AI/ML to Generate Medical Insights… While Maintaining Patient Data Security and Privacy

Machine Learning (ML) and Artificial Intelligence (AI) have come to the forefront of data analytics with the promise of generating new medical insights. However, for healthcare data, patient data security is paramount due to the GDPR and similar regulations. Traditional methods of data consolidation for Machine Learning/ AI modelling into a single warehouse or a data lake are often not possible even with anonymised data due to data protection rules. The new, emerging alternative is a federated in situ data analytics construct in which healthcare data are anonymised within the care facility and accessed via a secure cloud application. This paper demonstrates how this ensures data security within the original data domain, while allowing analytics for modelling and AI to be applied in a federated fashion.

Introduction – The Goal of Healthcare RWD is to Gain Insights in the Patient's Clinical Journey

Increasingly, understanding the patient journey through their healthcare system is being recognised as critical to modeling and understanding different patient outcomes and care metrics. Patients are complex, with not just a single factor between diagnosis, treatment, and outcome. Especially for chronic diseases, the social demographics, the pharmacogenomics , the time between diagnosis and treatment, the distance to a healthcare provider, existing comorbidities and health risks play in the longitudinal metrics of care and probable outcomes.

Healthcare has become multifactorial and is by nature an open system, with dynamic, sometimes unpredictable, and often chaotic behaviour. The recent and ongoing COVID-19 pandemic is a perfect example of this: the evolution and spread of the vaccine has not been predictable or anticipated, nor has the uptake of vaccine treatments been global or sufficient on a voluntary basis to prevent the further spread of the virus.

A doctor's first assessment of a patient often refers to any previous records possible with the first review looking for any immediate changes or reported conditions. However, this retrospective review, must now become more detailed and more extensive, especially considering multiple conditions or comorbidities the patient may present. This is often more than a single physician in a single visit can manage but is increasingly available through digital data sources (provided below) that physicians, caregivers, epidemiologists, and healthcare researchers can use to model, test, and validate various healthcare questions.

In their article on understanding the care pathway/patient journey, the authors point out that "Quantitative analyses carried out to generate deeper insights into any unmet needs and patient subpopulations that may benefit most from the new treatment… may involve:

- pharmacy claims data analyses
- electronic medical record database analyses
- retrospective patient chart reviews
- analysis of registry data
- cohorts or longitudinal studies

