Block Chain Use Case
Medical Data Transfer for Exploratory Use

Version	Date	Author
1.0	15-Jan-2018	Wulff, Moss-Pultz, O'Brian, Bergeron

Contents

Motivation

This document was assembled in response to participation at the industry-wide *Blockchain* meeting that took place at Pfizer Kendall Square on 1-Nov 2017. Teams were assigned to develop blockchain strategies to support a set of general industry use cases discussed at the meeting. Use cases that promote more efficient individual medical data sharing are described herein. In these cases, medical data can refer to patient data derived from electronic health records, individual-directed data generation (e.g. recreational genotyping via *Ancestry.com* or *23andMe*) and/or data derived from participation in clinical studies. The intent is to describe one or more industry processes that utilize individual data that can be positively disrupted by the application of blockchain technology to benefit of

life sciences research, industry and/or academic. In an ideal case, these use cases would positively impact individuals directly.

Participants

Nicole O'Brien (Pfizer), Jennifer Wulff (Pfizer), Sean Moss-Pultz (Bitmark), Jay Bergeron (Pfizer)

Blockchain Summary

Blockchain is a digital transaction management framework that enables decentralized multi-party exchanges. Originally designed for the *Bitcoin* crypto currency (https://bitcoin.org/bitcoin.pdf), the framework offers the potential to generally disrupt centralized processes that have historically required a third-party intermediary while ensuring trust between interacting parties. As such, Blockchain has been used, or proposed, to replace many entrenched transaction models including currency exchange, wire fund transfers and digital content investment/transfer including media and software. Additionally, contractual obligations can be, in certain blockchain implementations, embedded into transactions (i.e. *smart contracts*) to enforce condition-based responsibilities such as payment scheduling.

Blockchain relies on a network of independently-managed processing nodes that maintain identical copies of a transaction ledger and, at specific intervals, verify the state of the exchange by majority validation (or *"Proof of Work"*) of a solution to a computational puzzle based on a hashed conversion of the content of the records within the exchange. The integrity of a blockchain is dependent on a reasonably sized population of independent processing (a.k.a. ***"mining")*** nodes. A successful blockchain must nurture both a productive marketplace to attract a suitable number of participants to execute transactions and a suitable number of participants to maintain the platform. *Public* blockchains, such as Bitcoin do not limit participation and employ general incentives, such as monetary payments, to encourage the participation of miners. Private blockchains, which limit miners and participants to select individuals or entities, also exist.

Blockchain has been proposed as a solution to a many problems already supported by traditional digital technologies. Blockchain should be considered as an alternative to traditional methods where decentralization is desired and trust must be maintained across multiple parties.

This document explores the use of blockchain for the exchange of medical information directly between patients, or their brokers, organizations that generate such data (e.g. Patient care facilities, Pharmaceutical companies, etc.) and data consumers such as researchers interested in using data for therapeutic research.

Blockchain Generalized Implementation

Blockchain implementations create an *electronic immutable public* ledger of transactions that prove, assuming certain operational constraints are maintained, exceedingly resistant to malicious alteration.

New transactions (i.e. records) are added to an encapsulated collection (or *block*) of transactions. Extending a block with new transactions is a time-bound process and, once a time interval has expired,

the block is closed to new transactions. The closed block is appended to the ledger and a new block is started to accept new transactions. The ledger, therefore, is an aggregation (i.e. *chain*) of ordered closed blocks. The initial block of the ledger is instantiated in a specialized manner although all subsequent blocks are consistently created, manipulated and closed.

Identical copies of the ledger are maintained across a network of independent participants (a.k.a. *miners*). In addition to a copy of the complete ledger, all miners maintain an instance of the current active block. Each miner receives each new transaction, verifies that the transaction is valid based on prior transactions (e.g., accounts providing payment have sufficient funds) and, if verified, adds the new transaction to their individually maintained block. Once the block closes the miners collectively compare and verify the entire block (explained more fully below) and add the verified block to their ledgers. Each miner then opens a new block and begins a new round of receiving transactions. In this manner, each miner independently maintains an identical copy of the entire ledger.

