Is your real-world data research ready?

The purpose of real-world data (RWD) in life sciences is fairly straightforward. Fundamentally, it allows researchers and business leaders to better understand how medications, therapies and other interventions are performing. Are they safe? Are they effective? And if so, do they perform better than the alternatives?

Answering these questions is incredibly important to life sciences organizations and all health care stakeholders, from patients and health care providers to health plans, employers and government.

Before investing in this data, it's important to understand what goes on behind the scenes. How is it collected, where does it come from and what does it include and exclude? You want to be sure it accurately reflects the patient experience so the research and business decisions you make from it are as good as they can be. This is particularly true for electronic health record (EHR) data – still relatively new in the research field – which is the focus of this paper.

Below are some factors to consider as you make these evaluations:

Is it easy to use?

Is it reliable?

Is it relevant to your needs?

Is it longitudinal?

Optum

Is it easy to use?

Data extracted from EHRs tends to be very messy in its raw form. A great deal of time and effort is needed to standardize this data so it can be used for research and business intelligence purposes. When the data first arrives, it comes from multiple sources and in varying formats.

At Optum, we collect and aggregate data from 17 different EHR systems. It comes from more than 700 hospitals and 7,000 clinics. It encompasses 1,000 data sources, 3,000 medications and more than 800,000 procedure codes.

Workflows vary among providers and, because they are supported by myriad different technology vendors and solutions, the reality is that the way the data is collected, defined and stored in provider EHRs is almost always unique in some way. This high variability makes analyzing the data straight from the source extremely difficult.

Consider coding for a colonoscopy. There isn't just one uniform code representing this procedure within EMR systems. We found 80 unique local codes mapped to this single concept across multiple EMRs. They included almost 40 custom codes and more than 40 standard ICD-9/ICD-10, SNOMED, HCPCS and CPT® codes.

To overcome this challenge, Optum employs a multi-functional team of technologists and clinicians to understand and parse this data. The team includes experts in data extraction, semantic normalization and data science, in addition to nurse practitioners and physicians. Optum deploys this team in a systematic approach to extract data and apply a series of normalizations, thereby translating data to a consistent and analyzable form while maintaining its original meaning and credibility.

This results in data that is easier to sort, search, explore and analyze. It also allows for easier reporting and for advanced data visualization tools to be utilized.

Is it reliable?

The provenance of data is incredibly important for researchers and anyone making decisions based on it. We need to be able to confidently and reliably track each data point back to its source.

It's also important to know the data has not been incorrectly manipulated. For all the work done to standardize and normalize, it's crucial to keep the data as close to historical truth as possible. There must be processes in place to ensure accuracy and quality in the data.

Optum uses a metadata-driven approach to data warehousing that retains the local values. These mappings of local codes to standardized concepts are reviewed and updated every month. The validation process includes refining mapping criteria over time (as new information arrives) and verifying ensuing changes in counts through the lens of life sciences users. For example, did new mappings affect how carbon dioxide (CO_2) levels are displayed? And if so, what were the changes? Are they correct from a clinical standpoint, that is, validated by clinicians?

Optum applies a series of data standardization routines once the raw data is received. **These routines ensure duplicate data are removed, and fields and values are appropriately mapped.** They are repeated across approximately 150 data domains, consisting of approximately 4,000 unique concepts, and to date have been applied to approximately 2.3 million native codes found in the data.

Is it relevant to your needs?

The data needs to capture clinically relevant characteristics of disease, treatment and outcomes for it to be valuable and usable for research and business decision-making.

At its most fundamental level, RWD must enable researchers to obtain an objective answer to a specific question. In order to do this effectively, the data must contain as much relevant information as possible to answer the question and be organized in such a way that information can be readily extracted from it. In other words, it must be usable.

This can be challenging for a number of reasons. For the most part, the data in EHRs was not originally collected for the purposes of research and analytics, but rather to support clinical and administrative workflows for a health care delivery (provider) network. It is generated for billing, record keeping, insurance reimbursement, treatment outcomes, decision support, medical documentation and much more.

Moreover, the recorded information does not always fit in neat, discrete fields. But as long as those working with the data keep the context in mind, there is value in recorded information even with all the variance. It's crucial that the data specialists working with this data understand this as they do their jobs.

The key to addressing relevancy during the normalization process is working closely with medication/lab specialists, physicians and nurse practitioners to ensure the credibility and accuracy of all data mapping efforts. In addition, we involve our own life sciences experts from epidemiology, health economics and outcomes research (HEOR), commercial analytics and other product level experts. This helps us ensure that the data is not only credible and accurate but also relevant to researchers and business intelligence professionals in the industry.

The key to addressing relevancy during the normalization process is working closely with medication/lab specialists, physicians and nurse practitioners to **ensure the credibility and accuracy** of all data mapping efforts.

Is it longitudinal?

The greatest value of real-world data is the ability to follow and understand patient journeys. For it to be truly research ready, it must be longitudinal to allow users to understand how patients navigate the health care system. We need to understand more than what occurred just during treatment – we must know what happened before and after treatment as well.

