Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240643

Securing Reproducibility and Accountability in Distributed Healthcare Analytics: A Framework Based on Blockchain and Cryptography

Leonardo NUCCIARELLI[a,1], Benedetta GOTTARDELLI[a], Roberto GATTA[b] and Andrea DAMIANI[c]
[a] *Dip. di Diagnostica per Immagini, Radioterapia Oncologica ed Ematologia, Universita Cattolica del Sacro Cuore, Rome, Italy*
[b] *Dip. di Scienze Cliniche e Sperimentali, Università degli Studi di Brescia, Brescia, Italy*
[c] *Fondazione Policlinico Universitario Agostino Gemelli (IRCCS) Rome, Italy*
ORCiD ID: Leonardo Nucciarelli https://orcid.org/0009-0003-8085-1365, Roberto Gatta https://orcid.org/0000-0002-4716-9925

Abstract. The results and details of the clinical studies and research must be securely stored to ensure reliability, accountability, and prevent malicious misuse. To accomplish this, a secure method for storing metadata and study results is crucial. Also, a mechanism to ensure accountability for both data owners and researchers is needed. In this way, data owners and the scientific community can rely on and verify results and methods presented by researchers, while researchers can check the validity of the analyzed data and have proof of authorship for their work. A modular framework is presented in this paper, which utilizes blockchain and cryptography to store study results and metadata, along with proof of accountability. The framework has been tested within a privacy-preserving distributed analytics infrastructure.

Keywords. Blockchain, Distributed Analytics, Healthcare, Cryptography

1. Introduction

In this paper, a framework is introduced to tackle the issues of accountability and reproducibility in scientific research with a focus on the healthcare sector, where these factors hold great importance for a variety of reasons [1]. Indeed, Electronic Health Records (EHR) have been widely embraced, and the integration of machine learning and artificial intelligence in healthcare has increased the significance of these data repositories as valuable resources for research and technological progress [2]. The challenge is to manage data while respecting privacy and ownership [3]. Various approaches have emerged in recent years to address this need for privacy-preserving data mining: these include federated learning or distributed analytics [4], privacy-by-design infrastructures [5], blockchain [6], homomorphic encryption [7], and ensemble methods

[1] Corresponding Author: Roberto Gatta; E-mail: roberto.gatta.bs@gmail.com.

[8]. In such a scenario, there is a need to securely save study results and metadata, such as data structure and analysis settings, to ensure reproducibility, validation, and accountability [9,10]. Blockchain is an excellent solution to be able to ensure reliability and security through a distributed and immutable paradigm that has already found application in healthcare [11]. The work presented shows that integrating blockchain into a software infrastructure for data retrieval and analysis can enable the monitoring of all data alterations. This helps to safeguard patients and researchers against unauthorized access and misuse of confidential data.

1.1. Problem Formulation

The distributed analytics approach (meaning all privacy-preserving data mining approaches where the users cannot access the data) offers key points for privacy requirements by avoiding the need for anonymization, encryption, or perturbation through data non-disclosure. In a distributed analytics project, various key players are involved and here the interaction between researchers and data providers is examined. Assuming that a distributed data analysis project has been set up involving one of the technologies already mentioned, it is necessary to be able to securely and certifiably track and save the data being used, with metadata such as filtering and timestamps, along with the analysis results being produced. In fact, researchers, by definition, will not be able to have direct access to the data, just as study results could be misinterpreted in the absence of study details or could be modified. The previous efforts to regulate yielded only limited success [12,13]. Therefore, two essential requirements arise: firstly, study metadata and their results should be systematically stored in a secure and automated manner. Secondly, the origin of the results must be unequivocally reconstructed.
The framework outlined here implements an automatic, secure, and non-repudiable reporting of results and metadata. It demonstrates that integrating blockchain into a software infrastructure for data retrieval and analysis can enable the monitoring of all data alterations. This is crucial for safeguarding patients and researchers against unauthorized access and misuse of confidential data.

2. Methods

2.1. Environment Overview: the GEN-RWD Sandbox

The framework presented has been implemented and tested within the GEN-RWD Sandbox [14]. It allows for conducting analytical tasks on datasets in a privacy-preserving manner limiting user access to the actual data. Users can interact only with metadata from datamarts, including descriptions and variables and execute queries to filter datasets and extract desired cohorts. The infrastructure consists of three main modules: (i) Processor - installed at the data provider's premises, this module carries out computations. It accepts, as input, a *token* containing XML files with study descriptors: information about algorithm, data, and custom settings; (ii) Proxy - also installed at the data provider's premises, the Proxy manages user communications directed towards the Processor module; (iii) GUI - a web interface, deployable anywhere, where users can log in and execute analytical tasks.
Metadata and results storage was implemented using blockchain technology, specifically leveraging Hyperledger Fabric v2.4.7 [15]. The deployment was executed through a

Docker container and managed with Datome [16]. The choice of Hyperledger enables operating within a private blockchain environment, allowing experimentation in a secure and controlled manner. The Raft [17] consensus algorithm was used due to its simplicity. A single-node configuration sufficed for this study since the current focus does not extend to operational aspects.

