Blockchains and Data request to comment

Peter Wittenburg (Max Planck Computng and Data Facility, peter.wittenburg@mpcdf.mpg.de),
Wolfgang Kuchinke (University Hospital Düsseldorf, Wolfgang.Kuchinke@med.uni-duesseldorf.de)

Abstract

Blockchain is a new technology that has the potential for radical innovations in data management. Its use to improve transparency and increase auditability of data flows can significantly reduce transaction costs and increase efficiency. Yet the technology is in its early stages and limitations, challenges as well as possible technical and regulatory risks, need to be addressed. In general, questions remain about Blockchain's scalability, flexibility, and ability for privacy protection and data governance. One should keep in mind that Blockchain's primary value is the deployment of cryptographic mechanisms to reach consensus across different parties in the ledger. This eliminates the need for a central authority, like a Trusted Third Party, to create trust in a distributed system.
Here we discuss Blockchain technology (BCT) from the potential user perspective as integral part of research infrastructures and e-infrastructures to support the data life cycle of large research projects. In some cases we show that some aspects of BCT can more easily be implemented using already existing technologies, like Persistent Identifiers (PID) and metadata records. We focus on the area of medical research, clinical trials and health data to identify areas for a suitable use of BCT to improve especially the tracking and monitoring of data flow and transaction processes of sensitive data. And we point to problems of BCT with scalability, flexibility, and its ability to secure sensitive data protection.

Recently, blockchain technology (BCT) including its claims and expectations was recommended again as solving real problems in data driven research work, in particular for data challenges in medicine and clinical research. This motivated us to have a deeper look into BCT and look at some challenging situations in data science. Without being explicit in this paper we distinguish the cryptocurrency part, such as Bitcoin, from other instantiations of BCT, from Hyperledger etc. and focus on the second part.

BCT in short technical Terms

BCT is a distributed database approach; in detail it is a decentralized peer-to-peer network, where each participant maintains a replica of a shared append-only ledger of digitally signed transactions, with

- two different database components, one including the transaction data and/or smart contracts and another one storing the transaction sequence and pointer structures
- a special feature in so far as the hash (fingerprint) of an earlier block in the chain is being included in the data of the subsequent block
- a set of policy rules describing all kinds of actions that need to be carried out at certain events such as creation of a new block, replicating blocks once accepted etc.

A number of other features and functions of BCT can be mentioned such as permissions, encryption, node identities, certificates, replication, etc. but they can be found in other kind of solutions and are not specific for BCT.

There are many different ways how one can implement such a database system which all together defines a large solutions space. The good story about BCT is that it has narrowed down this solution space to a much narrower space - still allowing for some flavours: there is ready made software supporting BC, and there is an active developer community. One can implement public BC or a

private BC. A private BC requires an invitation and must be validated. Independent of the view on Bitcoin as a success or failure, it demonstrates that the concept is operational and addresses a few critical issues. Interested parties managed to create a momentum convincing many to invest in BCT, in particular in the financial sector.

Trust and Identities

The big question in internet based activities is how we can establish trust between actors at affordable costs in situations where data which needs to be protected is being accessed, exchanged or traded. We need to admit that there is currently little trust in the internet mechanisms[1] and that people have begun to admit to this fact and started to live with this in different ways. For dealing with certain goods such as "money" or "sensitive data" this is fatal and new approaches are urgently needed. So far, conventional solutions to create trust are the creation of independent Trusted Third Parties (TTP) or “Data Custodians”, solutions that are laboriously to integrate in data processes and costly. Here, BCT offers a mechanisms that automates trust and that functions independently from any TTP.

Fundamental for trust building mechanisms are clear identities of actors, mechanisms, which we are currently missing in the internet. With BCT it is possible to generate trust in processes in a widely untrusted environment. One way to achieve this is that BCT introduces identities of the participating nodes based on certificates. This is not new and certificate-based node interactions are being used in various scenarios. But in applications, for example, where everyone can become an actor (citizens), we still have the problem that confirmed identities are missing and BCT does not solve this problem either. For example, the introduction of personal certificates in the Grid world did not work out, since the resulting administrative overhead was not accepted by users. Apart from that, ORCID has been introduced in the science publishing world to map different spellings of names, but it only addresses identities of researchers. Finding simple ways that prevent manipulations in the digital domain will remain a challenge.

