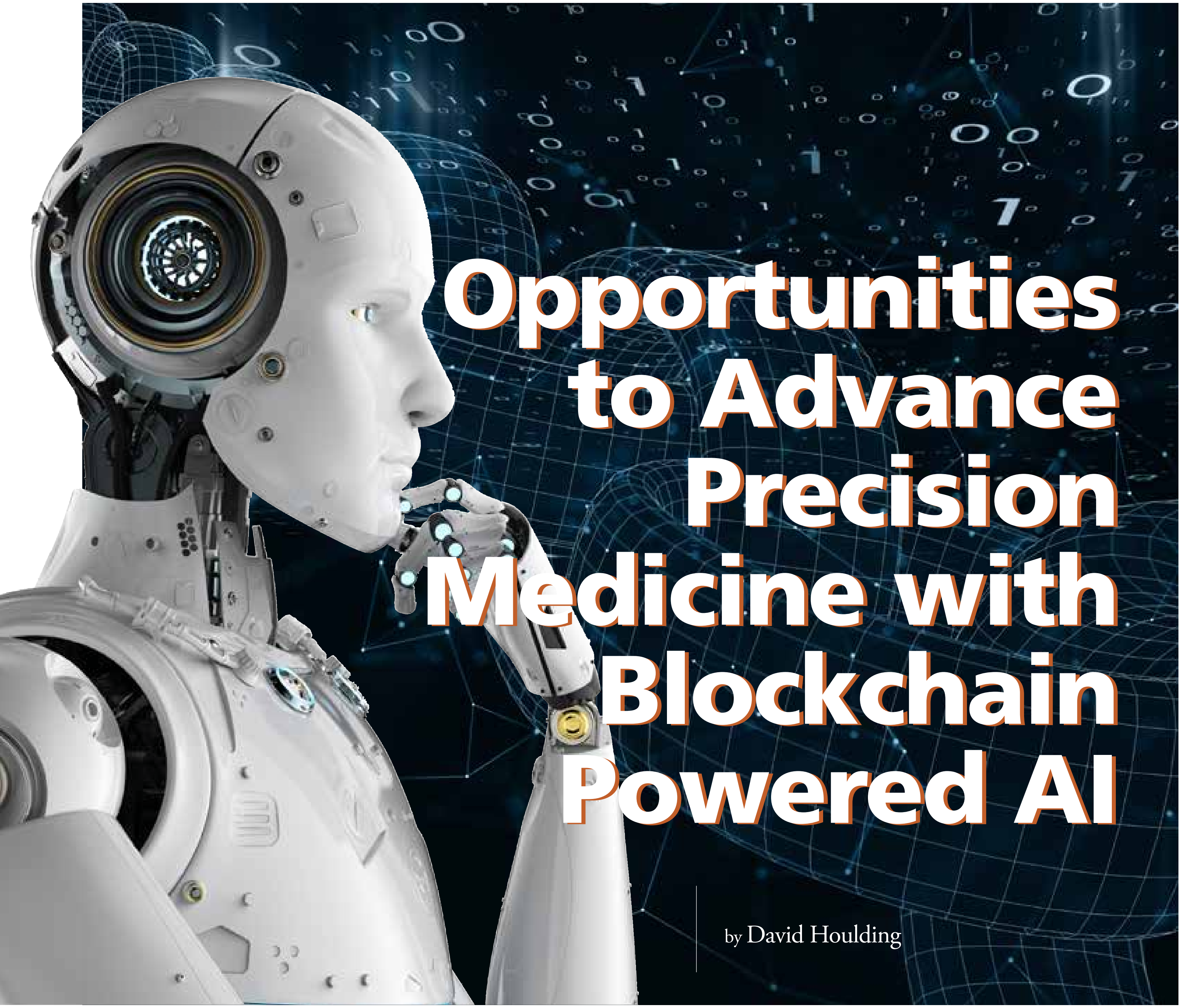
Opportunities to Advance Precision Medicine with Blockchain Powered AI



by David Houlding

AI shows great potential for precision medicine, however, the quality and feasibility of Artificial Intelligence (AI) has often been constrained by limited data. In general, we found that constraints are typically due to sourcing training data from just a single silo within a single healthcare organization. Blockchain offers a pathway forward to advance AI by enabling (and even incentivizing) data sharing and collaborating on the training and testing of shared AI models across a consortium of healthcare organizations. Here we explore how advanced precision medicine can result from the intersection of AI and blockchain – grounded in real genomics examples at several trailblazing organizations pushing the forefront of innovation.