The modularity of the proposed implementation allows for easy adaptation and reuse in other systems. Indeed, it was developed as a pluggable and agnostic software library in Python. For this purpose, an independent Graphical User Interface for technical team members to retrieve information from the blockchain (Figure 3) was developed.

A library [1](https://github.com/leocatnucc/distrib_block_crypto) was implemented to perform the workflow in Figure 1. It allows interacting with the blockchain via API and performing cryptographic tasks: Hash Calculation, Login/Logout from the Blockchain, Notarization in Blockchain, Blockchain Search by Filename or Identifier, Digital Signature, and Digital Signature Validation.

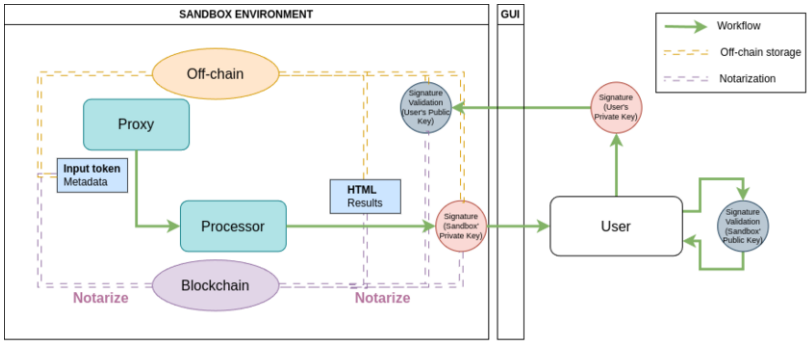

Figure 1. General Workflow Schema.

2.2. *General Workflow*

The general workflow can be summarised in the following steps:

- An initial notarization phase in the blockchain of the study settings, which in our case corresponds to the input token introduced earlier;
- Results are preliminarily notarized;
- The institution digitally signs the results;
- The users, download and verify the institution's signature using the institution's public key;
- The user digitally signs and uploads the results;
- The institution can verify the signature using the user's public key and acknowledge the user;
- The signed validated results are stored in the ledger;

The detailed workflow of the presented implementation, in Figure 1, is the following. When a user initiates a computational task, XML files are generated to record all information about the task. The module computes the hash of this data and preliminarily notarizes it in the blockchain. Upon completion of the computational process, a result

file is generated. Before making the result file available for display and download to the user, a hash of this record file is computed and notarized in the blockchain along with a copy signed with the infrastructure private key. Then, the user can download the results file from the GUI. The file is signed by the institution, allowing the user to verify its authenticity using the institution's public key. However, the result is not considered officially released by the institution until the user signs it with a personal private key and the signature is verified. Thus, after downloading the file and confirming the institution's signature, the user must sign it and re-upload it. The file signed by the user undergoes validation within the infrastructure using the user's public key, marking it as officially released by the institution. The hashes of all versions of the results, signed and unsigned, are also stored off-chain. The original copy is necessary for comparison with the copy signed by the users in the last validation phase. The other versions are useful for future comparison, when needed, with the hash present in the distributed ledger.Users can check the status of notarization and digital signs from the GUI as shown in Figure 2. If any of the signed files is altered, the validation with the corresponding public key will fail. All the results produced and signed are stored off-chain, along with the task metadata, for future cross-verification with the hashes in the blockchain.

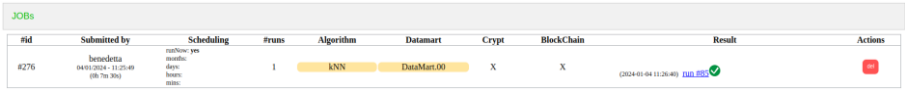

Figure 2. Completed and validated job with double digital signature as shown from the GUI.

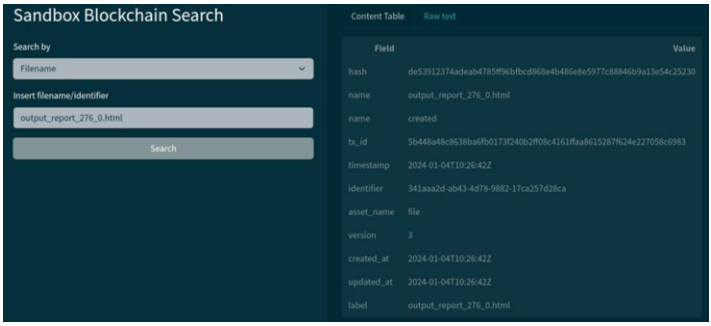

Figure 3. The utility GUI allows search for notarized assets' information.