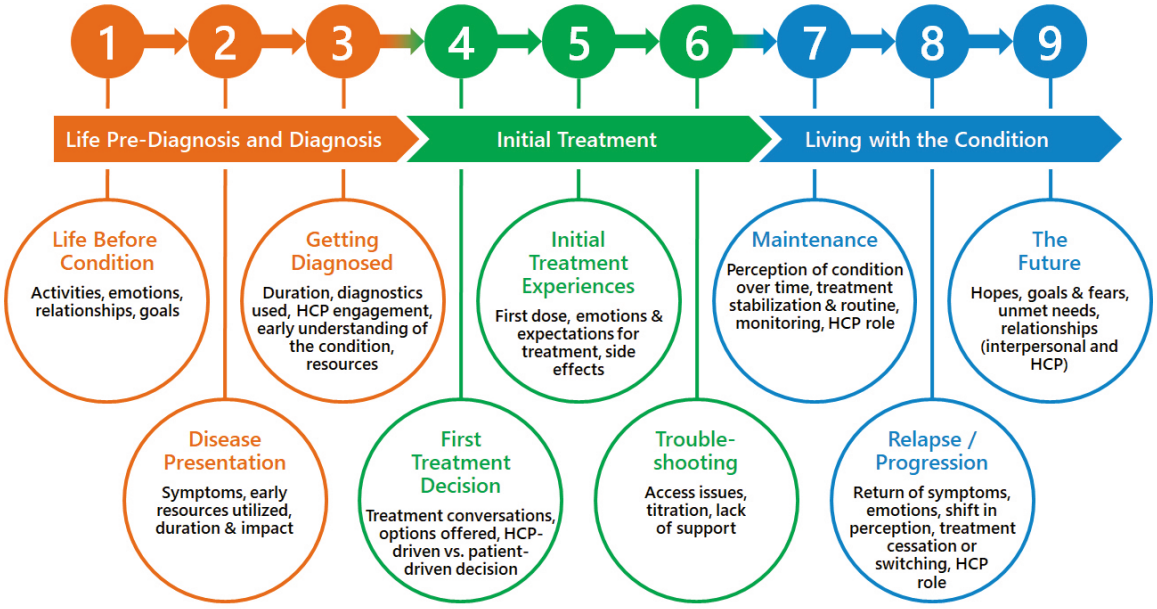


Figure 1[1] A typical care pathway starting with initial symptoms, diagnosis and progressing through treatment methodologies into various outcome metrics. The overall patient journey can be predictive to outcomes and overall patient care metrics.

- epidemiology or treatment pattern studies
- burden of illness studies."[1]

The advent of the digitalisation of patient care data has much improved the prospects of showing a physician a unified overview of a patient's medical care journey. However, the relatively recent General Data Protection Regulation (GDPR) in the EU is a herald for tighter personal data protections than before, and this has now to be accommodated for us to achieve this goal.

Two Regions, Two Models — How GDPR Has Changed What's Needed

With over 330 million citizens, the US block of healthcare data largely exceeds any other country, except China and India, and is the single largest high reimbursement country in the world. The EU plus 1 countries, which traditionally means the UK, France, Germany, Italy, and Spain, are also considered as high healthcare reimbursement markets, but both individually and collectively their size is still slightly less than the USA.

	Population
Germany	83,783,942
United Kingdom	67,886,011
France	65,273,511
Italy	60,461,826
Spain	46,754,778
TOTAL EU5	**324,160,068**
Total USA	331,002,651

Table 1: Population of the EU5 and the USA[2]

Although together slightly smaller than the USA, the disparate healthcare systems, languages, regulations, and standards of care in the EU, tend toward more fragmented healthcare data sources and RWD availability compared to the USA.

Therefore, the use of healthcare data across these two large blocks of high reimbursement healthcare patient populations, that are largely separate and independently regulated, has become highly challenging. When it comes to the use of real-world patient data for the generation of medical insights, HIPAA healthcare data regulations (in the USA) allows greater secondary use including implied consent, while GDPR regulations (in the EU) require explicit patient permission for data use and have strict provisions for personal and private data protections.

Digital Healthcare as the Next Forefront

While digital data is clearly important to healthcare, what comes into question is its application across disparate healthcare systems and regulations. There are increasingly more data and more capabilities unlocked by them. However, technological enablement across one healthcare system, such as the US, does not immediately mean global enablement across separate, disparate, and differently regulated healthcare systems. GDPR has a significant impact on data access and usage, not only within the EU but regions adopting similar regulations, and even within the US.

The globalisation of digital healthcare data does not necessarily mean equal access to healthcare data. Data access and use permissions differ across different parts of the world, and even country to country, disparate systems of healthcare are still the rule, and are just as fragmented as the general accessibility of healthcare itself on a global scale.

Our purpose is to examine the disparate systems for high reimbursement healthcare in the US and the EU, both of which have their pros and cons.

HIPAA and GDPR Differences in a Nutshell

In 2012, a now infamous story appeared of how the retailer Target identified from shopping patterns that a family's teenage daughter was pregnant.[3] Shopping habits and consumer data being used to segment and target clients based upon their purchasing patterns are now common not only in retail sales.[4] However, these purchasing patterns are also a key metric in credit card fraud protection programs,[5] in the same way that patient prescription patterns can be used to understand a patient's care journey. Social media companies such as Facebook, Google, and others readily use their consumer data to develop targeted consumer demographics for their own messaging and to sell advertising.

One of the key differences between HIPAA and GDPR is that the latter requires explicit permission from an individual for their personal data to be collected and used (unless covered by specific GDPR exemptions), as ownership intrinsically belongs to the individual. GDPR imposes significant penalties for collection and use of personal and private data without proper consent. Whereas in the US and other countries, the use of a data application on your smart device includes an inherent agreement in the Service-Level Agreement (SLA) code for the application that data can be collected and re-used, in the EU, the philosophy is that the consumer controls third party use of their own data and can decide to retract their permission at any time.

