www.impactjournals.com/oncotarget/ Oncotarget, 2018, Vol. 9, (No. 5), pp: 5665-5690

Research Paper

Converging blockchain and next-generation artificial intelligence technologies to decentralize and accelerate biomedical research and healthcare

Polina Mamoshina[1,2], Lucy Ojomoko[1], Yury Yanovich[3], Alex Ostrovski[3], Alex Botezatu[3], Pavel Prikhodko[3], Eugene Izumchenko[4], Alexander Aliper[1], Konstantin Romantsov[1], Alexander Zhebrak[1], Iraneus Obioma Ogu[5] and Alex Zhavoronkov[1,6]

[1] Pharmaceutical Artificial Intelligence Department, Insilico Medicine, Inc., Emerging Technology Centers, Johns Hopkins University at Eastern, Baltimore, Maryland, USA

[2] Department of Computer Science, University of Oxford, Oxford, United Kingdom

[3] The Bitfury Group, Amsterdam, Netherlands

[4] Department of Otolaryngology-Head & Neck Surgery, Johns Hopkins University School of Medicine, Baltimore, MD, USA

[5] Africa Blockchain Artificial Intelligence for Healthcare Initiative, Insilico Medicine, Inc, Abuja, Nigeria

[6] The Biogerontology Research Foundation, London, United Kingdom

Correspondence to: *Alex Zhavoronkov*, ***email***: *alex@insilico.com*

Keywords: artificial intelligence; deep learning; data management; blockchain; digital health

Received: *October 19, 2017* ***Accepted***: *November 02, 2017* ***Published***: *November 09, 2017*

ABSTRACT

The increased availability of data and recent advancements in artificial intelligence present the unprecedented opportunities in healthcare and major challenges for the patients, developers, providers and regulators. The novel deep learning and transfer learning techniques are turning any data about the person into medical data transforming simple facial pictures and videos into powerful sources of data for predictive analytics. Presently, the patients do not have control over the access privileges to their medical records and remain unaware of the true value of the data they have. In this paper, we provide an overview of the next-generation artificial intelligence and blockchain technologies and present innovative solutions that may be used to accelerate the biomedical research and enable patients with new tools to control and profit from their personal data as well with the incentives to undergo constant health monitoring. We introduce new concepts to appraise and evaluate personal records, including the combination-, time- and relationship-value of the data. We also present a roadmap for a blockchain-enabled decentralized personal health data ecosystem to enable novel approaches for drug discovery, biomarker development, and preventative healthcare. A secure and transparent distributed personal data marketplace utilizing blockchain and deep learning technologies may be able to resolve the challenges faced by the regulators and return the control over personal data including medical records back to the individuals.

INTRODUCTION

The digital revolution in medicine produced a paradigm shift in the healthcare industry. One of the major benefits of the digital healthcare system and electronic medical records is the improved access to the healthcare records both for health professionals and patients. The success of initiatives that provides patients with the access to their electronic healthcare records, such as OpenNotes, suggests their potential to improve the quality and efficiency of medical care [1, 2]. At the same time, biomedical data is not limited to the clinical records created by physicians, the substantial amount of data is retrieved from biomedical imaging, laboratory testing such

as basic blood tests, and omics data. Notably, the amount of genomic data alone is projected to surpass the amount of data generated by other data-intensive fields such as social networks and online video-sharing platforms [3]. National healthcare programs such as the UK Biobank (supported by the National Health Service (NHS)) [4] or global programmes such as the LINCS consortium (http://www.lincsproject.org/) and the ENCODE project (https://www.encodeproject.org/), provide scientists with tens of thousands of high-quality samples. However, while increased data volume and complexity offers new exciting perspectives in healthcare industry development, it also introduces new challenges in data analysis and interpretation, and of course, privacy and security. Due to huge demand for the treatments and prevention of chronic diseases, mainly driven by aging of the population, there is a clear need for the new global integrative healthcare approaches [5]. Majority of the recent approaches to personalized medicine in oncology and other diseases relied on the various data types including the multiple types of genomic [6-10], transcriptomic [11-13], microRNA [14], proteomic [15], antigen [16], methylation [17], imaging [18, 19], metagenomic [20], mitochondrial [21], metabolic [22], physiological [23] and other data. And while several attempts were made to evaluate the clinical benefit of the different methods [24] and multiple data types were used for evaluating the health status of the individual patients [25] including the widely popularized "Snyderome" project [26], none of these approaches are truly integrative on the population scale and compare the predictive nature and value of the various data types in the context of biomedicine. Introduction of new technologies, such as an artificial intelligence and blockchain, may enhance and scale up the progress in health care sciences and lead to effective and cost-efficient healthcare ecosystems.

In this article we first review one of the recent achievements in next-generation artificial intelligence, deep learning, that holds the great promise as a biomedical research tool with many applications. We then discuss basic concepts of highly distributed storage systems (HDSS) as one of the advantageous solutions for data storage, introduce the open-source blockchain framework Exonum and review the application of blockchain for healthcare marketplace. For the first time we introduce *half-life period* of analysis significance, models of data value for single and group of users and cost of buying data in the context of biomedical applications. WHere we also present a blockchain-based platform for empowering patients to ensure that they havetake a control over their personal data, manage the access priviledges and to protect their data privacy, as well to allow patients to benefit from their data receiving a crypto tokesscurrency as a reward for their data or for healthy behaviorand to contribute to the overall biomedical progress. We speculate that such systems may be used by the governments on the national scale to increase participation of the general public in preventative medicine and even provide the universal basic income to the their citizens willing to participate in such programs that will greatly decrease the burden of disease on the healthcare systems. Finally, we cover important aspects of data quality control using the recent advances in deep learning and other machine learning methods.

ADVANCES IN ARTIFICIAL INTELLIGENCE

While the amount of health-associated data and the number of large scales global projects increases, integrative analysis of this data is proving to be problematic [27]. Even high-quality biomedical data is usually highly heterogeneous and complex and requires special approaches for preprocessing and analysis. Computational biology methods are routinely used in various fields of healthcare and are incorporated in pipelines of pharmaceutical companies. Machine learning techniques are among the leading and the most promising tools of computational analysis.

Increased computer processing power and algorithmic advances have led to the significant improvement in the field machine learning. Although machine learning methods are now routinely used in various research fields, including biomarker development and drug discovery [28-31], the machine learning techniques utilizing the Deep Neural Networks (DNNs) are able to capture high-level dependencies in the healthcare data [32]. The feedforward DNNs were recently successfully applied to prediction the various drug properties, such as pharmacological action [33, 34], and toxicity [35]. Biomarker development, a design or search for distinctive characteristics of healthy or pathological conditions, is another area where the application of DNNs has led to significant achievements. For example, an ensemble of neural networks was applied to predict age and sex of patients based on their common blood test profiles [36]. Convolutional neural networks (CNNs) were trained to classify cancer patients using immunohistochemistry of tumour tissues [37]. And in early 2017, first neural network based platform, called Arterys Cardio DL, was officially approved by US Food and Drug Administration (FDA) [38] and is currently used in the clinic.

While the DNNs are able to extract features from the data automatically and usually outperform the other machine learning approaches in feature extraction tasks, one of the good practices is to select a set of relevant features before training the deep model, especially when the dataset is comparatively small. Algorithms such as the principal component analysis or clustering methods are widely used in bioinformatics [39]. However, these first-choice approaches transform the data into a set of components and features that may be difficult hard to

interpret from the perspective of biology. Supervised knowledge-based approaches such as pathway or network analysis, on the other hand, provide an attractive alternative, allowing for reduction of a number of input elements and preserving the biological relevance at the same time, which is crucial for the claimed to be hard to interpret "black box" methods such as DNNs. For example, Aliper and colleagues used signaling pathway analysis to reduce the dimensionality of drug-induced gene expression profiles and to train a DNN based predictor of pharmacological properties of drugs [33]. Selected pathway activation scores were compared to expression changes of over 1000 most representative, landmark genes. DNN trained on the pathway scores outperformed DNN trained on the set of landmark genes and achieved the F1 score, the weighted average of precision and recall, of 0.701 for the three drug pharmacological classes. In addition, signaling pathway-based dimensionality reduction allowed for the more robust performance on the validation set, while classifiers trained on gene expression data demonstrated a significant decrease in predictive accuracy on the validation set compared to the training set performance.

There are many promising machine learning techniques in practice and in development including the upcoming capsule networks and recursive cortical networks and many advances are being made in symbolic learning and natural language processing. However, the recurrent neural networks, generative adversarial networks and transfer learning techniques are gaining popularity in the healthcare applications and can be applied to the blockchain-enabled personal data marketplaces.

Generative adversarial networks

Generative adversarial networks (GANs) are among the most promising recent developments in deep learning. GAN architecture was first introduced by Goodfellow et al. in 2014 [40] and already demonstrated compelling results in image and text generation. Similar concepts were applied for molecule generation by Kadurin and colleagues [41]. A dataset of molecules with the different tumor growth inhibition (TGI) activity was used to train an adversarial Autoencoder (AAE), which combines the properties of both the discriminator and the generator. The trained model then was used to generate the fingerprints of molecules with desired properties. Further analysis of the generated molecules showed that new molecular fingerprints are matched closely to already known highly effective anticancer drugs such as anthracyclines. As a continuation of this work, authors proposed an enhanced architecture that also included additional molecular parameters such as solubility and enabled the generation of more chemically diverse molecules [42]. New model clearly showed the improvement in the training and generation processes, suggesting a great potential in drug discovery.

Recurrent neural networks

Electronic health records contain the clinical history of patients and could be used to identify the individual risk of developing cardiovascular diseases, diabetes and other chronic conditions [43]. Recurrent neural networks (RNNs), which are naturally suited for sequence analysis, are one of the most promising tools for text or time-series analysis. And one of the most advantageous applications of RNNs in healthcare is electronic medical record analysis. Recently, RNNs were used to predict heart failure of patients based on clinical events in their records [44]. Models trained on 12 month period of clinical history and tested on 6 months demonstrated an Area Under the Curve (AUC) of 0.883 and outperformed shallow models. Interestingly, analysis of cases that were predicted incorrectly, showed that networks tend to predict heart failure based on a patient history of heart diseases, for example hypertension. At the same time, most of the false negative heart failure predictions are made for cases of acute heart failure with little or no symptoms. Along with cardiovascular disease risk prediction, RNNs were also applied to predict blood glucose level of Type I diabetic patients (up to one hour) using data from continuous glucose monitoring devices [45]. The proposed system operates fully automatically and could be integrated with blood glucose and insulin monitoring systems.

While mobile health is an attractive and promising field that emerged recently, another exciting area of RNNs application is human activity prediction based on data from wearable devices. For example, RNN model called DeepConvLSTM, a model that combine convolutional networks and recurrent networks with Long Short-Term Memory (LSTM) architecture was applied on recordings from on-body sensors to predict movements and gestures [46]. Those technologies hold the most potential in distance chronic disease monitoring such as Parkinson's [47] and cardiovascular diseases [48].

Transfer learning

Being exceptionally data hungry, most of deep learning algorithms require a lot of data to train and test the system. Many approaches have been proposed to address this problem, including transfer learning. Transfer learning focuses on translating information learned on one domain or larger dataset to another domain, smaller in size. Transfer learning techniques are commonly used in image recognition when the large data sets required to train the deep neural networks to achieve high accuracy are not available. The architecture of CNNs allows transferring fitted parameters of a trained neural network to another network. Biomedical image datasets are usually

limited by the size of samples, so larger non-biological image collections, such ImageNet, could be used to fine-tune a network first. A CNN pre-trained on the ImageNet was further trained on magnetic resonance images (MRIs) of heart to outline the organ structure [49]. With an average F1 score of 97.66%, the proposed model achieved state-of-the-art cardiac structure recognition. Similarly, CNNs fined-tuned on the ImageNet were applied for glioblastoma brain tumour prediction [50].

One and zero-shot learning

One and zero-shot learning are some of the transfer learning techniques that allow to deal with restricted datasets. Taking into account that real-world data is usually imbalanced, one shot learning is aimed to recognise new data points based on only a few examples in the training sets. Going further, zero-shot learning intents to recognise new object without seeing the examples of those instances in the training set. Both one and zero-shot learning are concepts of the transfer learning.

Medical chemistry is one of the fields where data is scarce, therefore, to address this problem Altae-Tran and colleagues proposed a one-shot learning approach for the prediction of molecule toxic potential [51]. In this work, authors use a graph representation of molecules linked to the labels from Tox21 and SIDER databases to train and test models. One-shot networks as siamese networks, LSTMs with attention and novel Iterative Refinement LSTMs, were compared with each other, with graph convolutional neural networks and with random forest with 100 trees as a conventional model. Iterative Refinement LSTMs outperformed other models on most of the Tox21 assays and SIDER side effect. In addition, to evaluate the translational potential of the one-shot architecture, networks trained on Tox21 data were tested on SIDER, however none of the one-shot networks achieved any predictive power, highlighting the potential limitation in translation from toxic *in vitro* assays into the human clinic.

HIGHLY DISTRIBUTED STORAGE SYSTEMS

The recent explosion in generation and need for data has made it very necessary to find better systems for data storage. Among other requirements, the data storage systems should be better in terms of reliability, accessibility, scalability and affordability, all of which would translate into improved availability. While there could be many options for optimizing these requirements, HDSS has been found to be a very useful and viable option. Traditionally, a lot of technologies and techniques have been employed to store data since the development of computer systems, however, with the exponential increase in data demands and computing power, solutions like HDSS has become very important.

Basically, HDSS involves storing data in multiple nodes, which could simply be databases or host computers. Data stored in these nodes are usually replicated or redundant and HDSS makes a quick access to data over this large number of nodes possible. It is usually specifically used to refer to either a distributed database where users store information on a number of nodes, or a computer network in which users store information on a number of peer network nodes.

In recent years, storage failures have been one of the data handling challenges of higher importance, making reliability one of the important requirements for storage systems. HDSS, which allows data to be replicated in a number of different nodes or storage units and makes it protected from failures, has become very popular.

