

Prospective research topics towards preserving electronic health records in decentralised content-addressable storage networks

Toomas Klementi¹, Kristian Juha Ismo Kankainen¹, Gunnar Piho² and Peeter Ross¹

¹Department of Health Technologies, TalTech, Akadeemia Str 15A, 12618 Tallinn, Estonia

²Department of Software Science, TalTech, Akadeemia Str 15A, 12618 Tallinn, Estonia

Abstract

Data used in the digital economy are increasingly fragmented and scattered across data stores, with little or no interoperability between them. This makes reusing and analysing data difficult. The problem is especially salient in health care, where increasing volumes of data in different electronic formats are available. Reusing these healthcare-related data in various scientific analyses and research will bring about enormous benefits for individual patients and society. However, healthcare data re-use potential remains to be fully exploited because of semantic and technical data heterogeneity, mainly unstructured data nature as well as data sensitivity and protection reasons. In this workshop position paper, we present our preliminary results on the possibility to eliminate the barriers between health data silos using decentralised content-addressable storage networks for preserving electronic health records. We propose an innovative reference architecture for moving health data from the control of healthcare institutions to the complete control of data owners. We briefly evaluate the feasibility of the proposed solution by analysing both primary and secondary use scenarios of health data and sketch some preliminary research topics on data security, privacy, integrity, transparency and interoperability for the primary and secondary uses of electronic health records in decentralised content-addressable storage networks.

Keywords

EHR, electronic health record, secondary use, reference architecture, content-addressable storage networks

1. Introduction

Over the past decades, the healthcare industry has seen the wide adoption and use of electronic health record (EHR) systems. In 2016, the WHO reported that 27 WHO Member States in Europe have a national EHR system, 18 of which have legislation governing its use [1]. This means that an increasing amount of health data is available in electronic format.

Reusing health data could potentially be beneficial not only for providing immediate care for the patient but also for the greater good of society. PwC [2] was likely one of the first pioneers to highlight the importance of the secondary use of health data in medicine and bio-informatics.

HEDA-2022: The International Health Data Workshop, June 19-24, 2022, Bergen, Norway

✉ toomas.klementi@taltech.ee (T. Klementi); kristian.kankainen@taltech.ee (K. J. I. Kankainen);

gunnar.piho@taltech.ee (G. Piho); peeter.ross@taltech.ee (P. Ross)


🌐 taltech.ee/en/emed-lab (T. Klementi)

🆔 0000-0002-8260-526X (T. Klementi); 0000-0002-0551-927X (K. J. I. Kankainen); 0000-0003-4488-3389 (G. Piho);

0000-0003-1072-7249 (P. Ross)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons Licence Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

Routine clinical data are considered precious [3] and their secondary use is seen to be beneficial [4] for policy-makers, public health officers, scientists, clinicians, citizens and industry.

Although the need for health data reuse is widely recognised, actual progress in that area has been moderate. The reasons for this are the vendor-specific proprietary database schemes used by EHR systems, semantic heterogeneity and the sensitive nature of clinical data that sets legal and ethical restrictions on sharing. As stated in a survey from 2018 [5], due to the semantic heterogeneity of health data, we still do not have a unified approach and use the divide-and-conquer method instead. The review [6] conducted a year later concludes that no big-data analytics will happen without optimised data sharing and reuse, which we still lack despite differences in interoperability standards in the medical domain. The authors of [7] conclude that it is not enough to require data to be shared – sharing must be made easy, feasible and accessible, too.

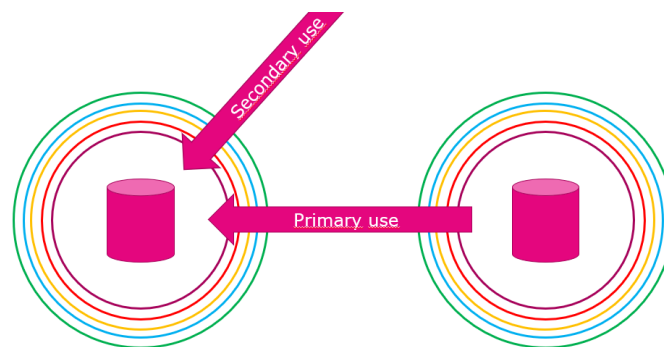


Figure 1: Health data silo problem

As sharing health data is not an easy task, can be costly and may bring about several data privacy and security risks, the institutions hosting the EHR data are generally not motivated and in most cases cannot even share the data. Therefore, all the data that have been stored in EHR systems largely stay inside the boundaries of the institutions that have collected them. This situation can be described (Figure 1) as the 'health data silo problem'.

