International Journal of Computer Trends and Technology
ISSN: 2231–2803 / https://doi.org/10.14445/22312803/IJCTT-V71I9P106
Volume 71 Issue 9, 41-58, September 2023

Original Article

Real World Data-Driven Transformation in Healthcare & Life Science: Evidence-Based Analytics, Machine Learning and AI Applications

Mayur Katariya[1], Snehal Tiwari[2]

[1]*Senior Delivery Solution Architect, Databricks, CA, USA.*
[2]*University of Southern California, CA, USA.*

Received: 18 July 2023 Revised: 30 August 2023 Accepted: 13 September 2023 Published: 30 September 2023

Abstract *- High demand for real-time and effective patient outcome-centered healthcare systems is increasing globally. Therefore, there is a pressing need to counter inefficiencies and advance care delivery to promote positive health outcomes. Real-World Data (RWD) and evidence (RWE) have a tremendous opportunity and potential to improve patient outcomes. Understanding how patients use prescribed medication accurately guides stakeholders across the healthcare and life science system in making lifesaving, real-time choices regarding patients' health. RWD identifies inefficiencies across the healthcare environment and fills gaps in information silos among the stakeholders throughout the healthcare & life sciences ecosystem. Also, RWD is being used by pharmaceutical and life sciences firms at all phases of the drug development lifecycle, from initial discovery to post-market. RWE can bring crucial empirical data to clinical investigations that a standard study cannot.*

This paper extensively researches the Real-World Synthea dataset, delves into the intricate process of generating patient data across various healthcare interaction points, offering a comprehensive view from the patient's perspective and showcases the different use cases that can be derived across the healthcare and life science systems such as patient demographics, treatment rates, medication adherence, comorbidity analysis, patient risk prediction, and disease progression tracking to improve patient outcomes. These use cases illustrate how RWD, when integrated with advanced analytics and artificial intelligence (AI), can drive informed decision-making, personalized patient care, and drug research and development advancements. It also incorporates actionable solutions using Medallion Architecture from Databricks Lakehouse. The integration of additional data sources, such as demographic data, genomics, claims data, and social determinants of health, is presented to enhance insights and improve patient outcomes. It concludes by emphasizing the profound impact of real-world data and the application of data analytics, machine learning, and artificial intelligence (AI) on reshaping healthcare systems, enhancing research endeavors, and ultimately paving the way for a future characterized by more efficient, patient-centric, and data-driven healthcare to improve patient outcomes and drive faster innovation across the drug lifecycle, leading to more promising prospects for patients and stakeholders alike.

Keywords *- Artificial Intelligence, Data Analytics, Real World Evidence, Real World Data, Databricks Lakehouse, Healthcare, Synthea Data, Cloud.*

1. Introduction

Real World Evidence defines the observational data collected from Raw-World Data such as electronic health records (EHR), disease and product registries, claims, patient-generated data, IoT sensors, wearables, medical imaging, and any piece of data acquired from other sources that can illuminate the effects of drugs used by actual patients, in real life such as IoT, research blogs and social media platforms. Understanding how patients use drugs accurately guides stakeholders across the healthcare system in making lifesaving real-time choices. Real-World data identifies inefficiencies across the healthcare environment and fills gaps in information silos among the stakeholders throughout the healthcare system.[1] Some stakeholders across the health systems are patients, payors, manufacturers, providers, and government entities. RWD can also provide essential insights to support more valuable care and propel better health outcomes among patients. The Food and Drug Administration (FDA). [2] states three primary uses of RWE. Firstly, it monitors drug post-market safety and the adverse effects before making regulatory decisions. Secondly, healthcare providers use RWE to support care decisions and create guidelines, strategies, and tools to assist clinical practice. Moreover, life sciences organizations use the information to aid clinical trial designs and observational studies to propel advanced drug discovery and patient care approaches. Additionally, payors benefit from RWE to assess the outcomes of treatment. Research has recently discovered that combining RWD and artificial intelligence can significantly help healthcare. This duo can assist in the optimization of clinical trial design and improve patient recruitment. AI can transform RWD into actionable information and propel better decision-making concerning medicinal products.

Some use cases across the drug lifecycle at the high level encompass.

1.1. Pre-clinic Research and Drug Discovery

RWE can help aid pre-clinical research in identifying the prevalence of the disease, the disease history, and possible treatment patterns that can lead to identifying unmet needs and the potential development of a new drug.

1.2. Clinical Trials Design and Patient Recruitment

RWE can be used to design clinical trials, optimize patients' selection criteria, and decide potential trial site areas after analyzing real-world patients' characteristics.

1.3. Market Authorization and Access

RWE can propel regulatory approval and help define the most efficient pricing model technique.

1.4. Comparative Effectiveness Research

RWE makes it easy to compare medical intervention strategies in real-world settings. It can help measure drug efficacy and safety across diverse patient populations and help in effective clinical decision-making and treatment guidelines.

1.5. Patient Safety Monitoring

RWE plays a crucial role in post-marketing surveillance of drugs and medical devices. It helps detect and evaluate adverse events, monitor long-term safety profiles, and identify potential drug interactions or risks not captured during clinical trials. This can also aid pharma companies in quickly identifying the risks/benefits of their medicines.

1.6. Patient Safety and Value Impact

Efficient and scientifically rigorous comparative safety/effectiveness analysis.

1.7. Health Economics and Outcome Research

RWE enables easy evaluation of the economic impacts, cost-effectiveness and quality of life outcomes associated with different health treatments and interventions, which is very valuable for payers, policymakers and providers when allocating resources and coverage.

1.8. Regulatory Decision Making

RWE is used by regulators to supplement traditional clinical trial data. It supports labeling decisions, promotes post-approval commitments and confirmatory trials, and aids in evaluating the actual benefits and dangers of medicinal products.

1.9. Real-Time Monitoring and Patient Engagement

RWE can be leveraged for real-time monitoring of patients using wearable devices, mobile apps, or telemedicine. It enables remote patient monitoring, early detection of adverse events, and personalized interventions, improving patient engagement and outcomes. Overall, Real-World Evidence plays a vital role in generating insights into the real-world performance of treatments, optimizing healthcare and life sciences decision-making, and improving patient outcomes. It complements traditional clinical trial data and offers a more comprehensive understanding of how interventions work in diverse patient populations and real-world settings.

This article showcases how we can perform data analytics on RWD datasets like Synthea datasets - synthetic, realistic (but not real), patient data, and associated health records - using a data platform like Databricks Lakehouse for multiple use cases. Databricks Lakehouse is a unified analytics platform that combines the best of both worlds - data warehouse and data lakes. The lakehouse can help combine an organization's RWD into one large and collaborative platform that supports most analytics and AI capabilities. By bringing health-related data in one place, organizations can effectively perform analytics to help inform data-driven decisions to comprehend the elements influencing positive and adverse health outcomes.

2. Data Description and Process Flow

2.1. Describe Data: What is Synthetic Data?

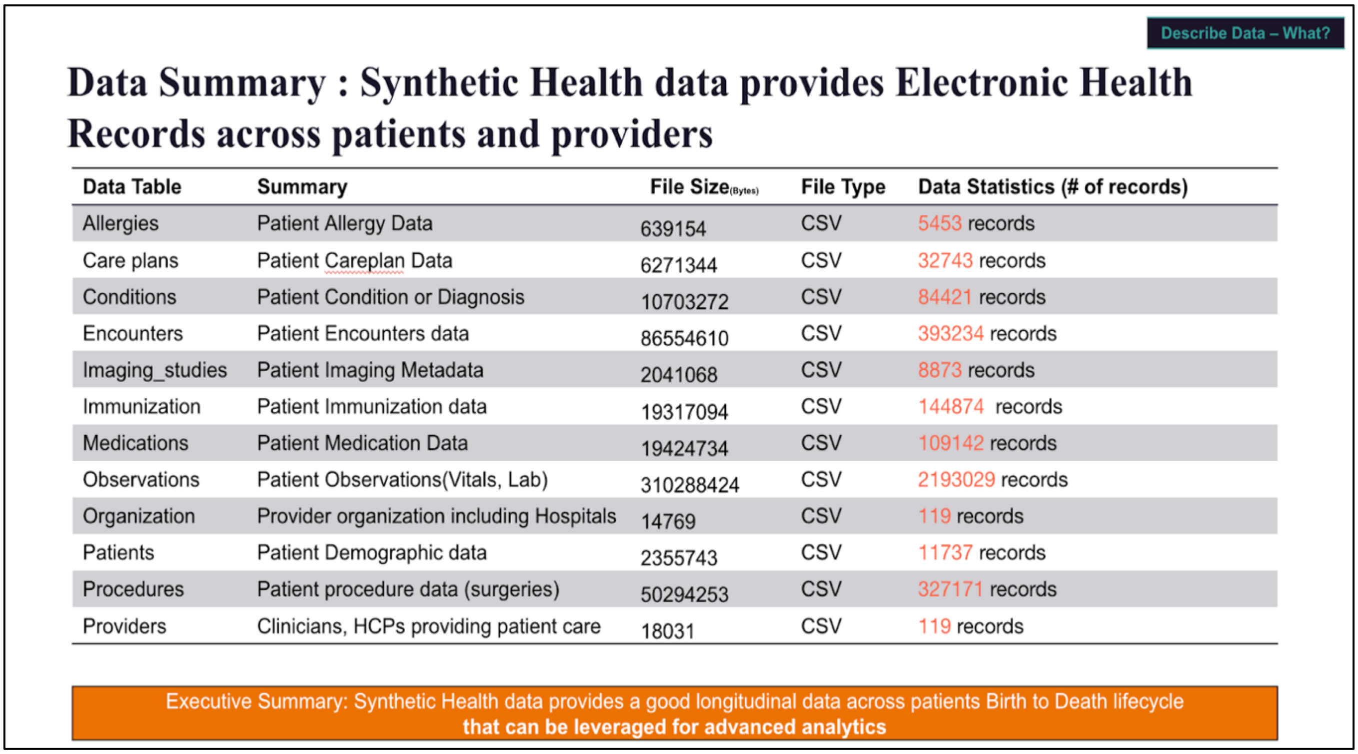
Gonzales et al. [3] define synthetic data as statistically generated microdata created by manipulating original data. Synthetic data upholds data effectiveness while guaranteeing the confidentiality and privacy of information. Synthetic data in healthcare can be an EHR dataset consisting of patient-sensitive data swapped with fake information to prevent replication and patient de-identification.

Below are some potential uses for synthetic data:

2.1.1. Hypothesis, Methods, and Algorithm Testing

Synthetic data can reflect on the format and structure of RWD. Information obtained is crucial in exploring the variables, assessment of data set feasibility, and hypothesis testing.

Additionally, knowledge of algorithms is critical for developing machine learning and AI. Information from the synthetic data is valuable for the robustness of the algorithms and testing of different methods and hypotheses of healthcare practices.[3]



Describe Data – What?