While incorporating new transactions into a block, the miners also compete to solve a non-trivial mathematical problem, the solution for which (at least in the Bitcoin block chain) requires a straight forward, but hardware intensive, brute force computation. The first, winning, miner to solve the problem closes their block and sends the solution to the other miners. If a majority of miners concur with the solution, i.e. confirm *proof of work* (for Bitcoin, the solution can be verified simply and rapidly), the block is incorporated into the ledger and the winning miner (at least in the Bitcoin blockchain) receives a payment. For the Bitcoin blockchain, the problem is based on a unique hash value generated from the content of the preceding block. As such, any alteration to a single miner's ledger, malicious or otherwise, will alter the hash of the preceding block and cause verification via proof of work to fail. Thus, as long as there are a diversity of miners, the majority of which are trustworthy, and the relative time for solving a block is reasonably small, the potential for illicit modifications to the ledger is exceedingly limited. In cases in which blocks are resolved simultaneously and verified separately by subsets of miners the ledger will bifurcate with an inconsistency in the most recent block. However, this bifurcation will (highly likely) be resolved upon the completion of the subsequent block by selecting the ledger having been verified by the largest number of miners.

For bitcoin, the parties to transactions are represented by encrypted IDs and each transaction is hashed and added to a specialized data structure, a Merkel Tree (https://en.wikipedia.org/wiki/Merkle_tree), within the block. The specialized data structure allows rapid access to the hashed entries and, through combination, creates a unique hash for the block as a whole.

In this manner:

1. The ledger can be made public
2. The ledger is protected from hardware malfunction, multiple miners provide failover
3. Transactions are reasonably protected from tampering
4. Transactions can be validated
5. Miners can enter or leave the block chain without impacting the integrity of the ledger

Assumptions:

1. There is a reasonably large population of miners (assuming reasonable incentive)
2. There are a majority of trusted miners
3. The block time is reasonably small (Bitcoin block time is typically ~6 minutes with the difficulty of the block problem automatically adjusted to maintain a consistent average block time)

The general blockchain model can, in principle, support any type of transaction and thus, has been proposed as a solution to many types of problems beyond those related to currency exchange.

Use Case: *Decentralization of Multi-Entity Research Collaborations*

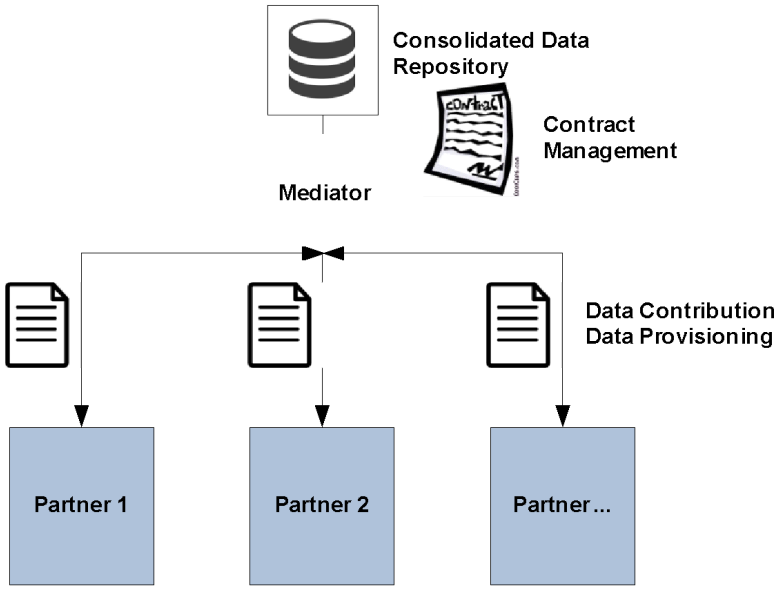
Background

This use case describes a scenario in which several competitors wish to partner to create an aggregated collection of individual medical data including clinical attributes (demographics, endpoints, phenotypes, etc.) and/or high dimensional datasets (gene expression, proteomics, etc.). Each partner will contribute proprietary data that is either already available or will be generated during the duration of the collaboration. The aggregated collection will allow each partner to conduct more informative and highly powered analyses. Each partner will be able to use these data for their own purposes although access to these data, as these data accumulate, will be contractually limited depending on the level of each partner's participation.