The silo problem causes issues both in primary and secondary health data usage. In the contemporary and mobile world, it is common practice for a patient to visit multiple healthcare service providers during their lifetime, leaving a digital trace in each of them. This results in patient data being scattered across numerous data silos. Each of these data silos is incomplete and may contain inconsistencies. Keeping the fragmented data synchronised between all the institutions that record EHR data is complex and is often based on national e-health systems like the one in Estonia [8] and those in other countries [9]. At the same time, the lack of common identification schemes makes matching the data in different databases impossible. Even if such a scheme existed, we still face the problem of completeness [10]. Getting a complete and comprehensive picture of patient data is challenging. This need is further emphasised as the focus of health care shifts from medical care towards preventing medical care. The increasing volumes of information are not created by the healthcare institutions themselves; rather they are recorded and gathered by various third parties, e.g. wearable devices, patient portals. Data produced this way are usually stored in third party proprietary databases, thus only increasing

the data silo problem.

We increasingly find ourselves in a situation where the need for a personal data repository under the owner's complete and sole control is obvious. These data can then be shared based on the requirements. Even if we were somehow able to guarantee perfect interoperability, traditional EHR systems are insufficient to deal with it. Furthermore, the entire rationale of hosting personal data at the clinical service provider is questionable. Since the data in the EU, according to the GDPR, legally belongs to the patient [11], why should it be stored at the hospital in the first place? A similar issue has been highlighted by the NGI DAPSI open call for data portability organisers [12]: *"As users of the internet, we sometimes feel as if our data is no longer under our own control. When some online service conveniently stores our data for us without even asking, it often turns out impossible to get it out in one piece later on. That is a rather hefty problem, if you consider how we do online these days. And it becomes increasingly urgent if you want to actually stop using a service – for instance because you realise that the company is seriously violating your privacy, doing business in an unethical way, overcharging you or just because you run into serious limitations of the service"*

A more natural solution proposes to keep the master copy of the complete data set close to the patient. In other words, instead of health institution-controlled EHR data, we propose a holistic, person-controlled, life-long electronic health record containing health data from various sources and any other health-related data of the person. Such a personal life-long electronic health record should be under the complete control of the person, with the ability to share it with healthcare service providers and other actors, when necessary. The challenges to creating such a life record are formidable. Since the main requirement is keeping the data private and under the complete control of the owner, an obvious candidate for such a store would be a local database on a personal smart device or desktop. Indeed, such solutions have been suggested, for instance, by the InteropEHRate project [13]. Although feasible, the proposed solution is far from ideal because of the risk of data loss in the event said personal device gets broken or lost.

Another solution has been proposed by the Solid Project [14]. The idea is to create a protocol that lets individuals store their data on the Internet in personal repositories or 'pods'. The disadvantage of this approach is the dependency on pod providers, over which the owner of the data has no control. Cloud databases are another candidate for the solution, but in the global context, it becomes unfeasible because of the enormous technical difficulties and risks associated with a database of such scale. The question is: who would host such a global health data cloud database? The latter would likely cause an insurmountable problem in terms of political agreement. A unique paradigm is needed – a solution that guarantees privacy, preservation, integrity, scalability and shareability and has no central control. Data silos must be dissolved, and data must be given back to their legal owners. In other words, the data owner must be the sole controller of their own data. Decentralised content-addressable storage networks may offer the solution we are looking for. To democratise data is a big challenge in contemporary society [15].

The rest of the paper is organised as follows. In Section 2, we provide an overview of decentralised content-addressable networks. Then, in Section 3, we present the possible primary and secondary use scenarios of patient EHR data in decentralised content-addressable networks. Next, in Section 4, we phrase the advantages of using EHR data in decentralised content-addressable networks. In Section 5, we draft the possible research topics with respect

to patient EHR data in decentralised content-addressable networks. Finally, Section 6 concludes the paper.

2. Content-addressable networks

One way to overcome the silo problem described is to use the nascent technology of decentralised content-addressable storage. To give an informal description of how it works, consider a paper shredder that works both ways – not only can it split a document into multiple tiny pieces, but it can also 'glue' the document back together given these original pieces. If we have a document that we want to store securely, we will put it through such a shredder and give each piece to a different person to keep. When we need to retrieve the document, we ask each person to give us back their piece and run them through the shredder 'backwards' to retrieve the original document. This guarantees privacy because no person other than the owner ever sees the full document, yet together the group acts as a document keeper.

A decentralised content-addressable network is a peer-to-peer network that operates in a very similar way – a piece of data that needs to be stored on the network is split into multiple pieces and each piece is given to a different node to store. For additional security, each piece can be individually encrypted, thus making it impossible for a node to even see the tiny piece of information it is storing. It is worth noting that nodes on such a network need not be trusted. The only thing expected of them is that they return the pieces they were given when asked to do so – that is, they adhere to the protocol. This is something that can be tested, with failing nodes being excluded from the network.

The network is decentralised in the sense that it lacks central authority. Each individual node is completely independent; it is the network as a whole that exposes the desired characteristics.