The impact on our story is that patients always have implicit rights to and ownership of their data. This impacts the use of personal and private information, as well as pseudonymised data, as these carry a potential risk of potential re-identification to protect the security of healthcare data.

Data Harmonisation versus Data Interoperability

US healthcare data are extremely harmonised and readily available through primary resources and a variety of data resellers as claims and EMR data. There is much more variation in the data from the big five European countries based upon local language and coding specifics, and subject to local and national data use restrictions in addition to the GDPR data privacy and protection provisions. US data use provisions allow sale and direct ownership transfer of data, whereas under the GDPR provisions and patients' rights, the actual transfer of ownership rights is not possible. This requires the creation of a structure for interoperability of data where harmonisation is not possible.

Data Sharing without Sharing

In the US, large-scale data aggregation is readily possible and common; however, in the EU, as we have seen, this is significantly more difficult due to tighter patient privacy regulations and has further challenges in data harmonisation across multiple languages, coding, and reimbursement practices. Increasingly, the best option is to maintain the original healthcare data *in situ* and develop searches and integration models across a federated network of originating data sites, regardless of whether they are government or healthcare institutional sites.

The result is that a federated network capability in the EU holds more potential for success (this aligns with the federated structure of the EU, itself) than the harmonised data and common data structure of the US. The EU is a hybrid system of negotiated agreements between governments and a supranational union and

therefore more like a federated network. The United States is a single constitutional federal republic. In some sense this mimics Alexis de Tocqueville's comments on the tyranny of the majority in the US.[6] The US healthcare system is the largest, has the most available data, and is the most easily accessible, and therefore sets the global benchmark. By comparison, the EU's structure and privacy provisions do not allow it similar capabilities or clout.

The Application of AI to Healthcare Data

The inherently differential approaches to patient rights and data protections, which lead to different approaches in both technology and data application use cases between the US and EU, are also evident in the application of ML/AI in Healthcare in the two regions.

In the United States, the large population living under limited privacy controls and a harmonised healthcare system means that healthcare data is plentiful and readily available, so traditional AI methods of data aggregation, pooling and sampling are possible. In the EU, multifarious coding, languages, ontologies, and data movement restrictions make traditional AI methods impractical. The solution in the region would have to be in the various alternatives to the transfer of data, such as Data Fabric technology and Federated Learning. The remainder of this article will discuss these approaches.

Data Fabric

Data Fabric is a data architecture methodology that creates relationships across various metadata points within disparate or even unconnected data sources, and thereby allows specific relationship maps to be created and followed. An example in the context of healthcare data would be the ability to link patient care patterns in localised electronic healthcare records (EHRs) with geographically specific pharmacy claims data, and then with overlapping physician registry information by Zip code, thereby enabling the mapping of specific physician–patient diagnoses and treatments with the direct costs of care by treatment centercentre and location. The specific connectivity between the disparate data sources does not have to be common data elements, or primary keys, such as in a data warehouse, but rather common metadata patterns that can be mapped or linked into a relational "fabric" across the differential data landscape.

An example of this would be a federated EHR network across different countries and languages where a composite phenotypic cohort model could be searched for across the differential metadata linkages.

Data Fabric Rare Disease Use Case

Rare diseases are by nature often undiagnosed, misdiagnosed, and untreated. In addition, rare diseases are often heterogenous in symptoms and therefore hard to diagnose, leaving room for ambiguity in diagnoses. The delay in diagnosis that arises makes it very difficult for patients, their families, and caregivers to manage their medical journey. Studies show that the impact of a rare disease is much wider than on the affected individual and represents a significant challenge for the healthcare system itself.[7]

In a survey of patients and caregivers in the USA and the UK, patients reported that it took on average of 7.6 years in the USA and 5.6 years in the UK to get a proper diagnosis, during which patients typically visited eight physicians (four primary care and four specialist) and received two to three misdiagnoses.[8] Of the 7,000 known rare diseases, 90% do not have an FDA-approved medication, which means patients must live with no treatment or go with off-label use of existing medicines to treat their symptoms.[9] Patients with rare diseases can live up to 20 to 30 years before diagnosis, or even entirely undiagnosed during their lifetime.[10,11,12]

The Data Fabric composite cohort model, which may represent several different phenotypic expressions of, for example, a rare disease, will link different symptoms, treatments, and combinations that an undiagnosed rare disease patient has had but not responded to. This can be used as a predictive and relational data fabric to potentially find rare disease patients that have never been correctly identified.