Real world examples of large scale multi-entity data sharing collaborations, as simulated in this use case, often use a third party intermediary, such as a law firm, to ensure that contract obligations are met by partners and that data contributed to the collaboration is maintained in a secure environment and used per the constraints of the agreement.

The mediator provides the following fee-based services:

1. Assists in establishing the contractual agreement across participants
2. Monitors/enforces contract compliance
3. Facilitates the entry of new partners/exit of existing partners
4. Manages the aggregated data set
5. Provisions data to partners per the terms of the agreement

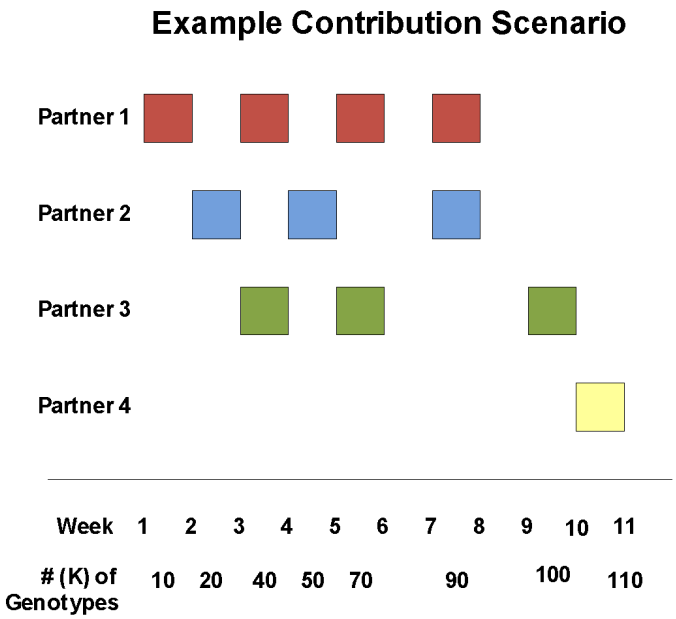


Although this arrangement tends to be operationally effective the overhead service costs are high and third party mediation can introduce delays in executing contract obligations, such as data transfer or on/off boarding partners. Reducing or eliminating these costs would be highly desirable. Additionally, storing the aggregated dataset in a technology environment separate from the individual partners introduces data security risks beyond that of data transfer and further complicates establishing a legal framework for the collaboration. This use case replaces the mediator with a blockchain to realize the benefits of decentralization.

Use Case Definition

A set of data sharing partners have agreed to provide up to 100K subject genotypes (these could be any type of data) over the course of one year. Partners gain access to the existing aggregated data set upon each incremental contribution of 10K genotypes. Each time a partner contributes 10K genotypes, the partner receives a set of keys that allow access to data files that currently exist in the collection. Partners cannot access data contributed after the time of their most recent 10K milestone unless they have contributed the full 100K complement of genotypes. Once a partner has contributed 100K genotypes they will be provided keys to all new contributions. Each partner agrees to load only genotypes having appropriate consent for medical reuse. Once contributed, genotypes cannot be removed from the aggregate collection except under extreme circumstances, such as a deletion request by an individual or an error with regard to consent.

The use case intends to create an environment in which partners are incentivized to contribute data on a consistent basis throughout the time frame of the collaboration. An example is provided in the figure below in which each box represents the completion of 10K genotype contribution.



In this example presented in the above figure, partners 1, 2 and 3 have all contributed genotypes by the end of week three. At this time partner 1 would have contributed 20K with partners 2 and 3 contributing 10K each. At the end of week three partners 1 and 3 have access to the entire 40K genotype collection (each having contributed during week three) while partner 2 has access to only the 20K which were part of the collection at 2. Partner 4 has no access to the data collection at week three. Partner 3 has access to the entire current collection at the end of week 9. This model can lead to potential disparity as, at week 11, partner 4, in their first 10K contribution, gains access to the entire content although they have contributed the least amount of data at this time. Additional rules that promote equity between a partner's data contribution and corresponding access privileges could easily be devised.

