Health Informatics Journal

Original Article

Rapid health data repository allocation using predictive machine learning

Health Informatics Journal
2020, Vol. 26(4) 3009–3036

DOI: 10.1177/1460458220957486
journals.sagepub.com/home/jhi
SAGE

Md Ashraf Uddin, Andrew Stranieri, Iqbal Gondal and Venki Balasubramanian
Federation University Australia, Australia

Abstract

Health-related data is stored in a number of repositories that are managed and controlled by different entities. For instance, Electronic Health Records are usually administered by governments. Electronic Medical Records are typically controlled by health care providers, whereas Personal Health Records are managed directly by patients. Recently, Blockchain-based health record systems largely regulated by technology have emerged as another type of repository. Repositories for storing health data differ from one another based on cost, level of security and quality of performance. Not only has the type of repositories increased in recent years, but the quantum of health data to be stored has increased. For instance, the advent of wearable sensors that capture physiological signs has resulted in an exponential growth in digital health data. The increase in the types of repository and amount of data has driven a need for intelligent processes to select appropriate repositories as data is collected. However, the storage allocation decision is complex and nuanced. The challenges are exacerbated when health data are continuously streamed, as is the case with wearable sensors. Although patients are not always solely responsible for determining which repository should be used, they typically have some input into this decision. Patients can be expected to have idiosyncratic preferences regarding storage decisions depending on their unique contexts. In this paper, we propose a predictive model for the storage of health data that can meet patient needs and make storage decisions rapidly, in real-time, even with data streaming from wearable sensors. The model is built with a machine learning classifier that learns the mapping between characteristics of health data and features of storage repositories from a training set generated synthetically from correlations evident from small samples of experts. Results from the evaluation demonstrate the viability of the machine learning technique used.

Corresponding author:
Md Ashraf Uddin, Internet Commerce Security Laboratory, School of Engineering, Information Technology and Physical Sciences, Federation University Australia, Mount Helen, PO Box 663, Ballarat VIC 3353, Australia.
Email: mdashrafuddin@students.federation.edu.au

Keywords
digital health record storage, Blockchain, security and privacy, Big Health data, classifier, stream data, electronic health record, quality of performance, machine learning, deep learning

Introduction

The management of health data is no longer exclusively regulated by clinicians but increasingly requires a level of consent from patients.[1] Patients can decide who can access, analyze, and exchange their health information more than ever.[2] For instance, a patient has a great deal of control over Patient-Generated Health Data (PGHD) created, generated and collected by themselves, such as vital signs or fitness data.[2] Managing PGHD requires effort, cost and time for assimilating the data and as a consequence PGHD is rarely integrated with other repositories.

Patients who shared their self-tracking data with service providers expressed their dissatisfaction with the level of the provider's engagement with the data.[3] Despite this, PGHD can enhance medical care if the data can be incorporated with current health data systems following data storage requirements. Broad categories of patient-generated health data[4] such as medication information, biometric tracking, behavioral tracking, environmental tracking, social interactions tracking, genetic information, mental health assessment, symptom tracking, reported outcomes, and legal documents have been identified in the literature.[5] However, few studies have examined the management of storage of various kinds of health data generated by patients.

Legislation has emerged in most jurisdictions regarding the storage of health data. Most legislation such as HIPPA[6] and AHPRA[7] are organization-centric. Healthcare professionals or agencies typically own the health data that is produced and gathered under their oversight. However, some jurisdictions such as GDPR[8,9] introduced in Europe, are consumer-centered regulations where the patient has complete control over health data and must consent to the collection of his or her health information, decide how long the health care professionals will hold the data, and where the collected data will be processed and stored.[10] Although the data protection regulations of the GDPR enable patients to have complete control over their health data, most users are unable to handle large quantities of data, understand the nature of the data collected, or various methods of processing and track their personal data in compliance with the GDPR requirements.[8,11]

The appropriate management of health data is necessary to protect the patient's privacy and confidentiality while ensuring that data is available to relevant stakeholders. Recent reviews have identified the security of health data to be a major issue, particularly with the emergence of data from power, and memory limited medical sensors[12,13] and many medical data repositories.[14] Currently, the huge volume of health data is stored in repositories managed by different types of organizations. We discuss seven such health record agencies below:

1. Governments Controlled EHR: A government-managed electronic health record (EHR) is a record of a patient's health events throughout the lifespan. Diverse healthcare providers have access to subsets of the data where access is controlled by patients to different degrees. For instance, My Health Record[15] run by the Australian Government provides patients with mechanisms to control access to the data except in the context of criminal investigations or national security. EHRs are typically regulated by national laws that prescribe constraints such as the requirement that data be stored within national boundaries.
2. Proprietary eHealth Cloud: Global entities including Microsoft[16], Google,[17] and Apple[18] have hosted health data repositories on publicly accessible Cloud storage medium. Though these global entities have struggled to maintain continuity of service, smaller-scale proprietary repositories are continuously emerging, offering public or private Cloud-based

medical records storage. Patients are often provided with a high degree of control of their data by proprietary eHealth Cloud providers. However, aggregated data can be on-sold by these providers to third parties, and the Cloud administrators can always access confidential data.

3. Technology managed Blockchain EHR[19–22] is a decentralized, tamper-proof ledger-based EHR in which a certain number of transactions are bundled into a Block to be reviewed by nodes called Miners prior to writing the Block in the current ledger. Figure 1 provides an overview of how Bitcoin Blockchain operates. Some startup Blockchain-based EHR projects such as Patientory,[23] and GEM[24] have recently been introduced. Access to data is completely controlled by patients with no exceptions for giving control for criminal investigations, system administrators or other entities. IBM estimated that 70% of healthcare leaders expect that Blockchain technology will enhance current clinical trial management, regulatory compliance, and promote a decentralized health record sharing system (HRS).[25] Blockchain supports the processing and exchange of health data without the need for third parties trust. Traditional health record systems maintain a database that is operated and maintained by a single agency. In contrast, the Blockchain database is available to all individuals, but a user can only access his or her information stored on the Blockchain. In Blockchain technology, miner nodes verify and validate transactions on a peer to peer network before committing those transactions in a ledger that is replicated amongst all participants of the system which guarantees the immutability and irreversibility of the recorded documents. Further, public cryptography applied in the Blockchain ensures data persistence, provenance, distributed data control, accountability and transparency. Blockchain leveraged health record system can accelerate collaboration, sharing, integration of health data across various health agencies, healthcare professionals and patients.[26,27]
4. Healthcare service providers' Electronic Medical Records (EMR): Most contact patients have with their provider leads to data being added to the provider's Electronic Medical Records (EMR). In most jurisdictions, providers own and control the storage and access to patient records though all providers must comply with regulatory requirements prescribed by acts such as the Health Records Act in Victoria.[28] Patients have varying levels of access to data stored in repositories managed by healthcare providers.
5. Insurance organizations' Health Database: Health records are often stored in repositories controlled by insurance agencies or related organizations. Patients typically have minimal control or access to data managed by insurance agencies. These health databases are mainly managed for billing and administration, but can also be utilized by researchers, health authorities and other stakeholders to promote observational studies.
6. Disease specific registries: Registries for cancer-related records were first launched in North America and Europe between 1940 and 1950, respectively.[29] The cancer registry holds studies, screening, and test findings related to various cancers such as skin, breast, and cervix, as well as malignant tumors to provide information on the occurrence of cancer incidence and control. The cancer registry gathers data from various health agencies on cancer cases diagnosed or treated. For example, the Australian Cancer Database (ACD)[30] contains data about all cases of cancer diagnosed in Australia since 1982.[31]
7. Patient controlled Personal Health Record (PHR): PHRs include patient-generated health records collected via consumer health apps, sensors, and wearable devices.[1] Health data can be hosted on storage systems entirely managed by patients. For instance, patients may collect their own blood glucose, ECG and other readings and store the data in a personal health record system they manage.

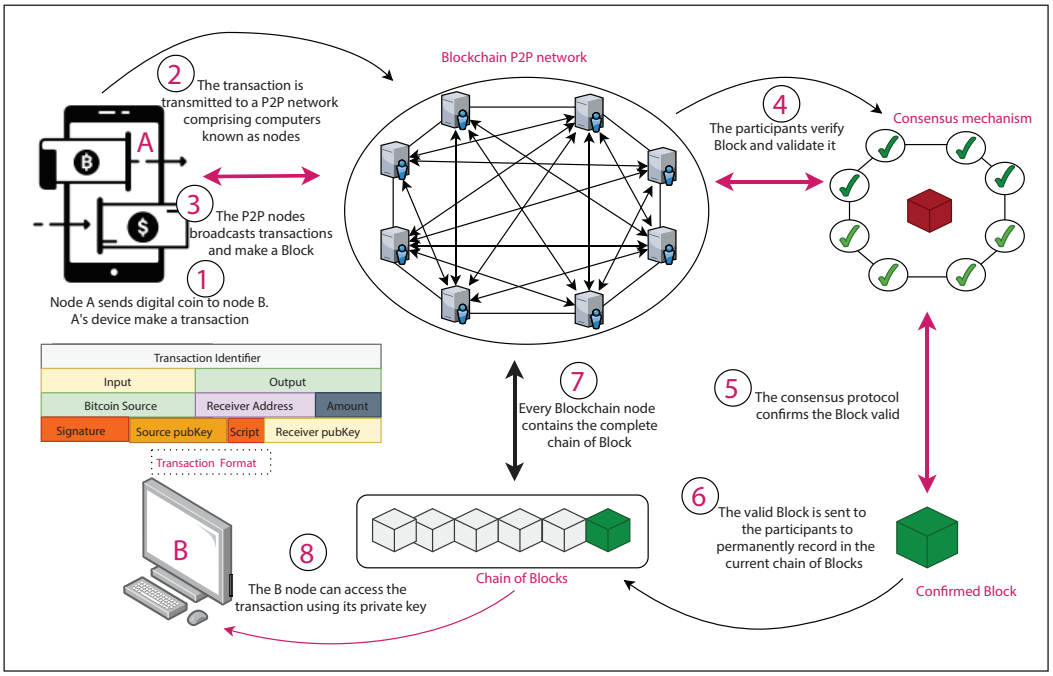

Figure 1. The basic operation of a Blockchain.
[1]A Blockchain's participant A wants to transfer some digital coins to another participant B. A's device needs to make a transaction using user's wallet. Participants can usually utilize their portable devices such as smartphone, laptop, and low-processing computer for making transactions. The transactions are digitally signed with A's private key and transaction contents are encrypted with the B's public key if necessary.
Next, [2]A's device transmits the transaction to a peer-to-peer network that comprises of high-processing devices also known as nodes. The Blockchain algorithms and protocols are implemented on this network.
After that, [3]the nodes on the Blockchain network replicate the transaction and broadcast it throughout the network. The nodes packed a certain number of transactions in a Block.
[4,5,6]All participants only bind the Block to the current chain of already verified Blocks when a miner node produces the new Block 's target hash code using Proof of Work method also known as computational puzzle. The verification process on the Blockchain, is called the consensus mechanism that varies in terms of computational cost and turnaround time.
Finally, [8]the B's device can access the transaction from the confirmed Block using its private key.