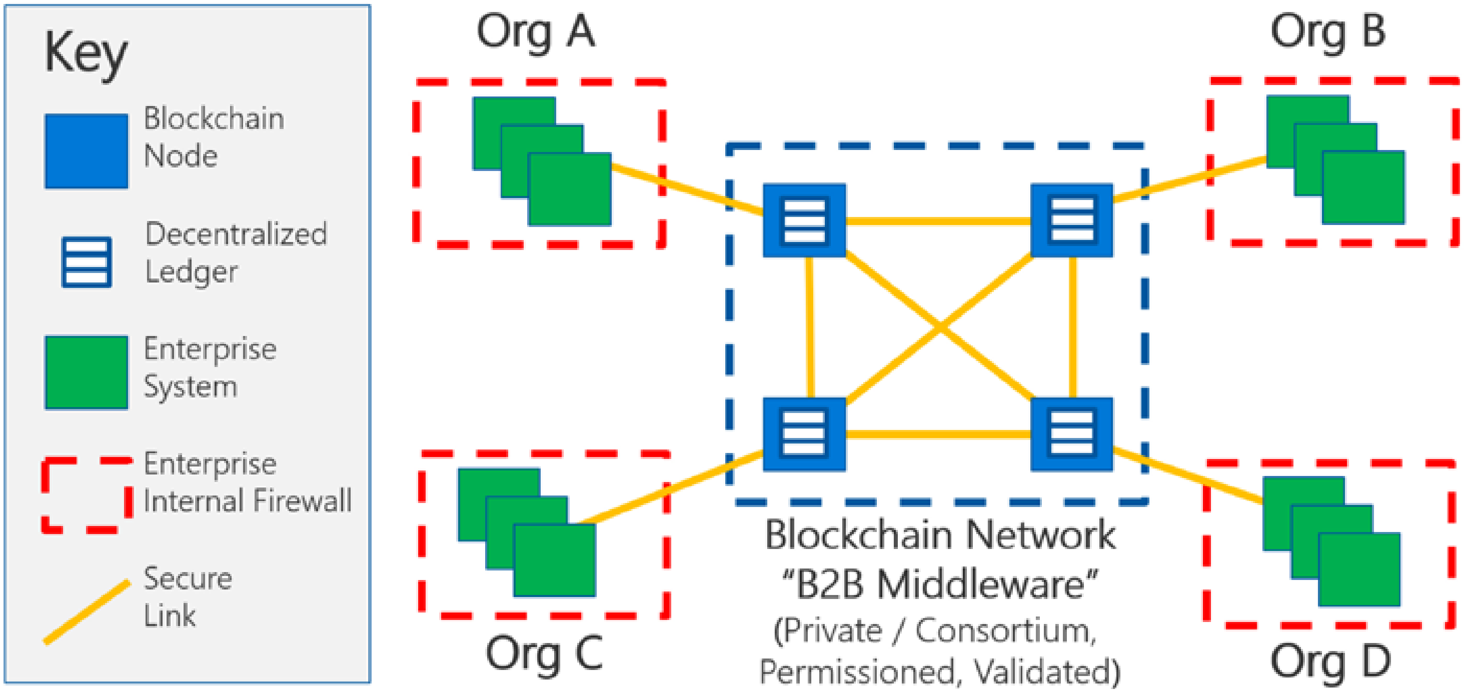


Figure 1: Model of a consortium of healthcare provider organizations collaborating via blockchain.