Data Summary : Synthetic Health data provides Electronic Health Records across patients and providers

Data Table	Summary	File Size(Bytes)	File Type	Data Statistics (# of records)
Allergies	Patient Allergy Data	639154	CSV	5453 records
Care plans	Patient Careplan Data	6271344	CSV	32743 records
Conditions	Patient Condition or Diagnosis	10703272	CSV	84421 records
Encounters	Patient Encounters data	86554610	CSV	393234 records
Imaging_studies	Patient Imaging Metadata	2041068	CSV	8873 records
Immunization	Patient Immunization data	19317094	CSV	144874 records
Medications	Patient Medication Data	19424734	CSV	109142 records
Observations	Patient Observations(Vitals, Lab)	310288424	CSV	2193029 records
Organization	Provider organization including Hospitals	14769	CSV	119 records
Patients	Patient Demographic data	2355743	CSV	11737 records
Procedures	Patient procedure data (surgeries)	50294253	CSV	327171 records
Providers	Clinicians, HCPs providing patient care	18031	CSV	119 records

Executive Summary: Synthetic Health data provides a good longitudinal data across patients Birth to Death lifecycle **that can be leveraged for advanced analytics**

Fig. 1 Synthetic health data presented as EHR record

2.1.2. Public Health Preparedness

Synthea data can be used to simulate disease outbreaks and assess the spread of the diseases. The information can be utilized to evaluate public health's effectiveness and prepare and develop response evidence-based emergency strategies.[3]

2.1.3. Medical Education and Training

Patient cases might be created using synthetic data for medical education and training. In a controlled setting, medical students and healthcare workers can practice clinical decision-making, learn about uncommon diseases, and improve their diagnostic and therapeutic abilities. [3]

2.1.4. Health Technology Development

Synthea data can be used to create and test electronic tools like Electronic Health Records (EHRs), medical imaging algorithms, and predictive analysis models. Similarly, it can play a vital role in creating decision support systems. [4]

For this article, we have used a Synthea-generated dataset and showcased how it can be used for supporting different RWD Use cases[5]. Synthea datasets provide good longitudinal data across patients' Birth to Death lifecycles that can be leveraged for advanced analytics. The figure above represents synthetic health data presented as EHR records. The Synthea dataset contains different data tables, including allergies, encounters, diagnostics tests, observations, etc. It is important to note that synthetic data is not the actual data of the patients. [6] However, they are statistically modified data that may resemble the original patient's information. The above table provides more quantitative details about different tables' data size details like file type, size, and statistics. Moreover, synthetic data includes tables representing different aspects of simulated healthcare information.

In the below section, we have described in detail about the data included in each of the tables along with some sample data view of all the tables after loading it into databricks for analysis.

Patient Table: This table contains demographic information about simulated patients, such as their unique identifier, gender, date of birth, race and Ethnicity.

Encounter Table: The encounter table captures patient encounters with healthcare providers. It includes details such as the encounter ID, patient ID, encounter type (e.g., outpatient, inpatient), encounter reason, date and time of encounter, and associated provider information.

Condition Table: This table stores information about the medical conditions or diagnoses affecting each patient.

Observation Table: The table contains recorded observations or measurements related to a patient's health. Measurements and observations made during patient visits may be recorded on the observation table. This may comprise measurements such as weight, temperature, blood pressure, heart rate, and any pertinent background data regarding the observations.

Medication Table: This table tracks the medications prescribed or taken by patients. It includes the medication ID, patient ID, medication name, dosage, start and end dates of medication use, and any related instructions or notes. Also, information on medications given to patients would be kept in this table.

It might contain details on the prescribed drug (name, dosage, frequency), the doctor who prescribed it, the start and stop dates, any adverse reactions that have been documented, and perhaps interactions with other drugs.

Procedure Table: This table captures information about medical procedures performed on patients. Records of patient medical operations would be kept in the procedure table. The type of procedure, the date and time it was carried out, the medical staff that assisted, any pre- or post-procedure instructions, and any complications or notes that could be included.

Allergy Table: The allergy table records information about known allergies or adverse reactions that patients may have. It includes the allergy ID, patient ID, allergy type (e.g., medication, food), specific allergen, reaction description, and related details or severity.

Immunization Table: This table contains data on immunizations administered to patients. It includes details such as the immunization ID, patient ID, immunization name, date of administration, and any associated notes or additional information. Also, data from patient immunization records, including information on vaccinations received, dates of administration, type of immunization, dosage, and any pertinent remarks, are likely to be kept in this table.

Organizations table: This table contains information on healthcare organizations providing patient care. It contains information on medical centers, clinics, hospitals, and other healthcare organizations. Names, locations, contact details, and provider affiliations of the organizations can be found.

Providers table: This table includes details regarding medical specialists like physicians, nurses, therapists, and other healthcare providers. The providers' names, specialities, license details, and contact information can all be found in this data.

Care Plans Table: The data in this table pertains to the patient care plans that have been developed. Care plans list the prescribed medications, procedures, and tasks healthcare professionals perform for a patient's condition. It might contain information on the kind of treatment, prescription drugs, sessions with therapists, and suggested follow-up appointments.[7]

Imaging Table: This table holds information on diagnostic imaging tests like X-rays, MRIs, CT scans, and ultrasounds. It contains information concerning the patient, the imaging method, the date and time of the procedure, and any conclusions or diagnoses drawn from the images.