There are several decentralised content-addressable storage networks currently available. Examples are IPFS [16], Swarm [17], Storj [18] and SIA [19]. Each of these have their own specifics; some of them probably cannot be called truly decentralised as they are operated centrally by private companies, but their fundamental technical operating principles are similar.

More precisely, a decentralised content-addressable storage network is a peer-to-peer network of nodes running open-source software that implements the network protocol. Each node is assigned an overlay address (as opposed to its underlay address, normally an IP address) from a certain address space, for instance 256-bit integers, and a metric is defined on the node address space. Often the Kademlia [20] XOR metric is used, which defines the distance between two nodes as a result of XOR-ing their overlay addresses and interpreting the result as an unsigned integer. When a piece of data, such as a file, is uploaded to such a network, it is split into multiple small chunks and a hash of each chunk is calculated belonging to the same value space as node overlay addresses. The distance between a chunk and a node can then be defined as the Kademlia distance between the node address and the hash value of the chunk. The network stores each chunk at its closest node. The chunk hashes themselves are organised in a Merkle tree that itself is stored on the network following the same rules as data chunks. The root hash of the tree becomes the hash of the data (file). Therefore, only the node that uploads the data sees the full copy; every other node in the network sees only the chunk(s) it stores. Download happens in the opposite direction – first the chunks comprising the Merkle tree are downloaded,

followed by the download of data chunks and reassembly of the original file. Chunks can optionally be encrypted before hashing, in which case it is practically impossible for any node to discover either the content or the owner of the chunk it stores. If the upload/download node is under the full control of the data owner, the privacy of the data is guaranteed despite being stored in a network of possibly untrusted nodes.

Since data on a decentralised content-addressable network are addressed by their hash, their integrity is guaranteed. Each node downloading a chunk can easily verify the integrity of its content by computing the hash. Content addressability also implies immutability – if the content of a chunk changes, so does its hash, and the network treats it as new data. This creates a natural versioning scheme where existing data is never overwritten. This versioning happens without duplication – chunks with unchanged content are never copied because their hash remains unchanged.

As nodes are free to join and leave the network at will, the network protocol should include measures to guarantee sufficient redundancy. This can be achieved by several means, but a simple method is to define the neighborhood of responsibility for each node as a sphere in the Kademlia address space, with radius r and the node's overlay address as its center. Each chunk would be stored not only at its closest node but also on all nodes belonging to the responsibility neighborhood of its closest node. Given that the churn rate of the network (def) is known, it is possible to lessen the probability of the loss of a data chunk as a result of churn by selecting sufficiently large r .

As long as the hash and encryption key (hashkey) are kept secret, data stored on the network remain under the sole control of the owner. The owner can give access to the data by sharing the hashkey. In the context of secondary use of health data, it is likely that the owner would not want to share the complete data, but perhaps a pseudonymised subset. This can be achieved, for instance, by locally running software that creates a special pseudonymised branch of the full dataset. By sharing the hashkey to that special branch, the user can share their pseudonymised data with anybody they wish. This effectively means sharing only a snapshot of the data. Any future changes to the original data remain private; they will not be visible in the snapshot until explicitly shared.

For such a network to be viable, a solid incentivisation scheme for node operators is needed. The most natural and versatile scheme is financial compensation. Users of the network would pay a certain amount for their data stored by the network, and node operators would get paid for their services. Ideally, any user of the network would also be running a node to securely interact with the network and to offset the costs of using its services at least partially.

It is also possible for any user to run a multitude of nodes as an independent revenue source. Existing implementations of decentralised content-addressable storage networks use either cryptographic tokens or fiat currency as their incentivisation mechanisms.

3. Primary and secondary use

We assume the primary use of health data when the use of data is associated with the health care and treatment of a person. If this data is used for any generalisations and analyses, we are talking about the secondary use of data.

In the primary usage scenario (Figure 2), all personal health data of citizens is stored on the decentralised content addressable network. The hashkey of the data is known only to the owner, thus putting the data under the sole control of the patient. During a visit to a hospital, physician, etc., the patient's data (either in full or in part) are made available to the service provider, either by disclosing the hashkey or possibly by exposing the data through an interface like HL7 FHIR. In the latter case, the hashkey never gets revealed. The service provider can interact with the data by reading them, modifying them or appending new data to them. When changes are written back to the storage, a new version branch is created with a different hash value, as described earlier. The patient confirms the changes by remembering this new hash value. The mechanism also protects the data from modification by unauthorised users should the hashkey be unintentionally disclosed. The hash of unauthorised modifications would not be remembered and the changes effectively lost.

Once the service provider has finished interacting with the data, it will delete them from its computer system and forget the hashkey. Of course, the service provider can and probably will record at least part of the data from the interaction with the patient in its own information system for business or other purposes. This may include personally identifiable data as long as it complies with legal restrictions. Effectively, the decentralised personal data act as a digital notebook the patient always carries. It is natural to assume that data contained in such a notebook is not limited to specific health data but includes all data on the subject gathered from various sources over the course of their lifetime. Instead of a health record, such data set could be called the life record of the person.