Each of the seven types of repositories for the storage of health data outlined above has different costs, security vulnerabilities, accessibility levels, usability features, and reliability track records. For example, Blockchain repositories avoid a trusted authority but are computationally very expensive. The government-run My Health Record prevents unauthorized individuals from sharing or disclosing patient data but has restricted capacity to store streamed data from sensors. Propriety Cloud eHealth repositories can provide patients with theoretically unlimited storage, but the retrieval of data can be slow.

The need to maintain privacy and confidentiality is often depicted as minimal requirements for all health data; however, in practice, health data is not equally sensitive for every patient at all times. A patient may generate her own ECG data for storage on a personal health record, allow indicators such as the ST segment to be copied to her cardiologist's record and be available to other providers through a government-operated EHR, however, rescind this when she attains a high public profile. The same patient may be compelled to accept that her provider will store her pregnancy test results but prefer that data should not be available to anyone else at all.

Health data can be thought to be disseminated among diverse agents managing storage repositories in such a way that the nominated storage medium reflects data management requirements including the quality of service, cost, volume, confidentiality, security, and privacy of data that the patient desires for each chunk of his or her data. Ko et al.[32] has taken one step toward this ideal by describing a hybrid execution model to store data defined as "sensitive" in a private Cloud and "non-sensitive" data in a public Cloud. This approach facilitates the processing of sensitive and non-sensitive data as defined by the user while preserving the user's privacy. However, this approach was not explicitly advanced for health data. Further, the communication between two kinds of Cloud platform causes long network delays and requires high bandwidth for data-intensive computation. Zhang et al.[33] advanced a hybrid Cloud platform within the same network to address the issue.

Artificial intelligence in healthcare has made it possible to automatically diagnose health data while streaming data from medical sensors, apps and devices. An algorithm can categorize specific health data, including ECG, blood pressure, and pulse rate as normal or abnormal based on a range of conditions, and the threshold set by healthcare professionals. For example, ECG wave having RR interval, QRS complex, and QT interval within the range of [0.12–0.20 s], [0.06–0.10 s], [0.30–0.44] respectively, and R-wave is less than or equal to 0.12 s is considered to be abnormal.[34] Abnormal data are usually clinically useful and important for potential research. To preserve abnormal data, Vaidehi et al.[14] proposed a multi-agent-based health monitoring system for elderly people using Body Area Sensor Networks. Four kinds of agents named Admin, Control, Query, and Data Agent manage health records where the Data Agent classifies the medical data as normal or abnormal. Normal data is filtered out, and abnormal data is compressed to handle Big data challenges in continuous patient monitoring. However, this approach assumes a single storage medium.

Huge amounts of health data are now generated, which necessitates diverse storage options.[35] Al Ghamdi and Thomson[36] explored different online storage systems and presented a case study for the storage of data generated from an oil company. Ghamdi identified storage-related challenges including energy consumption for operating and cooling storage, the capacity of repositories to cope with the growth of Big data, unused storage, the risk associated with data loss, downtime, and backup issues of different storage mediums. NetApp platform among NAS (Network Area Storage), SAN (Storage Area Network) and DAS (Direct Attached Storage), Cloud and Hadoop were suggested as storage mediums for Big data. A follow-up survey was conducted, which showed that NetApp facilitated data encryption, compression and solved the unused storage problem. Although they considered multiple storage mediums and assessed those against relevant criteria, oil company data, unlike health data, is relatively uniform, so no method for dynamically selecting a storage medium was proposed.

Many researchers[37–39] have developed methods for selecting suitable Cloud Service Providers (CSPs) to store consumer data, taking into account the performance and cost parameters of the CSPs. Ruiz-Alvarez and Humphrey[37,38] proposed a model that considered an application's requirements and user's priorities to choose a Cloud server among different Cloud Service Providers (CSP). They developed a mathematical model based on Linear Integer Programming with respect to storage computing cost and performance characteristics, including latency, bandwidth, and job turnaround. Yoon and Kamal[39] also proposed a Linear Integer Programming model that used processing time and cost to optimally allocate datasets to distributed heterogeneous Clouds. The Cloud service with high processing power minimizes the operational time but incurs high operational costs. Conversely, the Cloud service with low processing time minimizes the operational cost but increases the processing time. As in other work cited here, this work also focused on the performance assessment of different Cloud Service Providers but did not focus on mechanisms for selecting different types of health data repository. The more general problem of how best health data can

be disseminated among multiple health management systems based on data management requirements and patient preferences has still not been addressed.

Further, the Blockchain that promises security and privacy has prompted researchers to investigate it for the management of health data. However, Blockchain technologies are not an ideal solution for hosting Big health data due to its design. To address this issue, a number of researchers suggested merging traditional health databases with Blockchain-based eHealth and distributing data among them according to the user's choice and probable future data usage. For instance, Uddin et al.[40] advanced an architecture that places a software agent known as a Patient Agent that is aware of the patient's preferences, on hardware that could continuously make the storage repository decision on the basis of data sensitivity, context, significance, security, and access level. However, they did not describe a feasible model for making this decision. In addition, most of the focus by Ko et al.,[32] Zhang et al.,[33] and Uddin et al.[40] was on the development of improved cryptographic techniques to protect sensitive health data in the Cloud.

We extended these approaches by developing a model that can make the storage repository decision to select a repository amongst a range of repositories by taking into account a broader analysis of patient data beyond the "normal" or "abnormal" criteria Vaidehi et al.[14] adopts by also taking into account other factors such as data security, privacy, and QoP (Quality of Performance) requirements. In our work, we have considered data variations in terms of sensitivity, volume, and other factors in order to direct data to one or more of the health record management systems available.

Further, the state-of-the art works have not dealt with data storage requirements but rather focused on Cloud Service Providers (CSP) selection based on diverse criteria using optimization methods. We propose a novel health data storage recommendation model to distribute health data among multiple health repositories using machine learning.

Our work involves an automated health data storage recommendation model that suggests an appropriate storage repository by considering health data sensitivity, quality of performance, and patient's security and privacy preferences.

We describe relevant literature in the next section, our model after that, and evaluation trials in the results section before concluding the paper.

Related literature

The amount of health data has risen exponentially, with growing numbers of patients wearing bracelets and other medical IoT sensors. Each health record system cannot necessarily meet the requirements of Big data in terms of storage space, storage speed, storage structure, etc. Moreover, patients are at risk of losing important medical information[41] if the correct health record system is not selected.

In some studies, the health data generated from wearable sensors, and different medical apps were manually uploaded to personal health record systems which might have delayed the response from the caregivers. To address this issue, Andy et al.[42] and Peleg et al.[43] advanced the usages of patient-generated data by uploading it to commercial blood glucose monitors. Martinez et al.[44] developed an automated blood pressure cuff that channelled data to the HealthVault[45] hosted by Microsoft. Some research[46] has suggested filtering or compressing streamed data to fit into the electronic health record system. Hohemberger et al.[46] addressed the challenges of storing health data streamed from wearable sensors in EHR (Electronic Health Record) and proposed health data reduction policies that intended to save the heart rate of a patient in a specific range of ages.

The research in[47–50] advocated some action plans and standards to adopt an electronic health record system. However, these studies[47–50] did not develop any model to accommodate user's preferences and data storage requirements. Busis[47] urged healthcare professionals to follow three steps:

assessment, planning, and selection before adopting an electronic health record systems. Healthcare practitioners were recommended to recognize their requirements and affordability during the assessment process. In the planning steps, they would define their goals and identify priorities and barriers while choosing a health record system. Finally, many criteria for assessing a health record system such as time-saving, ease of use, billing, quality of service, and the ability to participate in a particular insurance plan are determined in the selection phase.

Weathers and Esper[48] emphasized that when choosing a specific Electronic Health Record, functional needs, troubleshooting, and optimization facilities should be taken into account. The author provided a checklist to follow before purchasing any electronic health record system. The checklist mainly covers on-site client meeting arrangements, site visiting, maintaining live workflow and others. Hart et al.[49] proposed 10 laws to follow before choosing a specific digital health data repository: future use, volume and access time of data, backup capabilities, and privacy protections, storage costs are important factors to be considered when choosing repositories for health data.[50]

Boonstra and Broekhuis and Ross et al.[51,52] described several obstacles faced by medical professionals and practitioners while adopting an electronic record system. Some of these are high implementation costs and maintenance costs, legal and technical problems like system complexity, lack of support staff, low customizability. Healthcare professionals and patients are usually not incentivized for using electronic health records, which has hindered wider adoption of Electronic Health record system. Further, patients and healthcare professionals' concerns regarding privacy and security have not been addressed to the extent they expect.[53]

Khan and Hoque[54] described the need to create a data warehouse for health data spread across a variety of sources, including clinics, hospitals, insurers, and patients. They proposed a broadly accepted conceptual and logical data warehouse model to store various types of geographically dispersed health data. They defined two data criteria: the amount of unstructured health data and confidentiality that the data warehouse model would tackle.

Hart et al.[49] emphasized that medical data should be stored in plaintext without filtration and compression. As the data analytics and processing method upgrade or change over time, future reanalysis and reproducibility may be possible to be carried out on the data to improve insights. Researchers can encounter difficulties in verifying potential empirical results, the validity of statistical models, and findings through studies using the derived data. However, the difficulty of maintaining raw data lies in protecting data integrity. The emerging Blockchain technology can provide a viable alternative to preserving raw data integrity. Blockchain can support the on-chain cryptographic hash code of the raw data to be maintained in a decentralized manner, which can validate the integrity of the raw data stored in off-chain.

Privacy in health informatics refers to an individual's right to monitor and control access and distribution of health data. Patients are often unable to fully control their health information, but they desire more control over their health information.[53] Individual's desire for privacy is influenced by their gender, age, the level of data sensitivity, and health conditions.[55] Some research[56] indicates women are more concerned about privacy than men. Yet Kenny and Connolly[55] concluded that males have greater privacy concerns regarding health data than females. Kenny and Connolly has described human characteristics, behaviors and experiences as driving factors in individuals' increasing concerns about privacy. The authors verified several hypotheses through their studies. For instance, individuals are hesitant to reveal sensitive information about health. Age has a positive influence on privacy matters, with older people having more concerns about privacy. An individual with a health condition typically has less privacy concerns, as they seek to benefit from health services.[55] Rahim et al.[57] provided a conceptual model for patient privacy preferences in the healthcare system. In the model, he identified four antecedents that positively

influence the patient's privacy in the healthcare environment. The antecedents described in the model include the needs for exchanging health data, the patient's faith in the EMR, the ease of access control in the EMR and patient's security awareness.