3. Conclusions

Storing information from studies in clinical trials and healthcare research is crucial for future researchers to benefit from the results. Maintaining accountability in the research process protects researchers and data owners from misuse of study results. The proposed framework addresses these issues by providing a pluggable, modular library for managing blockchain transactions and digital signatures.

The testing of the library took place in a privacy-preserving distributed analytics environment. Actual data computations were used to test the approach, and the module securely stores metadata and results in the blockchain, making trusted information easily accessible to users. The digital signature process is also in place to prevent unauthorized modifications of the results. This ensures that a malicious third party cannot replace the

correct result with a modified version. Similarly, the institution can be confident that users will not alter the released results.
Several possible limitations of the presented approach should be explored in the future. The single-node setup leaves the scalability topic in multicentric scenarios open. Also, assuring scalability could result in a decreased security level that must match with regulations depending on the geographical area. A future industrial implementation should test and report on these topics. Moreover, the exploration of Non-Fungible Tokens (NFTs) integration could strengthen the verification of authorship for the studies. Furthermore, the deployment of Smart Contracts could expand certain capabilities as decentralized applications. Ultimately, a plan for industrial development is necessary, including testing various types of blockchains and consensus algorithms, as well as measuring performance metrics such as latency and transaction throughput.

References

[1] Nucciarelli L, Gottardelli B, Gatta R. distrib_block_crypto; 2024. Available from: https://github.com/leocatnucc/distrib_block_crypto.
[2] Lee S, Xu Y, D'Souza AG, Martin EA, Doktorchik CTA, Zhang Z, et al. Unlocking the Potential of Electronic Health Records for Health Research. International Journal of Population Data Science. 2020;5.
[3] Consolidated text: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation);. Last accessed 19 January 2024. Available from: https://eur-lex.europa.eu/eli/reg/2016/679/oj.
[4] Sharma S, Guleria K. A comprehensive review on federated learning based models for healthcare applications. Artificial Intelligence in Medicine. 2023;146:102691.
[5] Zhang P, Kamel Boulos MN. Privacy-by-Design Environments for Large-Scale Health Research and Federated Learning from Data. International Journal of Environmental Research and Public Health. 2022;19(19).
[6] Singh Y, Jabbar MA, Kumar Shandilya S, Vovk O, Hnatiuk Y. Exploring applications of blockchain in healthcare: road map and future directions. Frontiers in Public Health;11.
[7] Munjal K, Bhatia R. A systematic review of homomorphic encryption and its contributions in healthcare industry. Complex & Intelligent Systems. 2023;9.
[8] Gu X, Sabrina F, Fan Z, Sohail S. A Review of Privacy Enhancement Methods for Federated Learning in Healthcare Systems. International Journal of Environmental Research and Public Health. 2023;20(15).
[9] Moher D, Glasziou P, Chalmers I, Nasser M, Bossuyt PMM, Korevaar DA, et al. Increasing value and reducing waste in biomedical research: who's listening? The Lancet. 2016;387(10027):1573-86.
[10] Tumber MB, Dickersin K. Publication of clinical trials: accountability and accessibility. Journal of Internal Medicine. 2004;256(4):271-83.
[11] Chukwu E, Garg L. A Systematic Review of Blockchain in Healthcare: Frameworks, Prototypes, and Implementations. IEEE Access. 2020;8:21196-214.
[12] DeVito NJ, Bacon S, Goldacre B. Compliance with legal requirement to report clinical trial results on ClinicalTrials.gov: a cohort study. The Lancet. 2020;395(10221):361-9.
[13] Miron L, Gonc¸alves RS, Musen MA. Obstacles to the reuse of study metadata in ClinicalTrials.gov. Scientific Data. 2020;7(443).
[14] Gottardelli B, Gatta R, Nucciarelli L, et al. GEN-RWD Sandbox: Bridging the Gap Between Hospital Data Privacy and External Research Insights with Distributed Analytics. Research Square. 2023. PREPRINT (Version 1) available at Research Square. Available from: https://doi.org/10.21203/ rs.3.rs-3816282/v1.
[15] Open, Proven, Enterprise-grade DLT;. Last accessed 18 December 2023. Available from: https://www.ibm.com/downloads/cas/0XMOQJNP.
[16] Complex processes made easy;. Last accessed 18 December 2023. Available from: https://www.datome.io/.
[17] Ongaro D, Ousterhout J. In search of an understandable consensus algorithm. In: Proceedings of the 2014 USENIX Conference on USENIX Annual Technical Conference. USENIX ATC'14. USA: USENIX Association; 2014. p. 305–320.