Federations and Network Size

BCT introduces the notion of a BC network which incorporates all nodes that agree on certain contracts. This is generally called a "federation" and many initiatives in the research domain have implemented federations to share for example HPC facilities or a common data space. It is common practice to sign agreements within these federations and exchange mission critical information such as certificate based identities, hot lines in case of errors, etc. Within such federations it is possible to define permissions of the different actors.

BCT makes use of smart contracts, i.e. agreements that are specified in terms of actionable rules making use of a specific formal language. This concept can already be found, for example, in complex document management systems, but until now most, if not all of the existing data federations in the research domain did not use this concept. Here we can learn from BCT.

Federations can be large, can be used to exchange up to terabytes of data and every node can be member of different federations where different rule sets apply. This situation addresses essential scaling issues, where BCT displays severe limitations. These are mainly due to the fact that encryption is being applied and that complex power consuming algorithms may be used to sort out acceptance of certain actions. Hyperledger addresses the scaling problem by allowing to create sub-networks, i.e. within the distributed ledger sub-networks can be built that only share part of the data. If federations of some kind would strictly apply encryption and implement some complex decision algorithm they would also suffer from heavy load on computers. In principle, BCT can

[1] There are many levels of trust. What is meant here is that the Internet is full of fake identities and fake news, has gaps with respect to security, etc.

accommodate only small amount of data, like identifiers, keys, hash values, data sets and any large scale data need external solutions.

Still, in an open data market solution, which will finally need to address the eminent needs in research and industry, BCT seems to be a closed network with its own narrow specifications. This implies that BCT should primarily be used where this "closed network" approach makes sense.

Identification, Chaining and Searching

In principle, BCT only knows two components: chains of blocks; this is its strength and also weakness depending on the application. In the research and industrial data domain the concept has now been widely accepted that every Digital Object (data, metadata, software, etc.) needs to have its own persistent identifier (PID) and needs to be accessible to be able to re-use it in different contexts. BCT does not incorporate this concept; in BCT one can only address the chain and process it. Of course, it is possible to add simple information units to each block and register, for example, a PID for each of them. However, this only makes sense if there would also be separate metadata that could inform applications about the specific content of each block. This would also allow nodes to create indexes that can be used for searching for information in the blocks based on the metadata. But as has been pointed out, this would create external instances about the information in the blockchain and could weaken its basic strengths of security and invariability.

In addition, chaining can be implemented in different ways and it doesn't need a BC mechanism for this. Let us assume that a repository makes systematic use of PIDs referring to each digital object as it is recommended by the FAIR principles, then it would be easy to add an attribute to the PID record to implement a chain. Giving this attribute well-defined types would tell machines how to use it. Of course, one could also add a "hash" attribute which machines could read as well as to test authenticity and repositories could even add this "hash" value within the data headers. This would rebuild using existing components the functions BCT already has incorporated. The big difference would be that the software that would rebuild this functionality would have to be certified to establish trust which would cost substantial efforts. In BCT the implicit assumption is that all crucial software has been verified and validated.

BCT and Metadata

It is apparent that BCT is not made for big data. The question then is, whether it can be used for example for handling metadata which consist normally of a limited set of key-value pairs. For dealing with metadata in research and industry many solutions have already been invented and employed by many initiatives world-wide:

- metadata should be open and being exchanged by a standard protocol - most using OAI-PMH and in future perhaps ResourceSync
- metadata are separate digital objects that have an identifier - most are using Handles/DOIs[2]
- metadata are described by a schema and include semantic concepts - both being openly registered
- metadata service providers harvest metadata, carry out some form of semantic mapping, do some indexing and offer search/browsing portals - all being based on widely used standards and tools

Of course, heterogenity with respect to structures and in particular semantics still is a huge problem and requires much effort to cope with. But these issues and issues, such as lack of quality, are independent of the database technology being used. For the topic of metadata management, BCT is too limited to account for the full potential of metadata that is necessary to exploit. Bluntly, we can

[2] DOIs are Handles associated with a specific business model

state that there is no need to yet use another technology such as BCT for dealing with metadata, in particular, when it follows a very much "closed" approach.

Provenance Tracking

One of the strengths of BCT is the built-in provenance, i.e. having a certain block allows the user to check its complete history. But the provenance model used by BCT is a very simple one. In data projects provenance tracking means to manage graphs, i.e. return back to a certain digital object and start a new track (experiment, study, etc.) in combination with others. BCT seems to be not made for complex provenance models.