Many studies[58–62] identified a wide range of parameters for evaluating Cloud services and proposed some guidelines that should be followed when choosing health records. We reviewed the following literature that advanced standards for assessing health record systems in order to design our proposed model.

Chang et al.[58] developed an objective mathematical framework for maximizing benefits with a given budget and cost to minimize the likelihood of CSP failure and improve availability. DP (Dynamic Programming) was used to select the best CSPs. The method maximizes the number of data blocks that survive when certain CSPs fail, or are subject to a fixed budget. Rehman et al.[59] put forward a framework for tracking the performance of CSP through feedback from users. Qu et al.[60] introduced a CSP selection process based on user feedback that includes four components; Cloud Selection Service, Benchmark Testing Service, User Feedback Management Service, and Aggregation Evaluation Service. Qu et al. defined the criterion for choosing CSP as subjective or objective. Cloud consumers give ratings as subjective criteria to the system, and third party trust supplies the system with measurable CSP performance as objective criteria. A simple Additive Fuzzy System which aggregates subjective and objective criteria were used to rank the available CSPs.

Lee and Seo[61] suggested a hybrid multi-criteria decision-making model for CSP selection in which, initially, decision-making factors were defined using a Balanced Scorecard (BSC) methodology. Secondly, critical decision criteria were extracted using the Fuzzy Delphi (FDM) process and thirdly, weight allocation for each decision-making criterion and CSP selection was carried out using FAHP (Fuzzy Analytic Hierarchical Process).

Halabi and Bellaiche[62] hierarchically identified a set of criteria for evaluating the security of CSPs, where security was subjectively and objectively evaluated using the Analytic Hierarchical Process (AHP). In order to comply with a CIA (Confidentiality, Integrity, and Availability), Halabi and Bellaiche[63] has also introduced a broker-based system that will fulfil the Service Level Agreement. They developed a CIA-based optimization function to identify CSPs with minimal user frustration for CIA. Halabi et al.[64] addressed online Cloud services allocations in view of global safety satisfaction. A linear optimization technique is used to formalize the resource allocation problem in relation to global security requirements. The linear optimization problem formulated was solved using a genetic algorithm.

Patient-centered health data with structural heterogeneity are produced at a particularly high rate, and high magnitude so needs to be stored and processed rapidly. Precision is crucial to extract useful insights from health data, but some sources generate vague and inaccurate data. Nonetheless, a distributed data management system can resolve these issues to some degree.[26]

The studies discussed above have explored diverse Cloud storage mediums. However, these studies did not develop machine learning-based mechanisms to meet user's preferences and data features and also did not design the selection of repositories considering various health data storage systems and data properties. Our approach for facilitating distributed health data management is outlined in the next section.

The health repositories recommendation model for health data

The storage recommendation system presented here assumes the patient is in control of the storage decision along the lines advocated by the "Gimme me my dam data"[65] movement. In many jurisdictions, health information generated by healthcare providers is owned and controlled

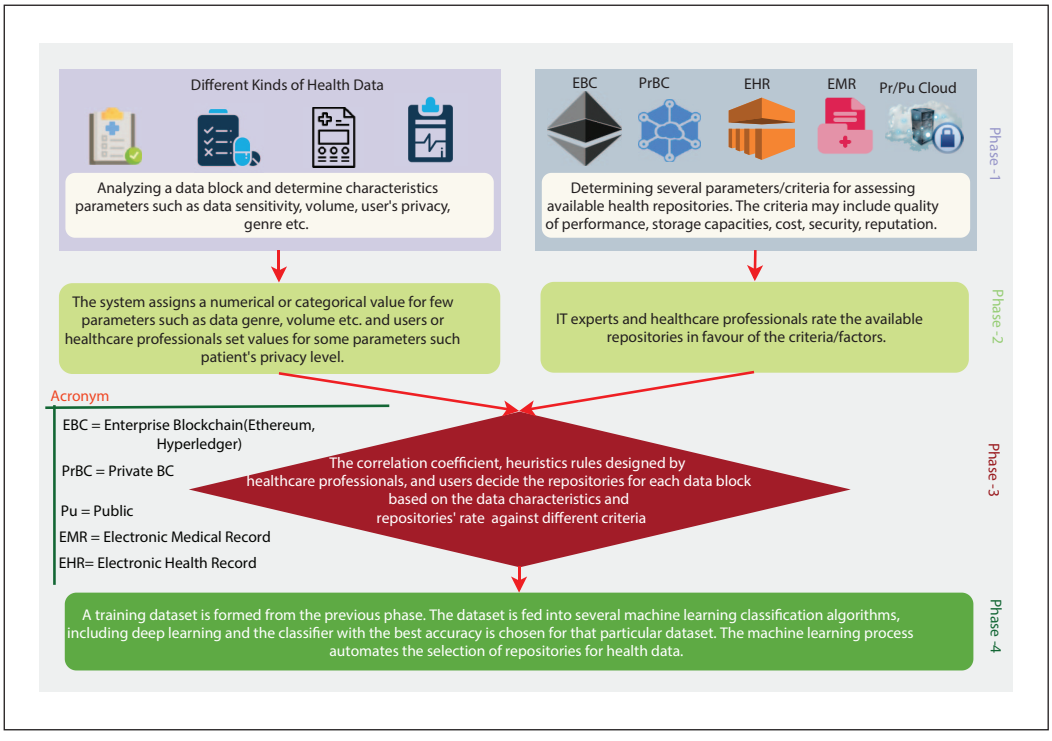

Figure 2. The high level view of the proposed recommendation model.

by a healthcare provider. However, as consumer health movements increase in popularity and increasingly patients generate their own data, storage decision's are assumed to become more pressing for patients. Further, as the quantum of streamed data increases, storage decisions must be made so frequently that manual consultation with the patient becomes cumbersome, and an automated process is required.

The process advanced here maps information about the storage requirements a patient has for a block of data to the storage features of repositories managed by diverse agents. However, a patient's data storage requirements vary enormously and cannot necessarily be pre-specified to cover all future patient contexts. This is managed by having a mapping manually specified by experts as a training set for a machine learning classifier to learn to generalize to a mapping that covers a wider set of patient contexts. Figures 2 and 3 show the overall approach developed here, explained in detail below. First, we describe a set of variables or features characterizing the requirements for storing a chunk of data – *the data storage requirements* which is illustrated in Phase-1 of Figure 2.

Some of the attribute's values for data storage requirements are declared to be numerical (range between 1 and 5) and some are categorical. Secondly, a dataset having these attributes or features is constructed where each instance reflects the specifications needed for storing a particular chunk of data (this constitutes Phase-2 of the model shown in Figure 2).

Next, the features that reflect characteristics of storage repositories called the *Health Repositories Evaluation Criteria* are calculated by adding the rating provided by an expertise group. This is presented in Phase-2 of Figure 2. Throughout this scenario, we are ranking five storage repositories against four standards. In Phase-3 of Figure 2, ultimately, statistical correlation, clinical heuristic rules (those rules can be created by the medical professionals or patients themselves), and user

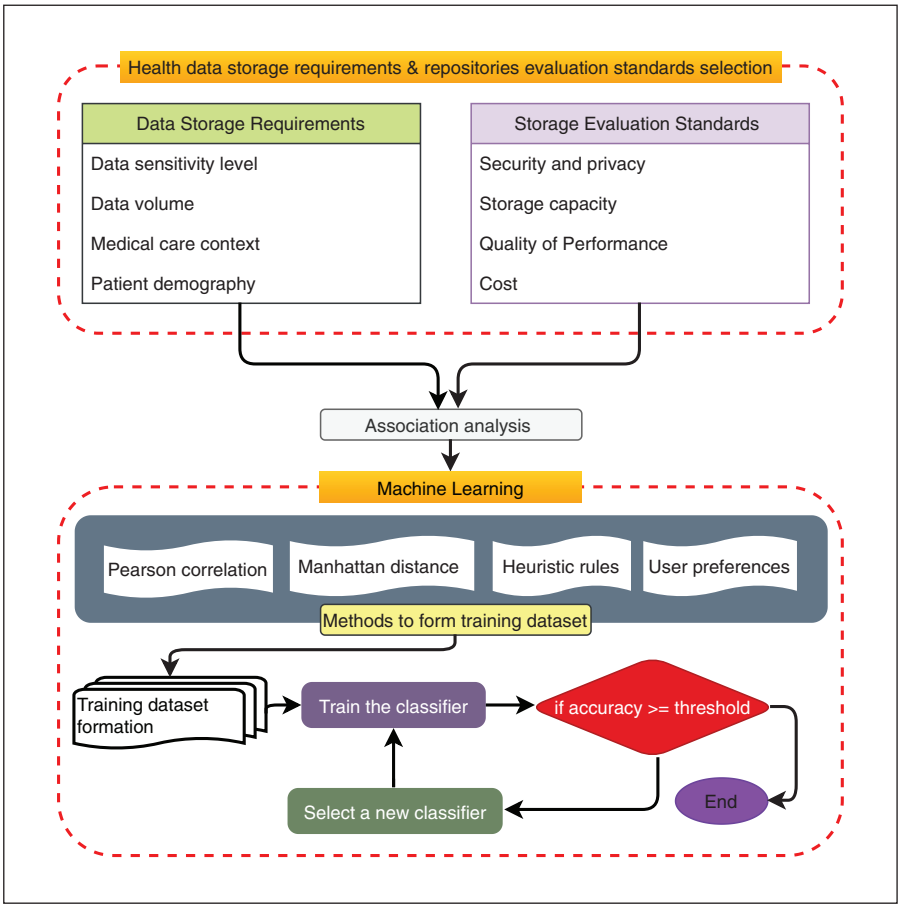

Figure 3. The health data storage recommendation systems.

preferences are used to decide the class labeling for each instance in the dataset. The experts or users may, in a real situation, allocate a storage repository (class label) to an instance that will be encoded using heuristic rules. The correlation coefficient is used to infer the class label of those instances for which a user's preferences or heuristic rules are not exactly matched.

In Phase-4 of Figure 2, a machine learning classifier trained with the sample dataset containing user and expert expectations can, therefore, generalize the mapping of data requirements to health repositories. The storage recommendation framework shown in Figure 3 comprises two parts: the selection of data storage requirements and assessment standards for health repositories, and Machine Learning. Each component is described below.

Data storage requirements and health repositories assessment standards selection

The upper part of the framework in Figure 3 includes features that reflect characteristics of the data to be stored called *Data Storage Requirements*, and features that reflect characteristics of storage repositories called the *Health Repositories Evaluation Criteria* and an association analysis between the two sets of features.

