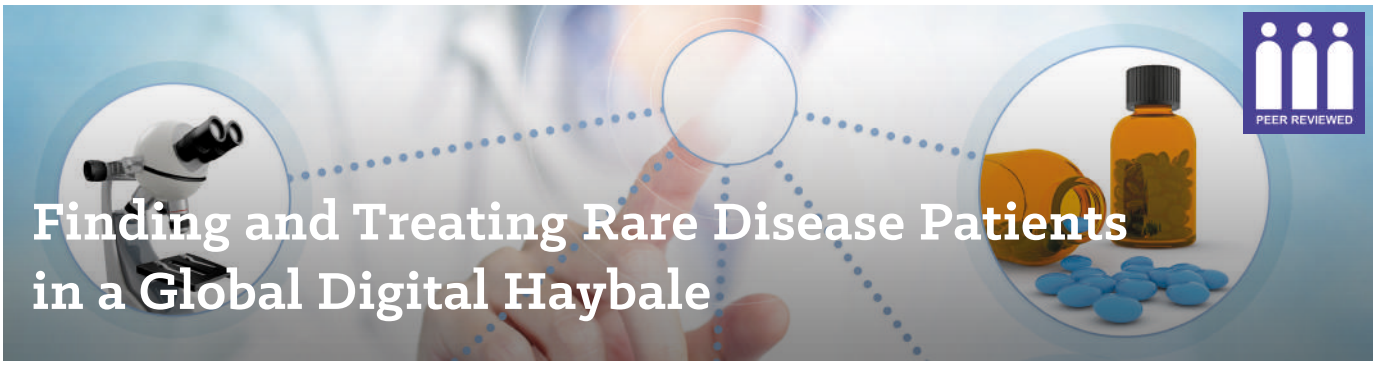

Finding and Treating Rare Disease Patients in a Global Digital Haybale

PEER REVIEWED

The amount of digital data we have generated, our digital footprint on this planet, is estimated to reach 44 zettabytes sometime this year, in 2020. That is 40 times more bytes than stars in the observable universe.[1]

That equates to 2.5 quintillion bytes of data created each day at our current pace, 90% of this over the last two years, by and for over 3.7 billion humans on the internet, with even more data to come from increasingly connectable devices via the Internet of Things (IoT).[2]

The Digital Solution has also become the problem: To quote Walt Kelly, "we have met the enemy and he is us" – and there is no off-switch. Our lives are increasingly, conveniently, digitally tethered to devices and social media, for work, as well as for news and entertainment. Source: ©Okefenokee Glee & Perloo, Inc. Used by permission. Contact: permissions@pogocomics.com

Healthcare: The Big Shark Tank

Our healthcare ecosystem has also gone the same way. Today, healthcare is no longer a house call by a local doctor who knows our history by simple familiarity and through their own paper records. Increasingly, our healthcare is through a network of distributed physicians and medical services that each have a puzzle piece in regard to our health status. For the services to integrate the complete picture of our healthcare journey, each of us must allow our data to be shared within this network. Multiply this across people, places, languages, regional and country regulations and one can start to envision the structural mayhem currently afoot in care as well as reimbursement.

The issue: we are all onions with layers, and the deeper you go, the more likely someone is going cry.

- Not all data is the same; some data is better than others.
- Raw data types give values that feed to aggregate metrics or measurements in context with other data points; that means a data hierarchy in which summary level results are derived from aggregate values which are derived from multiple single measurements.
- Data interpretation should be validated or will otherwise remain subjective. The quality, at each step (capture, analytics, and summary results generation), is critical, as garbage in will always equal garbage out.
- Ensuring personal and private data protection is critical and legally imperative.
- Regulations and standards are important to allowing interoperability while ensuring data privacy and security relative to standards and increasing legislation.

Healthcare Data

Healthcare data comes in all shapes and sizes, just like the patients and patient population it is derived from. Data sources vary widely; for example, individual metrics such as heartbeat and pulse from an Apple watch or individual EKG or genetic sequence, from patient registries, to electronic medical records, to claims data, to eCRF for patient chart review. Data can be classified as structured information, such as patient name, diagnosis codes and medications; or unstructured data, such as emails, audio recordings, and doctors' handwritten notes.

Increasingly, the challenge in our lives is to filter the background noise to what is important. It is the same concept in digital data management.

Data is therefore various and diverse. What is the best data? How is the best data acquired?

The best data is the most appropriate data, fit for the purpose intended.

- In digital biomarker development, IoT technologies are enabling objective, quantifiable, physiology and behaviour metrics through portables, wearables, and implantable and digestible technology vehicles. As an example, AI has been used to predict heart failure hospitalisation up to ten days in advance, using data from wearables.[3]
- In virtual clinical trials, your smartphone, watch or glasses could link you remotely to a study, with remote sensors recording data such as body temperature and blood glucose levels automatically to the study's electronic data capture (EDC) records.
- In the US, passage of the Patient Protection and Affordable Care Act (2012) also mandated adoption of electronic medical records over traditional paper files and reports by 2014, leading to electronic health records (EHRs) now being digital and better accessible to patients and their caregivers alike.
- Real-world data (RWD) is data on observed patient outcomes, derived from sources such as electronic health records, patient surveys, clinical trials, insurance claims, billing activities, and product and disease registries. Real-world evidence (RWE) is dependent on RWD and, as defined by the FDA, is "clinical evidence regarding the usage and potential benefits or risks of a medical product derived from analysis of RWD".