Sources of Data and Provenance Information is Key for Growth of AI

Artificial Intelligence (AI) and Machine Learning (ML) are data hungry, often requiring millions of high-quality data records to train a new model capable of running inference engines with an acceptable accuracy and low enough error rate. Many reasons contribute to the availability of training data and associated provenance information for AI in healthcare, including the volume, variety, and the velocity at which big data can be gathered. We cite these key factors in **Table 1** and analyze the impact of these factors in this article.

The potential net effect of these factors are *limited quality models, lower accuracy, and higher error rates*. In applications where errors can directly impact patient quality of care, or even patient safety, this is of paramount concern. This in turn makes the application of AI in many areas of healthcare currently impractical, especially where specialized AI models are required, since these must be trained from specialized training data, which is even more rare.

"BY ITS NATURE, CLINICAL DATA CONSIST OF MANY DATA SETS"

Building Trust in New AI Models

To ensure high quality healthcare and great patient outcomes, the AI model must be rigorously tested and the results validated before a new AI model can be trusted and used by a healthcare organization for patient care. Anything less risks adverse patient outcomes, liability concerns, and erosion of institutional trust. This process of building trust is slow, especially if each healthcare organization is doing this alone and, often, redundantly since many healthcare organizations are working on similar AI initiatives concurrently and without a Consortium Master Plan to coordinate collaboration.

What if there were way to enable, and even incentivize, collaboration across a consortium of healthcare organizations to share training data, ML models, test results, and validations of results in order to collectively advance and accelerate the development of AI for precision medicine?

Sharing Training Data Via Blockchain

By its nature, clinical data consist of many data sets. For example, the University of Washington Health Sciences Library classifies data in six key categories: Electronic health records, Administrative data, Claims data, Patient-/-Disease registries, Health surveys, and, finally, Clinical trials data. Genomic data would be in either an EHR or the Clinical trial data if the sequence data were collected as part of the trial itself. The concepts discussed in this article are applicable across all these categories, however, in this article we’ll go deeper into genomic data to illustrate key principles.