Again, as indicated above, one can implement provenance by using for example PIDs in a specific way. The difference again is that the software implementing provenance would have to be certified for applications with sensitive data.

Data Immutability

Data in BCT is immutable, i.e. one cannot modify information in historical blocks. This is ensured basically by two mechanisms: (1) in the actual hash all information about earlier hashes is encoded, i.e. to a certain degree the history is embedded in the actual code. (2) Changes are notified by the network participants and require agreements, i.e. on purpose it is made very difficult to change information.

The first point could also be solved by adding a hash value to the PID record, i.e. at all times people can check whether a given digital object is still the same. This does not address the second point of what needs to be done to find agreements about changes. Here BCT has a ready-made solution which other implementations would have to build in.

Data Preservation

As far as we can see, BCT does not address the issues of archiving or long-term preservation. At this moment BCT provides a database that is continuously accumulating new blocks and conceptually allows to summarise historical blocks. There needs to be a built in mechanism to extract parts of the history and to set it aside to be managed by a different system that also has to take care that content will not be changed. Besides this there is the general challenge of content stewardship, i.e. to maintain content in a way that it can be understood and used after many years of archiving. In these respects BCT seems to be more restricted than other database solutions.

Personal Health Data Example

It was suggested to use BCT for health data. Health data is sensitive data that needs special protection, according to the General Data Protection Regulation[3]; its processing is subject to usage rules and requires adequate technical safeguards. In the appendix we sketch typical data use cases in the health domain without claiming completeness, mention some requirements and discuss a few optional scenarios of applying BCT.

From the scenarios discussed in the appendix it seems that BCT cannot be a comprehensive solution to all health data problems, but it could indeed be used for some tasks complementing the current IT methods despite its limitations with respect to core tasks in processing health data. Here we summarise our view on BCT in the health data domain:

- In all cases where it makes sense to control sequences of actions between partners BCT could be used to document these and create an immutable record. In most of these health

[3] REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT: http://ec.europa.eu/justice/data-protection/reform/files/regulation_oj_en.pdf

use cases it makes sense to keep health data itself separate from the BC to not suffer from expensive acceptance algorithms.

- Companies may want to trade "clinical study data" and it seems that BCT is perfect for memorising any transactions, like trade interactions, since the BC transactions state all essential information: who has received which data under which conditions. This can be used in such a way that participants in a trusted network would receive the data sets and that BCT databases only contain the transaction records.
- BCT may contain encrypted identifier, creating an immutable record for provenance tracking.
- BCT are not suitable to handle larger sets of data (time series etc.) and it is also questionable whether it makes sense to use BCT for the metadata in scenarios where flexibility is required.

But keeping the "real" data outside of BCT would mean that it becomes necessary to do major actions, like data processing and analysis, outside of BCT resulting in several integration problems:

- BCT does not support the concept of metadata external to BC, i.e. it has an integrated view, as explained above. Therefore, finding suitable data for research, for example by re-using existing clinical study data sets, would always require to process whole blockchains and to have search functions as part of the assessed software stack at each node.
- Mechanisms integrated with BCT would be needed to check whether type-value data and the larger volume data have indeed been transferred as recorded in the BC information sequence.
- Analytic software will have to run outside of BCT because, i.e. there is no control of the proper and compliant usage of data, except one would invest quite some effort to extend BCT and the analytics SW by enabling the BC to record all metadata of the analytical processing of the data sets.
- The concept of data ownership legally implies that users (e.g. data controllers) need to be able to delete certain data, for example, delete health data on request of a patient as data owner. This, however, is hardly possible with ready-made BCT.

There seems to be a dilemma:

- If type-value data is being processed within the blockchains, there need to be methods in place for extracting them into an index to make data available and accessable via search portals and to be able to carry out statistical operations on them, etc. in this context, data will need to go outside of the controlled area of BCT to make it useful for future scenarios. But if one has to go off-BC, all security and provenance tracking features of BCT become only partially effective.
- If only the transaction information is stored in the BC the effect is almost the same, except that during exchange the special security features of BCT are not applied to the data.

