Title

Quorum-based model learning on a blockchain hierarchical clinical research network using smart contracts

Authors

Kuo, Tsung-Ting

Pham, Anh

Publication Date

2023

DOI

10.1016/j.ijmedinf.2022.104924

Peer reviewed

HHS Public Access

Author manuscript

Int J Med Inform. Author manuscript; available in PMC 2023 March 03.

Published in final edited form as:

Int J Med Inform. 2023 January ; 169: 104924. doi:10.1016/j.ijmedinf.2022.104924.

Quorum-based model learning on a blockchain hierarchical clinical research network using smart contracts

Tsung-Ting Kuo[*],

Anh Pham

UCSD Health Department of Biomedical Informatics, University of California San Diego, La Jolla, CA, USA

Abstract

Background: Collaborative privacy-preserving modeling across several healthcare institutions allows for the construction of more generalizable predictive models while protecting patient privacy.

Objective: We aim at addressing the site availability issue on a hierarchical network by designing an immutable/transparent/source-verifiable quorum mechanism.

Methods: We developed an approach to combine a hierarchical learning algorithm, a novel Proof-of-Quorum (PoQ) consensus protocol, and a design of blockchain smart contracts. We constructed QuorumChain as an example and evaluated the scenarios of site-unavailability during the initialization and/or iteration phases of the modeling process on three healthcare/genomic datasets.

Results: When one or more sites would become unavailable, HierarchicalChain could not function, whereas QuorumChain improved predictive correctness significantly (the full Area Under the receiver operating characteristic Curve, or AUC, improved from 0.068 to 0.441, all with p-values < 0.001).

Conclusion: By constructing a quorum to continue the modeling process, QuorumChain possesses the capability to tackle the situation of sites being unavailable. It inherits the capability of learning on network-of-networks, improves learning continuity, and provides data/software immutability, transparency, and provenance, which can be important in expediting clinical research.

[*]Corresponding author at: [*]9500 Gilman Dr, San Diego, CA, USA. tskuo@ucsd.edu (T.-T. Kuo).

CRediT authorship contribution statement

Tsung-Ting Kuo: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Software, Validation, Visualization, Writing – original draft. **Anh Pham:** Data curation, Validation, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijmedinf.2022.104924.

JEL Classifications:

Clinical Information System; Data Privacy and Security; Machine Learning

Keywords

Blockchain Distributed Ledger Technology; Privacy-Preserving Predictive Modeling; Clinical Information System; Data Privacy and Security; Machine Learning

1. Introduction

Collaborative predictive modeling across multiple healthcare institutions can increase sample size and the generalizability of constructed machine learning models.[1–3] To further protect patient privacy, many existing studies (Fig. 1A) proposed to only disseminate partially trained predictive models without exchanging observation-level patient data.[4–11] This approach could also attract more institutions to participate in collaborative learning initiatives.[12] Meanwhile, prior privacy-preserving modeling approaches mainly relied on a central server to direct the learning process and aggregate the models from each site, which may pose potential concerns related to the exchanged models including mutability, unverifiability, and opaqueness. In real life, such risks of data being unverifiable or internally mishandled have substantial consequences.[13–15] It is seen that the threat of internal tampering has been causing severe damage [16,17]. Therefore, decentralized solutions (Fig. 1B) were proposed to address these issues [18–24]. Specifically, these decentralized modeling methods adopt blockchain, [25] a peer-to-peer data sharing technology.

More recently, decentralized modeling on research network-of-networks was also developed to mitigate practical issues [26]. However, the previous design of such methods does not consider the situation that a site may be leaving during the learning process (Fig. 2A). As a result, all sites would need to wait until the maintenance is completed to continue the learning process. Such breakdowns may also happen in higher levels of the hierarchical network and thus delay the modeling procedure.[27–31].

Although a complete pause of the modeling process may be reasonable in some cases, those scenarios could be relatively rare. In practice, most incidents only involve one or few sites not being available at a given time (Fig. 2B). In this case, a “quorum” mechanism may be desirable to allow the learning process to continue, so that most of the records can still be used to build predictive models without being interrupted. Ideally, such a quorum mechanism should be immutable, transparent, and source-verifiable. Smart contracts,[32–34] which execute programs on a blockchain, possess these desirable features when compared to traditional off-chain software.[35–38,59] Therefore, adopting smart contracts for quorum computation could further improve the *software* immutability, transparency and provenance of the quorum mechanism.

The objective of this study is to address the site-unavailability issue on a blockchain hierarchical network, with three goals to improve the process of predictive modeling, including: (1) inheriting the benefits of privacy-preserving modeling on hierarchical

network-of-networks; (2) enhancing the continuity of model learning during site-unavailability events; and (3) retaining the immutability, transparency, and provenance for both data and software.

2. Material and methods

We developed QuorumChain to evaluate the design of the quorum mechanism for hierarchical research networks, which is the objective of this study. To further achieve our three goals, we leveraged the model learning components of HierarchicalChain, [26] developed a novel Proof-of-Quorum consensus algorithm, and adopted blockchain smart contracts. These three parts will be described in the following three subsections, respectively. The system implementation, evaluation datasets, and experiment settings will then be discussed in the latter half of this section.

2.1. HierarchicalChain decentralized model learning for network-of-networks

To inherit the benefits of privacy-preserving modeling on hierarchical networks, we leveraged the model learning parts of HierarchicalChain,[26] which is based on the GLORE logistic regression algorithm.[5] On one hand, we used the batch learning algorithm of HierarchicalChain to achieve the same level of prediction correctness when compared to its centralized counterpart [5,20]. On the other hand, we adapted the ensemble methods to aggregate the models learned in each level. The ensemble can be either horizontal (i.e., generating the prediction score of a site using the scores from all Level 1 models weighted averaged by the training data sizes), or vertical (i.e., generating the weighted-average prediction score of a site using the scores from each level related to that site) [26]. We developed and evaluated QuorumChain using both ensemble approaches.

2.2. The Proof-of-Quorum (PoQ) consensus learning algorithm

To enhance the continuity of model learning during the maintenance events, we designed a Proof-of-Quorum (PoQ) algorithm to check if most of the data records are available for learning. As illustrated in Fig. 3, even if a site is unavailable, given that quorums can still be formed in a level (a predetermined amount of data still available), the model for that level can be learned using the data from the remaining quorum sites. The checking of the quorum for the initialization phase was considered in Algorithm A.1 (all algorithms are illustrated in Appendix A.1), while the iteration phase was taken care of in Algorithm A.2 and A.4. During the model ensemble phase in Algorithm A.3, the non-quorum-consensus models were assigned a weight of zero, thus only the models with a consensus by a quorum were included in the computation. The PoQ algorithm was adapted from Proof-of-Hierarchy,[26] with three new smart contracts: *Quorum Contract* (for quorum checking), *Catalog Contract* (for obtaining smart contract addresses), and *Model Contract* (for model dissemination). Specifically, in the Quorum Contract, we computed and checked whether a quorum could be formed with majority of patient records (i.e., ≥ a pre-defined percentage threshold) within a certain time (i.e., ≤ a pre-defined count limit). Such a design allows continuous learning when a data-contributing party is taken off the network and is the main advantage of the PoQ Algorithm.