Data storage requirements

The requirements considered relevant for deciding which repository should best be used for a chunk of data have been selected from the literature and include sensitivity, volume, medical care context and patient demographic data:

- Sensitivity: Although all health-related data should be prevented from unauthorized access, some data can be regarded to be more sensitive to breaches than other data. The level of data sensitivity can be expected to vary from individual to individual, depending on their personal preferences and contexts. For example, data concerning a person's sexual orientation may be highly sensitive for one person in one context compared with another person in the same or different context. To illustrate, an ECG trace at one point may need to be kept extremely secure against unauthorized access for one patient but less so at another point in time.
- Volume: Is the data block a single small block as in a test result or is the data streaming forming huge datasets such as continuous streams including ECG, blood pressure, temperature, and oxygen level? This latter dataset requires health storage repositories that can support virtually unlimited storage, whereas static reports, medical diagnoses, and medication summaries are occasionally generated and do not need a storage medium with high capacity.
- Medical Care Context: Although many contexts patients find themselves in can be identified, a small number of contexts can be identified at a coarse-grained level. For this work, four contexts were considered sufficient to describe common medical care contexts: a palliative care context, emergency context, chronically ill context or non-chronic disease context. Medical care contexts can also be expected to vary from country to country. For example, in Australia, medical contexts might include front line care (GP), hospital care, emergency care, specialist care, allied care, elderly care and palliative care. Different care contexts can be served by storage repositories to different extents. For example, having health data stored in EMR managed by healthcare providers is more desirable during emergency or life-threatening contexts because it can be retrieved quickly.
- Patient Demographics: Data such as socio-economic status, profession, education, and nationality can play a significant role in the selection of a storage medium. For instance, storage cost may be particularly important for a person on a low income, whereas confidentiality may be very important for a person with a high public profile.

Health repositories evaluation criteria

Tables 1 and 2 illustrate features that distinguish four of the organizations that manage health data repositories described above. While many factors distinguish one manager from another, we limit our focus to four: security and privacy, performance quality, capacity, and cost. Each of the four main criteria has sub-criteria. Criteria related to the performance of a repository, such as downloading or uploading speed, data availability, and maintenance services are clustered as *Quality of Performance* criteria. Likewise, criteria related to security and privacy, such as the capabilities of preserving confidentiality, data integrity, and resistance to cyberattacks are listed as *Security and Privacy*. Figure 4 shows criteria and sub-criteria against which health repositories are assessed.

- Security and privacy: This includes confidentiality that represents the capacity for a storage medium to protect patient's data against inappropriate disclosure or tampering by the insider

Table 1. The strength and weakness of health repositories against criteria.

Criteria for evaluating health repositories	Government EHR	Blockchain EHR
Security and privacy	Government employees may access health data without patient's knowledge. EHR realizes legal compliance constraints enshrined in legislation.[66] EHR implementations are subject to rigorous audit process, minimizing the risk of data manipulation.	Blockchain EHR is a patient-driven data management technology that prevents unauthorized access to records. Blockchain EHR can anonymously process health records and guarantee information integrity by copying the entire ledger to multiple entities. However, a patient's privacy is breached if attackers can discover the data owner through content analysis.[67] Although Blockchain EHR withstands major cybersecurity attacks such as Denial of Service (DoS), Ransomware and single point of failure, it is susceptible to protocol related attacks such as a long-range attack, and mining attacks known as 51% attacks.
Storage capacity	Government EHR is a scalable storage management system but not suitable for streamed data. Although EHR facilitates an extensive archive of patient medical history with a high level of security, uploading streamed data to EHR is impracticable due to a large amount of data that needs to be stored over the time.[46]	Blockchain does not provide scalable storage facilities for mining Big health data on-chain as the record is required to be replicated in every participant.[68] However, off-chain data management in the Blockchain can meet this challenge.
Quality of performance	EHR maintains standardized and uninterrupted coordination services promptly.	Blockchain EHR can support cross border sharing of health data while preserving confidentiality and integrity. Users can access data from various points. However, slow processing and access to health data[69] due to limited scalability, legal and political compliance issues[70] can impact the quality of care.
Cost	Government EHR requires high implementation, maintenance and administrative costs that many national governments might not afford. However, government management of EHR maximizes cost-effectiveness and quality of care for the patient.[71]	Blockchain EHR alleviates many service costs, including employee wages, a legal fee but users have to contribute computational resources.

EHR: electronic health record; Blockchain EHR: Blockchain based electronic health record.

or outsider attackers. Some storage repositories have better cyberattack defenses than others. Additionally, some storage repositories can keep data accessible at all times or deliver data upon request, including unexpected disruptions like hardware failures, cyberattacks or natural disasters better than others. For some repositories, only the receiver and sender are involved in processing patient data, whereas others involve third-parties. Some repositories

Table 2. The strength and weakness of health record systems.

Criteria for evaluating health repositories	Proprietary eHealth cloud provider	Healthcare provider EMR
Security and privacy	Patient's identifier and health data is accessible by Cloud administrators[72] that threatens patient's privacy. It cannot guarantee the integrity of health data due to third parties' involvement in processing and providing storage.[73,74] Further, Cloud database is prone to many cyberattacks, including data breaches, prefix hijacking,[72] spoofing identity, trust management and non-repudiation among servers. However, top Cloud providers such as Microsoft, Amazon web Service safeguard customer's data from malicious attacks and facilitate the availability and access to data across multiple organizations located worldwide.	Insiders such as healthcare professionals, and support staff are associated with over half of recent health data breaches[75] in EMRs. EMRs are defenceless against different cyberattacks, including DoS, ransom, and single point of failure. Risks of information leakage during data dissemination. Laws and regulations bar the rapid sharing[76] of EMR data with other organizations from different countries. However, an organization managing EMR provides its healthcare professionals with instant access to each patient's history, allowing the practice to track patient history and identify patients who are due for visits, tests or screenings.
Storage capacity	Cloud virtually provides flexible and scalable storage to mine, manipulate, and analyze large health datasets.[70] However, Cloud servers may occasionally encounter operational failure causing unavailability of data.	EMR is built with limited storage capacity that accommodates health information from a single institution but not appropriate for continuously streamed data.
Quality of performance	Cloud causes some delays in handling massive numbers of entities depending on the quality of internet connections. However, Cloud facilitates seamless, and timely transmission and sharing[77] health data worldwide.	EMR system enables healthcare professionals to exercise consent exception in an emergency (insufficient time to pursue informed consent from a patient) which improves the quality of care. However, the EMR system provides inadequate interoperability while sharing health data across different health organizations due to their diverse security and access policies.[77]
Cost	Cloud offers a cost-efficient, more effortless scalable environment for storage and deployment of applications.	Most health organizations prefer on-premise storage which costs higher than Cloud-based storage options.

EMR: electronic medical record.

enable the patient to control access to his or her data to a greater extent than others (access control)

- Quality of Performance criteria includes processing speed that indicates the time of uploading, downloading and processing patient health data, interoperability refers to the ability of a storage medium to exchange data among different kinds of systems and software, and data transparency refers to the capability of a storage medium to ensure correctness, the legitimacy of the data source and the capacity to easily access and use data irrespective of source

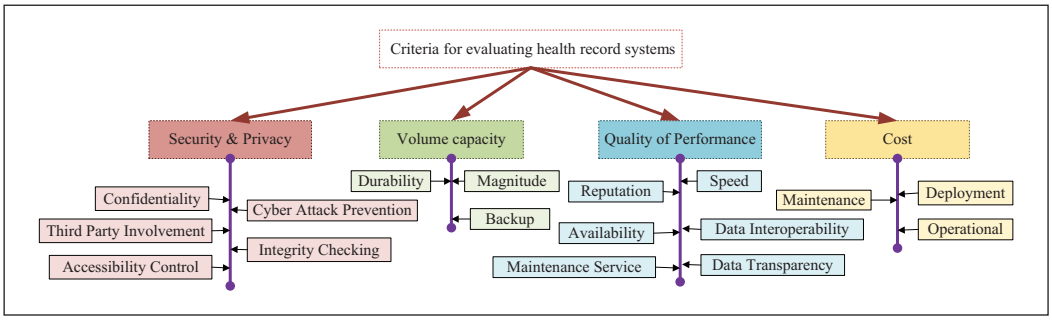

Figure 4. The hierarchical representations of health repositories evaluating standards.

and location. The storage organization's reputation represents the past history of the storage repository manager's ratings from bodies such as investors, customers, suppliers, employees, regulators, politicians, non-governmental organizations for its service.

- Storage capacity indicates the capacity of a storage repository to backup and archive data, and durability refers to the capacity for a repository to protect patient's health-related data from bit rot, degradation, and other long term corruptions.
- Cost involves deployment, and maintenance that indicates the action taken by a storage medium to retain or restore its service or machine, equipment, and service.

Tables 1 and 2 describe the strengths and weaknesses of four storage repositories against the four major criteria: security and privacy, quality of performance, and cost. Table 3 presents the assessment of five health data repositories against the sub-criteria under four major criteria. Values range from [1 to 5] for each feature. The ratings derive from the three of the authors' own judgements, as IT experts. Future research is planned to source the ratings from a wider group of IT experts and healthcare professionals. The single rating for criteria is calculated by averaging the ratings provided by the three authors. The rating in favor of a criterion for a health data storage repository is estimated according to equation (1).

$$r_{i,j} = \frac{\sum_{k=1}^{n} x_k}{n} \text{where } x_k = \frac{\sum_{c=1}^{m} r_c}{m} \quad (1)$$

$r_{i,j}$ indicates rate against a criterion i for a storage medium j. x_k indicates a rate given by each of m experts against the criterion i for the storage medium j. r_c represents rating given by an expert against sub-criteria. The radar graph depicted in Figure 5 visualizes the strength of five health repositories with respect to the four criteria.

The association between data features and repository evaluation standards

The proposed method aims to transfer medical data, particularly patient-generated health data to one of the health record systems that appropriately reflect the data requirements or user's preferences. Health data requirements outlined above are associated with storage evaluation criteria in a one to many relation where some associations are strong, and some are weakly related. Figure 6 shows the relationship between data storage requirements and storage evaluation criteria. The data

Table 3. Rating five health repositories against four criteria.

Evaluation criteria	Sub criteria	BC EHR	Cloud eHealth	EMR	PHR	EHR
Security and privacy	To what extent can the storage repository ensure data integrity?	4.65	2.85	2.90	3.40	2.85
	To what extent is the storage repository available 24/7?					
	To what extent can a third party access data?					
	To what extent can the storage repository withstand Ransomware, DoS, Insider Attacks?					
Storage capacity	To what extent can the repository support storage for Big data?	1.67	4.42	3.1	1.50	2.77
	To what extent can the repository facilitate processing of Big data?					
	To what extent can the repository facilitate storage for continuously streamed data?					
QoP	How fast can data uploading be?	2.00	3.67	3.52	3.17	3.52
	How fast can data retrieval be?					
	How fast can data processing be?					
Cost	How low is deployment cost?	3.83	4.05	3.44	1.73	4.40
	How low are maintenance costs ?					
	How low are service costs?					

BC EHR: Blockchain electronic health record; EMR: electronic medical record; PHR: personal health record; QoP: quality of performance.