The authors of a recent white paper explain that "Data fabrics are particularly useful for deep learning use cases because they reduce "fuzziness" that often results from algorithmic training across numerous types of data."[13]

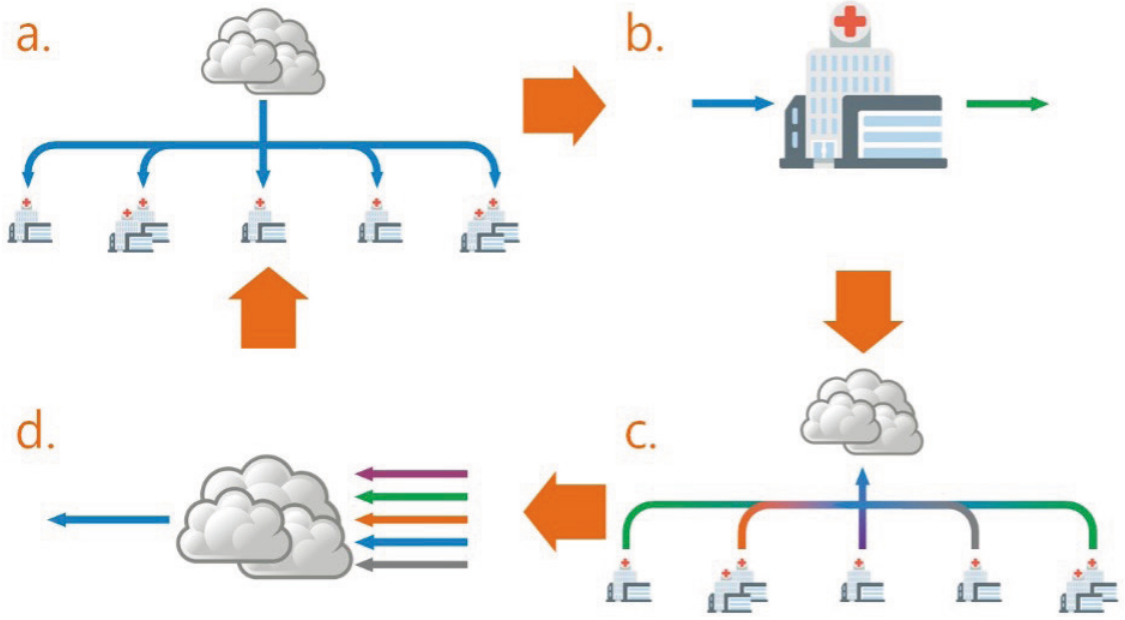


Figure 2:[14] *A schematic of federated learning where (a) a model is distributed across various nodes from the central cloud, to individual institutional incidences, with (b) as an example institution, wherein the model learns from the local data source. In step (c), the model improvements from all the various individual (b) nodes are shared back to the federated cloud architecture. In (d), the various models are consolidated before being redistributed again (a) in a repetitive cyclical pattern.*