2.3. The blockchain network and smart contract design

We used blockchain and smart contracts for QuorumChain. Blockchain has the benefits of data immutability, transparency, and provenance. These features are important to record models, as well as for various healthcare applications [39–45,60]. Therefore, we adopted blockchain instead of traditional distributed databases. QuorumChain adopted a permissioned blockchain with which only authorized sites can join the network to further protect privacy. Although the research network-of-networks are hierarchical, we use a single blockchain network for all participating sites to improve efficiency.

To retain immutability, transparency, and provenance for the software, we also adopted smart contracts. The software components of QuorumChain are demonstrated in Figure A.1 (in Appendix A.2), with three smart contracts: (i) Quorum Contract to compute quorums from sites; (ii) Model Contract to manage model and *meta*-data on-chain; and (iii) Catalog Contract to manage addresses of for the Quorum Contract and the Model Contract. Before the execution of QuorumChain, the three smart contracts were first deployed on the blockchain. Then, the PoQ Algorithm used the address of the Catalog Contract to locate Quorum and Model Contracts, and then started the model learning process of QuorumChain. The models were stored on-chain (Model Contract) and learned off-chain (PoQ Algorithm). The implementation of QuorumChain is explained in detail in Appendix A.3.

2.4. Datasets

The evaluation of QuorumChain was conducted on three datasets (Table 1). The first dataset is Edinburg Myocardial Infarction (“Edin”) with 1,253 samples.[46] The prediction target of this dataset is a binary outcome of the presence of disease (with 21.9 % patient records having positive ground truth labels). The Edin dataset includes nine binary covariates such as “Pain in Right Arm” or “Nausea”. The second dataset is Cancer Biomarker (“CA”) with 141 patients.[47] The outcome to be predicted is a binary indicator of the presence of cancer (with 63.8 % positive samples). The CA dataset contains two numerical values of biomarkers (i.e., “CA-19” and “CA-125”. These two datasets are both publicly available and contain no PHI data. We adopted Clostridium Difficile Infection (“C-Diff”) as our third and larger test dataset of 157,493 patients (more details in Appendix A.4).[48] The task is to predict the binary presence of infection, defined as having at least one positive C-Diff lab test within the data collection timeframe. The C-Diff dataset is highly imbalanced with only 1,541 (i.e., 1 %) of positive patients, and contains 25 covariates. The Institutional Review Board (IRB) at UCSD approved this study (190385X) on May 26, 2020.

2.5. Experiment settings

As the site-unavailability issue may happen during either the initialization or the iteration phases of the modeling process, to evaluate how QuorumChain performs, we simulated the site maintenance scenarios in both initialization and iteration phases of the predictive modeling process. We compared QuorumChain with HierarchicalChain, [26] a blockchain-based privacy-preserving predictive modeling approach designed for hierarchical research networks. Our hierarchical network was configured with the total number of participating sites $N = 4$, and the number of levels $H = 3$. For the three datasets (i.e., Edin, CA, and C-Diff), we randomly divided each one into four parts. For each part, the data were then

randomly split into 50 % training and 50 % test records, both containing at least one positive and one negative record. The evaluation process was repeated 30 times, with the blockchain network being reset before each trial, to collect the results. Our evaluation metrics include the full Area Under the receiver operating characteristic Curve (AUC) [49–51] with 2-sampled *t*-test (alpha = 0.05), the modeling success rate (i.e., the percentage of the trials completing the model ensemble learning process, out of the total 30 trials), the number of learning iterations, and the execution times, all averaged over four sites. Detailed hyper-parameter settings are described in Appendix A.5.

2.6. Site maintenance simulation scenarios

We simulated four scenarios (i.e., “Init-Ideal/Iter-Ideal”, “Init-Ideal/Iter-Practical”, “Init-Practical/Iter-Ideal”, and “Init-Practical/Iter-Practical”) for the site maintenance situations (details in Appendix A.6). “Init-Ideal/Iter-Ideal” refers to the scenario of an “ideal” site availability, during both initiation and iteration steps, with every site participating. “Init-Ideal/Iter-Practical” means all sites are available during initiation, but some sites are off the network during iteration, and the like. During such instances, for QuorumChain, if a quorum cannot be formed, the model construction is considered unsuccessful, and the AUC value was set to 0.5 (i.e., the lowest possible value). For HierarchicalChain, since there is no mechanism to deal with site-unavailability, we estimated its prediction correctness and modeling success rate indirectly based on the results of QuorumChain. That is, if the AUC of QuorumChain in a trial was different from that in the perfect “Init-Ideal / Iter-Ideal” scenario, in which QuorumChain is equivalent to HierarchicalChain, we regarded it as a HierarchicalChain’s unsuccessful model building trial, and HierarchicalChain’s AUC was set to 0.5. This way, we could still compare the two methods on a relatively fair basis.

3. Results

3.1. Predictive correctness

As shown in Fig. 4, the predictive correctness in terms of AUC for both QuorumChain and HierarchicalChain were the same in the “Init-Ideal/Iter-Ideal” scenario as expected, while QuorumChain outperformed HierarchicalChain in all other scenarios, for all three datasets and two ensemble methods (AUC improved from 0.068 to 0.441, all with p-values < 0.001, indicating statistically significant differences). Specifically, QuorumChain maintained relatively good performance in the “Init-Ideal/Iter-Practical” scenario, indicating that it is more resistant to the site-unavailability situation during the learning iterations. In contrast, HierarchicalChain suffered a higher correctness penalty under this scenario. The “Init-Practical/Iter-Ideal” and “Init-Practical/Iter-Practical” scenarios resulted in higher AUC drop for QuorumChain, showing the higher impact of the “Init-Practical” situation. The performance of QuorumChain in the “Init-Practical/Iter-Practical” scenario is either similar or slightly reduced as compared to the “Init-Practical/Iter-Ideal” situation, reassuring the resistance of QuorumChain under Iter-Practical. For details, see Appendix A.6. Overall, QuorumChain still performed better than HierarchicalChain in handling site-unavailability.

3.2. Modeling successfulness

The modeling success rate results in different scenarios are depicted in Fig. 5. The general patterns are like those of the prediction correctness. Compared to HierarchicalChain, QuorumChain was more likely to complete the modeling process in all non-perfect scenarios. Specifically, QuorumChain attained 100 % of modeling success rate in the “Init-Ideal/Iter-Practical” scenario, demonstrating high modeling endurance in this scenario.

3.3. Number of learning iterations

The learning iteration results of QuorumChain are shown in Table 2. QuorumChain constructed models with fewer iterations, reflecting the fact that some sites stopped learning in either of, or both initialization and iteration phases, thus the number of iterations was not increased. In general, the average number of iterations is the lowest in the “Init-Practical/Iter-Practical” scenario, followed by “Init-Practical/Iter-Ideal”. The results of the standard deviation of learning iterations are mixed on different datasets.

3.4. Execution time

As also demonstrated in Table 2, the total execution time reflects the number of iterations; it is the shortest in the “Init-Practical/Iter-Practical” scenario. In contrast, the per-iteration execution times are in general longer in the non-perfect situations; a situation with fewer learning iterations in general has a larger per-iteration time, implying the existence of an overhead time to form quorum. The longest per-iteration time is in the “Init-Practical/Iter-Practical” scenario. The execution time of HierarchicalChain is reflected in QuorumChain’s “Init-Ideal/Iter-Ideal.”.