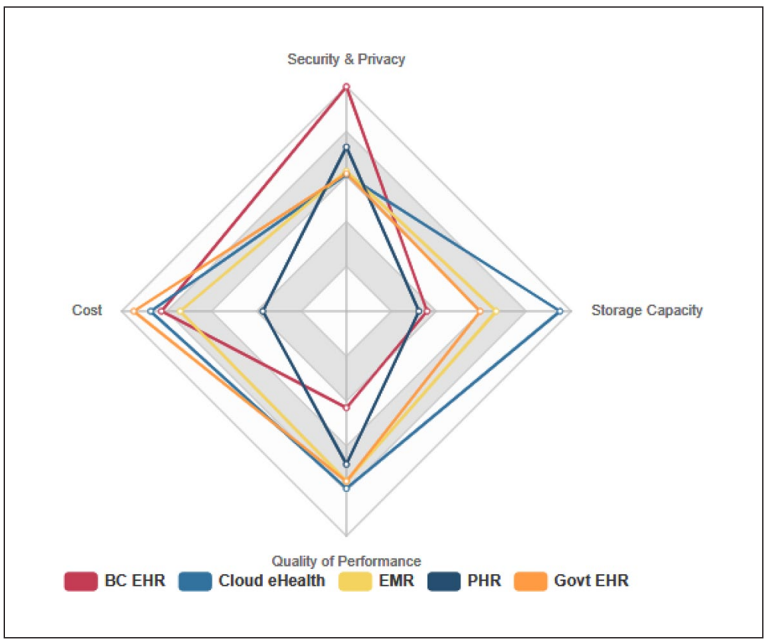

Figure 5. The strength of five health repositories in favor of four criteria.

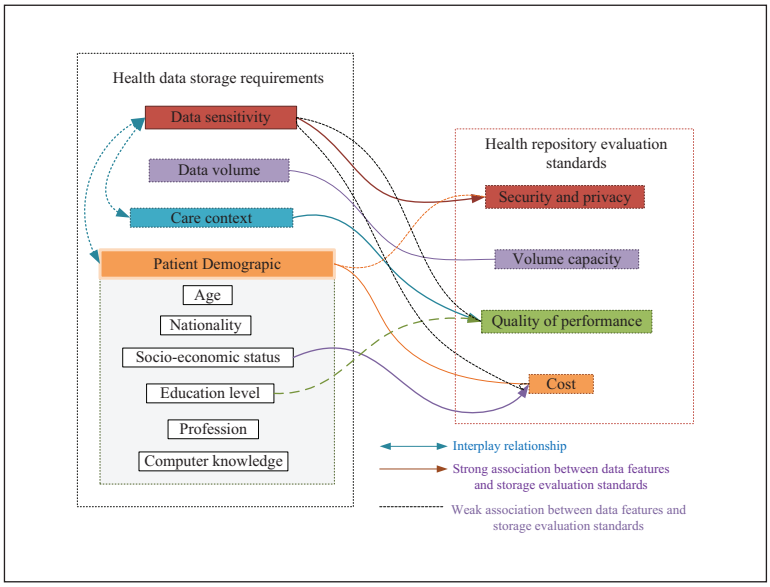

Figure 6. The mapping between data storage requirements and storage medium evaluation criteria.

features have interrelationships and effect one another. For example, a data block considered highly confidential may be submitted in plaintext format to a health record system for rapid processing. At the same time, a patient's demographic features (such as high social status or public profile) can make relatively low confidential data highly sensitive. Demographic data, such as education or technical experience, is likely to positively influence patient privacy concerns. So he or she can choose a particular storage repository that protects health data confidentiality.

Machine learning

This section describes how a training dataset that represents the mapping of the different data blocks to various health repositories is created for the machine learning algorithms. The system adopts supervised learning for dynamically suggesting health repositories for a particular data block. For this reason, we need to generate a training dataset with the label for each instance of the dataset.

Mapping between health data block and health repositories

We have taken into account a few methods to determine the class label (health repository) for each entity in the dataset. The approach includes correlation coefficient analysis, distance measurement, heuristic rules designed by healthcare professionals, and user preferences.

- *Mapping using correlation coefficient*: We specify several features for each data block to be assigned to a health repository. Some of these features are directly related to the data block, and some features are associated with the patient. The features might include the level of sensitivity, the magnitude or volume of data, data type, medical care context, and patient demographic information (nationality, profession, education and socio-economic status and income level).

Table 4. Relation between data storage requirements and repository evaluation criteria.

Data requirements	Remarks	Storage evaluation criteria
Data sensitivity	In general, all medical data is not labeled with the same level of sensitivity. For instance, ECG data for a person with a high public profile may be sought by so many commentators that the level of security required is extreme. The data sensitivity is intimately associated with the security and privacy capacity of a storage repository.	Security and privacy
Data volume	Large volumes of data should be channeled to a storage medium with high capacity, and low volume of data can be stored in a storage medium with lower capacity. So, data volume is linked to the storage capacity of a repository.	Storage capacity
Care context	Access to health data might tolerate a certain amount of delay depending on the types of care. For instance, the delay can be tolerated in normal care setting but not in an emergency setting. So, different levels of QoP need to be ensured on the basis of care status.	QoP (Quality of performance)
Socio-economic	A patient's demographic profile plays a role in deciding how much privacy, security and quality of performance a patient requires when selecting a health repository. In developing countries, the socio-economic status of a patient may be closely linked to the costs associated with a repository	Cost

Firstly, four features named data sensitivity, volume, medical care context and consumer's income level have a linear correlation with four attributes of health repositories: security and privacy, storage capacity, quality of performance, and costs associated with adopting a health record system. The association between data features and the criteria of the health repository is explained in Table 4.

Each data feature shown in Table 4 is assigned a value in the range [1 to 5]. For example, a specific health data block assigned to "higher confidential" has value 5 for the sensitivity feature, and medium one has value 4 for that attribute. Similarly, a data block with high magnitude has value 5 for the data volume feature and so on.

Pearson correlation coefficient is calculated to label an instance provided that other features do not have an impact on deciding the health repositories.

The Pearson correlation coefficient is presented in equation (2). We calculate the correlation coefficient between four features of a data block and four evaluation criteria of all health repositories. The repository with the highest Pearson coefficient with respect to features of a data block was considered best suited for that data block.

$$r_i = \frac{\sum_{j=1}^{m}(x_j - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{j=1}^{m}(x_j - \overline{x})^2}\sqrt{\sum_{j=1}^{m}(y_j - \overline{y})^2}} \tag{2}$$

Assuming that r_1, r_2, r_3, r_4 and r_5 are calculated between the set of data storage requirements (D) and the evaluation criteria of EHR (S_1), PHR (S_2), Cloud eHealth (S_3), Blockchain (S_4), and EMR (S_5), respectively.

The recommended storage (S_i) for a particular instance of the dataset D is estimated using equation (3):

$$S_i = \max(r_1, r_2, \ldots, r_n) \tag{3}$$

where $i = 1, 2, \ldots, n$ and $j = 1, \ldots, m$. n is number of storage mediums and m is the number of criteria.

However, if any instance with identical values for all the features appears in the dataset, the Pearson correlation coefficient cannot be calculated to discover the best-suited repository for that instance. In such cases, the Euclidean or Manhattan distance between data storage requirements and the criteria of all repositories is calculated to determine the best-fitted repositories for storing the data block.

Assuming that, the recommended repository (S_i) for a particular instance I that has identical value for all the features can be found using equation (4), and (5)

$$r_i = \min(\sum_{i=j}^{m} | x_{i,j}, -y_{i,j} |) \tag{4}$$

$$S_i = \min(r_1, r_2, \ldots, r_n) \tag{5}$$

where $i = 1, 2, \ldots, n$ and $j = 1, \ldots, m$

- *Mapping using experts' knowledge*: Secondly, healthcare professionals' decision, user's preferences and other features such as normal or abnormal patterns, patient profile status and other demographic factors can dominate in selecting an appropriate health data storage repository. For instance, unusual heart patterns in cardiovascular patients are likely to be clinically useful and should be stored in such a repository that enables rapid access by healthcare professionals. Data that is within normal ranges can often be stored in a low secured or inexpensive storage repository because it is unlikely to be of interest to future health care professionals, though may have minimal utility for future health research. Additionally, selection of health repositories also relies on the data block genre. For instance, in many countries such as Australia, USA, Europe, data related to a cancer diagnosis is uploaded to a cancer registry database.

Heuristic rules can best address the contexts discussed above. Patient specific heuristic rules can enable high-level user preferences (healthcare professionals) to be easily specified. The heuristic rules are set to take precedence over the correlation analysis method for nominating the most appropriate storage repository for a data block. Sample rules representing the authors' preferences are as follows.

1. **if** *Data* reflects *Normal patterns* **and** *Data volumes* are *high*, **then** *Storage medium* is *Cloud eHealth*
2. **if** *Data reflects Normal patterns* **and** *Data volume* is *small*, **then** *Storage medium* is *PHR*
3. **if** *Data reflects abnormal patterns*, **then** *Storage medium* is *EMR*
4. **if** *Public profile* is *high* **and** *Care context* is *normal*, **then** *Storage medium* is *Blockchain eHealth*
5. **if** *Public profile* is *high* **and** *Care context* is *emergency*, **then** *Storage medium* is *EMR*
6. **if** *Data genre* is *cancer*, **then** *Send a copy of data to the Cancer registry*

- *Mapping decided by users*: The decisions regarding how data is to be disseminated among multiple storage managers should be made in accordance with a user's preference. Different

Table 5. The sample training dataset for machine learning.

Data block	Sensitivity	DV	Care context	SES	PP	Data type	Storage medium
Block 1	4	2	1	4	Low	Normal	EMR
Block 2	1	4	5	4	Low	Normal	Cloud eHealth
Block 3	1	1	1	1	Low	Normal	BC_EHR
Block 4	2	2	2	2	Low	Abnormal	EMR

DV: data volume; SES: socio-economic status; PP: public profile.

users may have quite different choices regarding privacy, and the preferences may change depending on a diverse range of contexts.[78] The patient is expected to choose his or her health record systems depending on his or her health condition, demographic data (age, nationality), social profile or status, data type, sensitivity, and significance of data. For example, one user might give preference to having their vital signs data stored on healthcare providers storage for rapid access in an emergency setting but not in other contexts. Another patient may be a government employee who is reluctant to have their psychiatric record on a government-managed EHR. A patient with a low public profile may not need a level of high security for his or her ECG data. In contrast, a celebrity with a high profile may prefer his or her ECG data to be stored solely in a Blockchain. Most people may reveal their blood group, whereas individuals with a high public profile may be more reluctant to do so. Further, an individual's preference regarding the level of privacy and security may change over time. A young person may desire higher security and privacy than a palliative patient. The present study aims to incorporate user preferences regarding health record systems.