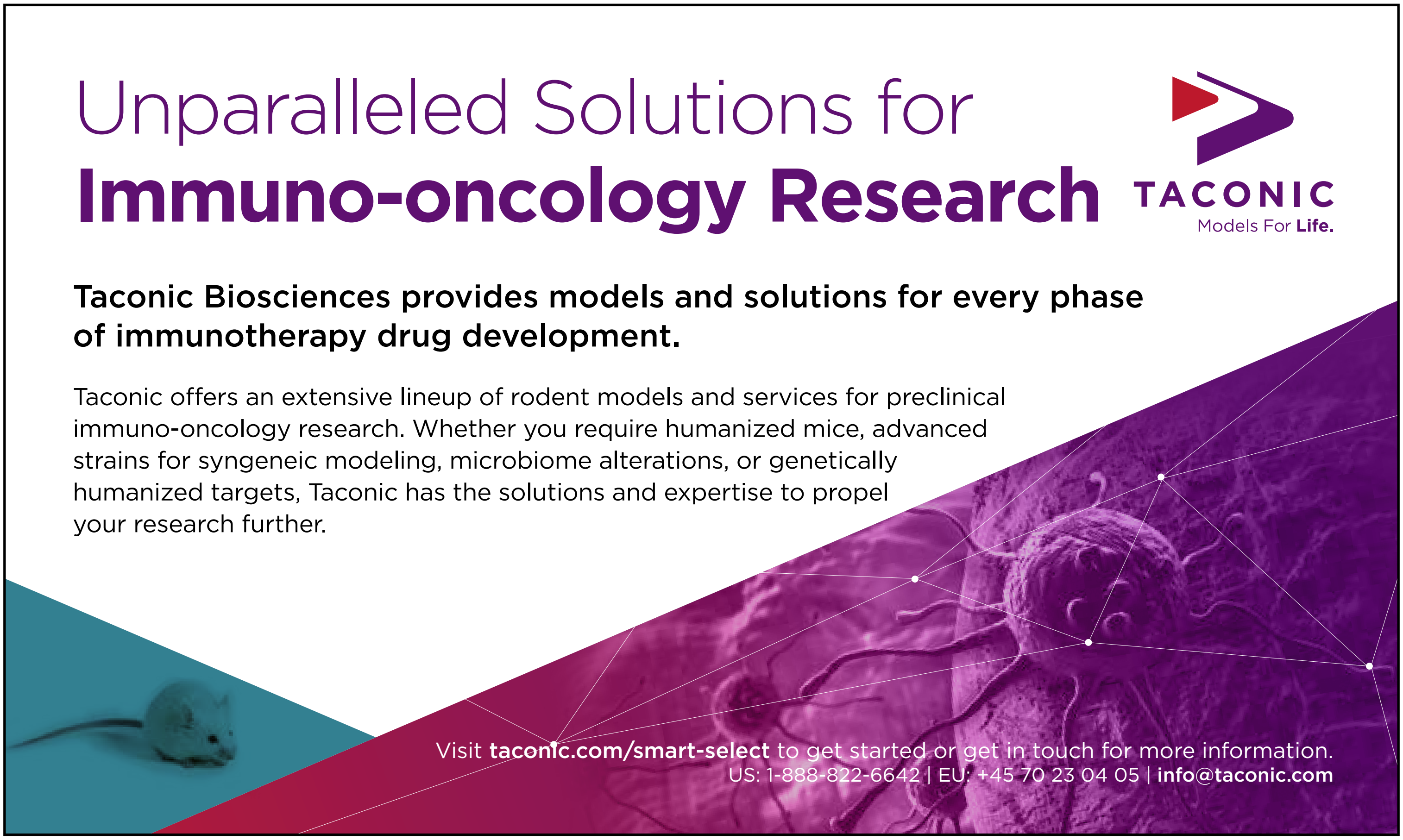
Data volumes in the region of millions of high-quality records are typically desirable to train an AI-ML model with acceptable accuracy. Sourcing this volume of high-quality data for AI models can best be done by a consortium of collaborating healthcare organizations, rather than each organization going it alone. A best practice with blockchain is to agree on data quality standards and limit data stored in the shared ledger of the blockchain to *minimal but sufficient* to support the target use case and achieve the desired business results. This is for both privacy, security, and compliance reasons and minimizing associated risk, as well as for performance reasons. Conversely, an approach of storing all data on the blockchain and figuring out later how to make use of it is discouraged.

Example 1.
Model Use Case of Consortia to Share Data

Consider a use case where a consortium of healthcare provider organizations is collaborating via blockchain to train a new genomic AI model to support precision medicine. The architecture of such a network is depicted in the **Figure 1**; note that this blockchain is a private consortium blockchain so it is not connected to any public blockchain, and all organizations accessing this blockchain are well-known and highly trusted across the consortium.

The vast majority of blockchain pilots happening today in healthcare use private consortium blockchains such as this. Each of the organizations in the consortium, and for each genomic data record they have, stores a metadata record on blockchain including: high level information about the genomic record; provenance information; a pointer to the source of this record (identifying the organization that has it, and a specific record ID); and a hash code that can be used to verify the integrity of the genomic data record. Other organizations in the consortium can then search the blockchain shared ledger; identify records of interest that are applicable to the specific AI model being trained; and can then initiate a secure direct peer-to-peer request to the source organization to retrieve the record, as depicted in **Figure 2**.

Once the record is received the recipient can verify its integrity using the hash code from the associated metadata record on the blockchain. This check ensures that the right record was retrieved and that and there has been no tampering with the record. **With this approach, the actual genomic data is never stored on the blockchain, avoiding associated privacy, security, compliance, and performance issues.** Furthermore, no data records are ever exchanged unless there is a compelling business need to do so, minimizing risks to data in transit, as well as network use for these particularly heavy types of data, where even just a single full genomic



data record can be more than a hundred gigabytes in size.

Example 2:
Model use case of Consortia Identifying New Drug Applications from Sharing Data

Pharmaceutical companies often check if a drug developed for one use may also be safe and effective for another use - for example, a drug for treating liver cancer may also be effective for treating breast cancer based on biomarkers or genomic signatures. Several such success stories already exist where cases have been found by companies from their own data. Blockchain has the potential to greatly accelerate this by enabling the pooling of data among companies participating in the consortium, and identification of alternative uses from pooled data versus only the data of a single organization.