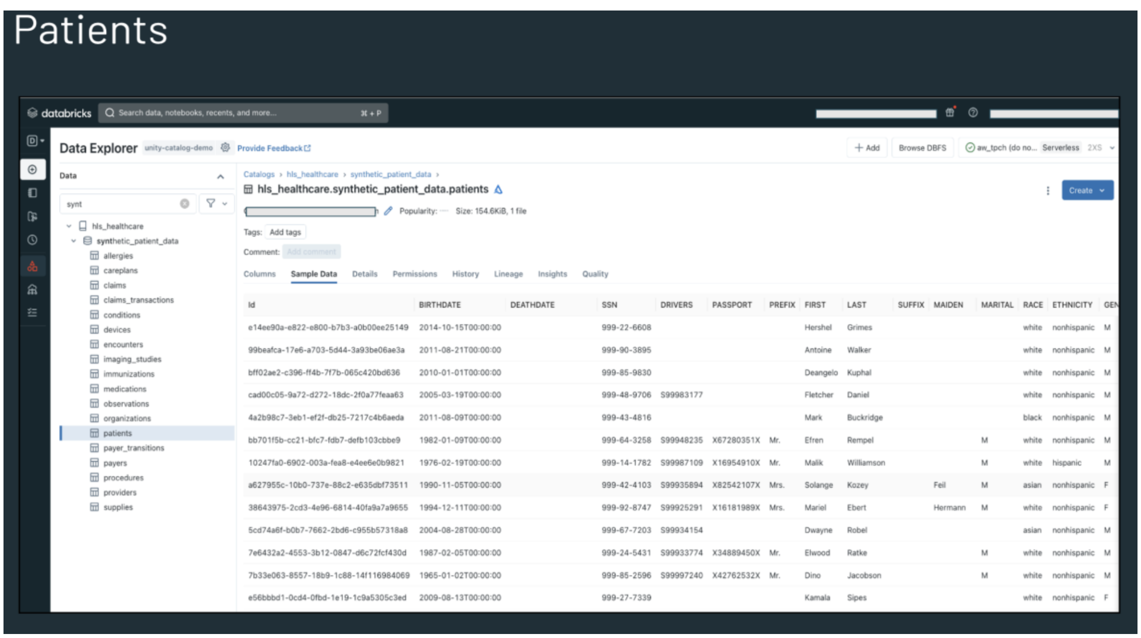


Fig. 2 Patient table sample view on databricks

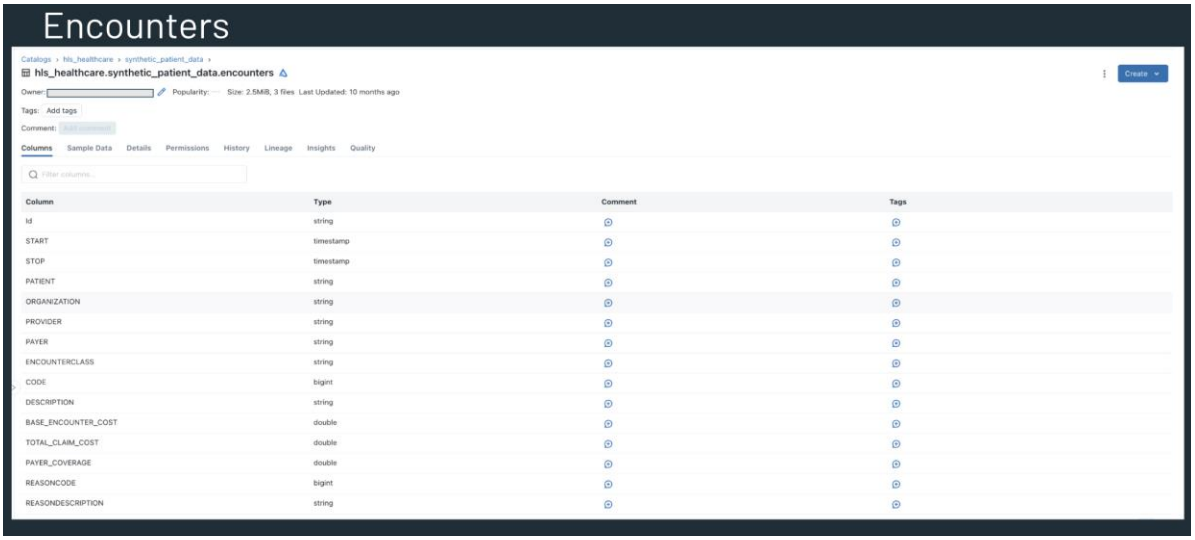


Fig. 3 Encounter table Sample Data view in Databricks

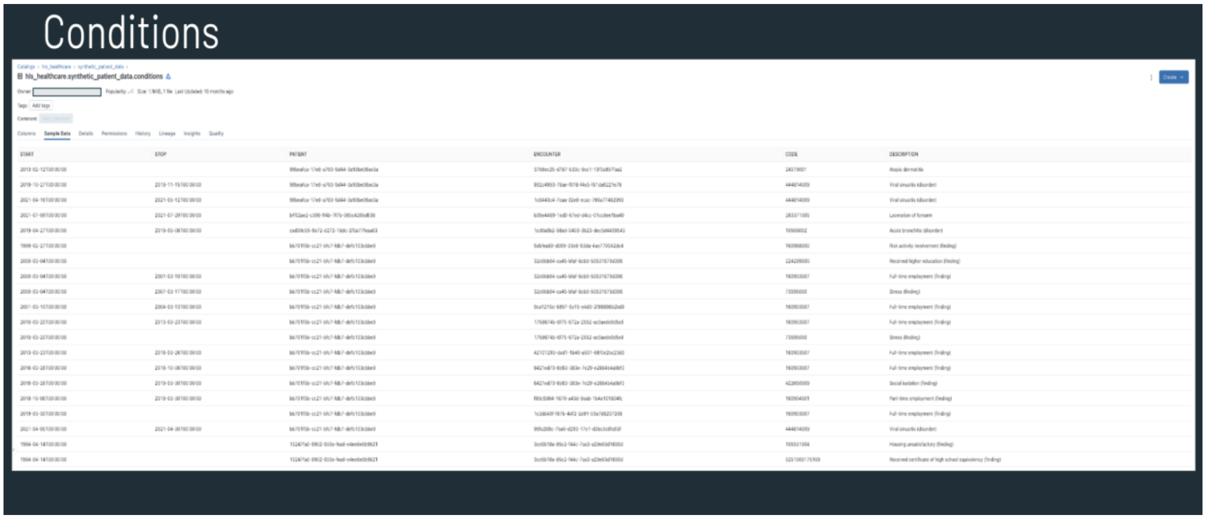


Fig. 4 Conditions table sample data view

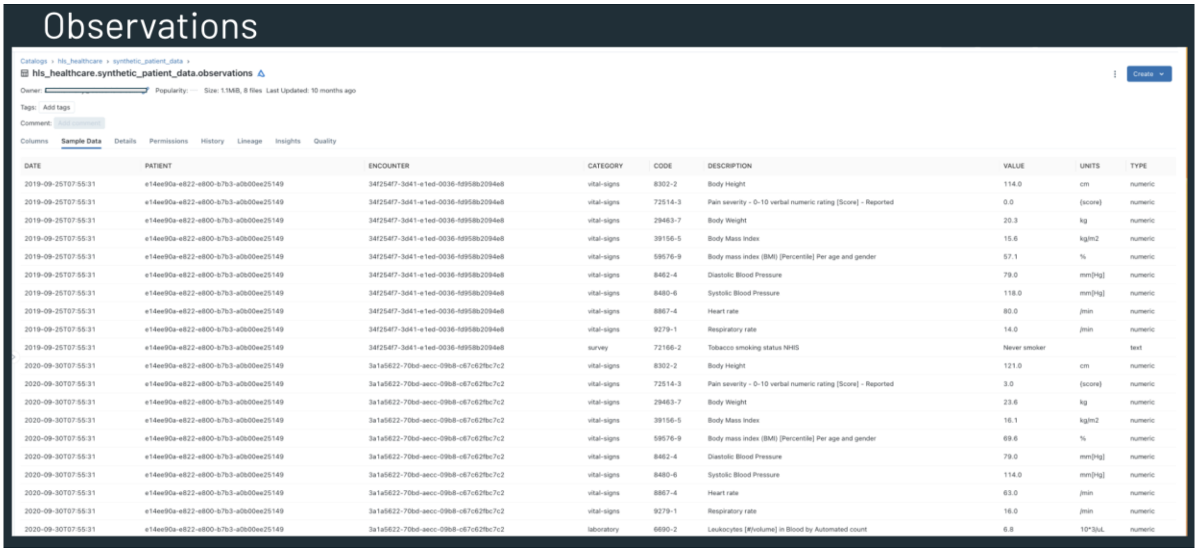


Fig. 5 Observation table sample data view

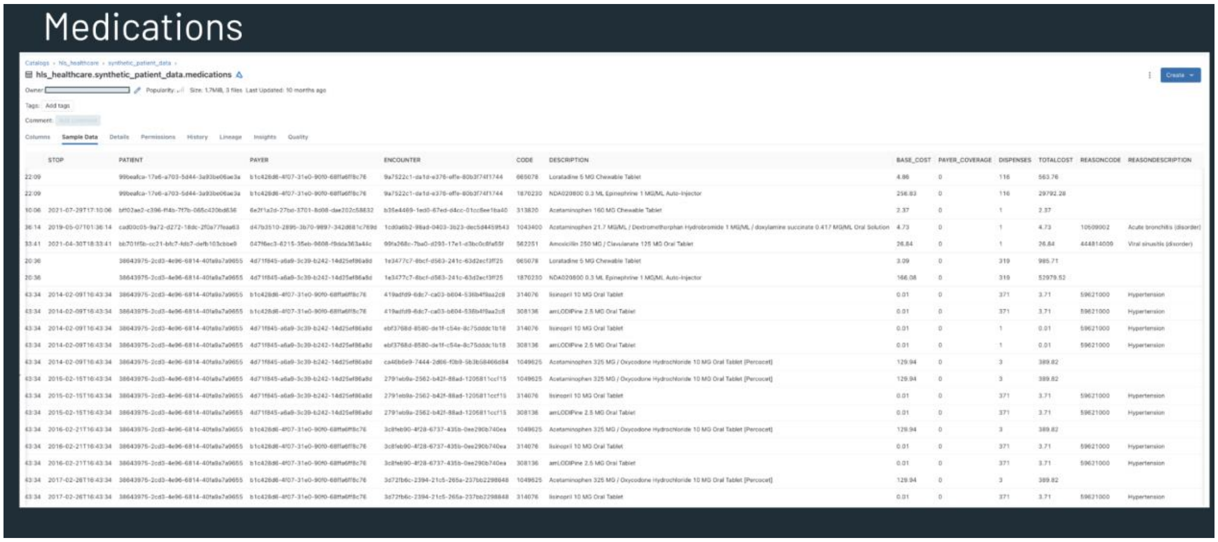


Fig. 6 Medication table sample data view

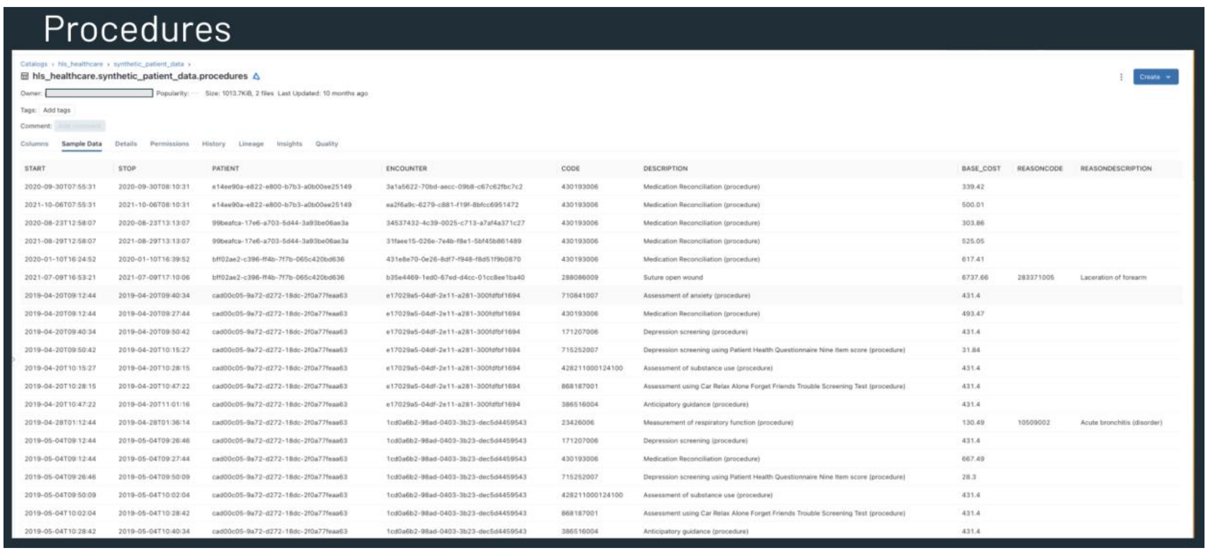


Fig. 7 Procedure table sample data view

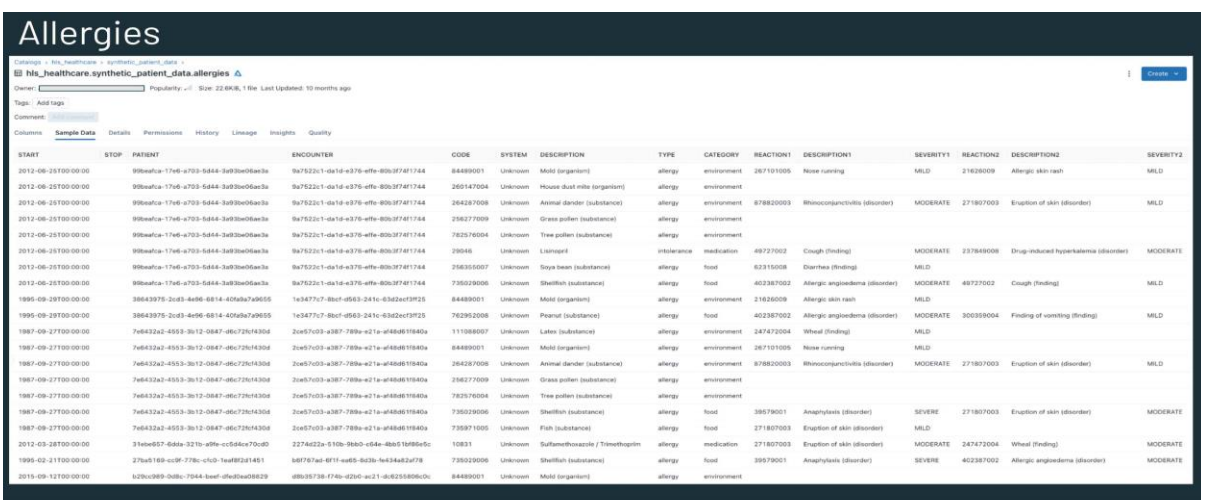


Fig. 8 Allergy table sample data view

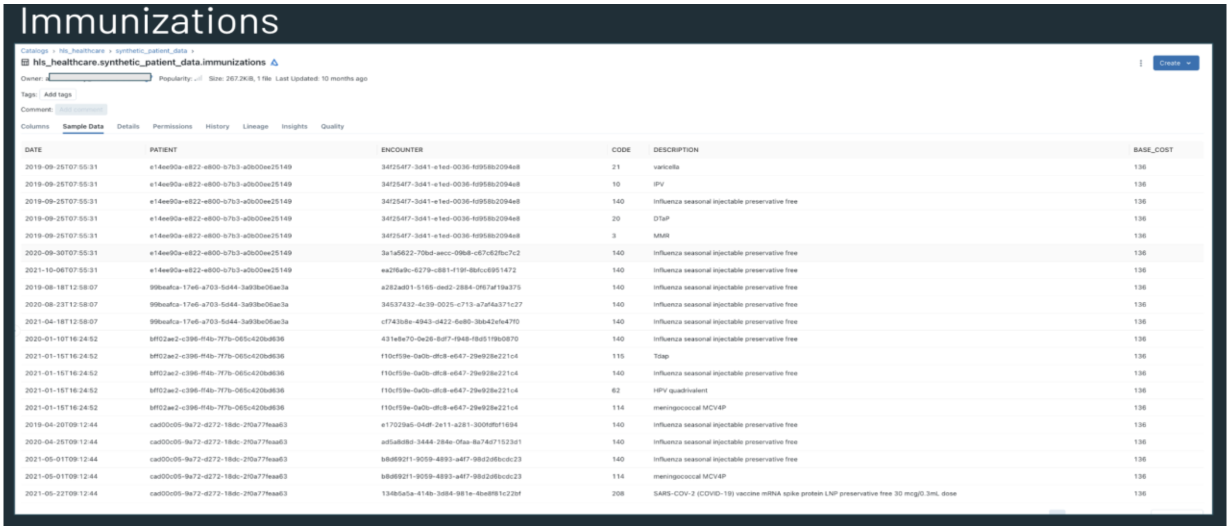


Fig. 9 Immunizations table sample data view

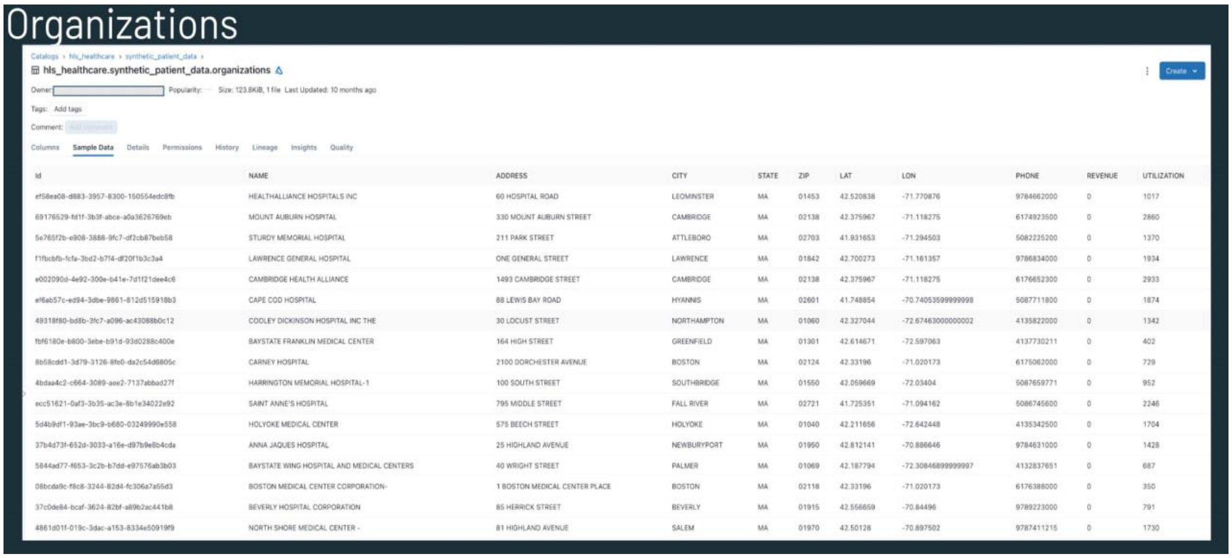


Fig. 10 Organizations table sample data view

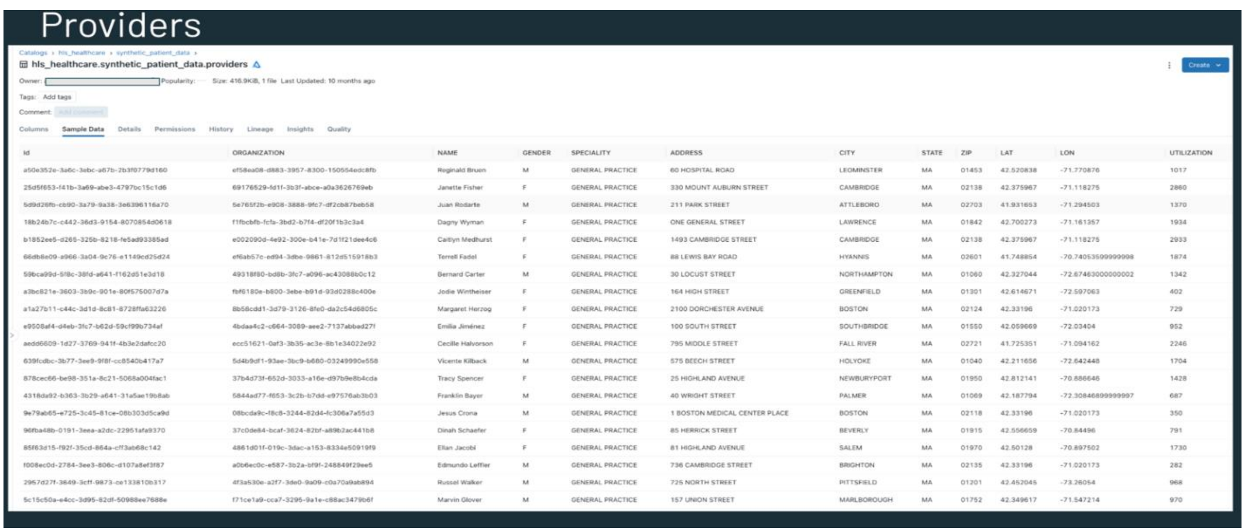


Fig. 11 Providers table sample data view

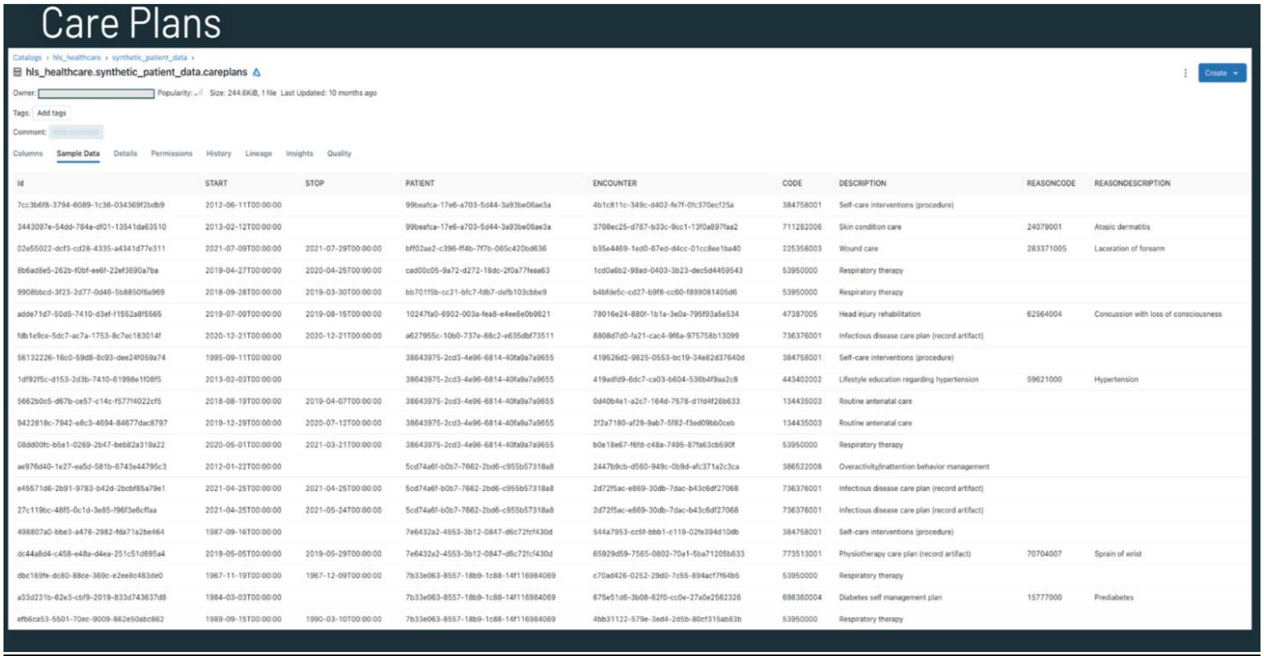


Fig. 12 Care plans table sample data view

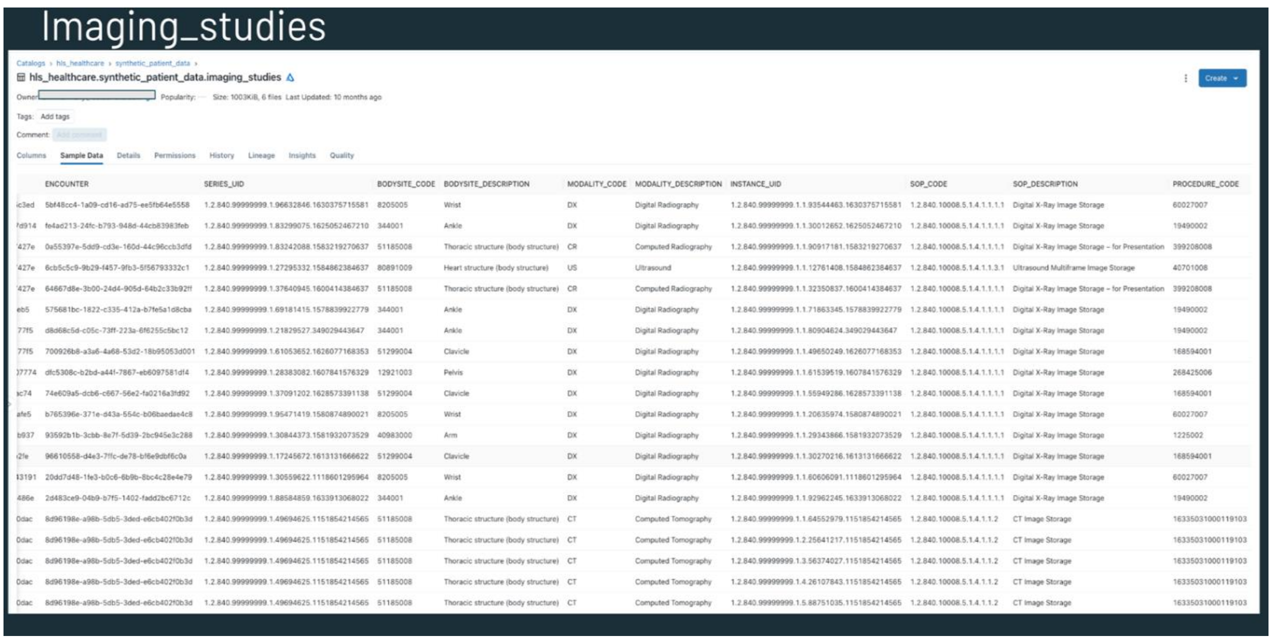


Fig. 13 Imaging_studies sample data view

3. Patient Data Generation Process Flow

3.1. Patient Process Flow

From a patient's perspective**,** the sample process flow of real-world data generation at the various stages when interacting with health professionals is illustrated below in Figure 14.