In summary, when working with health data, some software, which is not part of the trust-generating software stack of BCT will be used on data copies externally from BCT for data processing and analytic purposes. BCT may complement systems for secure and legally compliant data analysis, where a secure area is offered to process data, by providing an immutable record of all transactions performed on data. For the emerging re-usage scenario of stored and newly generated health data, search functions will be required, which also will have to be placed externally from BCT and which will require additional security structures. This future data usage scenario stresses a point, which has now been accepted widely in the research community and which is also being mentioned in the FAIR principles: Each digital object (DO) (be it type-value data, data from wearables and other sensors, brain scans, etc.) needs to be assigned a persistant identifier (PID) and fingerprint information to be able to check at all moments which digital asset is being accessed, exchanged, used etc. It has been overlooked in the discussion of BCT for health data that in case BCT technology is being used these features should be integrated.

The advantage for researchers in this scenario is that the employment of PIDs results in further options becoming available:

- The encrypted PID can be registered in a block.
- Each block (DO) can now be sealed in so far as the PID record could contain all required identities, in particular the one of the patient in anonymised form, to allow to check at all moments where the data originates from and who has ownership rights.
- Added fingerprint information to each DO ensures that the DO cannot be manipulated.

On the other hand, if PIDs would be assigned in the clinical data world to fulfill FAIR principles, to allow provenance tracking, etc., most of the functions where BCT seems to make sense could also be easily implemented by making use of advanced PID systems such as Handles. We need to accept however, that BCT is a ready-made technology, without the need for severe adaptations and without the need to implement a TTP structure.

Data Markets and BCT

Let us assume that people all apply the RDA DFT Core model[4] which basically states that each digital object has assigned a PID, is associated with metadata, has its bit sequences being stored in trustworthy repositories and that there are stable links between these entities; all together creating something like the ideal data world. Let's also assume that there are metadata providers offering digital objects for re-use, that their portals are somehow linked so that software agents can find them and that brokers take care of the metadata to include actionable re-use conditions. With these assumptions and a certain degree of harmonisation we can speak about a global cross-sector/discipline data market. Such a data market would create a huge stimulus for data sharing and meta-analysis of data across different research fields.

In many cases, data producers or data owners, as legal data controllers and owners of intellectual property rights, want to know who will have access to data, how data will be used, whether access conditions have been met, etc., i.e.; they want to know, whether data is being used according to their specifications, often stated in a Data Use Agreement, which could even include payments in case of commercial data trading. Re-usage could also happen within sequences of operations, i.e. someone aggregates or links data and creates new derived data based on this collection. Another one could re-use this resulting aggregate data for new purposes. Standards such as PROV[5] specify formats how provenance information should be structured and described.

We can see the following general needs:

- There needs to be a machine-readable way to specify where to find the provenance record of a specific digital object.
- Provenance records for data consisting of Open data and sensitive data linked together
- There needs to be a way for sensitive data to store "transaction" kind of metadata and make it findable as well.
- There need to be operators to check provenance records for correct re-use of data.

To implement the third point we could see the usage of the BCT to generate trust by tracking data ownership, data processing and financial transactions, and similar types of metadata for data sharing or trading purposes.

The basic problems we face today in the management of scientific and industrial data revolve around vast volumes of data rapidly accumulating in many different forms using different

[4] RDA DFT modell: http://hdl.handle.net/11304/5d760a3e-991d-11e5-9bb4-2b0aad496318
[5] PROV: https://www.w3.org/TR/prov-overview/

techniques across multiple disciplines. None of these core problems involve trust at first instance, which is the primary problem solved by BCT. There are specific domains in data management where BCT could be applied, but a proper integration with the major trends in scientific data management and processing will be the challenge.

Appendix: Birds View on Medical Data and Blockchains

The following text is a summarising sketch of major clinical data objects and sets with the goal to evaluate the suitability to use of blockchain technology to improve the handling and processing of health data. We only make a difference between clients' health records as they are maintained and used by doctors and those records that are being maintained and used for research projects which are often called “study data” or "trial data". It is known that all these records are managed by some information management system, that the specifications of such records are not yet harmonised despite all attempts and that the quality of the data is highly variable.