Making Sense of the Mess: Real-world Data to Real-world Evidence

The volume and diversity of digital data is exploding, and, in

healthcare, the number of electronic medical records is growing exponentially as technology makes this information increasingly available. Real-world data is increasingly accessible and useful for outcomes research and regulatory purposes. While clinical trial evidence remains the gold standard for evaluation of treatment efficacy, there is increasing interest and potential for converting RWD into real-world evidence that, through analysis and interpretation, can be used to inform healthcare decision-making.[4] RWD offers advantages over randomised controlled trials that are particularly useful for research and can be applied to healthcare decision-making. They include the availability of timely data at reasonable cost, large sample sizes that enable analysis of subpopulations and less common effects, and the representativeness of real-world practice and behaviours outside of a clinical trial setting. While RWE offers tremendous potential, it also presents very real risks, such as biases due to lack of randomisation, data quality, and the potential for spurious results due to data mining.

EHR Data: The Best Quality Clinically Validated Data

An electronic health record (EHR) is the systematised collection of patients' electronically stored health information in digital format. EHR systems are designed to store data accurately and to capture the statistics of a patient across time, thus relieving the need to share previous records with current and future caregivers. EHRs enable patient care to be more based on the entire healthcare network, instead of based on individual caregivers, thus allowing patients to be seen across their healthcare network and their conditions reviewed and treated by broader expertise, regardless of location.

EHRs contain patient demographic details, such as age and weight, as well as their medical history, including diagnoses, treatments, conditions, laboratory results, radiology images, and billing information. EHRs are often the best longitudinal record of a patient's journey, treatment, and diagnostic history. The upcoming 5G technology will offer the additional potential to better enable remote monitoring through wearables, telemedicine, and larger file transmission, such as medicinal images, which are often not part of medical record due to size. The challenge, with more data, is how to drink out of a firehose with increasing diameter and make utility of all the context now provided, without drowning in the data.

Data Resolution = fit to purpose: The value of any type of digital data, including medical data, comes from its resolution. The author is shown in a digital image, revealing at higher resolution the pixelation, or individual pixel bins, much like Seurat's pointillism made of different colour values, that make up the overall image in an electronic format. Source: author's own collection.

The exciting development which will bring us toward better healthcare is better resolution of data; digitally, specifically, but of the author, not so specifically (see picture above). Much like building a house, a foundation of digital clinical EHR data is the solid bedrock enabling better patient metrics and better outcomes. However, issues with EHR data are often skipped over and it is important to draw out the challenges in managing EHR data for patient insights:

Firstly, even though there are many global, regional and national initiatives to harmonise data, there also still exist a multitude of coding systems that need to be managed in parallel. There is, therefore, a need for thorough mapping and translation of codes between systems so that a common search strategy can be conducted across systems.

Secondly, coding may be harmonised, but the way that the EHR data is managed for the same conditions can vary greatly because of the richness of the coding systems and the different way diseases are understood and managed in different healthcare systems.

Thirdly, the way in which systems are implemented varies greatly, so the data can be managed in different systems and harmonised in different ways, again shifting the way in which the data is managed.

Lastly, for rare diseases, the coding is often up to 70% inaccurate, due to physicians being less familiar with those conditions and their coding, so a different strategy to traditional search methods needs to be considered.

Case Study: Rare Diseases – Enabling Better Identification and Diagnosis by Combining EHR Data Analysis and AI

Rare diseases, by nature of being rare, are often untreated, undiagnosed, or frequently misdiagnosed. In addition, rare diseases may present differentially – meaning patients don't all appear the same but are heterogenous and therefore hard to diagnose – which leads to a substantial delay in diagnosis, and makes it very difficult for patients, their families and healthcare givers to manage.

Studies show that the impact of rare disease is much wider than the individual affected and represents a significant challenge for the healthcare system itself.

In a survey of patients and caregivers in the USA and the UK, patients reported that it took on average 7.6 years in the USA and 5.6 years in the UK to get a proper diagnosis, during which time patients typically visited eight physicians (four primary care and four specialist) and received two to three misdiagnoses.[5] Of the 7000 known rare diseases, 90% do not have an FDA-approved medication, which means patients must go with no treatment or go with off-label use of existing medicines to treat their symptoms.[6] Patients with rare diseases can go up to 20 to 30 years before diagnosis, or even go entirely undiagnosed.[7,8,9]

Our company operates a platform on which a network of partner hospitals around the world make their patient EHR data query-able, with appropriate data privacy protections. When aiming to support the diagnosis of rare disease patients, this data is not enough to counter the data issues mentioned above. We therefore sought out technology which would identify phenotype, condition, and treatment models better.