Advances in HDSS

There have been a significant amount of progress both in the applications and the optimization of HDSS. However, some of the key challenges in HDSS applications are ensuring consistency of data across various storage nodes and affordability of the systems. These challenges have been addressed by many recent HDSS solutions, including distributed non-relational databases and peer network node data stores. This is for example, a case of peer-to-peer node data store implemented in blockchain.

Blockchain could be described as a distributed database that is used to maintain a continuously growing list of records. These records are composed into blocks, which are locked together using certain cryptographic mechanisms to maintain consistency of the data. Normally a blockchain is maintained by a peer-to-peer network of users who collectively adhere to agreed rules (which are insured by the software) for accepting new blocks. Each record in the block contains a timestamp or signature and a link to a previous block in the chain. By design, blockchain is made to ensure immutability of the data. So once recorded, the data in any given block cannot be modified afterwards without the alteration of all subsequent blocks and the agreement of the members of the network. Because of its integrity and immutability, blockchain could be used as an open, distributed ledger and can record transactions between different parties or networked database systems in an efficient, verifiable and permanent manner. It is also flexible enough to allow adding arbitrary logic to process, validate and access the data, which is implemented via so called smart contracts (components of business logic shared and synchronized across all nodes). This makes blockchain very suitable for application in healthcare and other areas where data is very sensitive and strict regulations on how data can be used need to be imposed.

DATA PRIVACY ISSUES AND REGULATORY BARRIERS

Data privacy issues

While data could be said to be the lifeblood of the current digital society, many are yet fully to grasp the need for appropriate acquisition and processing of data [52, 53]. Among the key concerns in the generation and use of data are privacy issues. This is even more important in healthcare, where a high percentage of personal health data generated could be considered private. In order to ensure propriety in the handling of data, there have been regulations and rules that guide processes such as generation, use, transfer, access and exchange of data. Although privacy has been recognized as a fundamental human right by the United Nations in the Universal Declaration of Human Rights at the 1948 United Nations General Assembly, there is yet to be universal agreement on what constitutes privacy [54]. As a result, privacy issues and regulatory concerns have often been topics of important but yet varied interpretations wherever data is generated and used.

Regulatory barriers

With the dawn of computing and constant advancements in tech, there have been massive amounts of data generated on daily basis, and a substantial amount of these data consists of information which could be considered private. Some regulatory efforts to ensure proper flow and use of these data could become barriers to meaningful development [52]. Among the key efforts to ensure that data is used within the appropriate standards, is establishment of the Health Insurance Portability and Accountability Act (HIPAA) of 1996 and Privacy Rule's minimum necessary standard [55]. While developers and researchers are usually keen to get down to work; analyzing, processing and using data, some barriers could make getting and using relevant data challenging [55-58]. While regulatory barriers like HIPAA are necessary to ensure appropriate use of information, they could delay developmental efforts, especially when meaningful work have to be done as fast as possible. For instance, HIPAA requires an institutional review board to approve the use of data, and this could simply introduce some degree of complexity to data use [57].

Most people believe that their medical and other health information is private and should be protected, and patients usually want to know how this information is being handled [59]. The transfer of medical records from paper to electronic formats could increase the chances of individuals accessing, using, or disclosing sensitive personal health data. Although healthcare providers and public health practitioners in the US traditionally protect individual privacy, previous legal protections at the federal, tribal, state, and local levels could be inconsistent and inadequate. Hence, the HIPAA was established to ensure health insurance coverage after leaving an employer, and also to provide standards for facilitating healthcare-related electronic transactions. With the aim of improving the effectiveness and efficiency of the healthcare system, HIPAA introduced administrative simplification provisions that required Department of Health and Human Services to adopt national standards for electronic healthcare transactions [60, 61]. Meanwhile, Congress realized that developments and advancements in computing and electronic technology could affect the privacy of health information. As a result, Congress added into HIPAA provisions that made the adoption of federal privacy protections for certain individually identifiable health information compulsory.

The HIPAA Privacy Rule (Standards for Privacy of Individually Identifiable Health Information) provides national standards for protecting the privacy of health information. Essentially, the Privacy Rule regulates how certain entities, also called covered entities, use and disclose individually identifiable health information, called protected health information (PHI). PHI is individually identifiable health information that is transmitted or maintained in any form or medium (e.g., electronic, paper, or oral), but excludes certain educational records and employment records [60, 62]. Among other provisions, the Privacy Rule:

1. gives patients more control over their health information;
2. sets boundaries on the use and release of health records;
3. establishes appropriate safeguards that the majority of health-care providers and others must achieve to protect the privacy of health information;
4. holds violators accountable with civil and criminal penalties that can be imposed if they violate patients' privacy rights;
5. strikes a balance when public health responsibilities support disclosure of certain forms of data;
6. enables patients to make informed choices based on how individual health information may be used;
7. enables patients to find out how their information may be used and what disclosures of their information have been made;
8. generally limits release of information to the minimum reasonably needed for the purpose of the disclosure;
9. generally gives patients the right to obtain a copy of their own health records and request corrections; and

10. empowers individuals to control certain uses and disclosures of their health information.

It is absolutely important to maintain the privacy and security of health data, and Regulatory barriers serve to ensure rightful handling and use of these sensitive information. However, the complexity and difficulty introduced by these barriers could hamper meaningful progress in the use of data [57, 59]. There is therefore the need to develop systems and procedures that would not only ensure the appropriate handling and use of data but that would also significantly facilitate the use of data for meaningful progress towards better health outcomes.

ADVANCES IN BLOCKCHAIN

The blockchain is a distributed database using state machine replication, with atomic changes to the database referred to as transactions grouped into blocks, with the integrity and tamper-resistance of the transaction log assured via hash links among blocks. The blockchain concept was introduced for Bitcoin in the context of decentralized electronic currency [63]. Blockchain is usually understood to be decentralized, jointly maintained by a plurality of independent parties (*maintainers*), with the security assumptions postulating that a certain fraction of these parties may be non-responsive or compromised at any moment during blockchain operation like Byzantine fault tolerance [64].

Here we briefly describe the key features of the public and private [65, 66] blockchain technology:

● Linked timestamping [67]: blockchain by design makes it possible to provide a universally verifiable proof of existence or absence of certain data or a state transition in the blockchain database. These proofs would be computationally unforgeable by third parties (i.e., anyone but a collusion of a supermajority of the blockchain maintainers), provided that underlying cryptographic primitives (hash functions and signature schemes) are computationally secure. Furthermore, accountability measures (e.g., proof of work or anchoring [68]) could make it prohibitively costly to forge such proofs for *anyone*, including the maintainers themselves, and provide long-term non-repudiation. Such proofs for small parts of stored data could be compact and do not need to reveal any other explicit information (only mathematically impersonal information).

● Blockchain uses a consensus algorithm [64, 69], which guarantees that non-compromised database copies have the same views as to the database state. In other words, consensus ensures that transactions in the log are eventually propagated to all non-compromised nodes and lead to the identical changes.

● Applied cryptography routines (e.g., public-key digital signatures [70]) are used to decentralize authentication and authorization of transactions taking place within the network. That is, transactions are created externally to the blockchain nodes, which limits the repercussions of a node compromise.

The blockchain users are commonly divided into three parts according to their roles:

● Maintainers of the blockchain infrastructure, who decide business logic on the blockchain. The maintainers store full replica of the entire blockchain data, thus have full read access to it and decide on the rules of transaction processing, and are active participants of the consensus algorithm on the blockchain, in other words, have write access to the blockchain. Importantly, the maintainers are bound with a formal or informal contract with the other users as to the business logic encoded in the blockchain. That is, the maintainers cannot set or change the transaction processing rules arbitrarily; indeed, they provide means for external users to audit the blockchain operation for correspondence to these rules.

● External auditors of the blockchain operation for example regulators, non-government organizations, law enforcement, who verify the correctness of the whole transaction processing in real time and/or retrospectively. Auditors are assumed to store replica of the entire blockchain data, or at least a logically complete portion of it, and read access to it to be able to perform complete audits. From the technical perspective, auditors do not participate actively in consensus, but otherwise are similar to maintainers in that they replicate the entire transaction log.

● Clients who are the end users of the services provided by maintainers. Each client may have access to a relatively small portion of blockchain data, but his/her software may utilize cryptographic proofs to verify, with reasonable accuracy, the authenticity of the blockchain data provided by maintainers and auditors.

For example, in Bitcoin, maintainers correspond to miners and mining pool software, auditors to non-mining full nodes, and clients correspond to simplified payment verification (SPV) wallets and, more generally, to client-side key management software. Generalizing Bitcoin network taxonomy, we will refer to nodes having read access to the entire blockchain as full nodes, which are subdivided into validator nodes and auditing nodes as per the roles described above; the software on the client side will be accordingly called client software.

By utilizing cryptographic accountability and auditability measures, blockchains could minimize trust and associated counterparty risk among participants in the system [71]:

● As transactions are cryptographically authorized by the logical originators of such transactions, blockchain eliminates the risks associated with the single point of failure posed by centralized authorization systems. Key management could be complemented with public key infrastructure that would tie authorization keys with real-world identities, if deemed necessary.

● Client-side data validation could allow reducing

the risks associated with man-in-the-middle attacks, including those when MitM is performed on the server side (e.g., by compromising the user-facing backend of the system). The client-side validation could further utilize secure user interfaces and key management (e.g., TEE capabilities in modern mobile platforms).

- The universality of cryptographic proofs provided to clients could allow to reliably convey them to third parties (e.g., use electronic receipts provided by a blockchain managing supply chain, for tax accounting purposes or as evidence in legal action). Furthermore, cryptographic soundness of proofs allows to definitively restore the blockchain state even in the case when the maintainers are entirely compromised.
- The availability of real-time and retrospective authorization tools with guarantees of data authenticity could reduce costs of auditing and monitoring processes. This, in turn, could allow counterparties to more accurately assess the contract risks, and/or allow regulators to more precisely estimate systemic risks.

Blockchains could be categorized by the level of access to the blockchain data [66]:

- In public permissionless blockchains, all blockchain data is public. Furthermore, the consensus algorithm is censorship-resistant (e.g., proof of work used in Bitcoin), which ensures that maintainers are free to enter and leave the system; i.e., write access to the blockchain is public, too. The maintainers' accountability in permissionless blockchains is achieved via economic means (e.g., prohibitively high cost of attacks in proof-of-work consensus).
- Private blockchains have a well-defined and restricted list of entities having read and write access to the blockchain (e.g., a group of banks, the regulator and law enforcement in a hypothetical banking blockchain). Notably, end users of services codified in the blockchain (i.e., bank clients in the example above) do not have any access to the blockchain data.
- Public permissioned blockchains restrict write access to the blockchain data similarly to private blockchains, but are engineered to be universally auditable and thus oriented for wide read access by end users. For the sake of brevity, in the following statement, we will refer to this kind of blockchains as permissioned.

EXONUM FRAMEWORK FOR BLOCKCHAIN PROJECTS

Exonum (https://exonum.com, from Latin *exonumia*, numismatic items other than coins and paper money) is an open-source blockchain framework oriented towards permissioned blockchain applications with wide read access to blockchain data.

Exonum employs service-oriented architecture (SOA) [72] and architecturally consists of three parts: services, clients, and middleware.

- Services are the main extensibility point of the framework, which encapsulate business logic of blockchain applications. An Exonum-powered blockchain may have a plurality of services; the same service could be deployed on a plurality of blockchains (possibly with prior configuration). Services have a degree of autonomy in that each service is intended to implement logically complete and minimum necessary functionality for solving a particular task; their interface allows reuse and composability. In blockchain terms, services implement endpoints for processing *transactions* (cf. POST and PUT requests for HTTP REST services), as well as for *read requests* (cf. GET endpoints for HTTP REST services) that retrieve persistent information from the blockchain state (for the definition of blockchain state, see below).
- (Lightweight) clients implement typical functionality of clients in SOA; they are intended to be the originators of most transactions and read requests in the system, and are correspondingly supplied with cryptographic key management utilities, as well as tools to form transactions and verify (including cryptographically) responses to read requests.
- Middleware provides ordering and atomicity of transactions, interoperability among services and clients, replication of services among nodes in the network (which is purposed for both service fault-tolerance and auditability via auditing nodes), management of service lifecycle (e.g., service deployment), data persistence, access control, assistance with generating responses to read requests, etc. That is, middleware reduces the complexity of the system from the point of view of service developers.

The main advantages of Exonum for the described application compared to alternative permissioned frameworks are as follows:

- Because of design of data storage structures for auditability, Exonum could make it easier for clients and auditors (incl. ones with incomplete read access to data) to audit the system both in real time (incl. intermittently) and retrospectively. Further, the list of auditors could be unknown beforehand, and could be scaled over the course of blockchain operation.
- Due to use of service-oriented architecture, the application could easily reuse services developed for other Exonum applications, add and reconfigure services utilized for the application, etc. The service orientation and direct use of common transports (such as REST + JSON) could allow to streamline integration of third-party applications into the ecosystem provided by the Marketplace. Furthermore, service orientation could theoretically provide costless interoperability with other Exonum-based blockchains. (Albeit this possibility is not currently realized by the Exonum framework, the middleware layer could largely alleviate interoperability efforts needed to be pursued by service developers.)
- Compared to permissionless blockchains and frameworks with domain specific language/virtual

machine indirection, Exonum provides substantially higher throughput capacity (order of 1,000 transactions per second), and ability to encode complex transactional logic, incl. interaction with external components.

● Exonum uses pessimistic security assumptions as to the validator node operation. The consensus algorithm employed in Exonum does not introduce single points of failure (e.g., dedicated orchestration / transaction ordering nodes). Furthermore, the set of validator nodes is reconfigurable, allowing to scale the security by adding new validators, rotating keys for validators, locking out compromised validators, etc.