Optum is able to provide this longitudinal view by drawing on expertise developed over decades of working with health information. We bring data sources together into a single, patient-linked, standardized view of the population. This allows us to build a comprehensive view of both an individual patient and the population from birth to death, without sacrificing quality and variable details. Extracting data from multiple EHRs delivers a better perspective of the entire patient journey. It's not just one chapter of the story. For example, with less longitudinal data, you may have information related to oncology treatment but not from the primary care physician or other specialists also involved in a patient's care.

In addition, we also apply natural language processing (NLP) to incorporate data from free-text fields (like physician notes) to offer an even richer view of the patient experience.

What is natural language processing?

Provider notes offer incredibly valuable information about what's happening between the patient and provider at the point of care. We use natural language processing (NLP) to make this data available in a structured way. This results in more usable data on symptoms, family history, biomarkers, medications and physician rationale, which might never be recorded in the structured fields.

We surface this critical and detailed information by leveraging our proprietary NLP system that extracts discrete information from the free-text provider notes within EHR data.

For example, if the goal is to identify patients with prostate cancer, the Optum NLP system identifies different semantic contexts and appropriately extracts the desired contexts into a structured format. The concepts are then able to be easily searched by users. Some examples of the contexts that occur within the notes are shown in Table 1.

Table 1: Sample of contexts for cancer statements

Sample text	Concept
Patient has stage II prostate cancer	Patient positive for prostate cancer
Negative for prostate cancer	Patient negative for prostate cancer
If prostate cancer is found, patient may require additional imaging	Hypothetical prostate cancer situation
Might be prostate cancer	Hedged prostate cancer statement
Prostate cancer is a common cancer among males	Prostate cancer not relevant to patient

The Optum NLP system leverages **best practices** in data science and automation.

Within the EHR data, there are millions of patients who have at least one solid tumor diagnosis code. Manually reviewing hundreds of millions of documents and manually extracting clinical data for research is not a scalable approach. Our NLP system offers an automated solution for providing insights from a large collection of medical notes that continues to grow each day.

The Optum NLP system leverages best practices in data science and automation. Our sophisticated system goes beyond term-matching and rules-based approaches by incorporating machine learning and deep learning in order to ensure the correct identification of the desired oncology context. The advantage of leveraging supervised machine learning models is the ability to accurately identify the appropriate contexts in an automated fashion over highly variable text. Our supervised machine-learning models are trained to identify broader patterns that are not explicitly and manually created by a human as a rule, but instead, the machine learns from a sample of labeled data that will then enable the system to generalize the relevant contexts.

The importance of privacy

A few words about this most important of topics. Often, we use data to support research and, where possible and permitted, we will look to de-identify PHI in compliance with HIPAA's de-identification requirements (HIPAA 45 C.F.R. 164.514(a)-(c)).

De-identified datasets can contain different information depending on their intended use of the de-identified data. Often different combinations of indirect identifiers (for example, race, geography, age) will be analyzed for inclusion depending on the research needs.

Additionally, the EHR data includes elements such as lab test results or drug names/doses that are entered into the EMRs as free-text fields. Some of the unstructured EMR information may contain identifiable attributes like a patient's name or phone number. These fields undergo additional scrutiny using both automated and manual methods so that we can identify the data and then remove or replace it.

Each release of de-identified data is assessed for compliance with the de-identification requirements and data quality standards. Once released, a de-identified dataset can be modified by summarizing the existing data elements, but additional data linking is not permitted without a review to ensure any data linked to the de-identified data maintains compliance with the HIPAA de-identification requirements.

Summary

Optum takes its responsibility to deliver research-ready data very seriously. We draw on data expertise developed over decades of working with health information, bringing data sources together into a longitudinal, standard view of the population. This allows us to build a comprehensive view of both an individual patient and the population from birth to death, without sacrificing quality and variable details. We extract data from multiple EHRs, apply natural language processing to ingest data from free-text fields (like physician notes), and follow an industry-leading data curation methodology.

The backbone of research is evidence, and repetitions or patterns in data make for good evidence. Systems that normalize data can help life sciences organizations address myriad health care challenges, like cost of care and identifying unmet medical needs. When pharmaceutical companies understand prescribing patterns, they can better determine where their new drug will fit in the treatment paradigm. Designing clinical trials with an EHR-based control cohort leads to trials that are more likely to be completed on time and on budget and gets life-changing treatments in the hands of patients sooner. Real-world data leads to the discoveries that can help us develop new interventions, empower providers and improve the lives of patients.

At Optum, we offer both depth and breadth of understanding from the health care system itself because we are an integral part of it. Optum is a part of UnitedHealth Group®— a Fortune 7 company.

Across Optum, we serve many constituents — government, health plans, health care providers, employers, life sciences organizations and consumers. These relationships give us an unmatched perspective, allowing us to forge connections, recommend best practices and increase efficiency in a way that is hard to match.

To learn how real-world data can help measure brand performance, please contact us.

optum.com/life-sciences

Optum

optum.com