Applying artificial intelligence analytics to compiled datasets:

Straightening out the long and winding patient journey

An interview with **Tom Sharrock, PhD and Frank Seo, Lifebit**

A PATIENT'S JOURNEY can seem a tortuous, random walk, full of sample collections, test results, more sample collection and tests, diagnoses, and treatments – and sometimes the journey is repeated (iteratively) on different paths! For some conditions, the diagnosis requires a battery of tests creating disparate datasets from several sources – imaging scans, blood tests, gene panels, biopsies, etc. The narrative overlaying the data gathering leg of the journey is that these data will lead to the correct diagnosis and selection of the most suitable treatment based on the patient's profile and test results, but data gathering itself can become a significant leg of the journey! In short, how to make the link from data to diagnosis to a precision medicine with the highest likelihood of a successful treatment.

How can the healthcare enterprise best use these datasets so that the journey ends with tangible benefits to patients and payers alike – for example, coordinated, value-based care plans and healthier patients?

Many sectors are involved in addressing this question. In this article, we ask leaders at Lifebit in to comment on how datasets can be harnessed to yield possible solutions. Lifebit works with large, compiled, diverse datasets for comprehensive data reviews and assessments. We pose the following questions to them to learn how artificial intelligence can use these large, compiled databases to advance the development of precision medicine for use in the clinic.

Data Compilation Approaches and Artificial Intelligence Applications to Healthcare

Q1. How does your organization define the respective types of database compilation structures deployed – integrated, federated, other? Why has your organization taken the chosen approach?

A. To achieve ultimate data security, it is critical that data remains within an organization's direct control. Fortunately, organizations are now turning to innovative federated solutions that virtually link distributed databases to enable the analysis of data where it sits, avoiding data movement across environments. This empowers researchers to run joint queries, analysis and AI algorithms over distributed data with federated compute resources.

Lifebit has pioneered the first fully federated multi-party genomic system that uses the zero-trust principle[1] to minimize potential threats and data compliance breaches. Our novel federated approach allows two or more parties in a distributed system to perform secure simultaneous analysis without exposing sensitive data to risks, applying highly optimized privacy-preserving and secure computational techniques to safeguard genomic data sharing and analysis.

Our federated system is in stark contrast to legacy systems, which require the regular movement of sensitive data into a centralized environment, such as a data lake or warehouse. These legacy approaches are expensive, redundant, slow, and are heavily reliant on security certifications to guarantee safe handling of classified data. With centralized approaches, the original data custodian loses control and ownership of their data and must trust the new data custodian (usually third-party vendors) to treat the data appropriately. We asked ourselves – why should data ever need to be transferred and copied to third-party environments in the first place? In fact, it doesn't need to be, as federation enables analysis of distributed data without it ever moving. This generates a positive, and often exponential, net effect that enables more questions to be queried on larger datasets, enabling healthcare providers and pharmaceutical companies to gain valuable and actionable insights."

Q2. Can you provide a short description of the strategy you use to query compiled data?

A. At Lifebit, we believe that data stays the most secure when it remains within its own local environment. For this reason, we never physically move and compile data. Instead, when Lifebit's technology runs analysis pipelines and Machine Learning algorithms, we access the disparate data in a federated manner. The system orchestrates analysis and computation directly in each data custodian's environment and combines the results in a federated manner, ensuring data security is never compromised, **Figure 1**. The advantage of federation, in this context, is that you get the same benefits as physically compiling the data, without the **data ever being passed or transferred between parties**. Therefore, there is no need to rely on third-parties to maintain patient confidentiality nor data security.