1. Requirements for Health Data

In the domain of health data, including data created in medical care as well as clinical research, comprehensive data governance frameworks exist. These frameworks are based on "Code of Practices", guidelines, standards, as well as local, national and international regulations. Access restrictions, identity authentication, data ownership, informed consent, usage agreements, contracts and data usage rules, all are used to guarantee the protection of privacy of the patient and the rights of the patient as data owner. Thus, the main aim of these frameworks is to create and maintain an environment of trust. Any technology employed in the health data domain, such as BCT, must satisfy the above requirements. A big hindrance for the usage of health data and other sensitive data is that research infrastructures and e-infrastructures have not established a comprehensive solution to deal with this sensitive data. There exist different forms of identity management, authorisation and access restriction procedures that are still neither user friendly nor foolproof. In this context, Blockchains may provide a stable process and provenance tracking mechanism that supplements the existing data protection frameworks. One problem hereby is that often additional tools are required to complement security frameworks. In the case of BCT, off-Blockchain facilities may be needed for user authentication, for patient anonymisation, for storage and analytics of sensitive data, etc. Such additional tools may compromise the simplicity and effectiveness of the closed BCT to handle specific subtasks such as recording specific events in a correct order for the monitoring of the course of a clinical study process. In this context, another problem seems to be if and how BCT based systems for clinical trials can be validated for GCP and legal compliance. It is a requirement of Good Clinical Practice that all systems that handle patient data in GCP compliant clinical trials must be subject of Computer System validation (CSV)[6].

A number of general requirements for implementing suitable technology for dealing with sensitive data follow from these descriptions:

- For each Digital Object (DO) such as for clinical trial data sets, data sets from wearables/sensors, brain scans, etc. an identifier needs to be assigned associated with fingerprint information. This allows checking at all moments which digital asset is being processed, exchanged, used, etc. Metadata is often associated as type-value pairs allowing to search for suitable data. In cases where this is not inherently the case, such as for time series, additional metadata should be assigned to be FAIR compliant and to enable re-use of the data sets. Personal information included in metadata must be protected. In addition, because of the included fingerprint the immutability of the data sets is guaranteed.
- Each DO or group of DOs (collection) needs to be sealed to allow checking at all moments where the data originated from, who has ownership rights, and what the allowed purpose of use is, etc.
- Sensitive data needs to be de-identied and ideally anonymised, however, for ethical reasons, there must be a way back to the individuum, who is the data provider and ownership must be clarified to enable, for example, data owners to let delete their DOs. Normally, this is done by a trusted third party (TTP) that is independent from the data owner and the data

[6] Ohmann C, Kuchinke W, Canham S, et al. Standard requirements for GCP-compliant data management in multinational clinical trials. Trials. 2011;12:85. doi:10.1186/1745-6215-12-85.

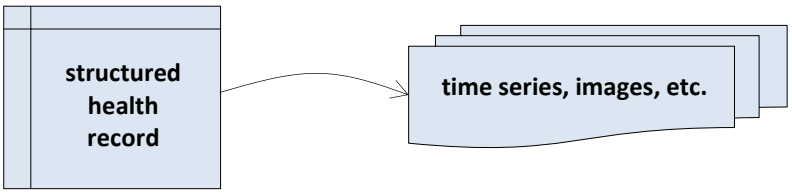
consumer. It is advertised that BCT may provide a mechanism by using identifiers including a personal key and encryption to generate trust without the need for such a TTP. We think it will be difficult to substitute an independent unit to habour the links of the different identifiers and pseudonyms by mechanisms that are based solely on a personal key.

- There need to be mechanisms that ensure that DOs cannot be exchanged/traded without the above mentioned crucial information and protection requirements being included.
- There need to be ways to allow data owners to specify the kind of use being acceptable and the (type of) actors that are allowed to carry out operations on the data. The data owner must be able to change these conditions.
- It would be ideal to ensure that only validated software may be used to carry out operations on data, but for complex operations this seems to be out of scope.

2. Health Records per Client

Doctors ask questions, make tests, measure variables, create time series and images, collect information, code data and give medications, all transactions are linked to a specific patient. This process results in structured health data records and where necessary include links to data of larger volume (time series, images, gene sequences, etc.).

It's the information system, like the EHR, that guarantees that the data and links are stable and that there are visualisations of data, like for time series and images. Just looking at the health records we can speak about small data volumes, however, they are exceeding the normal metadata size due to history tracking.



In general, patient data records only exist in the local information database instance, like the Hospital Information System (HIS), due to ethical and legal regulations which may vary per country. Ideally no one should change old records; in practice however, doctors gain new insights and may need to revise or annotate earlier statements.