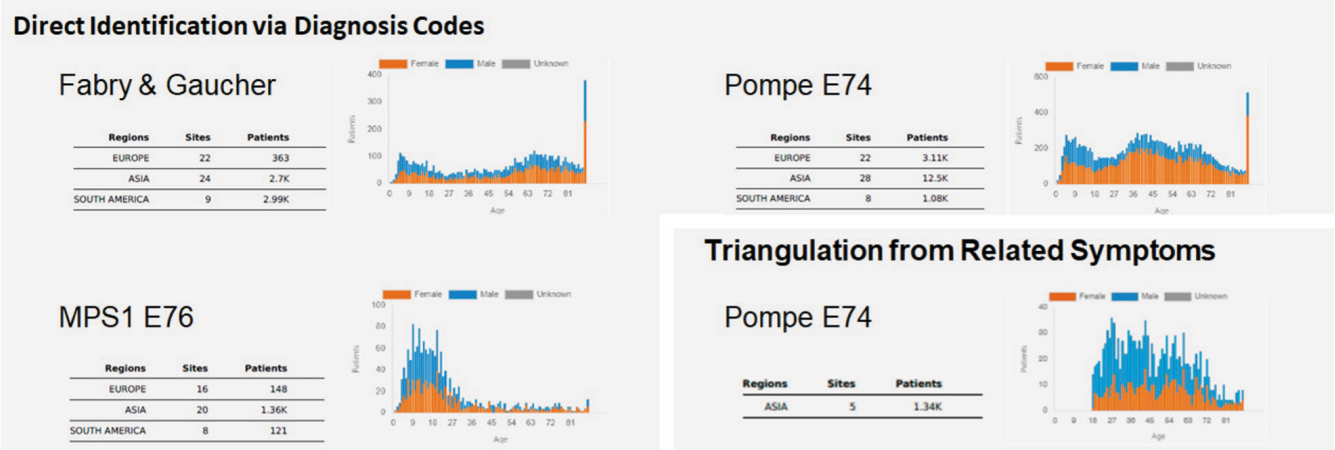


Figure 3: We applied these approaches to a sponsored search for Fabry, Pompe, Gaucher and Mucopolysaccharidosis Type 1 diseases in Turkey and in the United Arab Emirates, using an existing real-world data platform linking an international network of hospitals.[18] The figure above shows traditional direct identification through diagnostic codes, and bottom right, triangulation and predictive outcome metrics using related symptom or phenotype models to highlight underdiagnosed Pompe patients.

Federated Learning

Federated Learning is the architectural framework based upon a single global server with decentralised data across many differential client servers. The goal of Federated Learning is to apply discrete models on multiple client servers and allow them to iterate and learn across the disparate data centers and learn collectively through the central global server. The advantage of this methodology is that data are not aggregated or pooled but stay locally in the original host server and all that is transacted is the model outputs or learning from the federated framework. At the same time, models can adapt across disparate data centerscentres and iteratively learn across the aggregate without data pooling.[14]

The advantages to healthcare institutions as well as research sponsors for federated learning are significant, as highlighted below:

Healthcare Institutions

Data stays secure with no transfer rights.
Personal and Private information remain secure.
Data access is selectable.
Encourage cooperation.

Sponsors

Access to data that is otherwise not accessible.
Strict access to only anonymised records.
Data access is for specific purpose.
Creates collaboration.

Federated Learning Rare Disease Use Case

Typically, you need to look across around 10,000 features in a full EHR to find a relevant patient, reliably. With typical ML methods and standard toolsets, you expect to have to have around 50,000 "labels" (examples of patients with the disease in question) to allow the machine learning to generate a reliable model. As can immediately be gathered, this is an impossible threshold, as, by definition, the rare disease patients are very rare and often misdiagnosed, so "hidden" within the system.

The solution is to use the federated nature of the partner hospital network as the backbone for a federated learning layer over a cloud infrastructure. This is a breakthrough for patients that are likely to be held up in a lengthy "diagnostic odyssey," since the models can be adapted to the healthcare systems now on our own platform through our partnerships with hospitals around the world.

It is critical to construct prediction models which are both accurate and interpretable. As in all medical applications, it is essential that clinicians understand the basis for the predictions and recommendations of decision-support systems. One way to increase interpretability of the complex models produced by modern ML algorithms (e.g., deep learning, ensembles) is to identify which predictors/features are 'important' to the model's predictions and to quantify this importance. Within rare disease, this means looking at the patient clinical journey and identifying cognitive biomarkers, digital biomarkers and medical biomarkers that drive a mechanistically predictive rare disease model:

- **Cognitive** biomarkers are objective measurements that can be used to track the progression of a disease or the outcome of a treatment.[15]
- **Digital** biomarkers are where the actual data are informative in some way about the disease.
- **Physical** biomarkers are phenotypic features of the patients that are predictive of the disease.