Blockchain storage

Blockchain state in Exonum is a persistent key-value storage (KVS), where keys and values are, in most general case, byte sequences of an arbitrary length, with the defined operations:

● Put a value under a specified key (creating the key if necessary)

● Remove a key-value pair by the key

● Iterate over keys in the lexicographic order, including starting from a specific key.

Exonum allows to split the key space of the common KVS into the hierarchy of typed collections: lists, sets and maps, whereas items of the collections (or key-value pairs in the case of maps) are binary-serializable as per the Exonum serialization format. Operations on these collections are mapped to the corresponding operations of the underlying KVS. The two uppermost levels of hierarchy correspond to services and data collections within a specific service, respectively; i.e., the 2nd level of hierarchy are items of top-level service collections. Additional levels of hierarchy could be created by using collections as items of top-level collections.

Collections can be declared as *Merkelized*. Merkelized collections introduce a new operation, *hash* of the collection, which is the hash commitment to all its items (or key-value pairs in the case of map). This construction allows to create compact (logarithmic wrt the number of elements in the collection) cryptographic proofs of presence (and absence in the case of sets/maps) of items in the collection.

Similar to the hierarchical structure of collections within the blockchain described above, all Merkelized collections of a particular service could be committed to in a single hash digest (possibly, through one or more levels of indirection). Indeed, this hash digest could be calculated by creating a Merkelized meta-map of collection identifiers into collection hashes. Similarly, commitments of all services within the blockchain could be collected into a single blockchain-level hash digest, which would

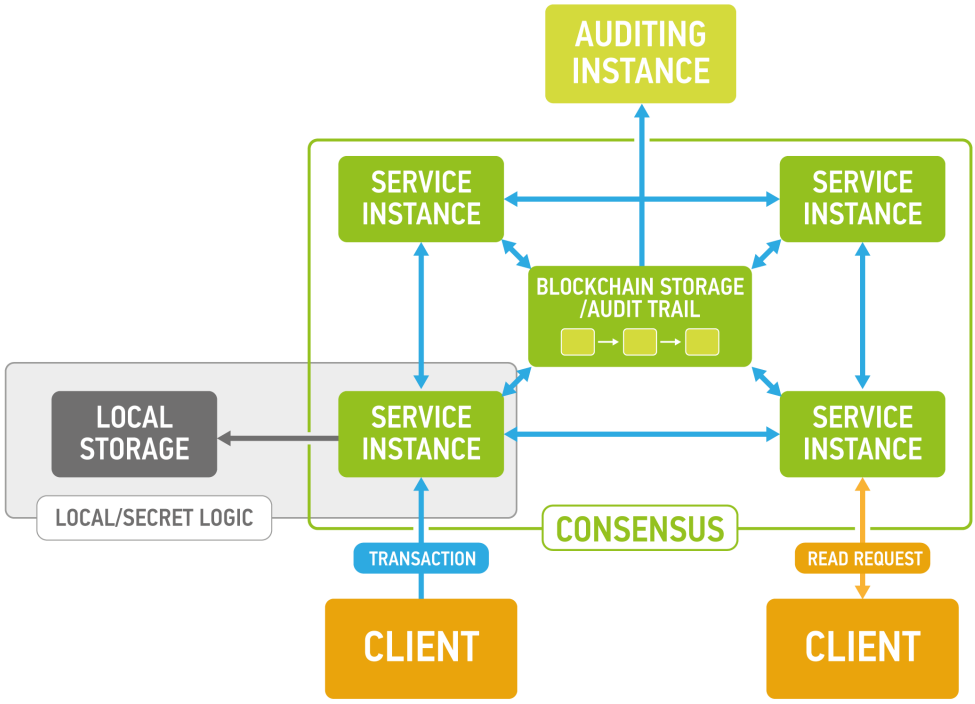

Figure 1: Exonum service design (each Service instance and Auditing instance has local replica of blockchain storage to ensure authenticity of the data and balance load).

Table 1: Distinguishing characteristics of Exonum service endpoints

Characteristics	Transactions	Read requests
Localness	Global (subject to consensus)	Local
Processing	Asynchronous	Synchronous
Initiation	Client	Client
REST service analogy	POST / PUT HTTP requests	GET HTTP requests
Example on the cryptocurrency service	Cryptocurrency transfer	Balance retrieval

commit to *all* data in all Merkelized collections within the blockchain state. For all intents and purposes, the resulting blockchain-level hash digest is the hash (commitment) of the entire blockchain state. This would allow to create proofs of existence or absence tied to this single hash as a root of trust.

In order to reduce risks of history revisions and equivocation, H_state may be anchored on a permissionless blockchain with strong accountability guarantees (e.g., Bitcoin), and proofs provided to clients augmented accordingly. Cf. notion of partial proofs in the OpenTimestamps protocol (https://opentimestamps.org/). Note that the anchoring scheme would allow to reliably assert statements about the blockchain state retrospectively, even if the blockchain itself has become unavailable (e.g., due to wide-scale compromise or collusion of the blockchain validators).

Network

Services may communicate with external world via 2 kinds of interactions:

- Transactions is the only way to change the blockchain state. Transactions are executed asynchronously, with their ordering and results of execution being subject of the consensus algorithm executed on the blockchain. For this reason, incoming transactions are broadcast among full nodes in the network
- Read requests allow to retrieve information from the blockchain state, which may be accompanied by the corresponding proofs of existence/absence. Read requests can be processed locally by any full node (or, more generally, by any node having sufficient read access to the relevant keyspaces of the blockchain state)

Transport layer

Due to universal verifiability of transactions and proofs, clients may connect to a single node for all requests. Note that maliciously acting node cannot forge proofs for read requests; but it could delay transaction processing by not broadcasting transactions received from the client. The transport protocol is not intended to be pinned by the Exonum specification; indeed, similarly to web services in frameworks such as Java EE and CORBA, the middleware layer is tasked with the responsibility to abstract transport layer functionality from service developers, so that invocation of service endpoints could be mapped to local method invocations. As of Exonum 0.2, RESTful JSON transport is supported for interaction of full Exonum nodes with clients, and TCP with a custom binary format is used in communication among full nodes.

Authentication and authorization

Transactions are necessarily authenticated by their originators with the help of public-key digital signatures to ensure their integrity, as well as real-time and retrospective universal verifiability. Public key infrastructure (PKI) could be built on top to achieve more complete non-repudiation and/or finely grained access control if necessary.

As read requests are local, authentication/authorization for them could be transport-specific, achieved, e.g., with web signatures (esp. for read requests implemented with the HTTP GET method) or by authenticating the communication channel (e.g., via client-authenticated TLS or Noise protocol).

In order to additionally boost security, service endpoints could be declared as *private*. Private endpoints could be compared with administrative interfaces in Web services; they are intended to process and manage local storage associated with a particular full node. Separation

of private endpoints could simplify access control and decrease attack surface; e.g., if the HTTP transport is used, private endpoints are mapped to a separate listen address compared to other endpoints.

Lightweight client

Most generally, a lightweight client in Exonum is a client-side library providing capabilities to communicate with full nodes (i.e., invoke service endpoints and receive responses) and cryptographically verify responses. A client could be complemented with key management capabilities and persistence of responses from full nodes; the former could be used for authentication of requests, and the latter could assist in non-repudiation and verifying consistency among different responses (e.g., monotonically non-decreasing blockchain height).

Consensus

To order transactions in the transaction log and agree on the result of transaction execution, Exonum utilizes an authenticated, leader-based Byzantine fault-tolerant (BFT) [73] consensus algorithm. The Exonum network would continue operating even if up to 1/3 validators are hacked, compromised or switched off. Hence, there is no single point of failure in the network; the whole process of transaction processing is fully decentralized.

The consensus algorithm works under the assumption of unforgeable public-key digital signatures and a partially synchronous network. Under these conditions, the algorithm provides safety and liveness as defined in [74], with safety not depending on partial synchronicity. Similar to other partially synchronous BFT algorithms such as PBFT or Tendermint, the algorithm uses three kinds of consensus messages - block proposals, prevotes and pre-commits (see Consensus Section of the https://exonum.com/doc/ documentation), which are authenticated by digital signatures to enable transferring of messages among validators and to improve non-repudiation.

Compared to other leader-based BFT algorithms, the algorithm used in Exonum has the following distinguishing characteristics:

- Unbounded rounds: Voting rounds have a fixed start time, but do not have a definite end time. A round ends only when the next block is received or committed locally. This helps decrease delays when the network connection among validators is unstable.
- Work split: Block proposals include only transaction hashes; furthermore, transaction execution is delayed; transactions are applied only at the moment when a node receives enough prevotes for a proposal. Delayed transaction processing reduces the negative impact of malicious nodes on the system throughput and latency.
- Requests: Requests algorithm allows a validator to restore consensus-related information from other validators by utilizing the fact that all messages are digitally signed. This has a positive effect on system liveness.

The validator set is reconfigurable; validators could be added or removed by the agreement of the supermajority of existing validators. The same procedure could be used for key rotation for validators.

Bitcoin anchoring

Exonum uses a BFT Bitcoin anchoring algorithm, which is packaged as a separate service. The algorithm periodically outputs the hash digest of a recent block on an Exonum blockchain, which commits to the entire blockchain state and transaction history, in a transaction on the Bitcoin blockchain. The anchoring transaction has a well-defined form and must be authenticated by a supermajority of validators on the anchored Exonum blockchain. Validators should use individual Bitcoin full nodes to get information from the Bitcoin blockchain in order to eliminate single points of failure associated with potential eclipse attacks [btc-eclipse] on the nodes. Anchoring transactions form a sequence; each next anchoring transaction spends an output created by the previous one. Authentication and chaining of anchoring transactions make the described anchoring procedure similar in its security characteristics to the paper-based anchoring described in [75].

COMBINATION AND TIME VALUE OF DATA

Health care providers around the globe are tracking patient encounters through an electronic medical records system, generating terabytes of patient medical records. This giant amount of medical data is a gold mine of health information.

Each type of data (basic blood test, basic urine test, MRI, electroencephalogram, electrocardiogram, genome, transcriptome, microbiome etc.) and their combinations have relevant value, depending on quality of the medical records and its biological significance for certain disease condition (Figure 2). Different types of medical data have their own predictive value, representative sensitivity, prediction rate and weight. Patterns reflecting the changes in patient condition are more readable when doctor operates complex information, presenting patient health state on different levels at the current period of time. Some of the data types, like pictures, videos, voice can also have substantial predictive value for medical condition. For examples, several research groups already studied the application of voice and speech pattern recognition for diagnosis of Parkinson's disease and its severity prediction

[76, 77]

Traditional diagnostics pipeline based on analysis combination of medical tests, especially when healthcare specialists try to diagnose serious and complex pathologies such as oncological, autoimmune, or neurodegenerative diseases. The combination of data types, especially the sum of low diagnostics data, provides a multi-level overview and better understanding of complex multifactorial conditions, and also leads to a faster diagnostics [78, 79]. Search and identification of suitable groups of biomarkers based on the multi-level data remains an important challenge. Taking into account different mechanisms of the disease development, biomarkers can acquire various forms. It is a common trend that various types of medical tests are being used for substantially broader diagnostic applications than those that were available at the start of their implementation.

Despite a large number of various diagnostic tests not all type of medical data is reasonable to use for the description of patient's health state in dynamics. For example, genome analysis provides an important information on heredity, but due to it relative stability has a low value for prediction of dynamic changes in the organism compared to epigenome [80] or transcriptome [81]. Sampling time is an important component of any medical analysis, which allows to accurately describe the state of the human health at the moment. Following the principles of 360º health introduced by the NHS [82], the more different parameters are analyzed at the same time, the more detailed and voluminous the overall picture is. One-time combination of date provides a very nutritious feed for artificial intelligence, allowing to create powerful algorithms of effectively and precisely detection different human health states.

In this paper we introduce a formal model that allows to evaluate data value that takes into account combination and time parameters of data. Based on the value model one may establish a proper *cost* for using the given data combination, and this in turn allows for creation of formal medical data transaction model to create medical data marketplace.

Data value model

Generally, data can be divided into the following categories: dynamic - reflecting the state of the organism at the time of sampling (blood test, transcriptome, epigenome, proteome, microbiome etc), and static - almost unchanged during the life of the patient (genome, fingerprint). Within a dynamic group, it is possible to differentiate rapidly changing data and gradually changing data.

In congenital genetic diseases, the records obtained in the first years of life are important as determining the further development of the disease (Figure 3.1), for the age-associated diseases, it is important to analyze the results obtained before the diagnosis was made (Figure 3.2), data role constant throughout the life of the patient (Figure 3.3).

Each personal biomedical record *R* could be viewed as a triplet (*type, time, quality*) where *type* is a categorical

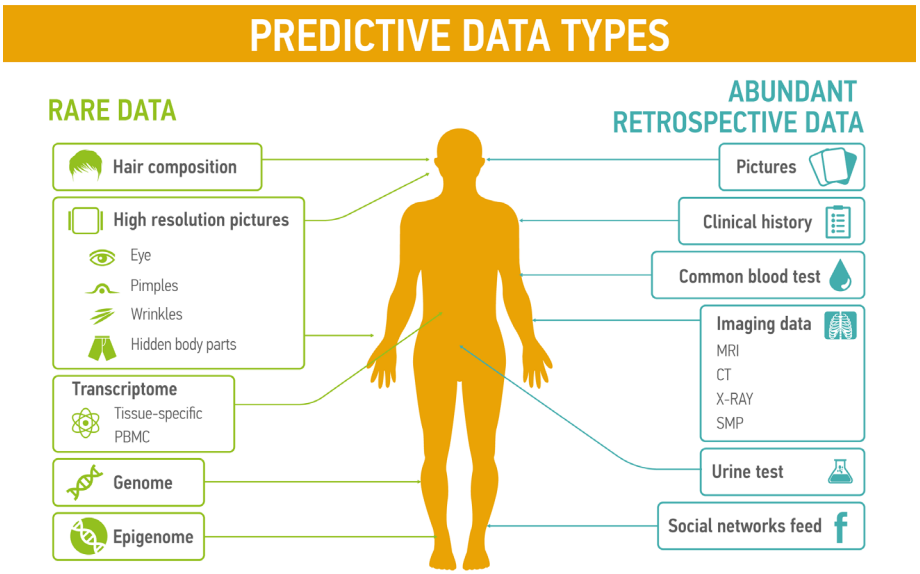

Figure 2: Predictive data types could be divided into two groups: rare data, such as the transcriptomic profiles, hair composition, or even novel data types that are not measured today and abundant retrospective data including common blood tests or the feed from social networks.

variable for a record type, *time* is a sampling time of patient's biomedical record (for example, when blood test was made) minus the patient's time of birth, *quality* is a nonnegative number reflecting record quality (generally, it could be a vector). One of the key attribute encapsulated in the *type* is the *half-life period* of analysis - characterizing the ½ duration of the relevance of the data. For example, according to one of largest medical practice and research center, Mayo Clinic, cholesterol check is valid for only five years or less if a patient at the higher risk of heart disease [83]. While, the genome profile is valid for the whole life of the patient, so genome analysis has longer *half-life period* than basic cholesterol blood test.