The Swiss company Volv Global is applying cutting-edge AI and machine learning technology to highlight possible rare disease patients. Their unique methodology not only ascertains patient cohorts at risk of disease, but also helps with trial recruitment, understanding patient journeys and can make assessment of real

market size for generating rationale to create new drugs to meet these important unmet needs.

Using this technology, we can address the challenges of developing computational models capable of detecting rare disease patients in population-scale databases such as electronic health records (EHRs) while addressing all the real-world challenges highlighted in the previous section. The issues with EHRs are non-trivial as the EHR is in fact a weak proxy for a patient phenotype when we are considering rare diseases.

Typically, one would need to look across around 10,000 features in the full EHR to find relevant patients reliably. With typical machine learning methods and standard toolsets, one would 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 the reader will note, this is an impossible threshold, as, by definition, the rare disease patients are very rare. Additionally, as we have seen, they are also often misdiagnosed, so "hidden" within the system.

Volv has overcome these obstacles with a novel, lightly-supervised algorithm that leverages unlabelled and/or unreliably-labelled patient data – which is typically plentiful – to facilitate model induction. Importantly, it can be proven that the algorithm is safe: adding unlabelled/unreliably-labelled data to the learning procedure produces models which are usually more accurate, and guaranteed never to be less accurate, than models learned from reliably-labelled data alone.

The methodology is not based on machine learning toolsets, but novel algorithm development with a validation and proof methodology built in.

This is a breakthrough for patients that are likely to be held up in a lengthy diagnostic odyssey, as the models can be adapted to the healthcare systems now on our own platform through our partnerships with hospitals around the world. Not only this, but Volv's remote learning and deployment capabilities mean that the combination allows us to build highly accurate and adaptive complex models reaching more and more patients as the platform expands.

Volv's methodology often has no examples of confirmed patients with disease to learn from (i.e. no "labels") and it is therefore useful to get a first 'gold-standard' input and validation of the model performance, which is done in a specialised part of the review methodology.

Using their novel techniques with extremely small sample sizes, that are typical to rare diseases and personalised medicines, Volv builds predictive diagnostic algorithms that outperform Human Clinical Diagnostic Performance by looking at data earlier in the patient journey and by identifying cognitive biomarkers, digital biomarkers and medical biomarkers that drive a completely new way to diagnose.

- **cognitive** biomarkers are where we are picking signals by way of things that are thought about the disease from the way it is handled or managed or classified within the clinical system
- **digital** biomarkers are where the actual data is 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 and subsequently derive clinically interpretable models.

It is desirable to construct prediction models which are both accurate and interpretable, as in 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 machine learning algorithms (e.g. deep learning, ensembles) is to identify which predictors/features are 'important' to the model's predictions and to quantify this importance. Alternatively, one could trade off model performance and interpretability, adopting a less accurate but easy-to-understand model structure (e.g. linear regression, decision tree). Unfortunately, neither of these options is very useful in medical domains:

- standard feature importance assessment methods are not appropriate for many medical informatics problems, such as modelling and analysing electronic health records (EHRs);
- implementing sub-optimal prediction models in high-consequence medical settings is hard to justify.

In Volv's process, they "learn" an interpretable model from a proprietary good/robust model and then assess the predictive importance of the features of this new, interpretable model. This delivers a model that can be utilised by a clinician, as it is developed in their language and terminology, and it has quantifiable predictive performance, derived by specialised analysis of their complex models. Importantly, we as humans can learn new things from these models that are novel.

One of the truly interesting things about the interpretable models that Volv produces is that they can sometimes be more than one, which may in fact mirror a clinical setting within which patients can find themselves. This is important as it means that the interpretable models are clinically relevant.

In summary, the technology collaboration with Volv allows us to **flag potential rare disease patients correctly and then work with their treating physicians to create the outreach programmes, test the patients for rare disease, and reach a correct diagnosis**. Volv's methods are shown to substantially outperform state-of-the-art models in patient-finding; Clinerion's patient data network allows healthcare systems to leverage their data in a secure and compliant manner to apply these models to the benefit of patients. Together, the two companies are enabling better healthcare for severe but undiagnosed conditions, one patient at a time.

REFERENCES

1. Desjardins J, How much data is generated each day? World Economic Forum, April 17, 2019, available at: www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bddf29f
2. Marr B, How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read, Forbes, May 21, 2018, available at: www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#1879786d60ba
3. Perkins J, PhysIQ and U.S. Veteran's Affairs Publish Breakthrough Study Predicting Heart Failure Hospitalization up to 10 days in Advance using AI, PhysIQ, February 25, 2020, available at: www.physiq.com/press-media/physiq-and-u-s-veterans-affairs-publish-breakthrough-study/
4. Real-World Evidence, ISPOR, available at: www.ispor.org/strategic-

initiatives/real-world-evidence
5. 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
6. 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)
7. 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
8. 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
9. 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

Douglas Drake

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

Email: douglas.drake@clinerion.com