Why vLife™?

The life science industry has been a late adopter of most innovative technologies, despite its large R&D budgets. One reason is that there are various types of players in the life science industry, from pharmaceutical manufacturers to providers, payers, research organizations, and other related businesses. Each of them generates terabytes of data, which is not made available for innovation and experimentation, partly because of strict regulatory and compliance standards and partly because of the extremely competitive environment in which these companies operate. The data ends up in silos in multiple locations in the health ecosystem. With vLife™, we are trying to address this issue.

What is vLife™?

vLife™ is a data-driven platform to engage, innovate, and operationalize solutions faster. The foundational layer of vLife™ is the life science data lake, which consists of datasets from electronic medical records (EMR)/electronic health records (EHR), claims, clinical trials, government agencies, human genomes, images, and medical devices made available as APIs to accelerate the pace of innovation efforts. Along with data, vLife ™ also contains ready-to-use IoT recipes, which help in rapid prototyping of IoT solutions.

What does vLife™ offer?

vLife™ has five offerings:

- Data as a Service
- Analytics as a Service
- AI as a Service
- Innovation as a Service
- IoT Recipes

What is Data as a Service?

The fundamental layer of vLife™ is the data lake. The data lake consists of data APIs for various datasets, such as EMR/EHR, claims, clinical trials, government agencies, human genomes, images, and medical devices. Data as a Service exposes all the data in the data lake in the form of APIs. You can register for vLife at ***http://vlife.virtusa.com/registration/*** and start accessing APIs.

What is Analytics as a Service?

Analytics as a Service provides pre-bundled analytics packages. It enables individuals to take a peek into the data that vLife™ offers. It lets you explore the datasets and extract meaningful insights.

What is AI as a Service?

AI as a Service provides pre-trained AI solutions. It empowers users to deploy their own machine learning (ML) or deep learning models on vLife™ datasets or on their own datasets.

What is Innovation as a Service?

Innovation as a Service helps companies to discover and experiment or prototype innovative solutions to complex business problems by helping them navigate through ambiguity. It helps you discover (identify and validate emerging technologies that are relevant and disruptive), define (help with a solution through rapid prototyping) and develop or deploy (prototypes into production and measuring value delivered). With vLife™, you just need to reach out to us with a business problem and we will take care of all the phases from identifying to productizing using our proven methodology.

What are IoT Recipes?

IoT recipes are reusable modules that let users accelerate their IoT solution development. IoT recipes help you in rapid prototyping of IoT solutions.

What is vLife's Open Innovation Platform (OIP)?

vLife's OIP provides a platform to accelerate AI software development by providing

a) A ready-to-use AI platform
b) Synthetic data or publicly available data in the form of APIs
c) Options to visualize and gather insights from the data
d) Pre-trained AI/ML models
e) IoT recipes for integrating devices into solutions

What type of data is currently available in vLife™?

vLife™ currently contains

- Publicly available data (from R&D organizations, government agencies, and academic institutions)
- Synthetic data (algorithmically manufactured patient data generated using tools like Synthea™)
- Acquired data (data brought from various aggregators like IBM MarketScan)

Are you authorized to use publicly available data?

Yes. Publicly available data is de-identified data used for academic and research purposes. The source of the data is government agencies, established research institutes, and academia.

What is de-identified patient data?

De-identified patient data is health information from a medical record that has been stripped of all "direct identifiers"—that is, all information that can be used to identify the patient from whose medical record the health information was derived. The Health Insurance Portability and Accountability Act (HIPAA) names 18 direct identifiers that are typically present in patient medical records. According to HIPAA, there are three acceptable ways to de-identify patient data. The first is the safe harbor option, in which all 18 identifiers are removed. The second is the statistical option, in which a retained statistician determines which of the 18 identifiers can be maintained without creating greater than a "very small" risk that the data could be re-identified. The third is the limited data set technique, in which the organization removes 16 identifiers and protects what remains with special security precautions.

Why is de-identified data not always the best option?

De-identified data, if available, is one option to address data sharing–related challenges, but the process of de-identification can be very complex and expensive. In addition, its use is often restricted, and in some cases, the quality of the data is marginal at best. Also, it is relatively easy to re-identify the data and link it to a real person, and if this happens, you could be in violation of the Health Insurance Portability and Accountability Act (HIPAA) and other laws.

What is synthetic data?