4. Discussion

4.1. Findings

Based on our results, our proposed QuorumChain can tackle the situation of sites being unavailable by constructing a quorum to continue the modeling process, with four specific observations: (1) The predictive correctness AUC was mainly related to the modeling successfulness, indicating that AUC corresponds to the amount of data available for learning. For example, comparing to “Init-Ideal/Iter-Ideal”, the “Init-Ideal / Iter-Practical” scenario had the same (i.e., 100 %) modeling success rate, while the AUC dropped for about 0.01 – 0.04. (2) QuorumChain is more resistant to the “Init-Ideal/Iter-Practical” scenario in terms of predictive correctness and modeling successfulness than the “Init-Practical/Iter-Ideal” scenario. In practice, the duration of the iteration phase is usually longer than that of initialization. Therefore, QuorumChain would be suitable to handle the more likely scenario that sites may become unavailable during the iteration phase. (3) We configured a relatively high site-unavailability rate (i.e., 25 % in the initialization phase, and 75 % at each hierarchy level in the iteration phase, assuming 100 learning iterations) to demonstrate the effectiveness of QuorumChain. In practice, the unavailability rates could be lower, leading to higher predictive correctness and modeling successfulness. (4) QuorumChain inherits the technical advantages of HierarchicalChain such as being able to handle hierarchical research networks and providing the exact predictive correctness when all sites are available. In

addition, QuorumChain offers novel features such as fault tolerance and immutable code. A comparison of QuorumChain and related studies is summarized in Table 3.

4.2. Limitations

The constraints of this study are as follows: (1) The scenarios regarding "site recovery" are yet to be investigated. In our experiments, we only simulated the site-unavailability scenarios in the initialization phase (assuming the site would remain down for the entire trial), and the iteration phase (assuming the site would "recover" at the next hierarchy level). In practice, just like a site can be unavailable in different phases, it can also recover in different phases. These scenarios warrant further studies and experiments. (2) QuorumChain is yet to be improved for scalability optimization. Although our results demonstrated that the learning iteration and execution time of the largest dataset C-Diff did not largely increase when compared to the other two smaller datasets Edin and CA, it only indicated the scalability in terms of the number of samples (i.e., C-Diff includes > 100 times more records than the other two datasets) but not necessarily the number of covariates as C-Diff, Edin and CA have 25, 9, and 2 covariates, respectively. (3) Different distributed technologies are yet to be leveraged and compared for implementation. For example, other blockchain platforms that support smart contracts [52,53] could also be tested. Non-blockchain systems [54–56] could potentially be adopted and evaluated as well.

From the above, it may seem that as compared to centralized solutions, distributed learning using blockchain technology could be less competitive when it comes to runtime. This should not be the single detrimental factor in considering the utilization of blockchain for healthcare data management purposes. In recent times, solutions have been proposed and explored to speed up blockchain transaction times with promising results. For example, Ethereum is considering proposals such as Proof-of-Stake (PoS) [57] and Sharding [58] to speed up transactions for permissionless blockchain networks, which may later be adopted by the permissioned blockchain used by QuorumChain. Such innovations point towards the feasibility of using blockchain to boost security in a wide range of health applications, among which QuorumChain stands to offer benefits.

5. Conclusion

We developed and evaluated QuorumChain, a quorum mechanism addressing the site-unavailability issue on a blockchain hierarchical network, with three main components: (1) a state-of-the-art hierarchical privacy-preserving modeling method to inherit the capability of learning on network-of-networks; (2) a novel PoQ algorithm to form quorums representing the majority of the patient records and thus improving learning continuity; and (3) a system design based on blockchain and smart contract to provide data/software immutability, transparency and provenance. Although there are no monetary incentives for the QuorumChain participants, the reward for each site to join is the improved prediction correctness through a more generalizable consensus model, a more robust model learning with the quorum mechanism, and more immutable/transparent/source-verifiable partial models during the modeling process. These benefits could be important for blockchain-based modeling methods to take another step further towards the real-world deployment

across health institutions to expedite clinical, genomic, and biomedical research. Potential future works include exploring the situation of site recovery (a site may keep going up-and-down intermittently), optimizing scalability (number of covariates and values of hyper-parameters), and adopting different distributed technologies (other blockchain platforms or even non-blockchain ones).

SUMMARY TABLE

What was already known on the topic	Collaborative predictive modeling can increase sample size and the model generalizability
	Privacy-preserving learning can train models collaboratively
	Decentralized solutions can address known security concerns especially on research network-of-networks
	Decentralized solutions can address known security concerns especially on research network-of-networks
What this study added to our knowledge	Development of QuorumChain addresses the site-unavailability issue on a blockchain hierarchical network
	Evaluation of QuorumChain demonstrates a more robust model learning protocol by using the quorum mechanism
	QuorumChain serves as a cornerstone of deploying collaborative modeling across institutions to expedite clinical/genomic/biomedical research

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors would like to thank Lucila Ohno-Machado, MD, PhD and Robert El-Kareh, MD, MS, MPH, Jihoon Kim, MS, Kai Post, MS, Tyler Bath, and Jeffery Tellew, for very helpful discussions, Michael Hogarth, MD, Andrew Greaves and Jit Bhattacharya, MS, for the technical support of the iDASH 2.0 cloud network, as well as Cyd Burrows-Schilling, MS, and Randi Sutphin for the technical support of the UCSD Campus AWS cloud network.

Funding

The authors were funded by the U.S. National Institutes of Health (NIH) (R00HG009680, R01EB031030, R01GM118609, R01HG011066, R01HL136835, RM1HG011558, U24LM013755, and U54HG012510), a UCSD Academic Senate Research Grant (RG084150), and the Graduate Division San Diego Matching Fellowship associated with San Diego Biomedical Informatics Education & Research (SABER) NIH National Library of Medicine (NLM) grant T15LM011271. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The use of the integrating Data for Analysis, Anonymization, and SHaring (iDASH) 2.0 and the UCSD Campus Amazon Web Services (AWS) cloud network was supported by Michael Hogarth, MD.

References

[1]. Navathe AS, Conway PH, Optimizing health information technology's role in enabling comparative effectiveness research, The American journal of managed care 16(12 Suppl HIT):SP44–7 (2010). [PubMed: 21314220]

[2]. Wicks P, Vaughan TE, Massagli MP, Heywood J, Accelerated clinical discovery using self-reported patient data collected online and a patient-matching algorithm, Nature biotechnology 29 (5) (2011) 411–414.

[3]. Landrum MJ, Lee JM, Benson M, et al. , ClinVar: public archive of interpretations of clinically relevant variants, Nucleic acids research 44 (D1) (2016) D862–D868. [PubMed: 26582918]

[4]. “Secure” log-linear and logistic regression analysis of distributed databases. International Conference on Privacy in Statistical Databases; 2006. Springer.

