BLOCKCHAIN APPLICATION NOTE #1

Healthcare
Intel Health and Life Sciences

Blockchains for Data Sharing in Clinical Research: Trust in a Trustless World

Authors

John Sotos
MD, MS
Intel Health and Life Sciences

David Houlding
MSc, CISSP, CIPP
Intel Health and Life Sciences

Data sharing is now a major theme in clinical research. However, effective, ethical sharing of clinical research data requires trust: research subjects must trust investigators to preserve their privacy, investigators must trust each other to give credit where credit is due, and the integrity of the entire research enterprise must be trusted by all.

Achieving this trust has been challenging. For example, when an international consortium of clinical trialists proposed that investigators be allowed to keep their raw data secret for two years after publication, patient advocacy groups quickly derided the delay as putting the career interests of researchers over the needs of ill patients.[1] Similarly, when faced with the fact that disclosing an individual's genome is always a disclosure of personally identifiable information, some investigators have suggested a retreat from long-held principles of privacy protection, to a new status quo where research subjects in genomic studies will simply have to surrender their genetic privacy[2]—as well as that of their close blood relatives.

Trust-failures like these threaten the growing momentum for widespread clinical data sharing. Can they be mitigated?

We believe so, based on the success of the notably untrusting financial industry in building several well-tested "alternate currency" systems—most notably Bitcoin—atop an emerging and disruptive software technology called blockchain.[3]

Fundamentally, blockchains are simply a time-stamped ledger of data and transactions. They create trust by being secure, public, permanent, and rigorously tamper-proof. Bitcoin is just one application built on a blockchain foundation. Others are possible.

By treating data-sharing as a series of transactions, specially crafted blockchain applications have the potential to flexibly and resiliently solve data-sharing's difficult trust problems. Below, we show how, at a non-technical level. Our overall aim is to demonstrate that information technology is a desirable, and perhaps necessary, ally in designing data-sharing regimes.

Basic Blockchain Concepts

As its name suggests, a blockchain is a sequence of blocks of digital data. Metaphorically, one may envision it as a sequence of boxes into which any digital information can be placed, permanently, irrevocably, and unalterably. Each box has a unique numerical address, and the possibly-encrypted contents of each box are time-stamped and visible to the outside world.

A typical blockchain is "distributed," meaning it exists in multiple copies spread across multiple computers, and any new box added to the blockchain is immediately copied to all copies of the blockchain. Hence, corruption of a single machine cannot alter data. This eliminates the single point of failure of a centralized database, improving the availability and resilience of the blockchain and also helping to build trust.

Blockchains create trust by making trust unnecessary. Their design assures that boxes cannot be removed, and that data in the boxes cannot be changed. No central authority is needed to operate the system or manage the addition of boxes. Anyone can start a blockchain.

Blockchain's open standards, off-the-shelf availability, increasing adoption, and nearly decade-long track record of stable operation make it an attractive foundation for new software.

Giving Credit when Sharing Data

Data-sharing agreements have long been used between investigators. We propose that such agreements, digitally signed by the investigators, be put in a blockchain box, to make them public, eternal, and unalterable. Several benefits would accrue.

First, explicit tracking of academic (or financial) credit becomes possible, as follows. Suppose an investigator who collected clinical data ("Collector") and a data-borrowing investigator ("Borrower") agree, in writing, that Collector will be a co-author on any resulting academic paper published by Borrower. In preparing such a paper, Borrower would add a sentence to the methods section akin to: "This study includes data originally collected from the 125 subjects in Collector's trial (citation), per the terms of agreement deposited in blockchain box NCC.1701." Reviewer, editors, and readers of the paper could then visit the specified blockchain box on the Internet, read the agreement, and confirm that the authorship conditions were properly fulfilled (Figure 1). The permanent visibility of the agreement makes it impossible for one party to unilaterally repudiate it.