Generating synthetic data

A training dataset is constructed using the above mentioned Pearson correlation coefficient, Manhattan distance and heuristic rules to train a classifier. Table 5 represents a sample training set where the data block features include sensitivity level, data volume (DV), medical care context, and socio-economic status (SES), public profile (PP), data type. These features' value range from 1 to 5. The class label for the first and second instance is fixed by using the Pearson correlation. Public profile and data type for these two instances are overridden because public profile value is low and data type is normal. In the fourth instance, data type is abnormal, which overrides the role of other features and the health data block is directed to healthcare professional providing Electronic Medical Record for having rapid health services.

We selected a supervised machine learning model over a rule-based expert system for suggesting health repositories for the following reasons. Large numbers of rules are required to be generated as features of data storage requirements increase. The machine learning algorithm can learn user's preferences about healthcare record systems under a diverse range of contexts. The supervised learner is trained with a pre-defined preference data to channel health data to available health repositories automatically. User's preferences cannot be encoded using generic rules because the user's preferences about health repositories are subjective and vary from individual to individual.

Rule-based AI (Artificial Intelligence) can infer conclusions in clearly defined and bounded situations. In contrast, ML (Machine Learning) can generalize conclusions along multiple dimensions, which can model more sophisticated behaviors than a sample matching. Selection of a particular storage repository for health data is stochastic. The healthcare professionals or users may prefer a health storage system under specific data storage requirements which might be challenging

Table 6. The confusion matrix.

	Condition positive	Condition negative
Predicted condition positive	True positive	False negative
Predicted condition negative	False positive	True negative

to represent using rules. A machine learning algorithm can produce the best-fitted output for the cases mentioned above.

Train classifiers

We formed a training dataset using the methods described in the *machine learning* section above. A sample of such training data is illustrated in Table 5. We assume that we have different data blocks that can contain discharge summaries, pathological results, psychiatric evaluations, and medical images or data continuously streamed from wearable sensors. In this experiment, our target is to investigate how well the classifiers learn the data distribution rules.

The four separate training datasets have size 500, 1000, 1500, and 2000 instances, respectively. The four datasets have been fed into five different classifiers to study the feasibility of a machine learning algorithm in selecting an appropriate storage medium. Five different classifiers trained here are Multilayered Perceptions (MLP), Random Forest (RF), J48, K-nearest neighbor (IBK), and Naive Bayes (NB). The classifiers are trained using a variable size of the synthetic dataset in Weka ToolKits[79] and evaluated in terms of the following metrics.

- Confusion matrix[80] shown in Table 6, also called contingency table, describes the results of classification. The upper left corner True positive is the number of entities being classified as true positive while those were true. The lower right cell False-positive represents the number of samples being classified as false negative while they were false. False-negative indicates the number of entities being classified as true although those were false. False-positive represents the number of entities being classified as true, although those were true.

$$\text{accuracy} = \frac{\Sigma\,\text{True positive} + \Sigma\,\text{True negative}}{\Sigma\,\text{total samples}}$$

$$\text{Precision} = \frac{\Sigma\,\text{True Positive}}{\Sigma\,\text{Predicted condition positive}}$$

$$\text{Recall} = \frac{\Sigma\,\text{True positive}}{\Sigma\,\text{condition positive}}$$

- MSE is measured by taking the square average of the difference between the data's original and predicted values. RMSE (Root mean square error) is the normal variance of the errors that occur while predicting on a dataset. This tests about how far from the actual output the forecasts were. RMSE is defined in mathematical terms as follows.

$$\text{RMSE} = \sqrt{\frac{1}{\text{N}}\sum\nolimits_{\text{i}=1}^{\text{n}}(\text{actual values} - \text{predicted values})^2}$$

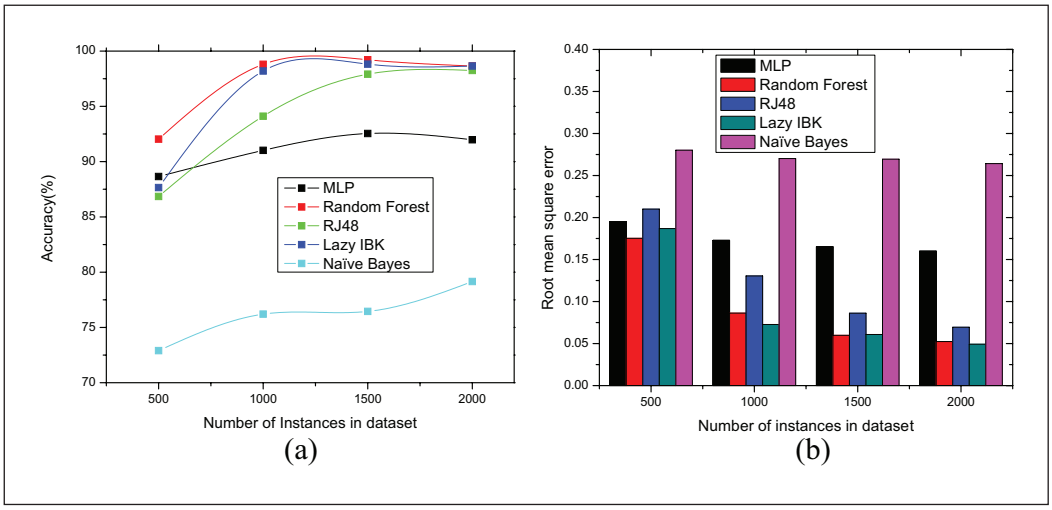

Figure 7. 10-fold cross validation. (a) Classifier's accuracy. (b) Classifier's root mean square errors.

- Receiver The Operating Characteristic Curve (or ROC Curve) is a plot of the true positive rate against the false-positive rate for the various possible diagnostic test cutpoints. ROC reveals the trade-off between sensitivity and specificity (a decrease in specificity will follow any rise in sensitivity). The more the curve follows the left border and the more closely the curve follows the top border of the ROC space, the more accurate the test.

The accuracy and ROC curve for both 10-fold cross-validation and percentage split are illustrated in Figures 7 to 10, respectively. The graph depicted in Figure 7(a) shows that Random Forest and Lazy IBK (K-nearest neighbor) classifiers offer higher accuracy with an increasing number of instances of the dataset in 10-fold cross-validation method. All the classifiers showed higher accuracy for the dataset having 1500 tuples because this dataset contains a balanced ratio of every class. Random Forest shows the highest accuracy of 99.21% and the next best classifier for this dataset is IBk that showed an accuracy of 98.82%. In contrast, all the classifiers with the dataset that has 2000 tuples showed a slightly lower accuracy largely because the dataset is imbalanced. The root mean square errors for 10-fold cross-validation is presented in Figure 7(b) where Random Forest and IBK are showing less RMSE in comparison to other classifiers.

On the other hand, the percentage split results depicted in Figure 8 present comparatively lower accuracy than 10-fold cross-validation. In percentage split, the dataset is partitioned into a training set (80%) and test set (20%) and classifier are trained once then all the classifiers showed low accuracy and high RMSE depicted in Figure 8(b).

The graph depicted in Figures 9 and 10 shows the Recall vs Precision and ROC curve for different classes in 10-fold cross validation method.

Deep learning is a subset of machine learning in artificial intelligence (AI). The deep learning networks are capable of learning unsupervised data that is unstructured or unlabelled. The datasets for rapidly recommending health repositories can be unstructured and unlabelled in a real situation. So, we adopted a deep learning approach for investigating the accuracy for our synthetic datasets. The synthetic dataset is fed into a deep learning model, and the model shows around 89% accuracy. The deep learning approach is modeled using Python with Keras framework. The

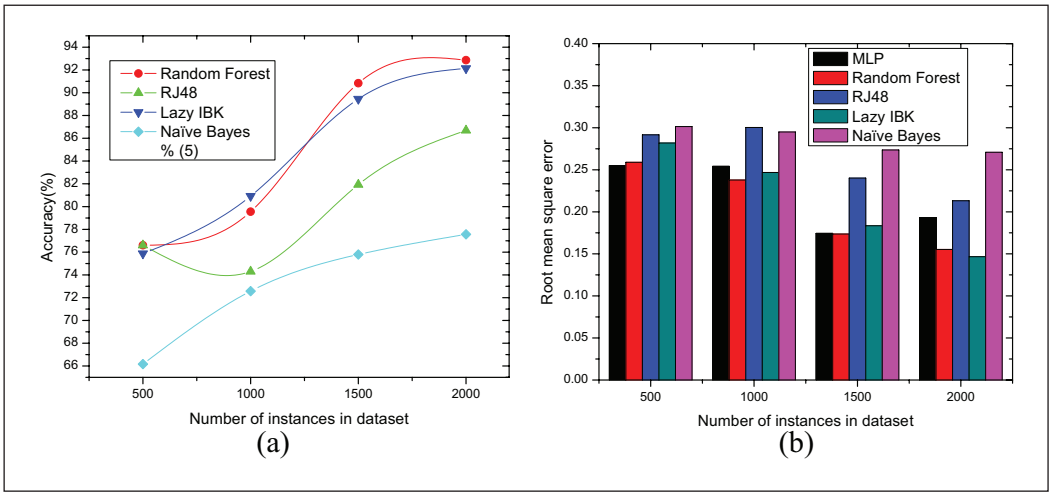

Figure 8. Percentage split (20% testset from training dataset). (a) Classifier's accuracy. (b) Classifier's root mean square errors.

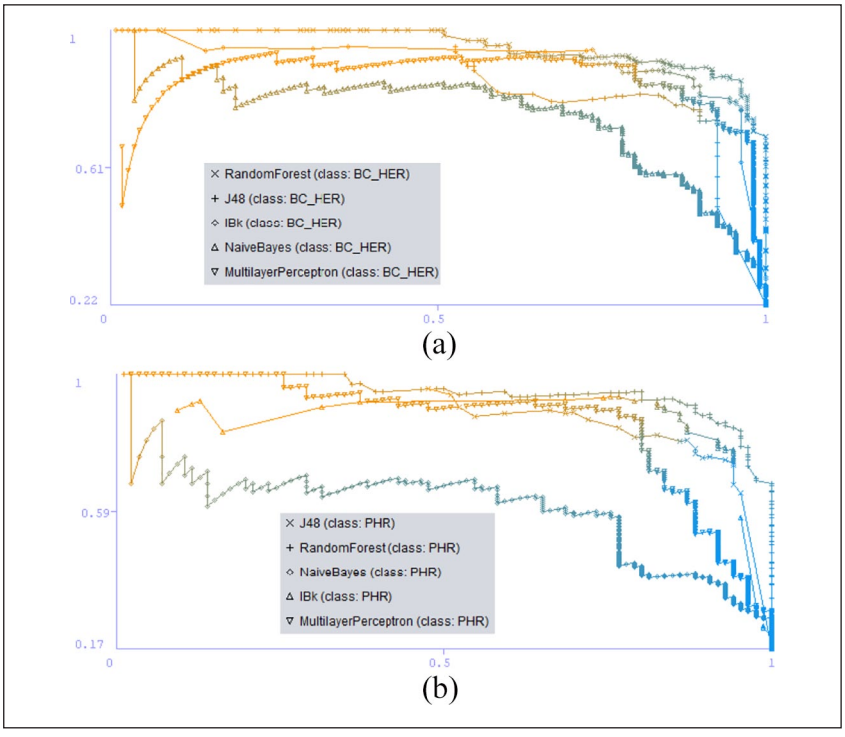

Figure 9. Recall versus precision. (a) Blockchain electronic health record. (b) Personal health record.