[5]. Wu Y, Jiang X, Kim J, Ohno-Machado L. Grid Binary LOgistic REgression (GLORE): building shared models without sharing data. Journal of the American Medical Informatics Association 2012;19(5):758–64 doi: 10.1136/amiajnl-2012-000862 [published Online First: Epub Date]|. [PubMed: 22511014]

[6]. Yan F, Sundaram S, Vishwanathan S, Qi Y. Distributed autonomous online learning: Regrets and intrinsic privacy-preserving properties. IEEE Transactions on Knowledge and Data Engineering 2013;25(11):2483–93 doi: 10.1109/TKDE.2012.191[published Online First: Epub Date]|.

[7]. Wang S, Jiang X, Wu Y, Cui L, Cheng S, Ohno-Machado L. Expectation propagation logistic regression (explorer): distributed privacy-preserving online model learning. Journal of biomedical informatics 2013;46(3):480–96 doi: 10.1016/j.jbi.2013.03.008[published Online First: Epub Date]|. [PubMed: 23562651]

[8]. El Emam K, Samet S, Arbuckle L, Tamblyn R, Earle C, Kantarcioglu M, A secure distributed logistic regression protocol for the detection of rare adverse drug events, Journal of the American Medical Informatics Association 20 (3) (2013) 453–461, doi: 10.1136/amiajnl-2011-000735[published Online First: Epub Date]|. [PubMed: 22871397]

[9]. A collaborative privacy-preserving deep learning system in distributed mobile environment. 2016 International Conference on Computational Science and Computational Intelligence (CSCI); 2016. IEEE.

[10]. Aono Y, Hayashi T, Phong LT, Wang L. Privacy-preserving logistic regression with distributed data sources via homomorphic encryption. IEICE TRANSACTIONS on Information and Systems 2016;99(8):2079–89 doi: 10.1587/transinf.2015INP0020 [published Online First: Epub Date]|.

[11]. Rocher L, Hendrickx JM, de Montjoye Y-A. Estimating the success of re-identifications in incomplete datasets using generative models. Nature Communications 2019;10(1):3069 doi: 10.1038/s41467-019-10933-3[published Online First: Epub Date]|.

[12]. Kim J, Neumann L, Paul P, et al. Privacy-Protecting, Reliable Response Data Discovery Using COVID-19 Patient Observations. medRxiv 2020: 2020.09.21.20196220 doi: 10.1101/2020.09.21.20196220[published Online First: Epub Date]|.

[13]. Klas ME, Blaskey S, Herald NN. A Numbers Game: The Florida COVID-19 data said one thing while Gov. DeSantis sometimes said another. 2020. https://www.miamiherald.com/news/politics-government/state-politics/article242937591.html (accessed February 3, 2021).

[14]. Wamsley L Fired Florida Data Scientist Launches A Coronavirus Dashboard Of Her Own. 2020. https://www.npr.org/2020/06/14/876584284/fired-florida-data-scientist-launches-a-coronavirus-dashboard-of-her-own (accessed February 3, 2021).

[15]. ThePodiatricOfficesOfBobbyYee. Notice of Data Incident Regarding the Podiatric Offices of Bobby Yee. 2018. https://www.prnewswire.com/news-releases/notice-of-data-incident-regarding-the-podiatric-offices-of-bobby-yee-300769294.html (accessed February 3, 2021).

[16]. U.S.Attorney’sOfficeDistrictofMinnesotaDepartmentOfJustice. Former It Employee Of Transcontinental Railroad Sentenced To Prison For Damaging Ex-Employer’s Computer Network. 2018. https://www.justice.gov/usao-mn/pr/former-it-employee-transcontinental-railroad-sentenced-prison-damaging-ex-employer-s (accessed February 3, 2021).

[17]. ManhattanDistrictAttorney’sOffice. D.A. Vance: Former Century 21 Employee Charged with Computer Tampering, Larceny For Breach of Company Data. 2020. https://www.manhattanda.org/d-a-vance-former-century-21-employee-charged-with-computer-tampering-larceny-for-breach-of-company-data/ (accessed February 3, 2021).

[18]. Kuo T-T, Hsu C-N, Ohno-Machado L ModelChain: Decentralized Privacy-Preserving Healthcare Predictive Modeling Framework on Private Blockchain Networks. ONC/NIST Use of Blockchain for Healthcare and Research Workshop. September 26, 2016 - September 27, 2016. Gaithersburg, Maryland, United States, 2016.

[19]. When Machine Learning Meets Blockchain: A Decentralized, Privacy-preserving and Secure Design. 2018 IEEE International Conference on Big Data (Big Data); 2018; December 10, 2018 - December 13, 2018. Seattle, WA, United States. IEEE.

[20]. Kuo T-T, Gabriel RA, Ohno-Machado L. Fair compute loads enabled by blockchain: sharing models by alternating client and server roles. Journal of the American Medical Informatics Association (JAMIA) 2019;26(5):392–403 doi: 10.1093/jamia/ocy180[published Online First: Epub Date]|. [PubMed: 30892656]

[21]. Kim H, Kim S-H, Hwang JY, Seo C. Efficient privacy-preserving machine learning for blockchain network. IEEE Access 2019;7:136481–95 doi: 10.1109/ACCESS.2019.2940052[published Online First: Epub Date]|.

[22]. Kuo T-T, Gabriel RA, Cidambi KR, Ohno-Machado L. EXpectation Propagation LOgistic REgRession on permissioned blockCHAIN (ExplorerChain): decentralized online healthcare/genomics predictive model learning. Journal of the American Medical Informatics Association (JAMIA) 2020;27(5):747–56 doi: 10.1093/jamia/ocaa023[published Online First: Epub Date]|. [PubMed: 32364235]

[23]. Kuo T-T. The Anatomy of a Distributed Predictive Modeling Framework: Online Learning, Blockchain Network, and Consensus Algorithm. Journal of the American Medical Informatics Association Open (JAMIA Open). 2020;3(2):201–08 doi: 10.1093/jamiaopen/ooaa017[published Online First: Epub Date]|. [PubMed: 32734160]

[24]. Kuo T-T, Pham A Detecting Model Misconducts in Decentralized Healthcare Federated Learning. International Journal of Medical Informatics 2022;158: 104658 doi: 10.1016/j.ijmedinf.2021.104658[published Online First: Epub Date]|.

[25]. Nakamoto S, Bitcoin: A peer-to-peer electronic cash system, Decentralized Business Review (2008:) 21260.

[26]. Kuo T-T, Kim J, Gabriel RA. Privacy-Preserving Model Learning on Blockchain Network-of-networks. Journal of the American Medical Informatics Association (JAMIA) 2020;27(3):343–54 doi: 10.1093/jamia/ocz214[published Online First: Epub Date]|. [PubMed: 31943009]

[27]. Weniger D Ramsey County says illegal ransomware hack compromised info of 8,700 clients. 2021. https://www.twincities.com/2021/01/29/ramsey-county-says-illegal-ransomware-hack-compromised-info-of-8700-clients/ (accessed February 3, 2021).

[28]. Womack B CyrusOne says six customers affected by ransomware attack. 2019. https://www.bizjournals.com/dallas/news/2019/12/05/cyrusone-ransomware.html (accessed February 3, 2021).

[29]. GuardianNews&MediaLimited. British Airways IT failure caused by ‘uncontrolled return of power’. 2017. https://www.theguardian.com/business/2017/may/31/ba-it-shutdown-caused-by-uncontrolled-return-of-power-after-outage (accessed Febrary 3, 2021).