These biomarkers are discovered by the model learning process, and we often find them out only as the model improves, so we can subsequently derive clinically interpretable models. The overall metric is to develop phenotypic models based upon actual patient journey that represent all the possible presentations, symptoms, and medical conditions, separately and in various combinations. The advantage of this approach over a federated network is that these models can be developed, shared, and learn (evolve) in a collaborative or data-private process for collaborative learning, institutional incremental learning, or cyclical institutional incremental learning.[16] The point is that federated learning enables incremental and progressive modelling and model learning across discrete datasets rather than nodes securely and effectively. By working with and across smaller datasets, the federated network creates a greater networked database.[17] Interestingly, Federated Learning has been shown to reach performances comparable to traditional centralised data model analytics across diverse therapeutic research topics (heart failure, diabetes, MIMIC-III, SARS-CoV-2, Avian Influenza, Bacteremia, Azithromycin, and Tuberculosis), while preserving privacy.

As in the case of Federated Learning to identify rare disease patients that have been underdiagnosed, we have applied these approaches in sponsored research for lysosomal storage disorders including Fabry, Gaucher, POMPE, and Mucopolysaccharidosis

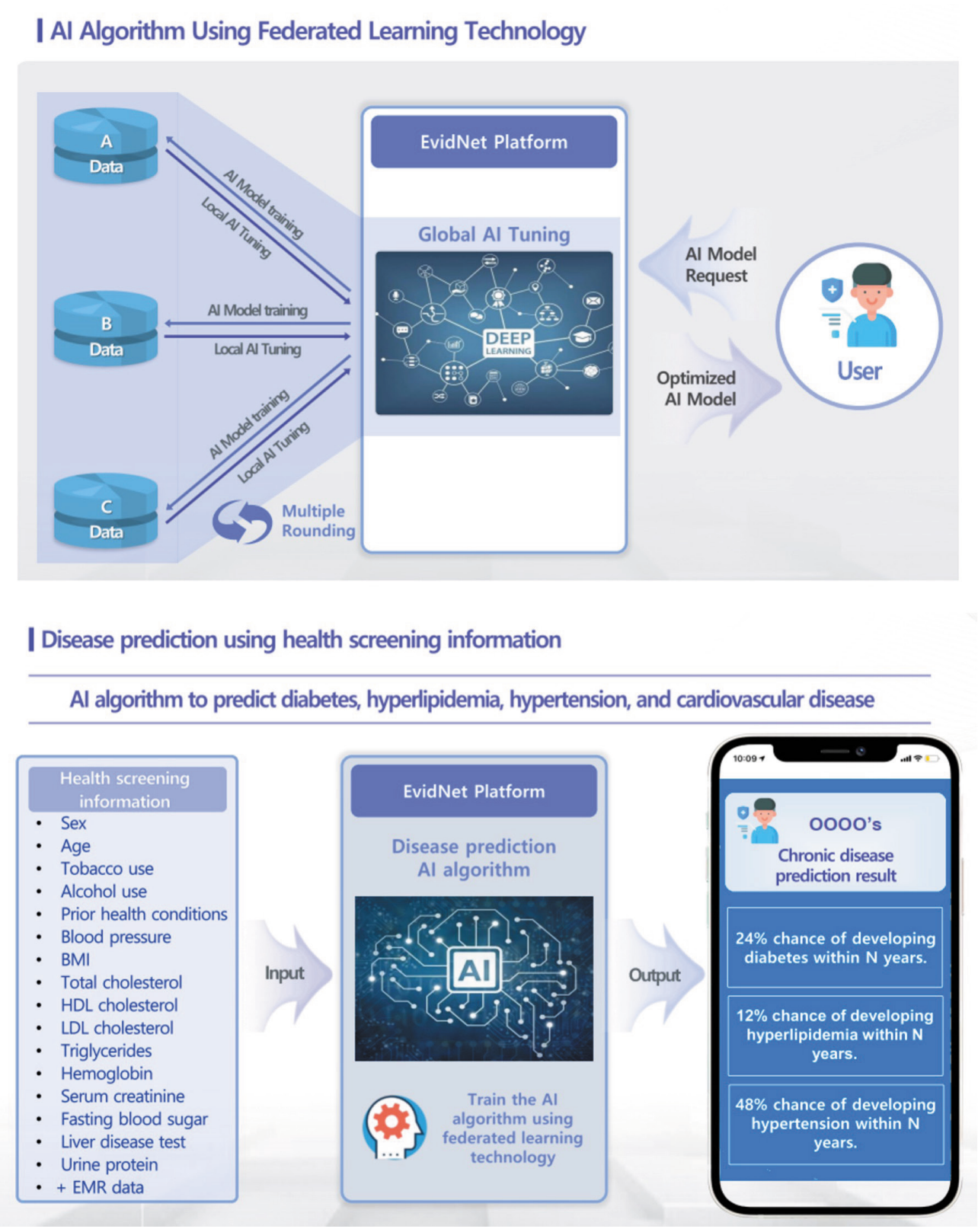


Figure 4: AI/ML predictive federated learning models developed by EvidNet across a Korean hospital network to predict chronic disease onset based upon healthcare screening metrics, with modelingmodelling and global AI optimisation based upon individual site model learning.