3.2. Symptoms Identification at Home

The primary reason patients visit a hospital or a pharmacy is sickness symptoms causing deterioration of health at their respective homes. Hence, the patient leaves home and goes to a doctor or pharmacy to seek medical help.

3.3. Patient Encounter and Registration

Patients interact with doctors when they visit a health facility or pharmacy. The patient's first step of the interaction usually involves surrendering important information such as their name, age, locality, gender, and, other times, contact details. The patients also explain the reason for their visit and some symptoms to offer further direction, all recorded in the patient table of the synthetic dataset.

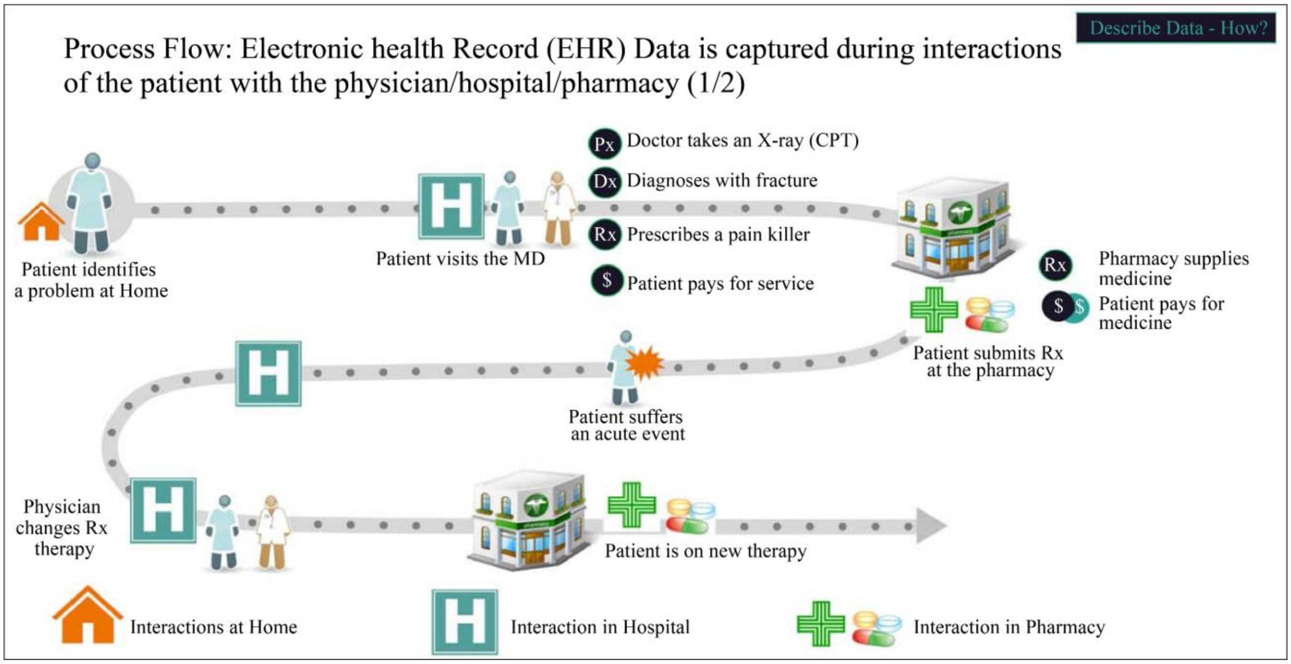


Fig. 14 Patient process flow

3.4. Medical History and Underlying Conditions

In this stage, the healthcare provider records the patient's previous medical history, underlying conditions, allergies, family medical histories, previous surgeries, and newly identified conditions. The condition table may record patients' existing or identified conditions.