[30]. Isidore C Delta: 5-hour computer outage cost us $150 million. 2016. https://money.cnn.com/2016/09/07/technology/delta-computer-outage-cost/ (accessed February 3, 2021).

[31]. PowellTribune. Campbell County Health awaits payment for ransomware attack. 2020. https://www.powelltribune.com/stories/campbell-county-health-awaits-payment-for-ransomware-attack,23609 (accessed February 3, 2021).

[32]. Buterin V, A next-generation smart contract and decentralized application platform, white paper 3 (27) (2014).

[33]. TheLinuxFoundation. Hyperledger Architecture, Volume II: Smart Contracts. 2018. https://www.hyperledger.org/wp-content/uploads/2018/04/Hyperledger_Arch_WG_Paper_2_SmartContracts.pdf (accessed January 4, 2021).

[34]. Yu H, Sun H, Wu D, Kuo T-T, Comparison of Smart Contract Blockchains for Healthcare Applications, American Medical Informatics Association, Bethesda, MD, AMIA Annual Symposium., 2019.

[35]. Mun Li M, Kuo T-T. Previewable Contract-Based On-Chain X-Ray Image Sharing Framework for Clinical Research. International Journal of Medical Informatics 2021:104599 doi: 10.1016/j.ijmedinf.2021.104599[published Online First: Epub Date]|. [PubMed: 34628257]

[36]. Kuo T-T, Bath T, Ma S, et al. , Benchmarking Blockchain-Based Gene-Drug Interaction Data Sharing Methods: A Case Study from the iDASH 2019 Secure Genome Analysis Competition

Blockchain Track, International Journal of Medical Informatics 154 (2021), 104559 doi: 10.1016/j.ijmedinf.2021.104559[published Online First: Epub Date]|. [PubMed: 34474309]

[37]. Tellew J, Kuo T-T. CertificateChain: decentralized healthcare training certificate management system using blockchain and smart contracts. JAMIA Open 2022;5(1) doi: 10.1093/jamiaopen/ooac019[published Online First: Epub Date]|.

[38]. Kuo T-T, Jiang X, Tang H, et al. iDASH Secure Genome Analysis Competition 2018: Blockchain Genomic Data Access Logging, Homomorphic Encryption on GWAS, and DNA Segment Searching. BMC Medical Genomics 2020;13(7):98 doi: 10.1186/s12920-020-0715-0[published Online First: Epub Date]|. [PubMed: 32693816]

[39]. Angraal S, Krumholz HM, Schulz WL, Blockchain technology: applications in health care, Circulation: Cardiovascular Quality and Outcomes 10 (9) (2017) e003800. [PubMed: 28912202]

[40]. Kuo T-T, Kim H-E, Ohno-Machado L Blockchain distributed ledger technologies for biomedical and health care applications. Journal of the American Medical Informatics Association (JAMIA) 2017;24(6):1211–20 doi: 10.1093/jamia/ocx068 [published Online First: Epub Date]|. [PubMed: 29016974]

[41]. Mettler M, Blockchain technology in healthcare: The revolution starts here, in: 2016 IEEE 18th International Conference on E-Health Networking, Applications and Services (Healthcom), IEEE, Munich, Germany, 2016, pp. 1–3.

[42]. Wong DR, Bhattacharya S, Butte AJ, Prototype of running clinical trials in an untrustworthy environment using blockchain, Nature communications 10 (1) (2019) 1–8.

[43]. B-fica: Blockchain based framework for auto-insurance claim and adjudication. 2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData); 2018. IEEE.

[44]. Sylim P, Liu F, Marcelo A, Fontelo P, Blockchain technology for detecting falsified and substandard drugs in distribution: pharmaceutical supply chain intervention, JMIR research protocols 7 (9) (2018) e10163. [PubMed: 30213780]

[45]. Grishin D, Obbad K, Estep P, et al. , Accelerating genomic data generation and facilitating genomic data access using decentralization, privacy-preserving technologies and equitable compensation, Blockchain in Healthcare Today (2018).

[46]. Kennedy R, Fraser H, McStay L, Harrison R. Early diagnosis of acute myocardial infarction using clinical and electrocardiographic data at presentation: derivation and evaluation of logistic regression models. European heart journal 1996;17(8): 1181–91 doi: 10.1093/oxfordjournals.eurheartj.a015035[published Online First: Epub Date]|. [PubMed: 8869859]

[47]. Zou KH, Liu A, Bandos AI, Ohno-Machado L, Rockette HE. Statistical evaluation of diagnostic performance: topics in ROC analysis: CRC Press, Boca Raton, FL, 2011.

[48]. Pham A, El-Kareh R, Ohno-Machado L, Kuo T-T. Early Prediction of Positive Clostridioides Difficile Test Results. AMIA Annual Symposium, 2021.

[49]. Lasko TA, Bhagwat JG, Zou KH, Ohno-Machado L, The use of receiver operating characteristic curves in biomedical informatics, Journal of biomedical informatics 38 (5) (2005) 404–415. [PubMed: 16198999]

[50]. Hanley JA, McNeil BJ, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 143 (1) (1982) 29–36. [PubMed: 7063747]

[51]. Davis J, Goadrich M. The Relationship Between Precision-Recall and ROC Curves. 23rd International Conference on Machine Learning (ICML). June 25, 2006 - June 29, 2006. Pittsburgh, Pennsylvania, USA, 2006:233–40.

[52]. Hyperledger fabric: a distributed operating system for permissioned blockchains. Proceedings of the thirteenth EuroSys conference; 2018.

[53]. Brown RG. The Corda Platform: An Introduction. Retrieved 2018;27:2018. https://www.corda.net/content/corda-platform-whitepaper.pdf.

[54]. Boyd S, Ghosh A, Prabhakar B, Shah D, Randomized gossip algorithms, IEEE/ACM Transactions on Networking (TON) 14(SI):2508–30 (2006).

[55]. Boyd S, Ghosh A, Prabhakar B, Shah D. Gossip algorithms: Design, analysis and applications. Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies. March 13, 2005 - March 17, 2005. Miami, Florida, USA: IEEE, 2005:1653–64.

[56]. Shah D Gossip algorithms. Foundations and Trends® in Networking 2009;3(1):1–125 doi: 10.1561/1300000014[published Online First: Epub Date]|.

[57]. TheEthereumCommunity. The Beacon Chain. 2021. https://ethereum.org/en/eth2/beacon-chain/ (accessed February 3, 2021).

[58]. TheEthereumFoundation. Shard chains. 2021. https://ethereum.org/en/eth2/shard-chains/ (accessed February 3, 2021).

[59]. Kuo T-T, Jiang X, Tang H, Wang X, Harmanci A, Kim M, Post K, Bu D, Bath T, Kim J, Liu W, Chen H, Ohno-Machado L. The evolving privacy and security concerns for genomic data analysis and sharing as observed from the iDASH competition. Journal of the American Medical Informatics Association. 2022. 10.1093/jamia/ocac165.