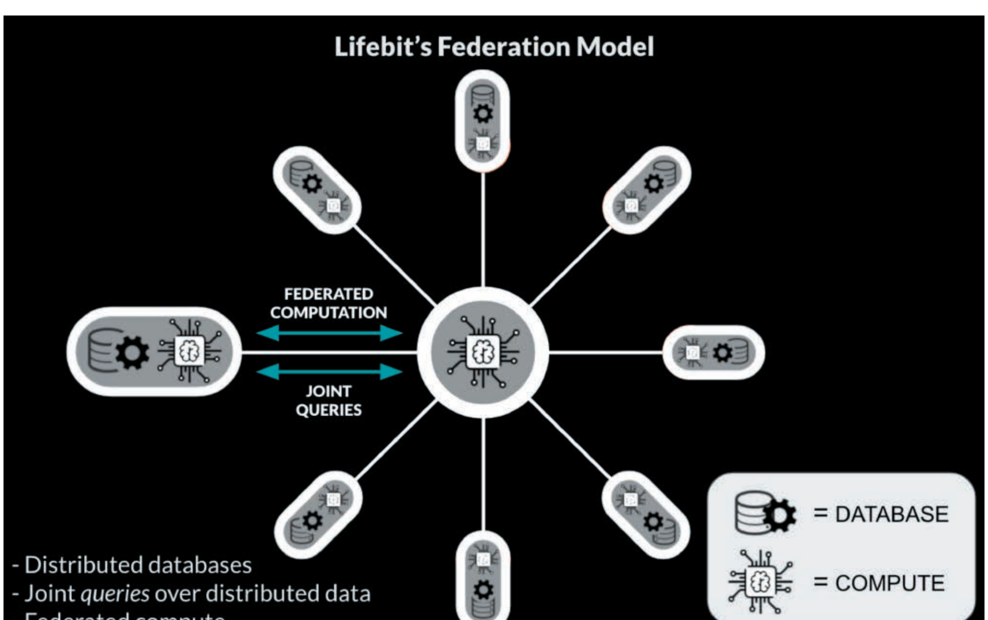

Figure 1: Schematic representation of Lifebit's federation model. The system orchestrates queries and computation directly in each data environment and combines the results in a federated manner.

Q3. Database design and data selection are critical to their utility. Can you address the design and selection criteria that go into your respective databases? In addition to genomic and clinical data, are billing datasets included as part of the compiled databases?

A. Lifebit has developed a proprietary database architecture that can be deployed in multiple tenancies. Lifebit supports multiple data modalities and is able to ingest any type of clinico-genomic, RWE, imaging, or billing data. Due to Lifebit's federated infrastructure, the data is never physically transferred into our own environment. Therefore, our database structure is secure by design and does not require data custodian responsibilities to change. At the same time this enables the use of any relevant dataset, both public and private, in order to best serve each project.

Q4. To what extent have cloud-based platforms been qualified or validated in trials or in practice, whether by the FDA or other regulatory agencies? Or by internal testing?

A. There are some well-known examples of cloud-based AI applications that are already being used in the clinic. These include medical imaging AI systems that help doctors diagnose heart problems and head triage AI systems for neurovascular emergencies. In the United States, once the efficacy of AI-based systems has been demonstrated, they currently require approval in the form of either pre-market clearance (510(k)), *de novo* classification, or pre-market approval. However, it has been noted that the current regulatory landscape needs adapting to be suitable and ready for the abundance of AI/ML software tools that are soon to be implemented in clinics.

Payers, Providers, and Patients

Q5. Can you address broadly how AI is applied to your compiled healthcare databases? For example, how are analyses eventually translated to precision medicine recommendations to a healthcare provider or to a patient to aid in a decision?

A. In reality, the industry is still a little ways off from precision medicine recommendations becoming commonplace within the clinic. However, national precision medicine initiatives such as Genomics England, focusing on patients with suspected undiagnosed rare disorders and hereditary cancers, will certainly help accelerate precision medicine efforts in the clinical context.

We are most excited about the prospect of combining genomics analysis and machine learning to help better diagnose rare diseases. This is an area of precision medicine where we expect platforms to be validated and deployed to the clinics within the next 3 to 5 years. In this context, federation is again essential, as genomic datasets, electronic health records (EHRs) and diagnostic test results, that are required for diagnosing rare diseases, are often locked in the location where they are generated. For instance, a patient's EHR data generally stays within their healthcare provider's software system, whereas their sequencing data remains within the sequencing laboratory's compute infrastructure. By using federated AI, disparate data can be integrated without ever having to move the data from their original location, resulting in patient insights without compromising patient security and privacy.