3.5. Drug Prescription and Pharmacy Visit

After the health provider diagnoses the patient, a treatment plan is usually chosen, mostly by prescribing medicine for the patient. The prescription details are usually recorded, including the name of the medication, its dosage, instructions, and duration. The patient visits a pharmacy outside or within the healthcare center to access the medication. The pharmacist records specific details such as the medication details and date provided. This information concerning medication is recorded in the medication table.

3.6. Laboratory and Diagnostic Tests

Sometimes, laboratory tests are recommended as part of the patient's evaluation process. After the patient has undergone the tests, the test type, values, dates, and results are recorded in the synthetic dataset's observation table and procedures table.

3.7. Follow-Up and Ongoing Care

Usually, doctors recommend follow-up visits to monitor the patient's progress with the prescribed medication. Follow-up and ongoing care occur after the initial encounter with the healthcare provider. Follow-up visits capture and reveal the patient's progress and possible further intervention. If the doctor identifies deficiencies in the initially prescribed medication, the doctor can recommend a substitute and continue to monitor the patient's progress. These details are recorded in the encounter table of the synthetic dataset.

3.8. Inpatient Visit

Inpatient visits can occur before or after follow-up and ongoing care. After the laboratory and diagnostic tests, the doctor may decide that the patient's condition is severe and requires close monitoring. In such a case, the doctor can recommend admitting the patient. Other times, the patient's condition may deteriorate during the follow-up visits. In such a case, the healthcare provider can also recommend that the patient be admitted and be closely monitored. The patient data flow process initiates and propels the generation of RWD from the patient's point of view. It presents the essential interactions between the patients and the healthcare provider, starting from their first encounter, diagnoses, treatment, prescribed medication, and other important information. This set of RWD provides life sciences organizations and healthcare providers with RWE to analyze and derive insights into different aspects of patient care and the entire performance of the healthcare system.

4. Architecture Diagram and ETL Flow

From the architecture perspective, this EHR data can be loaded into data bricks; below, figure 15 is a sample architecture of the overall data flow.

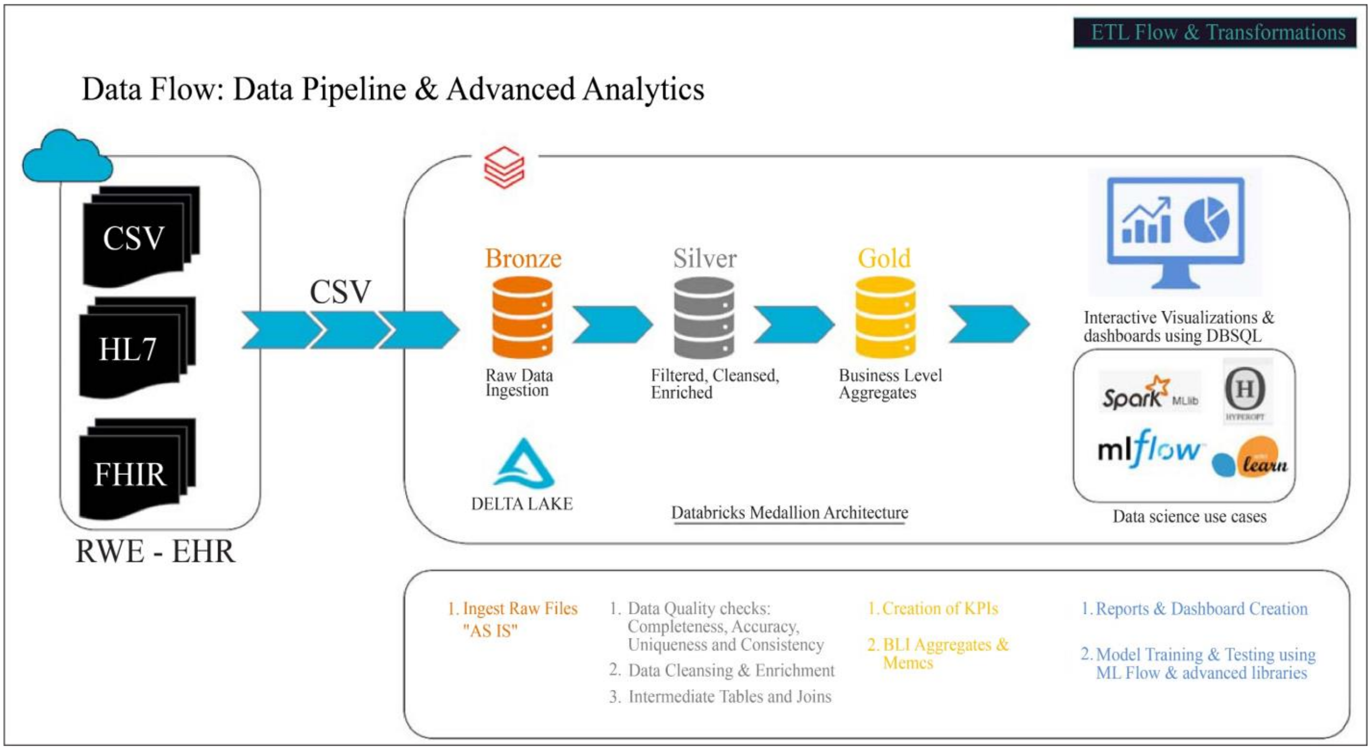


Fig. 15 Medallion architecture and data flow

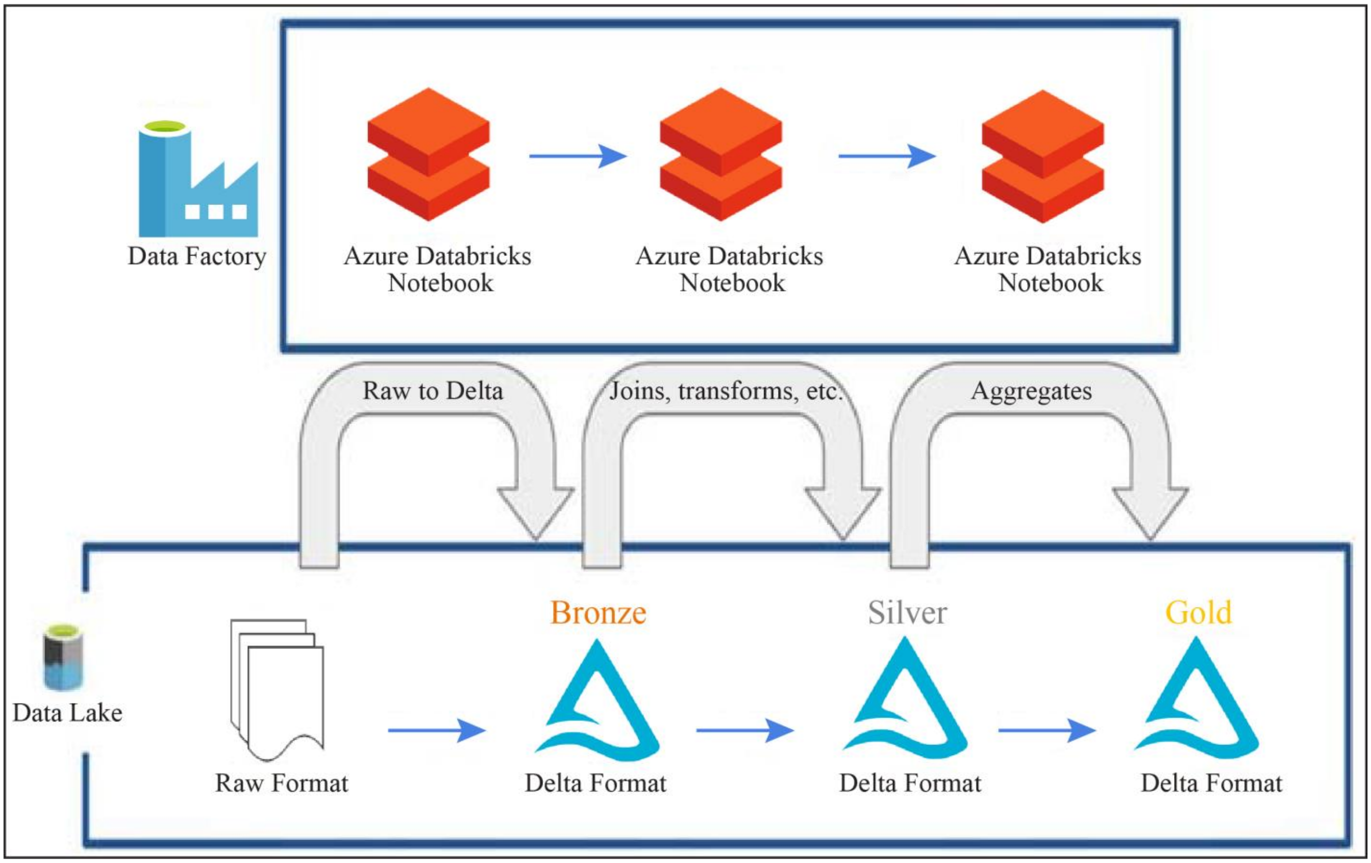


Fig. 16 Sample medallion architecture using azure databricks

The above figure indicates the high-level multi-hop architecture and data flow across various layers. The ETL process involves moving the raw data across different layers to ensure the data is valid and conforming to the system's needs. Databricks Lakehouse leverages Medallion Architecture, which helps organize data logically with structures (layers), namely bronze, silver and gold, that improve the data quality over time as it flows through each layer.

The bronze layer plays the most crucial part in the ETL process. It is a structural stage where the data is raw, where all the data is from the outside source land. In the Bronze layer, the data table is commonly structured as the external source (origin), with the table structure able to capture additional metadata columns that show the load date/time, process ID, etc. This layer provides a quick-change data capture with the ability to extract the historical archive sources of the data, known as cold storage. Moreover, it provides data lineage and an extension to data auditability, when necessary, without reading from the source system.[8] Silver layer, also known as cleansed and conformed data. At this layer, the data obtained from the first stage (Bronze) is matched, consolidated, conformed, and cleaned. The main aim of this layer is to ensure that the data captures the real information regarding the subjects of the data for storage, removing duplicates and cross-reference tables of data obtained from the bronze layer.

The Silver layer can provide an "Enterprise view" of all its key business entities, concepts, and transactions. (e.g., master customers, patients, payers, HCPs, non-duplicated transactions, and cross-reference tables). The Silver layer brings the data from different sources into an Enterprise view and enables self-service analytics for ad-hoc reporting, advanced analytics, and ML. It serves as a source for all the data personas to create further projects and analyses to answer business problems via enterprise and departmental data projects in the Gold Layer. Speed and agility to ingest and deliver the data in the data lake are prioritized, and a lot of project-specific complex transformations and business rules are applied while loading the data from the silver to the gold layer.

The figure below illustrates some sample data quality checks that can be performed while moving data from the bronze to the silver layer. This involves Uniqueness checks, which checks for this like Null checks, duplicates, unique column names to avoid conflicts, etc. Data Completeness checks involve Referential integrity checks, e.g., Ensuring every patient has at least one condition recorded or one visit recorded, etc. Consistency involves data type checks ensuring that different columns across different tables are standardized to the same data type and format, etc. Finally, Accuracy can involve customer business rules-related checks that can be embedded to ensure that it meets the required quality and constraints. e.g., the Death Date should not be before the Birth Date, the Visit End date should not be smaller than the Start date, the condition date should not be lesser than the birth date, etc. Since this layer has organized data, it enables data analysis for ad-hoc reporting and improves machine learning. It is a vital source for data engineers and scientists to develop further projects and solve different business problems through data projects in the Gold Layer. [9]

The gold layer is also known as the curated business-level tables. This layer contains the most consumption-ready data, and the information in this layer generally contains business-level metrics and KPIs ready to be consumed by data analysts. E.g., condition level Aggregates, cohort analysis, etc.