Exchanging Shared Models Rather Than Training Data

In a related approach to collaboration via blockchain, a shared AI-ML model can be passed around each organization in the consortium versus passing around genomic data. Furthermore, a shared AI-ML model can now be readily trained incrementally with records identified from metadata on the blockchain. Additional merits of this alternative approach are that privacy, security, and compliance risks are further mitigated **since no genomic data moves across organizational boundaries**, and network demand is vastly reduced since only models and metadata are passed around rather than heavy genomic data records.

Incentivizing Collaboration and Sharing Using Blockchain

The potential of blockchain goes beyond a platform for secure targeted sharing of information. It can also support cryptocurrencies and crypto-tokens which can be used to incentivize and reward collaboration and sharing. Such a capability can further accelerate the advancement of shared AI models and collective realization of the associated benefits when applied to precision medicine. Further, in any typical consortium there may be organizations of different sizes and different capacities for collaboration, and different volumes and types of data to share. This can sometimes introduce a challenge of fairness where larger organizations may feel they contribute much more value to the consortium via blockchain in terms of resources and data than smaller ones. Cryptocurrencies and crypto-tokens can help alleviate this challenge by pre-negotiated rewards for participating organizations proportional to their contribution through a Master Consortium Agreement (MCA).

Managing Security and Compliance

By design, blockchains afford excellent protection of the integrity of data stored on the blockchain. They also inherently protect the availability of the blockchain network since there is not central single point of failure. However, blockchain does not protect the availability of each blockchain node. As blockchains are used for mission-critical, production use cases, it becomes important to protect the availability of the blockchain nodes which serve as the onramps – offramps to the blockchain data "superhighway". This can be done with redundant nodes across availability zones, load balancing, and automatic failover.

Confidentiality of data can be protected first by keeping the variety and volume of data on the blockchain minimal but sufficient for the target use case, avoiding Personally Identifiable Information (PII) on the blockchain where possible. Private consortium blockchains can be used to limit access to only well-known and highly trusted organizations.

Table 1: Key Factors that Impact Data and Provenance Information for AI Training Sets

1.	**Organizational Boundaries and Silos Limit Data Availability**	Most AI initiatives in healthcare are stunted by a lack of available training data. This in turn comes from the fact that most initiatives source training data from only within a single healthcare organization, and often just a single enterprise system within that organization.
2.	**Privacy and Compliance Can Render Data Off-Limits**	Two issues further exacerbate this lack of availability of training data: patient concerns about privacy and security and healthcare organizations compliance requirements with regulations and data protection laws. In many cases, lack of patient or data subject consent and opt-in to AI research studies can keep certain types of data legally off-limits, further decreasing the availability of data.
3.	**Liability Concerns and Lack of Organizational Support**	Healthcare organizations concerns with the use of AI and ML (e.g., liability, culture, potential displacement of healthcare professionals) can limit their support for AI initiatives, further restricting the availability of training data for AI.
4.	**Lack of Data Provenance Information for Limits Quality**	Often data that is available has a lack of provenance information and is therefore of questionable quality, leaving AI data scientists in a quandary: either relax quality and data provenance requirements to get more training data and risk biased models, or enforce higher standards and accept less data that meets quality and provenance standards, in turn leading to inadequately trained AI models and suboptimal results.
5.	**Make a Master Consortium Agreement (MCA) and a Master Consortium Plan (MCP)**	To obviate some of the issues cited in this table, management leaders and healthcare professionals from across an organization should agree upon an MCA that addresses basic management responsibility, e.g., objective, business rules, consent forms, funding, etc.; an MCP provides for infrastructure (hardware), architecture (software), quality assurance, types of information to be gathered, protocols, etc. Finally, make sure to allow for sharing lessons learned and updating agreements and plans, including new data types coming online.

Encryption and blockchain side chains can also be used to further protect confidentiality and ensure only authorized access to data on the blockchain.