Health care lags other industries in information technology, data exchange, and interoperability. To close these gaps, developers require access to large repositories of high-quality health datasets for a range of secondary uses that have no clinical or medical implications, including software development, testing, and clinical training. However, access to real electronic health record (EHR) data is hindered by legal, privacy, security, and intellectual property restrictions. Where real datasets are unavailable, generating synthetic data is an alternative and better option. One can use any patient generator tool that models the medical history of synthetic patients to generate synthetic data. The output is high-quality synthetic—realistic but not real—patient data and associated health records covering every aspect of healthcare. The resulting data is free from cost, privacy, and security restrictions. It can be used without restriction for a variety of secondary uses in academia, research, industry, and government initiatives.

Why is synthetic data a better option when you can purchase data or de-identify patient data?

Access to real electronic health record (EHR) data is hindered by legal, privacy, security, and intellectual property restrictions. Since healthcare data exists in silos and is costly, sourcing and maintaining complete, high-fidelity patient data is slow, expensive, and error-prone. Cleansing data from multiple sources is time-intensive, and personal health information (PHI) compliance limits access and usage. Also, de-identifying patient data is complex and costly. Considering the above, synthetic data is a better option.

What is the difference between the terms coded, de-identified, and anonymous?

Coded: Direct personal identifiers have been removed (e.g., from data or specimens) and replaced with words, letters, figures, symbols, or a combination of these (not derived from or related to the personal information) for purposes of protecting the identity of the source(s), but the original identifiers are retained in such a way that they can be traced back to the source(s) by someone with the code. Note that a code is sometimes also referred to as a key, link, or map.

De-identified: All direct personal identifiers are permanently removed (e.g., from data or specimens), no code or key exists to link the information or materials to their original source(s), and the remaining information cannot reasonably be used by anyone to identify the source(s).

Anonymous data: Unidentified (i.e., personally identifiable information was not collected, or if collected, identifiers were not retained and cannot be retrieved) information or materials (e.g., data or specimens) that cannot be linked directly or indirectly by anyone to their source(s).

How can one generate synthetic patient data? What are we using in vLife™ to generate synthetic data?

A variety of synthetic data generation (SDG) methods have been developed across a wide range of healthcare domains, like Synthetic Electronic Medical Records Generator (EMERGE) as a methodology for creating EHRs, medical Generative Adversarial Network (medGAN) to generate realistic synthetic EHRs and Synthea™, a synthetic patient generator that models the medical history of synthetic patients. In vLife, we use Synthea™ to generate synthetic patient data.

How do you make sure the synthetic data is close to real data for applying ML models?

Virtusa, in collaboration with Fuse by CardinalHealth, uses a patient journey module generator called Proxi™ that learns or derives the patient journey from existing EMR/EHRs. This module is then used in a synthetic patient generator like Synthea™ to generate realistic but not real synthetic patient data. In addition to this, Virtusa, in collaboration with Fuse by CardinalHealth, has come up with a comparison tool/framework that compares the real data and synthetic data.

How do I access the data in vLife™?

Register for vLife™ at this registration link: http://vlife.virtusa.com/registration/

Can I download or move data to my own managed cloud?

No. Individuals can only access the data through the APIs. If you are a business, please contact vlifesupport@virtusa.com.

What if I have my own data? How do I leverage vLife™?

Use vLife's Open Innovation Platform to build and test out the solution/model and then expose it to your data.

Can I see some use cases of AI/ML models using vLife™ data?

Yes. Please refer to this source code:

- Github: https://github.com/Virtusa-vLife/Use-Cases
- Kaggle: https://www.kaggle.com/virtusavlife/kernels

Can I see some use cases of AI/ML models using vLife™ data?

CardinalHealth™: Virtusa collaborated with CardinalHealth™ to generate synthetic data using the Synthea™ synthetic patient generator for rheumatoid arthritis (RA) and simulated the US population for RA. The team also came up with a module generator, Proxi™, and a comparison tool to compare synthetic data with real data.

IBRI: The alliance of Virtusa and CardinalHealth™ collaborated with Indiana Biosciences Research Institute (IBRI). As a pilot, the team performed the task of scaling a small sample of EMR/EHR records provided by IBRI for type 2 diabetes and creating a large dataset using the module generator Proxi™ and synthetic patient generator Synthea™ to simulate the population of Indiana. The synthetic data was compared with real data using the comparison tool. An ML use case on type 2 diabetes is being built with the synthetic data now.

For further questions, write to **vlifesupport@virtusa.com**

www.virtusa.com

virtusa®
Accelerating Business Outcomes