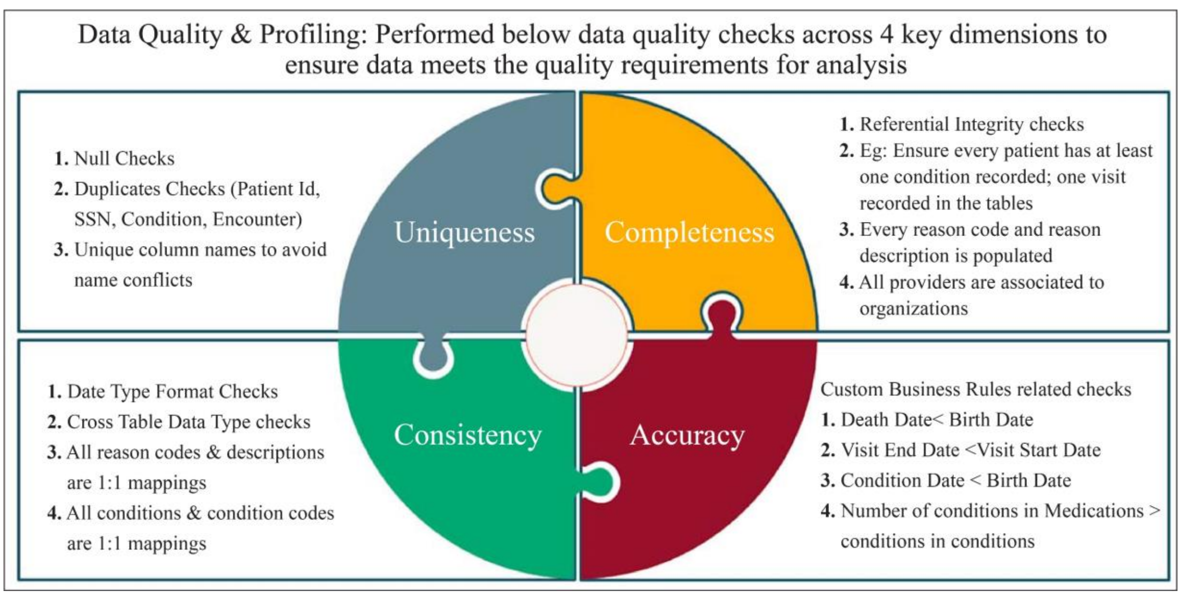


Fig. 17 Data quality and profiling

5. Use Deep Case Dive

5.1. Patient Demographics Dashboard and Insights

One of the main use cases that can be derived using RWD data is the creation of various visualizations and dashboards for analytics. [10] One such dashboard can be the patient demographic EHR visualization dashboard. Patients' demographic dashboards transform data into rich and informative visuals that can help provide valuable insights for data-driven decision-making in life sciences companies. These dashboards can help answer various key business questions around patient demographics, market opportunity, and key opinion leaders and ease decision-making for life sciences companies. [11]

The patient demographic dashboard can showcase various aspects of patient information, such as age distribution, gender, location, race, and ethnicity. It aids in answering questions related to healthcare trends, resource allocation, and targeted marketing strategies. For instance, it helps identify age groups requiring medical attention, pinpoint areas with high patient densities, understand ethnic and racial diversity, and analyze gender distribution. Moreover, patient demographic dashboards assist in patient segmentation by age, location, race, and gender. This segmentation enables healthcare and life science organizations to tailor their services and marketing to specific demographic trends. The data can also support creating culturally or demographically appropriate products and services for diverse communities, regional medical campaign targeting, and identifying potential gaps in accessibility to healthcare.

For instance, Distribution by race and ethnicity can be identified using patient demographic dashboards. To provide culturally competent treatment, it is essential to comprehend the ethnic and racial diversity of the patient group. It may additionally reveal differences in healthcare outcomes and access between various ethnic communities. Moreover, it can help allocate physicians and nurses who speak the local languages in each locality to reduce the communication barrier and provide easy access to healthcare.

Another key area to answer some key business questions is market analysis. It can help answer some key business questions like the most common or highly occurring conditions in the market, how many products/drugs/treatment options are available for a specific condition in each market, and the market share of various companies that are addressing similar conditions. Insights like these can help life science companies understand potential strategic market opportunities for investments and drug development. Moreover, it can also help identify key opinion leaders for a specific condition or drug that can help the commercial field teams target that physician for promotions or their education speaker programs accordingly.

Moreover, such dashboards can be shared to promote collaboration among KOLs, speeding up lifesaving discoveries and partnerships, enabling safe and open sharing of information and cooperation with organizations throughout the healthcare ecosystem and potentially enhancing how care is delivered [12]. In summary, leveraging patient demographic dashboards on Databricks allows for comprehensive insights into patient demographics and healthcare trends, aiding in decision-making, resource allocation, and targeted marketing strategies. The unified platform fosters collaboration among stakeholders, leading to improved healthcare services and advancements in research.

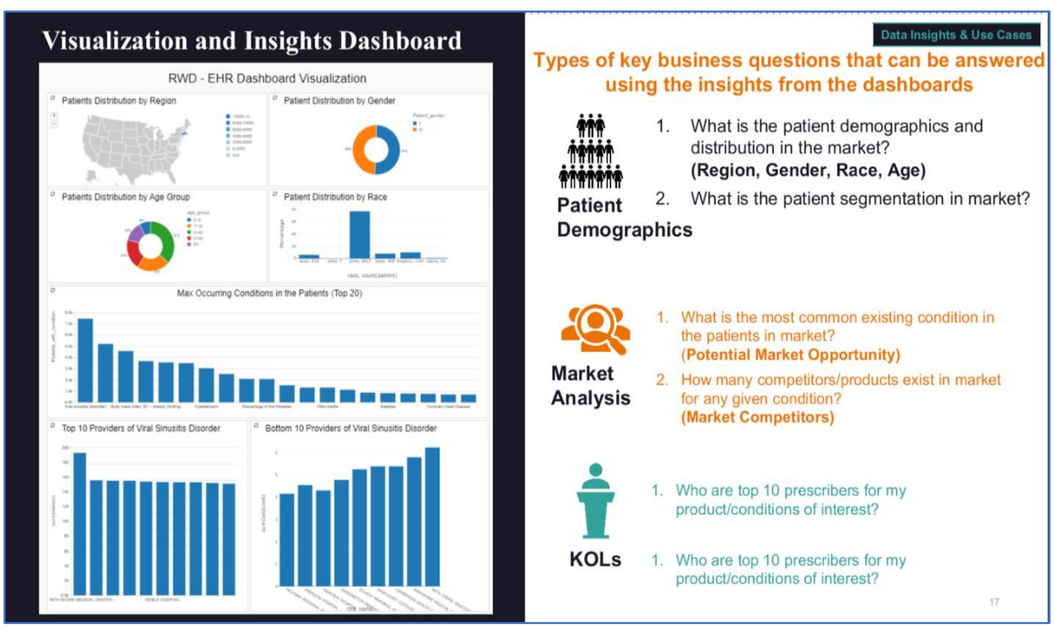


Fig. 18 Patient Demographic dashboard created in databricks

5.2 Treatment Rate, Medication Adherence & Treatment Progression Across Lines of Therapies

Another use case with which RWD data can help us understand the prevalence, treatment rate, medication adherence, and treatment progression across multiple lines of therapies for a given condition. Prevalence relates to the total number of individuals in a population who have a disease or health condition at a specific period of time, usually expressed as a percentage of the population. Treatment rate relates to the number of patients identified with a condition and prescribed medication. Medication adherence relates to the degree to which the patient's behavior corresponds with the agreed recommendations from a health care provider.[13].

This critical viewpoint reinforces remedial results, especially for patients influenced by constant conditions. Tracking medical adherence can reveal factors that promote or discourage patients from adhering to the prescribed medication. Healthcare providers can also track patient outcomes for patients who adhere to medication and those who fail to adhere to the prescribed medicine.

Details on patients' medical adherence can provide insight into strategies for encouraging patients to adhere more to medicine and reveal better methods of encouraging patients to follow the medication given. For instance, one factor influencing medication adherence among patients encompasses knowledge and understanding of the condition, the purpose of the medication, dosage, and possible side effects. [14] Patients with a comprehensive understanding of the medication are likelier to adhere to it.

The best results from medical therapies depend on patient adherence, mainly when several lines of therapy are being used. This is especially true in medical specialities like oncology, where patients frequently receive many lines of therapy; if their disease worsens or initial treatments are unsuccessful, better treatment options can be introduced. [15]

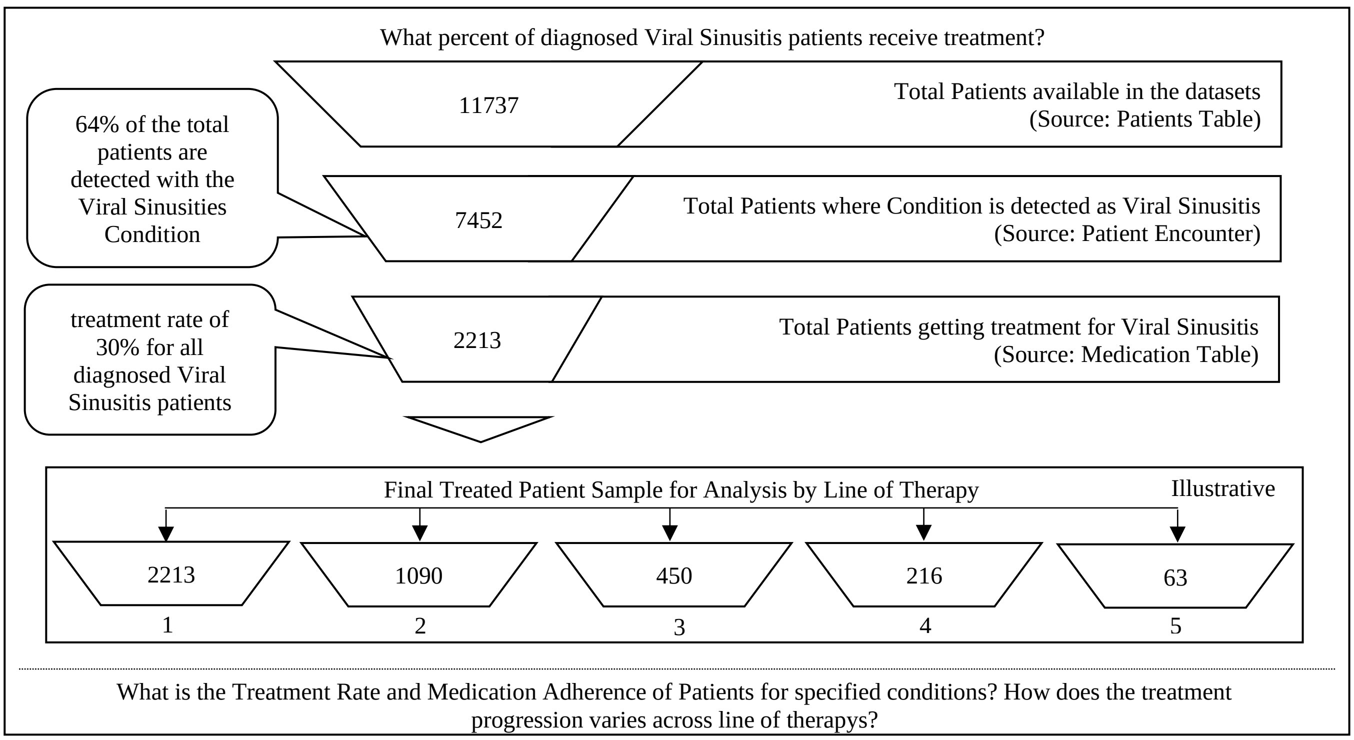


Fig. 19 Image showing sample medication adherence information that can be derived from this data

5.3. Comorbidity Analysis Interactive Dashboards

Another use case for which this real-world data (RWD) can derive insights is for comorbidity analysis. The comorbidity analysis interactive boards offer an interface for analyzing the co-occurrence of multiple health conditions or diseases in a population. This use case is a very common starting point and can help detect at-risk patients. In a clinical setting, we may look at comorbidities as a way to understand the risk of a patient's disease increasing in severity. From a medical coding and financial perspective, looking at comorbid diseases may allow professionals to identify common medical coding issues that impact reimbursement. In pharmaceutical research, looking at comorbid diseases with shared genetic evidence may give us a deeper understanding of the function of a gene. Moreover, these dashboards allow stakeholders like healthcare systems to visualize and research the relationship between diseases and their most common combinations. [9]. As a result, healthcare professionals and policymakers can acquire actionable insights into the prevalence, interactions, and association between different conditions. They can take some proactive actions to prevent future conditions using precision prevention, which is focused on using data to identify patient populations at risk of developing a disease and then providing interventions that reduce disease risk.