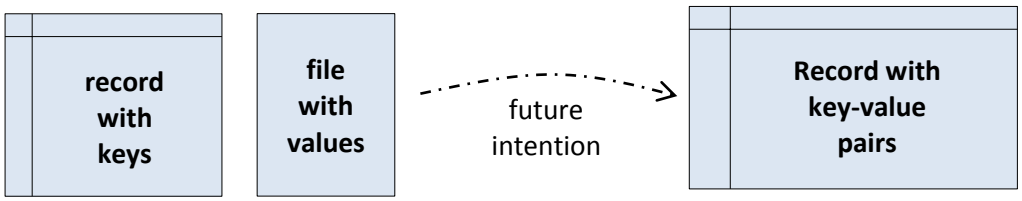
The major risks here are:

- doctors make wrong assumptions: a wrong diagnosis, apply wrong medication, etc.
- certain treatments are not happening in proper sequences etc.
- lack of quality control of data (e.g. missing data, units)
- data is being accessed or misused by unauthorised persons
- data is being sold without the patient being aware of it
- data is being exchanged between doctors and failures occur during the transfer

3. Clinical Research Data

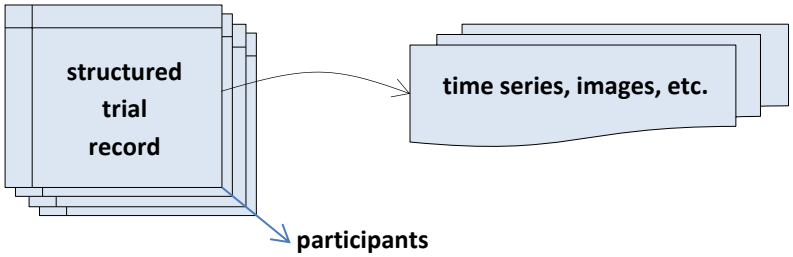
Here we speak about data collected in clinical studies and trials, which is often done in many sites across borders (e.g. multi-center, international trials). Investigational clinical trials take place in a regulated environment, where all operations are subject to regulatory compliance. On the other hand, non-interventional and epidemiological studies exist with a lesser degree of quality control and monitoring. For each study the set of observations and measurements is exactly defined as well as the methods of quality control and analysis are applied to guarantee a high quality of data and the protection of the trial participants.

After the last observation event in a trial the study data collection phase is finished and the data cleaning and analytics phase of the data will be carried out, after which the study is finished and study results will be published and / or submitted to a competent authority. Finally the study database will be archived for an extended period of time. Of course, all observations in general are done with pseudonymous data and blinding. This means, a key is used to indicate individuals and only a authorised person has the link between the key and the participant's ID. To harmonise studies pre-defined and validated data capture systems and data standards (e.g. CDISC) are being used.



This implies that for each study the data record structure can be specified and implemented in the used information system. The exchange of data is being done by using specified formats - currently most often separating between keys[7] and values. A new standard (ODM) integrating the two based on XML has been defined, but is not yet used widely.

The structure of the databases actually is similar to the own above.



Data is collected at many different sites in a multi-center trial. With the last observation event the data of the different sites is being collected in a single database, the trial database. Because local patient data is stored in different local information systems at each site and in the central study database, there exist two copies of all patient data guaranteeing that the data is not being manipulated. With BCT, one has a potential mechanism in place to guarantee the immutability of patient data without the need of two copies. The data in the central trial database is used for the final analytics[8]. Clinical trials are well planned and are carried out based on a special sequence specified in the study protocol. Theoretically, it should be impossible to re-calculate identities from the data set and all partners are seen as trustworthy. This is so, because in clinical trials conducted under the GCP standard, all participants have to sign the study protocol and agree that they act under consideration of Good Clinical Practice (GCP) and data protection. Often clinical trials are carried out by industry and for them study data possess value as intellectual property and need additional protection. For the analytics part the data set often is being sent to data centers that have suitable statisticians and sufficient processing capacity.

As in many other sciences, recently the wish came up to be able to combine data sets from different trials and conduct meta-analysis. This implies that "old" data is not seen as an archival object anymore, but as part of a repository the data of which are made accessible and can be used in different contexts for new research questions. In the EU project CORBEL[9], the research infrastructure

[7] This is often called metadata within the health domain.
[8] It seems to be practice to carry out early analytics to check whether the chosen methods are useful.
[9] http://www.corbel-project.eu/home.html