[60]. Kuo T-T, Zavaleta Rojas H, Ohno-Machado L. Comparison of blockchain platforms: a systematic review and healthcare examples. Journal of the American Medical Informatics Association (JAMIA). 2019;26(5):462–78. Epub March 25, 2019. 10.1093/jamia/ocy185. [PubMed: 30907419]

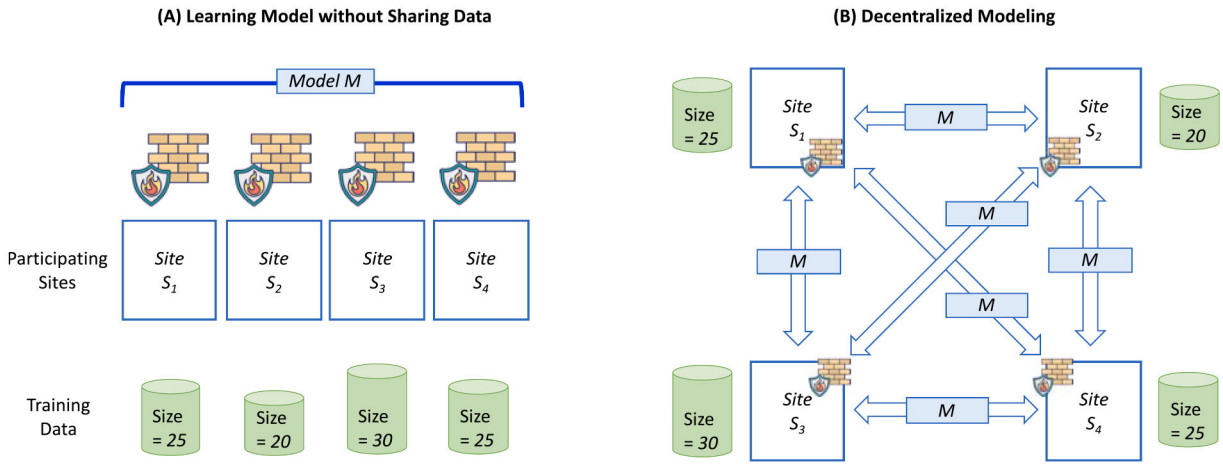

Fig. 1.

Comparison of privacy-preserving predictive modeling approaches. **A**. Privacy-preserving predictive modeling. The main principle of constructing privacy-preserving models is learning models without sharing data. Although each site uses their own data to create and update the model, the patient data is kept within each site and never disseminated. By exchanging only partially trained models with other sites, the model can be trained collaboratively to increase generalizability while still protecting patient privacy. **B.** Decentralized privacy-preserving modeling. Without a central server, the learning process obtain benefits such as immutability (i.e., the model recorded on chain cannot be tampered with), provenance (i.e., the source site of each model is verifiable), and transparency (i.e., every site can see all partially trained models).

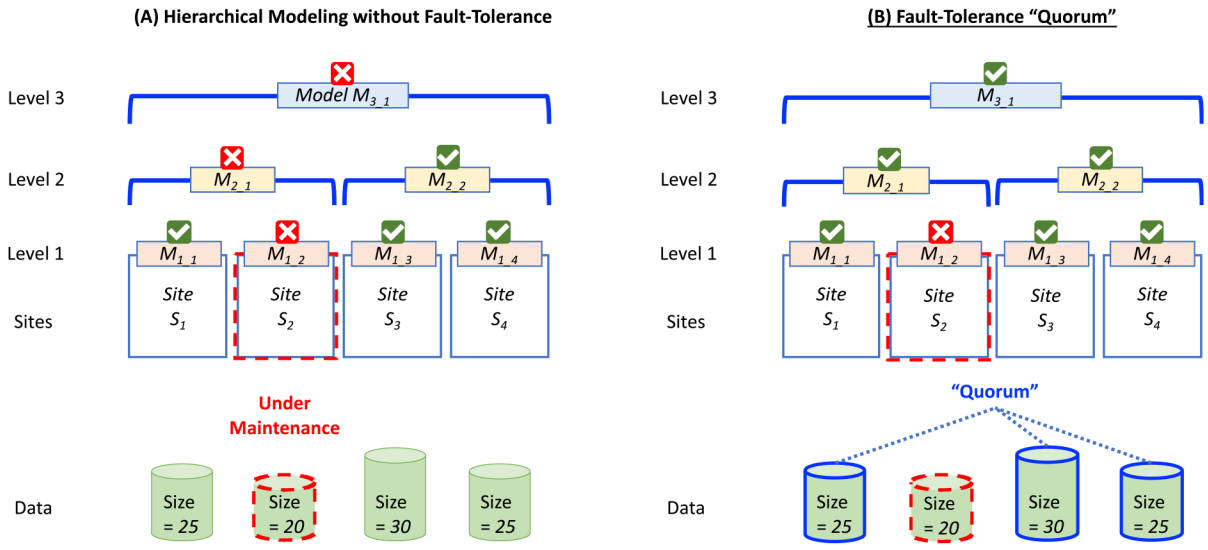

Fig. 2.

Hierarchical modeling. **A.** Hierarchical modeling without fault-tolerance. The models in each level (e.g., M_{2_1} and M_{2_2} for level 2) are first constructed, and then they are combined to improve prediction correctness. Although building models within a research network-of-networks is practical, the learning process could be delayed because a site (e.g., S_2) is under maintenance. In this case, models M_{1_2}, M_{2_1}, and eventually M_{3_1} cannot be built until site S_2 completes the maintenance and rejoins the network. **B.** Fault-tolerance "quorum". Although site S_2 may not be available to participate in the learning process and build M_{1_2}, it only contains 20 / (25 + 20 + 30 + 25) = 20 % of patient records. Therefore, the other sites (S_1, S_3, and S_4) form a "quorum" of 80 % of the patient records, and should be able to represent majority (e.g., ≥ 51 %) of the data to continue building models such as M_{2_1} and M_{3_1}.

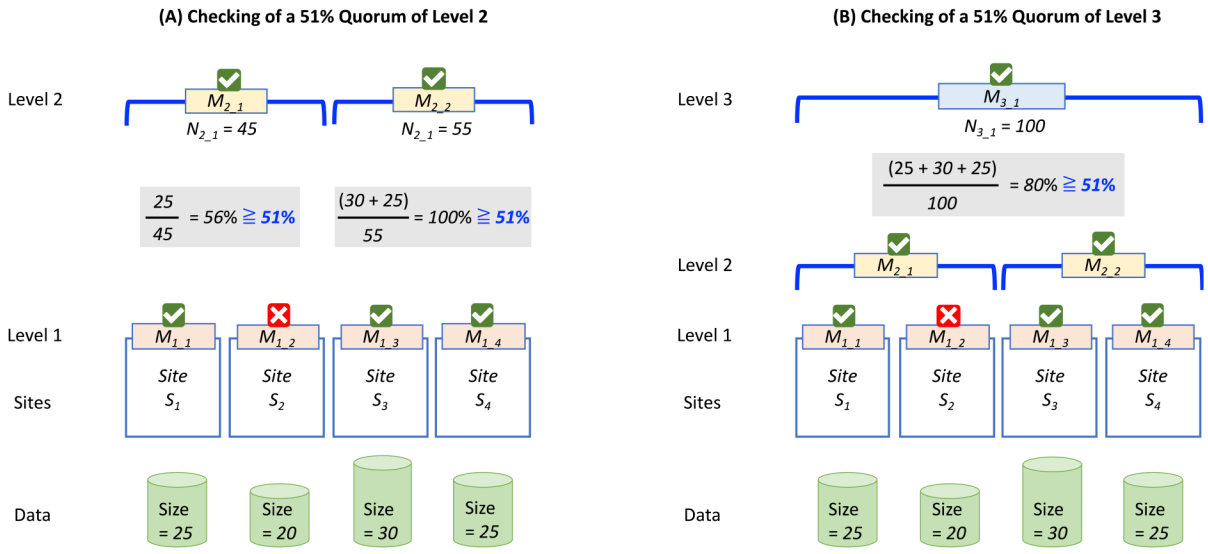

Fig. 3.