An intervention might include a digital app remotely monitoring at-risk patients, providing lifestyle and treatment recommendations, increasing disease status monitoring, or offering supplemental preventative. Thereby helping improve patients' health.

5.4. Patient Risk Prediction

RWD data can also be used for various data science-related use cases like predictive analytics. One such example is creating a patient risk scoring model using various statistical, data science, and machine learning models that can help predict the risk of specific condition occurrence in a patient.

Fig. 20 Interactive comorbidity analysis

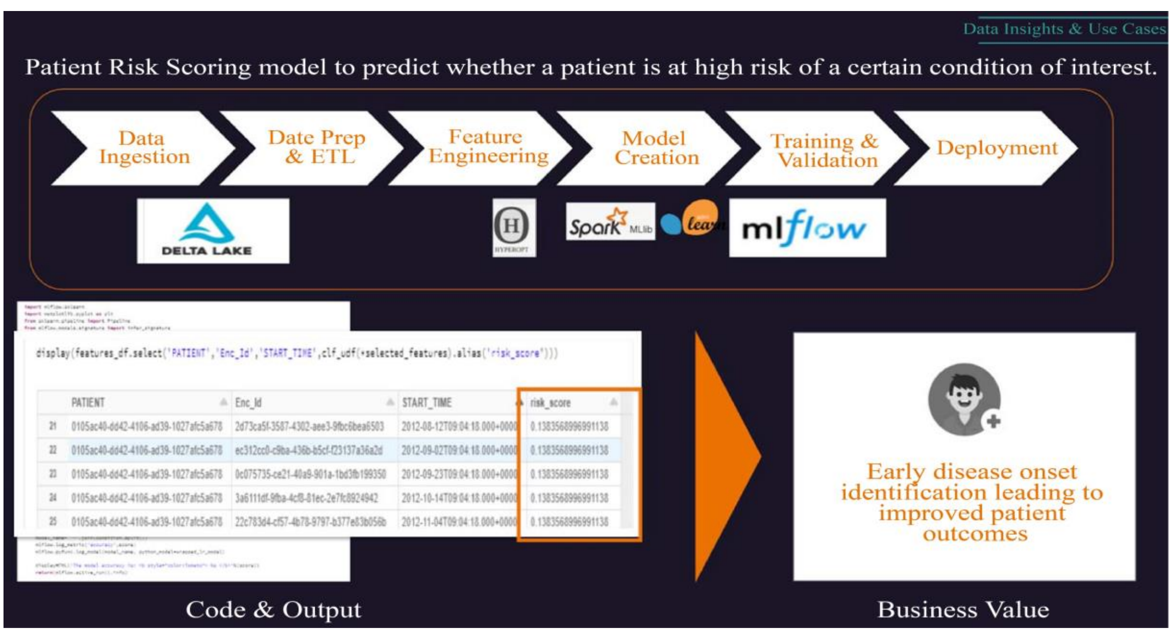


Fig. 21 Patient risk prediction

As shown in the above image (*Figure 21)*, the process encompasses analyzing elements such as lifestyle, genetic predisposition, demographic information, and environmental factors to estimate the patients' risk profile.

Healthcare systems can employ predictive models such as algorithms and statistical approaches in determining the patient's risk factors for experiencing a particular health outcome.[16] These techniques can be anchored on machine learning models, regression analysis, and other relevant mathematical approaches. These techniques use historical data to identify patterns and discover relationships between outcomes and risk factors, generating risk predictions for new patients.

5.5. Patient Condition Progression

A comprehensive understanding of a patient's disease progression can guide healthcare providers to gain actionable insights into the behavioral patterns of the ailment.[14] As a result, healthcare professionals can quickly identify similar patterns in other patients, causing better disease diagnosis, treatment, and improvement in disease association.

The healthcare providers also have the chance to identify specific conditions that cause the disease or lead to the further deterioration of the patient's health. These conditions include lifestyle, genetic predisposition, environmental factors, and comorbidities. Identifying such needs offers more precise and practical guidance in treating the patient and deciding further treatment interventions for the disease.

5.6. Prescriber and Patient Lead Identification

Healthcare providers can create such dashboards to help track patient treatment progression. The dashboards can also help track and identify field sales to identify specific field patient and physician leads at a given provider, targeting them accordingly with appropriate educational and promotional content. Thus, lead identification can help providers understand patients' preferences and behavior. This data assists in the identification of potential leads for potential patients for services and products. Such information guides the sales team to target the identified patients appropriately.

6. Data Expansion to Address Additional Comprehensive Use Cases

Integrating EHR data with additional data sources can assist life science companies, policymakers, healthcare professionals, and other related stakeholders understand the prevalence of certain diseases, patient medical adherence, and patient outcomes to propel more effective treatment or medical intervention methods. Below, Figure 24 highlights a few examples of additional data sources, use cases and outcomes.

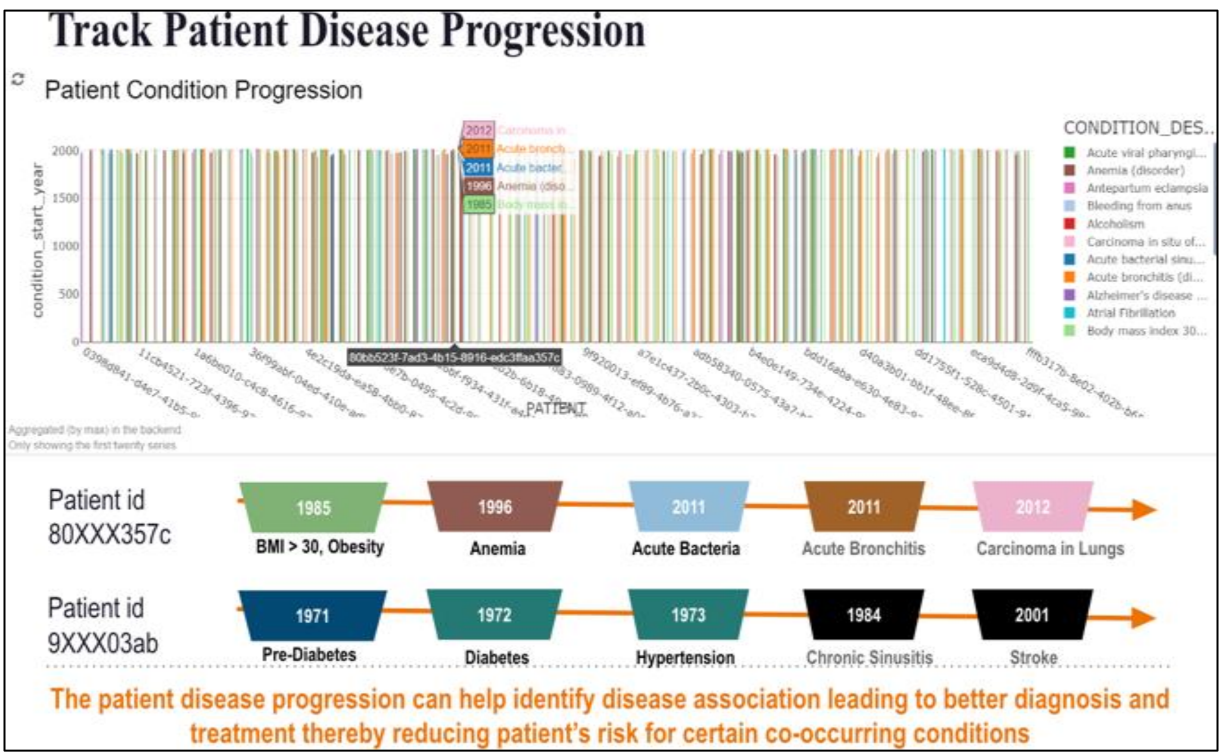


Fig. 22 Tracking patient disease progression

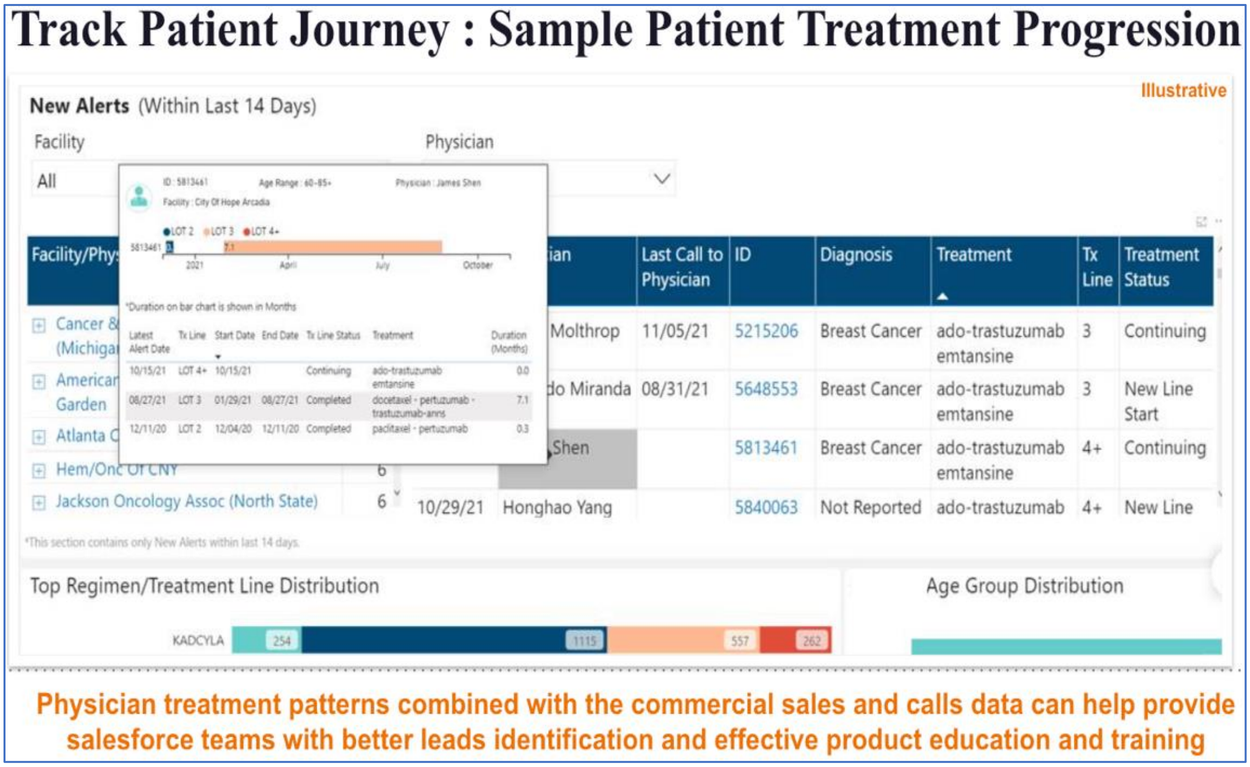


Fig. 23 Sample patient treatment progression

For instance, incorporating EHR Clinical data with demographic data can help health professionals gain insight into the relationship between patient demographics and specific health outcomes or conditions. Secondly, genomics and precision data can be used in single-cell sequencing to foster innovation for a better customer experience.