Secondary use (Figure 2) of data is any use of data outside the primary (original) use [21]. Secondary use of clinical data is a fast-growing field towards high-quality health care, improved healthcare management, reduced healthcare costs, population health management and effective clinical research [22]. Smart homes [23], clinical research [24] [25], diagnosis [26], supporting elderly persons [27] and real-time fraud detection in a medical benefit [28] are some of the few reported cases in which clinical and health data are reused.

However, according to a systematic analysis of literature conducted by Edmondson and Reimer [29], the four main challenges related to the secondary use of medical data are as follows: (a) insufficient data quality (completeness, correctness and currency); (b) data processing issues (extraction and transformation); (c) compliance with data protection requirements; and (d) limited data generalisation possibilities.

A case study [30] conducted in University Medical Center Utrecht (UMCU) in the Netherlands concludes that a modern technical infrastructure for reusing EHR data requires features for supporting (1) integrating data sources, (2) data preprocessing, (3) data storage, (4) collaboration and documentation, (5) integration with various software and tooling packages, (6) repeatability enhancement, (7) privacy and security enhancement, (8) data processing automation and (9) applications analysis.

By default, all personal health data in decentralised content-addressable storage networks are private, since only the owner knows the hashkey. If the data owner wants to share data, possibly for profit or charitable or other reasons, it is enough to make the relevant hashkey known to the third parties. However, it might not always be the case that a person would like to share identifiable personal data, even though a person has the full right to publish them according to data legislation.

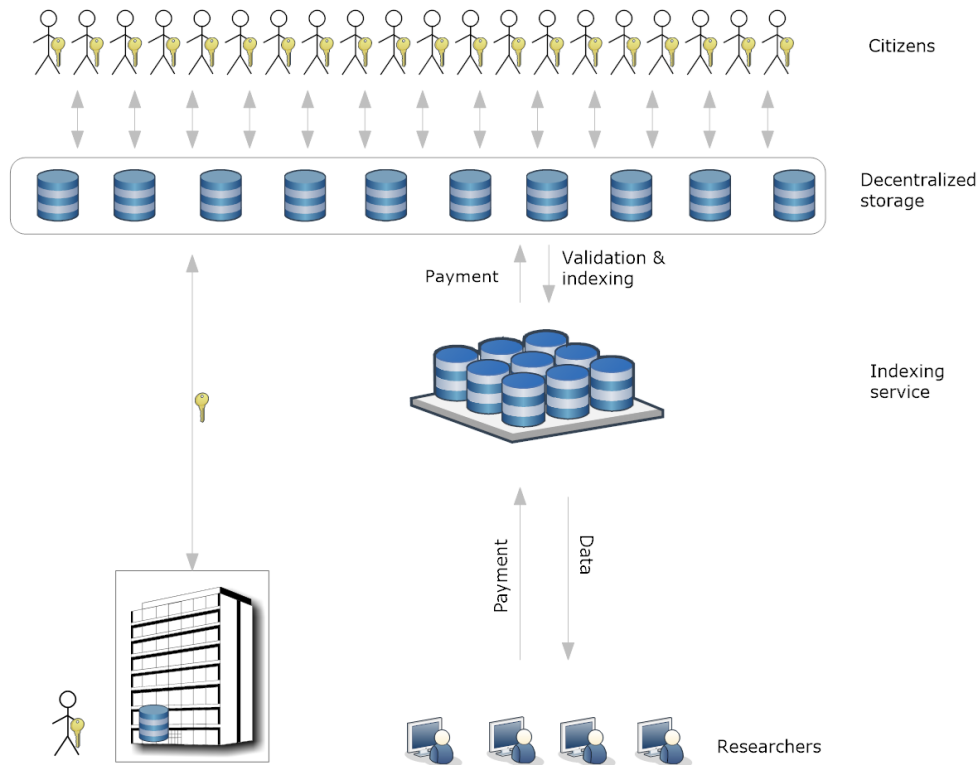


Figure 2: Primary and secondary use of health data in decentralised content-addressable storage

For anonymisation, data owners of data in decentralised content-addressable storage networks may locally run software that deletes identifiable attributes. Therefore, a data-owning person creates an anonymised version of their health data. This modified version of data constitutes a separate branch with its unique hashkey, and by sharing this hashkey, the anonymised data can be published while the original version remains private.

Data shared in this fashion are cannot be considered readily usable for analysis in a unified way with the health data of others. However, this interoperability and primarily semantic interoperability issue is also an issue with the health data in data silos, despite the vast number of healthcare-related interoperability standards [31] and solutions [32].