Intuition of QuorumChain with a 51 % quorum. **A.** Quorum of Level 2. Within the first sub-network of S_1 and S_2, if site S_2 is not available, data from S_1 alone can still form a quorum of 56 % (≥51 %), and therefore model M_{2_1} can be learned on data strictly from S_1. Meanwhile, the second sub-network of sites S_3 and S_4 form a quorum of 100 %, and therefore model M_{2_2} can be learned on data from both S_3 and S_4. **B.** Quorum of Level 3. In this case, sites S_1, S_3 and S_4 form a quorum of 80 %, thus the data from these three sites can be used to learn model M_{3_1}, even when site S_2 is still not available.

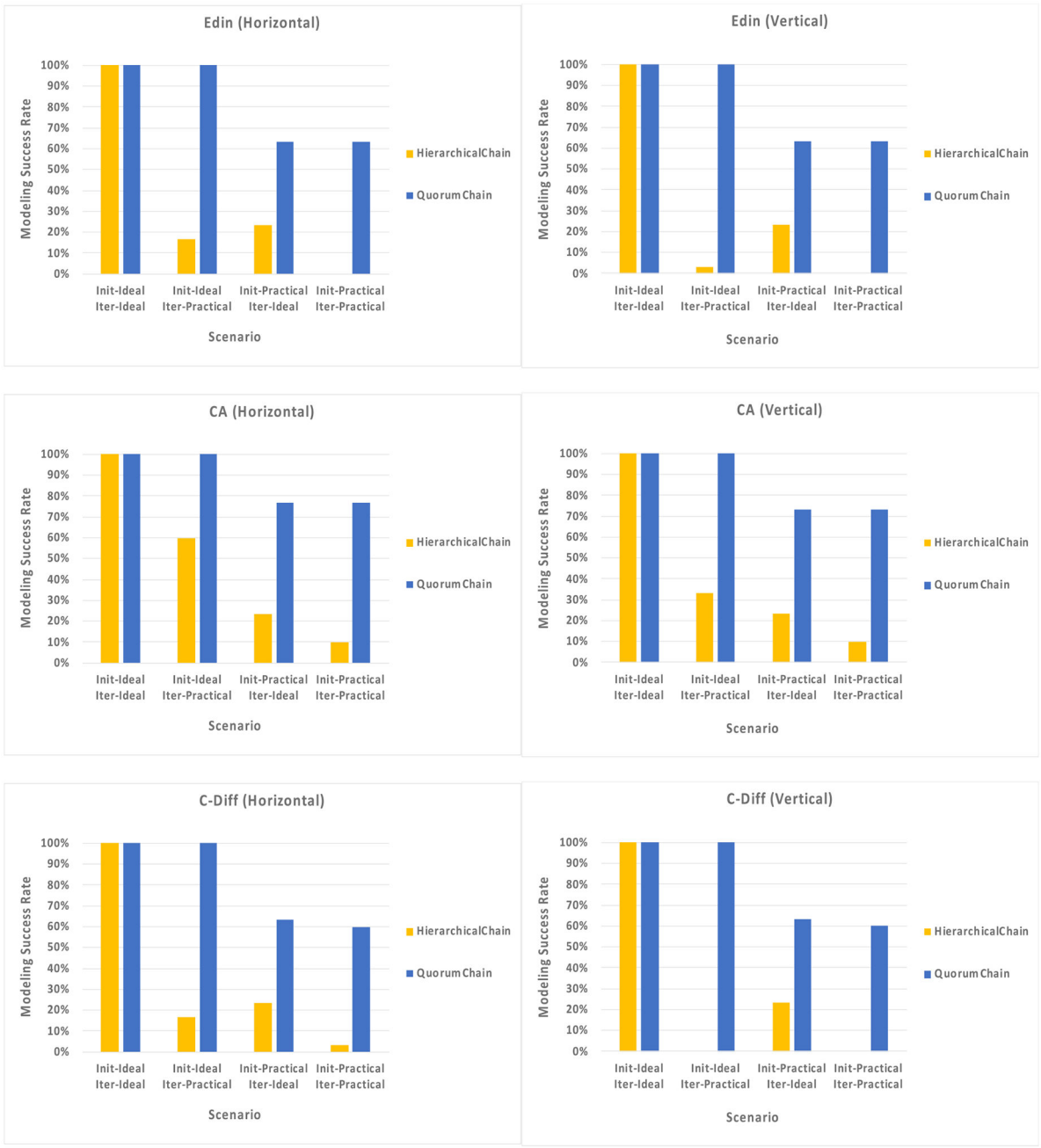

Fig. 4.

The predictive correctness results, including three datasets (Edin, CA, and C-Diff) and two ensemble methods (Horizontal and Vertical). We compared QuorumChain with the state-of-the-art HierarchicalChain method [26] in four site-unavailability scenarios: “Init-Ideal/Iter-Ideal”, “Init-Ideal/Iter-Practical”, “Init-Practical/Iter-Ideal”, and “Init-Practical/Iter-Practical”. The predictive correctness results are measured in the averaged full Area Under the receiver operating characteristic Curve (AUC).[49–51].