The dataset is a setDataset = $\{(\text{user}r_n, R_n)\}_{n=1}$ of N ($\text{user}r_n$, R_n) pairs, where *user* is a user profile. User profile *(Patient's profile)* - an attribute that includes information of ethnicity, date of birth, sex, diagnoses, blood type, medical prescriptions, vaccinations, chronic diseases, interventions, smoking and alcohol status, family relations, weight, height, geolocation. User profile refers to a hybrid attribute, since it includes both static (date of birth, ethnicity, sex, blood type) and dynamic parameters (diagnoses, smoking and alcohol status, weight, geolocation).

The dataset *Cost* is a function of a *Dataset* and it consists of two terms: the combination for a each single user and combinations for a set of same type records for a groups of users.

Cost for single user

$$Cost\ (user) = \sum_{k=1}^{\infty} \sum_{\{[i_1,\cdots,i_k]|1<m:i_1<i_m\ AND\ (user,R_{i_m})\in Dataset\}} f_k(R_{i_1},\cdots,R_{i_k}\ |user)$$

where k is a number of records in a combination, all records in the combination are for the *user* and are different, f_k is a cost function for a combination for k records.

k=1: R= (*type, time, quality*)

$f_1(R|user) = \Psi\ (type|user),\ \times\ quality\ \times\ \Psi\ (type|user)$

where

$\Psi\ (type|user)$ is a base value of given record type and user combintation. In the model we set it as a mapping of categorical parameter *type* to the positive numbers (*0, ∞*)

$\Psi\ (type|user)$ is a time value of record. It is a function $(0, \infty) \rightarrow (0, \infty)$

k>1: R_1= ($type_1$, $time_1$, $quality_1$), . . . R_K= ($type_K$, $time_K$, $quality_K$)

In case of combination of several records we keep the cost component structure similar to k = 1 case. This leads to the need to define base value, quality and time value for interaction component to the cost of several records.

$f_k(R_1, \ .\ .\ ., R_k|user) = \Psi_k\ (type_1,....,type_k|user)\ \times\ v_k(quality_1, \ .\ .\ .\ quality_k),\ \times\ \Psi_k\ (time_1, \ .\ .\ ., time_k, type_1,..., type_k|user)$

where

$\Psi_k\ (type_1,....,type_k|user)$ is a base value of addition due to interactions. It is a mapping of categorical

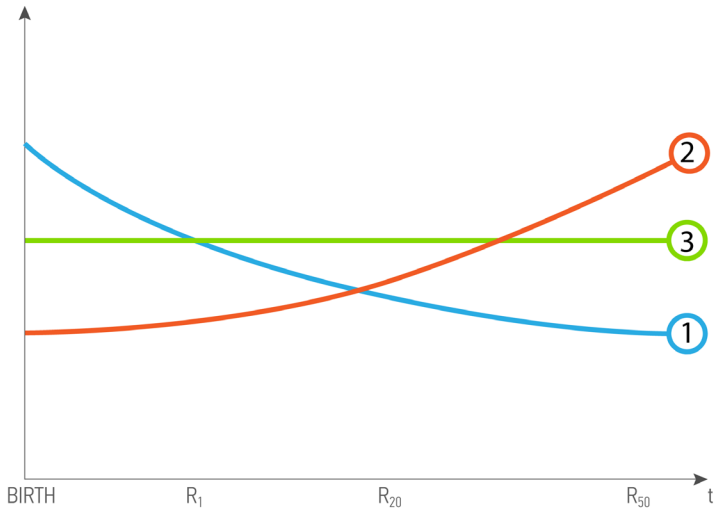

Figure 3: A possible scenario of data value dependence of age and health status of a patient. R- biomedical record, and index is patient's age. 1 - The curve of dependence for the R, when value decreases with time, and is most valuable in the young age 2 - The curve of dependence for the R, when value increases with time, and is most valuable in the old age. 3 - The curve of dependence for the R, when value is constant

parameters $type_1$,....,$type_k$ to the positive numbers $(0, \infty)$

$v_k(quality_1, \ldots quality_k)$ is a quality of combination of records for one user. It is a function $[0, \infty)^k \rightarrow [0, \infty)$ such that v_k is monotonic nondecreasing function of each input and $\lim_{quality, m \rightarrow 0} v_k(quality_1, \ldots quality_k) = 0$ for all $m = 1,...k$. The last property means that the adding a record R with zero *quality* does not change the cost of the *Dataset*. For example, $v_k = (\Sigma^k_{m=1} 1/ quality_m)^{-1}$

$\Psi_k(time_1, \ldots, time_k, type_1, \ldots, type_k | user)$ is a time value for a set of records. For a fixed set of $time_1, \ldots, time_k$ it is a function $[0, \infty)^k \rightarrow [0, \infty)$. For example, for each $type_m$ two time parameters T_m, $type_{m,o}$ and nonnegative function $w_m(t), t \geq 0$ is chosen. And

$\Psi_k(time_1, \ldots, time_k, type_1, \ldots, type_k | user) = max_t \, min_m = 1_{\ldots k} w_m\left(\frac{t - time_{m,0}}{T_m}\right)$

Figures 4 and 5 illustrate how the cost depending on records' time were done: for the combination of the same type (for example, blood tests made in different period of time) and for the combination of different types of data from the single patient (for example, blood test and transcriptome analysis). The greater intersection of time value curve is the greater combined records cost is. Data obtained it the same or short period of time have greater representation and predictive value, that's why we introduce term *time value of data*. *Time value of data* - an indicator that demonstrates the representative and predictive rate of the group value of data, based on the difference in the records' half-life time. It is relevant both: for a combination of data of one type and for a combination of data of different types.

Cost for records from group of users

The cost increases only for the multiple records of the same type from distinct users. Let us fix *type* of records to find their combination cost increase. . Let $user_{i1}$,....,$user_{ik}$ have records with type *type* in the *Dataset* Let $quality_{i1}, \ldots quality_{ik}$ be best corresponding qualities of users records with *type* in the *Dataset*. Then

$Cost(type, quality_{i1}, \ldots, quality_{ik}, user_{i1}, \ldots user_{ik}) = \gamma(K, type | user_{i_1}, \ldots, user_{i_k}) \times \frac{1}{K} \sum_{k=1}^{K} quality_{i_k}$,

where for a fixed *type* function $\gamma(k, type|user_{i1},...., user_{ik})$ has a fixed superlinear growth with k increase. For example, $\gamma(k, type|user_{i1},...., user_{ik}) = C \times K \times lnk$ or $\gamma(k, type|user_{i1},...., user_{ik}) = C \times K^{3/2}$

Every Dataset has it's own critical representative level, critical level depends on the type of data, their quality, and patient profile.

Cost of buying dataset

If customer wants to buy *Dataset* and already has bought some $Dataset^1$, then

$Cost(Dataset) = Cost(Dataset \; U \; Dataset^1) - Cost(Dataset^1)$

The payments for users data are also fairly distributed among users according to their contribution to the dataset cost and previous payments from the current customer. Thus, the application of the data value model

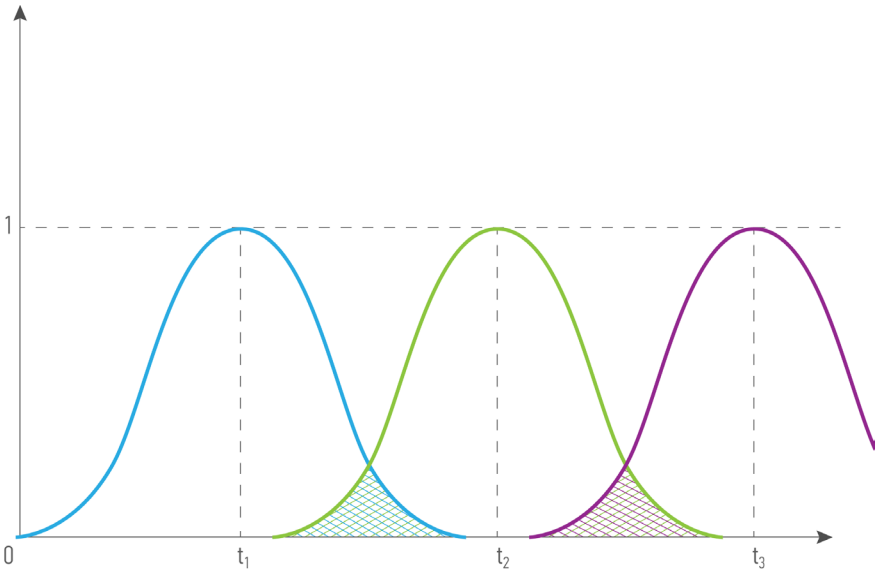

Figure 4: The cost of a combination of data (R_1, R_2, R_3) of the same type obtained in different periods of time from the single patient, where *type*1 = *type*2= *type*3, *quality*1 = *quality*2= *quality*3, *time*1 ≠ *time*2 ≠ $time_3$

makes it possible to convert *Big Data* into *Apprised Data.*

Family and relationship value of data

Many studies in healthcare require the data coming from the closely related subject from the same family or region and commonly involve Human data analysis is complicated redistricted experimental possibilities. However, these challenges could be overcome with a powerful and efficient design of data analysis. One of possible approaches is analysis of genetically close patients, twins, siblings or parents and offspring or colleagues and friends. , where observed effects are influenced by less number of features. Eventually,The biomedical data obtained for relatives is commonly more valuable than the same data obtained from the unrelated individuals.

Here we introduce a measure commonly used in the genealogy, the coefficient of relationship (r) between two individuals, also known as a coefficient of inbreeding, where a relationship between two subjects B and C is defined as

$r_{BC} = \Sigma p_{AB} p_{AC}$, where p is for path coefficients connecting B and C with common ancestor A.

and

where p_{AB} is defined as:

$p_{AB} = 2^{-n} \times \sqrt{\left(\frac{(1+f_A)}{(1+f_B)}\right)}$, where f_A and f_B are inbreeding coefficient for ancestor A and offspring B, respectively.

Given the fact, that humans population are genetically heterogeneous and usually , or random-bred, we could set thef_A=0 and a formula coefficient of relationship could be simplified:

$r_{BC} = \Sigma p\ 2^{-L(p)}$, where $L(p)$ is the length of the path p.

This way, r of parent-offspring is $2^{-1} = 0.5$, , and r of grandparent-grandchild is $2^{-2} = 0.25$.

And cost function of data could be modified as following:

$f_k(R_1, \ldots, R_k|user) = \Psi_k(type_1,\ldots,type_k|user) \times v_k(quality_1, \ldots quality_k), \times \Psi_k(time_1, \ldots, time_k, type_1,\ldots, type_k|user) + \lambda[\Psi_k(type_1,\ldots,type_k|user) \times v_k(quality_1, \ldots quality_k), \times \Psi_k(time_1, \ldots, time_k, type_1,\ldots, type_k|user)]$, where λ is a regularization coefficient equals to coefficient of relationship of users the system and could be set as following :

$\lambda \Sigma^m_{i=1} r_i$, where m is a number of users in the system and r is a coefficient of relationship between them.

For distant relatives, $r \rightarrow 0$ and almost will not contribute to the cost function of data, however, for a very close relative such as twins, r is equal to 1 and at the beginning, it will double the cost of data. The cost of data will grow with a number of close relatives that are using the platform and contributing their data.

PREDICTING PATIENT'S AGE TO EVALUATE THE PREDICTIVE VALUE OF DATA

Chronological age is a feature possessed by the every living organism and one of the most important factors affecting the morbidity and mortality in humans. The multitude of biomarkers linked to disease are strongly correlated with age. For instance, triglycerides, glycated hemoglobin (HbA1c), waist circumference, IL-6 increase

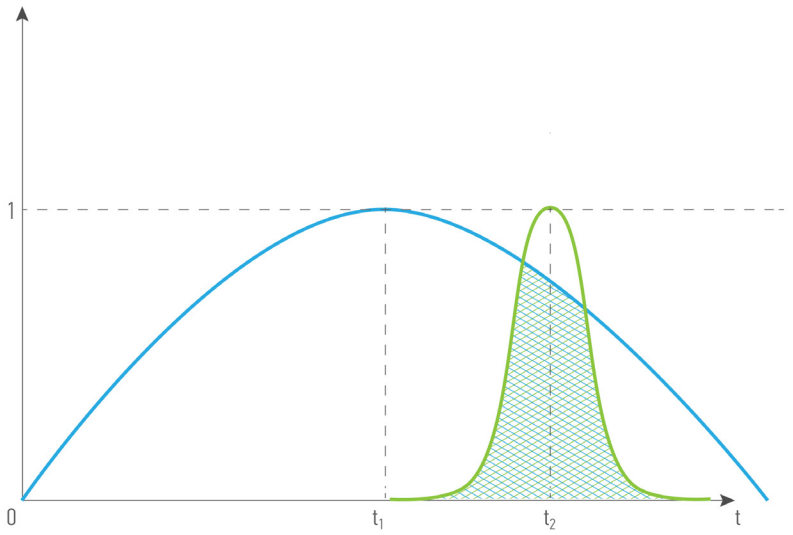

Figure 5: The cost of a combination of different data types (R_1, R_2) obtained in different periods of time from the single patient, where *type***1 ≠** *type***2,** *quality***1 =** *quality***2,** *time***1 ≠** *time***2** .

with age, but other parameters like albumin, IGF and creatinine clearance go in an opposite direction [84, 85]. Many efforts have been made to integrate biomarkers in various health/risk indexes like Healthy Aging Index [86, 87], Framingham Risk Score [88, 89], Frailty index [90, 91], Physiologic Index of comorbidities [92]. Ultimately, age is the closest estimate of a health status of a person. Hence, combining various biomarkers and linking them to age will provide the basis for platform able to provide integrative analysis of health status, assess data quality and even identify fake data. In addition, treating aging as a disease to train the deep neural networks to capture the most important biological properties of the age-related changes that transpire during aging using the deep neural networks facilitates for transfer learning on individual diseases using a much smaller number of samples. First proposed by Zhavoronkov et al in 2015 [93], this technique can be used to reconstruct the data sets with the missing or incorrect features.