Therefore, an indexing service needs to be created to gather the data shared by their owners and present them in a harmonious queryable layout. Such an indexing service could be considered a 'health data Google'.

An essential aspect of sharing data for secondary use is incentivisation. In the 21st century, data are the most valuable commodity and people who share them should receive fair compensation. Likewise, it would be natural for consumers to pay for the data they use. That fee could then be shared by the owner of the data and the indexing service provider. A variety of business models and compensation packages can be considered for payment for the secondary use of the person's health data , like health insurance, citizen's income or simply an ordinary purchase

and sale transaction.

A particular case of the secondary use of health data is the use of these data in a health emergency. It is possible to imagine each citizen having a specific branch of personal health data (e.g. patient summary [33]) containing all of the relevant information for an emergency case. The person might carry the hashkey for said data on a bracelet, on a chip card, in a smart app or in another form. In the event of an emergency anywhere in the world, these emergency health data containing information on medications, allergies, etc. will become immediately available to anybody with access to the hashkey.

An exciting option is to preserve health data in decentralised content-addressable storage networks not in personalised form but in anonymised form. For instance, we can use a person's age instead of dates and times when recording a health-related event or when conducting a measurement – if necessary, even to an accuracy of seconds or milliseconds from the time the person was born. Therefore, in this theoretically possible scenario, all people are born on the 1 January 0001.

Using a similar distance pattern analogously to the anonymisation of dates and times in health data, it is also possible to anonymise locations (geographic addresses). Hence, similarly to the anonymisation of dates and times, in the anonymisation of locations, all persons, according to anonymised EHR data, are born at location 0.0, and every other location in their EHR data is a vector from that person's birth location to a specified location. Consequently, all locations in the person's EHR data are anonymised; only by knowing a person's actual place of birth can these anonymised locations can be converted to real locations.

4. Advantages of EHR decentralisation

In this section, we enumerate the advantages of decentralised content-addressable networks and the reasons they should be used to preserve people's EHR data.

1. Data owners have full and exclusive control of their data. Only the person who knows the hashkey of the data can access them.
2. All of the health data of a person are stored decentrally in one logical place. Data are always complete and the problem of data fragmentation is eliminated.
3. Data owners have full control over which data they want to share with third parties. Multiple data sharing schemes can be envisioned.
4. Since data are addressed by their hash value, their integrity is guaranteed. Nodes of the network can easily verify the content of downloaded data by computing their hash and comparing it to the address.
5. Full versioning history is preserved. Every modification creates a new version of the data set with a different hash value. The old version remains intact and is addressable with the old value.

6. Data are protected against accidental loss. Decentralised networks offer sufficient redundancy, minimising the chance of data loss even in the face of a high churn rate (existing nodes leaving and new nodes joining the network).
7. Data de-duplication. Content addressability means that only a single (logical) copy of the same data exists on the network.
8. Global reach. Decentralised networks have no geographical limits. To access the data, only network connectivity and knowledge of the hashkey are needed.
9. Elimination of single point of failure. By design, the failure of a single node or even multiple nodes has no effect on the function of the network as a whole.

5. Main challenges and prospective research topics

A huge number of articles in medical informatics related to the secondary use of routine clinical EHR data are about how to get structured data from unstructured medical texts; how to make personalised medical data anonymous or pseudonymous or how to get consent to use those data; how to integrate data from separate data sources to achieve semantic interoperability; or how to ensure data quality in a secondary analysis. As stated in a survey [5], due to the semantic heterogeneity of health data, we still do not have a unified approach and use the divide-and-conquer approach instead. The review [6], conducted a year later concludes that no big-data analytics will happen without optimised data sharing and reuse, something we still lack despite various interoperability standards in the medical domain.

Therefore, in our understanding, data formalisation, semantic interoperability, depersonalising data and ensuring data quality are the main challenges to solve in secondary use, despite the locations of health data in centralised data silos or decentralised content-addressable storage networks. However, and we are advocating it, the health data locations' in decentralised content-addressable storage networks, under the data owners' full control and ownership, might be the most natural and the simplest way of tackling these challenges as well.

Next, we draft and explain possible research topics on preserving EHR in decentralised content-addressable storage networks.

R1: Data model

A recent study shows [34] that approximately 80% of medical data are preserved in free text. There is likely a good reason for this. Still, for secondary use, we somehow have to get structured (formalised) data from these unstructured medical texts; therefore it is quite expected that researchers are looking for ways to parse the necessary data from these texts.

The question here is how to format the person's EHR data preserved in the decentralised content-addressable network so that they can be displayed in various views (including human-readable free text), used by multiple systems or transferred to various systems using various communication protocols. For instance, each hospital and general practitioner the patient visits during their lifetime most likely does not use the same data communication protocols (e.g. HL7 v.2.7, CDA or FHIR), the same data reference models (e.g. HL7 RIM or openEHR RM) or

classifiers (SNOMED, LOINC or their different versions) or the same language (e.g. English, Estonian). It would be ideal if the data model did not force the use of any of the aforementioned formats, but was able to present the data in any of these formats on the fly.