data is based on seven input diameters with multiple classes. The model has three hidden layers where the first hidden layer has 100 output nodes that take input from seven input diameters, and the last hidden layer has five output nodes. The model is trained using 100 number of epochs, and

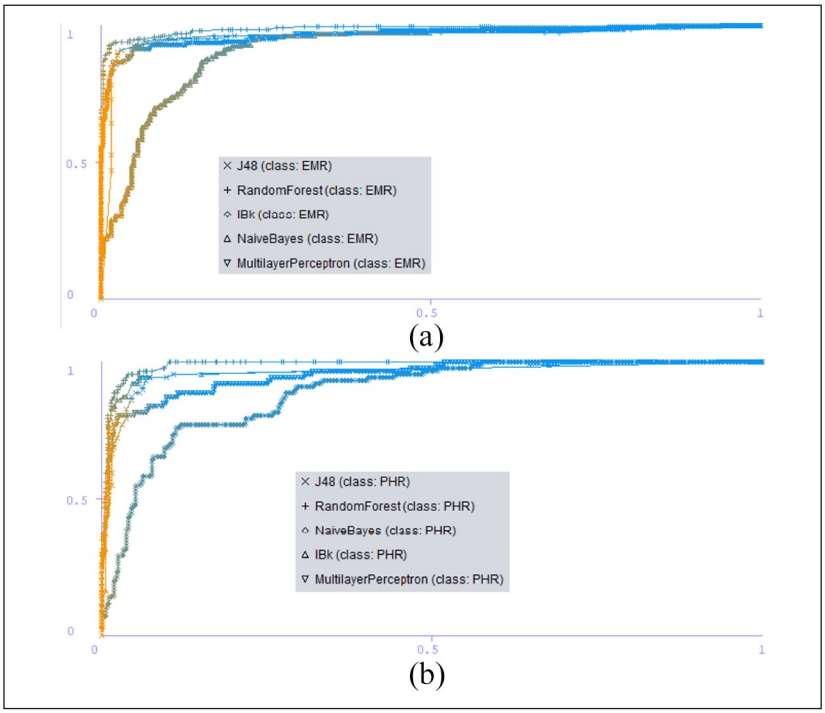

Figure 10. ROC curve. (a) Electronic medical record. (b) Personal health record.

Table 7. Accuracy of the deep learning model.

	Precision	Sensitivity or recall	f1-score
Cloud eHealth	0.93	1.00	0.96
PHR	1.00	1.00	0.96
EHR	1.00	0.93	0.96
EMR	0.85	0.96	0.90
Blockchain eHealth	1.00	0.92	0.96
Accuracy			0.89

the batch size is set 8. The Confusion matrix and the accuracy in terms of different metrics are presented in Table 7. Figure 11 shows the training loss and accuracy of the sample dataset where *X*-axis indicates the number of epochs and *Y*-axis indicates loss or accuracy.

The accuracy level of the classifier for the dataset demonstrates the feasibility of using machine learning or deep learning to learn the mapping between health storage mediums and a health data block.

With the rapid growth in the volume of health data that needs to be stored and accessed globally, this machine learning model could be an essential tool for improving the storage and access arrangements for the future. This method has the potential to enhance the consumer's ability to manage their health data storage and access, while also ensuring data stores are manageable from a size perspective. The ML model can assist with determining the most "fit for purpose" storage solution for different data assets.

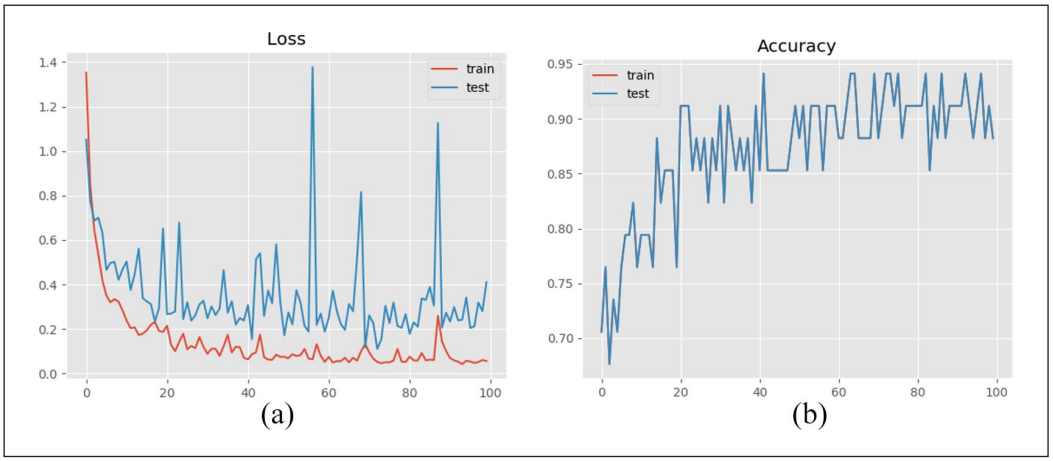

Figure 11. Result of deep learning model. (a) Training loss. (b) Training accuracy.

Adoption of new health repositories

In this paper, seven different health record systems are described as potential repositories for patient-generated health data. Five of the most prevalent repositories were investigated. With the advancement of medical technology, variations of health data are expanding, and new types of health record system can be expected to emerge. The proposed system supports new data variation and new health record in the following ways. First, the system asks IT and healthcare manager or professionals' rating for the latest health record in favor of a few criteria illustrated in Table 3. Secondly, the system revises the complete training dataset to relabel the instances. The addition of a new instance does not change the label of the old instances. The class label of the newly added instance is only required to be determined. The system needs only to re-train machine learning algorithms with the updated dataset.

Conclusion

As more repositories become available for preserving health data, patients will need to select the desired repository. Patients can be expected to avoid choosing a single repository for all their health data because their context of treatment, the pattern of data, legal constraints or personal preferences may change. Therefore, a selection algorithm needs to be developed to automate the storage decision. This is particularly important for continuously streamed health data. In addition, choosing the correct repository is complicated and needs professional knowledge of storage features for interoperability, data security and privacy, infrastructure availability, and regulatory issues. Our proposal to disseminate health data among various vendors will prevent the loss of confidentiality and ensure the privacy of medical records if they are stored in one repository. The automated storage recommendation model presented here can allocate health data blocks to a storage medium taking into account data types, data sensitivity, significance and QoP, patient safety and privacy required depending on the profile of an individual.

Acknowledgements

The authors are grateful to the anonymous reviewers for their valuable comments, and suggestions that improved the quality of the article.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The author(s) received financial support from the School of Engineering, Information Technology and Physical Sciences, Federation University Australia for conducting research, and publication of this article.

ORCID iD


Md Ashraf Uddin https://orcid.org/0000-0002-4316-4975

References

1. Plastiras P and O'Sullivan D. Exchanging personal health data with electronic health records: a standardized information model for patient generated health data and observations of daily living. *Int J Med Inform* 2018; 120: 116–125.
2. Cortez A, Hsii P, Mitchell E, et al. *Conceptualizing a data infrastructure for the capture, use, and sharing of patient-generated health data in care delivery and research through 2024 (white paper)*. Washington, DC: Office of the National Coordinator for Health Information, 2018.
3. Chung CF, Dew K, Cole A, et al. Boundary negotiating artifacts in personal informatics: patient-provider collaboration with patient-generated data. In: *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*, 2016, pp. 770–786.
4. Lordon RJ, Mikles SP, Kneale L, et al. How patient-generated health data and patient-reported outcomes affect patient–clinician relationships: a systematic review. *Health Inform J* 2020; 26(4): 2689–2706.
5. Demiris G, Iribarren SJ, Sward K, et al. Patient generated health data use in clinical practice: a systematic review. *Nurs Outlook* 2019; 67(4): 311–330.
6. Rovner J. Health insurance portability and accountability act (HIPAA). In: *Health care policy and politics A to Z*. CQ Press, 2009, pp. 93–96.
7. Bennett B, Carney T, Chiarella M, et al. Australia's national registration and accreditation scheme for health practitioners: a national approach to polycentric regulation. *Sydney L Rev* 2018; 40: 159.
8. Mulder T and Tudorica M. Privacy policies, cross-border health data and the gdpr. *Inform Commun Technol Law* 2019; 28(3): 261–274.
9. Rantos K, Drosatos G, Demertzis K, et al. Blockchain-based consents management for personal data processing in the iot ecosystem. *ICETE* 2018(2): 738–743.
10. Harison E. Who owns enterprise information? data ownership rights in europe and the us. *Inform Manage* 2010; 47(2): 102–108.
11. Zheng X, Mukkamala RR, Vatrapu R, et al. Blockchainbased personal health data sharing system using cloud storage. In *2018 IEEE 20th international conference on e-health networking, applications and services (Healthcom)*. IEEE, pp. 1–6.
12. Albahri A, Zaidan A, Albahri O, et al. Real-time fault-tolerant mhealth system: comprehensive review of healthcare services, opens issues, challenges and methodological aspects. *J Med Syst* 2018; 42(8): 137.
13. Isern D and Moreno A. A systematic literature review of agents applied in healthcare. *J Med Syst* 2016; 40(2): 43.
14. Vaidehi V, Vardhini M, Yogeshwaran H, et al. Agent based health monitoring of elderly people in indoor environments using wireless sensor networks. *Procedia Comput Sci* 2013; 19: 64–71.
15. Record MH. My health record, 2019. https://www.myhealthrecord.gov.au/.
16. Microsoft. healthvault, https://www.healthvault.com/en-us/ (accessed July 2019).
17. Mandel JC, Kreda DA, Mandl KD, et al. Smart on fhir: a standards-based, interoperable apps platform for electronic health records. *J Am Med Inform Assoc* 2016; 23(5): 899–908.