Proposal: ***Private Block Chain implementation***

A private blockchain is the proposed implementation for this use case. The actors in the use case include:

Partners
The entities that enter into an agreement to contribute and consume an aggregated medical data set

Individual (Contributors)
People who elect to contribute their own medical data to the collaboration for an incentive

Community Broker
An entity representing a community of individuals having the authority to contribute data for an incentive

The roles in this use case include:

Miner
Maintains and operates a node of the blockchain

Data Contributor
Transfers data to the aggregate collection

Data Consumer
Requests data from within the aggregate collection

Incentive Consumer
Receives direct incentives for the use of contributed data

The mapping of actors to roles is provided in the figure.

	Role	Miner	Data Contributor	Data Consumer	Incentive Consumer
Actor					
Partner		X	X	X	
Individual			X		X
Community Broker			X		X

Partners serve as miners
A private block chain would be implemented with each partner serving as a miner. As such, each partner would create separate mining and ledger environments and compete with respect to resolving proof of work. In this case, all partners know the state of data contribution for each partner and can assure the equivalency of ledgers. The use will run under the following constraints.

1. There is no incentive (payment, priority data access, etc.) for partners to compete to win a block. Partners are incentivized to build a large diverse dataset and this data sharing motivation should be strong enough to maintain data contribution. It is difficult to envision any advantages in promoting mining competition for this use case.
2. The expected time to solve the proof of work problem should be configured to roughly once per week. As there is no value to mining competition and that data contributions will likely follow a burst pattern, the weekly timeframe is surmised to be adequate.
3. Each partner implements consistent hardware to perform proof of work. As per 1. The proof of work hardware can be low performance given the lack of competition. However, hardware supporting verification and data transfer processes would need to be suitably robust and might be implemented differently across partners.
4. The nonce-based proof of work (as is implemented for Bitcoin) can be used for this blockchain. There is no compelling reason to adopt an alternative proof of work problem.

Data contribution and incentives

Although individual incentives are not likely to increase the speed of partner data contributions, such incentives may encourage individuals and community brokers to contribute data leading to a more diverse and powered collection. Although not required, individual and community broker contributors are a potentially worthwhile extension to the core collaboration model.

Partners as data consumers

Only partners will consume data in the context of the collaboration.

Implementation

In addition to the blockchain software operated by the miner, there are a number of other implementation considerations pertinent to the use case.

Ledger Record: Each record of the ledger represents a set of data for an individual and is described as follows.

1. **Contributor_ID:** A unique key that identifies a specific data contributor.
2. **Individual_ID:** An anonymous key to identify an dataset
    a. If multiple datasets are meant to be used in concert (clinical, genotype, proteomic, etc.) then these sets must share the same Individual_ID
    b. Datasets that are generated from the same individual (genotype, targeted sequencing, etc.) but are not intended to be used in concert will be associated with distinct Individual_IDs.
3. **Data Type:** Describes the type of data (genotype, proteomics, etc.)
4. **File Format:** Describes the format in which the file is provided (VCF, BAM, etc.)
5. **File Reference:** A unique identifier for each file that is used to access the file from the contributor
    a. This reference could be a hash of the file and could be used for parity QC when a file is received by a partner.
6. **Metadata:** A list of key/value pairs describing pertinent metadata. These metadata would be consistent for at least each data type and may include basic demographics
7. **Consent:** An acknowledgement that the individual from which these data are derived has consented for the research objectives associated with the collaboration. Although represented only as a Boolean in the use case, consent could be more specialized for individual types of research proposals.

Basic Block Chain Record

Contributor_ID (Long Integer)	**Individual_ID** (Long Integer)	**Data Type** Text/Integer Code	**File Format** Text/Integer Code	**File Reference** Long Integer/ Hash	**Metadata** Key/Value List	**Consent** Boolean

Data Sources

The data files are stored separately from the ledger. Each data contributor will maintain their file sets independently. A partner having earned the right to access a set of files would make an electronic

request to the pertinent contributor asking for specific files using the File Reference from the pertinent ledger record. The requestor would send a public key which the contributor would use encrypt the file are return the file to the requestor. It is expected that file transfer would be automated by default such that any/all files for which partners gain access would transfer without an explicit request at each proof of work resolution and block addition.