R2: Data quality

To be able to rapidly answer important clinical questions, the structure of electronic medical records and health administrative databases and data capture therein needs to be improved [35]. The quality of routine clinical data, such as data availability for variables of clinical interests, completeness within a clinical visit, missing and duplicate visits, variability of data capture systems [36], [37],[38],[39],[40],[41], are some of the concrete issues that need improvement. Therefore, mechanisms must be put in place for how the data are validated both technically and clinically (similarly the data are validated in clinical laboratories) before they are preserved in the EHR in a decentralised content-addressable network.

R3: Data interoperability

By data interoperability, we mean in particular the semantic interoperability of both data and knowledge [42]. Suppose that we have a simple, universal and human-readable as well as machine-readable data model for preserving EHR data in a decentralised content-addressable network. The question now is how to convert and present these data in various existing health and medical ontologies, taxonomies, terminologies, standards, communication protocols, etc. and so that the medical meaning is preserved and the semantic interoperability of systems is achieved on the fly using a federated, non-unified, interoperability approach [43].

R4: Primary use

Let's now suppose that we have both the data model and the data interoperability solution so that when a person visits a hospital or a general practitioner, that patient is able to present their data to clinicians. This now raises myriad issues. Whether and why clinicians should trust these data? How does the clinician enter new data into the collection? What happens if a person is unconscious and needs emergency assistance but is unable to present their data? Which data, if any, are hospitals and general practitioners preserving in their systems, in addition to patients' EHR data in a decentralized content-addressable network? And so on.

R5: Secondary use

The next myriad questions are related to the secondary use of health data. Even if a patient is able to make a copy of their medical data and publish it for the common good of society, the question remains as to how all of this works. How can those who need data request data? How do people know whether someone is interested in their data? How and in what form do people publish their data? Are we able to trust these data and why? Does it make things simpler or more complicated? Does it make collecting data for medical research more expensive or reduce the time and costs? And so on.

R6: Data security and privacy

In some cases, and especially in the medical domain, we should talk about pseudonymity, not anonymity. When EHR data are kept in decentralised content-addressable networks, the person has full control of their data and is able to publish part or all of the data. Still, good pseudonymisation and/or anonymisation tools and techniques are needed. Furthermore, several scenarios should be validated to ensure no risk of personal data leakage and that data protection rules are followed.

R7: Data integrity and transparency

When using data for any analysis from several data sources, we must ensure data transparency and integrity in order to trust the results achieved. Therefore, to find and evaluate data, appropriate data transparency and integrity methods and tools for EHR data in decentralised content-addressable networks are needed, both in the context of primary and secondary use.

R8: Linked data and machine learning

In the wider context, it would be interesting to investigate the potential role of decentralised content-addressable network as the foundation for the Giant Global Graph (GGG) - a term coined by Tim Berners-Lee in 2007 [44].

6. Conclusion

Despite the exponential increase in the volume of health data available electronically, one of the main obstacles to its reuse is the siloed architecture of our current data storage infrastructure. This paper proposes the use of decentralised content-addressable storage networks to dissolve data silos and give data control back to the data owner, thus facilitating the reuse of personal health data both for providing immediate care to the patient and for the common good of society. We presented a general description of the underlying principles of such networks and described how these decentralised content-addressable storage networks can be applied in both primary and secondary use scenarios. We also presented the possible future research topics to analyse and resolve before utilising this technology. Although the technology behind decentralised content-addressable storage networks is still in its infancy, it has the potential to revolutionise the way in which health information gets stored and shared.

As health care is increasingly turning away from dealing with the consequences of diseases and turning to general prevention, this paper also proposes the replacement of the traditional electronic health record with a more holistic personal electronic life-long health record, stored securely and privately on decentralised content-addressable networks.

Authors' contribution

T.K. designed the idea and wrote the manuscript with support from K.K. All authors contributed to the final version of the manuscript. G.P. and P.R. supervised the project.

Acknowledgments

This work in the project 'ICT programme' was supported by the European Union through the European Social Fund.

