchers.

In experiments where either the number of transactions per visit or the size of the transaction fees varies, it is observed that all associated patient and researcher costs increase, as one might expect. Consequently, researchers are able to acquire the data of fewer patients when these parameters are larger.

In the final experiments, in which the limit of the logistic model varies, in most cases the parameter demonstrates little to no relationship with the metrics evaluated. However, we observe that when patients sell their data to at least 75 percent of researchers, they can earn a reward equal to their expected annual costs: an unanticipated result.

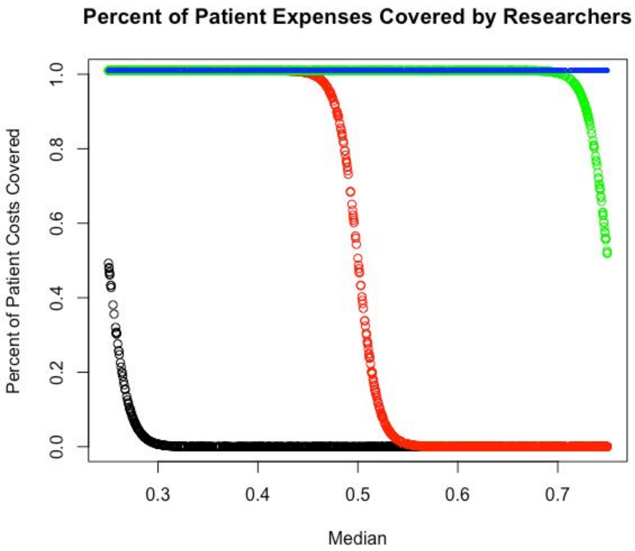

Fig. 2. Black, red, green, and blue curves correspond to a patient selling data to 25, 50, 75, and 100 percent of researchers respectively.

B. *Explanation and Discussion of Results*

The results of the mathematical simulations are significant because they reinforce our analysis that the Medibank protocol reliably ensures the integrity of the medical data accessible through the network. From a theoretical perspective, the system is robust against doctors and patients who might wish to collude in order to amass tokens because the rewards are designed to exactly offset the costs incurred by these participants. This principle is supported by empirical data in the following way. As the default, we let 1.01 be the limit of the logistic model, meaning that under certain circumstances the model would allow a patient to be paid up to 101 percent of their expected annual costs. This was done because it was expected that this value would only be achieved for unfeasibly large values *x* (i.e., the number of researchers to which a patient attempted to sell their data). However, the results of the logistic limit experiment demonstrated that once a patient sells their data to 75 percent of the medical researchers on the network they can receive payment equal to 100 percent of their expected annual costs. This result is significant because it allows for an implementation of the MediBank protocol to be a provably zero-sum incentivization for patients, thus disincentivizing the generation of fake data.

Furthermore, data from all experiments revealed that researchers pay more for data from patients who had the highest network utilization (i.e., those patients who visit doctors and hospitals most often, thus generating the most medical data). Thus, without rewarding patients for contributing extra data, which would incentivize corruption, the Medibank protocol ensures that researchers are able to acquire valuable data in a cost effective manner.

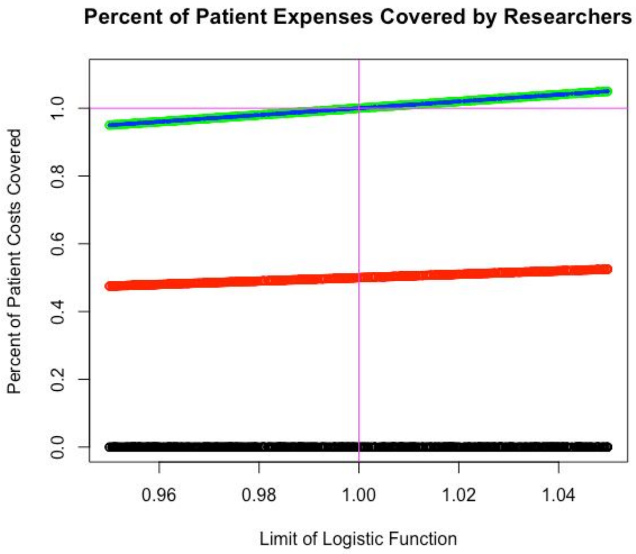

Fig. 3. Black, red, green, and blue curves correspond to a patient selling data to 25, 50, 75, and 100 percent of researchers respectively.

C. *Demo*

A video demo of our prototype can be found here: https://www.youtube.com/watch?v=CW43bgEZI0U. Additionally, our complete data analysis can be found in the accompanying R script: “modeling.R”.

V. Conclusion and Future Work

In conclusion, we have introduced and implemented a model for EMR sharing that is specifically designed to address the issues of allowing users full agency over their data, as well as providing proper incentives for long-term adoption and disincentives for posting fake information that would compromise the system’s integrity. Moving forward, we will explore the possibility of allowing patients to share their data with insurance companies in order to receive reduced premiums as a reward for “good health.”

Going forward we would like to implement the ability to store any file type of medical data. Currently, we only store the results of a pre-set form, but we would like to be able to adapt our system whatever file type an EMR is stored in (eg. PDFs, CSVs, etc). Additionally, an anonymization technique similar to that of the batch technique mentioned by McMahan et al. would allow us to progress towards truly anonymizing patient data contributions [7]. Finally, it would be helpful to utilize machine learning to find and eliminate fake data, as this would greatly benefit medical research as well; however, no method should ever produce false positives (our system should not take down real records in any scenario).

VI. Acknowledgements

We thank our research supervisor, Professor Kung, who provided insight and expertise that greatly assisted with our research. Furthermore, we would like to express our utmost appreciation for Marcus Comiter and Randall Raymond of Harvard University for not only guiding our statistical analysis but also assisting in the initial setup of our application.

VII. References

[1] A. Azaria, A. Ekblaw, T. Vieira and A. Lippman, "MedRec: Using Blockchain for Medical Data Access and Permission Management," *2016 2nd International Conference on Open and Big Data (OBD)*, Vienna, 2016, pp. 25-30.

[2] The Office of the Nat. Coordinator for Health Information Technology, "Report on health information blocking," U.S. Department of HHS, Tech. Rep., 2015.

[3] Konečný, J., McMahan, H., Yu, F., Richtárik, P., Suresh, A., & Bacon, D. "Federated Learning: Strategies for Improving Communication Efficiency.", 2016.

[4] Zyskind, G., Nathan, O., & Pentland, A. S. (2015). "Decentralizing Privacy: Using Blockchain to Protect Personal Data."

[5] Albeyatti, Abdullah. "Medicalchain", 2018.

[6] U.S. Department of HSS, "Hipaa administrative simplification: Regulation text," 2006.

[7] McMahan, Brendan and Daniel Ramage. "Federated Learning: Collaborative Machine Learning without Centralized Training Data", 2017.

[8] L. J. Conwell and J. W. Cohen. "Characteristics of Persons with High Medical Expenditures in the U.S. Civilian Noninstitutionalized Population", MEPS Statistical Brief #73 (Agency for Healthcare Research and Quality, March 2005), 2002.