Contracts

The following smart contracts will be in force for the use case. The software which executes the contracts would be embedded within the mining implementation.

File Quality Control: Upon the addition of a new record to the block chain the associated file will be available for a limited time for verification to ensure correct formatting. If the verification fails the record will not be added to the block. It is possible that each partner could independently verify each file or, if this is not feasible, the verification could be done through random assignment or as a process separate from the blockchain.

File Privileges and Access: Upon the resolution and addition of a new block the blockchain will update a table to assign new files to partners that have reached a milestone contribution. The table will exist in the block chain software and be maintained by each partner. Perhaps this listing could also be a component of the block hash to ensure that this table remains consistent across partners

Data Wallet

Each data contributor, including the partners, will use a separate application to submit new records to the block chain. Additionally, this *data wallet* could implement the request and transfer of data files from the contributor. Other capabilities of the wallet might include:

1. Pre-verification of data files prior to submission to the block chain
2. Anonymization or interoperability with an anonymization service
3. Accounting of files contributed, requested and transferred, including files rejected due to verification issues
4. Operational status of the blockchain
5. Block chain search (see below)
6. Automated file transformations
7. Refusal of data requests

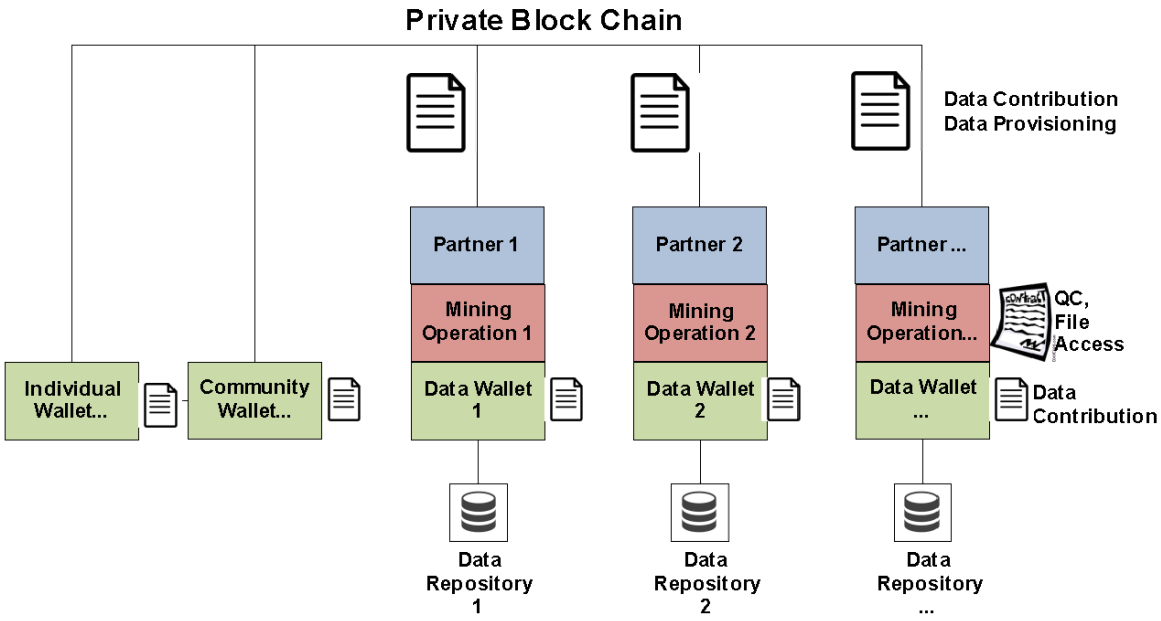
As contributions are managed via an independent wallet, it is straightforward to envision how additional data providers, including individuals and patient communities holding pertinent data, could contribute to the block chain. In these cases, transfer of files from individuals could be accompanied by an incentive cryptocurrency payment.

Given the potential of the blockchain to include contributors who are not partners the wallet would benefit from a search feature/interface (activated at least for data consumers) to interrogate file metadata, format and data type assignment (indexes could be implemented/maintained on the

mining software). Such a search feature would allow requestors to determine the data of most value to their research prior to making a request, especially important should requests require incentive payments.