Q6. How does your approach ensure that privacy and security are safeguarded in cloud structures, especially for the compiled clinical data and electronic health records?

A. The healthcare and life sciences industry is at considerable risk of cyber security threats, and account for approximately 40% of data breaches. Therefore, it is essential that large-scale precision medicine initiatives take substantial precautions when planning out their infrastructure, to minimize outside threats to their troves of data. In addition, regulations pertaining to personal health data, including genomic data, are becoming more strict due to the nature of the data – health data is immutable. Once there is a data breach, there is no going back and individuals cannot change their genetic makeup or health records to protect their privacy.

Acting as a cloud operating system built under the central pillar of federation, our Lifebit platform technology has unmatched security as it is the only system designed to run analysis over encrypted data in our clients' own environment. This basically eliminates unnecessary data movements and circumvents the need for inflexible centralized platforms which data must be transferred into.

Q7. Who owns the patients' data under these compiled systems? Or are parties' rights muddled by this process? Would patients be able to retain rights to their data under these systems?

A. Lifebit takes a unique approach and does not own the patient's data itself. Therefore, the rights of the data used for federated analysis always remain with the data custodian or the patient themselves. Lifebit simply provides access to the sensitive data in a secure and scalable manner allowing it to connect distributed and diverse data around the world.

Q8. Could this approach trace out a patient's journey – that is, a forward/backward timeline that includes hospital and doctor visits, prescriptions, and future steps? How might that journey be presented to a patient or a doctor?

A. Longitudinal datasets enable researchers to track the patient's timeline and understand their medical history in context. This enables our AI systems to recognize patterns and features during a patient's journey that are potentially indicative of a particular diagnosis. The temporal context of data adds massive value to AI systems as algorithms can utilize multiple data points, rather than a single snapshot, to generate clinically relevant insights.

Q9. Do healthcare providers recognize the potential value of this approach?

A. For AI applications in particular, we believe that diagnosis and treatment algorithms will become allies that assist, but never replace, healthcare professionals. Our vision is that when an individual enters the clinic, their genomic sequence, EHR and diagnostic testing results will be integrated into an AI-powered system that can help diagnose their specific condition. Furthermore, once a diagnosis is complete, the system will also recommend the most suitable personalised treatment for the individual. The technology advancements required for these systems are ready, however the next hurdle is the actual validation and implementation of these systems."

Q10. Would you like to make any final or summary observations?

A. Cloud-based AI applications have the potential to transform the patient journey. AI will help accelerate diagnosis and ensure patients receive the most suitable treatments at the best time. However, despite the technology being readily available, reaching and realising the full potential of AI powered care will require the healthcare industry to organise itself around topics like data access and standardisation, the ethical considerations of AI and regulatory standards of AI software solutions. JoPM

Dr. Tom Sharrock

Tom is the AI Engagement Manager at Lifebit. Tom works at the interface of Precision Medicine, Drug Discovery and AI, helping pharmaceutical companies implement, and realise the value of, AI technologies. Tom holds a PhD in Developmental Cell Biology from the University of Cambridge and a BSc in Biochemistry from the University of Bristol where he was awarded the Herman Watson Prize.

Frank Seo

Frank is the head of strategy at Lifebit and a recognised authority in life sciences software and data best practices. He has served in industry leadership positions across Strategy, Business Development, Professional Services and today focuses his energy on bringing precision medicine solutions to rare diseases via AI and Genomic data insights.

For further reading

CDC: Federated EHR Data for Public Health Surveillance, Including Cloud and Open APIs, https://www.cdc.gov/surveillance/pubs-resources/dmi-summary/ehr-data.html

FDA Poster: Enterprise Business Data Analytics (EBDA) Platform – FDA's Multifaceted Shared

Federated Analytical Ecosystem, https://www.fda.gov/media/142021/download

Keywords

Federated learning, edge devices, cloud storage, federated cloud structure, integrated databases, databases, blockchain, interoperability, data compilation, artificial intelligence, diagnostics hypothesis generation, fragmented data

Reference

1. Zero-trust is an IT security model that helps prevent data breaches by eliminating the concept of trust from an organization's network architecture.