All organizations connecting to the blockchain should have pre-negotiated terms for adequate security. An organization with inadequate security connecting to the blockchain risks introducing a weak link that can lead to a breach or other security incident that can impact not only that organization, but the whole blockchain consortium. To ensure adequate security, risk assessments and audits can be done proactively through an MCA for each organization in the consortium that is connecting to the blockchain, and the results of these assessments and audits shared across the consortium members to build trust. Should a risk assessment or audit uncover unacceptable risks, these risks can be proactively and collectively mitigated to ensure adequacy and minimize risks of security incidents or non-compliance with applicable regulations or data protection laws.

Improving the Quality of AI with Blockchain

The quality of inference depends on the quality of the AI model, which in turn depends on the quality of the training data used to create the model. Blockchain can be used across a consortium of healthcare organizations to track data provenance and audit log information that can in turn be used to improve the quality of AI for precision medicine. Data provenance, specifically the origin of the sample, the data, and any related parameters of acquisition (e.g., sample source and handling history, chain of custody, sequencing machine make, model, and serial number used to generate genomic data) can be tracked on blockchain and used with other metadata to search, discover, and filter out the most appropriate, highest quality data suitable to train an AI-ML model.

As noted above, data provenance can record the chain of custody of data where appropriate to ensure traceability. Audit log information about the training data used in a given model can also be tracked on blockchain, and, in the event a model is later found to be biased, the blockchain can be consulted to determine exactly what data was used to train and remediate. Where AI is used for inference to support precision medicine, the quality of patient care and patient safety depend on the integrity of the AI models. The integrity of training data, AI models, test results, validations of results, and so forth can be protected through hash codes on the blockchain immutable shared ledger, and these can be verified at any time to ensure absolute integrity and protect against unauthorized modification or deletion.

Closing the loop: Building Trust in Precision Medicine Models Faster

New AI models for precision medicine must be tested to ensure accuracy. A healthcare organization doing this alone may take months or years to adequately test and establish sufficient trust in a new model for it to be deployed and used for patient care. Blockchain enables a new method of collaboration where multiple healthcare organizations in a consortium can collaborate on testing a new model, each sharing inference results, validations of results, and enabling the consortium to perform more testing in a shorter timeframe, building trust in the new model faster, and enabling its use and realization of association benefits sooner. This kind of open, near real time, transparent collaboration on AI for precision medicine across a consortium can also detect errors or biased models faster, enabling remediation sooner and minimizing wasted effort that would otherwise be spent on a faulty model.

Engaging Patients and Incentivizing Participation in Research and Sharing of Data

The cost of high quality, full genome sequencing is falling rapidly. Concurrently, the rise of AI for precision medicine empowers healthcare professionals with new actionable insights in near real time, enabling them to deliver precision and personalized medicine in ways not possible just a few years ago. These opportunities cannot arrive soon enough, given the rampant increase in chronic diseases worldwide and the rising cost of healthcare - now at nearly 20% of GDP in the US! Many large employers now find that healthcare costs are the second greatest cost they face. Increasingly, such employers are opting to self-insure for the healthcare of their employees. These organizations are strongly motivated to provide proactive, preventive healthcare options that mitigate the risks of costly healthcare episodes that degrade the quality of life for patients, their employees.