Additionally, to further promote data contributions from independent communities or individuals, the wallet could provide information regarding how the data was used, i.e. for which research proposals, related disease areas and links to corresponding literature references.

Non partner contributors should also be able to refuse requests should they wish to no longer share their data. It may be worthwhile to reflect this in the ledger although the wallet should provide a suitable solution with respect to decisions that alter data sharing potential.



Activities

The following represent the basic activities performed by the actors in their various roles across the primary system applications, specifically the *Blockchain* software, *Data Wallet*s and

- A partner in the role of *Miner*, in a one-time action, launches the first block of the *Blockchain*.
- A partner in the role of *Miner*, hosting and operating the *Blockchain* software, participates in problem resolution and proof of work.
- A partner in the role of *Data Contributor*, through their *Data Wallet* application, identifies and submits data files to the block chain with one individual dataset corresponding to one record of the *Blockchain* ledger
- A partner in the role of *Data Consumer*, automatically via the *Blockchain* software interacting with the pertinent partner *Data Wallet*, requests and receives data files, from their peer partners, for which access privileges have been granted. The *Blockchain* software manages the encrypted transfer and quality verification.

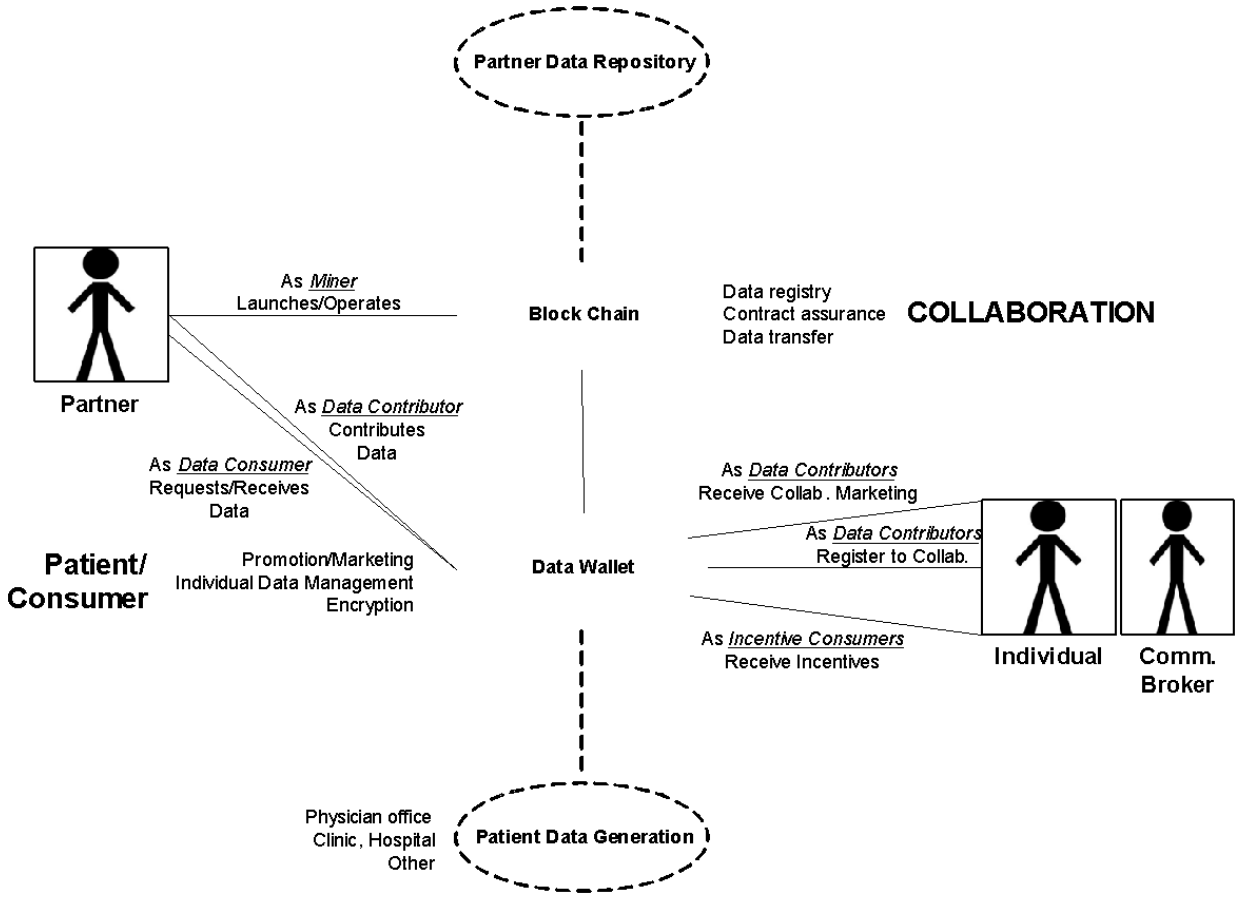
- A partner in the role of Data Consumer, via the BlockChain software, searches specific data sets of interest, particularly from *Individual* and *Community Brokers*, and requests and receives these data files and processes incentives through the *Blockchain* software interacting with the *Data Wallet* of the contributor
- The Partner(s), in the role of *Data Consumer*, will send marketing materials, including push notifications, announcing the collaboration, data desired and incentives (from potentially the *Blockchain* software or another marketing application) to the *Data Wallets* of *Individuals* and *Community Brokers*.