Type 1 (MPS1) in Turkey and the United Arab Emirates. The iterative models are being tested and now being clinically validated in studies, in which selected patients are tested for diagnosis as part of a clinical outreach study. The triangulation of potentially underdiagnosed patients using related symptoms/phenotypes/biomarkers, represents the effectiveness of learning to understand digital signals within the patient journey to better triage and flag patients who may not normally be identified as part of traditional patient screening for diagnostic review.

In a further example of Federated Learning, the Korean company EvidNet has developed a federated learning capability across their hospital EHR network. In this use case, the AI/ML algorithm is being used to develop disease prediction models based on health

screening inputs to predict the likely onset of chronic diseases. Figure 5 shows an example of the data flow, global optimisation of the AI/ML model and high-level model metrics.

Conclusions

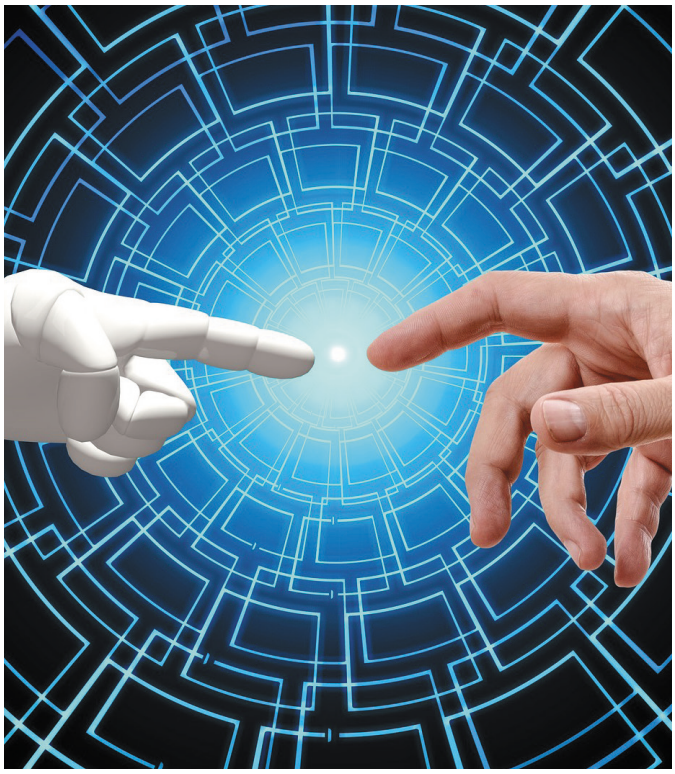
This article has discussed how disparate healthcare systems and access to data often dictate different strategic approaches to analysis and modelling. Circumventing these constraints can be achieved with Data Fabric and Federated Learning, in cases where data cannot be readily extracted or consolidated. These methods are proving to be effective and comparable to traditional centralised models, meaning that different approaches to data access and modeling can be effective and comparable. Likely, not one approach necessarily can work on a global scale, but each in concert can enable perspective and insight that contribute to our global vision and understanding of healthcare and patient care, specifically.

The larger aim of this review is to create the understanding that no single methodology is necessarily better, and that any solution approach needs to take into account access to data, heterogeneity within the data and the extent of harmonisation possible.

The global need is for digital enablement of healthcare, better insights, patient treatments and outcomes. The patient clinical journey is available across many EHR systems in a SMART hospital reference (https://healthcareglobal.com/hospitals/what-smart-hospital). The key is making this accessible and useful for patient care and stratification, not solely care reimbursement.