ECRIN[10] is establishing a pilot of a clinical trials data repository that contains individual-level patient data suitable for independent re-analysis and meta-analysis between different trials.
As in other sciences, such accessablity of data will require to point at and select specific observations, have an index of all studies and their metadata (keys) and a search portal. Often the documents and data of a clinical trial including the study protocol, the registration information, statistical analysis plan and data management plan, and publications are stored in different locations. Here identifiers (e.g. Handles/DOIs) and metadata can play an important role to link all DOs of a single clinical trial after it has been finished to make them findable. An area for the implementation of BCT could be Real World Data (RWD) in medical research. Clinical trials enriched by RWD coming from patient data repositories, sensors or data collection by PRO (Patient Reported Outcome). In PRO patients by means of their smartphones enter data about pain, quality of life, etc. Here too, a copy of the data does not exist and no means of an independent control of the absence of errors; this could be achieved by running blockchain based tracking during data entry. In addition, the integration of sensor data requires the tracking of data correctness, calibration state, software version used and provenance; all these information could be recorded by BCT.

The major risks here are thus:

- the quality of the collected data and the analytics may vary
- the sequence of the observations is not adhered to
- clinical trials are underpowered, requiring changes in the study protocol (e.g. eligibility criteria)
- data is being stolen or misused by unauthorised persons
- data is being exchanged between partners and failures occur during the transfer
- non-validated software and processes are used
- people do not use agreed software on the data set, incl. a re-calculation of identities

4. Large Studies, for example in Cancer Research

The example can be translated to other studies, such as studies about mental diseases, where large quantities of data are required to correlate disease phenomena with all kinds of basic data that may exhibit typical patterns.

Measuring massive gene sequences is done across people showing similar phenomena. Therefore, person characteristics and phenomena are stored typically together with gene and protein data to find evidences about patterns that may cause the health problems. Detecting typical patterns may lead to medication that would then lead to clinical trials (see above). Genome sequencing is one technology that has been recently used for clinical trials to get information about genes and markers to treat disease. Especially whole-genome sequencing will play an important role in personalized medicine, because clinicians will be enabled to identify increases in disease risk for individual patients. But access to such genotype-phenotype and pedigree data needs especially careful and strict control to prevent any misuse, like the identification of individuals or members of families.

Doing this kind of research means to

- collect much data from various participants and many steps of extraction and normalisation
- carefully describe characteristics of the participating persons[11]
- data needs to be integrated from different partners working in different labs, collecting data under different conditions and in different formats
- the consolidation of clinical data is necessary for the collection in a centralised clinical data repository (CDR), for reducing data complexity and clinical risk by enabling re-analysis of data

[10] http://www.ecrin.org/
[11] We assume here anonymisation according to the state of the art.

- laboratories have to standardise tests, methods, reference ranges and reporting, using reference materials and value assignment. Laboratory certification can guarantee the quality of test results.
- create collections of data from different groups of participants and carrying out a variety of complex operations in a sequence
- regroup collections or redo certain processing steps leading to complex graphs describing the experiments that have been carried out

There is a trend towards using flexible workflows to document all steps automatically and to use identifiers and metadata to describe all digital objects involved. Use of standardised reporting and cerification of processes are important and may be supported by BCT.

5. Usefulness of Blockchain Technology - few examples

In this chapter we want to elaborate on a few possible examples where BCT might be of use. This elaboration is not meant to be comprehensive, since in special cases different priorities might be of relevance.

5.1 Patients' Health Data Records

Doctors could be organised in federations to exchange patient data; for example, sending hospital data to the family doctor to continue treatment. Since one cannot foresee who needs to get which data either a huge federation is being set up or the doctors are part of incidentially created federations. Any technology that keeps track of who gets which data under which criteria is crucial.

BCT could eventually play a role here, if doctors' info systems would be part of a system of trusted nodes - a BCT network. In this case they could easily exchange data of all the patients that they share. Exchanging all data of all patients does not make sense and is not wanted. If one would use BCT one could use features such as encryption etc. to increase security. However, BCT as far as we can see would have to be extended to enable selective exchange of data, raising doubts about the usefulness of the standard shared ledger concept.

Another option would be to just share the "transaction information" that data of type X has been sent to doctor Y. Again the question needs to be solved, what the shared ledger exactly is and whether all nodes need to know which partners exchanged information. Our interpretation is that additional software would have to be written raising the question whether BCT is then useful at all. The pure information about transactions is not sensitive.

For all exchanges of data it is crucial to send the complete set of information, i.e. PIDs, metadata, usage specifications, etc. need to be transmitted as a "collection". BCT could be used to maintain records of what exactly has been exchanged, i.e. document, for example, that the receiver has received the full set of information.