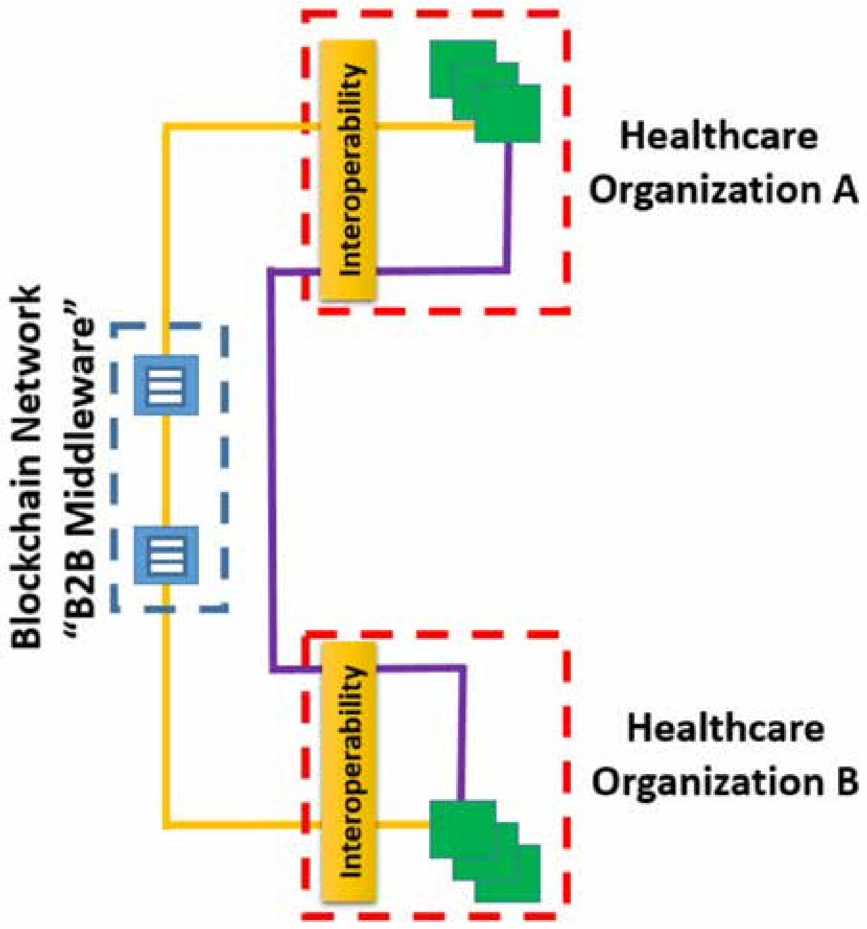


Figure 2: Diagram for a consortium Model that enables data storage, sharing, discovery, requesting, and retrieval.

Blockchain can engage and empower patients with privacy and control of their data. Via blockchain, data subjects can review their data, amend as needed, manage consent to opt-in or opt-out, or even request to be forgotten (requiring deletion of their data), a requirement of GDPR and similar data protection laws. Furthermore, blockchain cryptocurrencies or crypto-tokens may be used to incentivize patients to engage, and participate in clinical research studies, which can, in turn, increase data available for research and improved results.

Several examples of these capabilities already exist

Encrypgen provides a DNA Marketplace running on blockchain, empowering patients with their data, enabling them to list their data, make it visible to clinical researchers, and be paid for their data with cryptocurrency (called DNA) in case researchers wish to purchase access to their data. Further opportunities for patients to engage in precision medicine and genomic research are provided by Genomics Personalized Health, in collaboration with global genomics powerhouse Macrogen.

The collaboration provides employees of self-insured companies the opportunity to have their full genomes sequenced, learn from the results, and have the option to subsequently participate in the Encrypgen DNA marketplace to engage in further research. David Koepsell, CEO and Co-Founder at Encrypgen states:

> *"Genomic data is highly valued, with direct to consumer genetic testing companies earning hundreds of millions selling their customers' data. This data is valued as part of essential, basic*

> *research by pharmaceutical companies, especially when combined with meta-data supplied by users. New models, like ours, aim to eliminate the middleman and allow individuals better control and direct reward."*

Rapidly growing genomic data across the population empowers improved precision in defining genetic variants associate with individual traits. We note comments by two CEOs about the power of partnering for capabilities in this space. First, Ryan Kim, CEO at Macrogen USA, notes:

> *"Population-scale genomic sequencing is now possible with drastically reduced, and continually rapidly lowering cost of sequencing, and Macrogen's high-quality data analysis capacity in collaboration with Microsoft for both clinical and research purposes".*

And Michael O'Reilly, Co-Founder at GPH, adds:

> *"We are particularly excited to partner with the Microsoft team as they can assist with security, privacy, and compliance concerns associated with genomic data, both in the US as well as globally."*

Privacy and Compliance with Regulations and Data Protection Laws

It is paramount in these initiatives to maintain privacy of patient information, even empower patients with their information. Systems must meet such requirements to comply with regulations such as HIPAA (US), and data protection laws such as GDPR (EU) and GINA (US). Applicable regulations and data protection laws depend on the types of data handled, in particular PII (Personally Identifiable Information), and the geographical locations of the data*.