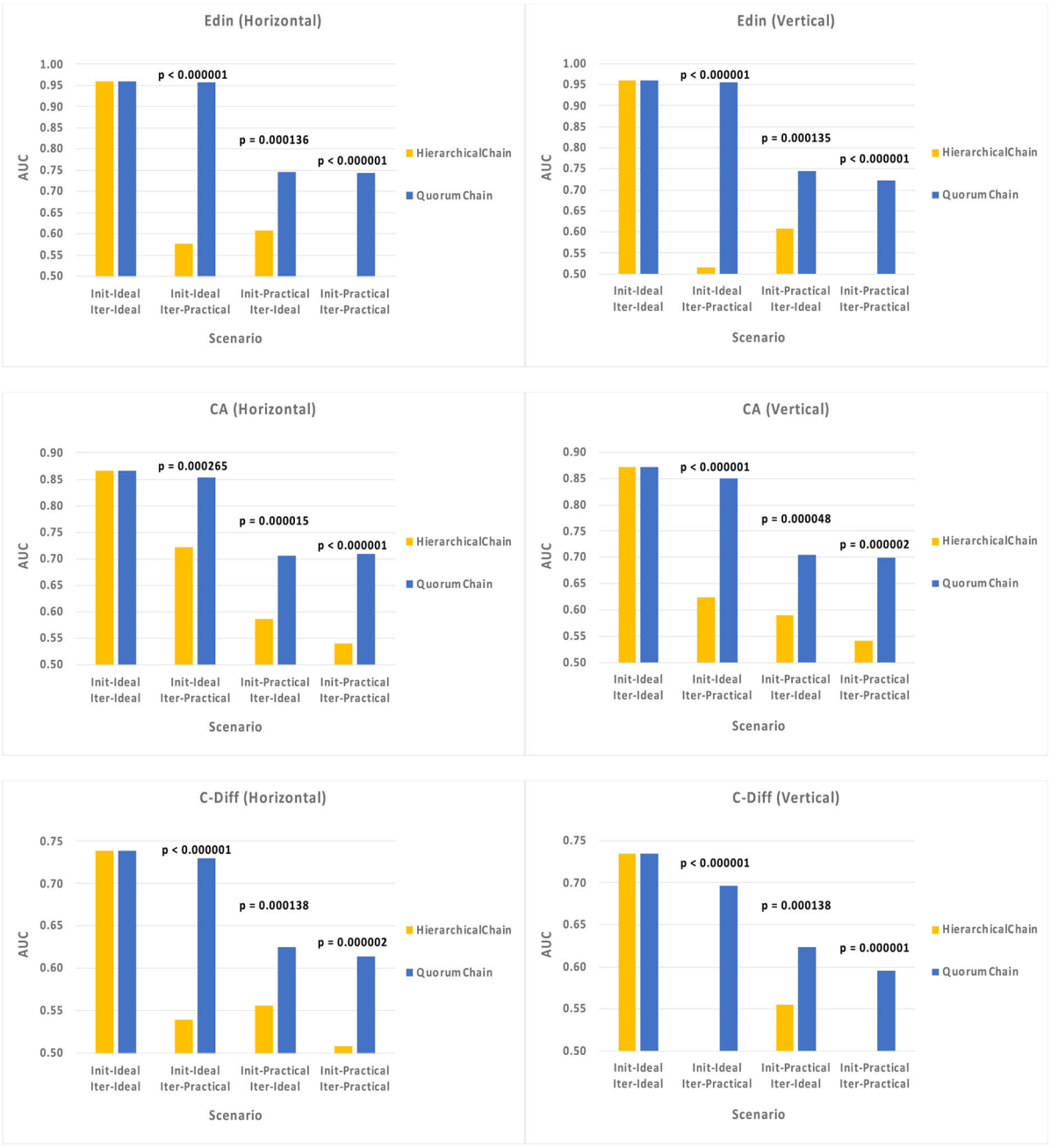

Fig. 5.

The modeling successfulness results. The modeling successfulness results are measured by the percentage of the trials that complete the model ensemble learning process out of the 30 trials in our experiment.

Table 1

Statistics and features of our test datasets, including their class distributions (i.e., the percentage of the positive/negative classes). The categorical covariates are converted into binary through dummy coding, and the numerical covariates are labeled with an asterisk symbol ("*"). The information for the Myocardial Infarction (Edin) and Cancer Biomarker (CA) datasets are adapted from prior studies.[722] For the Clostridium Difficile Infection (C-Diff) dataset, "Med" represents the number of medications.

Dataset	**Myocardial Infarction (Edin)**	**Cancer Biomarker (CA)**	**Clostridium Difficile Infection (C-Diff)**	
# of Covariates	9	2	25	
# of Samples	1,253	141	157,493	
Class Distribution	0.219 / 0.781	0.638 / 0.362	0.010 / 0.990	
Outcome	Presence of Disease	Presence of Cancer	Presence of Positive Lab Test	
Covariates	Pain in Right Arm	CA-19*	Gender - Female	Med - Analgesics & Anesthesia *
	Nausea	CA-125*	Gender - Unknown	Med - anti-Infective Agents *
	Hypo Perfusion	–	Race - American-Indian or Alaska-Native	Med - Antineoplastic Agents *
	ST Elevation	–	Race - Asian	Med - Biological *
	New Q Waves	–	Race - Black or African-American	Med - Cardiovascular Agents *
	ST Depression	–	Race - Native-Hawaiian or Other-Pacific-Islander	Med - Central Nervous System Drugs *
	T Wave Inversion	–	Race - Other-Race or Mixed-Race	Med - Endocrine & Metabolic Drugs *
	Sweating	–	Race - Undeclared	Med - Gastrointestinal Agents *
	Pain in Left Arm	–	Age - ≥ 65 years old	Med - Genitourinary Products *
	–	–	–	Med - Hematological Agents *
	–	–	–	Med - Miscellaneous Products *
	–	–	–	Med - Neuromuscular Drugs *
	–	–	–	Med - Nutritional Products *
	–	–	–	Med - Respiratory Agents *
	–	–	–	Med - Topical Products *
	–	–	–	Med - Others *

Table 2

The learning iteration and execution time results. QuorumChain is equivalent to HierarchicalChain in the perfect "Init-Ideal/Iter-Ideal" scenario (i.e., all sites are always available in both initialization and iteration phases). All times were measured in minutes.

Dataset	Scenario	Learning Iteration		Execution Time	
		Mean	Standard Deviation	Total	Per-Iteration
Edin	Init-Ideal / Iter-Ideal (equivalent to *HierarchicalChain*)	56.38	14.37	108.11	1.92
	Init-Ideal / Iter-Practical	43.37	12.63	99.99	2.31
	Init-Practical / Iter-Ideal	25.88	23.58	67.71	2.62
	Init-Practical / Iter-Practical	19.83	17.94	63.96	3.23
CA	Init-Ideal / Iter-Ideal (equivalent to *HierarchicalChain*)	20.60	12.60	69.77	3.39
	Init-Ideal / Iter-Practical	15.84	7.55	62.15	3.92
	Init-Practical / Iter-Ideal	11.61	11.72	56.32	4.85
	Init-Practical / Iter-Practical	8.93	9.87	49.41	5.54
C-Diff	Init-Ideal / Iter-Ideal (equivalent to *HierarchicalChain*)	86.51	2.69	214.67	2.48
	Init-Ideal / Iter-Practical	57.93	11.91	207.31	3.58
	Init-Practical / Iter-Ideal	43.38	35.51	144.74	3.34
	Init-Practical / Iter-Practical	29.53	25.06	127.54	4.32

Table 3

Comparison of QuorumChain with existing studies. Compared to its centralized counterpart, QuorumChain provides exact correctness when all sites are available during the learning process, with an added capability to continue learning in the site-unavailability scenarios (labeled with "*", an asterisk symbol), if a quorum can be formed.

Method	Blockchain Platform	Network		Learning		Immutability	
		Fully Decentralized	Hierarchical Topology	Exact Correctness	Fault Tolerance	Data	Code
GloreChain [20]	MultiChain	Yes	–	Yes	–	Yes	–
HierarchicalChain [26]	MultiChain	Yes	Yes	Yes	–	Yes	–
ExplorerChain [2223]	MultiChain	Yes	–	–	Yes	Yes	–
DML [21]	Hyperledger	Yes	–	–	Yes	Yes	Yes
LearningChain [19]	Ethereum	Yes	–	–	Yes	Yes	Yes
QuorumChain	Ethereum	Yes	Yes	Yes *	Yes *	Yes	Yes