Aging is also a continuous process gradually leading to loss of function and the age-associated diseases. The DNNs trained on the multi-modal data ranging from photographs, videos, blood tests, "omics", activity and even smell and sweat during aging capture the many biologically-relevant features about the group, individual, organ, tissue or even a set of molecules. These DNNs can be used to extract the features most implicated in aging and specific diseases to be used as targets or build association networks and causal graphs. These DNNs can also be re-trained on a much smaller number of data sets of specific diseases within the same data type or using the many types of biological data. Here we propose a high-level architecture fetaturing the various data types (Figure 6). First, for each data type we build a DNN predictor of chronological age for the reasonably healthy individuals. Individual DNNs will allow for the detection of outliers and data quality control. Then all individual DNNs will be used to train multi-modal one-shot learning DNN. This architecture allows not only for accurate age prediction, but also for feature importance analysis. Results of such analysis across all predictors will tell about the importance of each individual biomarker and may inform its relative 'cost'. Since many of the biomarkers related to age (Albumin, Glucose, Norepinephrine, WBC, Il-6, etc.) are measured routinely in the clinic in a separate tests of different degree of invasiveness it is important to know which ones are more predictive.

HEALTH DATA ON BLOCKCHAIN

One of the major problem for healthcare is data exchange and ability to use data in research and commercial projects. At the same time, healthcare sector requires to maintain a high standard of data privacy and

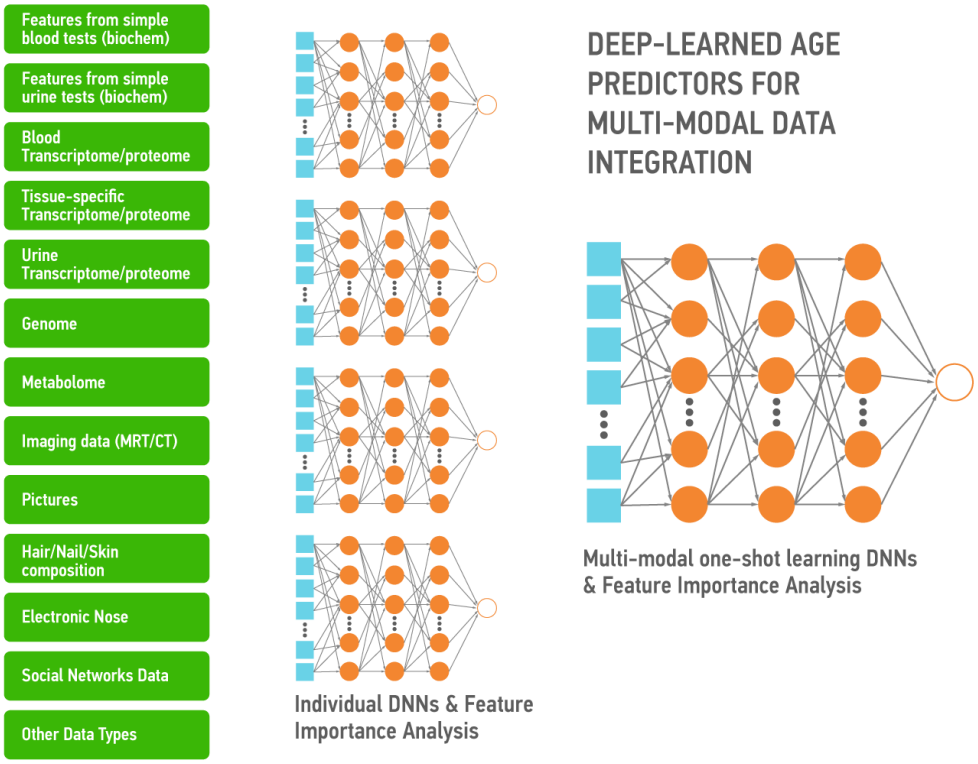

Figure 6: A simple depiction of the deep neural networks trained to predict the chronological age within the data type and using the features extracted using the feature importance and deep feature selection for multi-modal age predictors. These predictors may be used for data integration, verification and transfer learning.

security. Data breaches in healthcare storage systems can be especially costly because of HIPAA fines and reputation losses. Blockchain solutions as described later could reduce data breach risks by utilizing threshold encryption of data (meaning that cooperation of multiple parties is required to decrypt data), together with public key infrastructure (i.e., the use of asymmetric cryptography to authenticate communication with system participants). Gained a substantial attention in recent years T, the blockchain technology gained substantial popularity in recent years primarily due to the popularity of the Bitcoin crypto currency, was previously has been proposed as a medium for health care data storage solutions [94, 95] and as a tool for to improving thee transparency in clinical trials [96].

A blockchain-based system can dramatically simplify data acquisition process. They allow user to upload his data directly to the system and give his permission to use his data if it was bought through the system using transparent price formula determined by data value model. Also it would guarantee fair tracking of all data usage activity.

The promise of such solution is the opportunity for users to take ownership of their data and access priviledges and even allow them to sell their data directly to the consumers of data for the fair value of data. However, exchanging the data for currency may be problematic for many reasons including the need to perform a massive number of micro-transactions in multiple countries and among a large number of different types of the data market market participants.

Here we propose a new form of a utility crypto token called LifePound, which can be generated or mined by putting the data on the blockchain-enabled marketplace to facilitate for transactions and enable the novel incentive schemes.

The architecture of the proposed platform is described in Figure 7.

The clients of the marketplace and their goals are

- users: to store and sell their biomedical data and to receive advanced health reports from the results of data analysis
- customers: to buy data from users and to provide results of data analysis for users
- data validators: to check the data received from users

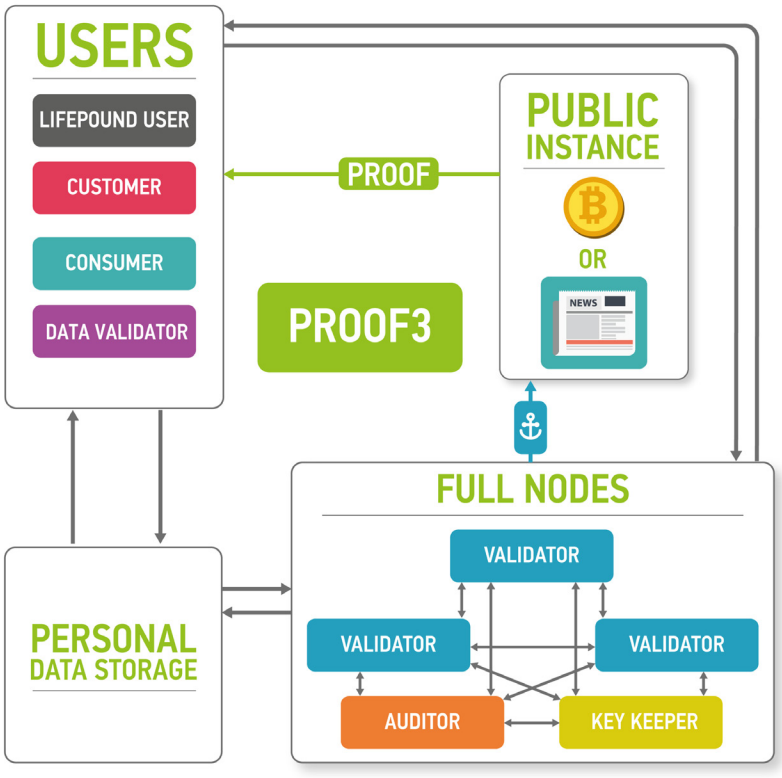

Figure 7: The marketplace ecosystem consisting of the four parts: blockchain part, data storage, users and public instances. The blockchain is used to process new blocks of transactions, store and send keys and audit itself. The data storage contains encrypted data. Users send and sell their data using the marketplace (users), validate data (data validators), buy personal medical data (customers) and use LifePound as a cryptocurrency (LifePound users). The system is not fully open, and the public instances are used in cryptographic proofs for users to guarantee the marketplaces functioning correctness.

● LifePound users: to use the cryptocurrency marketplace (possibly without any interaction with personal data).

Users are allowed to keep their data private and secured providing access to the data only for organizations whose paid for it and (optionally) staying as anonymous as possible. Customers intend to buy well-specified data samples which are aggregated from many users. To ensure the quality of the data provided by users, third party is needed - data validators, experts who are first buyers of the data. Data validators check the data quality and provide customers with a guarantee of user data validity. Interactions in the marketplace are registered on a blockchain in the form of transactions. Blockchain by itself does not contain any opened personal information. It contains hashes which could be used to timestamp and provide a reasonable level of non-repudiation for all actions at the marketplace. The former is achieved with the help of blockchain anchoring [52] and other accountable timestamping [59] techniques; the latter - with the help of digital signing and a blockchain-based PKI.

Blockchain full nodes and cloud storage are the remaining two parts of the ecosystem. Cloud storage could be an existing cloud storage, for example, Amazon Web Services (AWS), which allows for building HIPAA-compliant applications or Google Cloud Platform. One of the major reasons for integrating cloud storage into the ecosystem is to provide an off-chain storage solution especially for large biomedical data files, such as CT scans or MRIs, where the size of one data file could reach 50 Mb. The cloud storage may require authentication for read and write access to data, which in a preferable setup would be based on the PKI established on the marketplace blockchain. To ensure security and privacy, the data uploaded by users to the cloud storage would be encrypted on the user side using a threshold encryption scheme [97-100]. As the storage technology matures, it may be possible to replace cloud storage with the personal storage systems, where all the personal data would truly belong to the individual and also reside at the individual storage. The individuals also may be able to lend their data to the other parties for training purposes instead of selling the data.

Blockchain full nodes should be responsible organizations with an access to all information in the blockchain. They are divided into three subtypes:

● (Blockchain) validators: commit new blocks with transactions to the blockchain

● Auditors: audit the marketplace

● Key keepers: keep key shares according to a certain threshold encryption scheme necessary to decrypt user data in the storage. The precise protocol for key shares transmission and storage is out of the scope of this paper. In one possible setup, key keepers may have crypto-identities backed by a blockchain-based PKI, which would allow them to establish authenticated communication channels with other participants of the described protocol for key share transmission. In this setup, the keepers might use ordinary security mechanisms to guarantee at-rest security for the shares.

CLIENT WORKFLOW EXAMPLES

The intersections between different marketplace participants are illustrated in this subsection using several client workflows.

USER UPLOADS THE DATA

User chooses the data type and local path using system interface

1. User encrypts the data using a symmetric cipher (e.g., AES-256 in the CBC mode, or XSalsa20-Poly1305 authenticated encryption scheme used in libsodium [https://download.libsodium.org/doc/secret-key_cryptography/authenticated_encryption.html]). A Shamir's secret sharing technique [101] is then used to split the secret key to be distributed among key keepers, so that any K of key keepers together would be able to decrypt data, where K is a constant less than the number of key keepers N. The choice of constant K depends on the blockchain security model; as per Byzantine fault tolerance assumptions, $K > round\ (N/3)$.
2. User distributes key shares among key keepers, e.g., using a direct authenticated communication channel established with each keeper.
3. After user uploads encrypted data on a cloud it is consider to be LifeData.
4. User generates a transaction for data upload in order to notify ecosystem participants (in particular, data validators) that the upload has taken place. The transaction contains user's public key, data type, and a link to the data at the cloud storage.
5. User signs the transaction and broadcasts it to blockchain nodes.
6. The transaction is included into the blockchain via consensus algorithm.
7. Now data validators can buy this data for validation.

DATA VALIDATOR VALIDATES THE DATA

1. Data validator (DV) chooses the (batch of) unvalidated data and generates a transaction to buy it for the validation.
2. DV signs transaction and broadcasts it to the blockchain nodes.
3. The transaction is included into the blockchain via consensus algorithm. If the DV has enough

LifePounds to buy it for validation, the DV's LifePounds are sent to a validation smart contract and the workflow goes to step 4. Otherwise, the DV fails to validate data and goes to step 1.

4. Key keepers see an actionable data validation transaction in the blockchain. Each key keeper delivers stored key shares for each piece of data in the batch to the DV, e.g., via an authenticated communication channel.
5. DV uploads encrypted data from the cloud storage.
6. Once DV receives enough key shares from the key keepers, he decrypts the data.
7. DV validates data. The result is a vector of boolean values, signaling if corresponding pieces of data in the batch are valid or invalid w.r.t. the validation model used by the DV. The time for validation is limited. If validator failed to validate the data, the smart contract for data validation defaults to deeming all data in the batch valid.
8. DV forms and signs the transaction for data validation. The transaction contains the hashes of data and the validation result.