Conclusions

Potential capabilities of AI and Blockchain to impact precision medicine are highlighted in **Table 2**. Precision medicine provides a clear path to improve patient outcomes and engage and improve their experiences. AI is a powerful and critical tool required to derive actionable insights in near real time from the rising tsunami of data, including genomic data and more. AI is currently limited by data silos and limited sharing of data that could be used to train better AI-ML models. To power precision medicine to the next level, blockchain holds major potential to enable and incentivize data sharing and collaboration across a consortium of healthcare organizations. ■

Table 2: Highlights of capabilities and potential of AI and Blockchain for Precision Medicine described in this article

1.	AI has matured to become a powerful and trusted tool to derive actionable insights in near real time from the rising tsunami of data, including genomic data and more.
2.	Currently, AI is limited by isolated data silos; these data represent a pent-up potential that needs to be shared among silos and organizations to create the best AI-ML models.
3.	Blockchain technology holds the power to advance precision medicine to the next level by enabling and incentivizing data sharing and collaboration across a consortium of healthcare organizations. Individual could also benefit by taking part in incentive plans for data sharing.
4.	An inherent advantage in sharing data in a blockchain consortium is the ability to de-identify data. All parties benefit by advancing a technology platform with minimal risks for conflicts of interest.
5.	A shared AI-ML model can be passed around each organization in the consortium versus passing around genomic data. The shared AI-ML model can now be readily trained incrementally with records identified from metadata on the blockchain.
6,	Merits of this alternative approach are that privacy, security, and compliance risks are further mitigated since no genomic data moves across organizational boundaries, and network demand is vastly reduced since only models and metadata are passed around rather than heavy genomic data record
7.	Actual genomic data is never stored on the blockchain, avoiding associated privacy, security, compliance, and performance issues.
8.	Master Consortia Agreements should be developed for data sharing terms and Master Plans for system designs and protocols.
9.	Blockchains inherently protect the availability of the blockchain network since there is no central single point of failure.

David Houlding is the Principal Healthcare Lead on the Microsoft Azure Industry Experiences Team. David has more than 24 years of experience in healthcare spanning provider, payer, pharmaceutical, and life sciences segments worldwide, and has deep experience and expertise in blockchain, cloud computing, privacy, security, compliance, and AI / ML. David currently serves as Chair of the HIMSS Blockchain in Healthcare Task Force, a group of ~100 leaders from across healthcare worldwide, collaborating to advance blockchain in healthcare. David also currently serves as an advisor to both the British Blockchain Association and Lifeboat Foundation. David has led the successful creation and deployment of a wide range of solutions to help reduce the cost of healthcare, improve patient outcomes, experiences, and engagement. Prior to joining Microsoft in 2018 David served for over 10 years at Intel Health & Life Sciences where he was the Director of Healthcare Privacy & Security, responsible for enabling healthcare organizations worldwide to achieve compliance with regulations and data protection laws, and implement effective privacy and security programs. In his current role at Microsoft, David works with key partners and industry influencers to enable healthcare organizations make use of cloud computing and related technologies to reduce healthcare costs, and enable new transformative healthcare use cases to improve patient outcomes, leveraging strategic technologies such as such as AI / ML, blockchain, IoMT (Internet of Medical Things), and others. David has a proven track record for innovation with 5 patents granted by the USPTO. David currently holds the CISSP (Certified Information Systems Security Professional), and CIPP (Certified Information Privacy Professional) credentials, and has a Master of Applied Science in Data Compression and Digital Signal Processing from Simon Fraser University, Canada.

Disclaimer: This is not a summary of legal advice. Please consult your legal counsel for advice on your compliance requirements.