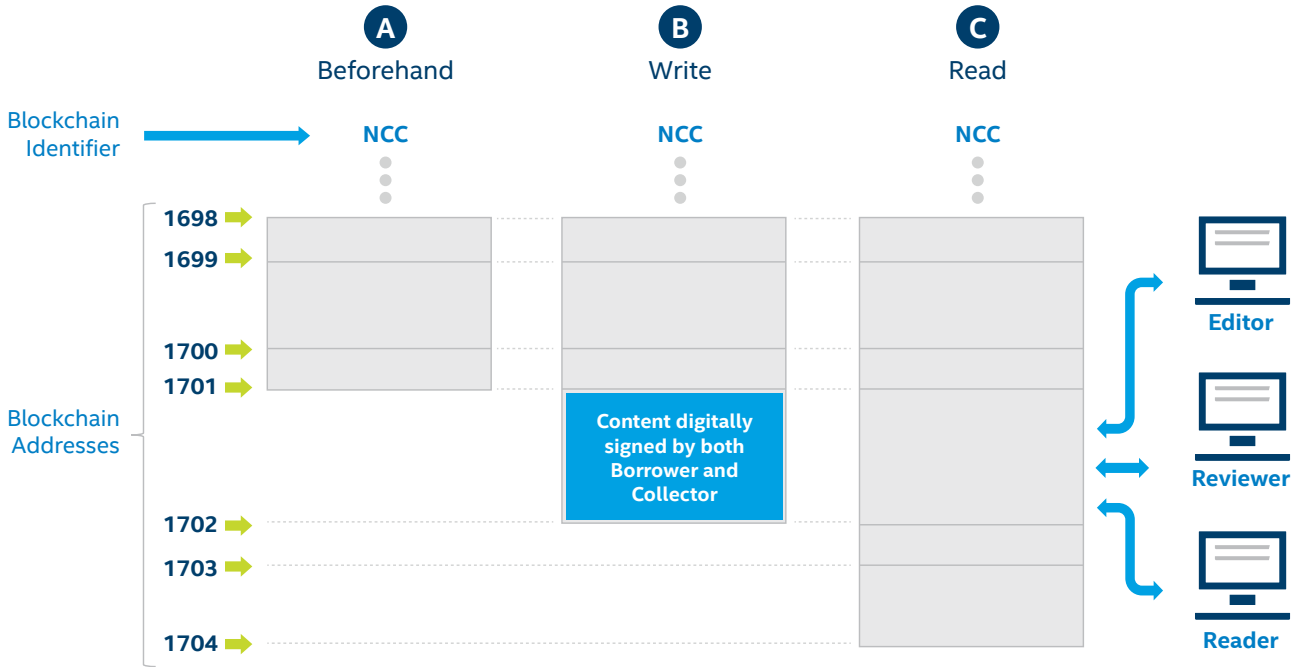

Figure 1. Schematic representation of a blockchain at three points in time. (A) A blockchain arbitrarily named "NCC" has 1700 blocks at this point in time, of which only the last three are shown. (B) A new block has been written to the NCC blockchain at location 1701: the contents of the new block have been digitally signed by two parties, which can be used to verify the origin of the contents. (C) The NCC block can now be read by anyone, including journal editors, reviewers, and readers.

Second, automatically-executing a funds transfer becomes possible if, instead of putting an English-language agreement into the blockchain, the investigators put executable computer software. For example, the investigators could deposit software that checks for specified pre-conditions A, B, and C, and, once they are met, immediately and without human supervision transfers a monetary payment from Borrower's financial account to Collector's.

Third, by adapting the funds-transfer example above, academic credit, as an abstract entity, could itself become a Bitcoin-like alternate currency, created ("mined") by publishing, and transferrable between investigators's accounts. This "AuthorshipCoin" (or whatever) would greatly increase the flexibility of possible publishing arrangements and credit-sharing.[4]

In all of these examples, the blockchain serves as a permanent public ledger of contracts between data collectors and borrowers—an open accounting system from which trust grows. Ideally, a social norm would arise, where any Borrower publication arising from analysis of shared data would cite a blockchain contract with Collector, or risk censure.

Preserving the Privacy of Research Subjects

Any investigator who shares genomic data will be sharing personally identifiable information (PII) about the research subjects. DNA sequences can be linked to real-world human identities[5] and, past a certain size, cannot be anonymized.

Fortunately, software offers an alternative: share transformed data, not raw data. As a trivial example, suppose Borrower wants to know the mean GC content of the 125 genomes in Collector's data set. As part of their data-sharing agreement, Borrower would deposit in the blockchain a software script to compute the answer. Collector would retrieve the script, run it against the genomes, then send the result to Borrower. The raw genomic data would never leave the control of Collector, thereby preserving the genetic privacy of the research subjects.

Several additional benefits are possible. Readers of any resulting publications could inspect the software code, to verify its correctness. Collector could compute a "hash" digest of the result, akin to a fingerprint, and put it in the blockchain, enabling others to detect if Borrower surreptitiously changed the result. If Collector later discovers and corrects an error in the data set, the script could be retrieved from the blockchain and re-run, perhaps automatically, with a notification sent to Borrower.

Such arrangements are not a panacea; they present their own issues. Scripts must be certifiably non-malicious, scripts must suit data formats, some calculations may be inefficient, PII could potentially leak in unusual circumstances (depending on the script), and, of course, overhead would be increased. Fortunately, however, these are technical challenges having technical solutions, and are far more tractable than the inter-human trust challenges they solve.

Conclusion

Data-sharing is universally perceived as an enabler of research collaboration and progress. Ideally, it should be ethical, incentivized, widespread, and frictionless. Although we are today far from that ideal, a public, distributed ledger of data-sharing transactions, implemented in blockchains, will bring us closer to it. These ideals will not come for free, but they will be worth the price.

To keep up with Intel's work with Blockchain, use this Google search:
`blockchain site:intel.com`

[1] International Consortium of Investigators for Fairness in Trial Data Sharing. N Engl J Med. 2016; 375: 405–407. See comments at: http://www.nejm.org/doi/full/10.1056/NEJMp1605654#t=comments

[2] Erlich Y, Williams JB, Glazer D, Yocum K, Farahany N, et al. (2014) Redefining Genomic Privacy: Trust and Empowerment. PLOS Biology 12(11): e1001983. doi: 10.1371/journal.pbio.1001983

[3] Yli-Huumo J, Ko D, Choi S, Park S, Smolander K (2016) Where Is Current Research on Blockchain Technology?—A Systematic Review. PLOS ONE 11(10): e0163477. doi: 10.1371/journal.pone.0163477

[4] As a simple example, explicit authorship shares could replace the traditional author-rank-listing, and could be used to compute AuthorshipCoin accrual. As a more complex example, access to a clinical data set could cost AuthorshipCoin. The cost could change over time: higher if accessed soon after Collector's initial publication derived from the data set, and lower as time passes. This need not restrict early access to well-published investigators. A prospective borrower with insufficient coin could sell a "future," e.g. pledge to give Collector 51% authorship on all resulting publications. The specifics are less important than the wide range of possibilities.

[5] Erlich Y, Narayanan A. Routes for breaching and protecting genetic privacy. Nat Rev Genet. 2014 Jun; 15(6): 409–421

Printed in USA 0317/PDF Please Recycle 335698-001US