If the result for a particular data item in the batch is "valid", then

- LifePounds from smart contract are distributed among the data submitters' accounts according to the data value model described in the previous section
- Validated data becomes available for sale on the platform

DV will get a share of the revenue from third parties purchasing data from the batch in the future (see the following section). In one setup, the share of the revenue allocated to the DV is a blockchain-wide parameter.

If the result for a particular data item in the batch is "not valid", then

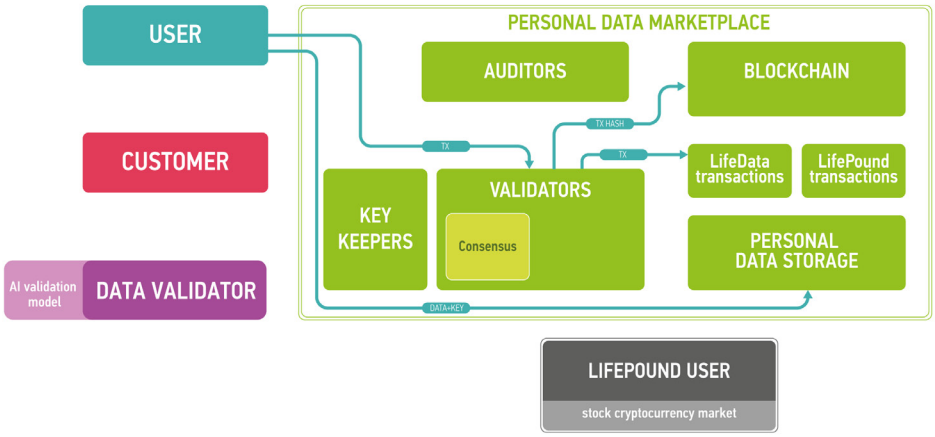

Figure 8: The workflow example for marketplace users. User uploads data and gets LifePounds as a reward (amount depends on the value of data).

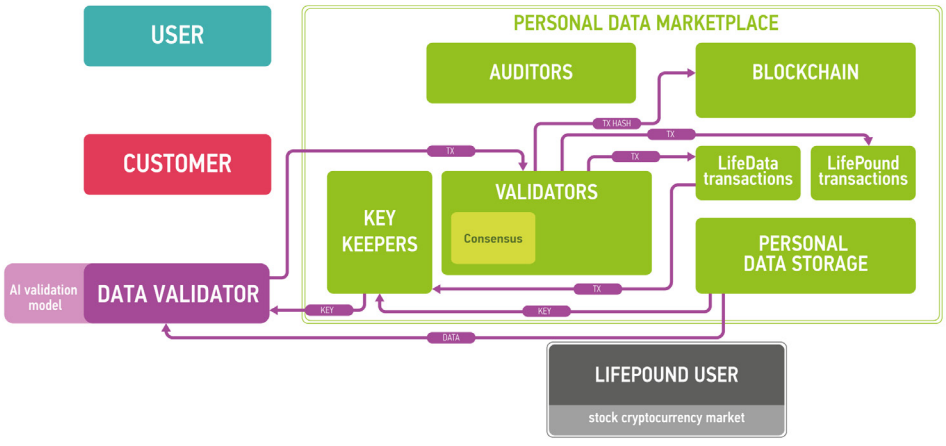

Figure 9: The workflow example for marketplace data validators (DV). Data validators are intermediate data buyers, who provide validation services to mine Lifepounds.

● LifePounds from smart contract are refunded to the DV's account

● Validated data is not available for sale at the platform.

Note. It is reasonable to have several DVs.

CUSTOMER BUYS THE DATA

1. Customer chooses the (batch of) validated data and generates a transaction to buy it.
2. Customer authenticates the transaction and broadcasts it to the blockchain nodes.
3. The transaction is included into the blockchain via consensus algorithm. If the customer has enough LifePounds to buy the specified data, the workflow goes to step 4. Otherwise, the customer fails to buy data and goes to step 1.
4. Key keepers see an actionable data purchase transaction in the blockchain. Each key keeper sends the key shares for all data in the batch and securely transmit them to the customer (e.g., via an authenticated communication channel).
5. Customer downloads encrypted data from the cloud storage.
6. Once the customer receives enough key shares from key keepers, he decrypts the data.

DATA SECURITY AND PRIVACY

One of the major challenges in the data-driven healthcare is data security and data privacy. Both the consumer, healthcare and research companies require the data of the many individuals to train their deep neural networks. The companies with the largest data sets acquire the data in the ways that may not be very transparent to the individuals and often the companies and the individuals do not understand the value of this data. The regulators often set up the barriers for the consumer data collection and storage substantially inhibiting the propagation of the recent advances in AI into the clinical practice (Figure 11).

The security of the described setup relies on the security of utilized crypto-primitives: the hash function and public-key signature scheme(s) utilized in the marketplace blockchain construction, as well as the symmetric cipher(s) and the secret sharing scheme(s) used for encrypting user data. Compared to centralized setups, the proposed scheme could allow to alleviate several attack vectors:

● Blockchain-based PKI for logically "well-known" users (e.g., blockchain validators, data validators, key keepers, etc.) could base on well-established measures for secure key management (e.g., key sharing, use of specialized hardware for key storage, etc.). These measures could be augmented with blockchain-based smart contracting (e.g., multi-signatures); further, blockchain could provide secure facilities for monitoring key revocation and issuance, which remain the weakest points for centralized PKI setups.

● The use of threshold encryption could allow to alleviate a single point of failure in long-term data storage. As data in the storage would be encrypted, the compromise of the storage would not lead to the data leakage. (Note, however, that access to the storage should be additionally restricted, e.g., by authenticating storage users with the help of the PKI established on the marketplace blockchain.) The compromise of a single key keeper likewise would not lead to the data compromise, as its key shares would be insufficient to decrypt data in the storage.

● The use of authenticated communication channels to transmit key shares could allow to achieve forward

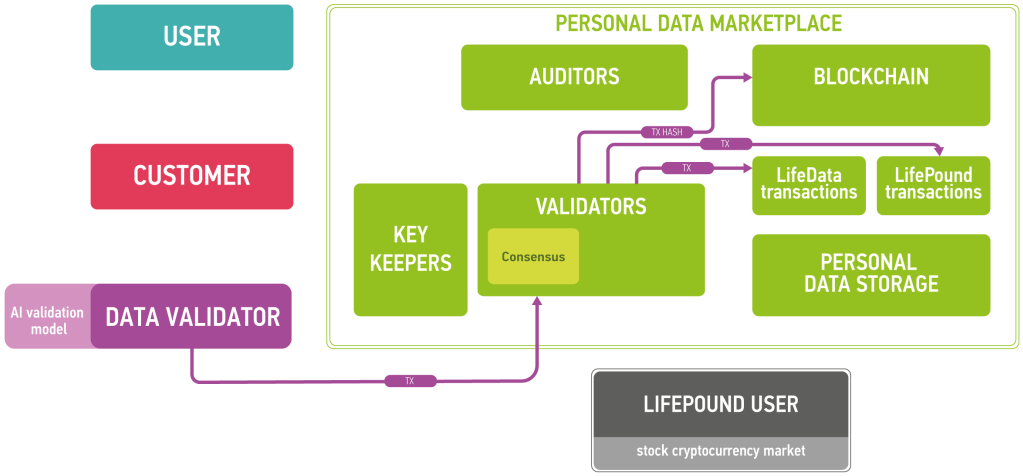

Figure 10: The workflow example for marketplace customers. Customers buy data for LifePounds. The data value model determines the data cost.

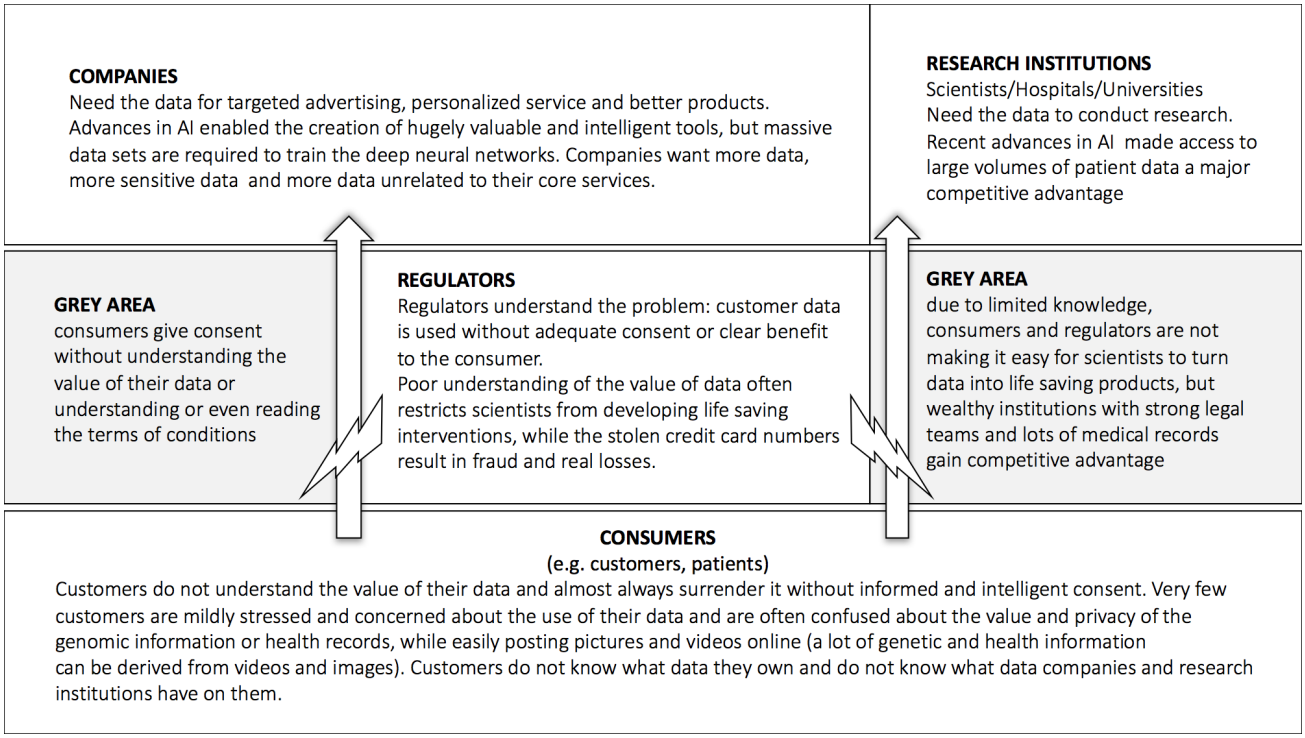

Figure 11: The flow of data from the individuals to the companies and research institutions. The introduction of the blockchain-based data ecosystem may help ensure that the individuals take control over their data and companies and research institutions may acquire data more freely reducing the need for the regulators to interfere.

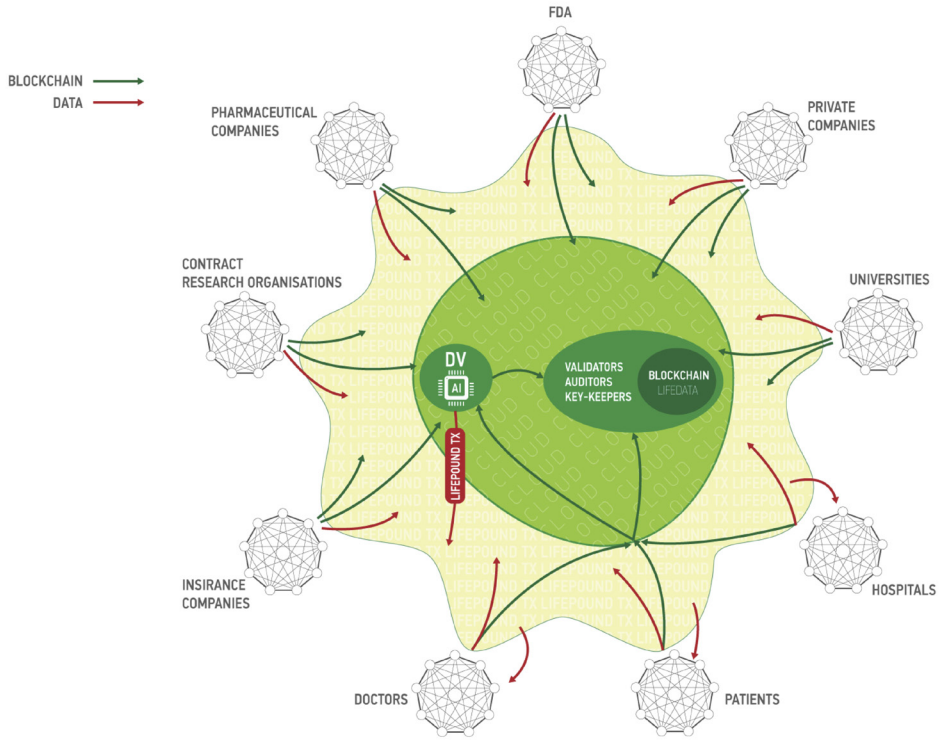

Figure 12: Proposed personal data-driven economy, where the individuals have full knowledge of and control over their data and are rewarded for generating new data and for providing the data for research or commercial purposes. Such ecosystem may allow the regulators including the Food and Drug Administration (FDA) and the pharmaceutical and consumer companies to exchange their data

secrecy, i.e., non-compromise of the encrypted user data even if the key keepers' long-term asymmetric keys would become compromised (incl. the compromise of the underlying asymmetric cryptosystem, e.g., with the advent of quantum computing).

The user-side key management for data encryption and authentication of blockchain transactions may be prone to various risks (e.g., a faulty random number generator resulting in generation of keys with insufficient entropy; inadequate security for long-term key storage; compromises of the user interface). Minimizing these risks requires careful design of the client software and supporting materials. Existing solutions for cryptocurrencies and/or generic key management may be adapted to reduce the risks.

The described setup does not concern data safety (in particular, protection against leakage) after the data has been purchased and transferred to the buyer. Such protection could be achieved with the help of existing security measures for data at rest and in use, and therefore is out of the scope of the present paper.