Single-cell sequencing models enable the analysis of individual cells, providing detailed insights into cellular heterogeneity, genetic variations, and gene expression profiles.[17] This information provokes a comprehensive understanding of disease mechanisms, identifies biomarkers, and creates personalized patient treatment.

Data recorded in claims can help conduct efficacy and competitive analytics, fostering improved patient outcomes and enhanced Salesforce effectiveness. Claims data can improve treatment outcomes by analyzing healthcare utilization patterns, medical adherence, and healthcare costs associated with specific treatment interventions.

As a result, the effectiveness of different interventions can be determined, and the most effective practices can be identified. This data can also enable easy comparison of other treatment options and assess the efficacy of this treatment intervention in the real world.[17] In competitive analytics, claims data can propel market insights by analyzing utilization patterns, prescribing practices, and treatment outcomes related to different interventions. This data helps to identify market trends, areas of improvement, and competitive positioning.

Medical deep learning with the help of artificial intelligence has been shown to help support medical diagnosis and graphically demanding workloads. Medical deep learning models can evaluate complicated medical images and data to help diagnose and make decisions using machine learning techniques and deep neural networks.

Additionally, analytics for IoT (Internet of Things) devices use sensor data from linked devices to enhance healthcare delivery and revolutionize care quality. Finding deviations from predicted patterns, enabling proactive interventions, and improving care delivery are all possible through anomaly detection.

Lastly, in social analytics, a more profound comprehension of the social and environmental aspects that affect patient health can be achieved by including SDOH data, such as housing conditions, access to transit, food security, or community features.[17] This information can pinpoint interventions, detect socioeconomic discrepancies, and enhance health equity.

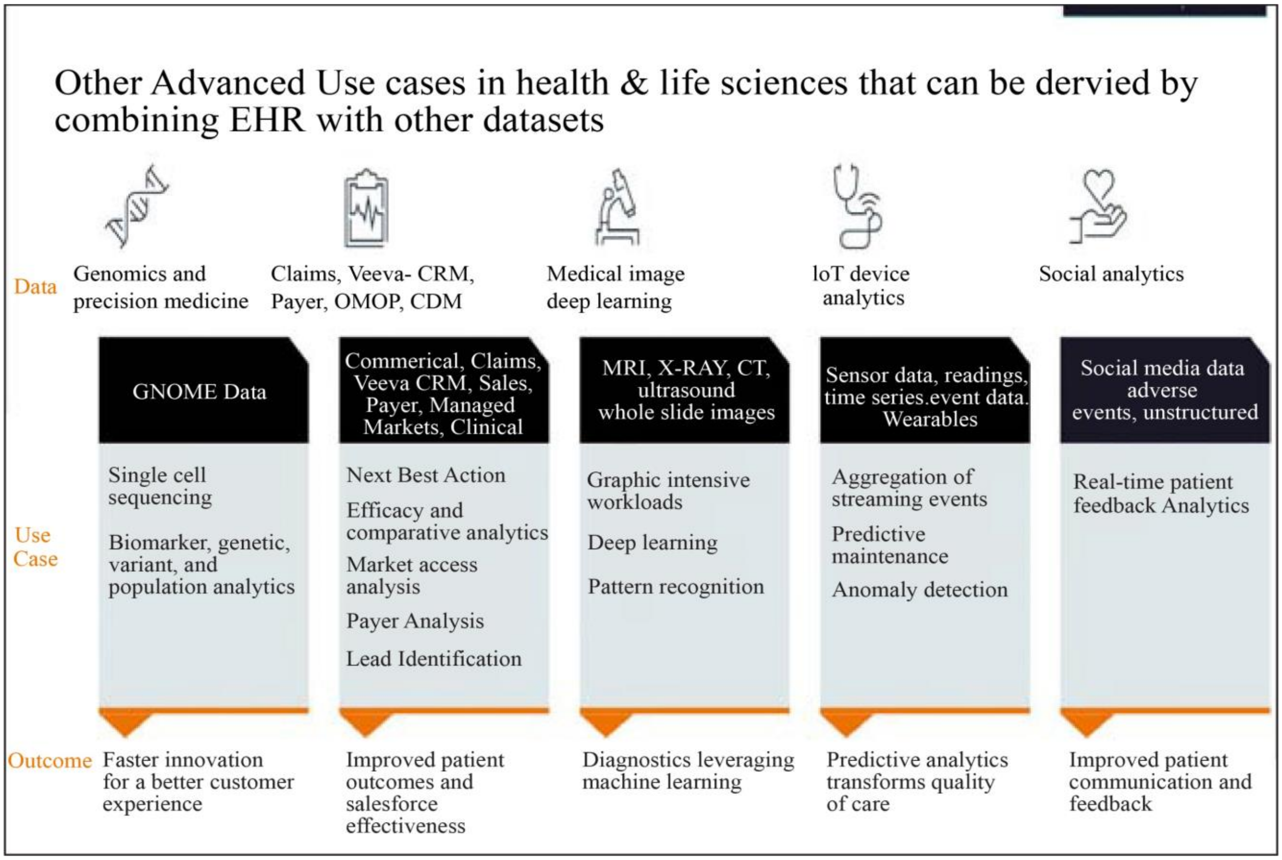


Fig. 24 Combining EHR with another data

7. Conclusion

In conclusion, this article sheds light on the enormous opportunity of Real-World data combined with AI and analytics technologies to change healthcare systems and the Research and Development of drugs. The research highlighted how the combination of RWD and advanced analytics platforms like Databricks could result in major gains in care for patients, productivity, and decision-making procedures throughout the healthcare and life science industry by examining several use cases.

The addition of practical options from Databricks Lakehouse has significantly underlined the viability and applicability of these use cases, laying the groundwork for their deployment in real-world circumstances.

Also, Databricks Lakehouse approaches and innovative artificial intelligence may successfully solve existing gaps and inconsistencies in healthcare systems. By utilizing the modern lakehouse platform, healthcare and life science organizations and policymakers may make better choices, optimize resource allocation, improve the patient experience, discover better and more advanced drugs, and consolidate operations, resulting in improved comprehensive healthcare delivery and drug research and development.

Therefore, adopting these creative ideas remains increasingly important as the Healthcare and life science business evolves and faces new problems. Efficient use of real-world data and modern technology can offer dramatic benefits, catapulting healthcare systems towards a future of better efficacy, cost-efficiency, and patient-oriented care. Thus, this study strongly argues for integrating RWD, machine learning, Databricks Lakehouse solutions, AI, and data analytics technologies into healthcare and drug research and development. By responsible adoption and adaptation of the above innovations, healthcare and life science players may prepare the path for a more promising and sustainable future in which the capability to fill gaps and errors in medical systems can be realized to the advantage of everyone pursuing high-quality medical care.

Acknowledgement

Author - Mayur Katariya extends his sincere gratitude to Ashish Upadhyay, Director of Field Engineering at Databricks, who asked the author to delve into a comprehensive analysis before joining Databricks that greatly influenced the creation of this article. The author is also indebted to ZS and its leaders for the invaluable professional experience in the Healthcare and Life Science domain, which served as a profound source of inspiration for this article.

References

[1] How Real-World Evidence Transforms the Entire Healthcare Ecosystem, DXC Technology, 2023. [Online]. Available: https://dxc.com/us/en/insights/perspectives/paper/how-real-world-evidence-transforms-the-entire-healthcare-ecosystem

[2] Blair Bean Robertson, FDA Advances Program for Real-World Evidence, The Regulatory Review, A Publication of the Penn Program on Regulation, 2023. [Online]. Available: https://www.theregreview.org/2023/02/27/bean-robertson-fda-advances-program-for-real-world-evidence/

[3] Aldren Gonzales, Guruprabha Guruswamy, and Scott R. Smith "Synthetic Data in Health Care: A Narrative Review," *PLOS Digital Health*, vol. 2, no. 1, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[4] Jake Gower, Christopher Lundeberg, Databricks and Technology Partners: Personalized Medicine with a Tailored Approach, 2023. [Online]. Available: https://www.databricks.com/blog/databricks-and-technology-partners-personalized-medicine-tailored-approach#:~:text=Integrating%20Cloud%2DAdjacent%20Technologies&text=Therefore%2C%20medical%20research%20institutions%20can,the%20best%20of%20both%20worlds./

[5] Synthea/README.md Synthea™ Patient Generator, synthetichealth/synthea, 2023. [Online]. Available: https://github.com/synthetichealth/synthea/blob/master/README.md

[6] Andre Goncalves et al., "Generation and Evaluation of Synthetic Patient Data," *BMC Medical Research Methodology*, vol. 20, no. 1, pp.1-40, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[7] R.N. Matt Vera Bsn, Nursing Care Plans (NCP), Ultimate Guide and List, 2023. [Online]. Available: https://nurseslabs.com/nursing-care-plans/

[8] What is a Medallion Architecture?, Databricks, 2023. [Online]. Available:https://www.databricks.com/glossary/medallion-architecture

[9] Dennis M. J. van de Sande et al., "A Review of Machine Learning Applications for the Proton MR Spectroscopy Workflow," *Magnetic Resonance in Medicine*, vol. 90, no. 4, pp. 1253-1270, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[10] Robby Nieuwlaat et al., "Interventions for Enhancing Medication Adherence," *Cochrane Database of Systematic Reviews*, vol.11, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[11] Kevin Lee, and Genpact, Patient's Journey Using Real World Data and its Advanced Analytics, 2023. [Online]. Available: https://www.pharmasug.org/proceedings/2023/RW/PharmaSUG-2023-RW-113.pdf

[12] Bill Zanine, Michael Sanky, and Adam Crown, The future of Healthcare Relies on Data Collaboration: How IQVIA and the Databricks Lakehouse Enable Better Outcomes, 2023. [Online]. Available: https://www.databricks.com/blog/future-healthcare-relies-data-collaboration-how-iqvia-and-databricks-lakehouse-enable-better.

[13] Wai Yin Lam, and Paula Fresco, "Medication Adherence Measures: An Overview," *BioMed Research International*, vol. 2015, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[14] Beena Jimmy, and Jimmy Jose, "Patient Medication Adherence: Measures in Daily Practice," *Oman Medical Journal*, vol. 26, no. 3, pp. 155-159, 2011. [CrossRef] [Google Scholar] [Publisher Link]

[15] Databricks for the Life Sciences Industry, Databricks, 2023. [Online]. Available: https://www.databricks.com/solutions/industries/life-sciences-industry-solutions

[16] Erin McNemar, What are the Benefits of Predictive Analytics in Healthcare?, HealthITAnalytics, 2022.[Online]. Available: https://healthitanalytics.com/news/what-are-the-benefits-of-predictive-analytics-in-healthcare

[17] R. Anjit raja, B. Nagarajan, and R. Dhanappriya, "Byzantine Neurobiological Phenomenon Analysis and Factors Prediction for Social Network based Adult's Suicides and Cyber Dismay by Hypercritical Machine Learning Techniques," *SSRG International Journal of Computer Science and Engineering*, vol. 4, no. 4, pp. 24-29, 2017. [CrossRef] [Publisher Link]