- *Individuals* and *Community Brokers*, in the role of *Data Contributor*, receive medical data into their *Data Wallet* from their physicians (corresponding medical offices, clinics and hospitals), recreational data generators (e.g. 23&Me, Ancestry, etc.), research organizations (clinical studies) and personal monitoring devices (Fitbits, etc.) into their *Data Wallets*.

- *Individuals* and *Community Brokers*, in the role of *Data Contributor*, register the collaboration *Blockchain* with their *Data Wallets*.

- *Individuals* and *Community Brokers*, in the role of *Data Contributor*, using their *Data Wallet* create a record on the Blockchain for each data set they wish to contribute. The *Blockchain* notifies the Data Contributor, via their *Data Wallet*, of either success or refusal of the record with reasons for refusal if applicable. A successful record will allow requests for these data.

- *Individuals* and *Community Brokers*, in the role of *Incentives Consumer*, receive, via their *Data Wallet*, incentives should a *Data Consumer* elect to use one or more of their datasets.



Proposal

Alternative 1: Simulation

The proposal put forward at the original meeting was to create a simulation of the blockchain described in this use case. Such effort would require extending a general blockchain open source implementation (such as Ethereum) to include (at least simulated) capabilities specific to the described scenario including smart contracts for data authorization and subsequent transfer. A remedial Data Wallet implementation would also be required to interoperate with the blockchain instance to enable a reasonable simulated system.

The base proposal may be enlightening and demonstrative of value, although there are no known impediments for applying a blockchain in this case. If a simulation is pursued then the project should be done as efficiently as possible using a shallow implementation for proof of concept only. If there is an example of an emerging collaboration that could benefit from this model perhaps the simulation could move forward to production implementation.

Alternative 2: Build the Data Wallet Application

As this use case was being developed it became clear of the critical role of a medical data wallet for connecting individuals with their own data, from patient care and recreational sources, as well as potential beneficiaries of these data. Although the Data Wallet could be simplified to essentially a secure personal medical information application, there are many nuances with regard to how the wallet is used within this use case that reinforce the potential benefits of endowing the wallet with a rich set of features to maximize the value of individual health data. The features include:

1. Standards for seamlessly interoperating with personal medical data sources (clinics, research facilities, personal accelerometers and health monitoring etc.) as well as research data consumers.
2. The ability to inform individuals of research interests that could benefit from their data and allow data consumers to promote their research proposals
3. Creation of a variety of incentives for individuals from micropayments to awareness of how their data is used and the impact of research to which individuals have contributed their data.
4. A security model that bolsters patient confidence with respect to their data privacy and ensures low risk data sharing.

Building a Data Wallet could be an excellent cross-industry effort to drive blockchain mediated medical and health data exchanges.