18. ios -health. https://www.apple.com/au/ios/health/ (accessed July 2019).
19. Christidis K and Devetsikiotis M. Blockchains and smart contracts for the internet of things. *IEEE Access* 2016; 4: 2292–2303.
20. Uddin MA, Stranieri A, Gondal I, et al. Continuous patient monitoring with a patient centric agent: a block architecture. *IEEE Access* 2018; 6: 32700–32726.
21. Hang L, Choi E and Kim DH. A novel emr integrity management based on a medical blockchain platform in hospital. *Electronics* 2019; 8(4): 467.
22. Uddin MA, Stranieri A, Gondal I, et al. Blockchain leveraged decentralized iot ehealth framework. *Internet Things* 2020; 9: 100159.
23. McFarlane C, Beer M, Brown J, et al. *Patientory: a healthcare peer-to-peer emr storage network v1*. Addison, TX: Entrust Inc, 2017.
24. Katuwal GJ, Pandey S, Hennessey M, et al. Applications of blockchain in healthcare: current landscape & challenges. *arXiv preprint arXiv*:181202776, 2018.
25. Hasselgren A, Kralevska K, Gligoroski D, et al. Blockchain in healthcare and health sciences–a scoping review. *Int J Med Inform* 2019; 134: 104040.
26. Mackey TK, Kuo TT, Gummadi B, et al. 'fit-for-purpose?'– challenges and opportunities for applications of blockchain technology in the future of healthcare. *BMC Med* 2019; 17(1): 68.
27. Mayer AH, da Costa CA and Righi RD. Electronic health records in a blockchain: a systematic review. *Health Inform J* 2019; 26(2): 1273–1288.
28. Carter M. Introducing health information privacy in victoria. *Priv Law P Rpr* 2000; 7(7): 130–131.
29. Coebergh JW, Van Den Hurk C, Rosso S, et al. Eurocourse lessons learned from and for population-based cancer registries in europe and their programme owners: improving performance by research programming for public health and clinical evaluation. *Eur J Cancer* 2015; 51(9): 997–1017.
30. Australian cancer database (acd). https://www.aihw.gov.au/about-our-data/our-data-collections/australian-cancer-database.
31. Ringland C, Arkenau HT, O'Connell D, et al. Second primary colorectal cancers (spcrcs): experiences from a large australian cancer registry. *Ann Oncol* 2010; 21(1): 92–97.
32. Ko SY, Jeon K and Morales R. The hybrex model for confidentiality and privacy in cloud computing. *HotCloud* 2011; 11: 8–8.
33. Zhang H, Ye L, Du X, et al. Protecting private cloud located within public cloud. In: *Global communications conference (GLOBECOM), 2013 IEEE*. IEEE, pp. 677–681.
34. Stantic D and Jo J. Detecting abnormal ecg signals utilising wavelet transform and standard deviation. In: *Proceedings of world academy of science, engineering and technology*. 71, World Academy of Science, Engineering and Technology (WASET), p. 208.
35. Stranieri A and Balasubramanian V. Remote patient monitoring for healthcare: a big challenge for big data. In: *Managerial perspectives on intelligent big data analytics*. IGI Global, 2019. pp. 163–179.
36. Al Ghamdi A and Thomson T. The future of data storage: a case study with the saudi company. *J IEEE* 2018; 6(1): 1.
37. Ruiz-Alvarez A and Humphrey M. A model and decision procedure for data storage in cloud computing. In: *Proceedings of the 2012 12th IEEE/ACM international symposium on cluster, cloud and grid computing (ccgrid 2012)*. IEEE Computer Society, pp. 572–579.
38. Ruiz-Alvarez A and Humphrey M. Toward optimal resource provisioning for cloud mapreduce and hybrid cloud applications. In: *Proceedings of the 2014 IEEE/ACM international symposium on big data computing*. IEEE Computer Society, pp. 74–82.
39. Yoon MS and Kamal AE. Optimal dataset allocation in distributed heterogeneous clouds. In: *2014 IEEE globecom workshops (GC Wkshps)*. IEEE, pp. 75–80.
40. Uddin MA, Stranieri A, Gondal I, et al. A patient agent to manage blockchains for remote patient monitoring. *Stud Health Technol Inform* 2018; 254: 105–115.
41. Yang Y and Chen T. Analysis and visualization implementation of medical big data resource sharing mechanism based on deep learning. *IEEE Access* 2019; 7: 156077–156088.
42. Andy YYL, Shen CP, Lin YS, et al. Continuous, personalized healthcare integrated platform. In: *TENCON 2012 IEEE region 10 conference*. IEEE, pp. 1–6.

43. Peleg M, Shahar Y, Quaglini S, et al. Mobiguide: a personalized and patient-centric decision-support system and its evaluation in the atrial fibrillation and gestational diabetes domains. *User Model User Adap Interact* 2017; 27(2): 159–213.
44. Martinez VI, Marquard JL, Saver B, et al. Consumer health informatics interventions must support user workflows, be easy-to-use, and improve cognition: applying the seips 2.0 model to evaluate patients' and clinicians' experiences with the conduit-hid intervention. *Int J Hum Comput Interact* 2017; 33(4): 333–343.
45. Liu LS, Shih PC and Hayes GR. Barriers to the adoption and use of personal health record systems. In: *Proceedings of the 2011 iconference*, 2011, pp. 363–370.
46. Hohemberger R, da Roza CE, Pfeifer FR, et al. An approach to mitigate challenges to the electronic health records storage. *Measurement* 2020; 154: 107424.
47. Busis NA. How can i choose the best electronic health record system for my practice? *Neurology* 2010; 75(18 Supplement 1): S60–S64.
48. Weathers AL and Esper GJ. How to select and implement an electronic health record in a neurology practice. *Neurol Clin Pract* 2013; 3(2): 141–148.
49. Hart EM, Barmby P, LeBauer D, et al. Ten simple rules for digital data storage. *PLoS Comput Biol* 2016; 12(10): e1005097.
50. Wilson G, Bryan J, Cranston K, et al. Good enough practices in scientific computing. *PLoS Comput Biol* 2017; 13(6): e100551.
51. Boonstra A and 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(1): 231.
52. Ross J, Stevenson F, Lau R, et al. Factors that influence the implementation of e-health: a systematic review of systematic reviews (an update). *Implement Sci* 2016; 11(1): 146.
53. Ben-Assuli O. Electronic health records, adoption, quality of care, legal and privacy issues and their implementation in emergency departments. *Health Policy* 2015; 119(3): 287–297.
54. Khan SI and Hoque AS. Towards development of health data warehouse: Bangladesh perspective. In: *2015 international conference on electrical engineering and information communication technology (ICEEICT)*. IEEE, pp. 1–6.
55. Kenny G and Connolly R. Drivers of health information privacy concern: a comparison study, 2016.
56. Joinson AN, Reips UD, Buchanan T, et al. Privacy, trust, and self-disclosure online. *Hum Comput Interact* 2010; 25(1): 1–24.
57. Rahim FA, Ismail Z and Samy GN. A conceptual model for privacy preferences in healthcare environment. In: *The 8th international conference on knowledge management in organizations*. Springer, pp. 221–228.
58. Chang CW, Liu P and Wu JJ. Probability-based cloud storage providers selection algorithms with maximum availability. In: *2012 41st international conference on parallel processing*. IEEE, pp. 199–208.
59. ur Rehman Z, Hussain OK, Parvin S, et al. A framework for user feedback based cloud service monitoring. In: *2012 sixth international conference on complex, intelligent, and software intensive systems*. IEEE, pp. 257–262.
60. Qu L, Wang Y and Orgun MA. Cloud service selection based on the aggregation of user feedback and quantitative performance assessment. In: *2013 IEEE international conference on services computing*. IEEE, pp. 152–159.
61. Lee S and Seo KK. A hybrid multi-criteria decision-making model for a cloud service selection problem using bsc, fuzzy delphi method and fuzzy ahp. *Wirel Pers Commun* 2016; 86(1): 57–75.
62. Halabi T and Bellaiche M. Evaluation and selection of cloud security services based on multi-criteria analysis mca. In: *2017 International conference on computing, networking and communications (ICNC)*. IEEE, pp. 706–710.
63. Halabi T and Bellaiche M. A broker-based framework for standardization and management of cloud security-slas. *Comput Secur* 2018; 75: 59–71.
64. Halabi T, Bellaiche M and Abusitta A. Online allocation of cloud resources based on security satisfaction. In: *2018 17th IEEE international conference on trust, security and privacy in computing and*

communications/12th IEEE international conference on big data science and engineering (TrustCom/BigDataSE). IEEE, pp. 379–384.
65. deBronkart D and Eysenbach G. Gimme my damn data (and let patients help!): The# gimmemydamndata manifesto. *J Med Internet Res* 2019; 21(11): e17045.
66. Faramondi L, Oliva G, Setola R, et al. Iiot in the hospital scenario: Hospital 4.0, blockchain and robust data management. In: *Security and privacy trends in the industrial internet of things*. Springer, 2019, pp. 271–285.
67. Kuo TT, Kim HE and Ohno-Machado L. Blockchain distributed ledger technologies for biomedical and health care applications. *J A Med Inform Assoc* 2017; 24(6): 1211–1220.
68. Wang S, Zhang Y and Zhang Y. A blockchain-based framework for data sharing with fine-grained access control in decentralized storage systems. *IEEE Access* 2018; 6: 38437–38450.
69. Amaraweera SP and Halgamuge MN. Internet of things in the healthcare sector: overview of security and privacy issues. In: *Security, privacy and trust in the IoT environment*. Springer, 2019, pp. 153–179.
70. McGhin T, Choo KKR, Liu CZ, et al. Blockchain in healthcare applications: research challenges and opportunities. *J Network Comput Appl* 2019; 135: 62–75.
71. Yaqoob S, Khan MM, Talib R, et al. Use of blockchain in healthcare: a systematic literature review. *Int J Adv Comput Sci Appl* 2019; 10(5): 644–653.
72. Rao BT. A study on data storage security issues in cloud computing. *Procedia Comput Sci* 2016; 92: 128–135.
73. Onik MMH, Aich S, Yang J, et al. Blockchain in healthcare:challenges and solutions. In: *Big data analytics for intelligent healthcare management*. Elsevier, 2019, pp. 197–226.
74. Sen J. Security and privacy issues in cloud computing. In: *Architectures and protocols for secure information technology infrastructures*. IGI Global, 2014, pp. 1–45.
75. Chernyshev M, Zeadally S and Baig Z. Healthcare data breaches: implications for digital forensic readiness. *J Med Syst* 2019; 43(1): 7.
76. Guo R, Shi H, Zhao Q, et al. Secure attribute-based signature scheme with multiple authorities for blockchain in electronic health records systems. *IEEE Access* 2018; 6: 11676–11686.
77. Azeez NA and Van der Vyver C. Security and privacy issues in e-health cloud-based system: a comprehensive content analysis. *Egyptian Inform J* 2018; 20: 97–108.
78. Trojer T, Katt B, Schabetsberger T, et al. The process of policy authoring of patient-controlled privacy preferences. In: *International conference on electronic healthcare*. Springer, pp. 97–104.
79. Witten IH, Frank E, Trigg LE, et al. Weka: practical machine learning tools and techniques with java implementations (Working paper 99/11). Hamilton, New Zealand: University of Waikato, 1999.
80. Khraisat A, Gondal I, Vamplew P, et al. Hybrid intrusion detection system based on the stacking ensemble of c5 decision tree classifier and one class support vector machine. *Electronics* 2020; 9(1): 173.