References

- [1] WHO, From innovation to implementation: eHealth in the WHO European Region, WHO Regional Office for Europe, 2016. URL: apps.who.int/iris/handle/10665/326317.
- [2] PWC, Transforming healthcare through secondary use of health data, PriceWaterhouseCoopers, 2009.
- [3] T. D. Wade, Refining gold from existing data, *Current opinion in allergy and clinical immunology* 14 (2014) 181.
- [4] W. O. Hackl, E. Ammenwerth, Spirit: Systematic planning of intelligent reuse of integrated clinical routine data: A conceptual best-practice framework and procedure model, *Methods Inf. Med* 55 (2016) 114--124. URL: doi.org/10.3414/ME15-01-0045.
- [5] B. Shickel, P. J. Tighe, A. Bihorac, P. Rashidi, Deep ehr: A survey of recent advances in deep learning techniques for electronic health record (ehr) analysis, *IEEE Journal of Biomedical and Health Informatics* 22 (2018) 1589--1604. URL: doi.org/10.1109/JBHI.2017.2767063.
- [6] X. Gansel, M. Mary, A. van Belkum, Semantic data interoperability, digital medicine, and e-health in infectious disease management: a review, *Eur J Clin Microbiol Infect Dis* 38 (2019) 1023--1034. URL: doi.org/10.1007/s10096-019-03501-6.
- [7] G. K. Hajduk, N. E. Jamieson, B. L. Baker, O. F. Olesen, T. Lang, It is not enough that we require data to be shared; we have to make sharing easy, feasible and accessible too!, *BMJ Glob Health* 4 (2019) 1550--1550. URL: doi.org/10.1136/bmjgh-2019-001550.
- [8] J. Metsallik, P. Ross, D. Draheim, G. Piho, Ten years of the e-health system in estonia, in: A. Rutle, Y. Lamo, W. MacCaull, L. Iovino (Eds.), *CEUR Workshop Proceedings*, volume 2336, 3rd International Workshop on (Meta)Modelling for Healthcare Systems (MMHS), 2018, pp. 6--15. URL: ceur-ws.org/Vol-2336/MMHS2018_invited.pdf.
- [9] J. Oderkirk, Survey results: National health data infrastructure and governance, *OECD Health Working Papers* (2021). URL: doi.org/10.1787/55d24b5d-en.
- [10] N. G. Weiskopf, S. Bakken, G. Hripcsak, C. Weng, A data quality assessment guideline for electronic health record data reuse, *eGEMs (Generating Evidence & Methods to improve patient outcomes)* 5 (2017) 14--.
- [11] EU, Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation) (text with eea relevance), *Official Journal of the European Union* 119 (2016) 1--88. URL: eur-lex.europa.eu/eli/reg/2016/679/oj.
- [12] NGI, Guidelines for applicants. DAPSI 3-rd open call for proposals, Technical Report, DAPSI, Data Portability and Services Incubator, 2021. URL: dapsi.ngi.eu/wp-content/uploads/DAPSI-Guidelines-for-Applicants.pdf.
- [13] InteropEHRate, Ehr in people's hand across europe, interopehrate.eu, 2020. EU's Horizon

- 2020 research and innovation programme grant agreement No 826106. Accessed: 2022-03-24.
- [14] Solid, Your data, your choice. advancing web standards to empower people, solidproject.org, 2021. Accessed: 2022-03-24.
- [15] Y. N. Harari, Why fascism is so tempting - and how your data could power it, ted.com/talks/yuval_noah_harari_why_fascism_is_so_tempting_and_how_your_data_could_power_it, 2018. Accessed: 2022-03-24.
- [16] Ipfs, Ipfs powers the distributed web, ipfs.io, 2022. Accessed: 2022-03-24.
- [17] Swarm, Swarm is a decentralised storage and communication system for a sovereign digital society, ethswarm.org, 2022. Accessed: 2022-03-24.
- [18] Storj, Decentralized cloud object storage for developers, storj.io, 2022. Accessed: 2022-03-24.
- [19] Sia, Decentralized storage for the post-cloud world, sia.tech, 2022. Accessed: 2022-03-24.
- [20] D. M. Petar Maymounkov, Kademia: A peer-to-peer information system based on the xor metric, <https://pdos.csail.mit.edu/~petar/papers/maymounkov-kademlia-lncs.pdf>, 2002. Accessed: 2022-03-24.
- [21] S. Canham, Long-term management of data and secondary use, in: S. Piantadosi, C. L. Meinert (Eds.), Principles and Practice of Clinical Trials, Springer International Publishing, 2020, pp. 1–30. URL: doi.org/10.1007/978-3-319-52677-5_286-1.
- [22] S. M. Meystre, C. Lovis, T. Bürkle, G. Tognola, A. Budrionis, C. U. Lehmann, Clinical data reuse or secondary use: current status and potential future progress, Yearbook of medical informatics 26 (2017) 38–52. URL: doi.org/10.15265/IY-2017-007.
- [23] L. M. G. Rodriguez, A. Ampatzoglou, P. Avgeriou, E. Y. Nakagawa, A reference architecture for healthcare supportive home systems, in: 2015 IEEE 28th International Symposium on Computer-Based Medical Systems, IEEE, 2015, pp. 358–359.
- [24] J. D. Patrick, P. Barach, A. Besiso, Information technology infrastructure, management, and implementation: the rise of the emergent clinical information system and the chief medical information officer, in: Surgical Patient Care, Springer, 2017, pp. 247–262.
- [25] J. W. McKeeby, P. S. Coffey, The importance and use of electronic health records in clinical research, Principles and practice of clinical research (2018) 687–702.
- [26] H. Ehtesham, R. Safdari, A. Mansourian, S. Tahmasebian, N. Mohammadzadeh, S. Pourshahidi, Developing a new intelligent system for the diagnosis of oral medicine with case-based reasoning approach, Oral diseases 25 (2019) 1555–1563.
- [27] M. Sousa, L. Arieira, A. Queirós, A. I. Martins, N. P. Rocha, F. Augusto, F. Duarte, T. Neves, A. Damasceno, Social platform, in: World Conference on Information Systems and Technologies, Springer, 2018, pp. 1162–1168.
- [28] I. Matloob, S. Khan, H. ur Rahman, F. Hussain, Medical health benefit management system for real-time notification of fraud using historical medical records, Applied Sciences 10 (2020) 5144.
- [29] M. E. Edmondson, A. P. Reimer, Challenges frequently encountered in the secondary use of electronic medical record data for research, Comput Inform Nurs 38 (2020) 338–348. URL: doi.org/10.1097/CIN.0000000000000609.
- [30] V. Menger, M. Spruit, J. de Bruin, T. Kelder, F. Scheepers, Supporting reuse of ehr data in healthcare organizations: The cared research infrastructure framework, in: HEALTHINF,