A last question is how patients as "owners" of their records can be integrated. They are not part of the BCT network, i.e. special authentication measures would have to be applied to make sure that the person acting is indeed the person it pretends to be. Until now, the personal key serves as an identifier in BCT; but this is insufficient for health data. If this problem would have been solved then it would make sense to give the patient the rights to say what can be shared with whom and for which purpose. Whatever technique would be chosen the risk exists that it would be complicated and overwhelm the capacities of most patients. Most patients already will have difficulties to manage their personal and private keys. Including a suitable rights management system to delegate authorisation is nothing new for databases and could be added to an electronic informed consent system.

5.2 Medical Research Data

Different scenarios could be thought of to use BCT.

Process Sequence Control

The sequence of a clinical trial is highly regulated and controlled. Trial data often must not only be collected in specified sequences but also in phases which require a high degree of synchronisation. A study protocol may be amended, this means a change in the procedure will be prescribed. At different sites this amendment or new versions of the study protocol will become effective at different time points (e.g. after the study investigator has been trained and has signed the new version of the study protocol). In this case, a blockchain of the time points when an amendment becomes effective in different sites in different countries and time zones will be a help to guarantee transparency of the trial conduct. All partners for a specific trial could be joining a BCT federation, i.e. share a ledger. Whoever, is starting/finishing a certain observation, completes a data entry step, signs a document belonging to a sequence/phase, needs to send an information to the partners so that all can check immediately at which step the other partners are and whether agreements are followed. One of these steps could for example be 20 patient enrolled, or 10 adverse events recorded. If this information would include some crucial information about the trial type and other metadata, then one could also make use of the specific trust & security features of BCT. BCT should be usable immediately for such an application. For each trial or trial phase a blockchain could be initiated to follow activities and in this way support the monitoring of clinical trials. Clinical trials that are conducted according to GCP have monitors who make sure that the data are collected and recorded properly and that the trial runs smoothly. They meet periodically with investigators and review their study records and they ensure that the reporting of adverse events is complete. Insight into the Blockchain would give monitors an overview over the study conduct at each site and may reduce the burden of meeting with investigators.

One could even imagine to exchange the individual results by using blockchain. If people need to be sure that the content has correctly been received at each node, then the CPU-intensive check feature of BCT needs to be used. But it would not solve two problems:

- high volume data (time series, images) do not fit with the BCT paradigm, i.e. separate mechanisms need to be used to exchange this kind of data
- at the end of a trial software will be used to run analytics on the whole data set, i.e. the data set needs to be exported and thus exists outside of BCT control.

And it would require expensive computer power for a check, which to us seems to be overkill. Instead, one could switch this feature off and simply use hash values to represent the content.

Version Control in Sequences

It will often be the case that certain observations have to be redone due to errors or unclear results, missing values, or wrong calibration of data output. In clinical trials, data management systems have so-called query generation systems installed. Entered data is subject to a check; for example, lies the entered value for systolic/diastolic blood pressure far higher of below 150 / 60 mm Hg. If the value does not lie in a certain defined range the investigator receives automatically a query to confirm the entered value. This would mean to replace a specific data subset within the whole data set. If this needs to be documented then one would get a relative complex graph. One could use BCT to follow each transaction, i.e. each replacement within the dataset. This would require that the data itself is not subject of the shared ledger, but that the ledger is used to carefully document all operations of all participants. This alone could be useful, however, it would not use BCT's core features.

Copy Control during Data Sharing

Assuming that a clinical research center has proper collaborations based on agreed terms to create trust federations to facilitate data exchange. On request centre A would send centre B some data subset for relevant analytics and is looking for a proper method to document who has received which data sets. There are a few comments to such a scenario:

- The centres that will want to receive data sets from centre A will change over time dependent on the type of research question, i.e. there will be changing federations.
- One term will probably be that centre B should not send data sets to third parties, i.e. that there is no extended sequence of exchanging data sets.

For these scenarios the use of BCT does not make much sense.

5.3 Studies with Large Data Sets

In these studies there is so much data and flexibility involved and skilled experts are involved that BCT would be a too limiting factor to be of great value. Also here collaborations change dynamically dependent on the study. The only application we can see at this moment is, to use BCT to document where and with whom data sets have been exchanged with. Due to the flexibility required we cannot see that the limited BCT would add efficiency; it would rather be seen as adding bureaucracy.