DEEP LEARNING FOR DATA QUALITY AND CONSISTENCY

While the DNNs are considered to have an exceptional generalization ability, they could be biased by the data they are trained on. Data quality is crucial for data-driven models; however, at the same time those models could be applied for data quality control and perhaps are the most suitable solutions for this task. First group of methods that could be utilized for healthcare quality check are unsupervised models aimed to detect anomalies that cluster far from the dataset. Deep autoencoders as unsupervised approaches which having as outputs input data itself and could be trained to reconstruct data also are suitable for anomaly detection. Poor-quality samples could be recognized as points with the highest reconstruction error [102]. Another set of approaches for the task are time-series based models, such as RNNs. Distribution of normal or good quality samples first is learned and then tested on a few next points in order to adjust model behaviour to the bias in the dataset not linked to the anomaly/poor quality samples. Anomalies in the data could also be linked to health conditions, so both those approaches could be used for pathology detection in health recordings [103, 104]. Finally, set of supervised techniques could be applied for data quality control [105]. However, one should take into account that supervised models require labelled dataset, which in case of anomaly detection will be highly unbalanced. Still, the problem could be solved with help of zero and one shot learning.

CONCLUSIONS

In this paper we presented the first attempt to assess the value of time and the combination value of personal data in the context of an AI-mediated health data exchange on blockchain.The value of the various types of data, combinations of the various data types, time value of one data type and time value of combination of data types is poorly understood and often debated. To address this problem, we foresee the emergence of a new profession "data economist" and creation of the health data economics research institutes. Recent advances in artificial intelligence enabled the creation of highly accurate predictors of biologically relevant features such as age, race and sex from very simple data types such as selfies, blood tests and such. The value of the various data types may depend on the application. For example, for the insurance companies, while the cost of data generation may be significantly higher for the genome compared to a selfie, the value of the recent picture of the patient may significantly exceed the value of the genome, since it may be more predictive of the patient's age, health status and mortality. However, the combination of these data types will be considerably more valuable than the value of these data types individually.

Blockchain and AI open new paradigms for health data ecosystems (Figure 12).

Blockchain technology enables the creation of a distributed and secure ledger of personal data, where patients are in control, own their data, and monitoring of access privileges and understanding of who looked at the data. Most importantly, blockchain technology allows for the creation of a data-driven marketplace, where patients can earn tangible rewards for making their data available to the application development community, pharmaceutical and consumer companies, and research institutions and generating new data through regular and comprehensive tests and checkups. Presently, only a few patients worldwide have the comprehensive data sets containing their clinical history combined with the genetic, blood biochemistry and cell count profiles, lifestyle data, drug and supplement use and other data types, because they do not see the value in this data and do not get tested regularly. On the other hand, the pharmaceutical and consumer companies alike are willing to pay substantial amounts for the large numbers of personal data records required to train their AI. These funds can be used to subsidize the regular testing by the patients, uncover the new uses for the various data types and develop sophisticated diagnostic and treatment tools.

Abbreviations

NHS, National Health Service; DNN, Deep Neural Network; CNN, Convolutional Neural Network; RNN, Recurrent Neural Network; AUC, Area Under the Curve; LSTM, Long Short-Term Memory; MRI, Magnetic Resonance Image; GAN, Generative Adversarial Network; HDSS, Highly distributed storage systems; HIPAA,

Health Insurance Portability and Accountability Act; PHI, protected health information; SPV, simplified payment verification; SOA, service-oriented architecture; KVS, key-value storage; PKI, Public key infrastructure; BFT, Byzantine fault-tolerant;

ACKNOWLEDGMENTS

We would like to thank Valery Vavilov, Marat Kichikov and George Givishvili of BitFury for their valuable advice and discussions that contributed to the development of the fully-distributed ecosystem. We would like to thank the developers of the Young.AI system and Anastasia Georgievskaya and Konstantin Kiselev for their valuable contributions to the sub systems leading to the development of the first prototype of the marketplace.

CONFLICTS OF INTEREST

The authors represent the commercially-oriented companies engaged in blockchain and artificial intelligence technology development and may directly benefit from the development of a blockchain-enabled personal health data exchange. The proposed blockchain- and AI-enabled ecosystem proposed in this article is highly speculative and the manuscript is submitted at the time when the association with the commercial blockchain technologies may lead to substantial profits or gains in research funding and academic reputation. The manuscript represents the opinions of the authors and not their respective institutions. The author declare a substantial conflict of interest.

REFERENCES

1. Earnest MA, Ross SE, Wittevrongel L, Moore LA, Lin CT. Use of a patient-accessible electronic medical record in a practice for congestive heart failure: patient and physician experiences. J Am Med Inform Assoc. 2004; 11:410-17.
2. Leveille SG, Mejilla R, Ngo L, Fossa A, Elmore JG, Darer J, Ralston JD, Delbanco T, Walker J. Do Patients Who Access Clinical Information on Patient Internet Portals Have More Primary Care Visits? Med Care. 2016; 54:17-23.
3. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, Iyer R, Schatz MC, Sinha S, Robinson GE. Big Data: astronomical or Genomical? PLoS Biol. 2015; 13:e1002195.
4. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, Liu B, Matthews P, Ong G, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 2015; 12:e1001779.
5. Zhavoronkov A, Litovchenko M. Biomedical progress rates as new parameters for models of economic growth in developed countries. Int J Environ Res Public Health. 2013; 10:5936-52.
6. Gleeson FC, Voss JS, Kipp BR, Kerr SE, Van Arnam JS, Mills JR, Marcou CA, Schneider AR, Tu ZJ, Henry MR, Levy MJ. Assessment of pancreatic neuroendocrine tumor cytologic genotype diversity to guide personalized medicine using a custom gastroenteropancreatic next-generation sequencing panel. Oncotarget. 2017; 8:93464-93475. https://doi.org/10.18632/oncotarget.18750.
7. Yi KH, Axtmayer J, Gustin JP, Rajpurohit A, Lauring J. Functional analysis of non-hotspot AKT1 mutants found in human breast cancers identifies novel driver mutations: implications for personalized medicine. Oncotarget. 2013; 4:29-34. https://doi.org/10.18632/oncotarget.755
8. Carpinetti P, Donnard E, Bettoni F, Asprino P, Koyama F, Rozanski A, Sabbaga J, Habr-Gama A, Parmigiani RB, Galante PA, Perez RO, Camargo AA. The use of personalized biomarkers and liquid biopsies to monitor treatment response and disease recurrence in locally advanced rectal cancer after neoadjuvant chemoradiation. Oncotarget. 2015; 6:38360-71.
https://doi.org/10.18632/oncotarget.5256.
9. Bennett CW, Berchem G, Kim YJ, El-Khoury V. Cell-free DNA and next-generation sequencing in the service of personalized medicine for lung cancer. Oncotarget. 2016; 7:71013-71035. https://doi.org/10.18632/oncotarget.11717.
10. Patel SP, Schwaederle M, Daniels GA, Fanta PT, Schwab RB, Shimabukuro KA, Kesari S, Piccioni DE, Bazhenova LA, Helsten TL, Lippman SM, Parker BA, Kurzrock R. Molecular inimitability amongst tumors: implications for precision cancer medicine in the age of personalized oncology. Oncotarget. 2015; 6:32602-32609.
https://doi.org/10.18632/oncotarget.5289.
11. Zhu Q, Izumchenko E, Aliper AM, Makarev E, Paz K, Buzdin AA, Zhavoronkov AA, Sidransky D. Pathway activation strength is a novel independent prognostic biomarker for cetuximab sensitivity in colorectal cancer patients. Hum Genome Var. 2015; 2:15009.
12. Artemov A, Aliper A, Korzinkin M, Lezhnina K, Jellen L, Zhukov N, Roumiantsev S, Gaifullin N, Zhavoronkov A, Borisov N, Buzdin A. A method for predicting target drug efficiency in cancer based on the analysis of signaling pathway activation. Oncotarget. 2015; 6:29347-56.
https://doi.org/10.18632/oncotarget.5119.
13. Bolotin DA, Poslavsky S, Davydov AN, Frenkel FE, Fanchi L, Zolotareva OI, Hemmers S, Putintseva EV, Obraztsova AS, Shugay M, Ataullakhanov RI, Rudensky AY, Schumacher TN, Chudakov DM. Antigen receptor repertoire profiling from RNA-seq data. Nat Biotechnol. 2017; 35:908-11.
14. Zabolotneva AA, Zhavoronkov AA, Shegay PV, Gaifullin NM, Alekseev BY, Roumiantsev SA, Garazha AV, Kovalchuk O, Aravin A, Buzdin AA. A systematic

experimental evaluation of microRNA markers of human bladder cancer. Front Genet. 2013; 4:247.

15. Di Meo A, Pasic MD, Yousef GM. Proteomics and peptidomics: moving toward precision medicine in urological malignancies. Oncotarget. 2016; 7:52460-74. https://doi.org/10.18632/oncotarget.8931.

16. Ionov Y. A high throughput method for identifying personalized tumor-associated antigens. Oncotarget. 2010; 1:148-55. https://doi.org/10.18632/oncotarget.118

17. Yin A, Etcheverry A, He Y, Aubry M, Barnholtz-Sloan J, Zhang L, Mao X, Chen W, Liu B, Zhang W, Mosser J, Zhang X. Integrative analysis of novel hypomethylation and gene expression signatures in glioblastomas. Oncotarget. 2017; 8:89607-19.
https://doi.org/10.18632/oncotarget.19171.

18. Lee D, Fontugne J, Gumpeni N, Park K, MacDonald TY, Robinson BD, Sboner A, Rubin MA, Mosquera JM, Barbieri CE. Molecular alterations in prostate cancer and association with MRI features. Prostate Cancer Prostatic Dis. 2017; 20:430-35.

19. Niklinski J, Kretowski A, Moniuszko M, Reszec J, Michalska-Falkowska A, Niemira M, Ciborowski M, Charkiewicz R, Jurgilewicz D, Kozlowski M, Ramlau R, Piwkowski C, Kwasniewski M, et al, and MOBIT Study Group. Systematic biobanking, novel imaging techniques, and advanced molecular analysis for precise tumor diagnosis and therapy: the Polish MOBIT project. Adv Med Sci. 2017; 62:405-13.

20. Alexander JL, Wilson ID, Teare J, Marchesi JR, Nicholson JK, Kinross JM. Gut microbiota modulation of chemotherapy efficacy and toxicity. Nat Rev Gastroenterol Hepatol. 2017; 14:356-65.

21. Sotgia F, Lisanti MP. Mitochondrial biomarkers predict tumor progression and poor overall survival in gastric cancers: companion diagnostics for personalized medicine. Oncotarget. 2017; 8:67117-28.
https://doi.org/10.18632/oncotarget.19962

22. Nielsen J. Systems Biology of Metabolism: A Driver for Developing Personalized and Precision Medicine. Cell Metab. 2017; 25:572-79.

23. Pretorius E, Bester J. Viscoelasticity as a measurement of clot structure in poorly controlled type 2 diabetes patients: towards a precision and personalized medicine approach. Oncotarget. 2016; 7:50895-907.
https://doi.org/10.18632/oncotarget.10618.

24. Radovich M, Kiel PJ, Nance SM, Niland EE, Parsley ME, Ferguson ME, Jiang G, Ammakkanavar NR, Einhorn LH, Cheng L, Nassiri M, Davidson DD, Rushing DA, et al. Clinical benefit of a precision medicine based approach for guiding treatment of refractory cancers. Oncotarget. 2016; 7:56491-500. https://doi.org/10.18632/oncotarget.10606.

25. Zhavoronkov A, Cantor CR. From personalized medicine to personalized science: uniting science and medicine for patient-driven, goal-oriented research. Rejuvenation Res. 2013; 16:414-18.

26. Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HY, Chen R, Miriami E, Karczewski KJ, Hariharan M, Dewey FE, Cheng Y, Clark MJ, Im H, et al. Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell. 2012; 148:1293-307.

27. Marx V. Biology: the big challenges of big data. Nature. 2013; 498:255-60.

28. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet. 2015; 16:321-32.

29. Lima AN, Philot EA, Trossini GH, Scott LP, Maltarollo VG, Honorio KM. Use of machine learning approaches for novel drug discovery. Expert Opin Drug Discov. 2016; 11:225-39.

30. Barardo DG, Newby D, Thornton D, Ghafourian T, de Magalhães JP, Freitas AA. Machine learning for predicting lifespan-extending chemical compounds. Aging (Albany NY). 2017; 9:1721-37.
https://doi.org/10.18632/aging.101264.

31. Vanhaelen Q, Mamoshina P, Aliper AM, Artemov A, Lezhnina K, Ozerov I, Labat I, Zhavoronkov A. Design of efficient computational workflows for in silico drug repurposing. Drug Discov Today. 2017; 22:210-22.

32. Mamoshina P, Vieira A, Putin E, Zhavoronkov A. Applications of Deep Learning in Biomedicine. Mol Pharm. 2016; 13:1445-54.

33. Aliper A, Plis S, Artemov A, Ulloa A, Mamoshina P, Zhavoronkov A. Deep Learning Applications for Predicting Pharmacological Properties of Drugs and Drug Repurposing Using Transcriptomic Data. Mol Pharm. 2016; 13:2524-30.

34. Wen M, Zhang Z, Niu S, Sha H, Yang R, Yun Y, Lu H. Deep-Learning-Based Drug-Target Interaction Prediction. J Proteome Res. 2017; 16:1401-09.

35. Gao M, Igata H, Takeuchi A, Sato K, Ikegaya Y. Machine learning-based prediction of adverse drug effects: an example of seizure-inducing compounds. J Pharmacol Sci. 2017; 133:70-78.