2019, pp. 41–50.

- [31] R. L. Richesson, C. O. Lynch, W. Hammond, Developing and promoting data standards for clinical research, in: *Clinical Research Informatics*, Springer, 2019, pp. 403–431.
- [32] J. N. S. Rubí, P. R. L. Gondim, Iomt platform for pervasive healthcare data aggregation, processing, and sharing based on onem2m and openehr, *Sensors* 19 (2019) 4283.
- [33] CEN, EN 17269, Health Informatics: The International Patient Summary, European Committee for Standardization, 2019.
- [34] E. Negro-Calduch, N. Azzopardi-Muscat, R. S. Krishnamurthy, D. Novillo-Ortiz, Technological progress in electronic health record system optimization: Systematic review of systematic literature reviews, *International journal of medical informatics* 152 (2021) 104507.
- [35] T. McGuckin, K. Crick, T. W. Myroniuk, B. Setchell, R. O. Yeung, D. Campbell-Scherer, Understanding challenges of using routinely collected health data to address clinical care gaps: a case study in alberta, canada, *BMJ open quality* 11 (2022) e001491. URL: bmjopenquality.bmj.com/content/11/1/e001491.full. doi:10.1136/bmjopen-2021-001491.
- [36] L. M. Kern, S. Malhotra, Y. Barrón, J. Quaresimo, R. Dhopeswarkar, M. Pichardo, A. M. Edwards, R. Kaushal, Accuracy of electronically reported “meaningful use” clinical quality measures: a cross-sectional study, *Annals of internal medicine* 158 (2013) 77–83.
- [37] A. Wright, A. B. McCoy, T.-T. T. Hickman, D. S. Hilaire, D. Borbolla, W. A. Bowes III, W. G. Dixon, D. A. Dorr, M. Krall, S. Malholtra, et al., Problem list completeness in electronic health records: a multi-site study and assessment of success factors, *International journal of medical informatics* 84 (2015) 784–790.
- [38] J. M. Madden, M. D. Lakoma, D. Rusinak, C. Y. Lu, S. B. Soumerai, Missing clinical and behavioral health data in a large electronic health record (ehr) system, *Journal of the American Medical Informatics Association* 23 (2016) 1143–1149.
- [39] K. Lucyk, K. Tang, H. Quan, Barriers to data quality resulting from the process of coding health information to administrative data: a qualitative study, *BMC health services research* 17 (2017) 1–10.
- [40] S. L. Feder, Data quality in electronic health records research: quality domains and assessment methods, *Western journal of nursing research* 40 (2018) 753–766.
- [41] E. C.-H. Wang, A. Wright, Characterizing outpatient problem list completeness and duplications in the electronic health record, *Journal of the American Medical Informatics Association* 27 (2020) 1190–1197.
- [42] R. K. Saripalle, S. A. Demurjian, Attaining semantic enterprise interoperability through ontology architectural patterns, in: *Computer Systems and Software Engineering: Concepts, Methodologies, Tools, and Applications*, IGI Global, 2018, pp. 705–740.
- [43] D. Chen, G. Doumeings, F. Vernadat, Architectures for enterprise integration and interoperability: Past, present and future, *Computers in Industry* (2008). doi:<https://doi.org/10.1016/j.compind.2007.12.016>, enterprise Integration and Interoperability in Manufacturing Systems.
- [44] Wikipedia, Giant global graph, en.wikipedia.org/wiki/Giant_Global_Graph, 2022. Accessed: 2022-05-23.