REFERENCES

1. Joyeux A, Olivaris R, Understand the care pathway/patient journey, available at https://rwe-navigator.eu/using-the-navigator-decision-support-tool/clarify-the-issues/understanding-the-patient-journey/
2. Statistics taken from https://www.worldometers.info/population/countries-in-europe-by-population/
3. Hill K, How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did, February 2012, available at https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/?sh=413c93b16668
4. Duhigg C, How Companies Learn Your Secrets, February 2012, available at https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html
5. Baboo SS, Preetha N, Analysis of Spending Pattern on Credit Card Fraud Detection, IOSR Journal of Computer Engineering (IOSR-JCE), e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 17, Issue 2, Ver. 1 (Mar–Apr. 2015), PP 61-64, available at https://www.iosrjournals.org/iosr-jce/papers/Vol17-issue2/Version-1/J017216164.pdf
6. Mansfield HC, Winthrop D, Alexis de Tocqueville, Democracy in America. Chicago: University of Chicago Press; 2000
7. Drake D, Finding and Treating Rare Disease Patients in a Global Digital Haybale, Journal for Clinical Studies, Volume 12 Issue 4, September, 2020.
8. Shire Report 2013, Rare Disease Impact Report: Insights from patients and the medical community, 2013, available at: https://globalgenes.org/wp-content/uploads/2013/04/ShireReport-1.pdf
9. Toth Stub, S, Conquering Rare Disease – Should taxpayers keep paying to develop drugs for unusual disorders?, 2020, available at: https://library.cqpress.com/cqresearcher/document.php?id=cqresrre2020012400&type=hitlist&num=0)
10. Mehta A et al., Fabry disease defined: baseline clinical manifestations of 366 patients in the Fabry Outcome Survey, European Journal of Clinical Investigation (2004), 34, 236–242
11. Mistry PK et al., Timing of initiation of enzyme replacement therapy after diagnosis of type 1 Gaucher disease: effect on incidence of avascular necrosis, British Journal of Haematology, 147, 561–570
12. Muenzer J et al, Ten years of the Hunter Outcome Survey (HOS): insights, achievements, and lessons learned from a global patient registry, Orphanet Journal of Rare Diseases (2017) 12:82



13. How Healthcare Organizations are Improving Time to Insights with Data Fabrics, TechTarget Custom Media White Paper, available at How Healthcare Organizations are Improving Time to Insights with Data Fabrics, available at https://www.delltechnologies.com/asset/en-us/solutions/industry-solutions/industry-market/how-healthcare-organizations-are-improving-time-to-insights-with-data-fabrics.pdf
14. Xu J, Glicksberg BS, Su C, Walker P, Bian J, Wang F, Federated Learning for Healthcare Informatics, Journal of Healthcare Informatics Research, November 2020, available at https://doi.org/10.1007/s41666-020-00082-4
15. Torous J, Keshavan M, A new window into psychosis: The rise digital phenotyping, smartphone assessment, and mobile monitoring, Schizophr Res. 2018;197:67-68. doi:10.1016/j.schres.2018.01.005.
16. Sheller MJ, Edwards B, Reina GA, Martin J, Pati S, Kotrotsou A, Milchenko M, Xu W, Marcus D, Colen RR, Bakas S, Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data, July 2020, available at https://doi.org/10.1038/s41598-020-69250-1
17. Sadilek A, Liu L, Nguyen D, Kamruzzaman M, Serghiou S, Rader B, Ingerman A, Mellem S, Kairouz P, Nsoesie EO, MacFarlane J, Vullikanti A, Marathe M, Eastham P, Brownstein JS, Aguera y. Arcas B, Howell MD, Hernandez J, Privacy-first health research with federated learning, npj Digital Medicine, September 2021, 4:132; available at https://doi.org/10.1038/s41746-021-00489-2
18. Drake D, Rudolf C, Strategies Toward Identifying Undiagnosed Rare Disease Patients, poster presented at Frontiers of Pediatric Genomic Medicine, April, 2021.

Douglas Drake

Douglas Drake, MS, MBA, is originally a life science researcher with a passion for digital enablement of better patient care. For over 30 years, Douglas has worked in various aspects of diagnostics, therapeutic research, drug discovery and global business development. He has broad experience in transformative technologies, data sciences, digital healthcare and applying these to improving patient engagement and the patient journey.