36. Putin E, Mamoshina P, Aliper A, Korzinkin M, Moskalev A, Kolosov A, Ostrovskiy A, Cantor C, Vijg J, Zhavoronkov A. Deep biomarkers of human aging: application of deep neural networks to biomarker development. Aging (Albany NY). 2016; 8:1021-33.
https://doi.org/10.18632/aging.100968.

37. Vandenberghe ME, Scott ML, Scorer PW, Söderberg M, Balcerzak D, Barker C. Relevance of deep learning to facilitate the diagnosis of HER2 status in breast cancer. Sci Rep. 2017; 7:45938.

38. FDA. The 510(k) Premarket Notification - Arterys Cardio DL. Guidance for Industry and Food and Drug Administration Staff. 2016. Available from:
https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K163253

39. Meng C, Zeleznik OA, Thallinger GG, Kuster B, Gholami

AM, Culhane AC. Dimension reduction techniques for the integrative analysis of multi-omics data. Brief Bioinform. 2016; 17:628-41.

40. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative Adversarial Networks. 2014. Available from: http://arxiv.org/abs/1406.2661

41. Kadurin A, Aliper A, Kazennov A, Mamoshina P, Vanhaelen Q, Khrabrov K, Zhavoronkov A. The cornucopia of meaningful leads: applying deep adversarial autoencoders for new molecule development in oncology. Oncotarget. 2017; 8:10883-90. https://doi.org/10.18632/oncotarget.14073.

42. Kadurin A, Nikolenko S, Khrabrov K, Aliper A, Zhavoronkov A. druGAN: An Advanced Generative Adversarial Autoencoder Model for de Novo Generation of New Molecules with Desired Molecular Properties in Silico. Mol Pharm. 2017; 14:3098-104.

43. Hivert MF, Grant RW, Shrader P, Meigs JB. Identifying primary care patients at risk for future diabetes and cardiovascular disease using electronic health records. BMC Health Serv Res. 2009; 9:170.

44. Choi E, Schuetz A, Stewart WF, Sun J. Using recurrent neural network models for early detection of heart failure onset. J Am Med Inform Assoc. 2017; 24:361-70.

45. Allam F, Nossai Z, Gomma H, Ibrahim I, Abdelsalam M. A Recurrent Neural Network Approach for Predicting Glucose Concentration in Type-1 Diabetic Patients. IFIP Advances in Information and Communication Technology. 2011; 0:254-59.

46. Ordóñez FJ, Roggen D. Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. Sensors (Basel). 2016; 16:115.

47. Pan D, Dhall R, Lieberman A, Petitti DB. A mobile cloud-based Parkinson's disease assessment system for home-based monitoring. JMIR Mhealth Uhealth. 2015; 3:e29.

48. Piette JD, List J, Rana GK, Townsend W, Striplin D, Heisler M. Mobile Health Devices as Tools for Worldwide Cardiovascular Risk Reduction and Disease Management. Circulation. 2015; 132:2012-27.

49. Margeta J, Criminisi A, Cabrera Lozoya R, Lee DC, Ayache N. Fine-tuned convolutional neural nets for cardiac MRI acquisition plane recognition. Comput Methods Biomech Biomed Eng Imaging Vis. 2015; 5:339-49.

50. Ahmed KB, Hall LO, Goldgof DB, Liu R, Gatenby RA. Fine-tuning convolutional deep features for MRI based brain tumor classification. Medical Imaging 2017: Computer-Aided Diagnosis. International Society for Optics and Photonics. 2017;101342E.

51. Altae-Tran H, Ramsundar B, Pappu AS, Pande V. Low Data Drug Discovery with One-Shot Learning. ACS Cent Sci. 2017; 3:283-93.

52. Cory N. Cross-Border Data Flows: Where Are the Barriers, and What Do They Cost? Available from: https://itif.org/publications/2017/05/01/cross-border-data-flows-where-are-barriers-and-what-do-they-cost.

53. Boonstra A, Broekhuis M. Barriers to the acceptance of electronic medical records by physicians from systematic review to taxonomy and interventions. BMC Health Serv Res. 2010; 10:231.

54. Kayaalp M. Patient Privacy in the Era of Big Data. Balkan Med J. 2017; 0:1.

55. Evans BJ, Jarvik GP. Impact of HIPAA's minimum necessary standard on genomic data sharing. Genet Med. 2017; 0:1.

56. Boonstra A, Broekhuis M. Barriers to the acceptance of electronic medical records by physicians from systematic review to taxonomy and interventions. BMC Health Services Research; 2010; 10:1:231. https://doi.org/10.1186/1472-6963-10-231.

57. Just E, Whitaker S. Addressing the Challenges of Translational and Clinical Research Data Management. Health Catalyst. 2015. Available from: https://www.healthcatalyst.com/addressing-challenges-clinical-research-data-management

58. Gottlieb LK, Stone EM, Stone D, Dunbrack LA, Calladine J. Regulatory and policy barriers to effective clinical data exchange: lessons learned from MedsInfo-ED. Health Aff (Millwood). 2005; 24:1197-204.

59. Office For Civil. Your Rights Under HIPAA. HHS.gov. US Department of Health and Human Services. 2017. Available from: https://www.hhs.gov/hipaa/for-individuals/guidance-materials-for-consumers/index.html

60. HIPAA Privacy Rule and Public Health Guidance from CDC and the U.S. Department of Health and Human Services*. 2003. Available from: https://www.cdc.gov/mmwr/preview/mmwrhtml/m2e411a1.htm

61. Health Insurance Portability and Accountability Act of 1996 Vol 104. Avaible from: https://www.congress.gov/104/plaws/publ191/PLAW-104publ191.pdf

62. Office for Civil Rights (OCR). 2015 [cited 2017 Oct 15]. Available from: https://www.hhs.gov/ocr/index.html

63. Nakamoto S. Bitcoin: A peer-to-peer electronic cash system. 2008. Available from: https://bitcoin.org/bitcoin.pdf

64. Lamport L, Shostak R, Pease M. The Byzantine Generals Problem. ACM Trans Program Lang Syst. 1982; 4:382-401.

65. Swan M. Blockchain: Blueprint for a New Economy. O'Reilly Media, Inc. 2015:1-152.

66. BitFury Group. Garzik J. Public versus Private Blockchains. 2015. Available from: http://bitfury.com/content/5-white-papers-research/public-vs-private-pt1-1.pdf.

67. BitFury Group. Digital Assets on Public Blockchains. 2016. Available from: http://bitfury.com/content/5-white-papers-

research/bitfury-digital_assets_on_public_blockchains-1.pdf.

68. BitFury Group. On Blockchain Auditability. 2016. Available from: http://bitfury.com/content/5-white-papers-research/bitfury_white_paper_on_blockchain_auditability.pdf.

69. Breitinger C, Gipp B. Proceedings of the 15th Int. Symposium of B, 2017. VirtualPatent-Enabling the Traceability of Ideas Shared Online using Decentralized Trusted Timestamping. 2017. Available from: https://www.gipp.com/wp-content/papercite-data/pdf/breitinger2017.pdf

70. Dwork C, Lynch N, Stockmeyer L. Consensus in the Presence of Partial Synchrony. J ACM. 1988; 35:288-323.

71. Stallings W. Cryptography and network security: principles and practices. Pearson Education India; 2006.

72. Erl T. Service-oriented architecture: concepts, technology, and design. Pearson Education India; 2005.

73. Pease M, Shostak R, Lamport L. Reaching Agreement in the Presence of Faults. J ACM. New York, NY, USA: ACM; 1980; 27: 228-34.

74. Kwon J. Tendermint: consensus without mining. 2015. Available from: http://tendermint.com/docs/tendermint.pdf

75. Buldas A, Lipmaa H. Schoenmakers - Public Key Cryptography B, 2000. Optimally efficient accountable time-stamping. Springer. 2000. Available from: http://link.springer.com/content/pdf/10.1007/b75033.pdf#page=304

76. Wu Y, Chen P, Yao Y, Ye X, Xiao Y, Liao L, Wu M, Chen J. Dysphonic Voice Pattern Analysis of Patients in Parkinson's Disease Using Minimum Interclass Probability Risk Feature Selection and Bagging Ensemble Learning Methods. Comput Math Methods Med. 2017; 2017:4201984.

77. Asgari M, Shafran I. Predicting severity of Parkinson's disease from speech. 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology. IEEE. 2010:5201-4.

78. Jiang F, Huang W, Wang Y, Tian P, Chen X, Liang Z. Nucleic Acid Amplification Testing and Sequencing Combined with Acid-Fast Staining in Needle Biopsy Lung Tissues for the Diagnosis of Smear-Negative Pulmonary Tuberculosis. PLoS One. 2016; 11:e0167342.

79. Yang SN, Li FJ, Liao YH, Chen YS, Shen WC, Huang TC. Identification of Breast Cancer Using Integrated Information from MRI and Mammography. PLoS One. 2015; 10:e0128404.

80. Jirtle RL, Tyson FL. Environmental Epigenomics in Health and Disease: Epigenetics and Disease Origins. Springer Science & Business Media; 2013. 302 pp.

81. Passos GA. Transcriptomics in Health and Disease. Springer; 2015. 344 pp.

82. Whitehead S. 360° of health data: Harnessing big data for better health. 2014. Available from: http://www.abpi.org.uk/our-work/library/medical-disease/Documents/360%20Degrees%20of%20Health%20Data.pdf

83. Mayo Clinic. Cholesterol test. Why it's done. Mayo Clinic. 2016 [cited 2017 Oct 15]. Available from: http://www.mayoclinic.org/tests-procedures/cholesterol-test/details/why-its-done/icc-20169529

84. Sebastiani P, Thyagarajan B, Sun F, Schupf N, Newman AB, Montano M, Perls TT. Biomarker signatures of aging. Aging Cell. 2017; 16:329-38.

85. Glei DA, Goldman N, Lin YH, Weinstein M. Age-Related Changes in Biomarkers: Longitudinal Data from a Population-Based Sample. Res Aging. 2011; 33:312-26.

86. Sanders JL, Minster RL, Barmada MM, Matteini AM, Boudreau RM, Christensen K, Mayeux R, Borecki IB, Zhang Q, Perls T, Newman AB. Heritability of and mortality prediction with a longevity phenotype: the healthy aging index. J Gerontol A Biol Sci Med Sci. 2014; 69:479-85.

87. Wu C, Smit E, Sanders JL, Newman AB, Odden MC. A Modified Healthy Aging Index and Its Association with Mortality: The National Health and Nutrition Examination Survey, 1999-2002. J Gerontol A Biol Sci Med Sci. 2017; 72:1437-44.

88. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. Circulation. 1998; 97:1837-47.

89. Fedintsev A, Kashtanova D, Tkacheva O, Strazhesko I, Kudryavtseva A, Baranova A, Moskalev A. Markers of arterial health could serve as accurate non-invasive predictors of human biological and chronological age. Aging (Albany NY). 2017; 9:1280-92. https://doi.org/10.18632/aging.101227.

90. Mitnitski AB, Mogilner AJ, Rockwood K. Accumulation of deficits as a proxy measure of aging. Sci World J. 2001; 1:323-36.

91. Nishijima TF, Deal AM, Williams GR, Guerard EJ, Nyrop KA, Muss HB. Frailty and inflammatory markers in older adults with cancer. Aging (Albany NY). 2017; 9:650-64. https://doi.org/10.18632/aging.101162.

92. Newman AB, Boudreau RM, Naydeck BL, Fried LF, Harris TB. A physiologic index of comorbidity: relationship to mortality and disability. J Gerontol A Biol Sci Med Sci. 2008; 63:603-09.

93. Moskalev A, Anisimov V, Aliper A, Artemov A, Asadullah K, Belsky D, Baranova A, de Grey A, Dixit VD, Debonneuil E, Dobrovolskaya E, Fedichev P, Fedintsev A, et al. A review of the biomedical innovations for healthy longevity. Aging (Albany NY). 2017; 9:7-25. https://doi.org/10.18632/aging.101163.

94. Kuo TT, Kim HE, Ohno-Machado L. Blockchain distributed ledger technologies for biomedical and health care applications. J Am Med Inform Assoc. 2017; 24:1211-

20.

95. Angraal S, Krumholz HM, Schulz WL. Blockchain Technology: Applications in Health Care. Circ Cardiovasc Qual Outcomes. 2017; 10:e003800.

96. Nugent T, Upton D, Cimpoesu M. Improving data transparency in clinical trials using blockchain smart contracts. F1000 Res. 2016; 5:2541.

97. Shamir A. Communications of the ACM A. 1979. How to share a secret. 1979. Available from: http://dl.acm.org/citation.cfm?id=359176

98. Blakley GR. Proceedings of the national computer 1979. Safeguarding cryptographic keys. 1979. Available from: https://pdfs.semanticscholar.org/32d2/1ccc21a807627fcb21ea829d1acdab23be12.pdf

99. Robling Denning DE. Cryptography and Data Security. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.; 1982.

100. Desmedt Y. Threshold cryptosystems. Advances in Cryptology — AUSCRYPT '92. Berlin, Heidelberg: Springer; 1992. pp. 1-14.

101. Menezes AJ, van Oorschot PC, Vanstone SA. Handbook of Applied Cryptography. CRC Press; 1996. 810 pp. https://doi.org/10.1201/9781439821916.

102. Vengertsev D, Thakkar H. Anomaly Detection in Graph: Unsupervised Learning, Graph-based Features and Deep Architecture. Available from: https://pdfs.semanticscholar.org/5049/920aeb54e481a865f9a9798b58706516fb10.pdf

103. Chauhan S, Vig L. Anomaly detection in ECG time signals via deep long short-term memory networks. 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA). 2015; 36678.

104. Wang K, Zhao Y, Xiong Q, Fan M, Sun G, Ma L, Liu T. Research on Healthy Anomaly Detection Model Based on Deep Learning from Multiple Time-Series Physiological Signals. Sci Program. 2016; 2016:1-9.

105. Revathi AR, Kumar D. An efficient system for anomaly detection using deep learning classifier. J VLSI Signal Process Syst Signal Image Video Technol. 2016; 11:291-99.